RESEARCH IN COMPUTING SCIENCE

ISSN: 1870-4069

Advances in Computer Science and Artificial Intelligence

Alexander Gelbukh
Michel Adiba
(Eds.)

Vol. 39

RCS
Research in Computing Science

Advances in Computer Science and Artificial Intelligence

Research in Computing Science

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 39** Octubre, 2008. 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.

Para la portada se usó la imagen cortesía del Prof. Akiyoshi Kitaoka, www.ritsumei.ac.jp/~akitaoka. 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*

Volume 39
Volumen 39

Advances in Computer Science and Artificial Intelligence

Volume Editors:
Editores del Volumen

Alexander Gelbukh
Michel Adiba

Instituto Politécnico Nacional
Centro de Investigación en Computación
México 2008

ISSN: 1870-4069

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

Indexed in LATINDEX and Periodica
Indexada en LATINDEX y Periodica

Printing: 500
Tiraje: 500

Printed in México
Impreso en México

Preface

With this special issue we celebrate the 50th anniversary of computer science in Mexico. The volume is structured into two parts and seven sections:

Computer Science

- Software Technology and Human-Computer Interfaces
- Workflow and Collaboration
- Networking

Artificial Intelligence

- Logic and Multi-Agent Systems
- Natural Language Processing and Information Retrieval
- Machine Learning and Data Mining
- Neural Networks, Image Processing, and Scheduling

This issue was prepared in collaboration with the Mexican Society for Computer Science (SMCC), Autonomous University of Baja California (UABC) and Regional Government of Baja California, Microsoft, Center for Computing Research (CIC) of the National Polytechnic Institute (IPN), French-Mexican Laboratory of Informatics and Automatic Control (LAFMIA-CNRS), National Council for Science and Technology (CONACyT, Mexico), National Council for Scientific Research (CNRS, France), Université Joseph Fourier (France), The Americas University (UDLA, Puebla, Mexico), National University (Colombia), Social Networks Research Center (SoNet RC) of the University of Central Europe (UCE) in Skalica (Slovakia), and Grenoble Informatics Laboratory (France).

We want to cordially thank all people involved in the preparation of this special issue. First of all our very special thanks go to the SMCC President Genoveva Vargas-Solar for her constant help and encouragement. We thank the SMCC Vice-President María Auxilio Osorio Lama, Marcela D. Rodríguez and Gabriel López Morteo of UABC, Antonio García Macías of CICESE, Alfredo Sánchez of UDLAP, and Oscar M. Rodríguez Elias of Universidad de Sonora.

We are grateful to Javier A. Espinosa-Oviedo, Hilda Massic-Fernández, and Sulema Torres for their help in the preparation of this volume. Finally, we appreciate the help of René Cruz, Cecilia Curlango Rosas, Ángel G. Andrade, Gloria E. Chávez, Aglay Saldaña Pacheco, Marlenne Angulo Bernal, Jorge Ibarra Esquer, Carmen Andrade Peralta, Ana Serrano, Pablo Navarro Álvarez, Laura Martínez Castillo, Omar Aguilar Villavicencio, Brenda L. Flores Ríos, Larisa Burtseva, Josefina Mariscal, María Luisa González, Linda E. Arredondo, and Juan de Dios Ocampo.

Alexander Gelbukh
Michel Adiba

Mexico City — Grenoble,
October 2008

Table of Contents
Índice

Page/Pág.

Computer Science

Software Technology and Human-Computer Interfaces

Workflow and Collaboration

Networking

Artificial Intelligence

Logic and Multi-Agent Systems

Natural Language Processing and Information Retrieval

Machine Learning and Data Mining

Neural Networks, Image Processing, and Scheduling

Computer Science

Software Technology and Human-Computer Interfaces

Case Study Evaluations for a Function Point Counting Improvement for Object-Oriented Projects

José Antonio Pow-Sang[1,2], Arturo Nakasone[1], Ricardo Imbert[2], and Ana María Moreno[2]

[1] Departamento de Ingeniería, Pontificia Universidad Católica del Perú, Peru
{japowsang, arturo.nakasone}@pucp.edu.pe
[2] Facultad de Informática, Universidad Politécnica de Madrid, Spain
{rimbert, ammoreno}@fi.upm.es

Abstract. Since the introduction of object-oriented (OO) development in industrial practice, many Function Point (FP) technique adaptations have been introduced to improve software size estimation in these kinds of projects. Current research work only deal with OO modifications to the previous version of the FP Counting Practices Manual (4.1). In this paper, we propose the use of the composition relationship analysis in classes to improve the rules included in FP Counting Practices Manual 4.2.1 for Internal Logic Files (ILF) and External Interface Files (EIF) identification. We also show the results obtained by applying our proposal in six case studies performed by practitioners and comparing against the results we obtained with undergraduate students. These results have proved to be at least equal in accuracy and consistency to the original FPA technique.

Keywords: Function Points, Object Oriented, UML, Conceptual Model.

1 Introduction

Function Point (FP) [11] [12] is a software measurement technique created by Allan Albrecht for IBM [3], and has gradually become a sounder alternative to other popular size metrics methods, such as source lines of code (SLOC), making it one of the most widely used techniques.

With the diffusion of the Unified Modeling Language [19], promoted by the Object Management Group, many object-oriented approaches to calculate function points have been proposed. Unfortunately, they do not consider some important specifications included in UML, such as the composition relationship between classes. For this reason, we propose in this paper an approach to calculate Logic Files (ILF and EIF) from an analysis class diagram that makes use of composition relationships. We have also tested our approach against the standard Function Points Counting Practices Manual, version 4.2.1 [12] proposed by the International Function Points User Group (IFPUG) obtaining interesting and promising results.

The rest of the paper is organized as follows: Section 2 describes the related work in the FP measurement technique area, Section 3 details our proposed rules to identify

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 5-18

Received 02/06/08
Accepted 31/07/08
Final version 15/09/08

logical files. Section 4 presents the background scenario for the empirical study; Section 5 shows the obtained results for each case study; Section 6 discusses those results. Finally, a summary and our plans for future research will conclude our paper.

2 Related Work

In order to cope with object-oriented software measurement, several methods to calculate FP are being promoted and used. These methods reformulate the IFPUG rules in terms of OO concepts to facilitate the function points counting process. The final result of the count using these kinds of techniques is similar to what is obtained by directly applying IFPUG Function Point Analysis (FPA). Fetcke [9] defined rules for mapping the OO-Jacobson method [15] to concepts from the IFPUG Counting Practices Manual 4.0 and the results obtained from three case studies have confirmed that these rules can be applied in a consistent way. Uemura et Al. [20] proposed FPA measurement rules for design specifications based on UML (Unified Modeling Language), developing a FP measurement tool. Cantone et Al. [6] and Caldiera et Al. [5] defined rules to map OO concepts to FPA and performed pilot studies to demonstrate the feasibility of their approaches. Finally, Abrahao et al. [1] [2] present a FP-based method called OOmFP and its evaluation through an empirical study.

All of the proposals described above define rules of mapping OO to concepts from older versions of IFPUG Counting Practices Manual (CPM). Although Abrahao considers composition relationship, some mapping rules to calculate Logic Files (LF) are not in accordance with the last IFPUG CPM (last version is 4.2.1).

Moreover, the majority of the proposals presented above calculate FP-Logical Files (LF) from a class diagram, but they only consider the aggregation relationship and not the composition relationship. UML [19] defines composite aggregation or composition as a stronger form of aggregation, which requires that a part instance be included in at most one composite at a time and that the composite object has sole responsibility for the disposition of its parts. In summary, composition presents a stronger dependence relationship than aggregation. Our approach includes both aggregation and composition relationship in order to identify files and its record element types (RET).

3 Rules to Identify Logical Files

The input for our proposed rules is the analysis class diagram included in the Jaaksi's method [14] or the domain model mentioned by Larman [17]. In this model transformation, we are considering the following relationships among classes: association, aggregation and composition. Generalization and association class are not included. Neither do we take into account the difference between ILF and EIF. Our rules only deal with the identification of LF and its number of RETs.

- **Rule 1**. Classes that are connected by a composition relation can be mapped together to one LF with 2 RETs. For example, in Fig. 1 you must consider one LF and 2 RETs.

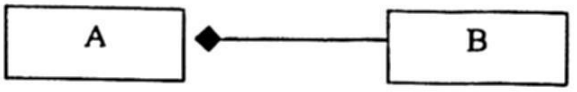


Fig. 1. Example of a composition relation

- **Rule 2.** If there are three classes A, B, and C and two of them (A and B) are connected by a composition relation as shown in Fig. 2, they must be mapped together to one LF for the composition relation with 2 RETs and another LF for class C.

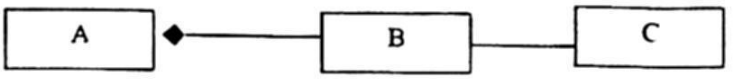


Fig. 2. Example of a composition and association relations

- **Rule 3.** If there are two classes, A and B, which are connected through an association or aggregation relation, and neither of them is connected by a composition one to a third class, one must follow the indications shown in Table 1. The table is an adaptation for OO from IFPUG CPM 4.2.1 [12].

Table 1. Rules to identify LF from classes without composition relationships

Multiplicity A	Multiplicity B	When this condition exists	Then Count as LFs with RETs and DETs as follows:
0..*	0..*	A and B are independent	2 LFs
0..1	0..*	A and B are independent	2 LFs
1	1..*	If B is independent of A	2 LFs
		If B is dependent on A	1 LFs, 2 RETs
1	0..*	If B is independent of A	2 LFs
		If B is dependent on A	1 LFs, 2 RETs
0..1	1..*	If A is independent of B	2 LFs
		If A is dependent on B	1 LFs, 2 RETs
0..1	0..*	A and B are independent	2 LFs
1	1	A and B are dependent	1LF, 1RET
0..1	0..1	A and B are independent	2 LFs
1..*	1..*	If B is independent of A	2 LFs
		If B is dependent on A	1 LFs, 2 RETs
1..*	0..*	If B is independent of A	2 LFs
		If B is dependent on A	1 LFs, 2 RETs

4 Experimental Design

For our case study scenario, we have considered the experimental software engineering suggestions made by Juristo & Moreno [16]. Our experiment is similar to one presented by Abrahao et al. [1] [2], and its goal is to empirically corroborate which method provides the best functional size assessment to identify logical files.

Using the Goal/Question/Metric (GQM) template for goal-oriented software measurement [4] we defined this experiment using the following parameters:

- **Analyze**: LF measurement
- **For the purpose of**: Evaluating our approach against the one proposed by the IFPUG CPM 4.2.1
- **With respect to:** Their accuracy
- **From the point of view of:** The researcher
- **In the context of:** undergraduate and graduate students.

The formulated research question was: Does our approach produce accurate measurements of LF at least equal to those found in the IFPUG FPA?

4.1 Variables Selection

Our independent variable was the method used by subjects to size a case study, and our dependent variable was the accuracy: the agreement between the measurement results and the true value.

To obtain a "true value" for comparison, we took similar case studies included in the IFPUG CPM 4.2.1.

4.2 Students Who Participated in the Experiment

We selected a within-subject design experiment; in other words, the students had to use both our method and the IFPUG CPM 4.2.1 method to determine LF for each case study. The subjects were randomly assigned to either one of two groups using the counterbalancing procedure with equal number of participants in each group. The methods were applied in a reverse order.

- Group 1: Our approach first and then the IFPUG CPM 4.2.1 method.
- Group 2: The IFPUG CPM 4.2.1 first and then our approach.

The undergraduate students who participated in the experiment were fourth year students of the Informatics Program at Pontificia Universidad Católica del Perú (PUCP) that were enrolled in the Spring '06 Software Engineering course. Table II presents a summary of the undergraduate students' knowledge and experience at the beginning of the experiment. The majority of the courses in the Informatics program at PUCP focus on software projects as applications of theoretical concepts, but they do not demand the utilization of estimation and planning techniques during their development.

The practitioners were students of the Postgraduate Diploma in Software Engineering at PUCP in 2006. Students had at least two years of experience in software projects and, although they used OO development tools at work, most of them were not used to applying OO analysis and design techniques. For this reason, these students had to take an OO analysis and design course previous to the elaboration of the experiment.

4.3 Materials and Case Studies

The materials used in the experiment are:

- Description of Case studies.
- Form to fill in the number of LF and its RETs for each case study.
- A questionnaire to know student's opinion about which technique was easier to apply.
- A summary of Coad's patterns [8]. We explained these patterns to the undergraduate and graduate students in previous courses, so they can elaborate their diagrams correctly.

Table 2. Knowledge and experience that the undergraduate students possessed at the beginning of the experiment.

Characteristic	Knowledge and/or Experience
Programming Languages / Programming Environment	Java, C#, Pascal, C and Prolog.
Database Modeling Techniques	Entity Relationship Diagram, IDEF 1X
Analysis and Design Techniques	Structured and Object-oriented
Project Management	Experience in managing developing projects that included short software programming projects (Work Team of 3 or 4 students). No previous planning and estimation experience.

The descriptions of the six case studies with their analysis class diagrams are the following:

- *Case Study 1*: The objective was to develop a Sales System that automates the registration of customers and the data of their invoices. The information of the client is: client code, name, address, and phone number. The information of the invoice is: number of the invoice, date, and total amount. In this case, clients without invoices and invoices without clients can exist in the system.

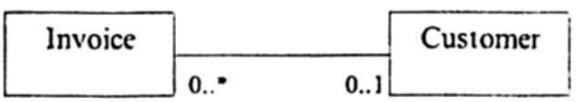


Fig. 3. Class Diagram for Case Study 1.

- *Case Study 2*: The objective was to develop a Sales System that registers invoices only. The information of the invoices is: number of the invoice, date, tax amount, and total amount. Additionally, the detail of the invoice includes: line number, product description, product quantity, product unit price and subtotal. The system also allows the registry of the types of products that are sold and their information: code, description and price.

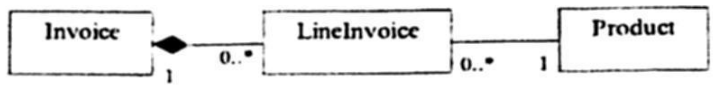


Fig. 4. Class Diagram for Case Study 2.

- *Case Study 3:* The objective was to develop a system that registers universities and their students. The information of the universities is: code, name, address, web page and telephone number. Students' information include: code, name, address, e-mail and telephone number. The system allows universities without students and students without universities. Students can belong to more than one university.

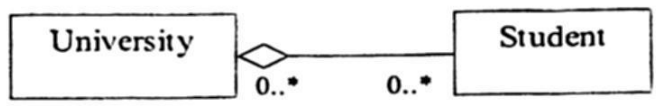


Fig. 5. Class Diagram for Case Study 3.

- *Case Study 4:* The objective was to develop a system that allows the registration of project plans and their activities. The system allows a project plan to exist without defined activities, but an activity must always be associated to a project plan exclusively.

Fig. 6. Class Diagram for Case Study 4.

- *Case Study 5:* The objective was to develop a system that allows the registration and maintenance of CD information which is defined by its code, title and duration. Information on each track of the CD includes: number of track, song name, artist and duration.

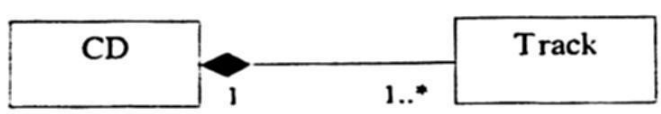


Fig. 7. Class Diagram for Case Study 5.

- *Case Study 6:* The objective was to develop a system that allows the registration of customer orders. Information for each order is: number, date and customer name. Additionally the detailed information for the order contains: line number, product code, product description and product quantity. The system also allows the registration of product information sold by the company, such as code and description.

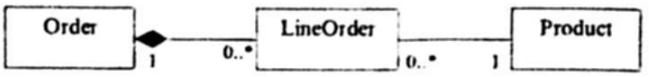


Fig. 8. Class Diagram for Case Study 6.

Further details of the case studies and used instruments can be found at:
http://macareo.pucp.edu.pe/japowsang/pf/composition.html

4.4 Tasks Performed in the Experiment

Table 3 shows the tasks carried out in the session by the students

Table 3. Tasks of the session carried out by the students

Number of Task	Group 1	Group 2
1	Reception of the case studies	
2	Reception of material with the explanation of our approach, Coad's patterns and form to fill number of LF and its RETs	Reception of material with the explanation of IFPUG CPM and form to fill number of LF and its RETs
3	Elaboration of analysis class diagrams and identification of LF	Elaboration of E-R diagrams and identification of LF
4	Delivery of completed forms back with the results	
5	Reception of material with the explanation of IFPUG CPM and form to fill number of LF and its RETs	Reception of material with the explanation of our approach. Coad's patterns and form to fill number of LF and its RETs
6	Elaboration of E-R diagrams and identification of LF	Elaboration of analysis class diagrams and identification of LF
7	Delivery of completed forms back with the results and reception of questionnaire	
8	Delivery of questionnaire	

The session lasted approximately one hour and the students performed forty and five minutes on average in all of the tasks. Although, our study did not include a timing analysis for each technique, we could observe that both groups used almost the same time to identify LF using both techniques. In addition, it is important to mention that we informed the students that the purpose of the questionnaire was to know their honest opinion about which technique was easier to apply.

5 Results

For each case study, we graded it with "1" (one) if the student correctly identified the number of LF and RET and "0" (zero) if he or she did it incorrectly.

Subsections 5.1, 5.2 and 5.3 present quantitative results and subsection 5.4 shows the results obtained from the questionnaire.

5.1 Undergraduate Students Results

Acknowledging the advantages of utilizing students in experiments [7], Table 4 presents the detailed results for each undergraduate student. For each case study, the first column shows the results obtained with the IFPUG CPM 4.2.1 technique and the second one with our proposal. Results presented in Table 4 were previously published in [18]

Table 4. Results obtained with undergraduate students

Student	Case Study											
	1		2		3		4		5		6	
1	0	1	1	0	1	1	0	0	0	0	1	0
2	1	1	1	1	1	1	1	1	1	1	1	1
3	1	1	1	1	1	1	1	1	1	1	1	1
4	1	1	1	1	1	1	1	1	1	1	1	1
5	1	1	0	1	1	1	1	1	1	1	0	1
6	1	1	0	1	1	1	1	1	1	1	0	1
7	1	1	0	1	0	1	1	1	0	1	0	1
8	1	1	0	1	1	1	1	1	1	1	0	1
9	1	1	1	1	1	1	1	1	1	1	1	1
10	1	1	1	1	1	1	1	1	0	1	1	1
11	1	1	1	1	1	1	1	1	1	1	1	1
12	1	1	1	1	1	1	1	1	1	1	1	1
13	1	1	1	1	1	1	1	1	1	1	1	1
14	1	1	1	1	1	1	1	1	1	1	1	1
15	1	1	1	1	1	1	1	1	1	1	1	1
16	1	1	1	1	1	1	1	1	1	1	1	1
17	1	1	1	1	1	1	1	1	0	1	1	1
18	1	1	1	1	1	1	0	1	0	0	1	1
19	1	1	1	1	1	1	1	1	1	1	1	1
20	1	1	1	1	1	1	1	1	1	1	1	1
21	1	1	1	1	1	1	1	1	1	1	1	1
22	1	1	0	1	1	1	0	1	0	1	0	1
23	1	1	1	1	1	1	1	1	1	1	1	1
24	1	1	1	1	1	1	1	1	1	1	1	1

A significance level of 0.05 was established to statistically test the obtained results. Since these results follow a non-normal distribution, the paired samples t-test could not be used and a non-parametric alternative was selected: the Wilkoxon signed rank test. The statistical hypotheses formulated to test both techniques are:

- H0: The distribution of the two samples is not significantly different.
- Ha: The distribution of the IFPUG CPM 4.2.1 sample is shifted to the left of the distribution of our proposed FPA sample.

Table 5. Wilkoxon signed rank test results for undergraduate students

Variable	Result
Observations	144
V	269.0
Expected value	1345.0
Variance (V)	50732.5
p-value (one-tailed)	<0.0001
Alpha	0.05

Since the computed p-value is lower than the significance level alpha=0.05 as shown in Table V, we can reject the null hypothesis H0 and accept the alternative hypothesis Ha. It means that we can empirically corroborate that our proposal produces more accurate assessments than the IFPUG CPM 4.2.1 approach.

5.2 Graduate Students Results

Eighteen students participated in the experiment. All of the graduate students, except one (he performed all case studies correctly using our approach, but only 2 case studies correctly using IFPUG), performed correctly the case studies using both techniques.

For these results, the same statistical test as the one utilized for undergraduate students results was used. The statistical hypotheses formulated to test both techniques are:

- H0: The distribution of the two samples is not significantly different.
- Ha: The distribution of the IFPUG CPM 4.2.1 sample is shifted to the left of the distribution of our proposed FPA sample.

Since the computed p-value is higher than the significance level alpha=0.05 as shown in Table VI, we cannot reject the null hypothesis H0 and reject the alternative hypothesis Ha. It means that we can not empirically corroborate that our proposal produces more accurate assessments than the IFPUG CPM 4.2.1. approach.

Table 6. Wilkoxon signed rank test results for graduate students

Variable	Result
Observations	144
V	0.0
Expected value	285.0
Variance (V)	0.0
p-value (one-tailed)	1.0
Alpha	0.05

Due the results obtained in Table 6, we changed the alternative hypothesis. The

new Ha was *the distribution of the two samples is significantly different*. Using the information showed in Table VI, we still cannot reject H0, which empirically corroborates that our proposal and the IFPUG CPM 4.2.1 approach produces the same accurate assessment.

5.3 Comparison of Quantitative Results

From what it is observed in subsections 5.1 and 5.2, there is a difference in results from both kinds of students. Some of the reasons for those differences are:

Undergraduate students knew about structured techniques, but they were more familiar with OO techniques because they participated in OO software development projects before the experiment. They obtained better results with our proposal because they had more experience with OO approaches.

Many of the graduate students apply only structured techniques at work, although they use OO development tools. Therefore, their experience with OO techniques was not as sound as the experience of undergraduate students (before the experiment they were required to take an OO analysis and design course). Due to their experience in structured techniques, graduate students obtained better results using IFPUG CPM 4.2.1 than undergraduate students.

Based on the results obtained with undergraduate and graduate students, we can conclude that our proposal produces at least the same accurate assessment as that of the IFPUG CPM 4.2.1.

5.4 Questionnaire Results

As can be seen from Table 3, students filled a questionnaire about the techniques used in the experiment. Two multiple-choice questions (results are presented in figures 9 and 10) and a comment section were included.

Fig. 9 shows the results of the question: Rules regarding composition (technique with classes) facilitate LFs and RETs identification compare to technique without classes? They clearly show that all of the undergraduate students and most of the graduate students consider that our proposal facilitates LF identification. It can be observed that 16.7% of graduate students think both techniques are the same, due to their experience with structured techniques at work.

Fig. 10 presents the results of the question *The technique with classes is easier to apply than the technique without classes?* From these results, it is shown that most of the students (graduate and undergraduate) think that our proposal is easier to apply than the IFPUG CPM 4.2.1 approach.

As it can be observed from Figures 9 and 10, the questionnaire results do not disagree with the quantitative results presented in previous subsections. Few students wrote comments, but all of them confirmed the results obtained in the multiple-choice questions.

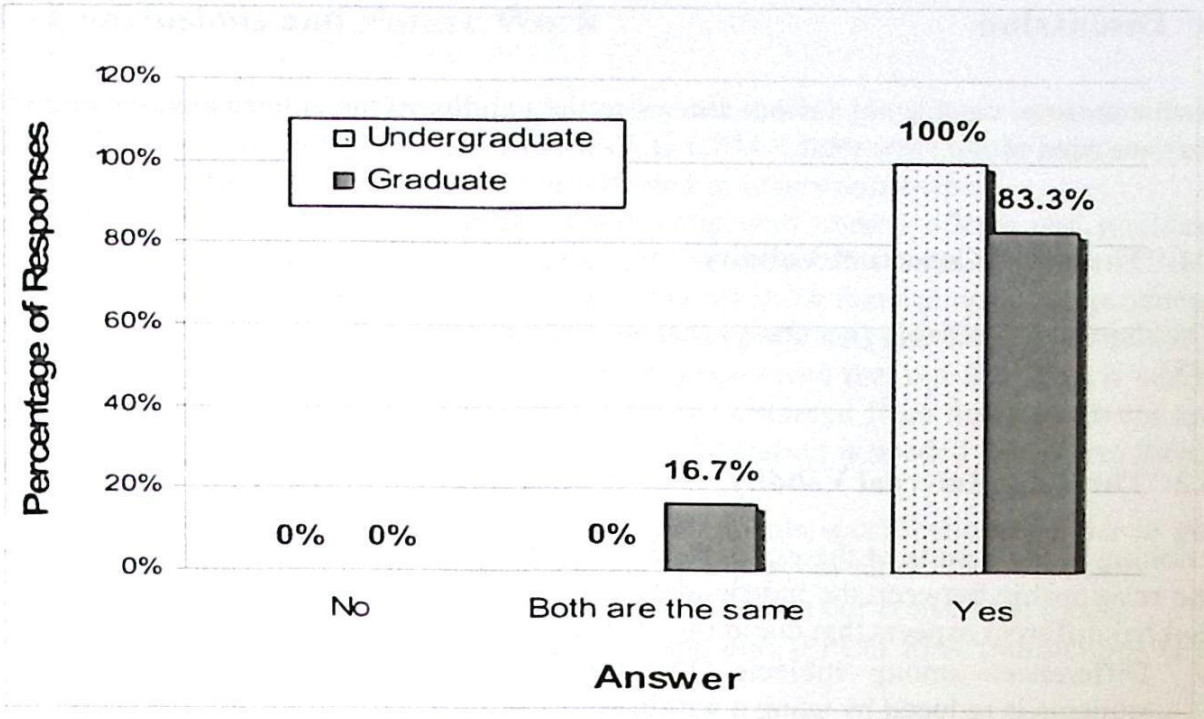


Fig. 9. Results of the question "Rules of our proposal facilitate LFs and RETs Identification?"

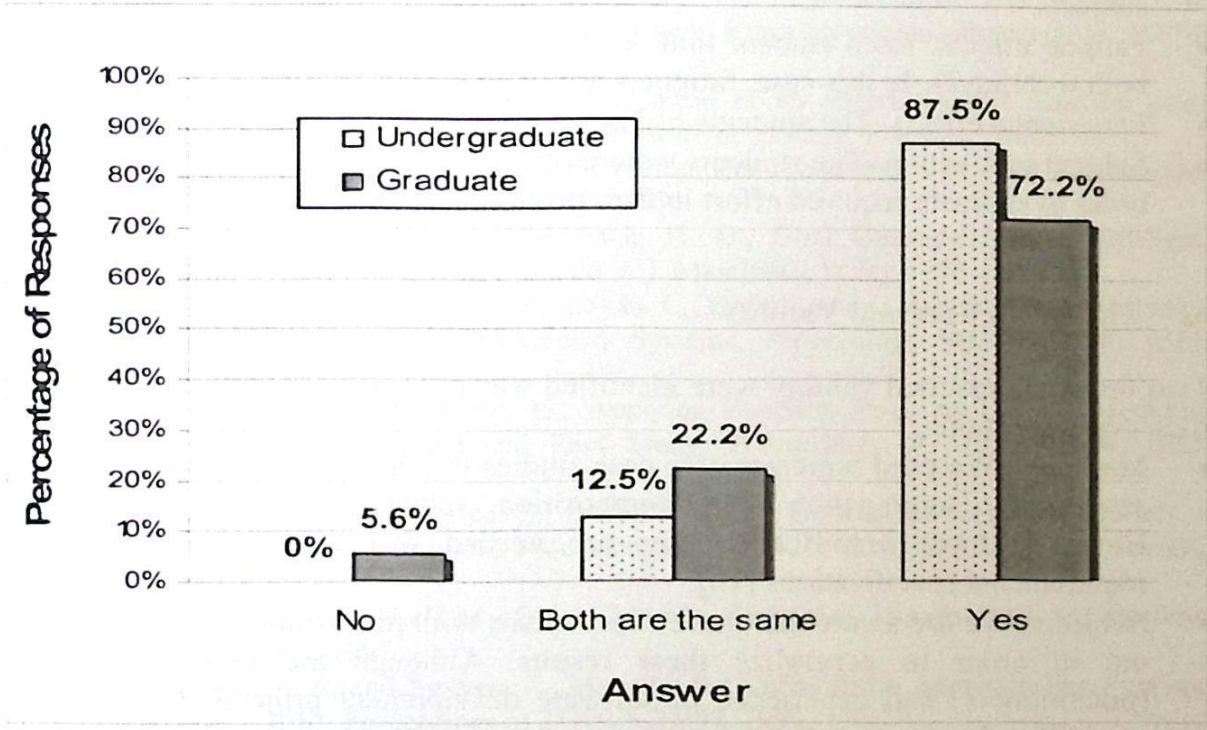


Fig. 10. Results of the question "The technique with classes is easier to apply than the technique without classes?"

6 Discussion

In this section, we discuss various threats to the validity of the empirical study and the way we tried to alleviate them.

6.1 Threats to Construct Validity

The dependent variable (accuracy) that we used is proposed in the ISO/IEC 14143-3 [13].

6.2 Threats to Internal Validity

Looking at the results of the experiment, we can conclude that empirical evidence for the relationship between the independent and the dependent variables exists. We have tackled different aspects that could threaten the internal validity of the study:

- Differences among subjects. The error variance due to differences among students is reduced by using a within-subjects design.
- Learning effects. The counterbalancing procedure (subjects were randomly assigned in two groups) cancelled the learning effect due to similarities and the order of application of both techniques.
- Knowledge of the universe of discourse. We used the same case studies for all subjects.
- Fatigue effects. Each student took 45 minutes on average per session to apply both techniques. In this case, fatigue was not a relevant factor.
- Persistence effects. The students had never done similar experiments before.
- Subject motivation. The students were motivated because they had to apply FP in order to estimate required effort in their projects assigned for the semester.

6.3 Threats to External Validity

Two threats to external validity were identified which limited the ability to apply any such generalization:

- Materials. We used representative case studies in which students had to identify association, aggregation and composition relationships between classes. However, more empirical studies are needed that make use of software requirements specifications [10].
- Subjects. We are aware that more experiments with practitioners must be carried out in order to generalize these results. Although the graduate students (practitioners) had experience in software development projects, they had not used FP technique in their projects before the experiment.

7 Conclusions and Future Work

This paper presented a conversion model to determine FP logic files using an analysis class diagram. This model considers the IFPUG CPM 4.2.1 rules and the composition relationship between classes that are not included in others approaches.

We also described two controlled experiments with undergraduate and graduate students in order to determine the accuracy of our approach compared to the IFPUG CPM 4.2.1 technique. The results of the experiments show that our approach produces at least equal results in accuracy when compared to the original IFPUG CPM 4.2.1 rules, and students (undergraduate and graduate) perceived that our approach is easier to use than the IFPUG FPA. Although the results obtained from the experiment are very encouraging, we are aware that more experimentation is needed to confirm them.

The future work regarding this research is:

- To conduct experiments with software requirements specification in order to get more results and opinions about the applicability of our model in the industry.
- To include generalization and association relations in our conversion model.
- To define rules to transform UML diagrams into IFPUG FPA transactions (EI; EO and EQ).

References

1. Abrahão, S., Poels, G., Pastor, O., Assessing the Reproducibility and Accuracy of Functional Size Measurement Methods through Experimentation, Proceedings ISESE 2004, IEEE Computer Society, 2004.
2. Abrahão, S., Poels, G., Experimental evaluation of an object-oriented function point measurement procedure, Information & Software Technology, Elsevier, 2007.
3. Albrecht, A. J., Measuring Application Development Productivity, in IBM Applications Development Symposium, Monterey, CA, 1979.
4. Basili, V. R., Caldiera, G. and Rombach, H. D., Goal Question Metric Paradigm, Encyclopedia of Software Engineering, ed. J. J. Marciniak, Wiley 1994.
5. Caldiera, G., Antoniol, G., Fiutem, R., Lokan, C., Definition and Experimental Evaluation of Function Points for Object-Oriented Systems, Proceedings METRICS'98, IEEE Computer Society, 1998.
6. Cantone, G., Pace, D., Calavaro, G., Applying Function Point to Unified Modeling Language: Conversion Model and Pilot Study, Proceedings of METRICS'04, IEEE Computer Society, 2004.
7. Carver, J., Jaccheri, L., Morasca, S., Issues in Using Students in Empirical Studies in Software Engineering Education, Proceedings METRICS'03, IEEE Computer Society, USA, 2003.
8. Coad, P., North,D., Mayfield,M., Object Models: Strategies, Patterns and Applications, Prentice-Hall, USA, 1997.
9. Fetcke, T., Abran, A. and Nguyen, T., Mapping the OOJacobson Approach into Function Point Analysis, IEEE Proceedings of TOOLS-23'97, 1997.
10. IEEE Computer Society, IEEE Std 830-1998, Recommended Practice for Software Requirements Specifications, The Institute of Electrical and Electronics Engineers, USA, 1998.
11. IFPUG, Function Points Counting Practices Manual (version 4.1.1), IFPUG: International Function Point User Group, Westerville Ohio, 2000.

12. IFPUG. Function Points Counting Practices Manual (version 4.2.1), IFPUG: International Function Point User Group. Westerville Ohio, 2004.
13. ISO. ISO/IEC 14143-3 - Information technology -- Software measurement -- Functional size measurement -- Part 3: Verification of functional size measurement methods, 2003.
14. Jaaksi, A., A Method for Your Object-Oriented Project. Journal of Object-Oriented Programming. Vol 10. No 9, 1998.
15. Jacobson, I., Object-Oriented Software Engineering. A Use Case Driven Approach. Addison-Wesley, USA, 1992.
16. Juristo, N., Moreno, A.M., Basics of Software Engineering Experimentation, Kluwer Academic Publishers. Boston, 2001.
17. Larman, C., Applying UML and Patterns: An Introduction to Object-Oriented Analysis and Design and Iterative Development. Third Edition, Addison-Wesley, 2004
18. Pow-Sang, J.A., Imbert, R., Including the Composition Relationship among Classes to Improve Function Points Analysis. Proceeding VI Jornadas Peruanas de Computación-JPC'07. Trujillo, Peru, 2007.
19. Object Management Group. OMG Unified Modeling Language, http://www.uml.org. USA. 2005.
20. Uemura, T., Kusumoto, S. and Inoue, K., Function Point Measurement Tool for UML Design Specification. in 5th International Software Metrics Symposium (METRICS'99), Florida, USA, 1999, 62-69.

Mathematic Formulae for Calculating the CAVE's Room Size

Marva Angélica Mora Lumbreras and Antonio Aguilera Ramirez

[1]Centro de Investigación en Tecnologías de Información y Automatización
Universidad de las Américas Puebla
Sta. Catarina Mártir, Cholula, Puebla, 72820, México.
{marva.morals, antonio.aguilera}@udlap.mx

Abstract. This paper is focused on the creation of mathematic formulae for calculating the room's size where a CAVE (CAVE Automatic Virtual Environment) of specific dimensions will be built. Furthermore, the design of mathematic formulae for computing the CAVE's size, when the room's size is already established. Four designs are explained. In the first design, the CAVE's walls are parallel to the room's walls. In the second design, the CAVE's walls are oblique to the room's walls. In the third design, the mirror and projector technique is used. Finally, the mirror and projector technique is used, too. The only difference is that the projectors are set on a strategic place.

Keywords: CAVE, projection, immersion.

1 Introduction

A CAVE is a system based on multiple projection' s screens that surround the viewer, in which projectors are directed to four, five or six of the walls of a room-sized cube and where generally a virtual environment is used.

The first CAVE was originally developed at the Electronic Visualization Laboratory at the University of Illinois at Chicago in 1992 [1], [2]. This CAVE used the mirror and projector technique.

Actually, this technique is used for many CAVEs, as in [3], [4] and [5]. Nevertheless, there is no reference that indicates a specific method for computing the minimum space required for building a CAVE. There are no references that may help us to compute the physical space when the mirrors are used in different positions.

2 Projection

In this research, different projectors were used for measuring the projected image's size. The results were similar. Every projector had an aspect ratio 4:3. The projector's length was not taken in account due to it is minimum size.

Fig. 1 shows the maximum projection's size, when the distance between the projector and the wall is of 3m, it is possible to notice that the width of the projected image is almost half of the distance between the projector and the wall. For this

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 19-26

Received 26/06/08
Accepted 04/08/08
Final version 09/09/08

reason, in the rest of this paper the projection's width will be taken as half of the distance between the projector and the wall.

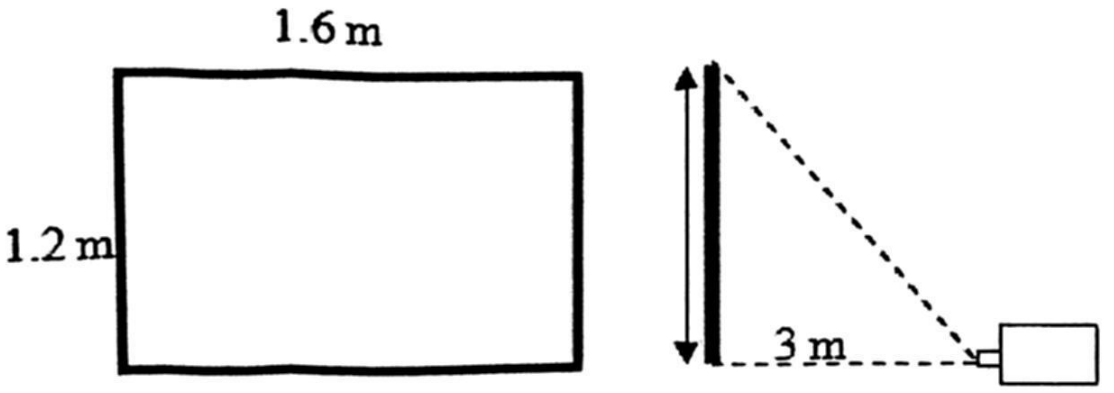


Fig. 1. Image's size obtained at a distance of 3 m.

3 Room Size

In order to achieve a good immersion level within the CAVE, its walls must be taller than the average height, so that people may not have the opportunity of looking over them. Using a screen of 2.10 m. height should be enough for most situations.

Considering the aspect ratio of 4:3, means that the projection height is 75% of its width [6], this example sets the screen width as P = 2.80 m. In this paper we use P, which avoids loss of generality.

3.1 Room Design #1

The first one is a straight approach that allows us to set the projectors behind screens. See Fig. 2.

$$W\_1 = 5P\,. \tag{1}$$

$$W\_2 = 3P\,. \tag{2}$$

$$R\_1 = (5P)(3P) = 15P^2\,. \tag{3}$$

This means that the room's width must be of at least 5P, while its length must be larger than $3P$ to allow the user's entrance into the cave.

In our example, where P = 2.80 m., the room size should be: R_1=117m^2.

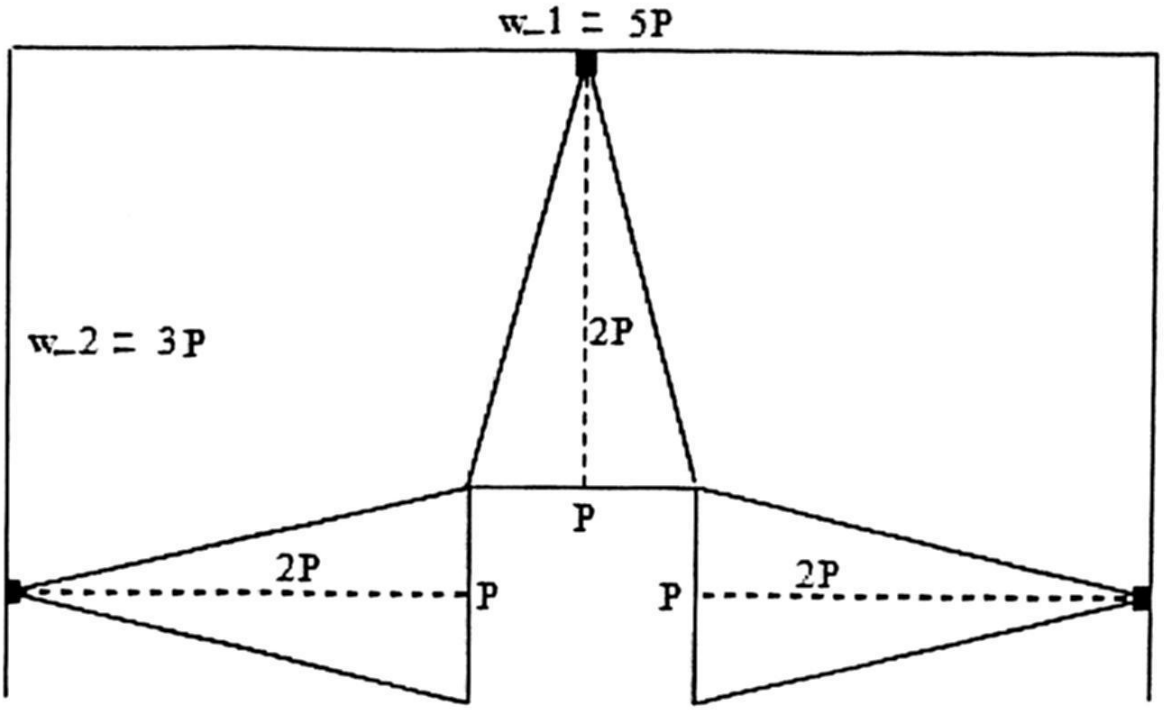


Fig. 2. Room size: CAVE walls parallel to room walls.

3.2 Room Design #2

In order to reduce the room's size, we can set the projectors at strategic places of the room. Corners may be used.

Fig. 3 shows that the distance from the center point of the room to any corner is: $\frac{5P}{2}$. Therefore the length of each wall is: $W=\frac{5P\sqrt{2}}{2}$.

$$R\_2=\left(\frac{5P\sqrt{2}}{2}\right)\left(\frac{5P\sqrt{2}}{2}\right)=\frac{25P^2}{2}. \tag{4}$$

Using P=2.80 m., the room's size is $R\_2=98m^2$

3.3 Room Design #3

As it was mentioned before, the projection's size depends on the distance between the projector and the wall, thus note that it is possible to reduce the room's size requirements by using mirrors.

There are two types of mirrors that are useful for back-projection systems such as: foil mirrors and surface glass mirrors. Normal mirrors are not usually used because they may produce double images [5].

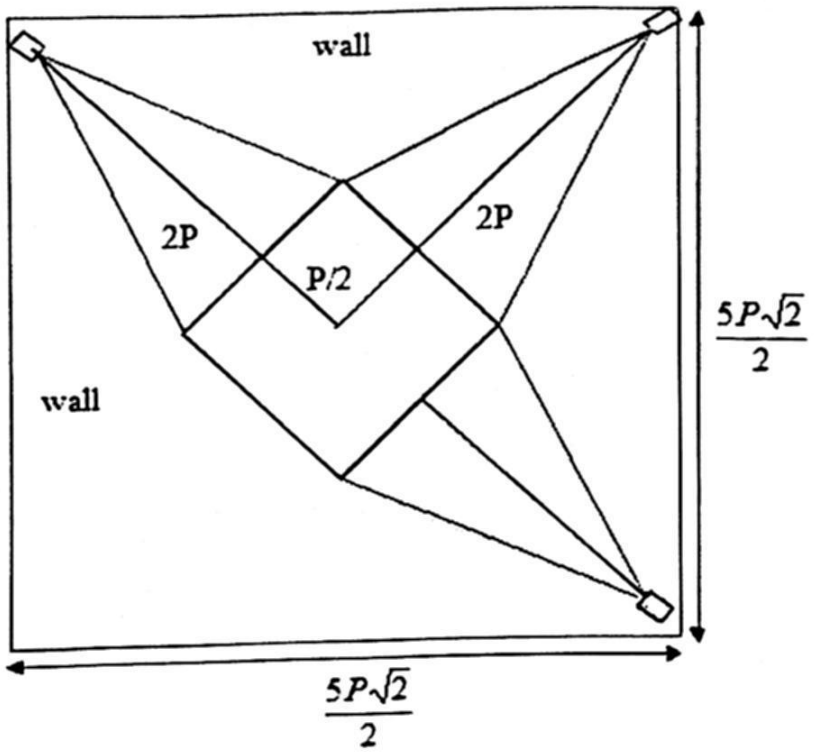


Fig. 3. Shows the room's size with the projectors placed in the corners.

Fig. 4 shows an example of an image's size obtained with a projector and a mirror. The strategy is to set the projector near the screen. The projection is sent to the mirror and the mirror reflects the projection onto the screen. The projection's distance and the horizontal projection's size are both equal to P.

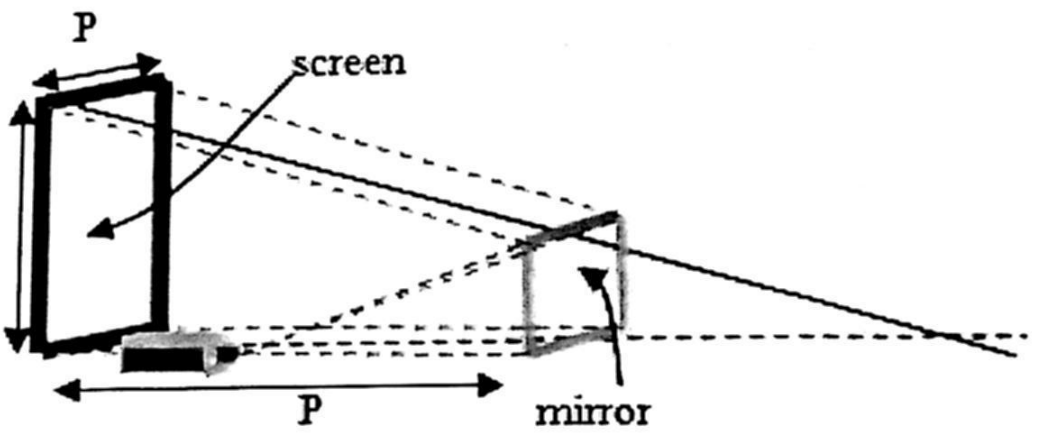


Fig. 4 A larger image size is obtained in smaller sized rooms using mirrors.

Fig. 5 allows us to see that the distance from the center point of the room to any corner is: $\frac{7P}{4}$, for that reason, the length of each wall is: $\frac{7P\sqrt{2}}{4}$.

In this design, the mirror's size is half the screens' size.

$$R\_3 = \left(7P\frac{\sqrt{2}}{4}\right)\left(7P\frac{\sqrt{2}}{4}\right) = \frac{49P^2}{8}. \tag{5}$$

It is necessary to be careful to place the mirror in the correct position because it is possible that the projection crosses a screen

In our example, where P = 2.80m., the room's size should be R_3= 48.02m^2

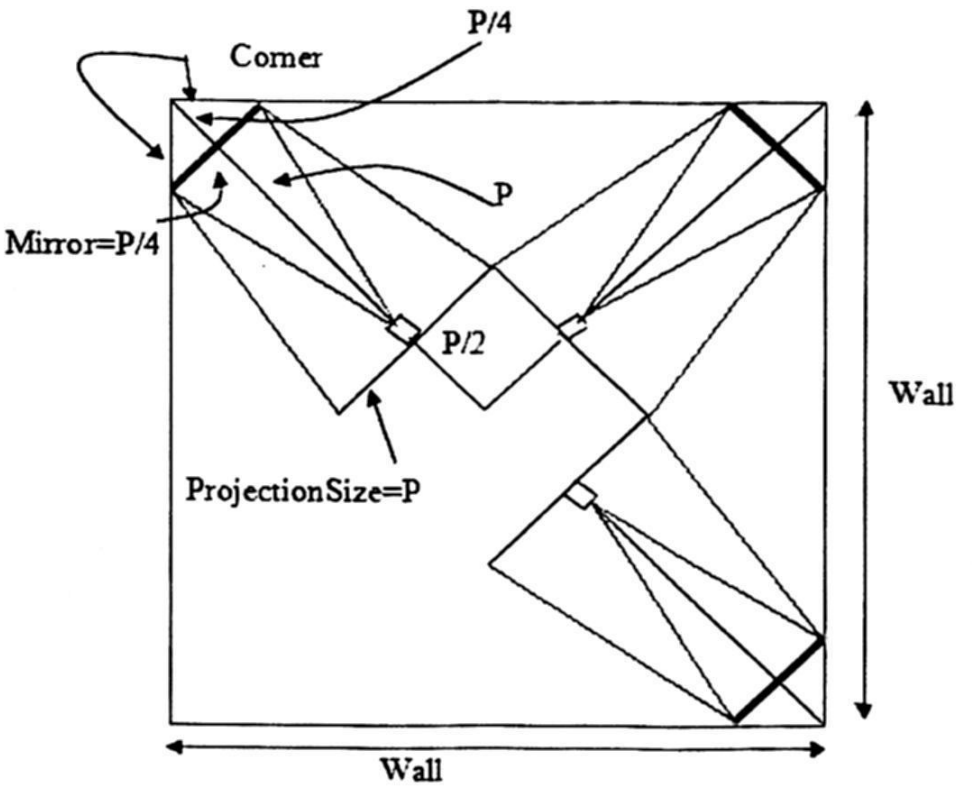


Fig. 5. Room size using mirrors.

3.4 Room Design #4

It is possible to find strategies for taking advantage of the mirrors. Fig. 6 shows that the mirror can be inclined towards the projector.

Fig. 7 allows us to see that the distance from the center point of the room to any corner is approximately $\frac{3P}{2}$, therefore the length of each wall is $\approx \frac{3P\sqrt{2}}{2}$

$$R\_4 \approx \left(3P\frac{\sqrt{2}}{2}\right)\left(3P\frac{\sqrt{2}}{2}\right) \approx \frac{9P^2}{2}. \tag{6}$$

In our example, where P = 2.80m., the room's size should be of $\approx$35.28m^2.

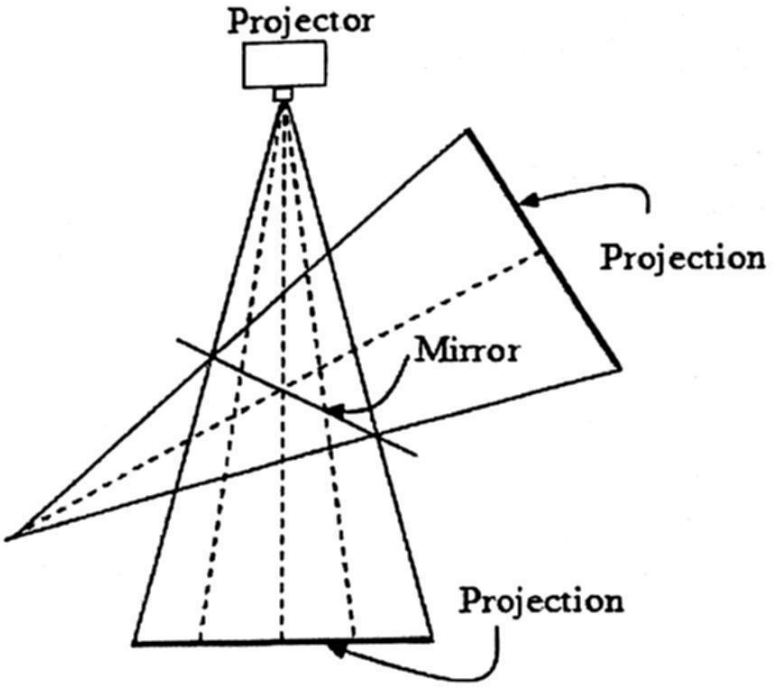


Fig. 6 Projection using an inclined mirror.

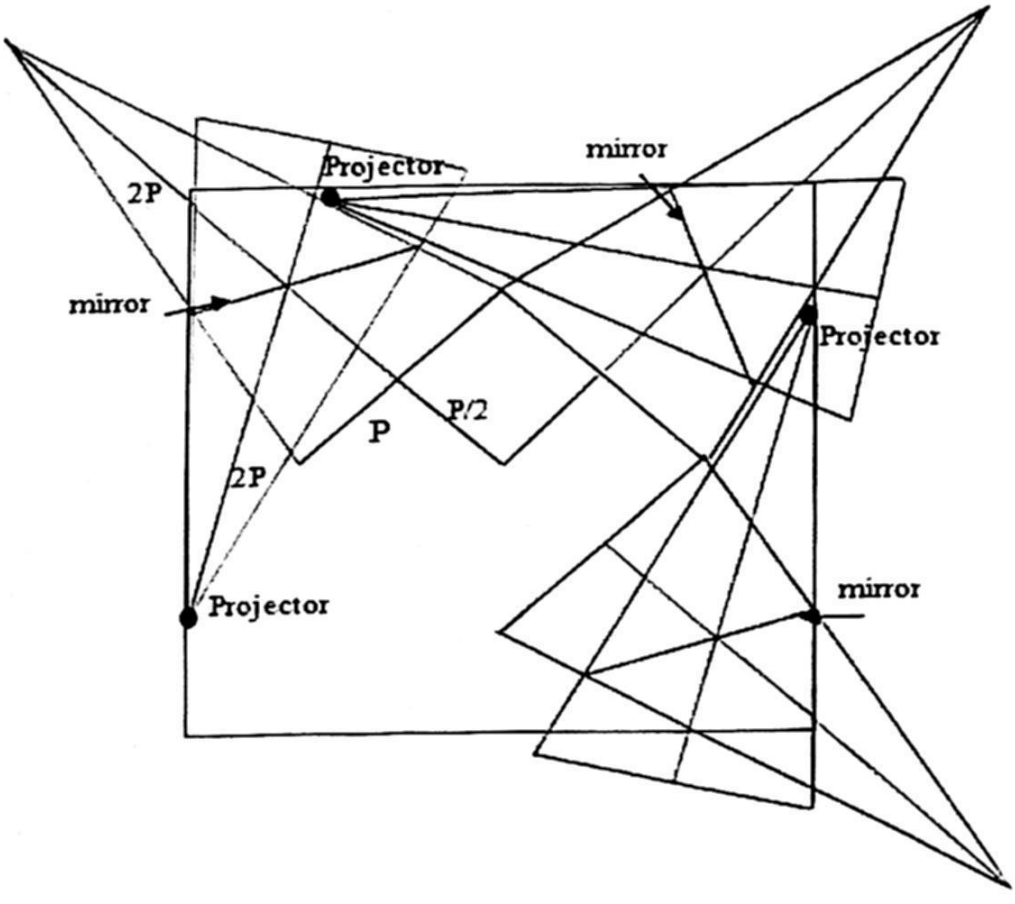


Fig. 7. Room Size using the mirror and projector technique, the CAVE's walls are in an oblique position against the room walls, placing the projectors in a strategic place

4 Calculating the CAVE's Size when the Room's Size is already Established

In the designs shown above, the room's size depends on the screen's size. Therefore, it is possible to calculate the CAVE's size when the room's size is known.

The CAVEs' sizes (C_2, C_3 and C_4) are calculated taking as a reference designs 2, 3 and 4:

If we know the room's size, we can calculate the CAVE's dimensions. Where P is the width and depth, ¾P is the height (H).

For design #2 and taking as reference formula (4), we can calculate C_2.

$$P = \frac{\sqrt{2}}{5} W . \tag{7}$$

$$H = \frac{3\sqrt{2}}{20} W . \tag{8}$$

$$C\_2 = \left(\frac{\sqrt{2}}{5} W \right) \left(\frac{3\sqrt{2}}{20} W \right) \left(\frac{\sqrt{2}}{5} W \right) . \tag{9}$$

For design #3 and taking as a reference formula (5), we can calculate C_3.

$$P = \frac{2\sqrt{2}}{7} W . \tag{10}$$

$$H = \frac{3\sqrt{2}}{14} W . \tag{11}$$

$$C\_3 = \left(\frac{2\sqrt{2}}{7} W \right) \left(\frac{3\sqrt{2}}{14} W \right) \left(\frac{2\sqrt{2}}{7} W \right) . \tag{12}$$

Finally, for design #4 and taking as a reference formula (6). we can calculate C_4.

$$P = \frac{\sqrt{2}}{3} W . \tag{13}$$

$$H = \frac{\sqrt{2}}{4} W . \tag{14}$$

$$C\_4 = \left(\frac{\sqrt{2}}{3} W \right) \left(\frac{\sqrt{2}}{4} W \right) \left(\frac{\sqrt{2}}{3} W \right) . \tag{15}$$

5 Conclusion

Four designs for building a CAVE with different characteristics were explained in this paper. Each design was evaluated with different projectors and the results were satisfactory.

Two contributions were given:

-The creation of mathematic formulae for calculating the room's size where a CAVE of specific dimensions would be built.

- The creation of mathematic formulae for calculating the CAVE's size when the room's size is already established.

Designs #1 and #2 used a great amount of space. Design #3 used the mirror and projector technique; the space used in this design was reduced. In my personal opinion, the best option was the last one.

With these designs we can demonstrate, that the technique of mirrors and projectors is effective.

References

1. Cruz-Neira, C., Sandin, D. J., And Defanti, T. A. 1993. Surround screen projection-based, virtual reality: the design and implementation of the cave. In Proceedings of the 20th annual conference on Computer graphics and interactive techniques, ACM Press, 135–142.
2. Dave Pape, Carolina Cruz-Neira, Marek Czernuszenko. CAVE User's Guide. Electronic Visualization Laboratory. University of Illinois at Chicago. 1997
3. Johan Ihr_en and Kicki J. Frisch. The Fully Immersive CAVE. Center for Parallel Computers in Stockholm, Sweden.
4. National Center For Supercomputing Applications, Beckman Institute, 2005, http://cave.ncsa.uiuc.edu/
5. Janne Jalkanen. Building A Spatially Immersive Display: HUTCAVE. Licentiate Thesis. Helsinki University of Technology (2000)
6. Aguilera, Antonio. Sistemas de Multidisplay: Técnicas y Aplicaciones (Multidisplay Systems: Techniques and Applications, written in Spanish). Master Thesis. IunivertecH (I.U.T.H.). Puebla, México (2008).

Workflow and Collaboration

Improving Knowledge Flow in a Mexican Manufacturing Firm

Oscar M. Rodríguez-Elias[1], Alberto L. Morán[2], Jaqueline I. Lavandera[3], and Aurora Vizcaíno[4]

[1]Departamento de Matemáticas. Universidad de Sonora. Hermosillo, Son., México
[2]Facultad de Ciencias. Universidad Autónoma de Baja California. Ensenada, B.C., México
[3]FAMOSA-Ensenada. Ensenada, B.C., México
[4]ALARCOS Research Group. Universidad de Castilla-La Mancha. Ciudad Real, España
omrodriguez@ciencias.uson.mx; alberto_moran@uabc.mx, jilavmac@efemsa.com; Aurora.Vizcaino@uclm.es

Abstract. Integrating Knowledge Management (KM) in organizational processes has become an important concern in the KM community. Development of methods to accomplish this is still being, however, an open issue. KM should facilitate the flow of knowledge from where it is created or stored, to where it is needed to be applied. Therefore, an initial step towards the integration of KM in organizational processes should be the analysis of the way in which knowledge is actually flowing in these processes, taking into account the mechanisms that could be affecting (positively or negatively) such a flow, and then, to propose alternatives to improve the knowledge flow in the analyzed processes. This paper presents the use of the Knowledge Flow Identification (KoFI) methodology as a means to improve a manufacturing process knowledge flow. Since KoFI was initially developed to analyze software processes, in this paper we illustrate how it can also be used in a manufacturing domain. The results of the application of KoFI are also presented, which include the design of a knowledge portal and an initial evaluation from its potential users.

1 Introduction

Nowadays Knowledge Management (KM) has captured enterprises' attention as one of the most promising ways to reach success in this information era [6]. In order to assist organizations to manage their knowledge, different strategies and systems (Knowledge Management Systems, KMS) have been designed. However, developing them is a difficult task; since knowledge *per se* is intensively domain dependent whereas KMS often are context specific applications. In the one hand, the lack of sophisticated methodologies or theories for the extraction of reusable knowledge and reusable knowledge patterns has proven to be extremely costly, time consuming and error prone [5]. On the other hand, an actual concern is that KM approaches should be well integrated to the knowledge needs of knowledge workers, and to the work processes of organizations [18]. Before developing a KM strategy it is advisable to understand how knowledge transfer is carried out by people in the different processes where the strategy will be applied. Once the forms in which knowledge is flowing

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 29-45

Received 04/06/08
Accepted 04/08/08
Final version 10/09/08

through a process have been recognized, it should be easier to identify the problems affecting that flow, and, as a consequence, to propose possible solutions to improve the flow.

In this paper, we illustrate the manner in which the Knowledge Flow Identification (KoFI) methodology [17] was used to analyze a manufacturing process, in order to improve its knowledge flow. The reason for engaging in this study was to assist a manufacturing organization in two main aspects: 1) to improve the training of highly competitive personnel, and 2) to promote organizational learning. The main concern was to develop a KM system to assists the human resources training process, by making useful information and resources available to the employees to promote self-learning and knowledge diffusion. The goal of this paper is to illustrate how the KoFI methodology can help to detect knowledge deficiencies in a manufacturing process, and can also help to design strategies to solve them; in this case a knowledge portal was designed. Hence, in the next section the manufacturing process where KoFI was used is described, after that in Section three we illustrate the different stages followed to improve that process. Then, in Section Four a knowledge portal, designed as a result from the findings obtained after applying the methodology, is described. Section Five depicts the results of a preliminary evaluation of this portal. A discussion of the results of this case study is presented in Section Six. Latter, in Section Seven our approach is compared to existent related work, to finally conclude in Section Eight.

2 The Manufacturing Process

In order to test the KoFI methodology it was used in an industrial company dedicated to the manufacturing of cans. We focused our work on a department where eight processes are carried out. It was decided to focus on one of the most important process: the one in charge of transforming the aluminum rolls into the first versions of the cans (known as "Formation area"). In this test 41 people were involved, including the department manager, the responsible of each area of the department, and the operating personnel, which were integrated by leader mechanics, productive processes mechanics, and machine operators.

It is important to highlight that the company has documented all its processes, and follows standards for documenting almost all its activities. Moreover it has an ISO9001-2000 certification. Because of this, detailed models of the processes were already available.

The data used to analyze the process was captured through interviews, and by analyzing documents and information systems. Nineteen employees were interviewed by using the long interview technique, but adjusting the interviews to the following format: the general data of those interviewed, the main activities performed, knowledge sources known by them, and their level of knowledge about the process. The duration of the interviews ranged from 30 minutes to 2 hours, depending on the level of responsibility of those interviewed. Additionally, a total of 119 documents and systems were also analyzed, of which 24 were discarded because they were duplicated.

3 Applying the KoFI Methodology

Before presenting the results of applying the KoFI methodology, we will present a brief description of the methodology. Then, we will focus on the results of the analysis of the manufacturing process.

3.1 Description of the KoFI methodology

The KoFI methodology was designed to aid in the analysis of software processes from a knowledge flow perspective [14, 17]. It was defined to assist in three main areas: 1) to identify, structure, and classify the knowledge that exists in the process studied, 2) to identify the technological infrastructure which supports the process and affects the knowledge flow, and 3) to identify forms with which to improve the knowledge flow in the process.

KoFI is orientated towards helping to analyze specific work processes. Therefore, it is necessary to define the specific process and model it. The process models are later analyzed following a four stage process, to finally identify and describe the tools which, positively or negatively, affect the flow of knowledge. Thus, based on the main activities of KoFI, the methodology is divided in three phases (see Figure 1):

- **The process modeling phase**, consisting of the definition and modeling of the process to be analyzed, using a process modeling language which provides elements to represent the knowledge involved in the process. It is recommended to model the process at different levels of abstraction. First, a general view of the process can be defined with a general and flexible process modeling technique. In our case, we have used an adaptation of the Rich Picture Technique [9, 14]. To perform a detailed analysis, a more formally constrained language should be used.
- **The process analysis phase**, which involves the identification and analysis of knowledge sources, topics, and flows, as well as the problems affecting the flow of knowledge. The main result of this phase is the definition of a knowledge map of the process, which can be structured towards the definition of an ontology of knowledge sources and topics, considering their relationships with other elements of the process, such as activities or roles.
- **The knowledge flow support tools analysis phase**, consisting of the analysis of the tools that might be useful knowledge flow enablers. To accomplish this phase, a framework has been proposed [16], which define four main steps to analyze information systems as knowledge flow enablers. First, the application domain of the system is defined. This includes identifying the use, scope, and domain of the knowledge managed. The second step consists of identifying the structure of such knowledge. Later, the third step focuses on defining the KM activities being supported by the tool. Finally, the fourth step consists of the definition of the main technical aspects considered important for the tools.

After the application of the methodology, information should have been obtained which is useful, for example, in defining the knowledge base of the process, discovering the problems affecting the flow of knowledge and the mechanisms through which knowledge is flowing, and making proposals to improve the knowledge flow.

In the present work, we focused on analyzing a manufacturing process in order to propose a KM system. Thus, for this study the methodology was applied only until phase two. That means we did not apply the final phase of the methodology. Therefore, in this paper we will focus on the process analysis phase. Detailed information about how to perform the other two phases can be found in [15] for the process modeling phase, and in [16] for the knowledge flow support tools analysis phase.

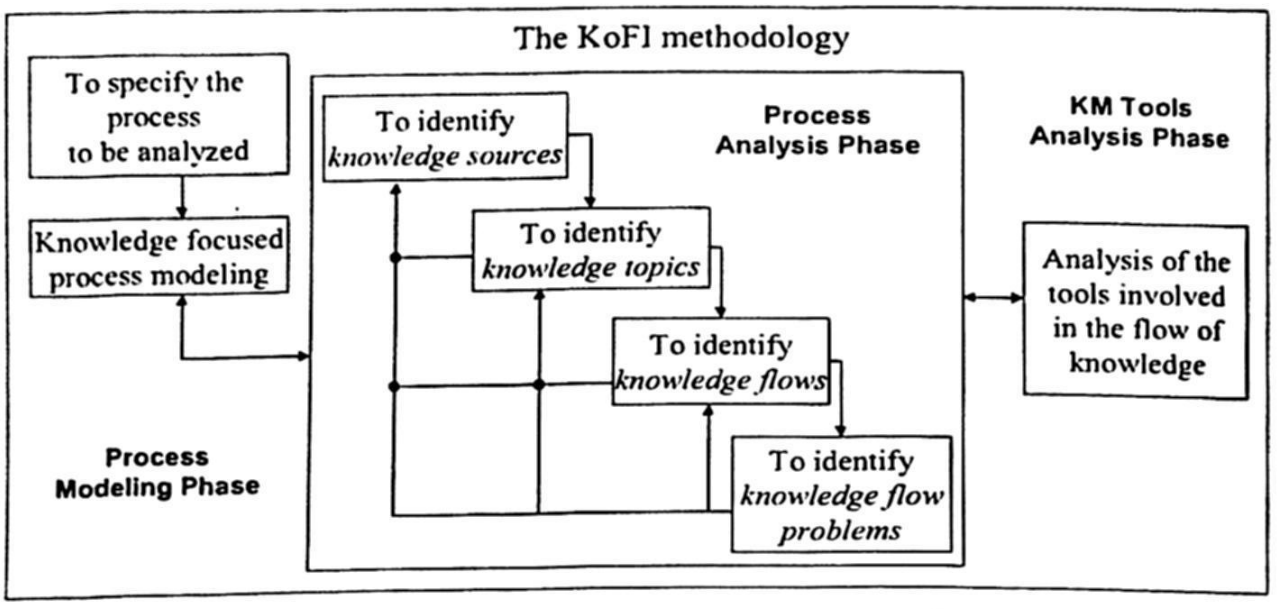


Fig. 1. The four steps of the process analysis phase of the KoFI methodology.

The process analysis phase of KoFI is composed of four steps as shown in Figure 1, which are performed in an iterative way, since each step might provide information useful for the others preceding it. Also, in that manner the products of each step would evolve to incorporate the new items found in each iteration. Next we briefly describe the four steps of the process analysis phase of KoFI.

According to the KoFI methodology, the first step of the analysis is to identify the knowledge sources involved in the process. This includes the identification of all those sources of information or knowledge that could be being used or could be useful for performing the different activities composing the processes. Those sources could include the people consulted by the personnel in charge of the process, such as their colleagues, external consultants or other experts; the information systems supporting the process, such as the intranet, simulation tools, etc.; or documents, such as memos, reports, tutorials, tools' user manuals, etc.

The second step focuses on the identification of the main knowledge topics or areas related to the activities performed in the process. For instance, knowledge required to perform the activities, or created from them. It is important to identify and classify the knowledge related to the sources found in the preceding step. An important result of this step might be the identification of important knowledge topics not stored anywhere, or that might be stored in sources not used or difficult to find. In short the identification of possible loss and misuse of knowledge.

These first two steps include the classification of the sources and topics found, which can be made through the definition of a taxonomy and an ontology of knowledge sources and topics. In fact, knowledge taxonomies are considered an important

initial activity towards the development of KM systems [12]. The ontology should make possible to relate the different sources to the knowledge that can be obtained from them, and vice versa, i.e. relate the knowledge to the sources from where it can be obtained, or where it is stored.

The third step focuses on identifying the manner in which knowledge is flowing through the process. To accomplish this, it is required to analyze the relationships between the knowledge sources and topics, to the activities of the process. This includes the identification of the activities where the topics and sources of knowledge are being generated, modified, or used. It is important to identify knowledge dependencies, such as knowledge topics generated in an activity and required in other; and knowledge transfers mechanisms, such as knowledge transferred from one activity to another through a document, or through an interaction between different roles or persons. This type of analysis can be useful to identify three main issues related to knowledge flow:

1. *important knowledge flow enablers*, that means, channels, sources, or information systems being used to facilitate knowledge flow;
2. *knowledge flow bottle necks*, that means, situations that could be negatively affecting the flow of knowledge; and
3. *knowledge that could be not flowing at all*, for instance, knowledge that is being lost because it is not stored anywhere, or knowledge not used because people ignore it exists.

Finally, the fourth step of the analysis consists of identifying and classifying the main types of problems detected and which affect the knowledge flow. KoFI proposes to do this by defining problem scenarios [14], a technique based on explaining a problem in the form of a story describing a common situation. Once described this common problem, one or more alternative scenarios are also proposed in order to illustrate the manner in which such a problem could be addressed. Those alternative scenarios would be latter useful to extract the main requirements to propose the KM strategy to follow, or the KM system to develop.

In the following subsections we describe the manner in which the four analysis steps of KoFI were carried out in the manufacturing company.

3.2 Identifying Knowledge Sources

In the first step of the analysis, the identified sources were very diverse. The identification of the sources was done through the interviews performed to the personnel of the company, and the analysis of documentation and information systems. To facilitate its management, and following the recommendations of the KoFI methodology, once the different sources were identified, we proceeded to classify them. To do this a taxonomy of knowledge sources was defined, which included four categories of sources:

1. **Documents**, group of all those sources which consist of physical or electronic documents. It includes three subcategories: a) *processes' documents*, grouping all the documents related to the processes followed in the studied company's area, b) *technical documents*, referring to specialized documents with information of the

tools and machines used in the process studied, and c) *organizational documents*, consisting of documentation of the organizations life and culture, such as organizational rules, or norms.

2. **Information systems**, refer to the sources consisting of information systems used in the company. This category includes two subcategories: a) *query systems*, consisting of all the systems used to search for information, and b) *transactional systems*, which refer to transactional applications used in the company.
3. **People**, groups all the different types of people involved in the process. It has been divided in four subcategories: a) *staff* which groups all the people working in the studied Company's area, b) *specialists* refer to people with specific knowledge who might be consulted by the staff, c) *external clients* to represent the final clients of the process, and 4) *internal clients* to refer to users or clients of the process whom work in other areas of the company.
4. **Others**, is a category used to group those sources not included in the preceding categories. Particularly it includes two subcategories: a) *problem analysis* which are tools used by the process' participants to analyze and solve problems, and b) *simulation tools*, which are tools available in the company to support the simulation of processes offline.

Each source was described by assigning it a unique identifier, a name, a description, its type and category, its location, its format, and the main knowledge topics which could be obtained from it. This last was useful for the following step which was the identification of the knowledge topics involved in the process, and to start relating the different knowledge sources to the knowledge that can be obtained from them.

3.3 Identifying Knowledge Topics

The identified knowledge topics were also very diverse, ranging from organizational behavior to special machine maintenance. In this step we did not focused on describing each topic in detail, but on identifying the main knowledge required in the process. The topics identified were classified, according to the utility of such knowledge in the activities of the process, in three categories:

1. **Product line activities** which includes knowledge about the operation of machines, about processes, and about quality of the processes and products. It is divided into four subcategories: a) *product quality*, consisting of knowledge about the specifications and attributes of products, and the inspection process; b) *machine maintenance*, consisting of knowledge about the applications and procedures for conducting machines' maintenance (preventive, corrective or predictive maintenance), and knowledge about the spare parts of machines; c) *operation*, which includes knowledge related to the management and operation of machines, equipment and tools used in the process; and d) *information technology (IT) application*, which consists of knowledge about the software systems for consulting and registering information about machine operation.
2. **Organizational culture**, is all that knowledge that employees must have about the company, its internal organization and norms, etc. It includes only one subcategory which is knowledge of the company.

3. **General knowledge** is a category defined to group all those topics and areas of knowledge that the employees might have, and which are not directly related to the process operation. It is subdivided into four subcategories: a) *resource management*, related to knowledge about the internal control procedures; b) *IT management*, refers to knowledge about the use of tools, and systems available; c) *personnel management*, groups knowledge and skills for managing people, such as leadership, personnel coordination, etc.; and d) *other individual knowledge* is where all those individual knowledge and skills such as writing skills, foreign languages knowledge, etc. are grouped.

Once identified, the main knowledge topics were described assigning them a unique identifier, a name, a description, its classification, and information to know where such topic could be useful, and why and how knowing it could benefit the organization or the person who knows about it. With the knowledge topics descriptions, a knowledge dictionary was developed for the process.

3.4 Identifying Knowledge Flows

In this step we modeled the knowledge required in each activity of the process, the knowledge that each role needs to perform these activities, and the knowledge sources consulted or generated in each activity, following an adaptation of the Rich Picture technique [9]. Figure 2 presents an example of this type of diagrams, in which the knowledge required in the "Lift trucks operation and management" process are represented. The figure shows the role in charge of such activity, the experience, skills and knowledge it provides to the activity, and the main source of knowledge used in the activity, which is an application for managing security rules and regulation of the company.

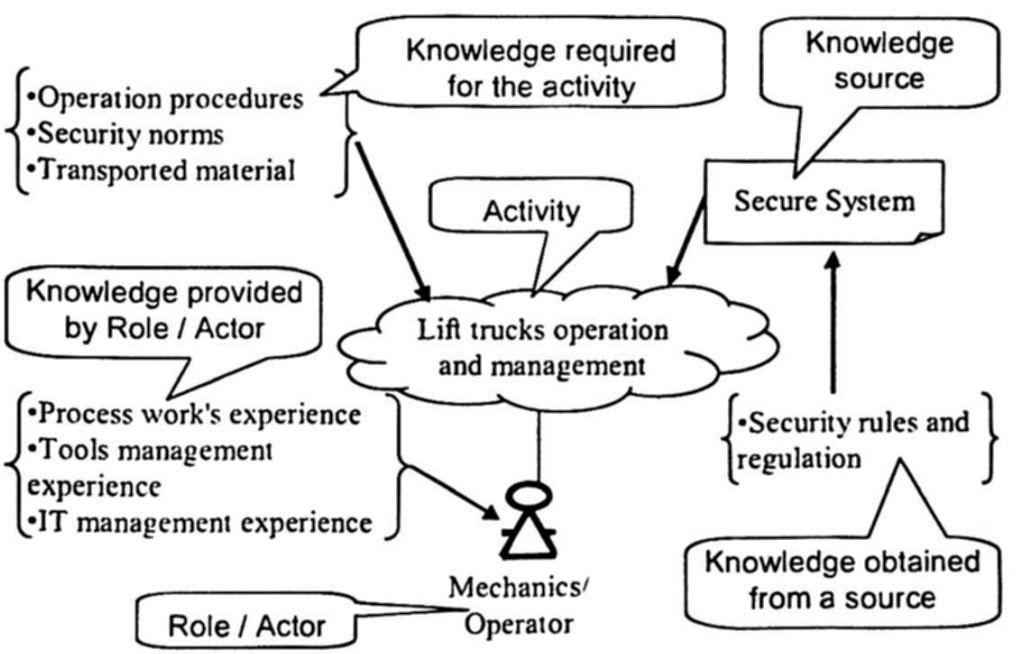


Fig. 2. Example of an adapted rich picture to analyze knowledge flows.

This type of models helped us to identify the relationships between the knowledge sources and topics, and the activities of the process. The above allowed us to create a knowledge meta-model (described in Section 4), which was used as the structure for developing a Knowledge Map useful to identify the knowledge that might be obtained from each source, and the activities in which the sources or the knowledge were being used or generated. This map was used in the construction of a Knowledge Portal (described also in Section 4). This portal was proposed to solve some of the main knowledge flow problems observed, as it is described next.

3.5 Identifying Knowledge Flow Problems

The final step of KoFI proposes to identify and classify the main problems affecting the knowledge flow in order to propose alternatives to minimize or avoid them. In our study, it was observed that some areas of the process were not well supported with documentation. For instance, there was not enough documented information on the use of certain mechanical and electrical tools; therefore, that knowledge resided only in people's experience. An additional problem was the identification of important knowledge sources that were not being used. Some reasons for the last were the difficulty for consulting some of the existent sources, either because they were unknown to the potential users, or because they were difficult to find by employees.

To address this problem, it was decided to develop a Knowledge Portal to facilitate the access to all the available sources, according to the areas, processes, or activities for which they are useful. Additionally, the portal would provide ways for pointing out to all those knowledge areas for which no sources exist. The last should be useful to identify all those areas for which knowledge sources should be created. Moreover, providing means for including those new sources easily was also a requirement for the portal. Additionally, it was also decided that the portal should provide access not only to documents, but also to other types of sources, such as information systems, or support tools, in order to promote the use of all the available types of knowledge sources of the company.

4 Design of the Knowledge Portal

In this section we describe the next: 1) a meta-model developed for structuring the knowledge map used into the portal, 2) the structure of such portal, and 3) the design of its user interface.

4.1 Meta-Model

The proposed meta-model, represented in Figure 3, comprises the knowledge types and sources involved in the knowledge generation and acquisition process. To develop it, we adapted a general knowledge sources and topics meta-model proposed as part of the KoFI methodology.

In the meta-model the knowledge concepts are integrated with the knowledge topics and sources. The knowledge concepts are required, generated or modified by the activities within the studied area, which are described as work definitions. In turn, these work definitions can be represented as processes, activities or decisions. Each knowledge concept/source association contains information about the knowledge level it requires. Finally, the available format and location for consulting each source are specified.

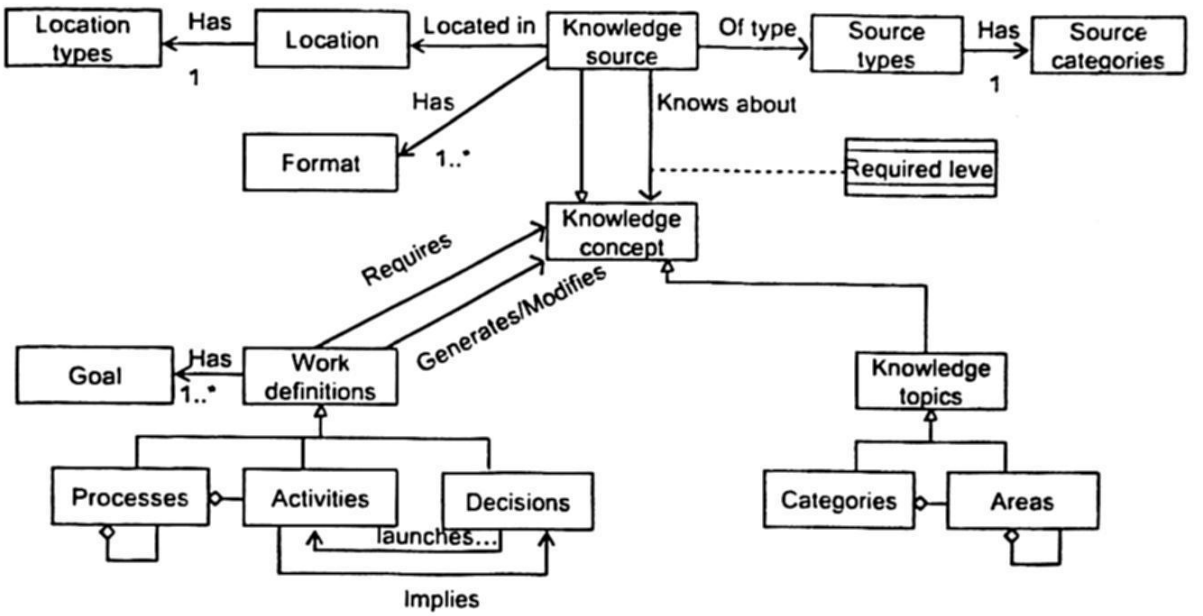


Fig. 3. Meta-model of knowledge types and sources.

4.2 Knowledge Portal Structure

The meta-model was used as a base to design the structure of the knowledge portal. Figure 4 shows the resulting general structure of the portal. This structure comprises a first level in which initial interfaces (pages) are accessible (e.g. home and registration pages). The second and third levels are pages which correspond to the manufacturing areas and sub-areas of the organization, respectively, according to the rich picture models developed during the analysis. The fourth level corresponds to pages on the processes that integrate each of the sub-areas identified from the involved knowledge flows. Finally, the fifth level presents all the identified knowledge sources for the specific process of the sub-area. This structure is representative of all and each of the manufacturing sub-areas, as identified during the analysis.

4.3 Knowledge Portal UI Design

The design of presentation and navigational features of the user interfaces (pages) also emerged from insights identified in the analysis and initial phases of design.

These include information about the identified knowledge flows, the main sub-areas of the organization, and the structure of the portal previously identified, which

resulted in the options included in the menus and main layout sections of the pages. These allow users to find the required information by simply identifying the specific area in which information is generated or required, and following the resulting navigational structure (area → sub-area → process) to locate the specific knowledge source, instead of just alphabetically (or randomly) browsing through the information.

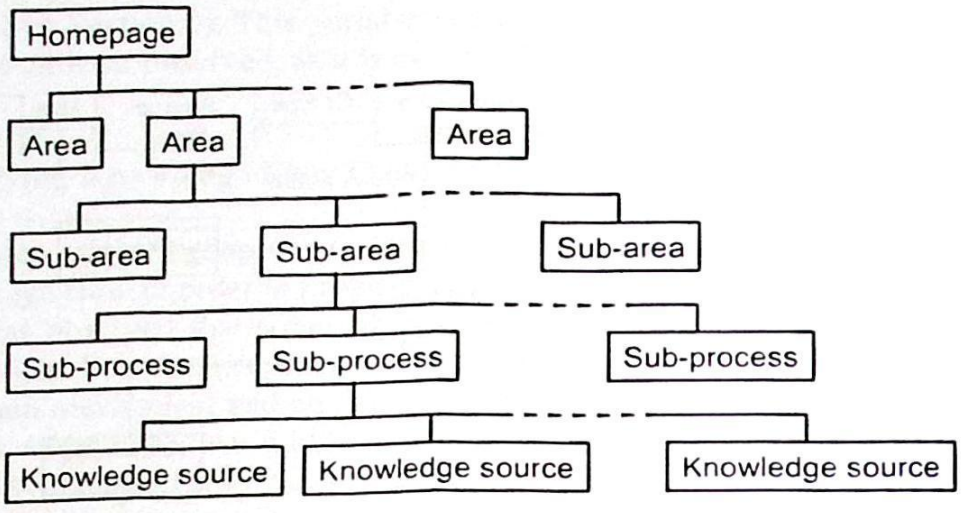


Fig. 4. General structure of the Knowledge Portal.

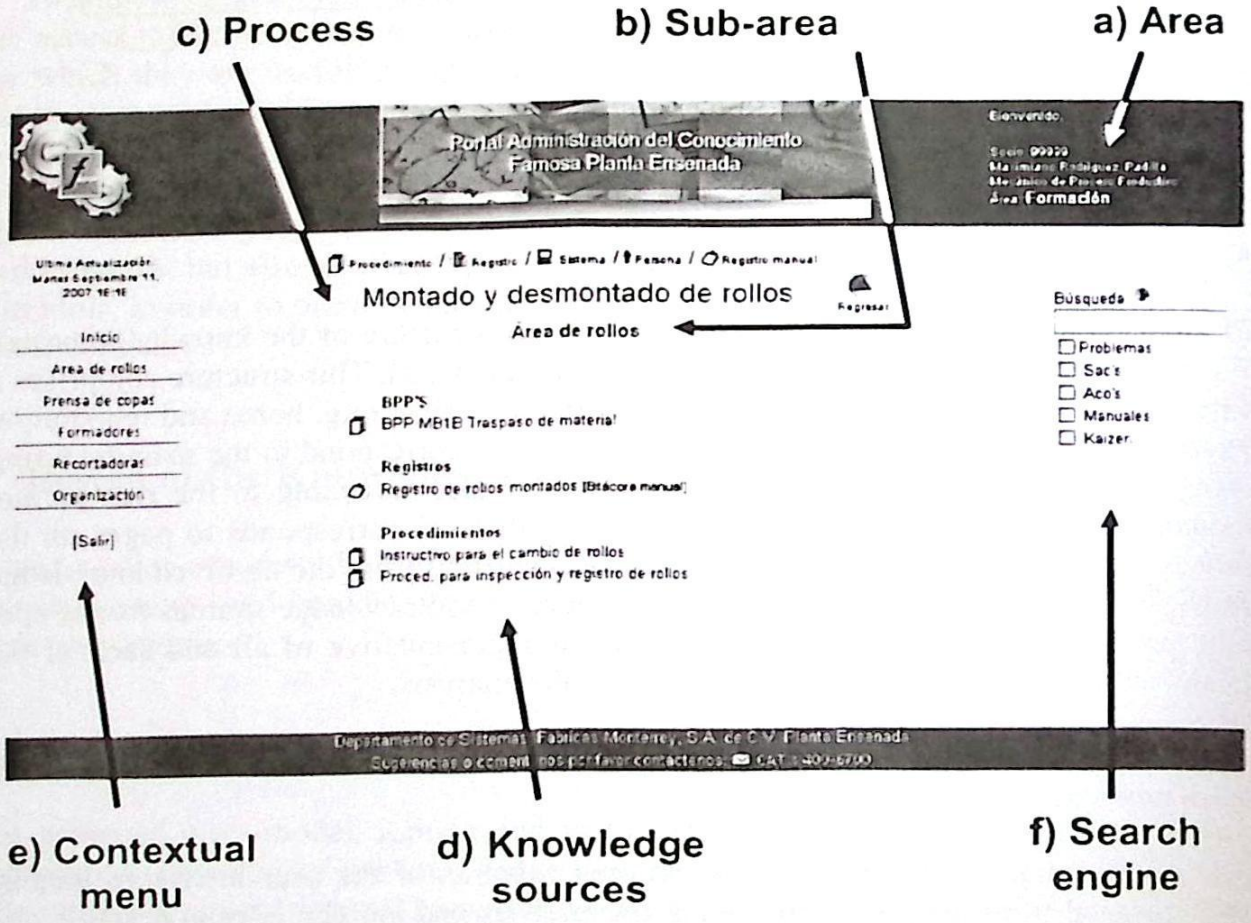


Fig. 5. Example of the page contents and layout of the Knowledge Portal.

Figure 5 depicts an example of the layout and content of a page from the current prototype for the "Formation" area.

The information provided includes the name of the manufacturing area being consulted (5.a), the name of the specific sub-area (5.b), the name of the selected process within the sub-area (5.c), and most importantly, links to knowledge sources (and types) available for that process (5.d).

Additionally, the page includes a "contextual" sub-area menu to facilitate navigation through the information (5.e), which is always available while the user stays in that particular sub-area of the portal. Also, it includes access to a search engine (5.f) which allows a search to be performed by simply specifying a keyword on the required topic, and optionally, the "places" in which the information should be searched for.

The interface in Figure 5 represents the final destination for users looking for a particular knowledge source whom, by following only three links (area → sub-area → process), arrive at the knowledge sources (either documents, systems or people) required to perform their intended activities.

Finally, this design adheres to the organization's established standard guidelines for this kind of applications.

5 Evaluation of the Knowledge Portal

We conducted a preliminary evaluation in one of the production areas to both determine the impact and acceptance level of the users on the system, and to provide support for the decision-making process concerned with the continuation of the system's implementation in other areas of the organization. The evaluation considered aspects concerning perception of usefulness and ease of use [4].

The evaluation consisted of

1. **an introductory session**, in which the system was presented to the users, and its functionality demonstrated to them. This included examples on how to search for and retrieve knowledge sources by means of navigating through areas, sub-areas and processes, as well as through the search engine; and
2. **the application of a questionnaire** containing 12 questions referring to perception of usefulness (6) and ease of use (6). Each evaluation session (induction and application of the questionnaire) was done in approximately one hour.

The subjects of the study were 41 employees of the "Formation" area for which the prototype was developed. The subjects included leader mechanics, process mechanics, operators and process engineers, whose participation was voluntary. The sample was divided into 4 groups according to the natural operative processes (3 groups of ten people and 1 of eleven). The application process of the evaluation was completed in three days.

5.1 Analysis and Discussion of Evaluation Results

The subjects had positive appreciations with regard to the knowledge portal, as it is reflected in their answers in the questionnaire. Figure 6 shows the answers to the

questions about the perception of usefulness of the tool. The users perceived that the portal would allow them to increase their productivity and to perform their tasks more easily (82.93% "Agree" in both cases), although some of them had doubts regarding the fact that this would increase their productivity (24.39% "Have Doubts"). Only one person (2.44%) "Disagreed" that the tool would help him/her to complete his/her tasks faster.

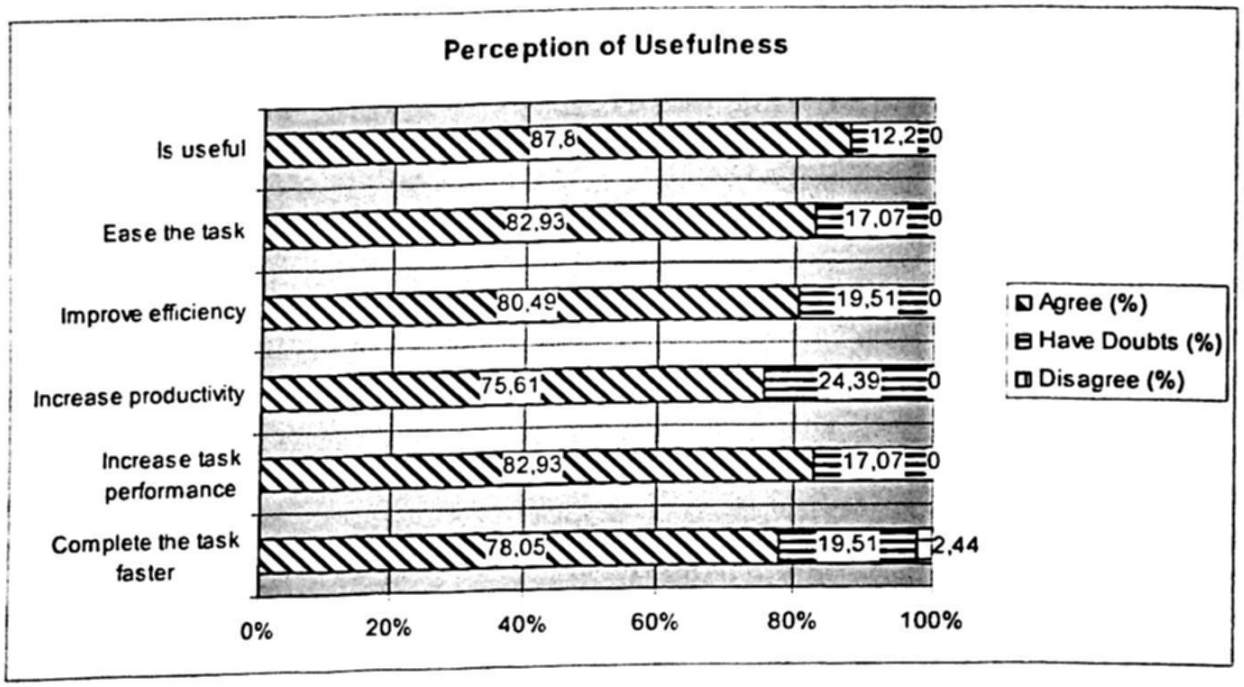


Fig. 6. Perception of Usefulness.

Figure 7 shows the answers to the questions about the perception of ease of use. As can be seen, although most of the users perceived that it was easy to learn to browse through the information (85.37% "Agree"), some had doubts concerning the ease of finding information (39.02% "Have Doubts"), and even more users had doubts concerning becoming experts on the use of the tool (46.34% "Have Doubts"). A possible explanation could be that a little more than a third of the users had doubts concerning the clarity of the presented interfaces, as well as about the interaction flexibility that these provide (34.15% in both cases).

In general, most of the users considered the knowledge portal as a useful (87.80% "Agree" – Figure 6) and easy to use tool (68.29% "Agree" – Figure 7) for the accomplishment of their work.

5.2 Additional Comments to the Evaluation

During the evaluation, we observed that results seem to be related to how much participants use information systems in their daily work. People with more experience using information systems, and particularly internet based systems such as web portals, were more positive about the knowledge portal. Therefore, it is recommended that for constructing the final version of the portal, to consider this situation in order

to identify how it can influence the final acceptance of the portal, particularly within the people who do not use computers regularly in the company.

Finally, formal evaluation of KM systems is a difficult concern since results emerge only in the long time, and are influenced by many external factors. Thus, it was out of the scope of this work to formally evaluate the usefulness of the knowledge portal. However, we expect to continue the evaluation during practice, when the final version of the portal is completed and in use in the company.

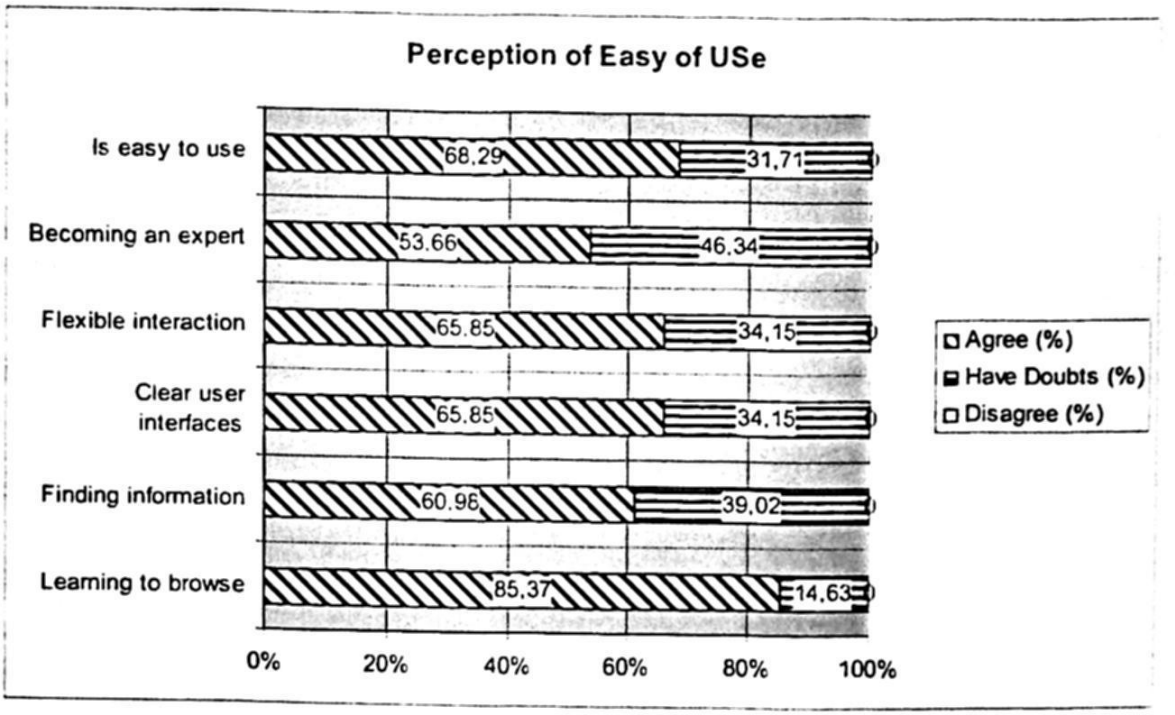


Fig. 7. Perception of Easy of Use.

6 Discussion and Lessons Learned

The KoFI methodology was initially developed to aid in the design of KM approaches to improve software processes. In this initial application domain, the methodology was useful to propose the design of a KM tool, and to structure and create a knowledge map of the studied process. It can be argued that software processes differ from manufacturing processes in that they have different knowledge requirements. In fact, software processes have been considered to be more knowledge intensive and dynamic than common industrial processes [13], such as the one studied here (the production of aluminum cans). Nevertheless, we have learned some important lessons from this study comparing it to the other studies we have done [14-17].

In the previous studies, we analyzed two software maintenance processes to identify knowledge management needs. These processes were not formally defined, so there was not a model of such processes. Instead, we required to construct the entire model of the processes from the interviews we made to the people responsible of such processes.

In the present study the processes were much more formally defined and documented. Nevertheless, it was decided to create models of the process to explicitly represent the knowledge sources and topics involved in it (as we exemplified in Section 3.3). The reason for this was that the process models were made with a common business process modeling language, which has not explicit representation of knowledge related issues. From the models we made, we were able to identify knowledge requirements and sources, which were not identified from the existent process models of the company. This observation has given us insights to argue that independently of how well defined and documented the process could be, if there is not an explicit representation of the knowledge and sources involved in the activities of the process, important sources and knowledge requirements could be lost or ignored during the analysis. Thus, in this case, the steps proposed by the methodology, provided a way for identifying important knowledge requirements and problems that were not identified before. Even though the process was well defined and documented from a manufacturing (or business) point of view.

Other important issue was that the strategies proposed as a result of applying KoFI in the different studies had different requirements. In the first studies (software engineering domain), we observed that the KM systems required to be more active than a traditional KM tool. It was observed that software engineers did not have enough time to capture or search for knowledge, neither for using new tools totally apart from the development environment they were working with. Thus, the systems required to be capable of performing some KM tasks in an autonomous way, and be integrated to the current tools being used by the engineers. In the present study we did not found this situation. However, an important issue was observed. In both types of studies (the previous ones, and this one), regardless of the type of KM approach required, a first step towards the development of such approaches was the definition of a knowledge map of the process, which consisted of the identification and definition of the main knowledge topic required in the process, the main knowledge sources, its relationships between them, and the activities performed. All this information was obtained during the application of the methodology. Form this observation we argue that creating a knowledge map of the process, in which the knowledge sources and topics are classified, and its relationships identified, should be an initial step, independently of which type of KM tools or strategies will be proposed.

Finally, we want to highlight the fact that, although the analysis of knowledge flow support tools was not a main concern during the present study, some important information systems being used as knowledge flow enablers were identified and considered into the portal. For instance, some reports and simulations that are obtained from external systems are accessible through the knowledge portal. To do this, the portal has some connections to other applications available in the company. Thus, this study helped us to confirm that the methodology is helpful in identifying knowledge flow support tools available for the process, in order to consider them as part of the KM proposals.

7 Related Work

Integrating KM into work processes is one of the main concerns of the KM community [18]. However, most of the organizational KM strategies are not well integrated to organizational processes [8]. In order to reduce this gap, some works have been proposed in literature. Perhaps the work most related to our own, is that of Kim et al. [7], whom have studied knowledge flows in a manufacturing firm through their own method including a special type of knowledge flow diagram. In their study, Kim et al. demonstrate that knowledge flow analysis can be a means towards understanding the relationships between the knowledge and the process studied. The main difference between KoFI and the proposal of Kim et al. is that the former was developed to be more practical, being not only a way to analyze a process, but also a means through which to propose practical solutions.

Other authors have also proposed approaches for modeling knowledge flows, or knowledge involved in work processes, i.e. [1, 11, 19]. However, most of the approaches we have found are either orientated towards developing specific KM systems, or require special tools or process modeling languages for their usage. Before proposing a specific approach for managing knowledge in an organization, it is important to analyze the organizations' work processes from a knowledge flow perspective [10], since supporting knowledge flow should be the main focus of KM [2]. Additionally, to have a successful integration of KM strategies into the organizational work process, we must consider the current technological infrastructure [3], which is an issue not addressed in the other works we have found in the literature.

Thus, the main contribution of our work is to use an approach which explicitly takes this observation into consideration. That means, we have not only cover most of which is covered by previous studies, we also have taking as an important concern the identification of the current technical infrastructure, in order to include such infrastructure as part of the KM strategies or systems proposed, and perhaps as the basis of them. We have illustrated how the KoFI methodology was used for proposing means to improve the knowledge flow in a manufacturing company. This should be accomplished not only by developing new systems, changing organizational culture, and so on, but also by integrating the current infrastructure and the actual work being done by the people in charge of the organizational processes.

8 Conclusion

In this paper we have illustrated the use of the KoFI methodology to analyze a manufacturing process in order to improve the flow of knowledge in it. The main result of the study was illustrating the usefulness of the KoFI methodology in a manufacturing setting; particularly for the design of a knowledge portal based on the real work structure of a company. Since KoFI was initially developed to be used to analyze software processes, this study has provided us with the initial evidence to argue that KoFI is open enough to aid in the design and construction of different types of KM approaches, and in different application domains. However, more case studies are required to continue evaluating the benefits and limitations of KoFI in different settings.

Other important result of this study is the knowledge portal per se, which integrates the knowledge sources available, and presents them to the users by following an organizational structure which emerges from the application of the different steps proposed by KoFI. Unfortunately, evaluating if the portal will allow the company to improve the training of highly competitive personnel and to promote organizational learning can only be made after a long period of time using the portal in real practice. However, the preliminary evaluation of this portal has led us to believe that it could help to accomplish this, since such a portal was considered to be highly useful and ease to use by the employees of the company. As future work, we are planning to apply the KoFI methodology to the analysis of all the remaining company's processes, in order to extend the use of the portal to the entire organization. This should help us to continue evaluating the benefits and limitations of KoFI and the portal.

Acknowledgments

This work is partially supported by UABC, project 0191 of the XI Convocatoria Interna de Proyectos, the MELISA project (grant PAC08-0142-3315) financed by the Consejería de Educación y Ciencias de la Junta de Comunidades de Castilla-La Mancha, Spain, and CALIPSO (TIN20005-24055-E) supported by the Ministerio de Educación y Ciencia, Spain. The authors acknowledge the support provided by FAMOSA-Ensenada, for the realization of this project.

References

1. P. Bera, D. Nevo, and Y. Wand, "Unravelling Knowledge Requirements through Business Process Analysis," Communications of the Association for Information Systems, vol. 16, pp. 814-830, 2005.
2. U. M. Borghoff and R. Pareschi, "Information Technology for Knowledge Management," Journal of Universal Computer Science, vol. 3, pp. 835-842, 1997.
3. T. H. Davenport, "Information Technologies for Knowledge Management," in Knowledge Creation and Management: New Challenges for Managers, K. Ichijo and I. Nonaka, Eds. New York, NY.: Oxford University Press, 2007, pp. 97-117.
4. F. D. Davis, "Perveived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," MIS Quarterly, vol. 13, pp. 319-340, 1989.
5. G. Gkotsis, C. Evangelou, N. Karacapilidis, and M. Tzagarakis, "Building Collaborative Knowledge-based Systems: A Web Engineering Approach," presented at IADIS International Conference WWW/Internet, Murcia, Spain, 2006.
6. K. Ichijo and I. Nonaka, "Knowledge Creation and Management: New Challenges for Managers." New York, NY.: Oxford University Press, 2007, pp. 335.
7. S. Kim, H. Hwang, and E. Suh, "A Process-based Approach to Knowledge Flow Analysis: A Case Study of a manufacturing Firm," Knowledge and Process Management, vol. 10, pp. 260-276, 2003.
8. R. Maier and U. Remus, "Defining Process-oriented Knowledge Management Strategies," Knowledge and Process Management, vol. 9, pp. 103-118, 2002.
9. A. Monk and S. Howard, "The Rich Picture: A Tool for Reasoning About Work Context," Interactions, vol. 5, pp. 21-30, 1998.
10. M. E. Nissen, "An Extended Model of Knowledge-Flow Dynamics," Communications of the Association for Information Systems, vol. 8, pp. 251-266, 2002.

11. G. Papavassiliou and G. Mentzas, "Knowledge modelling in weakly-structured business processes," Journal of Knowledge Management, vol. 7, pp. 18-33, 2003.
12. M. Rao, "Knowledge Management Tools and Techniques: Practitioners and Experts Evaluate KM Solutions." Amsterdam: Elsevier, 2005, pp. 438.
13. P. N. Robillard, "The Role of Knowledge in Software Development," Communications of the ACM, vol. 42, pp. 87-92, 1999.
14. O. M. Rodríguez-Elias, A. I. Martínez-García, A. Vizcaíno, J. Favela, and M. Piattini, "Identifying Knowledge Flows in Communities of Practice," in Knowledge Management: Concepts, Methodologies, Tools, and Applications, vol. 2, M. E. Jennex, Ed. Hershey, PA, USA: Idea Group Press, 2007, pp. 841-849.
15. O. M. Rodríguez-Elias, A. I. Martínez-García, A. Vizcaíno, J. Favela, and M. Piattini, "Organización de conocimientos en procesos de ingeniería de software por medio de modelado de procesos: una adaptación de SPEM," presented at VI Jornada Iberoamericana de Ingeniería del Software e Ingeniería del Conocimiento (JIISIC'07), Lima, Perú, 2007.
16. O. M. Rodríguez-Elias, A. I. Martínez-García, A. Vizcaíno, J. Favela, and M. Piattini, "A Framework to Analyze Information Systems as Knowledge Flow Facilitators," Information and Software Technology, vol. 50, pp. 481-498, 2008.
17. O. M. Rodríguez-Elias, A. I. Martínez García, J. Favela, A. Vizcaíno, and J. P. Soto, "Knowledge flow analysis to identify knowledge needs for the design of knowledge management systems and strategies: a methodological approach," presented at 9th International Conference on Enterprise Information Systems (ICEIS): special session on Business Intelligence, Knowledge Management and Knowledge Management Systems, Funchal, Madeira - Portugal, 2007.
18. W. Scholl, C. König, B. Meyer, and P. Heisig, "The future of knowledge management: an international delphi study," Journal of Knowledge Management, vol. 8, pp. 19-35, 2004.
19. M. Strohmaier and K. Tochtermann, "B-KIDE: A Framework and a Tool for Business Process-Oriented Knowledge Infrastructure Development," Journal of Knowledge and Process Management, vol. 12, pp. 171-189, 2005.

Modelling Regulated Social Spaces for Groupware Applications

Carmen Mezura-Godoy and Luis Gerardo Montané-Jiménez

University of Veracruz,
Av. Xalapa, Esq. Manuel Avila Camacho,
CP 91020, Xalapa Ver.
cmezura@uv.mx
lmontane@uv.mx

Abstract. Social regulation is an intrinsic aspect of collaborative work, although it is not considered in the most of cases in groupware applications. Regulation concerns the establishment of working rules and their negotiation. The MARS regulation model enables users to define their common workspace -arena, actors participating in the activity, roles assigned to the actors, interactions between actors and rules governing all the objects in the arena and also in several ones. This paper proposes a service to regulate groupware applications. The service implements a language to describe social aspects. In order to validate our approach, a prototype supporting the MARS model has been implemented.

1 Introduction

In Sociology, *social regulation* is the process which enables groups to create and modify social rules controlling their individual and collective actions[5]. Regulation enables people to describe how they wish to take part in the activity and the minimal conditions of work's execution: in other words, it enables people to create, negotiate and apply rules controlling their collaborative activity. Regulation does not imply a complete description of an activity, but rather a definition of minimal rules, and personal rights and responsibilities, in order to improve the group activity. Regulation inside a group changes during the activity. The rules described at the begining of the activity can be modified like a natural process of evolution of the activity; so, the work rules change along with the activity.

Activities group can be for instance: (a) communication with each other in order to exchange their ideas or points of view, (b) sharing information or their workspace and (c) coordinationof their activity (time, space and resources). Nevetheless, we can observe that people collaborate in several activities at the same time. For instance, members of a research team take part in meetings, in projects, in paper writing, etc., so people take part in several arenas.

Nowadays, groupware tools supporting collaborative work enable users to communicate, coordinate and cooperate in specific tasks [17, 8, 2, 13]. Nevertheless, these tools rarely incorporate social activity aspects. Moreover, they do not enable users to regulate their activity.

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 47-60

Received 02/07/08
Accepted 01/08/08
Final version 11/09/08

We consider important to include a model activity in groupware applications. So, it was proposed a new multi-arena regulation model [15, 14]. This model enables to describe activities in a single workspace and also in several ones.

On this paper we propose a regulation service for groupware tools. It enables developers of groupware applications to create new regulated applications from simple ones. This regulation service enables: to model several arenas and to execute them according to the regulation. In order to validate our model we implemented a software prototype. The prototype illustrates how a groupware application can be regulated, how end-users can describe and modify the regulation during the activity and how the service applies the regulation.

The rest of the paper is organized as follows. Section 2 surveys related work to establish the context of our research. Section 3 describes the MARS regulation model and its associated language. Section 4 presents a regulation service for groupware applications. Section 5 introduces our experimental software prototype. Finally section 6 summarizes the contributions of this paper and introduce some research directions.

2 Related Work

Differents works offer infrastructures allowing the introduction of some social aspects in groupware tools, under the term "coordination policies" *e.g.* [7,4, 11,12,16,10]. These policies enable to coordinate an activity in terms of access control to the production groupware space (private or shared workspace).

Even though, these systems offer the means to incorporate social aspects of an activity work (participants, roles, policies, rights). However, social aspects are defined at a low level of abstraction (coordination language). In order to enable users to define their activity, we believe that it is necessary to provide them with more powerful mechanisms to facilitate the definition of social aspects of collaborative activities.

Moreover, the models proposed by these systems focus only coordination aspects, rather than on the complete collaborative activity. For instance, the role concept is used to control access to shared resources (production space). For us, role is a collaboration concept: an actor in the activity has a particular role in the group (*e.g.* leader, coordinator, etc.).

In the works previously cited social aspects are limited to activity coordination. We believe that a model for collaborative activities must allow the definition of the activity and its context, which means enabling users to define group members taking part in the activity (their duties, rights and preferences), their social role, tools to facilitate doing their activity, and naturally groupwork rules. Including a complete model of the activity in groupware can aid end-users to use them better and adopt them more easily.

Works like, Locales [6], CooLDA [1], SeeMe [9] as well as participation model (PM) [3] propose activity models for groupware applications. These models are founded in works in social sciences and they offer concepts enabling users to define a group activity in a workspace. This workspace called, a "locale"(Worlds),

"support of activity"(CooLDA) or "arena"(PM), represents the group's activity and its context, *i.e* it has the activity components and it establishes necessary conditions for the activity execution. On the other hand the purpose of SeeMe is to support the early phases of developing concepts for socio-technical solutions and to document them with diagrams.

All these works take into account the evolutionary aspect of a group activity. Worlds and CooLDA are based on reflexive models to enable users to modify the activity model in runtime. CooLDA and PM consider redefinition of the activity as part of the activity itself.

An activity model must also make it possible to improve the design of groupware applications and, consequently allow users to better use them. Nevertheless, Worlds and CooLDA do not consider social aspects of activities, like people engagement in the activity and social work rules as PM does. Even though, the PM offers a social regulation model of collaborative activity, none of them enables the idea of "reusing" defined spaces in order to create more complex collaborative spaces.

3 The MARS regulation model and associated language

In this section we present, the MARS *Multi-Arena Regulation model* [15]. This model allows one to represent regulated group activities supported by groupware tools. These regulated group activities can be carried out by members of a group inside a collaborative space or in several ones.

3.1 Elementary concepts

A group activity is defined by interactions taking place in a collaborative space called *arena*. The users executing interactions are called actors. An *actor* represents a person, a software agent or a group. Throughout an activity, the actors handle and produce *objects*, such as documents, files, notes, etc. A family groups actors or objects having the same set of features (*e.g.* "writers", "readers", "books", or "papers" families). During the activity, actors and objects plays different *roles*, depending on the specific interaction they execute or take part. For instance during the "writing" interaction an actor plays the "writer" role, and the object handled in this interaction plays the "draft" role.

In order to regulate their activity, actors define scenarios for each interaction. A *scenario* describes how an interaction is carried out (operations or interactions that must be executed), who can participate in the interaction, and what objects can be handle. The scenarios represent the social protocols taking place in the arena.

In the following let $\mathcal{A}$, $\mathcal{O}$, $\mathcal{R}$, $\mathcal{S}$, $\mathcal{A}_{\mathcal{F}}$ and $\mathcal{O}_{\mathcal{F}}$, be the sets of *actors*, *objects*, *roles*, *scenarios*, *actors family* and *objects family* respectively, and α and β two functions returning respectively the family of an actor or an object.

3.2 Interaction, scenario and arena

An "interaction model" defines a regulated interaction inside the arena. It specifies all the families of actors and objects taking part in the interaction, roles attributable to actors and objects during the interaction, and the scenarios describing how the interaction can be carry out.

Definition 1 (Interaction model.)
An interaction model is a tuple $< n_I, E, A_f, O_f, R_s, S_s, \pi, \rho >$, *where* n_I *is the name of the interaction,* E *is a set of interaction states,* $A_f \subseteq \mathcal{A}_\mathcal{F}$, $O_f \subseteq \mathcal{O}_\mathcal{F}$, $R_s \subseteq \mathcal{R}$, $S_s \subseteq \mathcal{S}$, π *is a relation from* $A_f \rightarrow R_s$ *and* ρ *is a relation from* $O_f \rightarrow R_s$ □

Let us imagine the model for the interaction "to publish some document" defined as follows: < *to Publish*, {*active*, *finished*}, {*manager*, *collaborators*}, {*paper*, *document*, *image*}, {*publisher*, *published*}, {*scenario To Publish*}, {(*manager publisher*), (*collaborator*, *publisher*)}, {(*paper*, *published*)} >[1]. This model authorize "to Publish" interaction to "manager" and "collaborators", it limits the objects handled in this interaction to "papers", "images" and "documents", it defines the "publisher" role assigned to the "manager" and "collaborators" and the "published" role assigned to "papers", "images" and "documents". It specifies the "scenario To Publish" as the only scenario describing how this interaction can be carried out.

An "interaction", represents an interaction in execution. An interaction must always be according to an interaction model.

Definition 2 (Interaction.)
Given an interaction model $< n_I, E, A_f, O_f, R_s, S_s, \pi, \rho >$, *an interaction is a tuple* $< n_I, e, A, O, s, \sigma, \omega >$, *where,* n_I *is the name of the interaction,* $e \in E$, $A \subseteq \mathcal{A}$ *and* $\forall a \in A \alpha(a) \in A_f$, $O \subseteq \mathcal{O}$ *and* $\forall o \in O \beta(o) \in O_f$, $s \in S_s$, σ *is a relation from* $A \rightarrow R_s$, *and* ω *is a relation from* $O \rightarrow R_s$. □

An interaction representing the actor "carmen" publishing the "enc08" paper in the "writing" arena could be the following: < $toPublish_1$, *active*, *carmen*, *enc*08, $scenarioToPublish_1$,(*carmen*, *publisher*), (*enc*08, *published*) >.

A "scenario" describes how an interaction can be executed.

Definition 3 (Scenario.)
A scenario is a tuple $< n_S, Pre, Pos, S_s >$, *where* n_S *is the name of the scenario,* Pre *is a set of preconditions,* Pos *is a set of posconditions and* $S_s \subseteq \mathcal{S}$. □

A scenario for the interaction "to publish a paper" is defined as follows: < *scenarioToPublish*, {(*paper*, *finished*), (*dateOfPublication*< *deadline*)}, {(*paper*, *published*)} >. This scenario defines two preconditions. The first one evaluates the role of the paper, in this case it must be "finished" in order to be "published", and the second one evaluates the deadline. Finally, the poscondition estiblishes for the paper the role of "published".

[1] We identify in all our examples, actors, objects, roles, scenarios, interaction's models and interaction instances by their name.

An arena defines a group activity, actors, objects, interaction model and interactions.

Definition 4 (Arena.)
An arena is a tuple $< n_E, A_s, O_s, M_s, I_s >$, *where* n_E *is the name of the arena,* $A_s \in \mathcal{A}$, $O_s \in \mathcal{O}$, M_s *is a set of interaction models and* I_s *is a set of interactions.* □

An example of a writing arena is the following: $< writingArena, \{carmen, luis\}, \{enc08, image1, cscw07\}, \{toPublishIntM, toWriteIntM\}, \{toPublish_1, toWrite_2\} >$ where *writingArena* is the name of the arena, "carmen" and "luis" are the actors who can perform interactions in this arena, "enc08", "image1" and "cscw07" are tha accesibles documents (for publishing or writing interactions), "toPublishIntM" and "toWriteIntM" are the interaction models, the first one for publishing interaction and the second one for writing interaction, and finally "$toPublish_1$" and "$toWrite_2$" are the running interactions.

3.3 View and Complex Arena

A collaborative activity is defined inside an arena. Nevertheless, members of a work group collabortate with other groups or activities. So, people take part in several activities in different spaces, *i.e.* several arenas. For instance the members of a research team collaborate at the same time on projects, on writing articles or documents, on the organization of conferences or work meetings, etc. Each of these collaborative spaces is controlled by specific work rules.

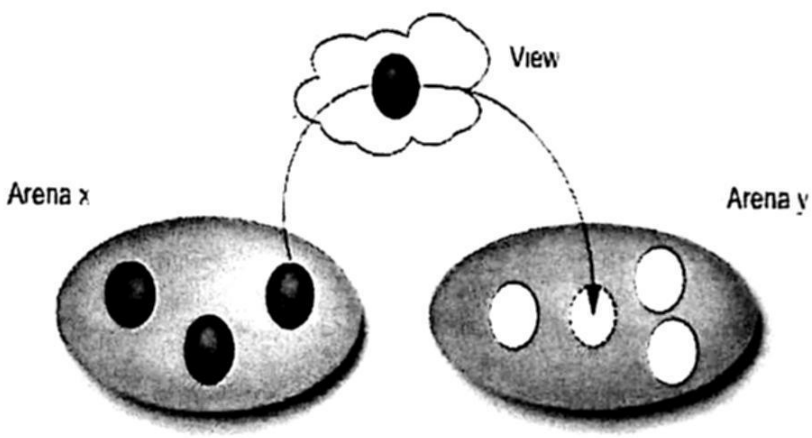


Fig. 1. Cooperation between arenas via the use of views.

A "view" defines, actors, objects and interactions that an arena can share with another (See Fig. 1).

Definition 5 (View.)
Given an arena $< n_E, A_s, O_s, M_s, I_s >$*, a view is a tuple* $< n_E, A_v, O_v, m_v >$*, where* n_E *is the name of the arena that produce the view,* $A_v \subseteq A_s$*,* $O_v \subseteq O_s$ *and* $m_v \subseteq M_s$*.* □

Let us imagine, that two arenas will cooperate to make their objects accessible to each other. The "libraryArena" produces a view with the objects it will share with the "writingArena". The view defined by libraryArena is: $< \{edgard, carmen, luis\}, \{LNCS2527, javaBeans, groupwareApplications\}, toBorrowIntM >$, where "toBorrowIntM" is the interaction that will be accessible from the writeArena, "edgard", "carmen" and "luis" are the actors with the possibility of borrowing a book from the libraryArena, "LNCS2527", "javaBeans" and "groupwareApplications" are the books from libraryArena accesible from the "writngArena".

Arenas importing remote objects from other arenas are called "complex arenas".

Definition 6 (Complex arena.)
Given an arena $< n_X, A_s, O_s, M_s, I_s >$ *and a view* $< n_Y, A_v, O_v, m_v >$*, a complex arena is a tuple* $< n_X, A_s \bigcup A_v, O_s \bigcup O_v, M_s \bigcup m_v, I_s >$*.* □

The "writingArena" is an example of a complex arena because it imported remote objects from the "libraryArena": $< writingArena, \{carmen, luis, edgard\}, \{enc08, cscw07, image1, image2, LNCS2527, javaBeans, groupwareApplication\}, \{toPublishIntM, toWriteIntM\}, \{toPublish_1, toWrite_2\} >$.

3.4 Arena and view operators

The arenas evolve according to the creation and the execution of interactions. For this reason, we defined operators which allow to manage arena and view objects. Each of these operators ensures the passage of an arena or a view from one coherent state to another, always respecting the arena regulation.

Table 1. Operators enabling the addition and deletion of actors, objects, interactions and model interactions to/from arenas.

Operator	Enter	Exit
addActor	(E, a)	$E' = (n, A \bigcup \{a\}, O, I, M)$
deleteActor	(E, a)	$E' = (n, A - \{a\}, O, I, M)$
addRobject	(E, o)	$E' = (n, A, O \bigcup \{o\}, I, M)$
deleteRobject	(E, o)	$E' = (n, A, O - \{o\}, I, M)$
addModelInt	(E, m)	$E' = n_E, A, O, I, M \bigcup \{m\})$
deleteModelInt	(E, m)	$E' = (n_E, A, O, I, M - \{m\})$
addInteraction	(E, i)	$E' = (n_E, A, O, I \bigcup i, M)$
addInteraction	(E, i)	$E' = (n_E, A, O, I - \{i\}, M)$

Table 1 summarizes the operations to allow adding and removing actors, objects, interaction models and interactions to/from an arena. For instance, for a given arena $E =< n, A, O, I >$ where: n, is the identifier of the arena, "A" is the set of actors, "O" the set of objects, "I" the set interactions and "M" the set of model interaction, the arena resulting after applying the operator "addActor(E,a)", is $E' =< n, A \bigcup \{a\}, O, I, M)$. The operation "addActor" verifies that the family of the actor to be added is defined in the arena.

We defined two operations to enable to cooperate arenas. These operators allow them to share their objects by export and import views.

Operator 1 (Export View.)
Given an arena $E =< n, A, O, I, M >$, *a set* $A_v \subseteq A$, *a set* $O_v \subseteq O$ *and a set* $M_v \subseteq M$, *the result of* $ExportView(n, A_v, O_v, M_v)$, *is the following view* $V =< n, A_v, O_v, M_v >$. □

Table 2. Operators enabling the addition and deletion of actors, objects and model interactions to/from views.

Operator	Enter	Exit
addViewActor	(V, a)	$V' = (n, A \bigcup \{a\}, O)$
deleteViewActor	(V, a)	$V' = (n, A - \{a\}, O)$
addViewRobject	(V, o)	$V' = (n, A, O \bigcup \{o\})$
deleteViewRobject	(V, o)	$V' = (n, A, O - \{o\})$
addViewModelInt	(V, m)	$V_M = (n, A, O, M_M \bigcup \{m\})$
deleteViewModelInt	(V, m)	$V_M = (n, A, O, M_M - \{m\})$

Table 2 summarizes the operations to allow adding and removing actors, objects and interaction models to/from a view. For instance, for a given view $V =< n, A, O, M >$, where: "n" is the identifier of the view, "A" is the set of actors to be exported, "o" is the set of objects and "M" is the set of interaction models, the view resulting after applying the operator "addViewActor(V,a)" is $V' =< n, A \bigcup \{a\}, O, M) >$

Operator 2 (Import View.)
Given an arena $E =< n_x, A, O, I, M >$, *and a view* $V = (n_y, A_v, O_v, M_v)$, *the result of* $ImportView(E, V)$ *is an arena* $E' =< n_x, A \bigcup A_v, O \bigcup O_v, I, M \bigcup M_v >$. □

Consider the following "writing arena": $writingArena =< \{carmen, luis\}, \{enc08paper\}, \{toWriteMInt, toReviewMInt\}, \{toWrite_1, toWrite_2\} >$. This arena defines an space for joint paper's writing. Inside this arena, "carmen" and "luis" can "write" and "review" a paper for the Enc2008. Now, we want to enable these actors to access the univertisty library, in order to ease the paper's documentation.

For that, the "libraryArena" exports a view (with actors, objects and interaction) to the "writing arena", as the following: $ExportView =< libraryArena, \{carmen, luis, edgard\}, \{cscw02, LNCS2517, criwg05\}, toBorrowIntMt >$. The operation $ImportView$ adds to the "writing arena" the objects of the library's view. The new "writing arena" is: $< writingArena, \{carmen, luis\ edgard\}, \{enc08paper, cscw02, LNSC2527, criwg05\}, \{toWriteMInt, toReview\ MInt, toBorrowMInt\}, \{toWrite_1, toWrite_2\} >$.

As we can see, in a multi-arena context, arenas have "local" and "remote" objects. So, in order to enable arenas handling with remote objects a propagation of operators is then necessary. We identified several constraints that must be satisfied during the application of the operators in order to ensure coherence between arenas [14].

3.5 Collaborative Regulation Language: CoRaL

In order to describe scenarios with the MARS model a language named CoRaL was proposed. Remember that a scenario defines how an interaction can be execute, it specifies through pre and pos conditions: i) who can participate in the interaction, ii) what objetcs can be manipulated, iii) what role has an actor or object during the interaction and in same cases iv) has references to another scenarios.

Let be E, the set of all strings representing names of: arenas, actor, families of actors and objects, roles and interactions. In the following we describe the syntax of CoRaL lenguage in BNF notation.

<statement> ::= <expression> <operator> <expession> ";"
<expression>::= <keyword> ":{"<elements> "}"
<expression>::= ! <keyword> ":{"<elements>"}"
<operator>::= :: | ->
<keyword>::= "Arena" | "Interaction" | "Actor" | "Actor family" |
"Object family" | "Object" | "Role"
<element>::= any string on E

In CoRaL a "statement" describes pre and pos conditions. A "statement" is formed by two expressions and an operator. The operator "::" is used to represent preconditions and the operator "->" is used to represent posconditions. An "expresion" is formed by a "keyword" and "element", where the "keyword" is a string identifing a component of the model (*e.g.* Arena, Interaction or Actor) and the "element" is a string representing an identifier of E elements (*e.g.* "to Publish" or "carmen").

For instance, the precondition *Actor:{"carmen"} ::Arena:{"writingArena"}*, states that carmen must be defined as an actor in the arena, in order to execute an interaction inside the "writingArena" arena.

4 Regulation service

In order to regulate a groupware aplication we introduce a regulation service based on the generic regulated architecture proposed on [14]. This architecture proposes the construction of a regulated collaborative application from two components: the component application and the regulation's component (See Fig. 2). At the application layer we observe the functions of the groupware applications on regulated way and at the regulation layer are described the collaboration spaces (arenas) and rules associated to all the possible interactions (scenarios). Whenever a user invoke a function, a request to the service of regulation is carried out. At the regulation layer the rules are verified and the corresponding scenario is executed.

The regulation service is composed by four components: (i) arena manager, (ii) CoRaL parser, (iii) regulation engine and (iv) regulation observer.

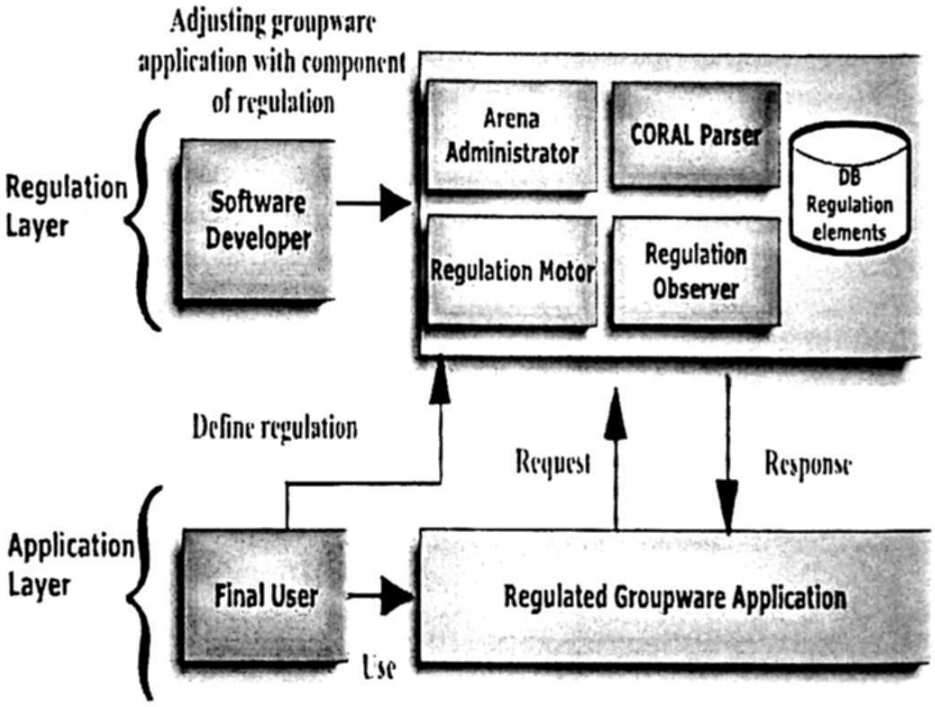


Fig. 2. Regulation service for groupware applications.

- **Arena manager:** Through this component it is possible to define regulation elements such as: arenas in simple or complex form, actors participating on arenas, family of actors/objects, interaction models and views. It includes funtions to add/delete/update elements inside arenas. It also allows to describe scenarios for interactions. The scenario includes, the actor participating on the interaction, the roles asigned to the actors and the sequence of individual actions executed by actors.

- **CoRaL parser:** The parser verifies that the scenarios described for each interaction are lexically and syntactically correct. The parser acts by request of the regulation engine or the arena manager.
- **Regulation engine:** The engine creates, deletes, up-dates and executes interactions according to the scenarios described using CoRaL. For that, it applies operators defined on section 3. So, it must verify the origin of the objects taking part in the interaction in order to know what regulation need to be applied (local or remote). If the interaction concerns remote objects, the regulation engine must send the interaction to the other arena and in the other arena it must apply its own regulation.
- **Regulation observer:** This component implements an observer of all the actions executed inside the arenas according to the rules established by the actors. It could be a software agent that can propose new rules for the group activity based on its observation and analysis.

5 Prototype

In order to demostrate the feasibility of our approach, we developed a prototype implementing the MARS model. This prototype was developed with the Microsoft .Net Framework 2.0 and works with he IIS Web server and SQL Server Express. The regulation service was implemented using a MVS architecture. This regulation service was proved on a groupware application developed at the Faculty of Computer Science of the University of Veracruz. This application, named GroupwareFei, offers tools for communication, coordination and cooperation such as: chat, forum, group agenda and space to share documents. It also offers functions to allow users to create a group and to add/delete group members (See Fig. 3).

In order to regulate the application it was necessary: i) to identify the functions of the application to be regulated, ii) to adapt the application in order to request the regulation service and iii) to define the scenarios using CoRAL.

5.1 Idetifying functions and adapting the application

The functionalities idetified were "to create a group", "to modify a group", "to share a document", "to communicate using chat", "to send messages to a group or a partner" and "to use a group agenda". Each one was associated to the regulation service in order to define the regulation model *i.e* the arena and all its elements. For instance, the function "to create a group" was associated to the operation "create arena". The association is specifiedt by developers.

5.2 Defining scenarios

The *arena manager* offers an interface that allows the creation of all the instances of the regulation model and the definition of the scenarios. Both actions are carried out by end-users. The instanciation of the regulation model is representing

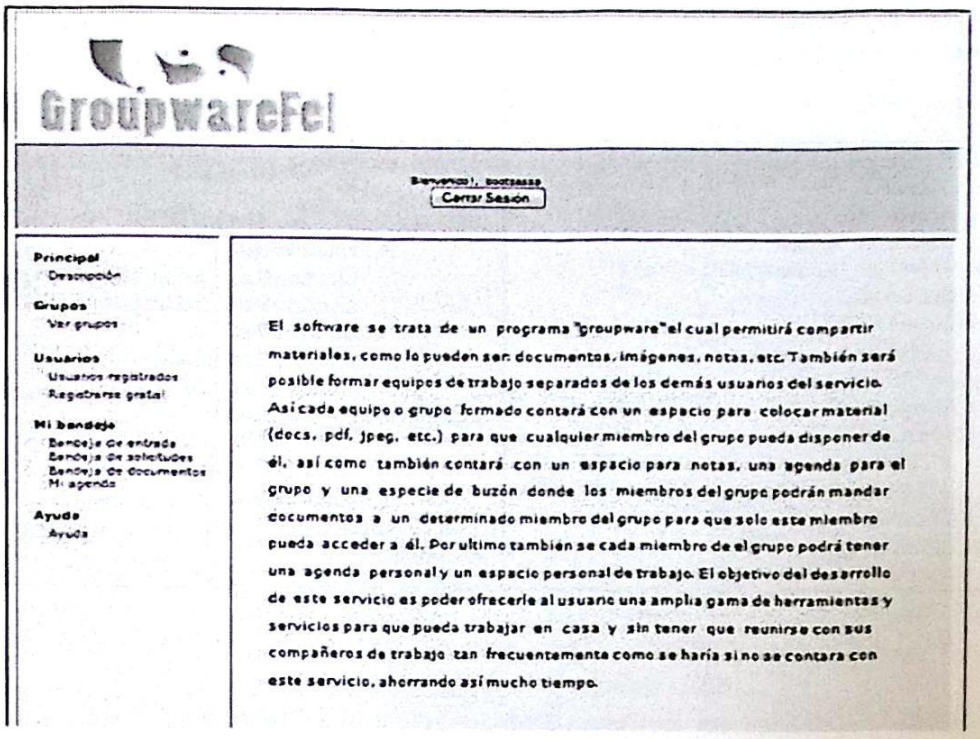


Fig. 3. Regulated GroupwareFei application.

internally using XML. In Fig. 4 it is possible to observe the definition of a "writing arena". It contains families of actors, actors partcipating inside the arena, objects, roles assigned to actors and objects, interactions and the scenarios.

The *parser* validates the scenarios described by the users. The scenarios must be described according to the elements of the arena. For instance, in the following we describe some preconditions of a particular scenario:

1. *Actor family:"manager" -> Interaction:"to Publish"*;
2. *Actor family: "collaborator" -> !Interaction:"to Publish"*;
3. *Actor family:"visitor" -> !Interaction:"to Publish"*;

We observe that the possibility to publish a document is only defined for the "manager".

5.3 Execution of the regulated application

The *regulation engine* is called by the application when a functionality is executed. This engine (i) verifies the interaction associated to the functionality in the regulation model, (ii) verifies the elements associated to the interaction and (iii) execute the scenario. The end-users execute the application according to the group rules, they can modify the elements of the arena. For instance: to add /delete an actor, to modify a scenario, to change roles, etc.

```
<Arena idarena="1"
name="Writing Arena ">

  <ActorsF>
    <ActorF idactorf="1"
name=" Administrator " />
  </ActorF>
  <Actors >
    <Actor idactor   ="1"
idactorf="1" name="Luis" />
  </Actors>
  <Objects>
    <Object  idobj="1"
name="Archivo">
  </Objects>
  <Scenarios>
    <Scenario idscenario="1" />
    <Scenario idscenario="2" />
  </Scenarios>
  <Interactions>
  <Interaction idinteraction ="4"
idscenario="2" name="
            ScenarioToPublish     " >
        <ActorsFRef>
          <ActorFRef  idactorf="1">
            <ActorFRef idactorf="2">
              <ActorFRef
idactorf ="3">
            </ActorsFRef>
        <ActorsRef >
          <ActorRef idactor="1"/>
          <ActorRef idactor="2"/>
        </ActorsRef >
        <RolesActancial>
          <RolActancial
idroleact="1" name="
            To Publish "/>
          <RolesActancial/>
      </Interaction>
  </Interactions>
</Arena>
```

Fig. 4. Definition of an arena for GroupwareFeiUV

6 Conclusions

Regulation is a natural social aspect of collaboration. In order to collaborate, people need to establish minimal rules, personal rights, preferences, availabilities and responsibilities in order to carry out their activity. Nevertheless, groupware tools rarely incorporate regulation. We believe, that if groupware applications are designed to support group activities, they must consider the model of this activity. This has two advantages: on the one hand, it make easy to developers the implementation of these applications and, on the other hand, it enables end-users to better adapt and use them.

This paper proposed a regulation service based on MARS, a multi-arena regulation model. This model enables to define the necessary interactions to carry out an activity in a single arena, but also it enables to define interactions taking place in several arenas. In order to develop regulated groupware applications, is also proposed an architecture composed by a regulation and aplication layer.

In order to validate our approach, we implemented a .net-based prototype. This prototype implements the multi-arena regulation model and controls the execution of interactions according to arena regulation. It allows to model the functions of the groupware applications in terms of interactions, to instaciate the regulation model and to describe scenarios for interactios. Developers adapt the functions of the application and end-users instanciate the model and define the scenarios. In this prototype the regulation service only includes the arena manager, the CoRaL parser and the regulation engine.

We observe that interaction can play an active role on the activity, so we explore the possibility to incorpore MAS technology not only for interaction but also for modelling the regulation observer. Our future work includes also modelling scenarios in a more detailed way in order to regulate complex interactions.

Acknowledgements

The authors thank the anonymous reviwers of this paper for their useful comments. This work has been supported by PROMEP (Project number: PROMEP/103.5/08/1073).

References

1. L. A. and B. Gregory. Inter-activities management for supporting cooperative software development. In *Proceedings of the Fourtheen International Conference on Information Systems Development ISD 2005*, pages 11–20. Springer Verlang, 2005.
2. Bscw: Be smart cooperate worldwide, 2008. http://public.bscw.de/en/index.html.
3. M. Christian, V. Laurence, F. Christine. and D. G. LDL: a language to model collaborative learning activities. In *ED-MEDIA 2006*, 2006.
4. M. Cortez and P. Mishra. DCWPL : A programming Language for Describing Collaboration Work. In *ACM Conference on Computer Supported Cooperative Work, CSCW'96*, Cambridge MA, USA, November 1996.
5. J. Couet and A. Davie. *Dictionnaire de l'essentiel en sociologie.* Edition Liris, Paris, France, 1998.
6. C. Dave. *The Locales Framework: Understanding and Designing for Wicked Problems.* Kluwer Academic Publishers, 2003.
7. W. K. Edwards. Policies and roles in collaborative applications. In *CSCW '96: Proceedings of the 1996 ACM conference on Computer supported cooperative work*, pages 11–20, New York, NY, USA, 1996. ACM.
8. Groove: Microsoft office groove, 2007. http://office.microsoft.com/es-es/groove/default.aspx.
9. T. Herrmann. SeeMe in a nutshell. the semi-structured socio-technical modeling method. Report, Department Information and Technology Manangement (IMTM), University of Bochum, Germany, 2006. https://web-imtm.iaw.ruhr-uni-bochum.de/pub/bscw.cgi/0/208299/30621/30621.pdf.
10. I. Jahnke, C. Ritterskamp, and T. Herrmann. Sociotechnical roles for sociotechnical systems: a perspective from social and computer science. In *AAAi Fall Symposium, 8. Symposium: Roles, an interdisciplinary perspective*, Menlo Park/California(USA): AAAI Press, 2005. AAAI.
11. D. Li and R. Muntz. COCA: collaborative objects coordination architecture. In *CSCW '98: Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pages 179–188, New York, NY, USA, 1998. ACM.
12. D. Li, Z. Wang, and R. R. Muntz. "got coca?" a new perspective in building electronic meeting systems. In *WACC '99: Proceedings of the international joint conference on Work activities coordination and collaboration*, pages 89–98, New York, NY, USA, 1999. ACM.
13. R. Li and D. Li. A new operational transformation framework for real-time group editors. *IEEE Trans. Parallel Distrib. Syst.*, (3):307–319, 2007.
14. C. Mezura-Godoy. *Une architecture pour le support de la régulation dans les collecticiels.* PhD thesis, Université de Savoie, Chambéry, Francia. 2003.
15. C. Mezura-Godoy, S. Talbot, and M. Riveill. Mars: Modelling arenas to regulate collaborative spaces. *Lecture Notes in Computer Science*, 2806/2003:10–25, June 2003.

16. M. R. Morris, K. Ryall, C. Shen, C. Forlines, and F. Vernier. Beyond "social protocols": multi-user coordination policies for co-located groupware. In *CSCW '04: Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 262–265, New York, NY, USA, 2004. ACM.
17. Zimbra: Zimbra collaboration suite (zcs), 2008. http://www.zimbra.com/.

Spatial Data Integration for e-Government Workflow Processes

Catalina Aranda-Castillo, Rafael Ponce-Medellín, and Gabriel González-Serna

Centro Nacional de Investigación y Desarrollo Tecnológico, Interior Internado Palmira s/n, 62490 Morelos, Mexico
{aranda06, rafaponce, ggonzalez}@cenidet.edu.mx

Abstract. Workflow systems had evolved to offer greater functionalities that could satisfy the organizations' needs and requirements, actually also taking advantage of the spatial information that can be taken from geographic information systems, with the benefits that this entails. However, new tendencies from government agencies' in some countries, such as Mexico, have sought the inclusion of open source software applications to perform their activities, appearing the problem that workflow systems that incorporate geo-referential information are expensive. This paper presents a comparison between different open source workflow systems: from them, Bonita workflow was selected to be modified to integrate into it the use of geographic information within the modeling process, describing the followed steps to modify the ProEd process definition tool, for the incorporation of a map viewer inside the executable process of the system. Finally, the advantages of incorporating geographic information in a workflow system focused on e-Government processes are discussed.

Keywords: Distributed systems, workflow management, Web application, Geographic Information System.

1 Introduction

The workflow systems had searched automation of the politics and procedures from business processes, associating people and work groups to the activities, for the management of the works and activities to be realized in an organization, making with this the possible the cooperation among different people and groups. Therefore, the incorporation of a workflow system inside enterprise and institutions for their process management has repercussions in the service and attention to the client, and in the efficacy and productivity they have in their activities, due to the speed up of the transactions and the control over the managed data. The grown of an enterprise is a key element in respect to the competitiveness against others in the business market, independently of the referred sector.

These types of systems had improved their characteristics in order to satisfy the actual organizational requests and necessities; between the more remarkable claims are: Web access, a graphic modeling for the business processes, distributed

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 61-81

Received 24/06/08
Accepted 01/08/08
Final version 11/09/08

architectures, standards adoption that guarantee the interoperability between diverse applications, etc.

Another really important characteristic didn't previously mentioned and that is relevant for the purposes of this article, is related to the functionality need of the workflow systems to include a geographic information support. This necessity could be covered through the integration of the functionalities available by the now so popular geographic information systems (GIS).

The relevance of using geographic information relays in the manner of represent the information with higher level of detail, bringing more elements to the person in charge for a more confident decision taking; with the traditional workflow systems, this wasn't enough feasible.

An example of a sector which had taken advantage of the workflow systems related to geographic information, is the cadastre of the government agencies, this is due to most of the transactions realized inside them need to incorporate the use of maps and sketches.

This last, in conjunction with the actual government necessities about dispose a system that allows to manage their process, with the goal of enhance the attention to the users, and at the same time trying to fulfill with the objectives of different austerity edicts (such as minimize costs related to software acquisition, material resources and stationery) have propitiated the development of this investigation.

According with the organization activity profile, the kind of procedures and information used and managed should vary, so some organizations must require the manipulation of geo-referential information (e.g. as the ones that requires to use and show a map to define a point of interest, or as a reference) in one or more different activities that compose a process.

A problem presented in this type of processes is that they can't be modeled by the majority of the actual *workflow* systems, because in general, they don't consider an interaction with components of geo-spatial services during the modeling phase, nor in the final Web application generated; meanwhile, the workflow systems that consider such characteristics, in fact are expensive and focused to particular activities, without the chance to modify or adopt them to the particular necessities of an organization.

This situation causes a minimum or also null automation of the processes that are needed and implemented inside the enterprises, generating: show or bad customer services, delays in the response times for requests and solicitudes, development of expensive ad-hoc systems and low scalability in the management of their procedures, excessive time consumption to modify (if possible) the part of the system that controls a specific procedure, etc.

In the present paper the integration of geo-referential information functionality inside an open source workflow system for e-government processes is presented. For the modeling of processes that involve geo-referential information, are presented the steps that have been followed to modify an open source tool that haves a form editor, and the adding of the necessary components for the handling of geo-referential information. So in this way, a workflow system that considers geo-spatial data could help in reality for a better decision-taking.

The sections of this article are distributed as next: section II addresses the problematic that motivates the perform of this project; to establish a context of the technologies implicated in this, an introduction about the geographic information

systems, the workflow systems and e-government, are presented in sections III, IV and V, respectively.

Section VI presents some examples of workflow system for particular applications that have some integration with geographic information systems. Section VII presents some representative workflow editors, their characteristics and a comparative between them.

The overall process to realize the implementation of the project proposed to solve the identified problem is described in section VIII. Finally, conclusions related are presented in section IX.

2 Workflow Situation in Government Dependencies

Some organizations such as government dependencies need to achieve their transactions and negotiations using geographical information (geo-referential data), as part of the information needed to accomplish a service request, being this the basis over that the decision taking to accept or reject a request is made, for example, construction licenses procedures, water intake request and outflow reports, among others. Studies as [1] shows the relevance of integrating geographic information systems in e-Government activities, as well as the social and political characteristics that must be taking into account for this kind of developments.

With the recent tendencies that have been occurred inside this kind of organizations, where the maximum reduction of costs is wanted in which is referred to software acquisition, it has been opted to adopt open source solutions (such as GIS works as [2] or workflow ones as [3] or [4]), being desirable not only specific-work frameworks but also to count with workflow systems that have under such characteristics the combine and manipulation of geo-referential data.

Meanwhile, comparatives between workflows such as [5] are made to compare the characteristics and benefits among these systems, that can help in the definition of workflow processes, such as in the case of e-government ones.

The problem is that the few systems which let process this kind of information are expensive because of the acquisition and payment of software licenses, meanwhile on the other hand, the open source workflow systems don't consider the use of information of this nature, or they do it in a restrictive way.

As a direct consequence, this type of processes can't be modeled with the actual open source workflow systems, having consequences this in a few or null automation of the transactions offered by this organizations, causing among other situations: a bad customer service, delays on the response times to a request, expensive ad-hoc systems development which would result with a low scalability for the procedure management, as well as an excessive time consumption to modify and give maintenance to the system that controls the process.

Some works have presented the integration of geographic information for e-government activities, such as cadastre (i.e. Online Cadastre Portal [6]) or maintenance of public road (i.e. Road maintenance management system, RMMS, [7]), among others; but these had been in the majority of the cases ad-hoc implementations

that search solving for specific problems, and that are not focused on the integration of open standards.

Because of the consideration of the previous reasons, this paper propose to modify an open source tool for process definition, which integrates the necessary components to the workflow system architecture, with the purpose of being able to model business processes which includes the use of geo-referential information.

In next sections are presented a brief overview of the geographic information systems and its use, as well as a description of the workflow systems, being the understanding of these systems the basis for the geographic information use inside a workflow system.

3 Geographic Information System (GIS)

A Geographic Information System (GIS) according to 1 is a system for the management, analysis and visualization of geographic knowledge, structured in different information sets, such as: interactive maps, geographic data, geo-processing models, data models and metadata, where the interactive maps provide an interactive vision of the geographic information, giving to the user the necessary tools for the interaction with this information.

In 8 is mentioned that a GIS is conformed by the following components:

1) Hardware equipment. This is the computer hardware in which the geographic information system operates; it is composed by general use equipment and specialized one.
2) Software. These are the set of functions and software tools that are used to analyze, store and display the geographic information. Between these are founded the database management system (DBMS), the graphic user interface, tools for data input and data manipulation, tools for geographic search, analysis and visualization.
3) Data. The most important part of a GIS. The system is in charge of integrating spatial data with other data resources and even it can utilize the most common database managers to administrate the geographic information.
4) Human resource. This refers to the personal who operate, manage and develop over the system, and that are also responsible for accomplish the decision taking.
5) Procedures. A geographic information system operates according a well structured plan, with clear rules, such as the models and operative practices features of each organization.

Those components let process and display the geographic information in a digital way, generating outputs according to the necessities of each different user group.

A GIS has a variety of different applications, as the ones mentioned in 10 and 11, highlighting among them: installations management, cadastre, urban design, transport services, geographic marketing, natural resources management, civil protection works, archaeological deposit studies, scientific research, education, automated cartography, territorial management, social equipment, digger resources, transit engineering, demographic studies, planimetry, 3D digital cartography, among others.

The use of geographic information can complement a variety of aspects for the model making of work flows, in the cases where the use of this kind of information be primordial, it's because this that it's necessary to know the characteristics and the components that conform a workflow system.

4 Workflow Systems

The Workflow Management Coalition (WfMC) is an international organization who has developed standards for workflow systems, which have been used for the communication between the components of a system. It's in this way that the workflow systems that adopt this standardization are available to interoperate between them.

A workflow system, according to the WfMC 11 is defined as: "*A system that defines, creates and manages the execution of workflows through the use of software, running on one or more workflow engines, which is able to interpret the process definition, interact with workflow participants and, where required, invoke the use of IT tools and applications*".

4.1 Reference Model

This organization had published a reference model [13] that specifies a framework for the workflow systems, identifying their characteristics functions and interfaces.

In Fig. 1 is shown the interfaces and components that can be found in a workflow system architecture.

Those components are described next:

Process Definition Tool. It is used to create a description about the processes in a resolvable computer way. This tool can be based in a language for formal process definition, in an interaction model between objects or only by a set of rutted rules for information transfer between the participants [14].

Workflow Engine. The workflow engine is software that provides the control for the executable environment for an instance of the workflow. Commonly, it provides facilities to: interpretation of the process definition, maintain control for the instances of the processes: create, activate or finish them, among others; to let navigation across the workflow activities, to bring up support for the user interaction, to let the user controls the data and applications and to summon external applications [15].

Workflow Service Representation. This is used for the interpretation of the process description and it is in charge of the control of the different instances of them, the establishment of the activities sequence, the adding of elements to the task user schedule, and for the invocation of necessary applications. All these tasks are made for one or more workflow engines, which are responsible for the management of the execution for the different instances of a group of process.

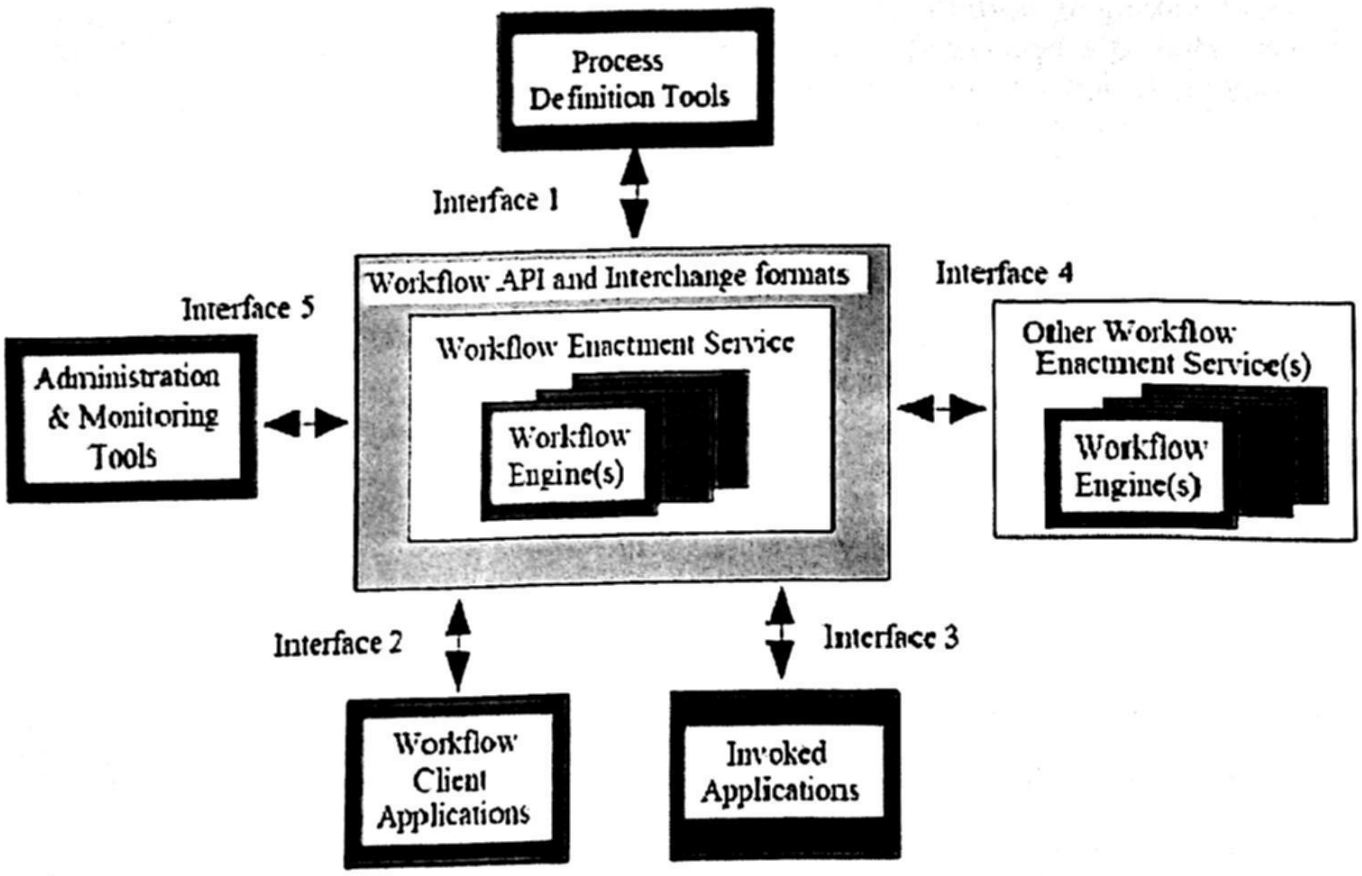


Fig. 1. Workflow reference model - components & interfaces.

Workflow Application Program Interface (WAPI). The WAPI can be sought has a set of API's (Application Program Interface) and functions for data interchanging, supported by the workflow representation service; these lets the interaction between the workflow representation service with other resources and applications.

It's important to denote that the open source workflow systems, in general, have adopted the reference model previously described. The next section describes some popular systems that incorporate the functionality of the geographic information systems together with workflow systems.

5 e-Government

The e-Government approaches let the citizens be informed, in a regional, state or federal scope, giving them access to a variety of services offered in a digital way. This can also be used by the Government as a way to guarantee and take advantage of the information accessed and its use.

Between the advantages of the government digitalization, is the transparency and speed up of different processes, as tax declarations, licensing requests, passports or

any transaction, making use of the actual information technologies, in an integrated process of continuous innovation.

The characteristics that an e-Government project must consider, includes: that it be open, according to the international Internet standards, totally oriented to the society, integrating the different government services in their business and jurisdiction processes towards a complete integrated system.

Four main kinds of activities take place in these systems [16]:

- Show information on Internet: notifications, regulatory services, holidays, etc.
- A two-way communication environment, in cases between the government agency and the citizen, a business or also another agency.
- Conducting transactions, e.g. conducting services and grants.
- And for governance, like in the voting and electoral campaigning.

Actually, some countries' government has been adopting open source software, as alternatives against proprietary software, in a way to minimize costs and to take more control over the applications, being possible to adapt them to the particular governmental necessities.

It's because this that is necessary to adapt and implement the necessary changes of the existing platforms, in order to give them a higher functionality, according to the newest and increasing necessities of the society, as is the case of the use and integration of geospatial information in some agencies' transactions.

6 Workflow Systems Integration with GIS: Representative Examples

6.1 GeoPISTA

GeoPISTA [17] is a territorial information system used in city halls; it lets implement the necessary functionalities for the territorial management: urban planning, cadastre, habitant census, contaminant activities, patrimony, infrastructures, activity licensing, urban guides, etc.

It lets elaborate small processes or work flows (such as user control, documentation flow, date management, etc.) and also lets control the documentation and/or observation storage generated by any element from the application [18]. It has been programmed in Java, characterizing the application as a multiplatform one.

GeoPISTA Components:

In [19] are mentioned four different elements that integrate GeoPista's structure: databases, data servers, GIS clients and modules, described next:

1) **Data server**. It works with the PostgreSQL database management system, using the PostGIS extension, used to handle with the geographic information. The database is structured according to a predefined model for each application ambit (infrastructures, referente information, patrimony, etc.) and over such model the rest of the applications are supported.

2) **Internal basic services.** The cartography administrator is in charge of present the geographic information in the form of maps, as well as to manage the users with different permissions and has the task to resolve concurrency problems of data access. On the other hand, the map server lets publish geographic information on Internet.
3) **Clients.** They let the user to navigate across the geographic information, to make queries over this and to analyze it, to edit the data and also print maps. GeoPISTA lets connect some common commercial solution GIS (ArcGis, MicroStation and AutoCAD) with its system.
4) **Specific modules.** A set of modules that brings assistance in the municipal management, in such scenarios as urban planning, construction licensing procedures, infrastructures, contaminant activities inspection, among others.

6.2 The Dorado

The Dorado [20] is a digger concession system, which has as objective to make a more efficient activities management, from a request income until the granting and posterior following of the audition step.

Dorado Component Architecture

The system is based on a client/server environment for the data management, from the spatial data to the administrative processes. It lets the manipulation of information external to the system in a transparent and automatic way for the integration between the application and the operative system.

The processes coordination is through the use of the M&B Process product. The management of the spatial data model uses development applications under MapObjects (ESRI) object libraries.

Functionalities

The system is divided in five modules, which are described next:

1) **Client attention module.** It provides tools that let administrate and coordinate all the requests that income from the public, in general, or by the digger business customers.
2) **Arrangement module.** It lets income and validate the taxes of the requests, the work assignation between the different functionaries and to take control of the time that spends each process. The spatial component fulfills with the functions of cadastre actualization, including spatial validations, as much as in the incomes in the areas as well as in the technical evaluation processes.
3) **Audition module.** It lets register, follow and evaluate the activities of daggering exploration and exploitation. It manages the control of concessions and the administration for gold, diamonds and precious jewels merchants; also, it covers the related to taxes, penalties and digger rights extinction.
4) **Executive statistic information module.** It presents the business information to the ministry high directive.
5) **Maintenance module.** It provides tools that give support to the system processes, such as the actualization of information and the management of the

multi-user platform of the spatial data; it also lets have process for a massive data charge used for the actualization for cadastre data.

6.3 ArcCadastre

ArcCadastre [21] is a system for the management of cadastre and geographic information, such as the map generation based on GIS systems; it manages cadastre measurements and field information.
It has been developed in cooperation with ESRI Inc. and is based in the following platforms:

- ArcGIS
- Survey Analyst (used for planimetry and calculus functions).
- ME Objects, de Safe Software Inc., utilized to import and export data in different file types.

The activities are managed and monitored within help of a work flow; the easiest of them is a control list, meanwhile another form lends the user through the different steps in a hierarchical order. The most complicated flows requires more personalization, this is obtained through programming modifications with VB or C++. The information is stored in geographic databases. ArcCadastre gives support to store information in many databases managers, such as MS Access, Oracle, IBM, Informix and the Microsoft ones.

7 Workflow Open Source Systems

7.1 Workflows Standardized under the WfMC Model

JaWE / Shark. JaWE (Java Workflow Editor) is a graphical editor of workflow processes, based on Java and XML, and is compatible with the WfMC specifications. It works with the XPDL processes definition language (XML Process Definition Language) 12.

Shark is a workflow engine based on the WfMC and OMG (Object Management Group) specifications. It uses the XPDL definition process language, and can be used in different environments such as a Web application or swing applications, and can be installed as a CORBA service or it can be accessed by client applications through CORBA ORB or by an EJB container [23].

Bonita. Bonita workflow 24 accomplishes with the WfMC reference model specifications 13. It is developed according with the J2EE specification, it's distributed under a LGPL license and it uses the XPDL process definition language.

It uses JOnAS application server (Java Open Application Server) 25, which is developed also under the J2EE specifications. It's in charge of the security management and the messaging with other services.

Bonita's environment is based over the Web, so it can be accessed through any Web browser.

To realize the process model is used the ProEd process definition tool, which is incorporated in Bonita. The modeled process is stored in XPDL format, which will be interpreted by the workflow engine.

7.2 Workflows non-Standarized under WfMC Model

JBoss jBPM. JBoss jBPM is interoperable with all the integration technologies based on J2EE, such as Web services, Java messaging, J2EE, JDBC (Java Database Connectivity) and EJBs (Enterprise JavaBeans) connectors [26].

The main components of this tool are:

- jbpm-server. A preconfigured JBoss application Server.
- jbpm-designer. An Eclipse plugin for process modeling in a graphical way. It provides a program model oriented to process with jPDL process definition language jPDL (jBOSS Process Definition Language).
- jbpm-db. A compatibility package for the database. JBoss jBPM can be configured with databases as: Oracle, MySQL, Hypersonic SQL, PostgreSQL, among others, and can be implemented over any application server.
- jbpm. Component developed under Java (J2SE) for the process management definitions and the execution environment for the running of the process instances.

Intalio. Intalio is an open source system for the business process management; it accomplishes with the J2EE specification and is developed under a MPL license (Mozilla Public License). It utilizes the BPEL language (Business Process Execution Language), it generates Web services and includes a rule engine and a Web user interface 27.

The definition processes tool used is a development environment based over Eclipse; it lets that a BPMN model (Business Process Modeling Notation) can be converted to an executable process, without writing code, this can be achieved through a combination of proprietary generating code algorithms.

Workflow Systems Comparative

Table 1 shows a comparative between the previously described workflow systems. In this, are evaluated the following aspects: WfMC model adoption, Web based user interface, if it counts with a definition process tool based on a Web application, if it incorporates a formulary editor, the process definition languages supported, if it is implemented under the J2EE specification, and the type of distribution license it has.

The following characteristics were considered for a integration between a workflow system with a GIS:

WfMC model adoption. The proposed model by the WfMC establish a set of interfaces that lets an interaction between the components of a workflow system, one of them (interface 3, from the Fig. 1) allows the communication with external

systems, being a crucial point for the developers of a project if the functionalities of a GIS are wished to be used inside a workflow system.

Form editor. This point refers to the characteristic about if a tool used for process modeling incorporates a form editor, which be used to define and manipulate the forms. This kind of component is necessary to take into account, because for a integration with a GIS is necessary to add to the forms a special class of data type: a geo-spatial data component.

Database interoperability. The workflow interoperability with different database managers is a relevant characteristic to be considered, due mainly because the geo-spatial information used to reference a geographic place or zone must be stored into a manager that supports this special kind of data.

License. The type of license of distribution for a workflow system should be considered, due that if the selected tool is going to be modified, is necessary to get the source code, as well as the freedom to make the appropriate changes to workflow's components.

The relevant characteristics that were taken into account for the study case of this paper were:

· **Web client.** The workflow system must incorporate a web client or a Web process manager console, with the purpose that this can be acceded from the majority of the clients.

Web modeling. It is needed that the tool proportioned by the workflow system could also be acceded via Web.

Process definition language. The language used for the process definition must be a standardized language, because it needs to be interpreted by other workflow engines, towards an interoperation between different systems that could collaborate between them.

Adoption of the J2EE specification. The J2EE specification defines the guidelines for the development of multilayer distributed systems, making this characteristic desirable in a workflow, as a evaluation criteria.

Table 1. Open source workflow systems comparative.

System	WfMC model adoption	Formulary editor	Database interoperability	License	Web client	Web model	Process definition language	J2EE
JaWE/ Shark	✓	✓		LGPL	✓		XPDL	
Bonita	✓	✓	✓	LGPL	✓	✓	XPDL	✓
jBPM		✓	✓	LGPL	✓		jPDL	✓
Intalio		✓	✓	MPL	✓		BPEL, BPMN	✓

As can be seen in the previous table, the workflow system that more satisfies the previously explained requirements is Bonita workflow; Bonita also provides a set of mechanisms called *hooks*, which let connecting this workflow system with external systems. These mechanisms were provided thanks to the adoption of the WfMC reference model.

8 Solution Method

To visualize the benefits that involves the implantation of the presented workflow system, one practical example are the governmental dependencies, in which the most of the paperwork and procedures that are offered to the citizens, that imply the use of geo-spatial data, such as the ones that requires the use of maps or sketches to set the specific location of a required service; examples of this type of procedures are: requirements for construction licenses, water inlets, water leak, among others.

The existing problem in these dependencies is the way in which the procedures for the services are realized, mainly because at the moment of making the registration of a request, this is usually filled by hand, and after this, it is assigned to the respective personal in charge for the following of that request. The previous situation causes an inaccuracy in the information proportioned by a user, when defining the geographic location of the required service; also the high consumption of stationer resources, caused by filling errors, by the number of realized requests, the accidental lose of requests, delayed delivery in the request's responses, bottle necks, among others.

To solve this problematic (briefly described in section II), the following solution method was developed:

8.1 Workflow Open Source System Analysis

For this activity, it had been developed an analysis about the main open source workflow systems, which are available through Internet. The analysis results are presented in section 7.

Considering the previous evaluation, the Bonita workflow was selected to be modified with the pertinent changes with the purpose of integrating into it the necessary modules for geographic information functionality and process.

8.2 Bonita Workflow Analysis

The Bonita workflow administrative console is named jiapAdmin, which is a Web application developed using JSP (Java Server Pages) in which are managed the basic system configurations and the process management. It handles four different user roles, described next:

1) **Administrator**. It is in charge of modifying the basic data configuration of the workflow engine, for the users administration and their configuration. It is the user with higher privileges.

2) **Designer**. It can access to the process administrator to create or modify processes models using the ProEd workflow editor. It can manage the process models (such as import different process in XPDL format or erase process).
3) **Operator**: It is in charge of establish user preferences, display, refold or start process, finish process instances, access to information about the process instances, start, finish or cancel an activity in a specific instance, configure and bring access to the logs and process instances historical.
4) **User**. It is the user with less privilege, its activities are focused in starting workflow process, he can start, stop and cancel activities and display the finished activities which are still visible.

jiapAdmin communicates with the workflow engine through the APIs provided by Bonita.

Architecture

Bonita workflow architecture is presented in Fig. 2, in this can be identified the two principal components of it: the ProEd process editor and the workflow engine.

The first component is the ProEd definition process tool (Process Editor), which was developed using java language. It lets make models in a graphic way, using the BPMN standard (Business Process Modeling Notation). It saves the model processes in XPDL format.

ProEd incorporates a formulary generator called xformeditor, which is based on XForms, a markup language for Web formularies that separates the data from the logic of the presentation.

The second component is the workflow engine, which is executed over JOnAS application server; This one is in charge of bring the necessary components (Web container, EJB container, message services, security, etc.) to give support to the logic business and data access functions.

Bonita workflow incorporates three APIs, with through them, other applications, as jiapAdmin can interact with the workflow engine. These APIs are described next:

- **User's API.** It provides total control over the process in execution, for example, to start or finish an activity. It can retrieve automatically the identity of a user in the context of the J2EE security.
- **User registry API.** It lets create and modify the user properties inside Bonita system.
- **Project API.** It incorporates functions necessary to define a workflow process, such as the creation of activities, transitions, roles, actions, etc.

According to the realized functions, the APIs can call to the following beans, which are situated in the EJB container:

- User session register. It provides an interface that is used to user creation and management, as well as groups creation.
- Project session. It provides an interface used to process creation, node (activities) and edges (transitions) definitions, and for properties listing and modifications.
- User session. It implements commands and petitions related to: user's projects, lists of task to do, execution of activities, start/finish/cancel process commands.

- Engine session bean. It implements the states machine and controls the processes executions.
- Container Manager Persistence (CMP). It matches the fields from Bonita to their respective table inside the database.
- XPDL session. Module that analyses a XPDL file. During the analysis process of the XPDL file, the module calls directly to the session project bean of the API. In fact, this task will call a java client, which wills carry out a call to the XPDL session bean, responsible of the analysis and interaction with Bonita's API [29].
- Message controller. It implements the notification of the changes in the definition and execution inside a workflow process. Each user interaction is notified to Bonita's nucleus and a JMS (Java Message Service) event is launched. It is also in charge of redirect the messages, through email or by instant messaging.

The authentication mechanism is realized through JAAS (Java Authentication and Authorization Service), a standard way to configure the security of a J2EE application.

The process data and process instances are managed by a database. JOnAS utilizes by default the HSQL database manager, however, this can be replaced.

Formgenerator utilizes Chiba [28] as a formulary processor, which is a JavaBean application which can be integrated inside the application. This is summoned each time a form is displayed, for the register of information about an activity or process.

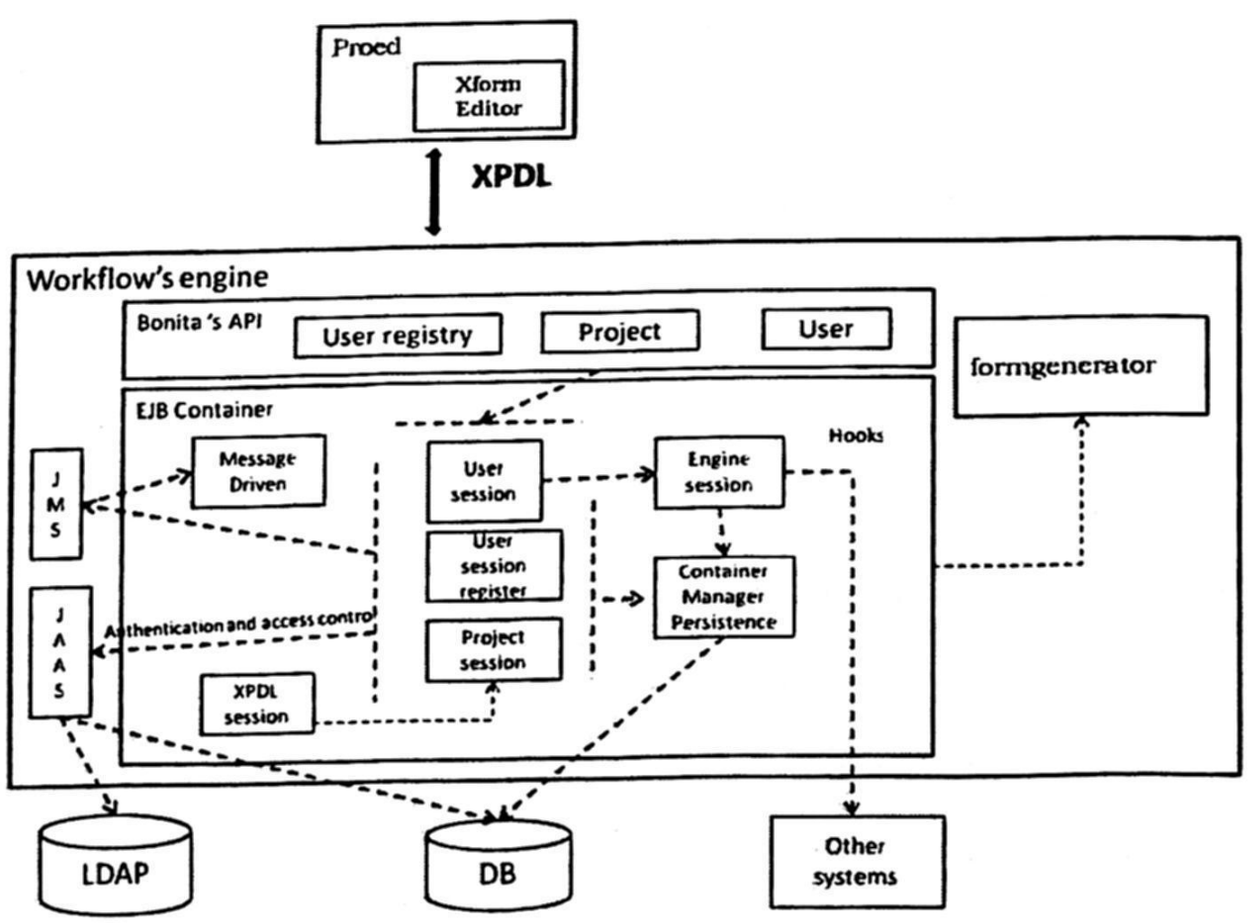


Fig. 2. Bonita workflow architecture.

8.3 Solution Design

As a result of Bonita's workflow architecture and functionality analysis, it was discovered that the ProEd process editor incorporates the xformeditor form maker, and that the Bonita's workflow engine has the formgenerator module for the management of the formularies which uses the Chiba form process. By taking into account that Chiba supports JavaScript code [30] and that the library used to create OpenLayers map visors is coded in JavaScript, it had decided that the best solution consists in modify the ProEd process definition tool, for adding into it a geo spatial component, as well as the develop of a Web manager console that can proportionate a friendly interface against the final user. Coming next the proposed solution design is described.

ProEd process editor Tool modification

It's necessary to make the pertinent changes of the ProEd process editor and the xformeditor formulary generator, to add the modules that let insert a map viewer inside a formulary. These modules are responsible for: the management of the geo-spatial component as an attribute inside the business process, the insertion of a script for the map viewer inside the form, and the petitions making to the map server for the display of available map catalogs to the user. The map viewer is in charge of managing the following properties:

- Map Server direction (URL).
- The requested maps themselves.
- Map layers to be showed over a map.

The maps viewer is developed over JavaScript, using the OpenLayers library; this brings the following activity facilities over a showed map:

- Zoom in.
- Zoom out.
- Pan.
- To specify the layers to be showed.
- Set POIs (Point Of Interest) over a map.

Web Process Manager Console Modification

This step consists on the modification of the Web process manager console, which is used to interact with Bonita workflow. This application needs a usable and intuitive interface against the final user. As a result of that, the process manager console must follow usability user interface principles, focused on a easiest and simple use for the final user.

The manager console consists on a Web application based on the MVC (Model-View-Controller) design pattern, which lets separate the data, the user interface and the control logia as three different components (see Fig. 3).

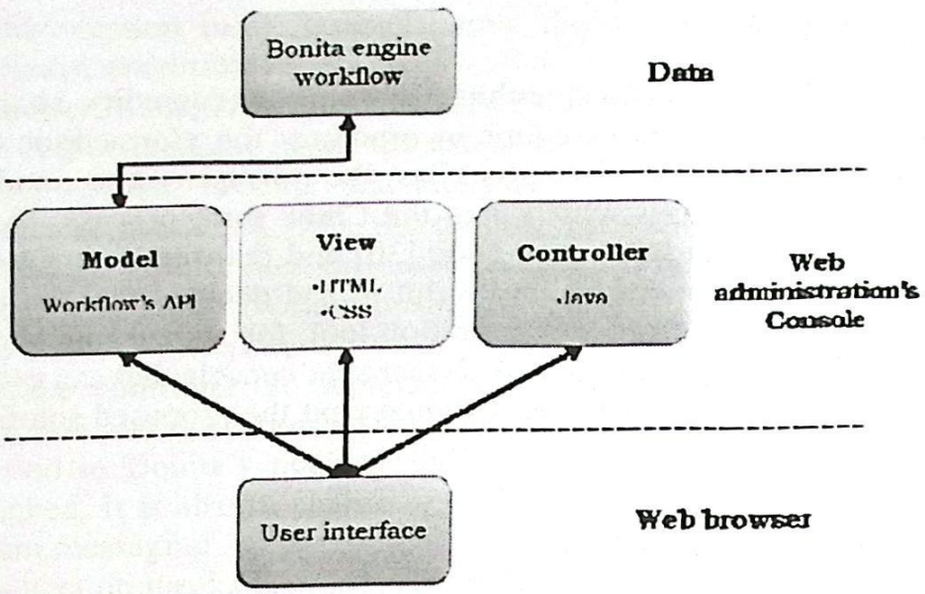


Fig. 3. Web process manager console architecture.

8.4 Implementation

This activity is composed by the modification of the ProEd process definition tool, adding as a spatial component a map viewer inside the xformeditor formulary editor.

The modification to the Web application was made using JSP technology and struts, inside the Eclipse's integrated development environment.

Once realized the respective modifications to the ProEd's tool for process definition, and the ones for the process management console, the architecture shown in Fig. 4 was obtained:

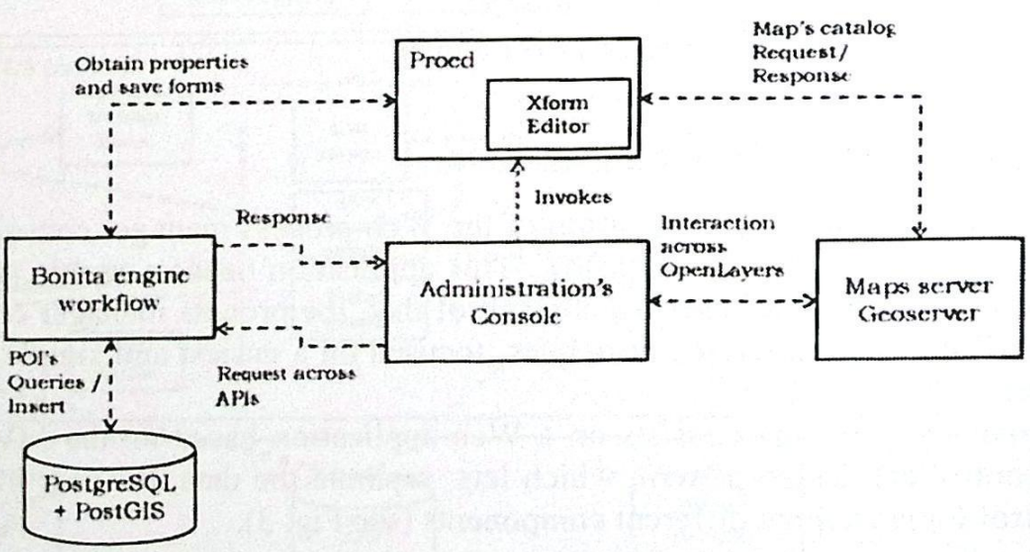


Fig. 4. Architecture for the definition and management of processes interconnecting a form editor (*ProEd*) and a map server (*GeoServer*) for geo-spatial data.

As can be seem in the previous picture, through the manager console, the modified ProEd process editor can be called to realize the model of the process. With this, ProEd has acquired the functionality to realize petitions to Geoserver, and obtain a catalog of the available, and that these maps could be displayed to the user, selecting the needed ones, and display them within the map viewer each time a process is executed. Once selected the maps and their layers to be displayed, an according script is created inside the form, and it's saved inside the workflow engine.

When a process is imported and deployed inside the workflow engine, it can be executed, and when this happens, the management console summons the respective forms for the process. When the form is loaded, the defined map viewer inside it realizes a map request to the map server, requesting also the layers to be displayed; this communication is made in real time, using the OpenLayers library.

The insertion of POIs is made through a *hook* component, which is proportioned as a part of the Bonita workflow engine. The hook is programmed to retrieve the coordinates of the POI, and insert them into a spatial database, as well as retrieve them in the cases that they already exist in any of the instances of a process.

One example of the new functionality that the workflow engine has acquired, is shown in Fig. 5; in this figure can be observed an instance of a process in execution, in which a spatial component has been defined, showing a loaded map of Cuernavaca City, in Morelos, Mexico. Such as has been previously explained, the map is loaded from Geoserver map server, in real time, using the OpenLayers library.

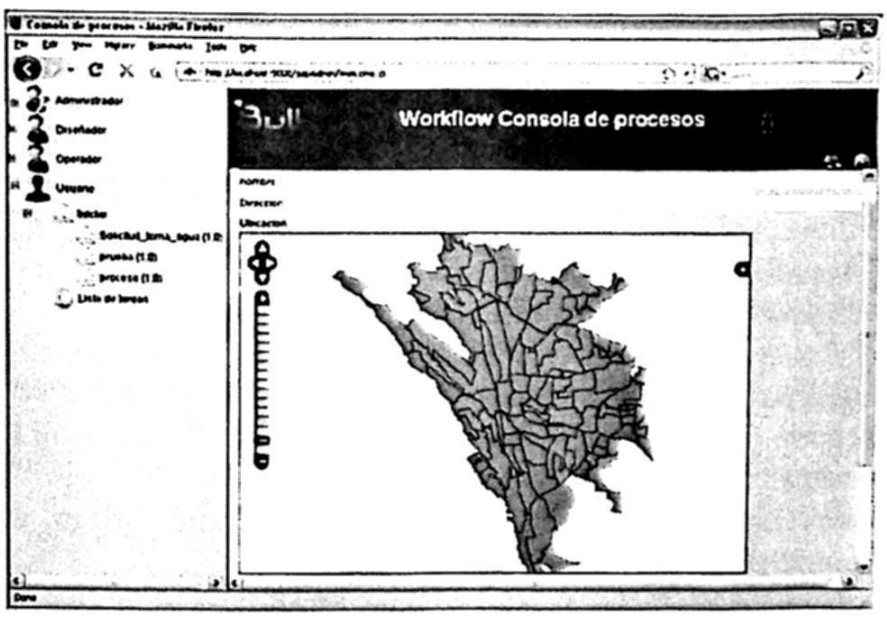


Fig. 5. An example of the geo-spatial functionality in an open source workflow engine (in this case, based on Bonita).

8.5 Testing

The development of this activity consists on conducting the implementation of the government dependencies process, any one in which it could be necessary the

management of geographic information, letting the user to introduce a geographic localization for a specific request.

Also, it was developed a test plan based on the IEEE 829-1998 standard [31], used for software testing; through this one, the tool functionality for the process definition and the developed Web application are verified.

The characteristics to be tested are listed next:

- Connection to the map server. This test tries to verify that PorEd editor correctly can establish a connection with the map server.
- Interaction with the map server. This test is focused on verify that the ProEd editor can request the catalog of the available maps on the map server.
- Forms definition with spatial data. This is oriented to verify that the created formularies with ProEd tool will be correctly stored in the workflow engine and that these could contain geospatial components.
- Map server connection. This is focused to test that the embedded map viewer inside a formulary of an execution activity, could establish a communication with the map server specified in the activity properties during the process modeling.
- Map interaction. The map visor capacity is verified to be used in operations such as: zoom in, zoom out, pan and selection of layers to view.
- Spatial database interaction. The capacity to do insertions and consults over the spatial database is verified, with the purpose of POIs storing and retrieval for each process instance.

Finally, the implementation of this integration between GIS and a workflow engine into an organization such as town government dependencies could prevent the problems and errors exposed at the beginning of this section, also with the following advantages:

- Cost reduction over stationary consumption, when the requirements are made through electronic format.
- Reduction over the error margin in the requirements filling, due to if the geographic location is managed through electronic ways, it can have a better precision for locating zones and places in a digital map, with more confident coordinates and a more realistic level.
- Improvement of the service quality offered to the citizen, at reducing the response times for the requests.
- An accessible system that can be accessed through a Web browser, so, it doesn't need specialized installation software for each computer that needs to access to the system.
- The payment of licenses is avoided, thanks to the use of open source products, working in favor of saving costs, and supporting government's austerity decrees.
- Time reduction used to make modifications over the process modeling.
- Use of a tool with a friendly user interface, that can be adopted to the particular necessities of a process, letting to the users' system the execution of their work, at following the required service solicitudes.
- Interoperability with other tools that fulfill the WfMC specifications.

- A scalable system that lets establish communication between systems through the use of hooks..

9 Conclusions

It had been adopted that the use of open source software, particularly in the case of workflow systems, due to a financial perspective, they can save licenses payments, against the proprietary software competitors; it also allows to make software adaptations for the particular necessities of an organization, because commonly the source code is available; its use doesn't no imply any dependency with a particular operative system; it ensures the permanence and reusability of the information, thanks to the use of open and standardized formats, and no use of closed standards that lose their currency with the time.

The study and analysis of diverse open source workflow systems were presented; from these, Bonita workflow was been selected as a very interesting choice, due to its characteristics which are offered as a set of APIs used to establish a communication with the workflow engine, the information access via Web and by the use of standardized notations used to make the processes models.

The geographic information support in Bonita workflow is achieved through the modification of the ProEd process definition tool, what is made by the incorporation of the necessary classes that allows defining a spatial component (map viewer) during the form definition in the process model step.

The use of a workflow system with the previously mentioned characteristics in this paper, could allow disposing of a set of functionalities in a government agency that lets create a totally scalable, manageable and reusable business process. This is obtained through the use of a graphic language used for the business process definition, by the support for the manage of geo-referential information during the process definition, by the automatic generation of the Web application used to manage the defined processes and finally, by the support of the Web application for geo-referential information in a manner of map views that helps for a more founded decision taking.

References

1. Aziz M.: Integration of GIS and e-Government. GIS Development –Middle East, Vol 2 Issue 2, Yashi Media Works, New Delhi (2006), pp 16-20.
2. Souza C., Leite F., Rodrigues E., Paiva A.: Using Open Source GIS in e-Government Applications. Electronic Government, LNCS, ISSN: 0302-9743, ISBN: 978-3-540-22916-2, (2004), pps 418-421.
3. Wolter C., Plate H., Herbert C.: Collaborative Workflow Management for eGovernment. Database and Expert Systems Applications, DEXA 07, Vol 3, Issue 7. (2007), pps 845-849.
4. Becker J., Algermissen L., Niehaves B.: Processes in E-government focus: A procedure model for process oriented reorganisation in public administrations on the local level. LNCS, ISSN: 0302-9743, ISBN: 3-540-40845-2. (2003).

5. Wohed, Petia, Russell: Patterns-based Evaluation of Open Source BPM Systems: The Cases of jBPM, OpenWFE, and Enhydra Shark. BPM Center Report BPM-07-12, BPMcenter.org, (2007).
6. Radwan M., Bishr Y., Emara B., Saeh A., Sabrah R.: Online Cadastre Portal Services in the Framework of e-Government to Support Real State Industry in Egypt. Pharaohs to Geoinformatics. Egypt. (2005).
7. Trewina P. Y., Dickson K. W.: Alert-Driven E-government Service Management: A case Study on road Maintenance Management System. IEEE Asia-Pacific Conference on Services Computing. (2006), pps 160-167.
8. ESRI España Geosistemas (ESRI). ¿Qué es un SIG?. http://www.esri-es.com/index.asp?pagina=285
9. Carmona A., Monsalve R. J.: Sistemas de información geográfica. http://www.monografias.com/trabajos/gis/gis.shtml
10. Delgado J.: Integración de información económica y territorial. http://unpan1.un.org/intradoc/groups/public/ documents/CLAD/clad0047804.pdf
11. Tinoco R.: Definición y algunas aplicaciones de los sistemas de información geográfica. http://www.monografias.com/trabajos14/informageogra/informageogra.shtml#ap
12. Workflow Management Coalition: Terminology & Glossary. (1999) http://www.wfmc.org/standards/docs/ TC-1011_term_glossary_ v3.pdf
13. Hollingsworth D., The Workflow Management Coalition: The Workflow Reference Model. http://www.wfmc. org/standards/docs/tc003v11.pdf
14. Marroquin W.: Workflow y UML. http://www.willydev.net/descargas/articulos/general/ WorkFlowUML.pdf
15. Canseco V.: Breve Introducción a los Sistemas Colaborativos: Groupware & Workflow. Universidad Tecnológica de Mixteca. http://mixtli.utm.mx/~resdi/breve_ introduccion_a_los_sistemas_colaborativos.pdf
16. Maureen Brown M.: Electronic Government. Jack Rabin (ed.). *Encyclopedia of Public Administration and Public Policy*, Marcel Dekker, (2003) pp.427-432.
17. Diputación Provincial de Valencia: Aplicación GeoPISTA. http://www.dival.es/ isum/Main?ISUM_ID=Left&ISUM_ SCR=serviceScr&ISUM_CIPH=ntnBCl!WtMj1UE-TUO0yiuzUr6iUUu 6GOofSGyqoevU#.
18. GeoPISTA: GeoPISTA (Product paper). http://www.geopista.com/fileadmin/docs/Hoja%20de%20 producto.pdf
19. GeoPISTA: Guía GeoPISTA. (2006). http://www.pistalocal.es/ppal/Main?ISUM_ID=Content&ISUM_SCR=externalServiceScr &ISUM_CIPH=NO4nynRwy0OE6lYBQZAIU3ppTzJTAb8hGkD54Dj8nH1tyPpcHA1Ml IlFVSaz763SmXxMkI3utXtp3StuJ5zpcaqdhzkhW6GgKRsKp19ysnlzVv2uQ5eOEHkS-0M2GeO8QjJdRD M3Kj3yNMMdXNTOK9PMB4GMvOMe
20. Bastidas Ramírez C.: Integración de tecnologías de información cliente/servidor, workflow y GIS una experiencia nacional. http://gis2.esri.com/library/userconf/latinproc99/ponencias/ ponencia39.html#_Toc447014972
21. Lantmateriet.: ArcCadastre Product. (2007). http://www10.giscafe.com/link/display_detail.php?link_id= 19442
22. Together Teamlosungen: Open Source graphical XPDL Java Workflow Editor". http://www.enhydra.org/workflow/jawe/ index.html
23. Together Teamlosungen.: Java Open Source workflow engine based on XPDL. http://www.enhydra.org/workflow/shark/ index.html
24. Bull R&D.: Bonita, The Open Source Workflow project. http://wiki.bonita.objectweb.org/xwiki/bin/view/Main/
25. ObjectWeb: JOnAS Java Open Application Server. http://wiki.jonas.objectweb.org/xwiki/bin/view/Main/WebHome

26. Red Hat, Inc., JBoos: Diseñador Gráfico de Proceso JBoss jBPM". (2006). http://docs.jboss.org/jbpm/v3/ spanish/jbpm-gpd-installation-spanish.pdf.
27. Intalio: Product Benefits. http://www.intalio. com/products/ designer/
28. Blachon M.: Cutomizing XForms. OW2 Consortium. (2007). http://forge.objectweb.org/forum/ message.php?msg_id=8373
29. Valdés Faura M., Blachon M.: The BONITA Workflow System XPDL Support" (Version 1.0). BULL R&D: (2006). http://wiki.bonita.objectweb.org/xwiki/bin/download/Main/Documentation/bonitaXPDL.pdf
30. Chiba: http://chiba.sourceforge.net/
31. Software Engineering Technical Committee of the IEEE Computer Society: IEEE standard for software test documentation. USA. E-ISBN: 0-7381-1444-8, ISBN: 0-7381-1443-X. (1998).

REC: Improving the Utilization of Digital Collections by Using Induced Tagging

J. Alfredo Sánchez, Omar Valdiviezo, Emmanuel Aquino, and Rosa Paredes

Laboratory of Interactive and Cooperative Technologies
Universidad de las Américas Puebla
{j.alfredo.sanchez,omar.valdiviezo, rosa.gpe.paredes}@gmail.com,
aquinoperezemm@hotmail.com

Abstract. REC is a software environment designed to generate recommendations of potentially useful web resources based on tags assigned by a community of users. Recommendations prompt users to access resources that otherwise would go unnoticed, thus promoting improved data exploitation. REC implements the notion of "induced tagging", a technique devised to improve the effectiveness of social bookmarking for digital libraries by having a group of specialists tag resources as part of their job, in addition to tagging performed by the community. This paper describes the initial experiences with the introduction of induced tagging into an actual learning community, which has motivated an implementation of REC that is embedded in a popular social networking environment. We have gathered initial evidence that indicates that REC mediates an effective collaboration that promotes resource discovery and results in an improved utilization of digital library collections to support academic activities.

Keywords: Tagging, social networks, recommendations, digital libraries.

1 Introduction

Tagging clearly has become a ubiquitous mechanism that supports both personalized and collaborative organization of varied information spaces. Also known as social bookmarking, tagging is present in an ever increasing variety of contexts: photo and video sharing environments such as Flickr and YouTube, citation databases such as CiteULike, many electronic newspapers and popular blog services, to name a few cases, provide handy mechanisms for users to assign keywords, or tags, that function as metadata that describe documents or multimedia objects. When tags are assigned by a sufficiently large number of users, they can be used to effectively support browsing or searching very large information spaces. Also, tag collections can be regarded as determining dynamic classifications of digital objects that depend only on the keywords chosen by all kinds of users, rather than the more conventional controlled vocabularies used by specialists. In that sense, tags are said to originate "folksonomies".

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 83-93

Received 01/07/08
Accepted 01/08/08
Final version 12/09/08

In addition to proprietary tagging mechanisms that are embedded in systems such as those mentioned above, a significant number of services have been developed that help users assign tags to any web resources and provide them with tools to manage their tag collections and to take advantage of tags assigned by all users in a community. The most popular of such services is arguably del.icio.us, which only a few months ago boasted three million registered users and about 100 million bookmarked unique URLs[1]. Other services that specialize in social bookmarking include Mister Wong (mister-wong.com) and Simpy (simpy.com).

The success of social bookmarking has prompted a lot of work aimed to investigate, for example, how tagging occurs, how tag collections are structured, how close folksonomies are to formal classifications generated by specialists, or how tags can be used to enhance information retrieval. Our group has devised a technique termed induced tagging, which is based on the idea that a group of specialized users within a community can be in charge of tagging as part of their jobs. We posit that this will have a positive impact on the size and quality of the tags collection and, as a result, on the usefulness of social bookmarking to support the discovery of digital resources and to improve the exploitation of large information spaces.

We have developed REC, a software environment that implements induced tagging and we report in this paper on our experiences with its introduction into an actual community, the challenges we have faced for its adoption and how these experiences have influenced our redesign of REC.

In what follows, we discuss work that is related to our research. Then, we introduce induced tagging with more detail, as well as its implementation via REC. Our initial experiences with the use of REC are presented next, followed by a discussion of the changes to our initial version of the software. Finally, we present an overview of our ongoing and future work as well as the conclusions that can be drawn from the work we have conducted so far.

2 Related Work

Among the studies of the issues and impact of tagging and its potential for generating useful recommendations or to improve the effectiveness of information retrieval, the following are salient.

An early but still current study of the structure of collaborative tagging was conducted by Golder and Huberman [3]. Among other findings, they reported regularities in user activity, tag frequencies, kinds of tags used and bursts of popularity in bookmarking. Stable patterns for tags assigned to certain URLs are attributed to imitation and shared knowledge. As our work progresses, we plan to use this work as a reference for analyzing the dynamics of our data sets.

The work by Stoilova and colleagues [8] developed a similarity measure for generating recommendations based on the personal bookmark files of users who donate them to help improve search for a web community. Their method takes advantage both of the hierarchical structure of the bookmark files of individual users,

[1] http://www.techcrunch.com/2007/09/06/exclusive-screen-shots-and-feature-overview-of-delicious-20-preview/, last accessed on August, 2008.

and of collaborative filtering across users. Though this work is not developed in the context of a tag management system, it provides leads to ways in which recommendations can be improved. In our current stage, the similarity measures we are applying are based only on the number of tags, the assigned ratings and the types of users.

For a survey of tagging systems and a discussion of their potential for knowledge organization and discovery, an interesting source is [4].

What motivates users to actively participate in social bookmarking is among the most important issues that are being studied by researchers. Marlow and colleagues [5] suggest five main motivators for tagging: future retrieval, contributing to the visibility of a resource, attracting attention, expressing opinions, competing with other users and leaving persistent marks. Out of these motivations, the first three (or a combination of them) appear as having particular importance for generating useful tags, possibly to be used as the basis for recommendations. The other two (competing and leaving marks) are features that could be promoted to increase user motivation. A related taxonomy of incentives, only oriented more specifically to image tagging, has been suggested by Ames and Naaman [1], whereas the various roles played by users when tagging has been explored by Thom-Santelli and colleagues [10].

A taxonomy of tagging styles that is particularly relevant for our work was proposed by J. Cañada [2]. After analyzing patterns in del.icio.us and Flickr, he found four such styles: (1) Selfish tagging, when users assign tags that are related only to their personal context and most likely are not meaningful to others; (2) friend-oriented, when tags are familiar only to a closed circle of people; (3) altruistic, when the assigned tags are consciously selected as the most descriptive and generally accepted; and (4) populist, when the assigned tags are enticing but intentionally deceive other users. Evidently, the greatest benefit for the community occurs if users tag altruistically, but it is also true that incentives need to be devised in order for this to happen, as most well-intentioned users tend to tag for themselves or only for a few people. Our approach aims to address precisely this need for incentives, as explained in the following section.

3 Rationale for Induced Tagging

We defined induced tagging [7] as a kind of social bookmarking with two key characteristics: (1) a well-defined group of participants are knowledgeable on the available resources and the background of the user community; and (2) tagging is required as part of that group's regular responsibilities as a reference team.

The concept of induced tagging has been proposed to take advantage of the shift that is occurring in the role of information professionals, particularly reference librarians. Personnel at the reference desk generally become knowledgeable and capable of locating and recommending resources from vast and dynamic collections in a timely manner. Increasingly, those resources are part of web-accessible digital collections. In the process of helping users, staff and users often discover resources that might be useful for supporting current or future tasks. It should be possible for information experts to bookmark those resources and share their findings with their

colleagues and the entire community they serve. Since these information experts are also familiar with the needs of the community and the terms most commonly used, altruistic tagging should happen naturally.

In induced tagging, there is a service component that involves mandatory tagging from information experts. Policies would be needed to require information experts in an organization to participate in collaborative tagging. Although all users are encouraged to tag, having a specialized group that does altruistic tagging continuously and applies tags consistently for extended time periods as part of their job, addresses concerns on the advantages of controlled vocabularies as well as incentive issues. Moreover, given the familiarity of the information experts with the needs of their user community, schemes can also be devised to facilitate the generation of personalized recommendations that are based on the tags assigned to relevant resources.

A critique to induced tagging is that it may contravene a popular view that is often used to explain the success of social bookmarking, which is best represented by the so-called wisdom of crowds [9]. According to this view, the collective intelligence that results from aggregating imperfect judgments (tags assigned by regular users) typically outperforms the judgments of experts (tags assigned by information professionals).

In contrast, recent experiments by Razikin and colleagues [6] show that the wisdom of crowds theory is not consistently supported. Their experimental evidence indicates that tags assigned by experts tend to be better descriptors for resource sharing and information retrieval than those assigned by the general public.

Our view is that induced tagging should get the best of both worlds: by overcoming the lack of incentives for altruistic tagging, the process should get a significant boost resulting in an accelerated growth of the collection of tags. Information experts know the vocabulary most commonly used by the user community, which should result in tags that are helpful for information retrieval and discovery, which in turn should attract community members to use existing tags and assign new ones. In the long run, the tags universe will reflect the wisdom of the community.

We believe induced tagging is especially well suited for stable learning communities such as those of universities or research centers that use digital libraries intensively and afford a regular staff that supports their information inquiries. With appropriate tools, tagging can become part of the regular support offered to users for their information retrieval needs.

We have developed REC, an environment for collaborative tagging that supports induced tagging and generates resource recommendations based on tags. REC is described in the following section.

4 REC

REC is an Ajax-based platform developed to explore induced tagging. REC provides a toolbar that can be added to a web browser so users may label resources in a minimally disruptive manner while they navigate around the web. The dialog box that pops up when the user is browsing the web and selects the "tag" option from the REC toolbar is illustrated in Figure 1. In this case, the user "waldo" is using the tags

"ENC", "SMCC" and "Computer Science" to describe a website, and is assigning a five-star rating to this site.

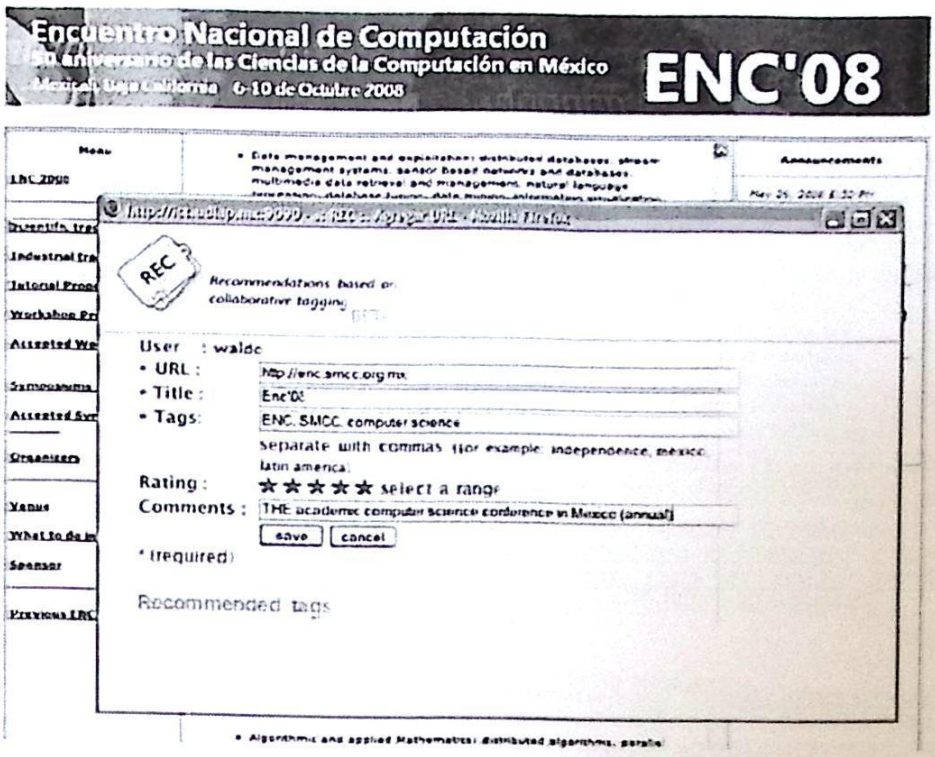


Figure 1. Tagging a website in REC.

We are particularly interested in studying how induced tagging works in the context of digital libraries, but REC actually can be used on any web-accessible resources. Additionally, users may manage tags and request recommendations using the main REC interface, which is illustrated in Figure 2 (a complete interface in English is still in progress at the time of this writing).

As illustrated, various tabs in REC allow the user ("waldo", in this case) to access recommendations that result from specific queries, as well as his tagged resources and the tags he has assigned. Additionally, at the bottom of this interface, REC displays a list of the most popular web resources that have been tagged by the user community. In the figure, as the user types a keyword, REC suggests tags that start with the characters being typed. The total number of resources that have been assigned each tag is indicated in the list. The user is also indicating that resources with at least a two-star ranking are preferred. After selecting "consumer behavior" from the list of suggestions, the user is presented with a list of resources that have been tagged and rated by information experts or other users. This is illustrated in Figure 3.

In Figure 3, the four resources displayed (out of the seven that were indicated in the list of suggestions in Figure 2), in this case scholarly papers in subscribed digital libraries, have been tagged as related to "consumer behavior" and rated with two or three stars. The title of each paper is displayed along with other tags that also have been used as descriptors. What is important here is that each of these tags can also be clicked upon, resulting immediately in a new list of resources, possibly with new tags that can lead to other resources, and so on. In Figure 3, if the user clicks on "social

behavior", which is a tag assigned to the first document in the list, he obtains a new set of three resources, as illustrated in Figure 4. At any moment during the interaction with REC, the user may choose to peruse the suggested documents or to continue to explore the related tags.

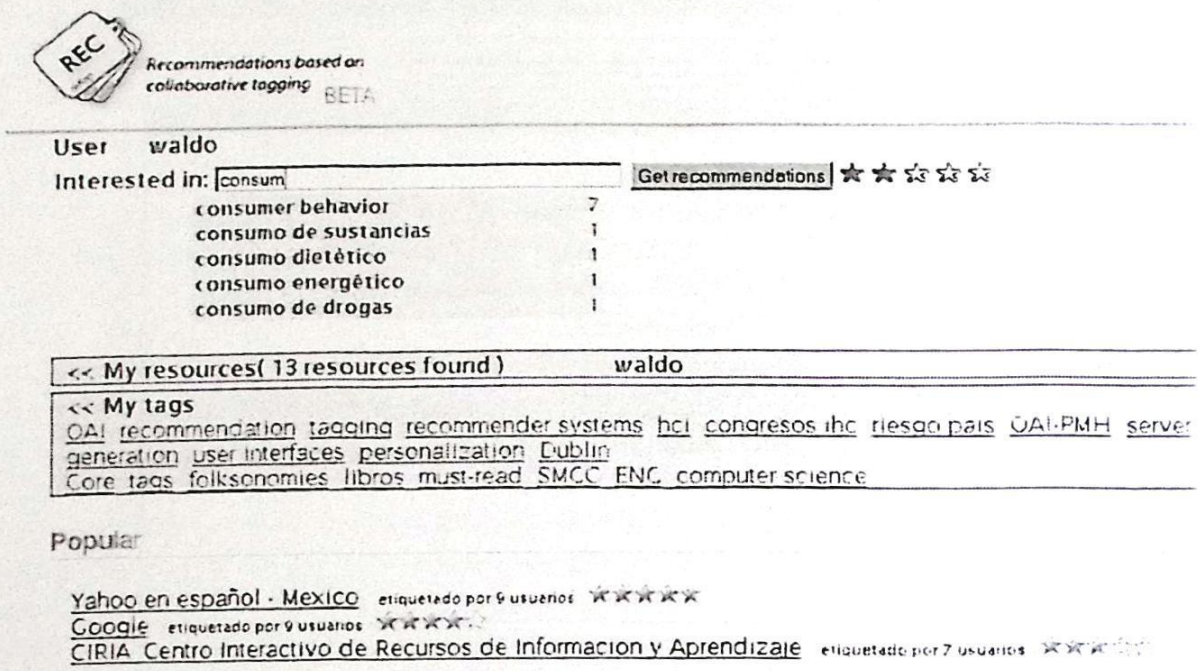


Figure 2. Main interface of REC.

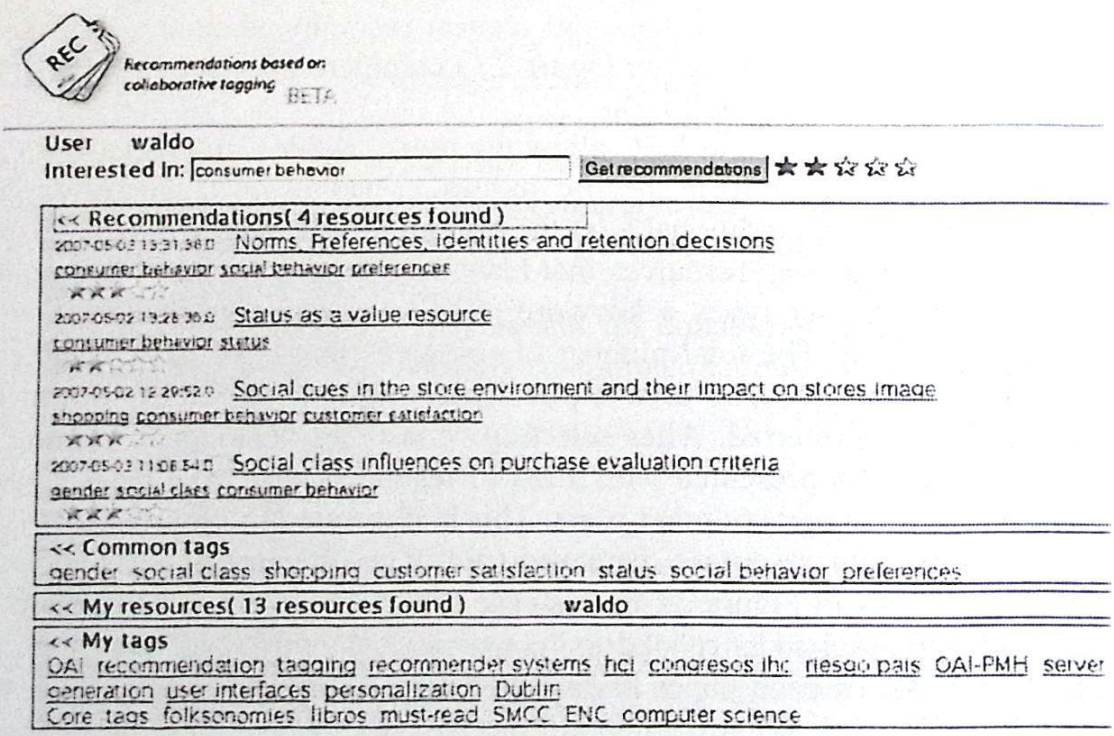


Figure 3. Obtaining recommendations in REC.

Resources tagged by a particular user may also be selected. A convenient list of all related tags in the list of resources is provided as an additional tab, and so is a list of all the resources tagged by the user. Lists of tags are presented as clouds in which font size is proportional to the weight of each terms. Term weight is determined as a combination of global term frequency, resource rating, and user category (regular or expert). All sections of the interface are dynamically updated upon selection of any of its elements (e.g., specific users or tags).

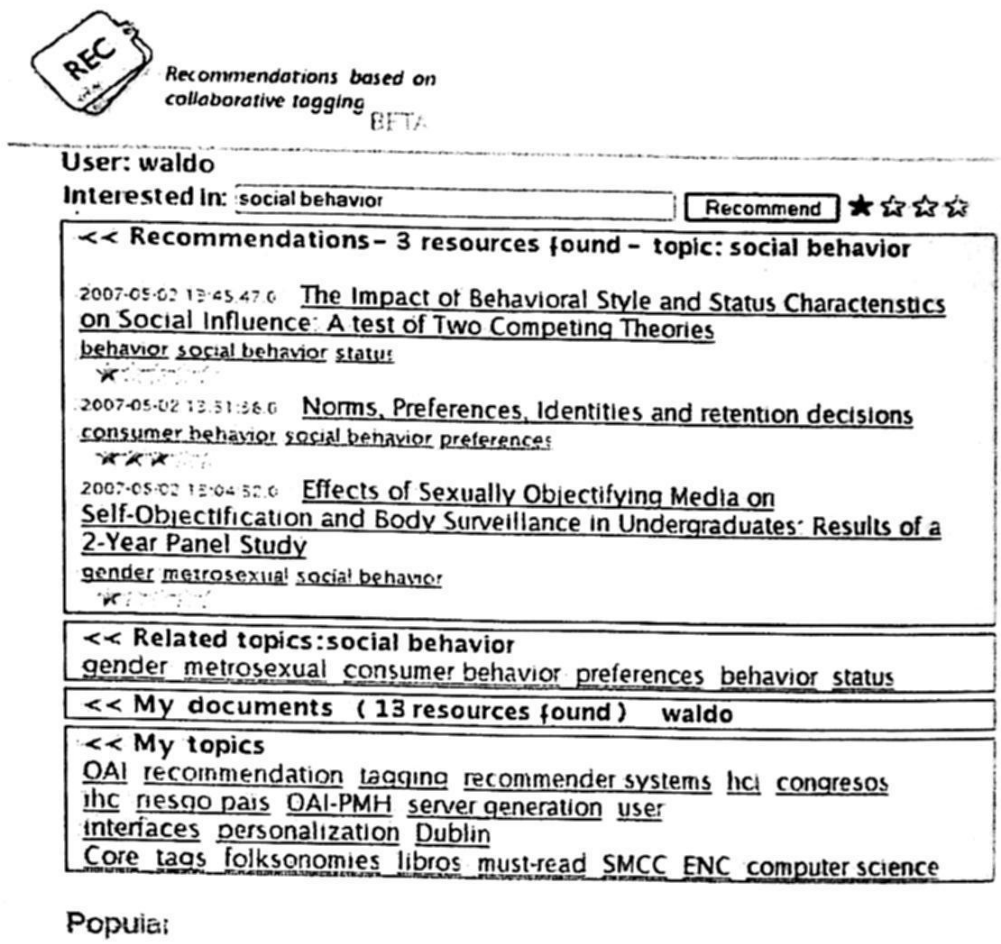


Figure 4. Exploring the tag space in REC.

5 Experiences with the Adoption of REC

REC is open to the public for evaluation and feedback at http://ict.udlap.mx. In order to assess its potential, we presented REC to our information experts and provided basic training for them to start labeling resources in the digital collections built or subscribed to by our institution.

Over the period of about six months, six information experts have been tagging resources using REC in addition to their regular duties, which include assisting users at the reference desk and via a virtual reference environment. We also have had students use REC as part of project assignments in a first-year college course called "Information Culture", which is taken by students from all majors and is intended to

provide orientation for them to learn about the information services and resources provided by the university. Although students also have tagged resources, they mostly have been users of existing tags. Usability studies of REC's interface have been conducted by students of courses on human-computer interaction. We also have a number of additional users who have registered to explore the tagging environment.

Collections provided by different vendors typically are available through highly heterogeneous interfaces. Finding related documents on those collections as part of a student or faculty research project often becomes a time-consuming and frustrating task. When tags are assigned to those documents by the information experts (or other users), REC becomes a uniform interface that establishes tacit links that are used to discover resources comprised by diverse collections that otherwise would go unnoticed. Participating users have been able to fetch documents carefully selected by information experts, by just following links recommended by REC, and remaining essentially unaware of the specific collections that contain the documents. Obtaining such documents normally would entail formulating complex queries. Tag clouds have resulted also in a discovery mechanism, as they invite users to explore relevant tags and their associated resources.

Although the size of our tags collection is still modest, our observations have helped shape new features and requirements for our tagging environment to provide support for improved exploitation of vast digital collections. Out of over 11,000 tags that describe more than 4,150 web resources, about 76% have been assigned by information experts. This portion of the tag collection has a high descriptive value and is expressed in the terms most likely used when requesting recommendations, as our staff has been consciously considering the users' needs in the tagging process. This shows that the role of designated personnel is fundamental for describing the collections. Given that each information professional has an area of expertise, the available collections are being covered rapidly, but there is still little overlap among the tags being assigned. Thus, the current average number of tags per resource is only 2.17 (standard deviation of 1.79). We expect these figures to change significantly as REC becomes more popular in our community.

One of the requirements derived from our initial observations was that participants wanted to be able to recommend resources to specific users. They also wanted to be able to define groups of users and suggested interface features that were similar to other familiar environments, such as those in their social networking systems. This suggested the version of REC described next.

6 REC on Facebook

We believe the combination of social networks and induced tagging can provide an even more powerful environment for exploiting information spaces. Not only are our users familiar with various social networking systems, but they also are keen on extending their network of contacts, which could just as well include information experts that could help them accomplish academic endeavors without having to leave their social environment of choice.

We developed a version of REC that can be added as an application of Facebook, one of the most popular social networking systems. Users of Facebook who install REC can obtain or give recommendations within the same familiar environment. More specific registration as a user of REC is needed only to enable tagging and is required only the first time a user tags resources. Both versions of REC share the same databases, so tags and recommendations managed by either of them are equivalent.

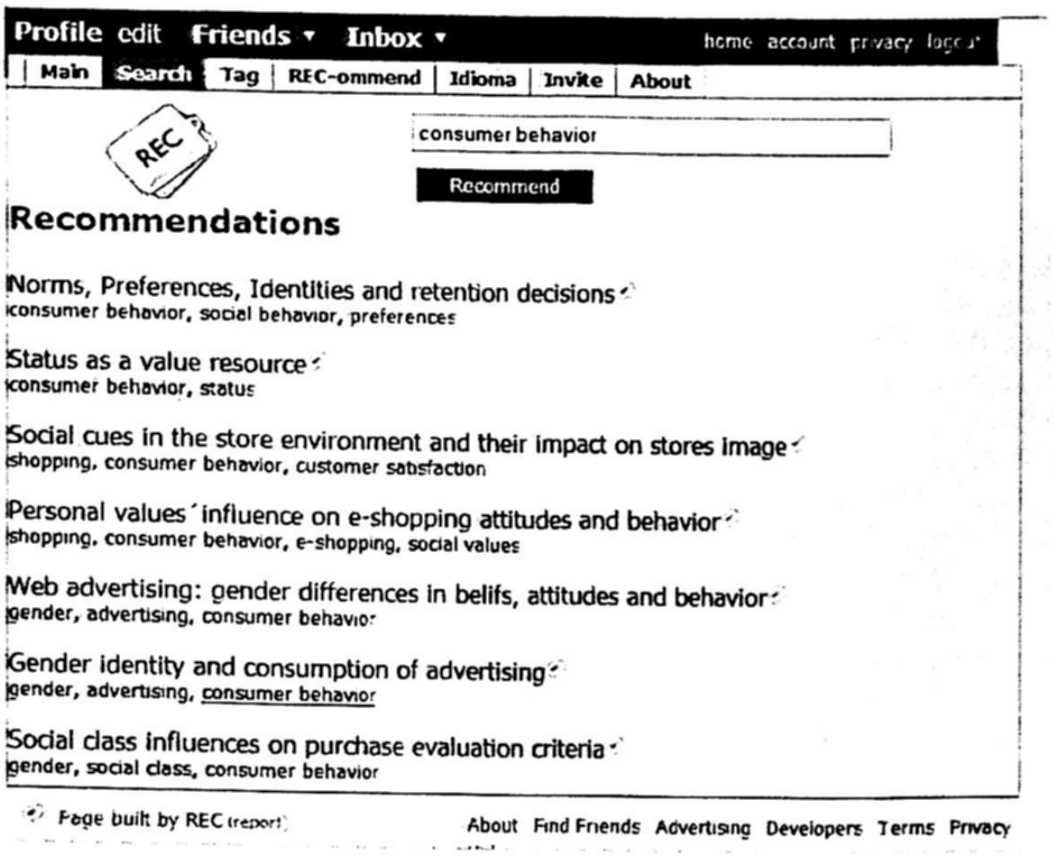


Figure 5. REC as a Facebook application.

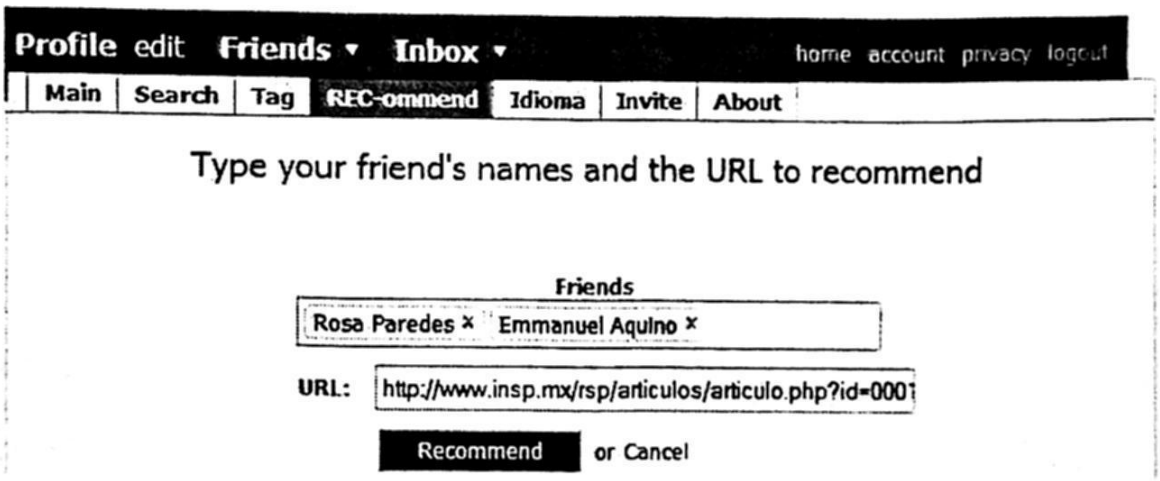


Figure 6. Recommending resources to friends.

Figure 5 illustrates the interface of REC as a Facebook application. In this case, the recommendations displayed correspond to those in Figure 3; only they have not been filtered by rating. As illustrated in the figure, other functionality offered by REC is available via tabs that include options for tagging resources, recommending resources to friends, switching to another language (Spanish or English in the current version), and inviting friends to become users of REC. By clicking on the small icon that appears next to each document in the list, users can generate recommendations for their friends. Figure 6 is an example of how recommendations of resources can be sent to Facebook friends.

7 Ongoing and Future Work

At the time of this writing we are deploying REC for its use by our entire learning community, which comprises about 7,000 students and 500 faculty members. We also are sharing the software for installation by member institutions of the Open Network of Digital Libraries (ONeDL, http://www.onedl.org.mx) and the Institutional Network for Library Cooperation (also known as the Amigos Group, http://ciria.udlap.mx/amigos). Whereas REC is, as mentioned earlier, open to the general public, we particularly are interested in conducting formal experimentation and further observation of the impact of induced tagging in the context of digital libraries and academic communities. We will conduct longitudinal studies aided by surveys, questionnaires and activity logs.

We also are working on functionality enhancements such as improved visualization of the tagging spaces, so users can discover relationships among the resources in the digital collections by relying on graphical representations. This visualization will provide filtering mechanisms so users may include or leave out tags by date, tagger or Internet domains, among other criteria.

Both on the web-based and Facebook versions of REC, we plan to provide options for users other than the "official" information professionals to participate as domain experts. In the realm of social networks, we plan to produce an implementation of REC that takes advantage of OpenSocial[2], a specification that defines a common API for social applications across multiple websites . OpenSocial is being implemented by other popular social networking service providers, such as Friendster, hi5, LinkedIn, MySpace, and orkut. Our goal is to make REC accessible from as many social networking platforms as possible.

8 Conclusions

We have made progress in exploring the notion of induced tagging, a technique for accelerating the construction of tag collections that may effectively support the exploitation of vast digital collections. Progress includes the implementation and refinement of a social bookmarking environment called REC, which has been used by

[2] http://www.opensocial.org, last accessed on August, 2008.

a group of information experts to assign tags to scholarly web resources. Their tagging style is altruistic as they are familiar with the vocabulary and needs of the target user community and tagging has been included in their regular duties. Our initial experiences with REC motivated the development of a version of this software that can be installed as an application of Facebook, the popular social networking environment. By combining collaborative induced tagging and social bookmarking we are contributing to the development of tools that will support users in at least four areas: (1) exploiting the information resources they have available, (2) organizing their information spaces in a personalized manner; (3) discovering digital resources that otherwise would remain hidden, and (4) participating in the social construction of a tag collection that reflects the wisdom of the community.

Acknowledgments. We would like to thank the group that participated at the April 2008 Seminar Series of the Information Systems and Interactive Systems Group at the Manchester Business School, organized by Víctor M. González, for valuable feedback on the notion of induced tagging.

References

1. Ames, M. and Naaman, M.: Why we tag: motivations for annotation in mobile and online media. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (San Jose, California, USA, April 28 - May 03), pp. 971-980. ACM, New York, NY (2007)
2. Cañada, J.: Tipologías y estilos en el etiquetado social. Blog entry, available at http://www.terremoto.net/tipologias-y-estlos-en-el-etiquetado-social/, (2006). (In Spanish).
3. Golder, S., Huberman, B. A.: The structure of collaborative tagging systems. HP Labs Tech. Report. (2005). (http://hplabs.com)
4. Macgregor, G., McCulloch, E.: Collaborative tagging as a knowledge organisation and resource discovery tool. Library Review, 55, 5 (2006)
5. Marlow, C., Naaman, M., Boyd, D., and Davis, M.: HT06, tagging paper, taxonomy, Flickr, academic article, to read. In: Proceedings of the Seventeenth Conference on Hypertext and Hypermedia (Odense, Denmark), pp. 31-40. HYPERTEXT '06. ACM, New York, NY (2006)
6. Razikin, K., Chua, A. Y. K., Lee, C. S., Goh, D. H.-L.: Can social tags help you find what you want? In: Proceedings of the European Conference on Digital Libraries, (ECDL 2008, Aarhus, Denmark). *In print* (2008)
7. Sánchez, J. A., Arzamendi-Pétriz, A., Valdiviezo, O.: Induced tagging: Promoting resource discovery and recommendation in digital libraries. In: Proceedings of the Joint Conference on Digital Libraries (JCDL 2007, Vancouver). pp. 396-397 (2007)
8. Stoilova, L., Holloway, T., Markines, B., Maguitman, A. G., and Menczer, F.: GiveALink: mining a semantic network of bookmarks for web search and recommendation. In Proceedings of the 3rd international Workshop on Link Discovery (Chicago, Illinois, August 21 - 25, 2005). LinkKDD '05, pp. 66-73. ACM, New York, NY (2005).
9. Surowiecki, J.: The Wisdom of Crowds, Doubleday, New York (2004)
10. Thom-Santelli, J., Muller, M. J., and Millen, D. R.: Social tagging roles: publishers, evangelists, leaders. In *Proceeding of the Twenty-Sixth Annual SIGCHI Conference on Human Factors in Computing Systems* (Florence, Italy, April 05-10), pp. 1041-1044, ACM, New York, NY (2008)

Networking

First Experiences with BlueZ

Sukey Nakasima-López[1], Francisco Reyna-Beltrán[2], Arnoldo Díaz-Ramírez[2], and Carlos T. Calafate[3]

[1] Graduate School, Cetys Universidad
Calzada Cetys s/n, Mexicali, Mexico
sukey.nakasima.lopez@gmail.com
[2] Department of Compuer Systems, Instituto Tecnológico de Mexicali
Ave. Tecnológico s/n, Mexicali, Mexico
freyna, adiaz@itmexicali.edu.mx
[3] Department of Computer Engineering, Polytechnic University of Valencia (UPV)
Camino de Vera s/n, Valencia, Spain
calafate@disca.upv.es

Abstract. The low cost and low battery consumption of *Bluetooth* devices, allied with a plethora of novel functionalities, has promoted the widespread adoption of this technology. In this paper we enter into the *Bluetooth* Technology, that like an emergent technology, it is conceived to be the best option for the wireless communication, combined to the ample range of possibilities of development of applications, positioning itself like the technology of the future for the movable devices. Our contribution consists of impelling the development of mobile applications under this technology and shares our experiences. We have chosen *BlueZ*, a protocol stack with License *GPL* for *Linux*, and using *C*, *C++* and *Qt* as programming languages. We hope to promote its use and show its tools for application development at basic level for new *BlueZ/C++* users.

Keywords: Bluetooth, BlueZ.

1 Introduction

The necessity of communication nowadays is a prevailed subject, for businesses or for personal reasons, it is always necessary to be in constant communication. In an enviroment where the technological devices become indispensables in our lifes, because of little effort that these require, the facility of its uses and in some cases its low costs, it is difficult to imagine us without a cellphone, palmtop, laptop or any other device that allows us to be in contact with the exterior world. At this life-rhythm, it is elementary to be able to free us of cables and use wireless connections of short-range in order to facilitate the demand of connectivity between the devices, so *Bluetooth* is the simplest option.

Bluetooth is a wireless standar available in the whole world, it connects mobile telephones to each other, laptops, MP3 players and a lot of other devices. This

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 97-114

Received 14/06/08
Accepted 01/08/08
Final version 11/09/08

technology provides great efficiency and saving of costs for home and businesses users; allowing the replacement of cable, the facility to share files, wireless synchronization and connectivity to internet, also, thanks to its great acceptance, a *Bluetooth* device can be connected with almost any other compatible device in its proximities eliminating the borders anywhere of the world.

In other way, for the best advantage of all these qualities already mentioned about *Bluetooth*, it is elemental have the necessaries tools that help to the efficient development of applications for this technology. In this sense, *BlueZ* is an alternative to consider, starting from the fact that it is free-software and included in the Linux core since the version 2.4. Unfortunately almost does not exist documentation about *BlueZ* or about the APIs to development applications with *BlueZ*.

In this article we describe briefly the features of the *Bluetooth* technology and we will show the *BlueZ* protocols stack by Linux, thus examples of its uses, so that any user who begins to developing under this standard counts by a minimal reference and understands better its operation.

2 Bluetooth technology

The *IEEE 802.15.1* standard [8], also known as *Bluetooth*, is an open standard for the wireless connectivity that allows the data and voice transference between the communication devices and PC's giving facility to the users for create *Wireless Personal Area Networks (WPANs)* and *Ad Hoc networks*, impelling a greater integration of the *Bluetooth* technology to MANET networks (*Mobile Ad Hoc Network*).

2.1 Bluetooth features

Bluetooth operates in a free license band, enlarging the possibilities of its use, this industrial, scientific and medical band (ISM) is between 2.4 and 2.485 Ghz, using an extended spectrum, frequency hops, full-duplex signal in a nominal rate of 1600 hops/sec. Actually there are three available "classes", the *Bluetooth* devices have a rank operation from 3 - 300 feet (1-100m), depending the class of the device and adapted for the needs of the user.

In order to reduce to the minimum any interference, the *Bluetooth* technology makes use of a capacity denominated adapted frequency hops (AFH); which was designed to detect other devices in the spectrum and to avoid the frequencies that are in use, at this way the signal hops between 79 frequencies in 1 Mhz intervals, generating a high degree of immunity to interferences, respecting to the voice and data transferences, *Bluetooth* supports up to 3 synchronous channels of voice of 64kbls each one and asynchronous data up to 723,2 kb/s asymmetric or 433.9 kb/s symetric [3].

2.2 Bluetooth profiles

So that device can use the *Bluetooth* wireless technology, it must know how to interpret the *Bluetooth profiles*, as shown in Figure 1 that describes the different possible applications. These profiles are guides that specify the adjustment of stack parameters as well as the required features and procedures so that the devices equipped with *Bluetooth* technology communicate to each other. Thanks to the exclusive concept of "profiles", it is not necessary to install controllers in the *Bluetooth* devices. All the profiles supported by *Bluetooth* are defined in its web site [9].

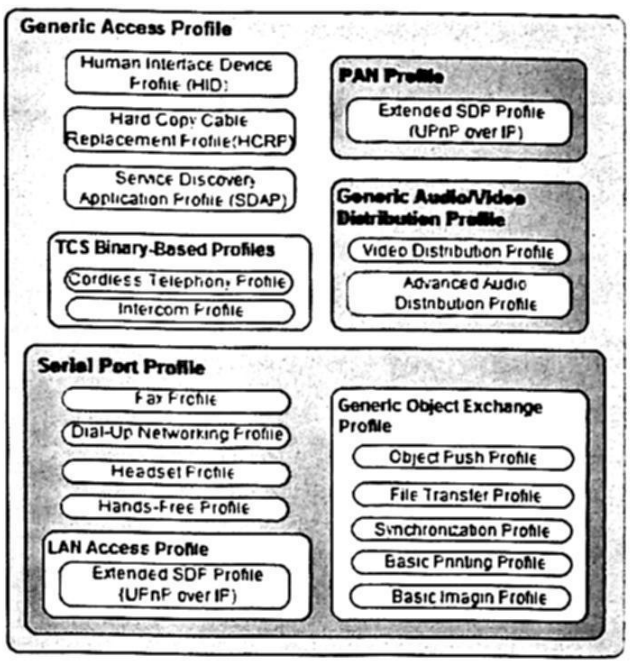


Fig. 1. *Bluetooth profile structure.*

2.3 Network topology

Piconets y scatternets. The *Bluetooth* wireless links are formed in the context of a Piconet. Piconet is a group of two to eight devices that occupy the same physical channel (unique for each piconet), consisting of a single master device and one or more slaves, where slave devices are synchronized to the same clock and a specific pattern of frequency hops provided by the master device. Besides, a device can belong at the same time as more of a Piconet, creating what is called a Scatternet, as seen in Figure 2.

Operational procedures. *Bluetooth* uses a search procedure (*Inquiry*) to discover or to be discovered by other near devices, as well as to discover the services that these nodes offer. Followed to this, the paging procedure is used to contact between both devices.

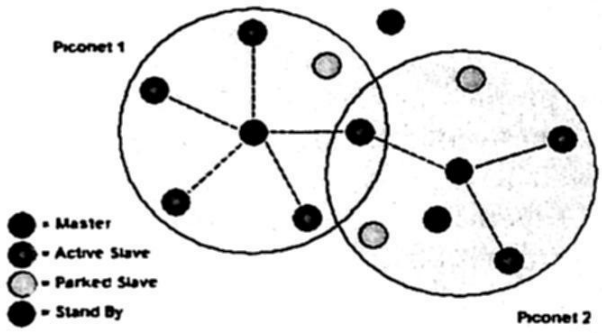


Fig. 2. *Bluetooth scatternet diagram.*

2.4 Bluetooth specification

The core specification. The core specification define all the layers of the *Bluetooth protocols stack* (Figure 3), which is structured in four layers with associated protocols defined by the specifications or *Bluetooth* profiles. The three inferior layers radio, baseband and Link Manager are grouped in a subsystem denominated *Bluetooth Controller (Module)*, while the *L2CAP* layer, services layer and the superior layers are known as *Host Bluetooth*. This grouping of the core layers requires a physical communication interface between the *Bluetooth* controller and the *Host Bluetooth*, this interface is known as *HCI (Host Controller Interface)*. The *Bluetooth* specification makes possible the compatibility between different *Bluetooth* systems through of the definition of protocol messages that interchange it between the equivalent layers. Also it determines a common interface between the controllers and *Bluetooth* Hosts to make compatible the different subsystems.

Protocols stack. The *Bluetooth protocol stack* is divided in two zones, each one is implemented in different processor:

The *Bluetooth Module (hardware)* is the responsible of the tasks related to the information transferences through radio frequency interface. The *Host Bluetooth (software)* is the responsible of the part related to the layers superiors of connection and application.

On the layer of specific *Bluetooth* protocols, each manufacturer can implement his propietary protocols layer of application. by this way, the open specification of *Bluetooth* expands the number of applications that can be benefited from their capacities. Even though, the *Bluetooth* specification demands that, in spite of the existence of different propietary application protocol stacks, the interoperability must exist between devices that implement different stacks.

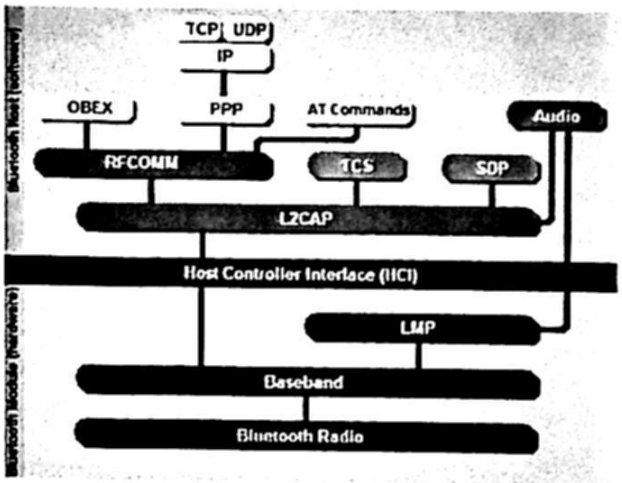


Fig. 3. *Bluetooth Protocols Stack.*

3 BlueZ

In the following paragraphs the main characteristics of *BlueZ* and their tools will be commented. There is special emphasis in the pand daemon beacuse it is the base of the development.

3.1 BlueZ definition and features

BlueZ [7] is the *Official Linux Bluetooth Protocol Stack*; it was developed initially by *Qualcomm* and at present is released under the *General Public License (GPL) [4]* that means that can be copied, studied, modified and redistributed freely.

BlueZ is part of the Linux official kernel since the version 2.4, therefore, included in any modern distribution of Linux; it will be not necessary to install something. Some of the main characteristics of *BlueZ* are:

- Flexible, efficient and modular architecture.
- Support for multiple *BlueZ* devices.
- Multitask data processing.
- Hardware abstraction.
- Standard socket interface to all layers.
- Multi-platform: x86 (single and multi-processor), SUN, SPARC, ARM, PowerPC, Motorolla, DragonBall.
- Operation in all the Linux distributions: RedHat, Debian, Suse, etc.
- Great quantity of supported devices (PCMCIA, UART, USB).
- Supports *L2CAP*, *SDP*, *RFCOMM* and *SCO*.
- Availability of a *Bluetooth* emulator and devices of configuration and test.
- Support for the following profiles of use: GAP, DUN, LAN, SPP, PAN, Headset, OBEX (FTP), OBEX (OPP).
- Mailing lists of participating, with developers anywhere in the world contributing to the support and programming with *BlueZ*.

The main disadvantage of *BlueZ* is that there is not enough documentation. This can be solved directly studying the source code, but this takes a lot of time and could be complicated for the new users.

3.2 BlueZ packages

BlueZ is distributed in a set of packages, although the core depends on the distribution of kernel Linux that we are using. For the previous versions to version 2.4 that do not include the *Bluetooth* functionality exists patches [6]. Besides the support of the core, the packages that can be used based on the final needs of the users are:

bluez-libs: Necessary libraries for the development of applications and the rest of *BlueZ* packages and applications that link dynamically to the libraries.
bluez-utils: Control applications for the *Bluetooth* devices. Necessary to make inquiry or general communications.
bluez-sdp: It contains the libraries, tools and the *SDP* server (sdpd) that conform all the *SDP* functionality .
bluez-pan: Programs, daemons and scripts necessaries for the profiles DUN, LAN and BNEP-PAN.
bluez-hcidump: Useful orders to debug and to study the general operation of the devices *Bluetooth* using *HCI*.
bluez-hciemu: It contains the emulator. It allows the programmers to test their code without a real *Bluetooth* device.
bluez-bluefw: It contains the firmware of several kind of *Bluetooth* devices.

All these packages can be downloaded from the official *BlueZ* site [7] in the section of downloads.

3.3 BlueZ tools

As already commented, the documentation of this protocol stack is non-existent and therefore the knowledge about this API has been realized from the own source code. Fortunately, the *BlueZ* core comes accompanied by a set of tools that allows to execute the *Bluetooth* functions implemented in the protocol stack from a shell or console orders. The first step to determine which are the functions that interest to us is studying of these tools:

hciconfig: Configure local *Bluetooth* devices.
hcid: *HCI* interface daemon.
hcitool: Link manager with other *Bluetooth* devices, detection of remote devices and name resolutions among others fuctions.
hcidump: Local sniffer for the *HCI* traffic, either incoming or outgoing, by the *Bluetooth* device installed in the system.
l2ping: Send request "echo request (ping)" in *L2CAP* level.

sdptool: *SDP* manager, dicovering of *Bluetooth* services in remote devices.
sdpd: Daemon of the service discover protocol *SDP*. It manages to provide access to the local *Bluetooth* services.
rfcomm: Manager of connections rfcomm.
pand: Manager of PAN (*Personal Area Network*) connections.

For greater information about command-functions that provide us these tools, we can review the manuals that linux provides to us in the console. For example, look the fuctions that sdptool privide to us:

```
#: man sdptool
```

The command-functions provided by these tools are relatively very simple to use, the manual provides information of each command, as well as of its syntax.

4 Sockets

Sockets [2] are a fundamental tool in the communication between devices, throughout the next examples we will use sockets in repeated occasions to make connections. For such reason is good idea to know a little bit of sockets implementation before getting to program with them.

The function `socket()` returns a socket descriptor, which we will be able to use soon for calls to the system. If it returns -1 , an error has taken place.

```
int socket(int Dominio,int Tipo,int Protocolo);
```

- Domain: It defines the property to a group of socket that we want to use, that is, you can use AF_INET (for protocols ARPA of Internet), AF_UNIX (protocols that allow internal communication of the system) and AF_BLUETOOTH (protocol for the communication between devices that support this technology).
- Type: It refers about to the class of socket that we want to use, is this of datagrams UDP or data stream TCP. We will use SOCK_SEQPACKET (is used to indicate a socket with reliable datagram-oriented semantics where packets are delivered in the order sent).
- Protocol: It indicates the protocol that will allow us the information transference (BTPROTO_L2CAP).

Once obtained the socket descriptor, will be necessary to associate it with a port, for this reason, we will make use of the function `bind()` [4].

```
int bind(int fd, struct sockaddr *my_addr,int addrlen);
```

Already established the relation between socket and port we can to make a connection through the function `connect()`. This function is used to connect to a port defined in a IP address.

```
int connect(int fd, struct sockaddr *serv_addr,
int addrlen);
```

Now to be able to obtain that our device remains awaiting for some other device can be connected to, it will have to use the function `listen()`, which has main function remain awaiting the incoming connections.

[4] For more information about the functions parameters, see the reference [2].

```
int listen(int fd,int backlog);
```

Already having a function that is awaiting for a connection, when finally some device is connected, is necessary to accept this connection, for it we will use the function `accept()`.

```
int accept(int fd, void *addr, int *addrlen);
```

We already have all the previous elements, you will be able to use of the functions `send()`, `recv()`, `write()` and `read()` for the exchange of information through sockets descriptors .

For the development of the next applications, we have made use of the protocol *L2CAP*, that allows the interchange of packages with or without connection-oriented, but as we mentioned previously are diverse protocols of which you can make use according to your needs.

5 Bluetooth applications examples

The development is located for Linux platforms, doing use of the *BlueZ* protocol stack *BlueZ* that comes included in the linux kernel. For which, it is elementary that the core of the application to be developed entirely in language *C++* programming, since it is the used one in *BlueZ*, whereas the graphical interface is recommended to be realized in *C++* language using the library *Qt* de Trolltech [1].

As previously mentioned (Section 3), the *BlueZ* protocol supports different transport protocols as *RFCOMM, L2CAP, SCO*, etc., which use a programming structure (interface) based on *sockets* (Section 4) to be able to communicate, Unlike those protocols, *HCI* made it easier to use thanks to functions and specific-use commands that it provides us. In the subsequent examples the protocols *L2CAP, SDP* and *HCI* are used, showing the difference between its uses and capabilities.

This chapter starts from a simple scaning application to a client/server using a *SDP* service. Logically programming in *C++* for *BlueZ* has a much broader scope of which we will present, but, by the propose of our work some examples at basic level for new *Bluetooth-BlueZ* and *C++* users are shown.

5.1 Searching nearby devices (scan)

The scan is a primary function at the time of schedule with *Bluetooth* devices, because it provides us information of the nearby devices with which we can interact. this tool is defined by default as a *BlueZ* tool (hcitool scan), also is recommended to understand its operation on the code, since it is highly required in most applications that we develop. A way to get this code is exploring the *BlueZ* libraries [5], where not only the scan function is, we can find all the functions defined by the API (Section 3.3); is necessary to identify the .c file with the

[5] It is recommended not to modify the original libs, you can download them from www.bluez.org/downloads and make with them your tests.

functions, isolate the code and make our tests to help us to understand its functioning.

The following example [5] will search for nearby *Bluetooth* devices, providing us its name and *Bluetooth* address. This program will use *HCI*, in subsequent examples were used *L2CAP* and *SDP*.

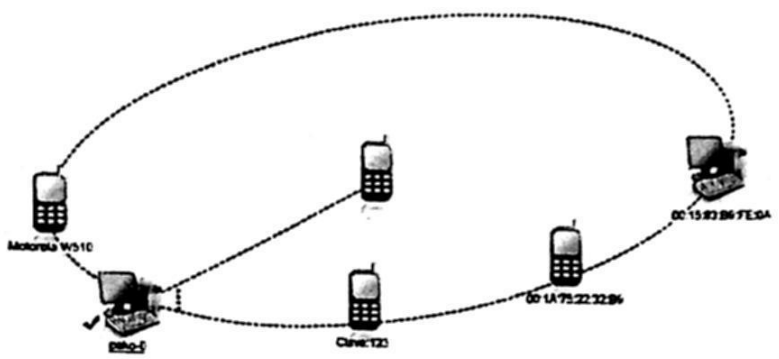


Fig. 4. *Scanning nearby devices.*

Note: *The explanation of the examples code will be given at the end of each.*

scan.c

```
#include <stdio.h>
#include <stdlib.h>
#include <unistd.h>
#include <sys/socket.h>
#include <bluetooth/bluetooth.h>
#include <bluetooth/hci.h>
#include <bluetooth/hci_lib.h>
int main(int argc, char **argv)
{
 inquiry_info *ii = NULL;
 int max_rsp, num_rsp;
 int dev_id, sock, len, flags;
 int i;
 char addr[19] = { 0 };
 char name[248] = { 0 };

 dev_id = hci_get_route(NULL);
 sock = hci_open_dev( dev_id );

 if (dev_id < 0 || sock < 0) {
 perror("opening socket");
 exit(1);
 }

 len = 8;
 max_rsp = 255;
 flags = IREQ_CACHE_FLUSH;
 ii = (inquiry_info*)malloc(max_rsp * sizeof(
 inquiry_info));
 num_rsp = hci_inquiry(dev_id, len, max_rsp, NULL,
 &ii, flags);

 if( num_rsp < 0 ) perror("hci_inquiry");
```

```
for (i = 0; i < num_rsp; i++) {
ba2str(&(ii+i)->bdaddr, addr);
memset(name, 0, sizeof(name));

if (hci_read_remote_name(sock, &(ii+i)->bdaddr,
sizeof(name), name, 0) < 0)
strcpy(name, "[unknown]");
printf("%s %s\n", addr, name);
}

free( ii );
close( sock );
return 0;
}
```

To compile the program is necessary to link the *Bluetooth* library for use its functions.

```
# gcc -o scan scan.c -lbluetooth
```

Running the program...

```
# ./scan
```

There are differents predefined structures for schedule *C++* with *BlueZ*, these structures will be required for some functions and we must know its purpose. One of the most often used is bdaddr_t; which is referred to store and manipulate the *Bluetooth* devices addresses.

```
typedef struct {
 uint8_t b[6];
} __attribute__((packed)) bdaddr_t;
```

These addresses can be converted between strings and bdaddr_t structures or the opposite through the following functions:

```
int str2ba( const char *str, bdaddr_t *ba );
int ba2str( const bdaddr_t *ba, char *str );
```

When we schedule, is possible to have multiple *Bluetooth* devices in our computer, therefore, it is required to specify which *Bluetooth* adapter we are going to use for allocating system resources. These adapters are identified by a number starting from 0.

If you already know the *Bluetooth* local adapter (*Bluetooth address*), the following function returns the resource number of the *Bluetooth* adapter address passed in as a parameter.

```
int dev_id = hci_devid( "00:32:56:27:6B:9A" );
```

But, if only we have a *Bluetooth* adapter or no matter which we use, the following function (Null) retuns the resource number of the first *Bluetooth* adapter available.

```
int dev_id = hci_get_route( NULL);
```

Once the *Bluetooth* adapter is defined, is required to open a socket using *hci_ open_ dev*, this function opens a socket connection to the microcontroller (for controlling it) on the specified local *Bluetooth* adapter. "Attention",the socket is not a connection to a remote *Bluetooth* device.

```
socket = hci_open_dev( int dev_id );
```

If there are errors when opening the socket, the function returns -1 and sets errno [6], if there are no problems, returns a handle to the socket.

When the socket is already open, the scan (*inquiry*) starts using the function:

```
int hci_inquiry(int dev_id, int len, int max_rsp,
```

[6] Most of the functions return -1 in case of error.

```
const uint8_t *lap, inquiry_info **ii, long flags);
```

This function does not even use the socket, we use the resource number of the *Bluetooth* adapter (int dev_id), the duration of the inquiry (len * 1.28sec.), the maximum number of responses (max_rsp), a structure for storing the info of inquiry (inquiry_info ** ii) and flags for indicate whether or not to use previously discovered device information or to start a fresh (IREQ_CACHE_FLUSH: cache flushed, 0: results of previous inquiries may be returned). If there is not an error, the devs parameter are stored in an predefined array of inquiry_info structures.

It is often easier to identify devices by its friendly name (nickname) than by its direction, hence, the function hci_read_remote_name provides us the remote device name in a char (name). Is necessary to indicate the socket and the *Bluetooth* device address.

```
int hci_read_remote_name (int socket, const bdaddr_t
 * ba, int len, char * name, int timeout);
```

finally free the memory used by * ii and close the socket.

```
free( ii );
close( socket );
return 0;
```

5.2 Basic client-server

The example above applies only the scanning *HCI* function, now we will see the incorporation of functions *L2CAP* to demonstrate how to establish an *L2CAP* channel and transmit a string of data like a server-client application.

server.c

```
#include <stdio.h>
#include <string.h>
#include <sys/socket.h>
#include <bluetooth/bluetooth.h>
#include <bluetooth/l2cap.h>
int main(int argc, char **argv)
{
 struct sockaddr_l2 loc_addr={0}, rem_addr={0};
 char buf[1024] = {0};
 int s, client, bytes_read;
 socklen_t opt = sizeof(rem_addr);

 // allocate socket
 s = socket(AF_BLUETOOTH, SOCK_SEQPACKET, BTPROTO_
 L2CAP);

 // setting parameters
 loc_addr.l2_family = AF_BLUETOOTH;
 loc_addr.l2_bdaddr = *BDADDR_ANY;
 loc_addr.l2_psm = htobs(0x1001);

 // bind socket to port 0x1001 of the first
 //available bluetooth adapter
 bind(s, (struct sockaddr *)&loc_addr, sizeof
 (loc_addr));

 // put socket into listening mode
 listen(s, 1);

 // accept one connection
```

```
client = accept(s, (struct sockaddr *)&rem_addr,
&opt);

ba2str( &rem_addr.l2_bdaddr, buf );
fprintf(stderr, "accepted connection from %s\n"
, buf);
memset(buf, 0, sizeof(buf));

// read data from the client
bytes_read = read(client, buf, sizeof(buf));

if( bytes_read > 0 ) {
printf("received [%s]\n", buf);
write(client, "hello client!", 15);
}
close(client);// close connection
close(s);
}
```

client.c

```
#include <stdio.h>
#include <unistd.h>
#include <sys/socket.h>
#include <bluetooth/bluetooth.h>
#include <bluetooth/l2cap.h>
int main(int argc, char **argv)
{
struct sockaddr_l2 addr = { 0 };
int s, status,bytes_read;
char server[18], buf[1024] = { 0 };

if(argc < 2) {
fprintf(stderr, "usage: %s <bt_addr>\n",
argv[0]);
return 1;
}

//asignate the server address to server
strncpy(server, argv[1], 18);

// allocate a socket
s = socket(AF_BLUETOOTH, SOCK_SEQPACKET,
BTPROTO_L2CAP);

// set the parameters for connection
addr.l2_family = AF_BLUETOOTH;
addr.l2_psm = htobs(0x1001);
str2ba( server, &addr.l2_bdaddr );

// put socket into listening mode
listen(s, 1);

// connect to server
status = connect(s, (struct sockaddr *)&addr,
sizeof(addr));

// if connect succefully
if( 0 == status ) {

// send a message to server
status = write(s, "hello server!", 15);

// read data from the client
bytes_read = read(s, buf, sizeof(buf));
if( bytes_read > 0 ) {
```

```
    printf("received [%s]\n", buf);
    }
    }

    if( status < 0 ) perror("Error:");
    close(s);
    return 0;
}
```

For connecting we have to open a *L2CAP* socket;

```
s = socket(AF_BLUETOOTH, SOCK_SEQPACKET, BTPROTO_L2CAP);
```

The first parameter should still be AF_BLUETOOTH, but the next two parameters should be SOCK_SEQPACKET and BTPROTO_L2CAP, respectively.

L2CAP sockets use the struct sockaddr l2 addressing structure:

```
struct sockaddr_l2 {
 sa_family_t l2_family;
 unsigned short l2_psm;
 bdaddr_t l2_bdaddr;
};
```

The first field, l2 family, should always be AF_BLUETOOTH, l2_bdaddr denotes the address of either a server to connect to, a local adapter and port number to listen on, or the information of a newly connected client, depending on context. and the l2_psm field specifies the *L2CAP* port number to use.

Running the client program we must indicate the server address to connect, for example:

```
#: ./client 00:0C:78:31:FC:8C
```

If we don't know the server address, we could find it using:

```
#: sdptool scan
```

5.3 Registering and searching a service

When a *Bluetooth* device offers a service (application) acts as a server, making it necessary to specify the type of service being offered and register it; thus the remote devices (clients) searching for this service could identify the service and request it to a correct device (server).

The Service Discovery Protocol (SDP) defines the way by which a client can discover the services availables on *Bluetooth* devices, as well as their attributes (of services). In the following examples *SDP* tools are used for register and search *SDP* services.

The first program registers a service named called bluefriend (the name does not matter), while the second program will search for this service.

The SDP daemon. Every *Bluetooth* device typically runs an *SDP* server that answers queries from other *Bluetooth* devices. In *BlueZ*, the implementation of the *SDP* server is called sdpd, and is usually started by the system boot scripts. Sdpd handles all incoming *SDP* search requests.

Registering a service. Registering a service with sdpd involves describing the service to advertise, connected to sdpd, instructing sdpd on what to advertise, and then disconnecting. For make it easier the service has the *UUID* 0xABCD.

registering.c

```
#include <bluetooth/bluetooth.h>
#include <bluetooth/sdp.h>
#include <bluetooth/sdp_lib.h>
sdp_session_t *register_service()
{
 uint32_t service_uuid_int[] = { 0, 0, 0, 0xABCD };
 const char *service_name = "Bluefriend";
 const char *service_dsc = "An Matching Aplication";
 const char *service_prov = "BF Server";

 uuid_t root_uuid,l2cap_uuid, svc_uuid;
 sdp_list_t *l2cap_list = 0, *root_list = 0,
 *proto_list = 0, *access_proto_list = 0;

 sdp_record_t *record = sdp_record_alloc();

 // set the general service ID
 sdp_uuid128_create(&svc_uuid, &service_uuid_int);
 sdp_set_service_id( record, svc_uuid );

 // make the service record publicly browsable
 sdp_uuid16_create(&root_uuid, PUBLIC_BROWSE_GROUP);
 root_list = sdp_list_append(0, &root_uuid);
 sdp_set_browse_groups( record, root_list );

 // set l2cap information
 sdp_uuid16_create(&l2cap_uuid, L2CAP_UUID);
 l2cap_list = sdp_list_append( 0, &l2cap_uuid );
 proto_list = sdp_list_append( 0, l2cap_list );

 // attach protocol information to service record
 access_proto_list = sdp_list_append(0, proto_list);
 sdp_set_access_protos(record, access_proto_list);

 // set the name, provider, and description
 sdp_set_info_attr(record, service_name, service_
 prov, service_dsc);

 int err = 0; sdp_session_t *session = 0;

 // connect to the local SDP server (BDADDR_LOCAL),
 //register the service record, and disconnect
 session = sdp_connect( BDADDR_ANY, BDADDR_LOCAL,
 SDP_RETRY_IF_BUSY );

 err = sdp_record_register(session, record, 0);

 // cleanup
 sdp_list_free( l2cap_list, 0 );
 sdp_list_free( root_list, 0 );
 sdp_list_free( access_proto_list, 0 );
 return session;
}
int main()
{
 sdp_session_t* session = register_service();
 sleep(5);
 sdp_close( session );
 return 0;
}
```

A way to check whether the service there has been satisfactorily registered is through the command `sdptool browse`, this command shows all services registered in the local *Bluetooth* device.

Searching a service. Once the service is registered, the next step is that a client device finds it, therefore requires a *SDP* connection to the remote device (server) to find the service with the *UUID* desired, we the remote sdp server will return a list of services founded with the specified *UUID*.

searching.c

```
#include <bluetooth/bluetooth.h>
#include <bluetooth/sdp.h>
#include <bluetooth/sdp_lib.h>
int main(int argc, char **argv)
{
 uint32_t svc_uuid_int[] = { 0, 0, 0, 0xABCD };
 uuid_t svc_uuid;
 int err, num;
 bdaddr_t target;
 sdp_list_t *response_list = NULL, *search_list,
 *attrid_list;
 sdp_session_t *session = 0;
 str2ba( "00:0C:78:31:FC:6B", &target );

 // connect to the SDP server on the remote machine
 session = sdp_connect( BDADDR_ANY, &target,
 SDP_RETRY_IF_BUSY );

 // specify the UUID we're searching for
 sdp_uuid128_create( &svc_uuid, &svc_uuid_int );
 search_list = sdp_list_append( NULL, &svc_uuid );

 // specify that we want a list of all the matching
 // applications' attributes
 uint32_t range = 0x0000ffff;
 attrid_list = sdp_list_append( NULL, &range );

 // get a list of service records (UUID 0xABCD)
 err = sdp_service_search_attr_req(session, search
 _list, SDP_ATTR_REQ_RANGE, attrid_list,
 &response_list);

 sdp_list_t *r = response_list;
 num = sdp_list_len(r);

 //if there isn't an error and there are
 //services found
 if((err==0) && (num> 0)){
 printf("\nService(s) found in %s.\n", target);

 // go through each of the service records
 for (; r; r = r->next ) {
 sdp_record_t *rec = (sdp_record_t*)
 r->data;
 printf("found service record 0x%x\n",
 rec->handle);
 sdp_record_free( rec );
 }
 sdp_close(session);
}
```

5.4 Using all the examples

Finally there is a program using the tools displayed in the previous examples, consists of 3 steps:

1. As a first step the device server records the service (registering.c)
2. Server application (server_blue.c) to listen the clients requests for the service and then exchange information.
3. Client application (client_blue.c) to search for nearby devices and search the service in each one, if the service is found, the client shares information with the server, otherwise continues with the search.

As the following examples are a compilation of previous (practically the same code), the code will not be shown again, only the little changes applied that are explained below.

server_blue.c is practically the same as server.c, the difference is that server_blue.c has a service registered, and besides, now we want to exchange a structure data instead strings, we only have to apply the following code lines in server.c.

```
typedef struct data
{
 char name[50];
 int id;
}my_data,rec;
```

we create rec and my_data to store the data received and my data.

```
struct data my_data,rec;
//setting my data
strcpy(my_data.name, "Client");
my_data.id=24;
```

we receive the client data that is stored in rec,

```
bytes_read=read(client, &rec, sizeof(my_data));
```

print the data received

```
printf("Hello my name is %s and my id is %d \n",
 rec.name, rec.id);
```

and send my_data to client:

```
write(client,&my_data,sizeof(my_data));
```

Exchanging structure is not so complicated, we only have to be careful when you create structures that match the types of data with which we are using.

client_blue.c applies the basic principle to identify nearby devices through a scan (scan.c), in each device search a service (searchin.g), and if it find it, a connection is created (client.c) to performs an action (exchange structures). It is necessary to place the code of the 3 examples in one [7]. The explain is below:

The first step is to declare the struct data my_data and rec used in server_blue.c, then we scan for nearby devices using the scan.c code

```
num_rsp = hci_inquiry(dev_id, len, max_rsp, NULL,
 &ii, flags);
```

where, for each remote device found:

```
for (i=0; i < num_rsp; i++)
```

a *SDP* connection is made using the *Bluetooth* address provided (searching.c)

```
session = sdp_connect(BDADDR_ANY, &(ii+i)->bdaddr,

 SDP_RETRY_IF_BUSY)
```

[7] Both programs can be run on a single computer (server.c and client.c or server_blue.c and client_blue.c), just identify the role that each device plays.

for search the service with the *UUID* 0xABCD.

```
err = sdp_service_search_attr_req( session, search_list,
 SDP_ATTR_REQ_RANGE, attrid_list, &response_list);
```

if the service is found,

```
if((err==0) && (num> 0))
```

a *L2CAP* socket connection is made using the same *Bluetooth* address (server.c),

```
s = socket(AF_BLUETOOTH, SOCK_SEQPACKET, BTPROTO_L2CAP);
status = connect(s, (struct sockaddr *)&address,
 sizeof(address));
```

and then exchange data structures:

```
if(write(s,&my_data,sizeof(my_data))==-1){
 perror("write");
}
bytes_read = read(s, &rec, sizeof(my_data));
```

The loop ends with this remote device and proceeded to continue with the next on the num_rsp list, until all go up and exit the program.

There are too much applications we can develop with these basic examples. It is necessary to mention that these programs are at a very basic level, for which there are many tools and opportunities offered by *C++* and *BlueZ* that aren't in this paper. The purpose of this paper is to involve new users in programming with *Bluez / C++* in a quick and easy way, forthcoming work is intended to bring this approach to one much more advanced .

6 Conclusions and future work

The proliferation of mobile communication devices at low cost and low power consumption has opened the possibility of developing applications that take advantage of new models of communication, such as the ad hoc networks. The *Bluetooth* technology is having a greater presence in the market for mobile devices. Unfortunately there is little documentation about it. This article presented the characteristics of *Bluetooth* and the *Bluetooth* protocol stack for Linux called *BlueZ*, as well as some of the most important utilities for their use. As future work is developing an application that uses *BlueZ* and programming languages C, *C++* and *Qt* that serve as reference for the development of future applications and future *BlueZ* programmers.

References

1. Jasmin Blanchette and Mark Summerfield. *C++ GUI Programming with Qt 3.* Prentice Hall in association with Trolltech Press, 2004.
2. Douglas E. Comer and David L. Stevens. *Internetworking with TCP/IP, Vol. III: Client-Server Programming and Applications, Linux/Posix Sockets Version.* Prentice Hall, November 2000.
3. P. D. Garner. Mobile bluetooth networking: Technical considerations and applications. *3G Mobile Communication Technologies*, page 274, June 2003.
4. The Free Software Fundation Inc. GNU General Public License. *http://www.gnu.org/copyleft/gpl.html.*
5. Albert S. Huang and Larry Rudolph. *Bluetooth Essentials for Programmers.* Cambridge University Press, 1st edition, September 2007.

6. Linux Kernel Patches. *http://www.holtmann.org/linux/kernel.*
7. BlueZ Official Web Site. *http://www.bluez.org.*
8. The IEEE 802.15.1 Standard. *http://www.ieee802.org/15/pub/ TG1.html.*
9. Bluetooth's Official Website. *http://www.bluetooth.com.*

Towards an Emergency Domain Name System Based on a Peer-To-Peer Network

Carolina Del-Valle-Soto, Iván Razo-Zapata, and Carlos Mex-Perera

Center for Electronics and Telecommunications
ITESM, Campus Monterrey
Av. Eugenio Garza Sada 2501 Sur, Col. Tecnológico
Monterrey, N.L., CP 64849 Mexico
{carolinadvs@yahoo.com.mx,carlosmex,razozapata}@itesm.mx

Abstract. In this work the performance of a Domain Name System (DNS) over P2P (Peer-to-Peer) networks is studied. We propose a double ring structure of P2P networks to provide a DNS. The architecture is called eDNS (emergency-DNS). eDNS can be used as an emergency mechanism in case of loss of availability of the original DNS. eDNS provides security and robustness for finding information under attacks. The performance of the proposed architecture was measured with the percentage of lost queries and the average number of hops in each query considering scenarios with a number of compromised nodes. Simulations show how the performance of the DNS is affected under attacks. According to the results, it is observed that the eDNS architecture overcomes the damage caused by attacks improving the availability of the DNS resources.

Key words: Domain Name System, Peer-To-Peer, DNS

1 Introduction

The main purpose of the DNS is to translate domain names into IP addresses. When a user wants to download a web site, he/she types in the Internet browser the corresponding domain, *i.e.* www.google.com, then the DNS will try to obtain the IP address. The DNS structure is based on a hierarchical inverse tree model (see Figure 1), where the origin is known as "root". Below that level, there are domains defined as TLDs (Top Level Domains). Such domains are classified as gTLDs (generic TLDs such as .com, .net, .org, .edu, .gov, .mil) and ccTLDs (country-code TLDs such as .uk, .se, .es, .mx). In order to have data origin authentication and authenticated denial of existence of domain names for DNS, Domain Name System Security Extensions (DNSSEC) can be used. DNSSEC was designed to protect from attacks like DNS cache poisoning [2].

Although DNS is a distributed scheme, it could be affected by some kind of attacks trying to damage critical TLDs, therefore, attenuating the performance.

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 115-126

Received 01/07/08
Accepted 01/08/08
Final version 08/09/08

In this way, it is desirable to apply another distributed schemes, such as Peer-to-Peer (P2P) architectures for supporting and protecting the DNS.

The hierarchical DNS is highly sensitive to latency and it has a significant challenge as to serve efficiently. In addition, the same hierarchical organization makes DNS have a disproportionate burden on the different levels of the hierarchy. The higher nodes are highly susceptible to a denial of service attacks, which causes a vulnerable security system as a whole. Another problem is that DNS servers are required for domestic outlets, incurring high administrative costs because DNS administration has to be manual [11]. That is why we want to establish a hybrid proposal that combines the concepts of a P2P network efficiently to improve the shortcomings that presents the DNS and thus, obtain a search more quickly and securely through a network bit vulnerable. So, our motivation is to make a robust architecture against attacks that combines the advantages of DNS with the advantages that presents a P2P network. This can control various DNS attacks such as denial of existence. Thus, an attack on the root of the DNS or any attack on the second level of the hierarchy, such as domain TLDs can be compensated through a double tier P2P-based structure. Then, this structure is called eDNS and it will be shown that can be a good mechanism for scenarios such above where a fast, secure and reliable DNS is desirable.

P2P architecture is a growing concept in the world of network technologies and the Internet. In a P2P system, distributed computing nodes of equal roles or capabilities exchange information and services directly with each other. The goal of a data-sharing system is to support a search protocol and retrieve data found on user caches. The information searches are a fundamental and problematic aspect of P2P environments because they aim for ensuring that there is not a single point of failure and offering scalability and dynamism [3].

P2P systems can be classified into 3 categories: *Centralized* systems , which have a central directory server to which users (clients) make requests. However, these systems have a single point of failure which makes them highly vulnerable [6]; as an example of these systems is Napster. The *decentralized* systems do not have a central server, but on the contrary, the nodes form a network among themselves and thus petitions are sent. Among decentralized networks designs, some are *structured*; in which there is a close link between P2P topology and the location of data; for example, Pastry [9], Tapestry [14], Skipnet, CAN [8] and Chord [10]. Other decentralized P2P systems are *unstructured*; where there is no link between the topology and location of information [1] that is, there is no precise control between the topology of the network and the place where the data in it, for instance Gnutella and Freenet.

This work is carried out under an architecture platform defined by Pastry, using DHT (Dynamic Hash Tables), where each node has a unique identifier assigned randomly, each object has also allocated a selected identifier and it is stored in the node whose identifier is numerical closest to the object ID, called *"home node"*. The routing is done docking with the prefix resource identifier node. Each request is traveling across the network with a certain number of hops up to find the best coupling of the digits of their ID with a specific node.

Thus, the search space is reduced exponentially. Then, the petition routes is under $O(log(b)N)$, where N is the number of nodes in the table and b is the numerical basis used in the system [11].

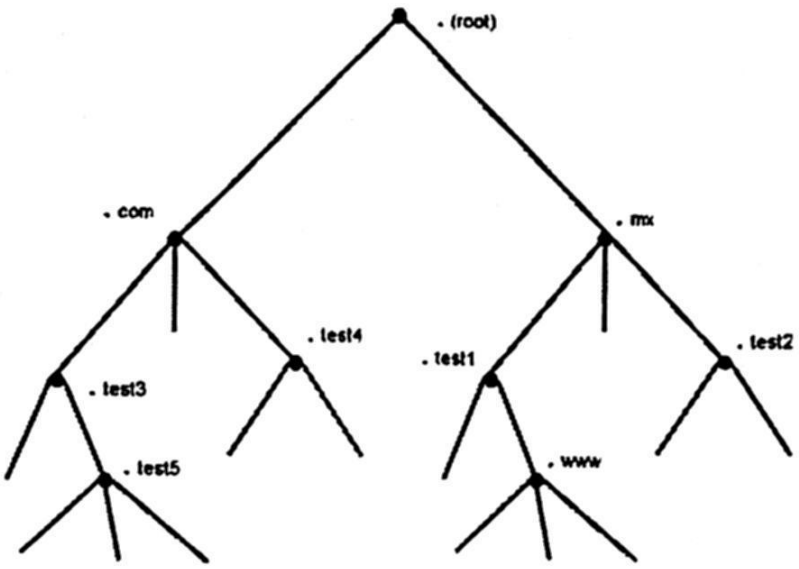


Fig. 1. Structure DNS Tree.

But the most problems with P2P data-sharing is that the networks adds overload, latency and low security. Besides it is susceptible to malicious activity, an attacker peer can insert itself into the network at the various points or a single point and forwards a lot of queries, or discard them altogether too. Also, a secure P2P network is challenging because the network topology is dynamic and nodes are on-line and off-line very often times. Thus we have problems like signatures to verify, legitimate packets and drop packets that do not pass the verification because P2P networks are very susceptible to malicious activity like previously mentioned. For that reason we propose a double tier structure for a better distribution of information combined with DNS sub-trees, this is a hybrid system that takes advantages of both strategies, centralized and decentralized schemes [12].

The rest of this paper is structured as follows. In the next sections, we describe related work and the concepts with peer-to-peer architecture over two tiers with DNS. After that, we present the proposal and then we evaluate the performance of the network through simulations and give results considering scenarios with a number of compromised nodes. Afterwards contributions are summarized. Finally, last section depicts the future work.

2 Concepts

Neither fixed clients nor fixed servers are present into a P2P network, but it contains nodes that behave both as clients and servers. P2P networks manage and

optimize the use of bandwidth of all users leading to a better workload balance. Thus, this result gives more efficiency in connections and transfers than with some centralized conventional methods where a small number of servers provides total bandwidth and shared resources for a service or application. Among the most common and potential applications of P2P networks are sharing and search files (most widespread application of such networks), distributed file systems, Internet telephony systems, alternatives to the conventional distribution of films and television programs and scientific calculations which process databases, among others [1].

Some of the most common topologies for networks are: star, mesh, backbone, ring, and combinations between them. In this work the P2P networks are based on ring topology, where the nodes are connected in a closed circuit without a central system in the network, like Figure 2. In addition, P2P networks can work under protocols such as: Tapestry, Pastry, CAN and Chord [8–10, 14], which define how the information is distributed and how the searches must be performed. For instance, in our proposal we have used Pastry.

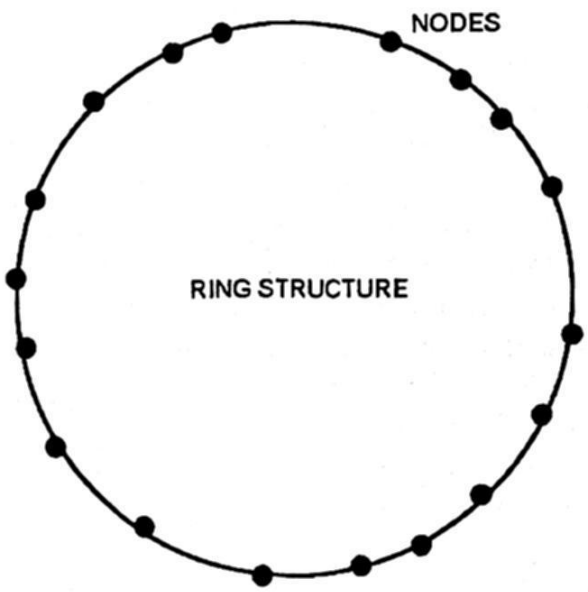


Fig. 2. Structure of a P2P ring.

In recent studies, it was shown that P2P networks were very much employed to share information or give hints as to where said information could be found. In the work presented by Pfeifer *et al.* [5], he shows a more specific architecture for P2P networks, either using DHT tables or hash functions is shown. Besides, information organization and safeguard is achieved in a much easier way through node creation. These may, without harm to the system, connect or disconnect themselves on a random basis. Data and network information caching is often used to reduce response time in P2P networks and this is where the problems of cache attack or cache poisoning become significant. In the work by Yang and Garcia-Molina [12], the respective benefits and drawbacks of centralized and

decentralized forms of P2P architectures are presented. As it can be seen in the article, the centralized form of P2P networks has some serious disadvantages, the most conspicuous appear when a node changes places or disconnects itself. The other nodes (except for the central one) are not notified and the queries will receive neither up to date information nor a reliable copy of it.

3 Related Work

Previous investigations were focused on resource record security in DNS and they helped to make the DNS tree safer and more reliable. The article by Ayumu Kubota *et al.* [4] refers the implementation of DNS in a P2P network without loss of hierarchy. Also, it referred to a topology with several tears in a hierarchical tree form that resembles to DNS structure with important features such as redundancy and number of hops to be made to reach a destination, and here is where it comes to a close similarity with our work because the replication in areas seeking to minimize the number of hops to get to satisfy a query. In addition, it refers to possible attacks that can be made on the network, such as a malicious node can know the well-defined destination and therefore it can know their neighbors and thus, all of the neighborhood.

The article by Ramasubramanian and Sirer [7] studies the security problem as an improvement in a specific network rather than as a mandatory requirement for the structuring of that network. Thus, here DNS plays an important role. Besides, this network is better structured thanks to DHT tables. In our work are also used DHT tables for quick search of information, plus giving a better organization of nodes in the ring.

The studies that have been made regarding P2P networks focused their investigation on analysis of a single ring, however, the work presented by Haiyun Luo *et al.* [13] mentions a dual structure that operates separately . So what we want to achieve with our work is maintaining a double-ring, where the two rings operated jointly and the workload of the petitions does not fall on one of them, but they share the job features such as: information capacity , nodes, time to answer, among others. In short, in our proposal we want to provide an efficient solution of two rings that operate in tandem distributing the burden on the network. In addition, we exploit the advantages that provides DNS regarding security and the possible future use of DNSSEC.

4 Proposal

We propose to use an hybrid system where elements of both, pure P2P and client/server systems coexist. Currently, hybrid file-sharing systems have better performance than pure systems because some tasks (like searching) can be done much more efficiently in a centralized manner. We propose a two-tier design with application layered on pastry and dynamic hash tables (DHTs), in our assessment, would comply with this goal. Besides, we distribute the information according to popularity of DNS resources and in this manner we balance

the workload among the nodes of both tiers. Also, with this design we provide decentralization and self-organization.

Thus, our proposal consists of the arrangement of two rings which will be the topology for eDNS. In such a way, these rings form two P2P networks in which actions such as data sharing and data update are performed. One tier has servers and the other tier has clients and both tiers are connected. For simplicity, we will refer them as servers and clients rings, respectively. Note that these names come from the original roles of such nodes in the classical DNS, however once they are working in a P2P architecture all of them must behave as peers. Consequently, the workload balance will be improved, once the nodes that form the ring will have replicates of the resources that are more commonly requested. Thus, allowing for more efficient and faster request solution.

The servers ring consists of nodes that are responsible for "small" DNS trees. This helps to improve the information search as it can take advantage of the DNS can employ, for example, DNSSEC for greater security.

Figure 3 shows how eDNS works. When a client tries to solve a domain name, it sends a query to the clients ring, if no answer is given then the query is send to the servers ring. Due to DNS sub-trees in the servers tier, this scheme is more reliable than an architecture with only clients ring.

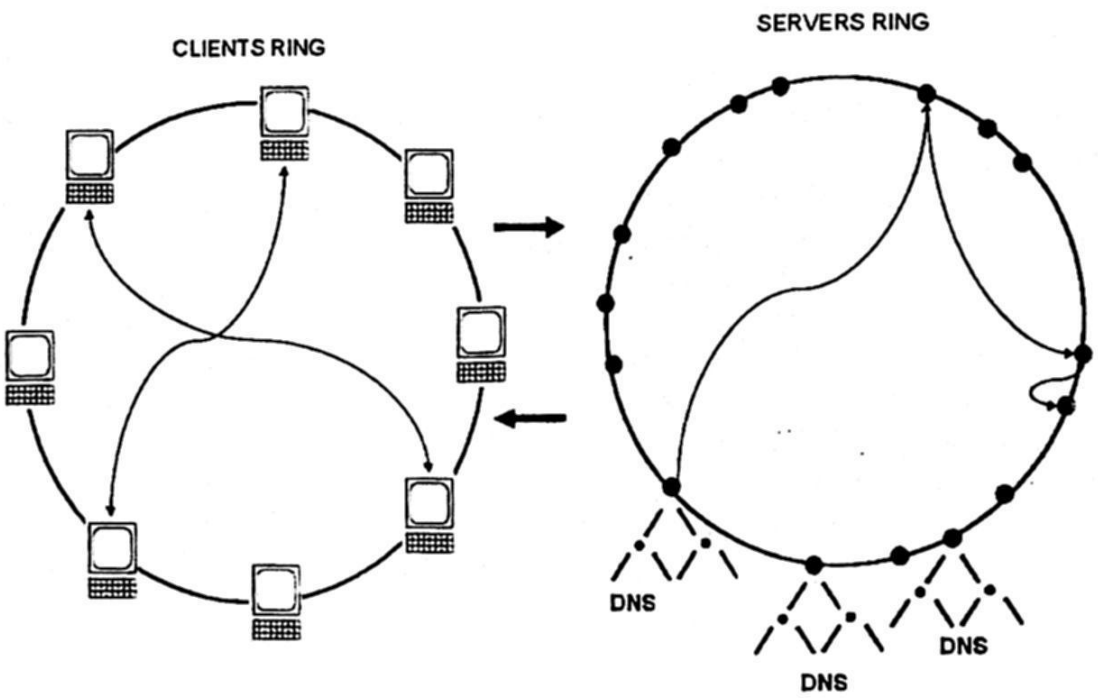


Fig. 3. Clients and servers organize to form double peer-to-peer network with DNS.

Figure 4 shows resources organization in the P2P tier, which is explained with an example, we have a specific domain: *example.com*, and its hash id is 0121, which will be located in a node having a hash id close to itself. So, resources are located according to their neighbors' hash id to form a logic P2P tier. Besides,

in Figure 3 we can see the relationship between servers and clients, and servers with DNS tree.

In the servers ring each node is an "DNS island", *i.e.* is a small DNS tree. These nodes are created as characteristic domains based on the ccTLDs (country domains, such as .mx, .co) and gTLDs (generic Internet domains for organizations such as: .com, .net). The advantages that this form of ring organization has is that it provides delegation of areas and the possibility of establishing "islands of trust", all these characteristics of DNSSEC. Then, these trees in each node allow greater ease of finding information, better organization of resources and the possibility of using characteristics and advantages of DNSSEC. As disadvantage may be that keys and information updates could be made only through each island of trust independently.

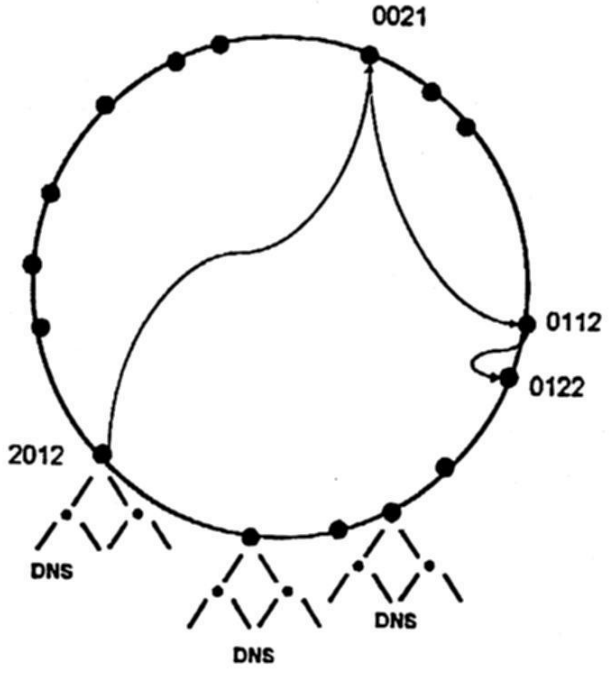


Fig. 4. Self-organization with DHT tables.

Note that servers nodes are normally more robust, while clients belong to computer users, who frequently connect and disconnected all the time.

5 Simulations and Results

5.1 Preliminary Considerations

Due to the importance of illustrating the performance of the eDNS architecture, it was necessary to feed the eDNS with realistic petitions. Therefore, DNS petitions were captured from one DNS server which belongs to one university network. These petitions were captured in the DNS server, and subsequently the

set of petitions was computed for getting knowledge about the required domain names and their frequency.

In this way, two files were yield. The first one with the set of petitions sorted by their arriving time (*Time-DNS*), and the second with 13818 domain names sorted by their frequency (*Frequency-DNS*). Consequently, the eDNS was built using the information contained into the second file and the petitions into the eDNS were performed following the same distribution as in the first file.

5.2 Building the eDNS architecture

The DNS information is distributed into the clients tier using DNS lists. Therefore, each node has a list with domain names. Since both, the store and search of domain names into the ring are performed using the Pastry protocol, each node has domain names whose hash id is similar to the node's hash id.

On the contrary, in the server tier, DNS information is stored using the concept of gTLDs and ccTLDs. For domain names with gTLD and ccTLD, such as *www.google.com.mx*, one node will be built using the combination of gTLD, ccTLD and the domain name, in this way the node will have the *google.com.mx* domain. On the other hand, if the domain name has only either gTLD or ccTLD, such as *www.mty.itesm.mx* or *www.network.ieee.org*, the domain name will have only the combination of the domain name with the corresponding gTLD or ccTLD, in this case the nodes will have *itesm.mx* and *ieee.org* respectively. Therefore, the resources that are common domain are stored in a single node, forming a DNS tree. This allows a more quick efficient search, and the possibility of exploiting DNSSEC advantages.

The petitions were conducted with the ring of clients and servers operating jointly. The basic operation of this scheme is as follows: Whenever a query arrives to the eDNS, the query is passed to the clients ring which responds either with a positive answer or with a negative answer. If the clients ring responds positively so the resource was reached, on the other hand if the answer was negative, the clients ring passes the query to the servers ring and the servers ring will respond to the petition.

Furthermore, in order to test the robustness of the eDNS, some malicious nodes were introduced into both rings. The coalitioned nodes are malicious nodes that drop petitions, consequently they make that information can not pass through them, *i.e.* they are a barrier in the searching process and also play like failures into the eDNS architecture.

5.3 Results

Figure 5 shows the percentage of lost queries according to the percentage of coalitioned nodes in each tier, for instance: 0, 0.03, 0.3, 3, 30 and 45 percent of coalitioned nodes. The blue-triangled line shows the performance of the clients tier and the red-circled line shows the performance of the servers tier. As observed, the clients ring lost more messages than servers ring. It is because server-ring nodes clusters more domain names in their sub-trees. For example, if the

node with the domain name *mty.itesm.mx* in the clients ring was lost, also the petition asking for *mty.itesm.mx* will be lost, and the message will be redirected to the node with the *itesm.mx* domain in the servers ring.

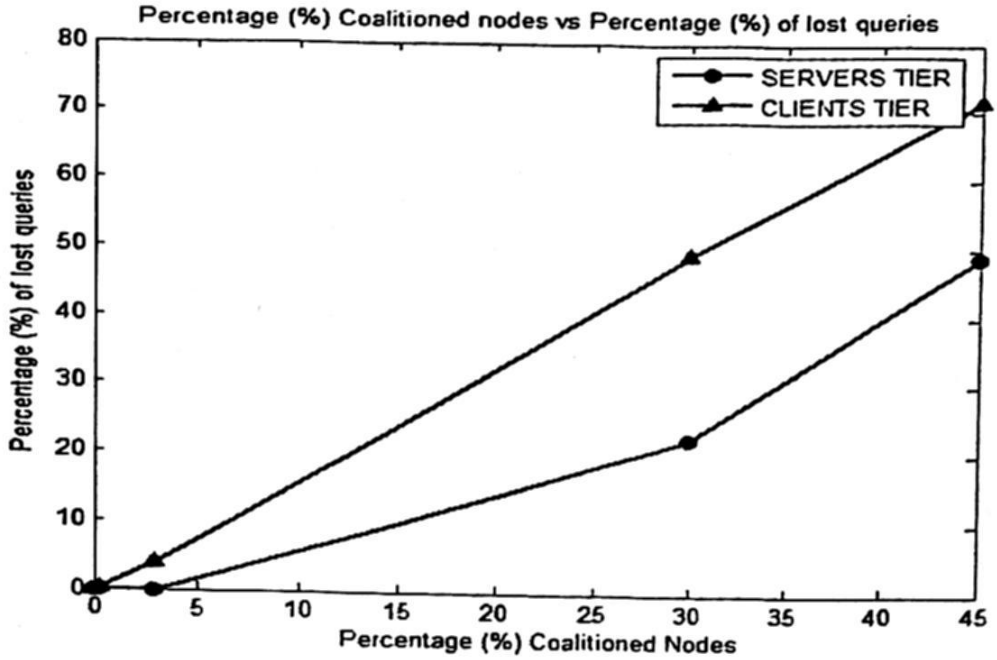


Fig. 5. Percentage of lost queries

On the other hand, Figure 6 shows the average number of hops in the clients tier (triangled line) and the servers tier (circled line) for an equal number of coalitioned nodes, including: 0, 0.03, 0.3, 3, 30 and 45 percent of coalitioned nodes. Also this figure depicts a tradeoff between clients ring activity and servers ring activity, therefore the less activity in clients ring the more activity in the servers ring. In this way, whenever a client-ring node gets a negative answer for a petition, it will ask for the domain name to the servers ring, so increasing the workload in the servers ring.

Figure 7 shows tables 1, 2, 3 and 4. Tables 1 and 2 show the percentage of lost queries according to the percentage of coalitioned nodes in each tier, as well: 0, 0.03, 0.3, 3, 30 and 45 percent of coalitioned nodes. On the other hand, Tables 3 and 4 show the average number of hops in the clients tier and the servers tier for an equal number of coalitioned nodes. Table 1 presents the result of a test with the more frequent resources, *i.e.* more popular requests, and where the nodes in the clients tier share their caches and no shared caches and it shows the percentage of lost queries. In table 2 results obtained from a test with less frequent resources are given, *i.e.* less popular requests, and where the nodes in the clients tier share their caches and no shared caches and it shows the percentage of lost queries. Table 3 shows a test with more frequent resources and where the nodes in the clients tier share their caches and no shared caches

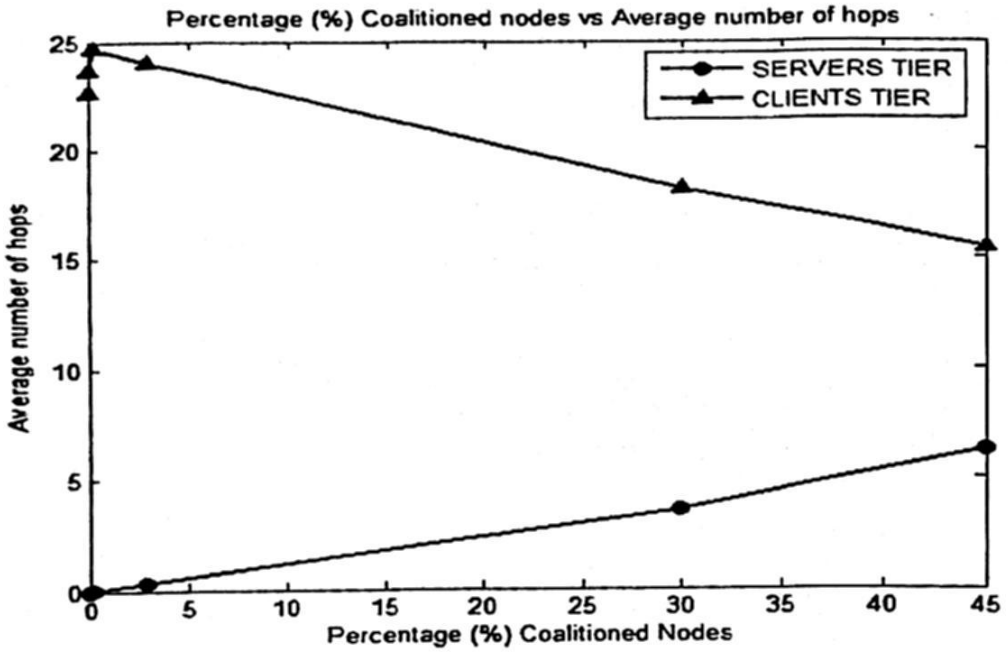


Fig. 6. Percentage of lost queries

and it reports the average of number of hops. Table 4 shows a test with less frequent resources and where the nodes in the clients tier share their caches and not share caches and it presents the average of number of hops.

6 Conclusion

As findings of this study we can say that regarding the results, the client tier always has the largest number of hops of lost queries. This is because the server tier nodes has only two domain levels, maximum three, which makes the search for more general information. Analyzing comparatively tests with more popular and less popular domains with shared caches and not shared caches (Tables 1, 2, 3, 4), we can see that when nodes do not share their caches the delay in meeting a request becomes larger, this shows that the number of hops increases for the graphic when the cache was not shared between the nodes.

7 Future Work

As future work we intend to make an architecture that responds to changes in the cache nodes and also to be robust in terms of requests for emergency. We will try the issue of denial of existence as a form of direct attack to DNS. Through a complete interplay of the two P2P rings will be offset the burden on the network as best as possible. Also we want to do experiments with the cache nodes, making the shared cache to work by geographic regions and the possibility of generating preferences on users. It is also expected to take advantage of the

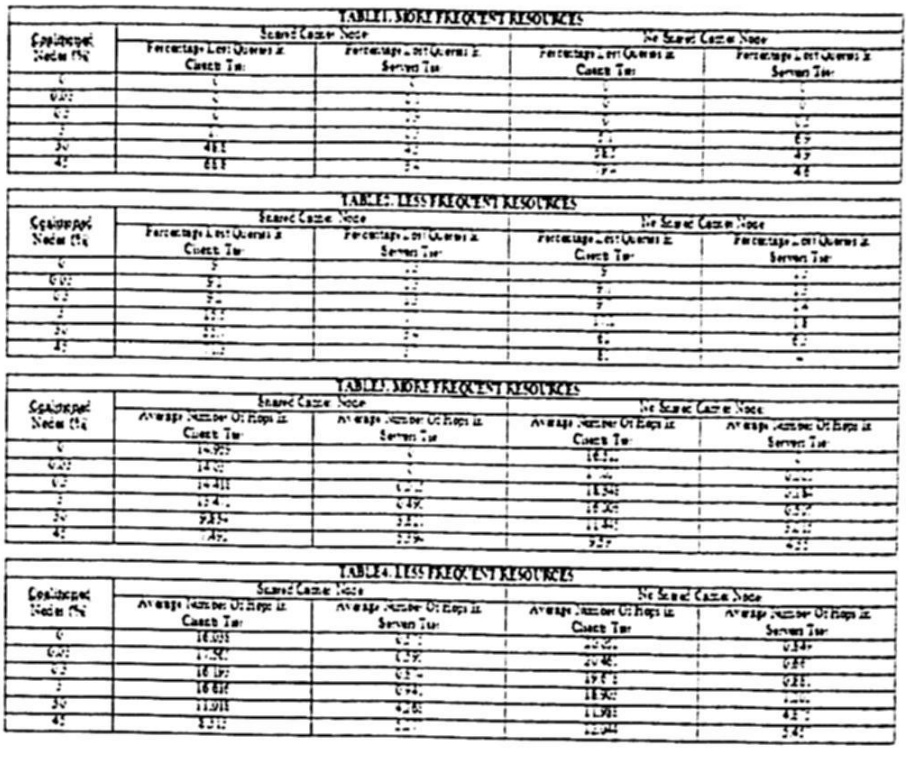


Fig. 7. Percentage of lost queries

characteristics of DNS along with DNSSEC for changing keys on a chain of trust, providing greater robustness and reliability of the information. Further simulations will work with a more real cache, where nodes have characteristics in common and can be related under specific parameters. All this will lead to a more wisely exploit of the advantages that P2P network offers in conjunction with the advantages of having a DNS chain of trust, where there cryptographic keys and information are updated.

Acknowledgment

Authors would like to thank to Catedra de Biometricas y Protocolos Seguros para Internet and Google.

References

1. Scott Shenker Edith Cohen. Replication strategies in unstructured peertopeer networks. *SIGCOMM'02*, agosto, 2002.
2. Yong Wan Ju, Kwan Ho Song, Eung Jae Lee, and Yong Tae Shin. Cache poisoning detection method for improving security of recursive dns. *ICACT*, 2007.
3. Junjiro KONISHI, Naoki WAKAMIYA, and Masayuki MURATA. Design and evaluation of a cooperative mechanism for pure p2p file-sharing networks. *IEICE TRANS. COMMUN.*, pages 2319–2326, September 2006.

4. Ayumu KUBOTA, Yutaka MIYAKE, and Toshiaki TANAKA. Secure host name resolution infrastructure for overlay networks. *PAPER Special Section on Networking Technologies for Overlay Networks, IEICE TRANS. COMMUN.*, 2006.
5. G. Pfeifer, C. Fetzer, and T. Hohnstein. Exploiting host name locality for reduced stretch p2p routing. *Sixth IEEE International Symposium on Network Computing and Applications (NCA 2007)*, 0, 2007.
6. Edith Cohen Kai Li Scott Shenker Qin Lv, Pei Cao. Search and replication in unstructured peer-to-peer networks. *Department of Computer Science, Princeton University.*
7. V. Ramasubramanian and E. Gun Sirer. Beehive: O(1) lookup performance for power-law query distributions in peer-to-peer overlays. *Dept. of Computer Science, Cornell University.*
8. Sylvia Ratnasamy, Paul Francis, Mark Handley, Richard Karp, and Scott Shenker. A scalable content-addressable network. In *ACM SIGCOMM'01.*, 2001.
9. Antony Rowstron and Peter Druschel. Pastry: Scalable, decentralized object object location and routing for large-scale peer-to-peer systems. In *IFIP/ACM.*, 2001.
10. Ion Stoica, Robert Morris, David Karger, M. Frans Kaashoek, and Hari Balakrishnan. Chord: A scalable peer-to-peer lookup service for internet applications. In *ACM SIGCOMM'01.*, 2001.
11. Emin Gun Sirer Venugopalan Ramasubramanian. Beehive: O(1) lookup performance for power-law query distributions in peer-to-peer overlays. *Dept. of Computer Science, Cornell University.*
12. Beverly Yang and Hector Garcia-Molina. Comparing hybrid peer-to-peer systems. *Proceedings of the 27th VLDB Conference*, 2001.
13. Hao Yang, Haiyun Luo, Yi Yang, Songwu Lu, and Lixia Zhang. Hours: Achieving dos resilience in an open service hierarchy. *Computer Science Department, University of California, Los Angeles.*
14. Ben Y. Zhao, John Kubiatowicz, and Anthony D. Joseph. Tapestry: An infrastructure for fault-tolerant wide-area location and routing. Technical Report UCB/CSD-01-1141, U.C. Berkeley, April 2001.

Artificial Intelligence

Logic and Multi-Agent Systems

Optimizing Type-1 and Type-2 Fuzzy Logic Systems with Genetic Algorithms

Nohe Ramon Cazarez-Castro[1], Luis T. Aguilar[2], Oscar Castillo[3], and Antonio Rodriguez[1]

[1] Universidad Autónoma de Baja California, Facultad de Ciencias Quimicas e Ingenieria, Tijuana BC 22390, Mexico,
nohe@ieee.org,
WWW home page: http://www.uabc.mx
[2] CITEDI-IPN, Tijuana BC 22510, Mexico
[3] Instituto Tecnologico de Tijuana, Tijuana BC 22414, Mexico

Abstract. Genetic Algorithms (GAs) are proposed as optimization method for tuning Membership's Functions (MFs) parameters of Type-1 and Type-2 Fuzzy Logic Systems (FLSs). The problem is to find the optimal MFs parameters to achieve a desired behavior in a closed-loop system. The case of study of the output regulation of a servomechanism with backlash is presented. Simulations results illustrate the effectiveness of the optimized closed-loop systems.

Key words: Fuzzy Control, Fuzzy Logic Systems, Genetic Algorithm, Type-2 Fuzzy Logic, Optimization

1 Introduction

The design of FLSs is a heavy task that we face every time that we try to use Fuzzy Logic (FL) as a solution to some problem, the design of FLSs implies at least two stages: design of rules and design of MFs.

There has been a lot work published in the design of Type-1 FLS using GA, [1] presents GAs as optimization method for control parameters, both to Type-1 Fuzzy Logic Control (FLC) as other control strategies, in [2] GAs are used to optimize all the parameters of a Type-1 FLC, in [3] a hybridizing of Neural Networks and GAs are presented to optimize a Type-1 FLC, a Hierarchical GA (HGA) is proposed in [4] to optimize rules and MFs parameters of a Type-1 FLC, in general an extended list of references can be obtained from [3] and [4].

Type-2 FLSs [5] are a generalization of Type-1 FLS [6–8], which allow us to deal with the uncertainty induced into a mechanical system by noise, frictions, backlash, etc. In the few last years a growing interest in the research of theories and applications of type-2 FLS can be seen from the academic and industry sectors. In [9] is presented an extended review of Type-2 Fuzzy Logic Systems (FLS) in control applications, and in [10] is presented a comparison between Type-1 and Type-2 FLCs, giving conditions for the use of each one.

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 131-153

Received 23/06/08
Accepted 01/08/08
Final version 11/09/08

As Type-2 FLSs are relatively new, and this type of FLS has more parameters to be optimized, some options have emerged to optimize some of its parameters. In [11] was proposed a Particle Swarm Optimization (PSO) method to optimize parameters of the primary MFs of Type-2 FLS. The Human Evolutionary Model (HEM) is proposed in [12] to the optimization of interval Type-2 MFs, HGAs are proposed in [13] to optimize gaussian MFs.

Type-2 FLS allow us to deal with uncertainty, but this uncertainty must to be modeled in form of Type-2 MFs, which can carry a new problem in the designing of FLCs. In [14] it is show that making a uniform modification to the MF's parameters to a certain limit, the closed-loop system keep some properties like stability, but will lost or gain in some others like performance.

GAs [15], are derivative free optimization methods that have been used in a wide range of issued, particulary in [16], GAs are presented as a class of optimization methods for FLSs.

In the present paper, the output regulation problem is studied for an electrical actuator consisting of a motor part driven by DC motor and a reducer part (load) operating under uncertainty conditions in the presence of nonlinear backlash effects, is presented as case of study. The objective is to drive the load to a desired position while providing the boundedness of the system motion and attenuating external disturbances. Due to practical requirements [17], the motor angular position is assumed to be the only information available for feedback.

The paper is organized as follows. Type-2 Fuzzy Sets and Systems are presented in Section II. A Hybrid Genetic - Fuzzy optimization approach of a Type-2 FLC is presented in Section III. Simulations results are presented in Section IV, and in Section V the conclusions are presented.

2 Fuzzy Sets and Systems

2.1 Type-1 Fuzzy Sets and Systems

A Type-1 Fuzzy Set (FS), denoted A is characterized by a type-1 MF $\mu_A(x)$ [15], where $x \in X$, i.e,

$$A = \{(x, \mu(x)) | \forall x \in X\} \tag{1}$$

where $\mu(x)$ is called *membership function* of the fuzzy set A. The MF maps each element of X to a membership grade (or membership value) between 0 and 1.

Type-1 FLSs are both intuitive and numerical systems that maps crisp inputs into a crisp output. Every Type-1 FLS is associated with a set of rules with meaningful linguistic interpretations, such as:

$$R^l : \text{IF } x_1 \text{ is } A_1^l \text{ AND } x_2 \text{ is } A_2^l \text{ THEN } w \text{ is } B^l, \tag{2}$$

which can be obtained either from numerical data, or from experts familiar with the problem at hand. Based on this kind of statements, actions are combined with rules in an antecedent/concecuent format, and then aggregated according

to approximate reasoning theory, to produce a nonlinear mapping from input space $U = U_1 \times U_2 \times \ldots U_n$ to output space W, where $A_k^l \subset U_k$, $k = 1, 2, \ldots n$, and the output linguistic variable is denoted by w.

A Type-1 FLS consist of four basic elements (see Fig. 1): the ***Type-1 fuzzyfier***, the ***fuzzy rule-base***, the ***inference engine***, and the ***Type-1 defuzzifier***. The ***fuzzy rule-base*** is a collection of rules in the form of (2), which are combined in the ***inference engine***, to produce a fuzzy output. The ***Type-1 fuzzifier*** maps the crisp input into Type-1 (FS), which are subsequently used as inputs to the ***inference engine***, whereas the ***Type-1 defuzzifier*** maps the Type-1 FSs produced by the ***inference engine*** into crisp numbers.

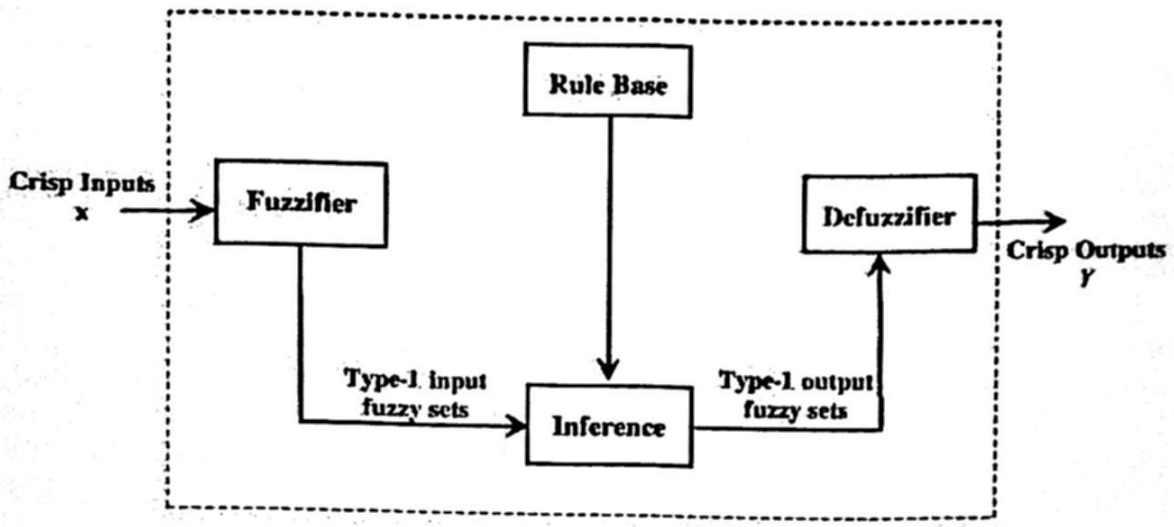


Fig. 1. Type-1 Fuzzy Logic System.

FSs can be interpreted as MFs U_x that associate with each element of x of the universe of discourse, U, a number $\mu_X(x)$ in the interval [0,1]:

$$\mu_x : U \rightarrow [0, 1]. \tag{3}$$

2.2 Type-2 Fuzzy Sets and Systems

As the Type-1 FS, the concept of Type-2 FS was introduced by Zadeh [6–8] as an extension of the concept of an ordinary FS (Type-1 FS).

A Type-2 FS, denoted $\tilde{A}$ is characterized by a Type-2 MF $\mu_{\tilde{A}}(x, u)$ [18], where $x \in X$ and $u \in J_x \subseteq [0, 1]$, i.e,

$$\tilde{A} = \{((x, u), \mu_{\tilde{A}}(x, u)) | \forall x \in X, \forall u \in J_x \subseteq [0, 1]\} \tag{4}$$

in which $0 \leq \mu_{\tilde{A}}(x, u) \leq 1$. $\tilde{A}$ can alse be expressed as follows [18]:

$$\tilde{A} = \int_{x \in X} \int_{u \in J} \mu_{\tilde{A}}(x,u)/(x,u) \tag{5}$$

where $J_x \subseteq [0,1]$ and $\int \int$ denotes union over all admissible x and u [18].

J_x is called primary membership of x, where $J_x \subseteq [0,1]$ for $\forall x \in X$ [18]. The uncertainty in the primary memberships of a Type-2 FS $\tilde{A}$, consists of a bounded region that is called the *footprint of uncertainty* (FOU) [18]. It is the union of all primary memberships [18].

A FLS described using at least one Type-2 FS is called a Type-2 FLS. Type-1 FLS are unable to directly handle rule uncertainties, because they use Type-1 FSs that are certain. On the other hand, Type-2 FLSs, are very useful in circumstances where it is difficult to determine an exact, and measurement uncertainties [5].

It is known that Type-2 FS let us to model and to minimize the effects of uncertainties in rule-based FLS. Unfortunately, Type-2 FSs are more difficult to use and understand that Type-1 FSs; hence, their use is not widespread yet.

Similar to a Type-1 FLS, a Type-2 FLS includes ***Type-2 fuzzifier***, ***rule-base***, ***inference engine*** and substitutes the ***defuzzifier*** by the ***output processor***. The ***output processor*** includes a ***type-reducer*** [5] and a ***Type-2 defuzzifier***; it generates a Type-1 FS output (from the *type-reducer*) or a crisp number (from the ***Type-2 defuzzifier***). A Type-2 FLS is again characterized by IF-THEN rules, but its antecedent and consequents sets are now of the Type-2, see (6). Type-2 FLSs can be used when the circumstances are too uncertain to determine exact membership grades. A model of a Type-2 FLS is shown in Fig. 2.

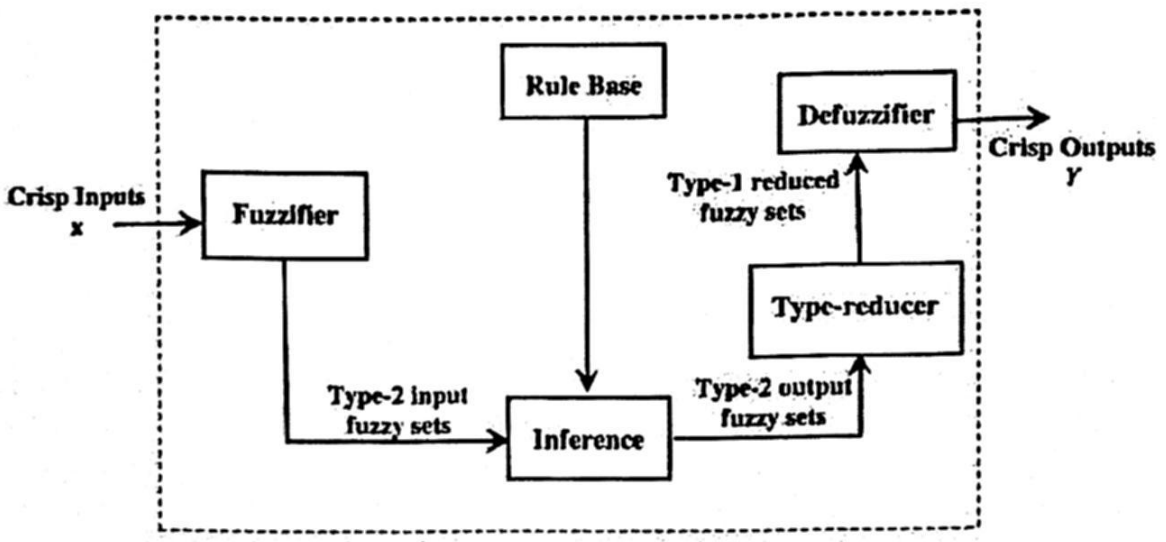


Fig. 2. Type-2 Fuzzy Logic System.

$$R^l : \text{IF } x_1 \text{ is } \tilde{A}_1^l \text{ AND } x_2 \text{ is } \tilde{A}_2^l \text{ THEN } w \text{ is } \tilde{B}^l, \quad (6)$$

Note that for both Type-1 FLS and Type-2 FLS ***rule-base*** we are describing the IF-THEN rules following the Mamdani's [19] type.

3 Genetic Algorithms

GAs are derivative-free optimizations methods based on the concepts of natural selection and evolutionary process [15]. They were first proposed and investigated in [20]. As a general-purpose optimization tool, GAs are moving out of academic sectors and finding significant applications in many areas. Their popularity can be attributed to their freedom from dependence on functional derivatives and their incorporation of other characteristics reported in [15].

The main idea of a GA is to maintain a ***population*** of solutions of a problem that evolves over a time through a process of competition and controlled variation. Each individual in the population represents a candidate solution to the specific problem, and each individual has associated a *fitness* to determine which individuals are used to form (by sexual reproduction and mutation) new ones in the process of competition.

The sexual reproduction of GAs consists basically in a **Selection Process** [20], where a set of individuals are selected to be passed through a **Crossover Operation** [20], which consist in to take a pair of individuals and interchanging its gens from one (or more) random selected cross point to the end of the chromosome. **Mutation** [20] consists in to change one or more randomly selected gens of the chromosome in some of the selected individuals.

The ***objective function*** [20] of a GA is a the value (*fitness*) that the method must maximize or minimize.

4 Case of Study: Backlash Problem

In the present paper, the output regulation problem is studied for an electrical actuator consisting of a motor part driven by DC motor and a reducer part (load) operating under uncertainty conditions in the presence of nonlinear backlash effects. The objective is to drive the load to a desired position while providing the boundedness of the system motion and attenuating external disturbances. Due to practical requirements [17], the motor angular position is assumed to be the only information available for feedback.

The study is motivated by a PEGASUS robot manipulator (see Fig. 3) installed in the Robotics & Control Laboratory of CITEDI-IPN where the backlash problem occurs due to the chains and gears transmission elements. Measurements are provided from the motor side while the links, attached into the load of the motor, must be positioned at the desired point.

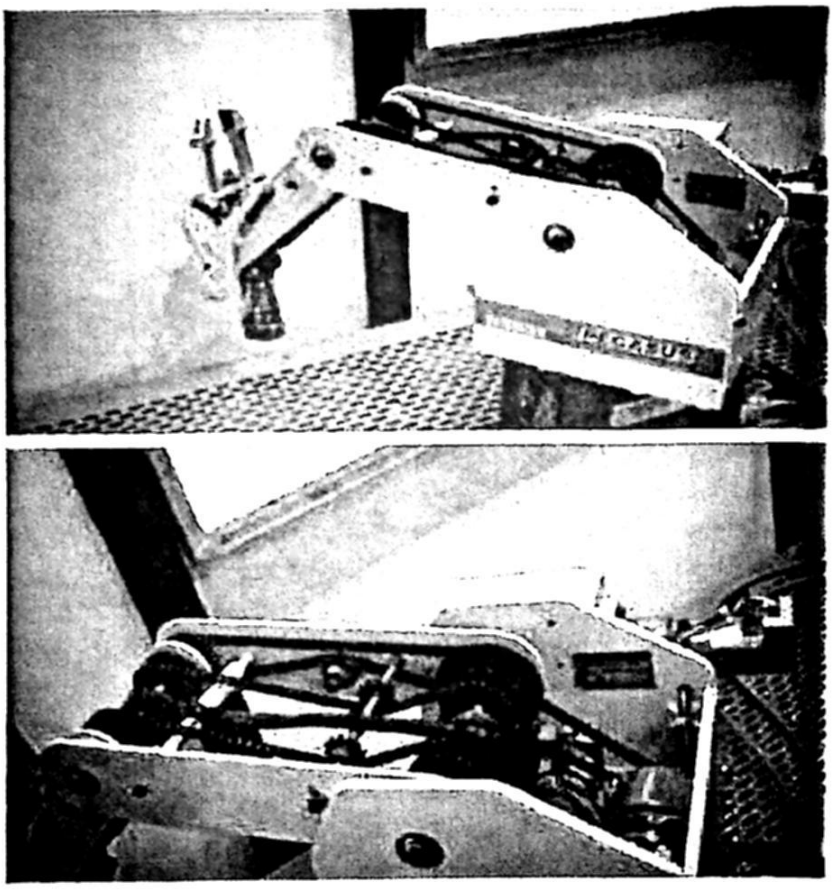


Fig. 3. PEGASUS robot manipulator of Robotics & Control Laboratory of CITEDI-IPN, and view of the problem in question given by each degree of freedom.

4.1 Dynamic Model

The dynamic model of the angular position $q_i(t)$ of the DC motor and $q_o(t)$ the angular position of the load are given according to

$$\begin{aligned} J_0 N^{-1} \ddot{q}_0 + f_0 N^{-1} \dot{q}_0 &= T + w_0 \\ J_i \ddot{q}_i + f_i \dot{q}_i + T &= \tau_m + w_i, \end{aligned} \tag{7}$$

hereafter, J_0, f_0, $\ddot{q}_0$ and $\dot{q}_0$ are, respectively, the inertia of the load and the reducer, the viscous output friction, the output acceleration, and the output velocity. The inertia of the motor, the viscous motor friction, the motor acceleration, and the motor velocity denoted by J_i, f_i, $\ddot{q}_i$ and $\dot{q}_i$, respectively. The input torque τ_m serves as a control action, and T stands for the transmitted torque. The external disturbances $w_i(t)$, $w_0(t)$ have been introduced into the driver equation (7) to account for destabilizing model discrepancies due to hard-to-model nonlinear phenomena, such as friction and backlash.

The transmitted torque T through a backlash with an amplitude j is typically modeled by a dead-zone characteristic [21]:

$$T(\Delta q) = \begin{cases} 0 & |\Delta q| \leq j \\ K\Delta q - Kj\text{sgn}(\Delta q) & \text{otherwise} \end{cases} \tag{8}$$

with

$$\Delta q = q_i - Nq_0, \tag{9}$$

where K is the stiffness, and N is the reducer ratio. Such a model is depicted in Fig. 4. Provided the servomotor position $q_i(t)$ is the only available measurement on the system, the above model (7)-(9) appears to be non-minimum-phase because along with the origin the unforced system possesses a multivalued set of equilibria (q_i, q_0) with $q_i = 0$ and $q_0 \in [-j, j]$.

To avoid dealing with a non-minimum-phase system, we replace the backlash model (8) with its monotonic approximation:

$$T = K\Delta q - K\eta(\Delta q) \tag{10}$$

where

$$\eta = -2j\frac{1 - \exp\left\{-\frac{\Delta q}{j}\right\}}{1 + \exp\left\{-\frac{\Delta q}{j}\right\}}. \tag{11}$$

Coupled to the drive system (7) subject to motor position measurements, it is subsequently shown to continue a minimum phase approximation of the underlying servomotor, operating under uncertainties $w_i(t)$, $w_0(t)$ to be attenuated. As a matter of fact, these uncertainties involve discrepancies between the physical backlash model (8) and its approximation (10) and (11).

4.2 Problem statement

To formally state the problem, let us introduce the state deviation vector $x = [x_1, x_2, x_3, x_4,]^T$ with

$$x_1 = q_0 - q_d$$

$$x_2 = \dot{q}_0$$

$$x_3 = q_i - Nq_d$$

$$x_4 = \dot{q}_i$$

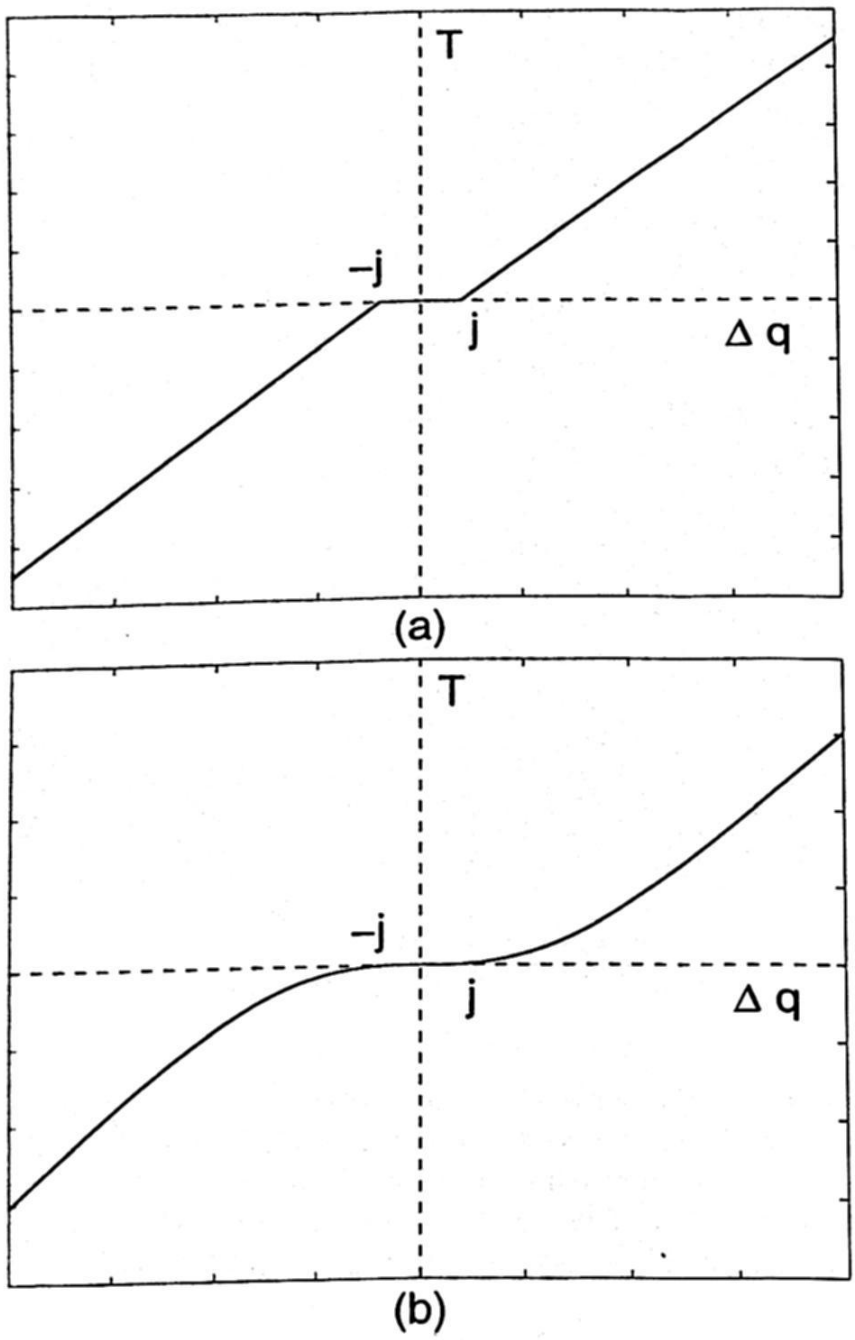


Fig. 4. a) The dead-zone model of backlash; b) The monotonic approximation of the dead-zone model.

where x_1 is the load position error, x_2 is the load velocity, x_3 is the motor position deviation from its nominal value, and x_4 is the motor velocity. The nominal motor position Nq_d has been pre-specified in such a way to guarantee that $\Delta q = \Delta x$, where

$$\Delta x = x_3 - Nx_1. \tag{12}$$

Then, system (7)-(11), represented in terms of the deviation vector x, takes the form

$$\begin{aligned}
\dot{x}_1 &= x_2 \\
\dot{x}_2 &= J_0^{-1}[KNx_3 - KN^2x_1 - f_0x_2 + KN\eta(\Delta q) + w_o] \\
\dot{x}_3 &= x_4 \\
\dot{x}_4 &= J_i^{-1}[\tau_m + KNx_1 - Kx_3 - f_ix_4 + K\eta(\Delta q) + w_i].
\end{aligned} \tag{13}$$

The zero dynamics

$$\begin{aligned}
\dot{x}_1 &= x_2 \\
\dot{x}_2 &= J_0^{-1}[-KN^2x_1 - f_0x_2 + KN\eta(-Nx_1)]
\end{aligned} \tag{14}$$

of the undisturbed version of system (6) with respect to the output

$$y = x_3 \tag{15}$$

is formally obtained by specifying the control law that maintains the output identically zero.

The objective of the Type-2 FLC output regulation of the nonlinear driver system (7) with backlash (10) and (11), is thus to design a FLC so as to obtain the closed-loop system in which all these trajectories are bounded and the output $q_0(t)$ asymptotically decays to a desired position q_d as $t \to \infty$ while also attenuating the influence of the external disturbances $w_i(t)$ and $w_0(t)$.

5 Fuzzy - Genetic Architectures

In this paper we use a GA to optimize the parameters of the MFs of a Type-1 FLS and a Type-2 FLS, this optimization is performed assuming that each FLS have a preestablished fuzzy *rule-base*.

By the knowledge that we have about the systems of our case of study, we propose the seven rules of Table 1, where can be seen that we select two input variables (`error` and `change of error`) and one output variable (`control`), each one of this input and output variables are granulated in three linguistic interpretations (MFs), this linguistic interpretations are: *negative* (n), *zero* (z) and *positive* (p).

The next two subsections describe the architectures for each one of the optimization approaches.

5.1 Fuzzy - Genetic Architecture for Type-1 FLS

To make the Type-1 Fuzzy - Genetic optimization we are considering triangular MFs to each one of the linguistic interpretation in which we granulate the three variables of the FLS, we encode each individual in a 27 real gens chromosome [22], where are represented the three parameters of each one of the three MFs

Table 1. Fuzzy IF-THEN rules

No.	error	change of error	control
1	n	n	p
2	n	p	z
3	n	z	p
4	p	p	n
5	p	n	z
6	p	z	n
7	z	z	z

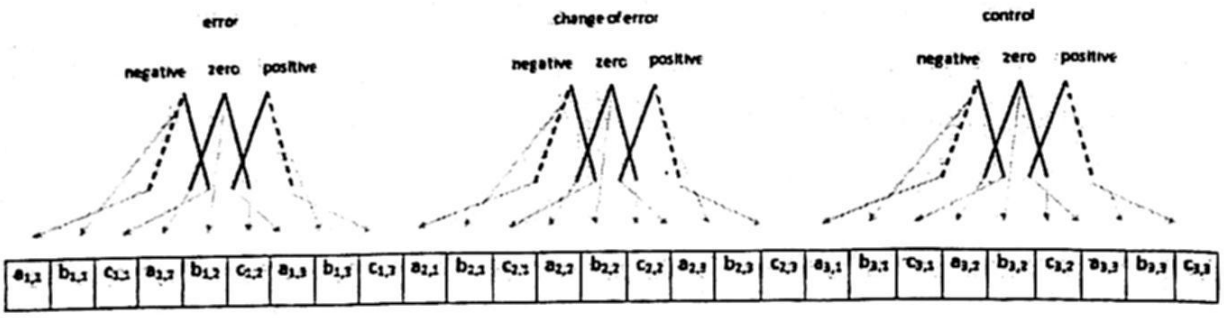


Fig. 5. Genotype for Type-1 FLS.

of each one of the three variables of the Type-1 FLS, this is called a genotype [22] of the population, see Fig. 5 for an schematic representation.

In this case we select the ***objective function*** of equation (16). To achieve our optimization problem, we must to minimize this ***objective function***.

$$\text{fitness}_i = min(mean|error|) \tag{16}$$

The set of parameters of the GA are shown in Table 2.

5.2 Fuzzy - Genetic Architecture for Type-2 FLS

To make the Type-2 Fuzzy - Genetic optimization we are considering Triangular Interval MFs to each one of the linguistic interpretation in which we granulate the three variables of the FLS; following the proposed in [23]-[24] we need six parameters for each Triangular Interval Type-2 MF, that is, we need to encode a total of 54 parameters for each individual (Type-2 FLS) of our population, to make this encoding we design a chromosome structure of 54 consecutive real gens, where are represented the six parameters of each one of the three Triangular Interval Type-2 MFs of each one of the three variables of each Type-2 FLS (individual) of our population. Fig. 6 show an schematic of the genotype [15] of our Type-2 Fuzzy - Genetic optimization approach, where the subindex of each gen represents the parameter number of the MF in question and the superindex represent the number of the gen in question, remember that each Triangular

Table 2. Parameters of the Genetic Algorithm

Parameter	Value
Representation	real
Population size	10
Selection method	Roulette [15]
Cross method	two points
Rate of cross	0.8
Mutation method	Gaussian
Rate of mutation	0.1
Elitism [15]	2
Generations	100

Interval Type-2 MF is represented by six parameters (gens) which means the left hand side of Fig. 6 represents the MF *negative* of input variable `error`, and the right hand side of Fig. 6 represents the MF *positive* of the output variable `control`.

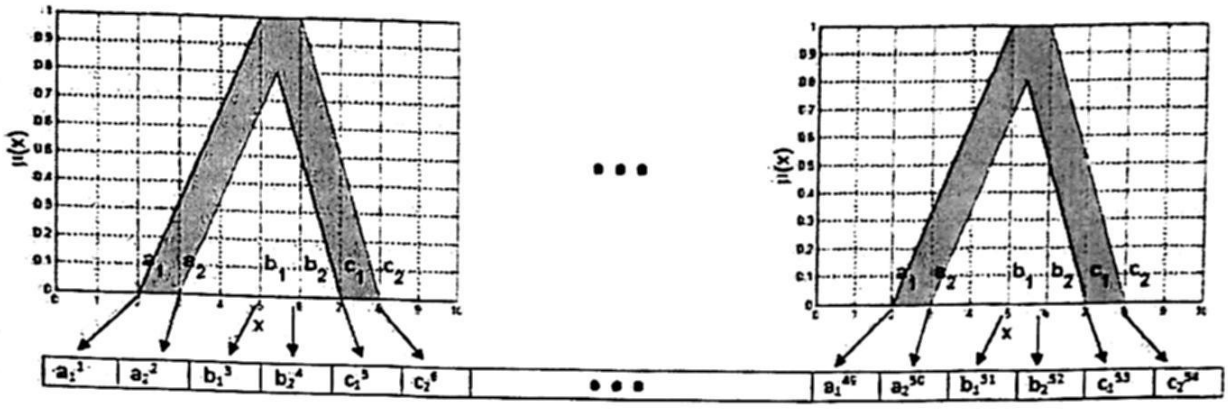


Fig. 6. Genotype for Type-2 FLS.

Our optimization problem is again a minimization, and with the ***objective function*** [15] of equation (16) we express that we want to minimize the mean error in our solutions space.

The set of the GA parameters are the same shown in Table 2.

6 Simulations Results

To perform simulations we use the dynamical model (7) of the experimental testbed installed in the Robotics & Control Laboratory of CITEDI-IPN (see Fig. 7), which involves a DC motor linked to a mechanical load through an imperfect contact gear train [21]. The parameters of the dynamical model (7) are in Table

3, while $N = 3$, $j = 0.2$ [rad], and $K = 5$ [N-m/rad]. These paremeters are taken from the experimental testbed.

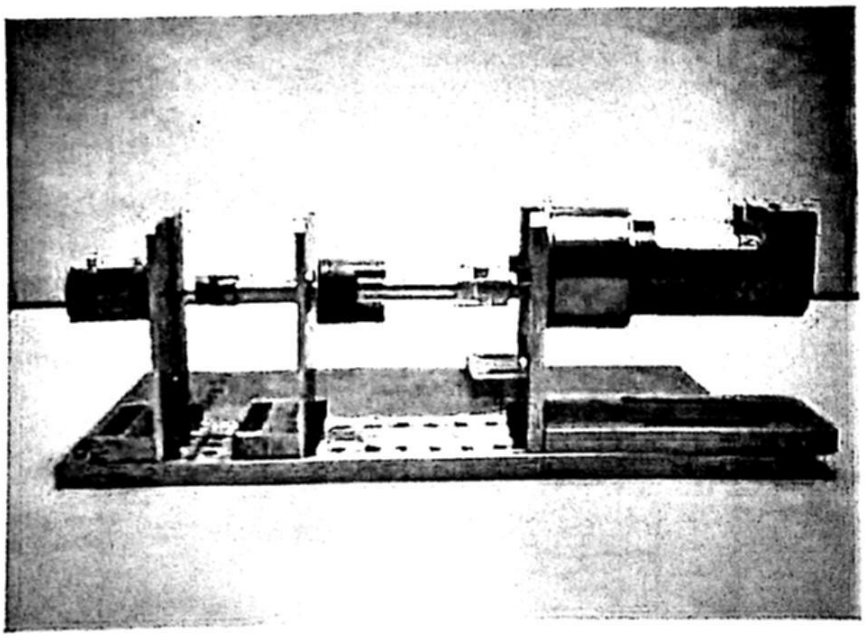


Fig. 7. Experimental test bench.

Table 3. Nominal parameters.

Description	Notation	Value	Units
Motor inertia	J_i	2.8×10^{-6}	Kg-m^2
Load inertia	J_o	1.07	Kg-m^2
Motor viscous friction	f_i	7.6×10^{-7}	N-m-s/rad
Load viscous friction	f_o	1.73	N-m-s/rad

The input-output motion graph of Fig. 8 reveals the gear backlash effect of the system.

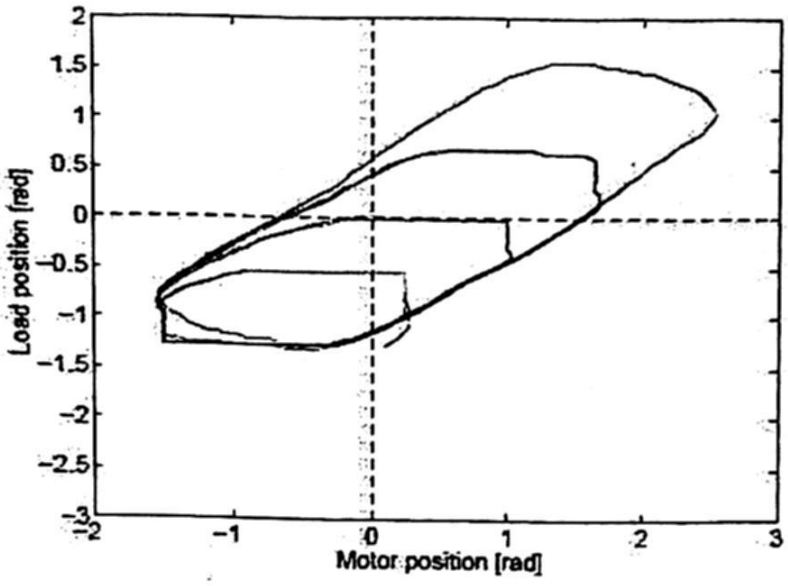


Fig. 8. Backlash hysteresis before compensation.

6.1 Numerical Solutions for the Type-1 FLS - Genetic Architecture

The GA was implemented using the Genetic Algorithm and Direct Search Toolbox [25], each individual (Type-1 FLS) of the population was tested in a closed-loop system modeled in Simulink, in that model we consider the angular motor position as the only information available for feedback. In the simulations, the load was required to move from the initial static position $q_0(0) = 0$ [rad] to the desired position $q_d = \pi/2$ [rad]. In order to illustrate the size of the attraction domain, the initial load position was chosen reasonably far from the desired position. The GA was executed in a PC Computer with Intel Pentium processor of 2.4 GHz and 512 Mb of RAM. Four executions conclude satisfactory and the solutions are concentrated in Table 4.

The GA was executed in about 26 hours (see Fig. 9 for details of convergence), producing the results shown in Table 4, Table 4 include the settling time of each individual, and as can be seen, all the individuals converge to a same fitness, this can be caused because there are just 27 gens in the chromosome (see Fig. 5) and the **Crossover Operation** makes than all individuals merge in the same one, moreover, the ***objective function*** (16) is not designed to considerate this case of situations.

Table 5 shows the chromosome of the resulting individual of Table 4, and its phenotype is in Fig. 10, its surface of control is in Fig. 11 and Fig 12 shows the system's response for this individual.

![h]

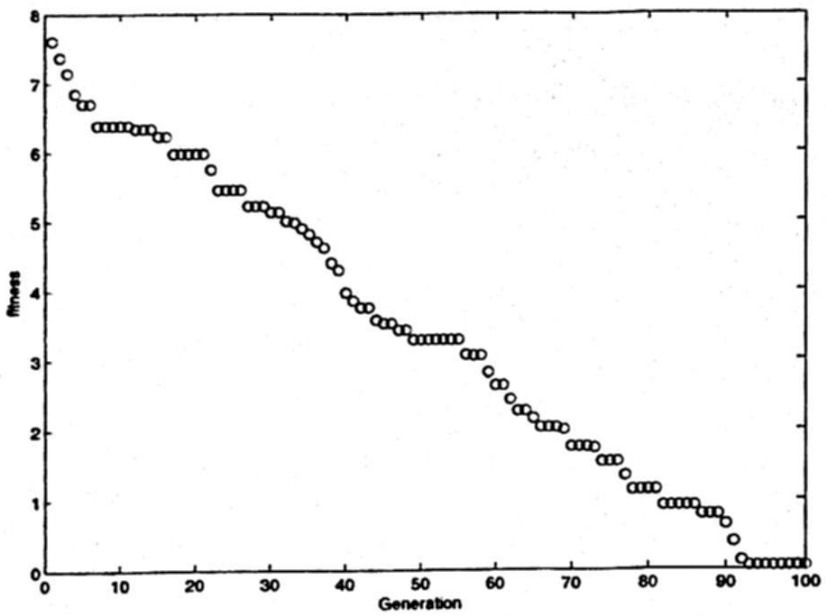


Fig. 9. Evolution of GA for the Type-1 FLS - Genetic Architecture.

Table 4. Genetic Algorithm Results for the *best* execution of the Type-1 FLS.

Individual	Fitness	Settling Time
1	0.0519	38.2158
2	0.0519	38.2158
3	0.0519	38.2158
4	0.0519	38.2158
5	0.0519	38.2158
6	0.0519	38.2158
7	0.0519	38.2158
8	0.0519	38.2158
9	0.0519	38.2158
10	0.0519	38.2158

Table 5. Data of the *best* individual of the Type-1 FLS - Genetic Architecture.

Variable	Membership Function	a	b	c
error	negative	-1.5000	-1.0000	-0.5377
	zero	-0.1173	-0.1173	2.3416
	positive	-0.5380	1.0000	1.5000
change of error	negative	-1.5000	-1.0000	0.0258
	zero	-0.4957	0.6905	0.6905
	positive	-0.5658	1.0000	1.5000
control	negative	-1.5000	-1.0000	-0.6166
	zero	-0.6619	0.7550	0.7550
	positive	0.0800	1.0000	1.5000

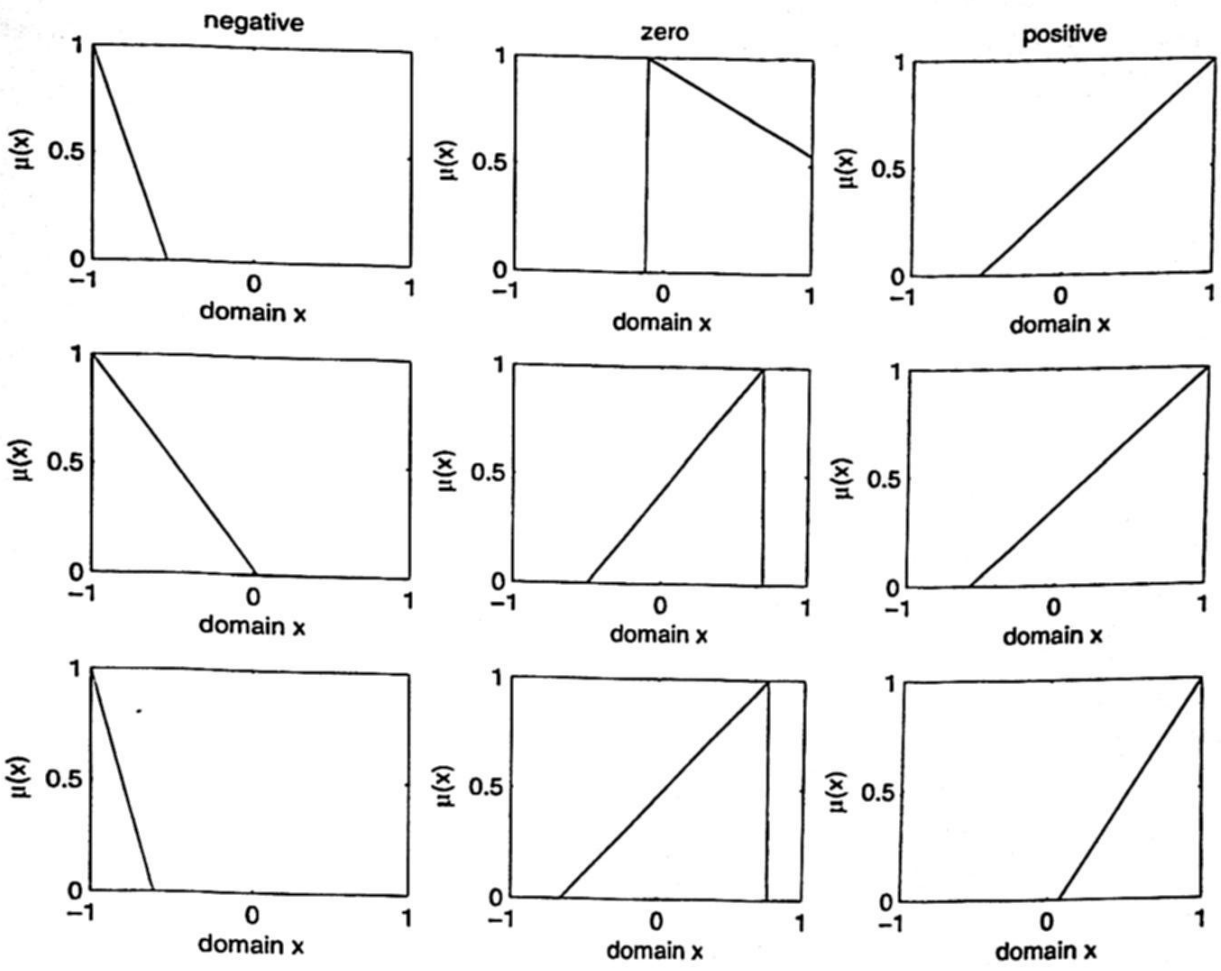


Fig. 10. Phenotype of the *best* individual of the Type-1 FLS - Genetic Architecture.

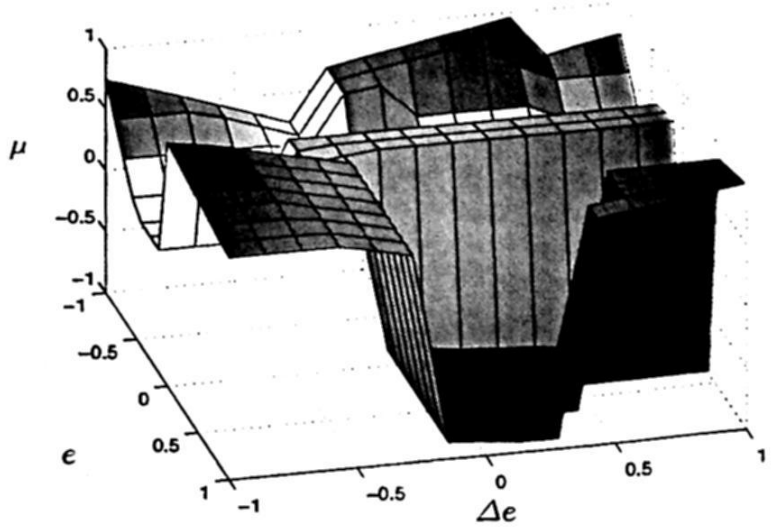


Fig. 11. Surface of control of the *best* individual of the Type-1 FLS - Genetic Architecture.

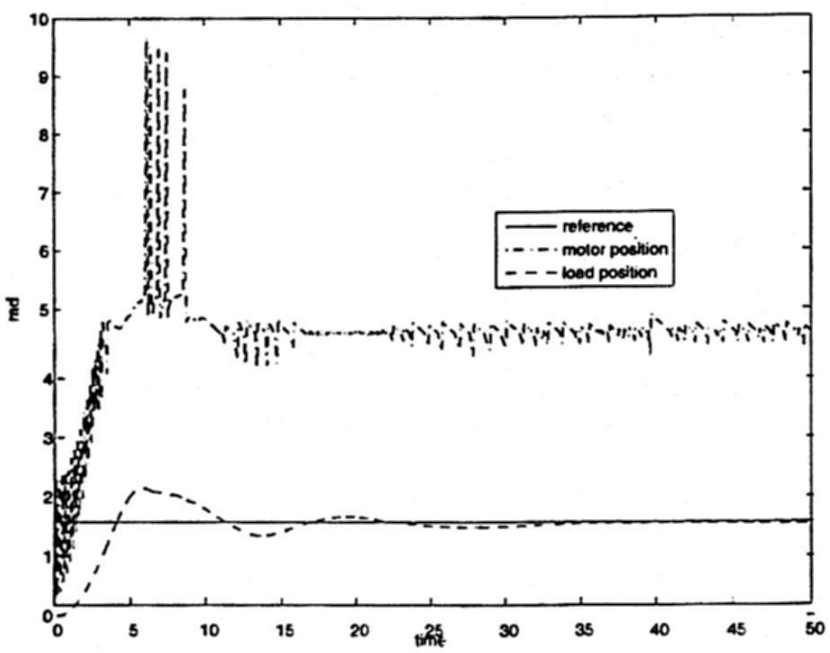


Fig. 12. Simulation result of the *best* individual of the Type-1 FLS - Genetic Architecture.

6.2 Numerical Solutions for the Type-2 FLS - Genetic Architecture

The GA was implemented using the Genetic Algorithm and Direct Search Toolbox [25], each individual (Type-2 FLS) was implemented in the Interval Type-2 Fuzzy Logic Toolbox reported in [23]-[24], and each individual of the population was tested in a closed-loop system modeled in Simulink, in that model we consider the angular motor position as the only information available for feedback. In the simulations, the load was required to move from the initial static position $q_0(0) = 0$ [rad] to the desired position $q_d = \pi/2$ [rad]. In order to illustrate the size of the attraction domain, the initial load position was chosen reasonably far from the desired position. The GA was executed in a PC Computer with Intel Pentium processor of 2.4 GHz and 512 Mb of RAM.

The GA was executed in about 200 hours (see Fig. 13 for details of convergence), producing the results shown in Table 6, including the settling time of each individual, and as can be seen, a best fitness do not mean a best settling time, this because the ***objective function*** (16) is not designed to achieve this performance measurement, moreover, some of this results can give a good performance in simulation, but in physical applications can be destructive for the mechanisms.

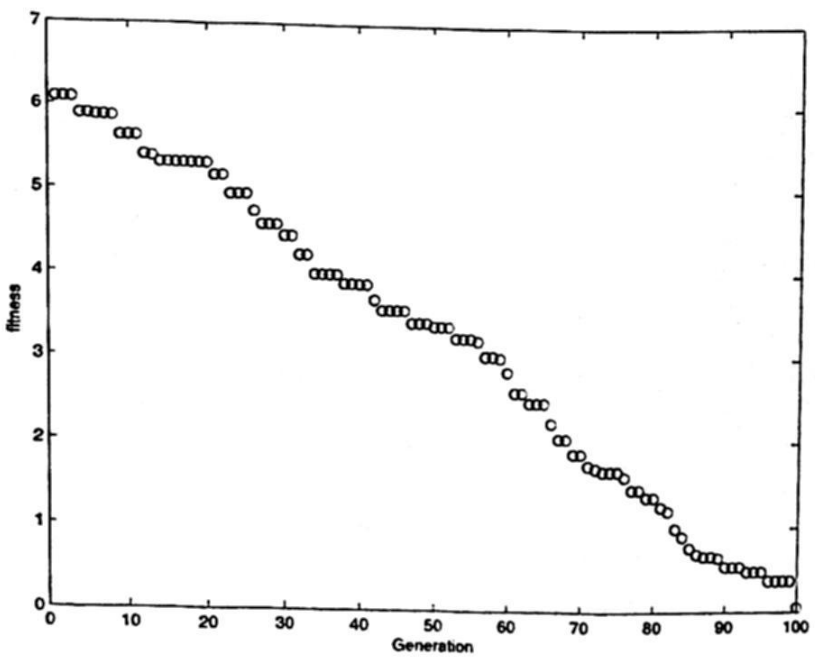


Fig. 13. Evolution of GA for the Type-2 FLS - Genetic Architecture.

From Table 6, the individual number six is the best solution of the GA, this because it have the best fitness, Table 7 shows the chromosome of individual number six who's phenotype [15] that is depicted in Fig. 14. Fig. 15 shows its surface of control and Fig. 16 shows the system's response for this individual number six.

Table 6. Genetic Algorithm Results of the Type-2 FLS.

Individual	Fitness	Settling Time
1	0.3032	17.1036
2	0.8898	24.1193
3	0.8593	15.3227
4	0.3046	17.0761
5	0.3809	16.0508
6	0.3004	15.2611
7	0.3308	16.4957
8	2.2201	22.9921
9	0.6113	20.8435
10	2.3717	35.2812

Table 7. Data of the *best* individual of the Type-2 FLS - Genetic Architecture.

Variable	μ	a1	b1	c1	a2	b2	c2
	negative	-1.0000	0.2176	0.2176	0.3012	0.4293	0.4293
error	zero	-0.1982	-0.1982	-0.1912	-0.4260	0.8378	0.8378
	positive	0.2669	0.2669	1.0000	-0.5479	-0.5479	1.0000
	negative	-1.0000	0.3504	0.6897	0.0926	0.9659	0.9659
change of error	zero	-0.6963	-0.6963	-0.1490	-0.4170	0.0000	0.4005
	positive	0.2611	0.4405	1.0000	-0.9561	-0.9561	1.0000
	negative	-1.0000	0.4586	0.4586	-0.6673	0.7247	0.7247
control	zero	-0.4571	-0.4571	0.7067	0.0446	0.0446	0.0446
	positive	0.6220	0.9432	1.0000	-0.0605	-0.0605	1.0000

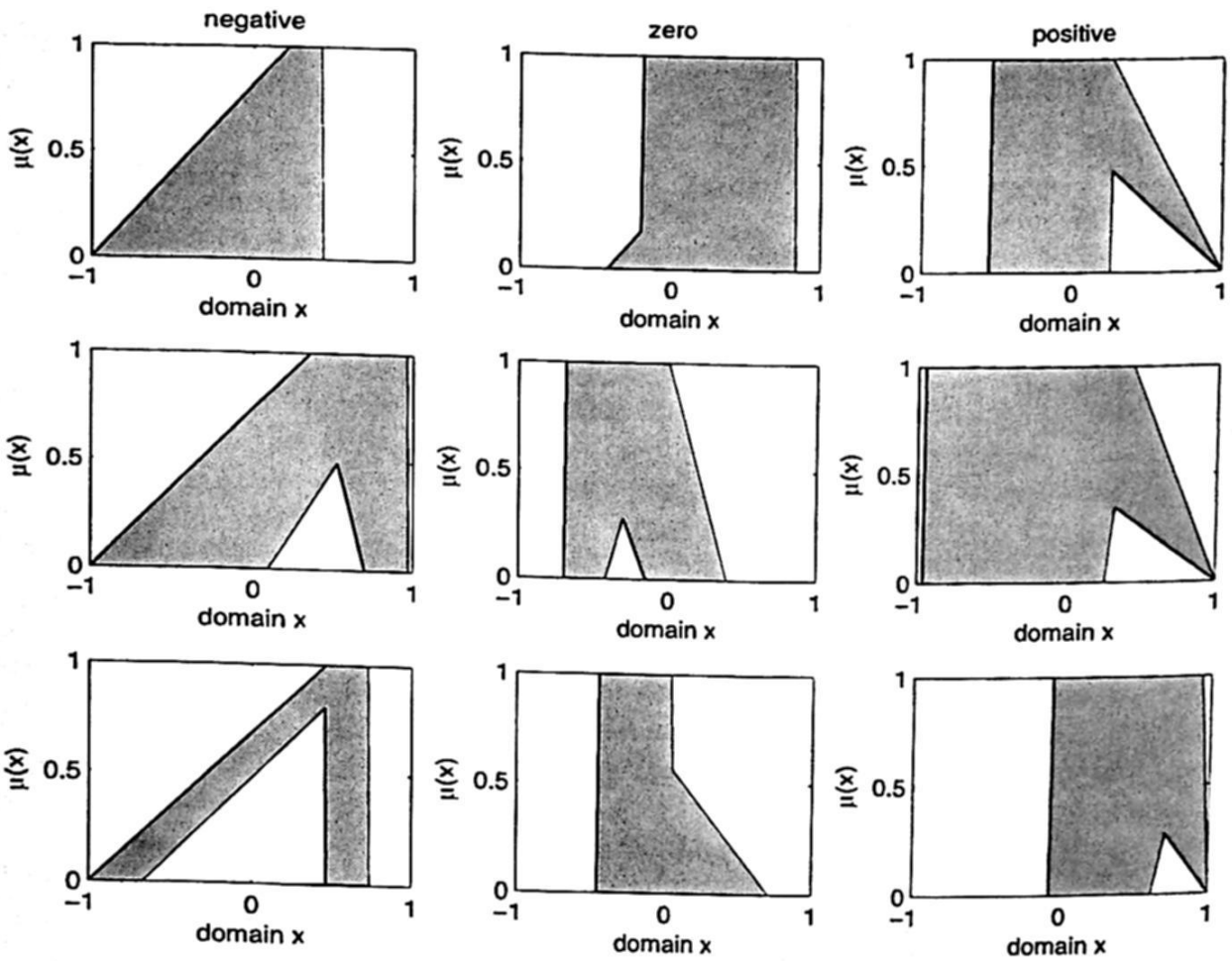


Fig. 14. Phenotype of the *best* individual of the Type-2 FLS - Genetic Architecture.

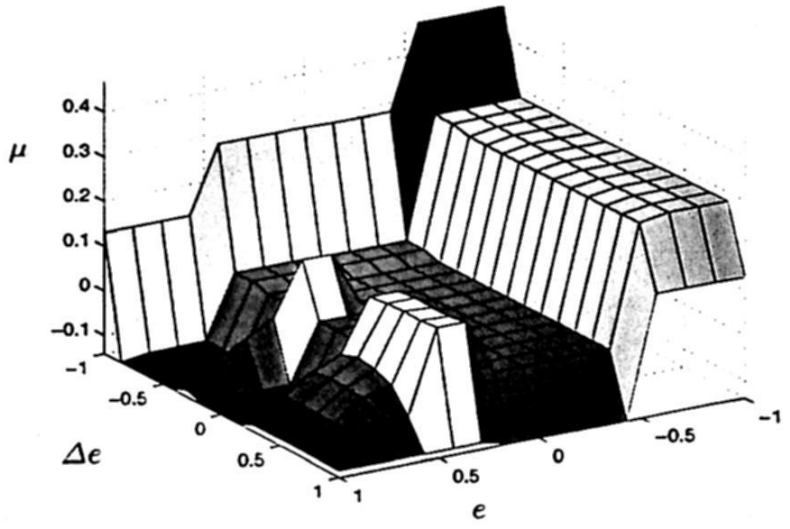


Fig. 15. Surface of control of the *best* individual of the Type-2 FLS - Genetic Architecture.

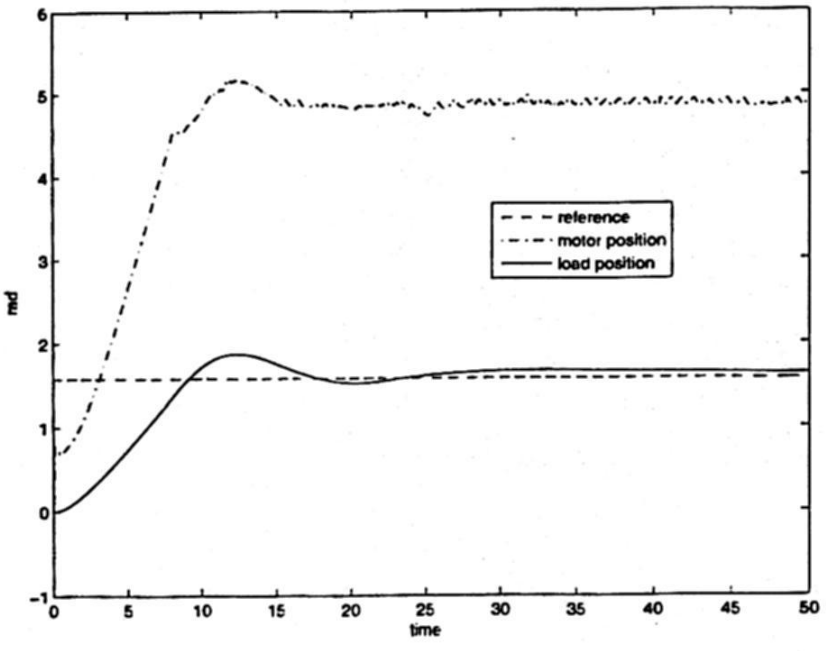


Fig. 16. Simulation result of the *best* individual of the Type-2 FLS - Genetic Architecture.

7 Conclusion

The main goal of this paper was to develop hybrid approaches combining GAs with Type-1 and Type-2 FLSs for the design and optimization of Type-1 and Type-2 FLCs. The optimized Type-1 and Type-2 FLCs were designed for the case of study of the output regulation of a servomechanism with backlash, giving to the GA a heavy task because the nonlinearity of the proposed problem, and a heavy task test to the Type-1 and Type-2 FLCs for the high uncertainty in the case of study.

The GA was implemented and performed in a satisfactory fashion, giving a whole family of solutions of Type-1 and Type-2 FLCs to the problem in question, some of the resulting solutions give us best performance that other ones.

In this paper we are reporting explicitly just one of the solutions for Type-1 and Type-2 FLCs because of space limitations. The solutions reported in this paper are evidently different each from another, confirming that as in the Type-1 FLS, Type-2 FLS can be obtained from human expertise, and in this case, each solution can be a representation of the expertise of different experts.

The combination of a GA and Type-1 and Type-2 FLSs results to be a good method for the proposed problem, but the time necessary to run each generation of the GA is too big, and maybe we can find another set of solutions faster by trial and error, but those results most probably will not be optimal.

If we compare the resulting best optimized Type-1 and Type-2 FLCs, is evidently that the Type-2 FLC is better that the Type-1 FLC, this because from 12 and 16 we can see that the response of the closed loop system is more soft from optimized Type-2 that from optimized Type-1 FLCs, and in 12 we can see that the optimized Type-1 FLC is giving a motor behavior that may be destructive to the physical mechanism and this do not occur with the optimized Type-2 FLC. Moreover, the settling time of the response of the closed-loop systems with the optimized Type-1 FLC is twice as long that the responte of of the closed-loop systems with the optimized Type-2 FLC.

Comparing the necessary time to the convergence of the Type-1 and Type-2 FLCs approaches, optimize the Type-2 FLC needs almost eight times that the time needs to optimize the Type-1 FLC.

GAs has been proved in this approach that are a good method to optimize MF's parameters in the designing of Type-1 and Type-2 FLSs, however, we must be aware of the price we pay for using GA, that is, we must to be agree in spend so much computation resources, like memory and CPU time.

References

1. Grefenstette, J.: Optimization of control parameters for genetic algorithms. IEEE Trans. Syst. Man Cybern. **16** (1986) 122–128
2. Lee, M., Takagi, H.: Integrating design stage of fuzzy systems using genetic algorithms. Fuzzy Systems, 1993., Second IEEE International Conference on (1993) 612–617 vol.1

3. Melin, P., Castillo, O.: Intelligent control of complex electrochemical systems with a neuro-fuzzy-genetic approach. IEEE Transactions on Industrial Electronics **48** (Oct 2001) 951–955
4. Castillo, O., Lozano, A., Melin, P.: Hierarchical genetic algorithms for fuzzy system optimization in intelligent control. Fuzzy Information, 2004. Processing NAFIPS '04. IEEE Annual Meeting of the **1** (27-30 June 2004) 292–297 Vol.1
5. Mendel, J.: Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions. Prentice Hall, Upper Saddle River, NJ (2001)
6. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning-1. Information Sciences **8** (1975) 199 – 249
7. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning-2. Information Sciences **8** (1975) 301 – 357
8. Zadeh, L.A.: The concept of a linguistic variable and its application to approximate reasoning-3. Information Sciences **9** (1975) 43 – 80
9. Hagras, H.: Type-2 flcs: A new generation of fuzzy controllers. IEEE Computational Intelligence Magazine **2** (2007) 30–44
10. Sepulveda, R., Castillo, O., Melin, P., Rodriguez-Diaz, A., Montiel, O.: Experimental study of intelligent controllers under uncertainty using type-1 and type-2 fuzzy logic. Inf. Sciences **177** (2007) 2023–2048
11. Al-Jaafreh, M., Al-Jumaily, A.: Training type-2 fuzzy system by particle swarm optimization. Evolutionary Computation, 2007. CEC 2007. IEEE Congress on (25-28 Sept. 2007) 3442–3446
12. Sepulveda, R., Castillo, O., Melin, P., Montiel, O., Aguilar, L.: Evolutionary optimization of interval type-2 membership functions using the human evolutionary model. Fuzzy Systems Conference, 2007. FUZZ-IEEE 2007. IEEE International (23-26 July 2007) 1–6
13. Castillo, O., Huesca, G., Valdez, F.: Evolutionary computing for optimizing type-2 fuzzy systems in intelligent control of non-linear dynamic plants. Fuzzy Information Processing Society, 2005. NAFIPS 2005. Annual Meeting of the North American (26-28 June 2005) 247–251
14. Castillo, O., Aguilar, L.T., Cazarez, N., Cardenas, S.: Systematic design of a stable type-2 fuzzy logic controller. Journal of Applied Soft Computing **8** (2008) 1274–1279
15. Castillo, O., Melin, P.: Hybrid Intelligent Systems for Pattern Recognition Using Soft Computing: An Evolutionary Approach for Neural Networks and Fuzzy Systems. Springer-Verlag, Berlin (2005)
16. Cordón, O., Herrera, F., Hoffman, F., Magdalena, L.: Genetic fuzzy systems: Evolutionary tuning and learning of fuzzy knowledge base. World Scientific, Singapore (2001)
17. Lagerberg, A., Egardt, B.: Estimation of backlash with application to automotive powertrains. Proc. of the 42th Conf. on Decision and Control (1999) 135–147
18. Mendel, J., John, R.: Type-2 fuzzy sets made simple. IEEE Transactions on Fuzzy Systems **10** (2002) 117 – 127
19. Mamdani, E.H., Assilian, S.: An experiment in linguistic synthesis with a fuzzy logic controller. Int. J. Hum.-Comput. Stud. **51** (1999) 135–147
20. Holland, J.M.: Adaptation in Natural and Artificial Systems. University of Michigan Press, Ann Arbor, MI (1975)
21. Aguilar, L.T., Orlov, Y., Cadiou, J.C., Merzouki, R.: Nonlinear H_∞-output regulation of a nonminimum phase servomechanism with backlash. Journal of Dynamic Systems, Measurement, and Control **129** (2007) 544–549

22. Goldberg, D.E.: Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA (1989)
23. Castro, J.R., Castillo, O., Martinez, L.G.: Interval type-2 fuzzy logic toolbox. Journal of Engineering Letters **15** (2007) 89–98 online version.
24. Castro, J.R., Castillo, O., Melin, P.: An interval type-2 fuzzy logic toolbox for control applications. In: Proc. FUZZ-IEEE 2007. IEEE, London (2007) 61 – 67
25. : Genetic Algorithm and Direct Search Toolbox User's Guide, The Mathworks Inc. Available: http://www.mathworks.com. (2007)

Computing Pragmatic Similarity in Distributed Interaction Systems

Maricela Bravo

Morelos State Polytechnic University, Cuauhnáhuac 566, Texcal,
Morelos, México, CP 62550
mbravo@upemor.edu.mx

Abstract. A Distributed Interaction System (DIS) consists of a set of autonomous software agents, capable of communicating among each other with cooperation and coordination purposes. To achieve their goals, all software agents in a DIS must exchange messages following a protocol. Currently there are research efforts to provide standard mechanisms to achieve cooperation between multiple heterogeneous DIS. However, there are still some issues that must be solved in order to fully automate integration and inter-operation between them. One of this problems is interoperation In this paper we present a novel approach for computing pragmatic similarity for multiple DIS, our objective is to support system developers and integrators with specific information concerning the pragmatic similarity among multiple software agents. We describe a case study to show the applicability of our approach.

Keywords: Distributed Interaction Systems, Pragmatic Similarity Measure.

1 Introduction

A DIS consists of a set of autonomous, independently developed software agents, which have communication capabilities for cooperation and coordination among them. To achieve their goals all software entities in a DIS, exchange messages following an interaction protocol. Currently Internet-based environments have gained more atention than ever, and many software entities independently developed are being deployed on the Web, causing the problem of heterogeneity. Heterogeneity in DIS has many causes: use of different hardware and operating systems, use of different programming languages for implementation, different data base management software, different naming techniques, different data and code formats, etc. Considering this inherent DIS heterogeneity, to provide software entities with mechanisms to be adaptable at run time to interoperate in dynamic and complex environments such as Internet, represents an open problem. Therefore, research efforts to model interactions between different software entities, and evaluate the behavior similarity among them, will benefit the design selection of an interoperable solution for DIS over Internet.

Many authors have approached this problem from a data-based perspective, which we will refer as the data approach. We consider this is a good solution approach,

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 155-170

Received 05/07/08
Accepted 01/08/08
Final version 14/09/08

which has proven good results. However, this data approach has been mainly used for heterogeneous information sources integration: such as distributed data bases, vocabularies, ontologies, taxonomies, etc. This data approach would be enough if Internet was populated only with information sources, but it is not. Indeed, Internet has more than only data; it has programs, agents, Web applications, etc. We will refer to this kind of sources as interactive software entities. These interactive software entities have to communicate to each other through the exchange of messages, following interaction protocols or rules.

In this paper we focus on the pragmatic aspect of interactive software entities, in particular we propose an approach for computing pragmatic similarity between software entities, which use different interaction protocols. Our approach is based on the analysis of transition functions extracted from interaction protocols modeled with Finite State Machines (FSM). The aim of this research is to provide software programmers or system integrators with mechanisms to compute pragmatic similarity among interactive software entities, in order to identify or propose a good solution for dynamic interoperation.

To evaluate our solution we have implemented a Web-based interaction environment, into which different software entities can be deployed to execute communications among them. Our environment incorporates an Ontology-based translator to help software entities whenever there is a misunderstanding. We also proposed a pragmatic measure to evaluate a priori the level of similarity. We executed a series of experiments to show that the resulting set of pragmatic relations improve interactions between heterogeneous software entities.

The rest of the paper is organized as follows. In section two we present a brief description of representation formalisms which have been used for modeling, simulating and implementing interactions in DIS. In section three we describe the pragmatic representation of software entities. In section four we present an example to show the applicability of our approach. In section five we describe the implemented interaction environment. In section six we present the experimental results. In section seven we present related work and finally, in section eight we conclude.

2 Representation Formalisms

There are various formalisms reported in literature to model DIS, for example Petri Nets, Colored Petri Nets, Pi-calculus, AUML, BPEL, OWL-S.

Petri Nets. Petri Nets were first introduced by Carl Adam Petri, as a result of his Ph.D. thesis in 1962 [1]. Petri Nets have been used to analyze and verify systems in different areas of science, such as artificial intelligence, concurrent systems, control systems, analysis of networks, etc. Petri Nets represent a traditional formalism for modeling interactions and concurrency. A Petri Net is a directed, connected, bipartite graph with annotations, in which each node is either a place or a transition. Tokens are in places, when there is at least one token in every place connected to a transition, then that transition is enabled.

Colored Petri Nets. Colored Petri Nets [2] are based on Petri Nets, but the have added properties. Tokens are not simply blank markers, but have data associated to

them. A color in a token represents a schema or type specification. Places are sets of tuples, called multi-sets. Arcs specify the schema they carry, and can also specify Boolean conditions. Arcs exiting and entering a place may have an associated function which determines what multi-set elements are to be removed or deposited. Boolean expressions, called guards, are associated with the transitions, and enforce some constraints on tuple elements. Colored Petri Nets are equivalent to Petri Nets, but the richer notation of colored Petri Nets makes them more suitable for modeling interactions with more information.

Pi-Calculus. Is a process algebra presented in [3]. It is a formalism for modeling concurrent processes, whose configurations may change as the process executes over time. In Pi-Calculus the fundamental unit of computation is the transfer of a communication link between two processes. The simplicity of the Pi-Calculus is because it includes only two kinds of entities: names and processes. These entities are sufficient to define interaction behavior.

AUML. Agent Unified Modeling Language is a representation formalism which facilitates the visual development of multi-agent systems, with emphasis on agent conversations. It was first introduced by [4]. AUML is concerned mainly with interaction diagrams for conversation modeling. Interaction diagrams mainly extend the OMG definition of UML sequence diagrams with the possibility to express explicitly concurrency in the sending of messages.

BPEL. Business Process Execution Language [5] is a formalism used for specifying the composition of Web services. It was created to standardize interaction logic and process automation between Web services. BPEL is a convergence of language features form WSFL and XLANG. However, this language lacks well defined semantics, which makes it difficult to reuse and compose.

OWL-S. OWL-S [6] is one of the standards for the description of Web services. It includes a process model for Web services. Each process is described by three components: inputs, preconditions and results. Results specify what outputs and effects are produced by the process under a given condition.

FSM. FSM [7] represents a powerful formalism for describing and implementing the control logic of an interaction system. They are suitable for implementing communication protocols, control interactions and describe transitional functions. FSM mainly consist of a set of transition rules. In the traditional FSM model, the environment of the machine consists of two finite and disjoint sets of signals, input signals and output signals. Also, each signal has an arbitrary range of finite possible values.

The above described representation formalisms are useful to model interactions among distributed software entities. However, not all have the same purpose and facilities. Some are good for modeling interactions formally (Petri Nets, Colored Petri Nets, Pi-Calculus and AUML), others have tools for verification and simulation (Petri Nets and Colored Petri Nets), others are good for executing and implementing composite processes (BPEL and OWL-S). But for the aim of this work we have selected FSM because they offer a simple manner of implementing transitional functions in order to compute similarity in pragmatics of represented protocols.

3 Representing Pragmatics

One of the main problems we faced when trying to evaluate the pragmatics of a software agent, was the selection of the appropriated interaction protocol representation formalism. As it was described in Section 2, there are various formalisms reported to achieve this goal. However, some are useful for modeling (UML diagrams, Pi-Calculus), some others are good for executing processes (BPEL, OWL-S), some are good for simulating interaction processes (Petri Nets and Colored Petri Nets), but we needed to use a formalism easy to implement and therefore to compute. We also needed a formalism which would help us in the automatic discovering of functionalities, which is our work in progress. Thus, we selected FSM for this reason.

A **FSM** is a tuple (S, *I*, *O*, *ft*, *fo*, *s0*), where
S is a finite set of states,
I is the set of inputs,
O is the set of outputs,
ft is the transitional function,
fo is the output function and
s0 represents the output state.

For the purpose of our work we have adapted this FSM definition in order to represent pragmatics of interaction systems. In particular we have adapted the transition functions to represent initial states, and final states generated by an incoming message.

The methodology we followed for extracting and representing pragmatics from a software entity is described next.

a) Acquisition of software entities protocols and communication messages details. Currently this process is executed manually through a Web-based environment.

b) FSM drawing. This process requires extensive human interaction, because we need to build FSM in order to obtain transition functions from them.

c) Transition functions generation. This process consists of analyzing the FSM diagrams to identify arcs and define the transitional functions.

In order to compute pragmatic similarity between different interaction protocols we established a common set of states, to allocate all the messages and transition functions in those states. Therefore, we adopted the proposal of Müller [8]. Müller specifies that any communication protocol consists of three general states: ***start***, ***react*** or ***complete*** depending on the moment in the FSM when the primitive is issued, but we added another the ***modify*** state to be more specific about a conversation between software entities.

4 An Example

In this section we present an example to show the applicability of our approach. The objective of this case is to represent and measure similarity among a set of interaction protocols from three different software entities using their communication message details.

Given a DIS integrated with three software entities e_1, e_2, and e_3, each with its own interaction protocol ***IP***.

$$DIS = \{ IPe_1, IPe_2, IPe_3 \}$$

For each IP there is a set of primitives which are used as communicative acts to send and receive messages among them. The set of interaction primitives for each IP are as follows.

IPe_1 = {*Initial_Offer, RFQ, Accept, Reject, Offer, Counter_Offer*}

IPe_2 = {*CFP, Propose, Accept, Terminate, Reject, Acknowledge, Modify, Withdraw*}

IPe_3 = {*Requests_Add, Authorize_Add, Require, Demand, Accept, Reject, Unable, Require-for, Insist_for, Demand_for*}

Based on the set of primitives described we manually generated the FSM and defined the set of transition functions to compute similarity. In Fig. 1 we present the state transition diagram of interaction protocol of software A. In Table 1 we describe the set of transition functions of software A.

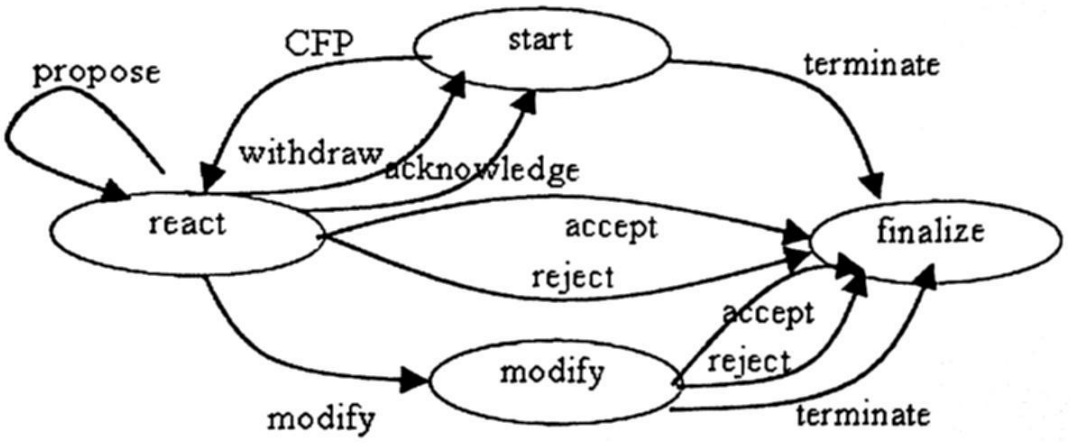


Fig. 1. Interaction protocol for software A

Table 1. Transition functions of software A

Transition functions of interaction protocol of software A
ft(start, CFP) = react
ft(react, Propose) = react
ft(react, Acknowledge) = start
ft(react, Modify) = modify
ft(react, Withdraw) = start
ft(react, Reject) = finalize
ft(react, Accept) = finalize
ft(modify, Reject) = finalize
ft(modify, Accept) = finalize
ft(react, Terminate) = finalize
ft(start, Terminate) = finalize

Fig. 2 shows the state transition diagram of software B. In Table 2 we present the transition functions of interaction protocol of software B.

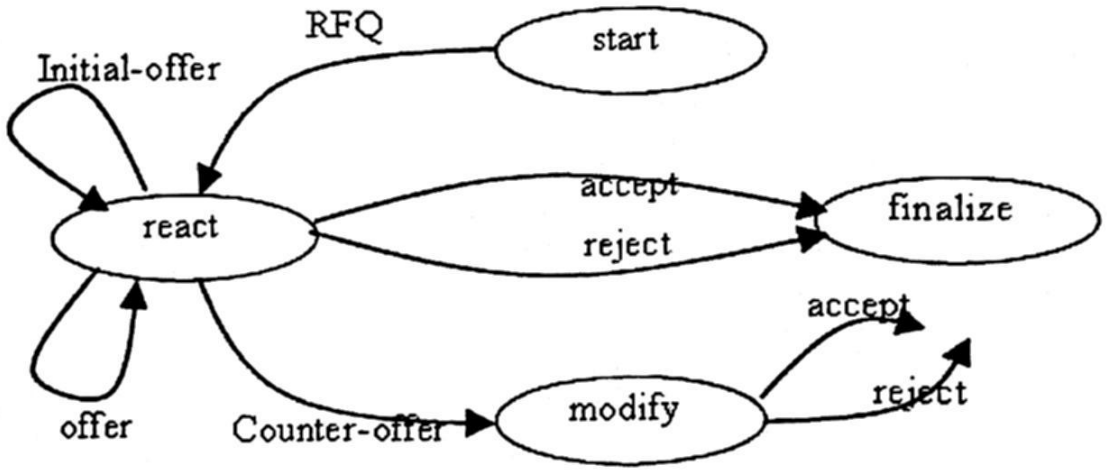


Fig. 2. Interaction protocol of software B

Table 2. Transition functions of software B

Transition functions of interaction protocol of software B
ft(start, RFQ) = react
ft(react, Initial-offer) = react
ft(react, Counter-offer) = modify
ft(react, Offer) = react
ft(modify, Accept) = finalize
ft(modify, Reject) = finalize
ft(react, Reject) = finalize
ft(react, Accept) = finalize

Fig. 3 presents the state transition diagram of software C. Table 3 shows the transition functions of interaction protocol of software C.

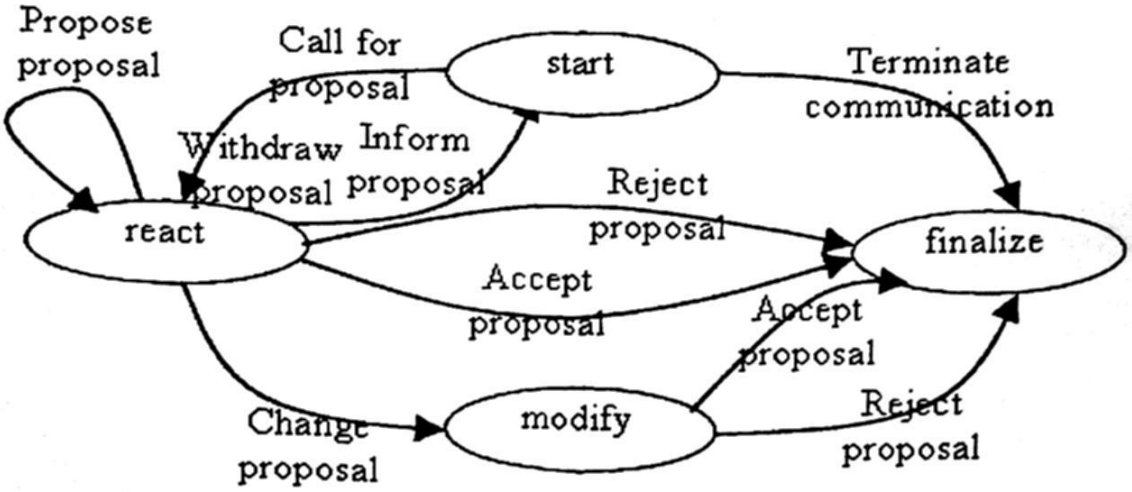


Fig. 3. Interaction protocol of software C

Table 3. Transition functions of software C

Transition functions of interaction protocol of software C
ft(start, Call for proposal) = react
ft(react, Propose proposal) = react
ft(react, Inform Proposal) = start
ft(react, Change Proposal) = modify
ft(react, Withdraw Proposal) = react
ft(react, Reject Proposal) = finalize
ft(react, Accept Proposal) = finalize
ft(modify, Reject Proposal) = finalize
ft(modify, Accept Proposal) = finalize
ft(start, Terminate Communication) = finalize

4.1 Computing Similarity

To compute pragmatic similarity we need first to calculate the total number of different interaction links and identify the set of pairs of software entities that will interact among them.

a) Number of Interaction links

Considering a set of n software entities, the possible number of peer to peer interaction links among them is n^2. However, as we are evaluating heterogeneity, we need to extract the number of interaction links where software entities are equal, which is n. We also considered that a interaction link between software entities (a, b) has the same heterogeneity as an interaction link of

software entities (b, a), thus we reduced the number of different interaction links dividing by 2.

$$CL = (n^2-n)/2$$
$$CL = (3^2-3)/2 = 3 \quad (1)$$

B) Set of Different Interaction links
Considering a set of software entities, the set of different interaction links is given by:

$$\boldsymbol{DCL} = \{ (a_1, a_2), (a_1, a_3), \ldots, (a_i, a_j) \}$$
$$\boldsymbol{DCL} = \{ (A, B), (A, C), (B, C) \} \quad (2)$$

Algorithm for Computing Pragmatic Similarity

Our algorithm is based on the equivalence definition of FSM. Given two FSM, two states that belong to each FSM are said to be equivalent if for a given initial state and a given input message, the resulting state is the same. We adapted this definition, but we are considering that the input messages may be syntactically different, and then we are computing similarity among these different messages as follows: if their initial states and their final states are equal, then the syntactically different messages are pragmatically similar. To compute pragmatic similarity we implemented an array-based algorithm. For our case study we implemented three arrays, each with three columns which represent: the initial state, the input primitive and the final state. The algorithm is executed for each different communication link (ai, aj), where ai represents the array of agent i.

The result of this algorithm is a set of relations, which will help as the basis for a translation approach solution.

```
For each transition function ft of a_i
 For each transition function ft of a_j
  If (a_i[initial-state] is equal to a_j[initial-state])
     and (a_i[final-state] is equal to
             a_j[final-state])
             if (a_i[input-primitive]
                is-different-syntactically to
                 a_j[input primitive])
                   a_i[input-primitive]
```

```
is-similar-pragmatic to
a_j[input primitive]
```

After obtaining the resulting set of similar functions, we have to evaluate them, in order to check inconsistencies. We defined only similar pragmatic relations for primitives that are syntactically different. We did not established differences as relations, because this kind of relations will not support interoperability. However, they are important to measure heterogeneity and to propose another solution based on a learning approach. Results of this process are shown in Tables 4, 5 and 6. To define relations we used the form:

$$\boldsymbol{REL}(S_i, P_i, S_j, P_j)$$

where

S_i is the software issuer of primitive P_i
S_j is the software issuer of primitive P_j

Table 4. Set of relations of interaction link between software entities A and B

IL(A, B)
REL(A, CFP, B, RFQ)
REL(A, Propose, B, Initial_Offer)
REL(A, Modify, B, Counter_offer)
REL(A, Propose, B, Offer)
REL(A, Terminate, B, Reject)
REL(A, Terminate, B, Accept)

Table 5. Set of relations of interaction link between software entities A and C

IL(A, C)
REL(A, CFP, C, Call for proposals)
REL(A, Propose, C, Propose proposal)
REL(A, Modify, C, Change proposal)
REL(A, Withdraw, C, Withdraw proposal)
REL(A, Acknowledge, C, Inform proposal)
REL(A, Accept, C, Accept proposal)
REL(A, Reject, C, Reject proposal)
REL(A, Terminate, C, Terminate communication)

Table 6. Set of relations of interaction link between software entities B and C

IL(B, C)
REL(B, RFQ, C, Call for proposals)
REL(B, Offer, C, Propose proposal)
REL(B, Counter_offer, C, Change proposal)
REL(B, Accept, C, Accept proposal)
REL(B, Reject, C, Reject proposal)
REL(B, Initial_Offer, C, Propose proposal)

4.2 Pragmatic Similarity Measure

Another important result is a pragmatic similarity measure, which will help to analyze more precisely the level of similarity of a set of software entities. In this section we describe this measure.

The pragmatic similarity measure is a ratio which results from dividing the number of equivalent functions by the total number of transition functions from participating software entities protocols.

a) Number of transition functions

The total number of transition functions is obtained from the sum operation of all sets of transition functions.

$$NTF = |TFa_1 + TFa_2 + ... + TFa_n| \quad (3)$$

b) Number of equivalent functions

The total number of equivalent functions (NEF) results from the sum of the resulting set of the algorithm.

c) Pragmatic similarity

The pragmatic similarity results from dividing the number of equivalent functions by NFT, which is the ratio that will serve as an indicator for evaluating pragmatic similarity.

$$Pragmatic\ similarity = NEF / NFT \quad (4)$$

Table 7. Pragmatic similarity measure between each interaction link

IL	*NEF*	*NTF*	*Pragmatic Similarity*
A, B	6	19	.31
A, C	8	21	.38
B, C	6	18	.33

The resulting pragmatic similarity ratios for the three interaction links is shown in Table 7.

Therefore a ratio of 1 indicates a fully pragmatic similarity between software entities, while a ratio of 0 indicates impossibility of interoperation between both software entities.

5 Interaction Environment

To evaluate our approach we implemented an interaction environment populated with the set of software entities described in Section 4. We also implemented a translator module which is invoked whenever is necessary. The general architecture for the execution of interaction protocols is illustrated in Figure 4. This architecture has an intermediary program which is responsible for initialization of interaction processes, sending and receiving messages form both software entities, and recording the communication until an ending message is issued by any of the participants.

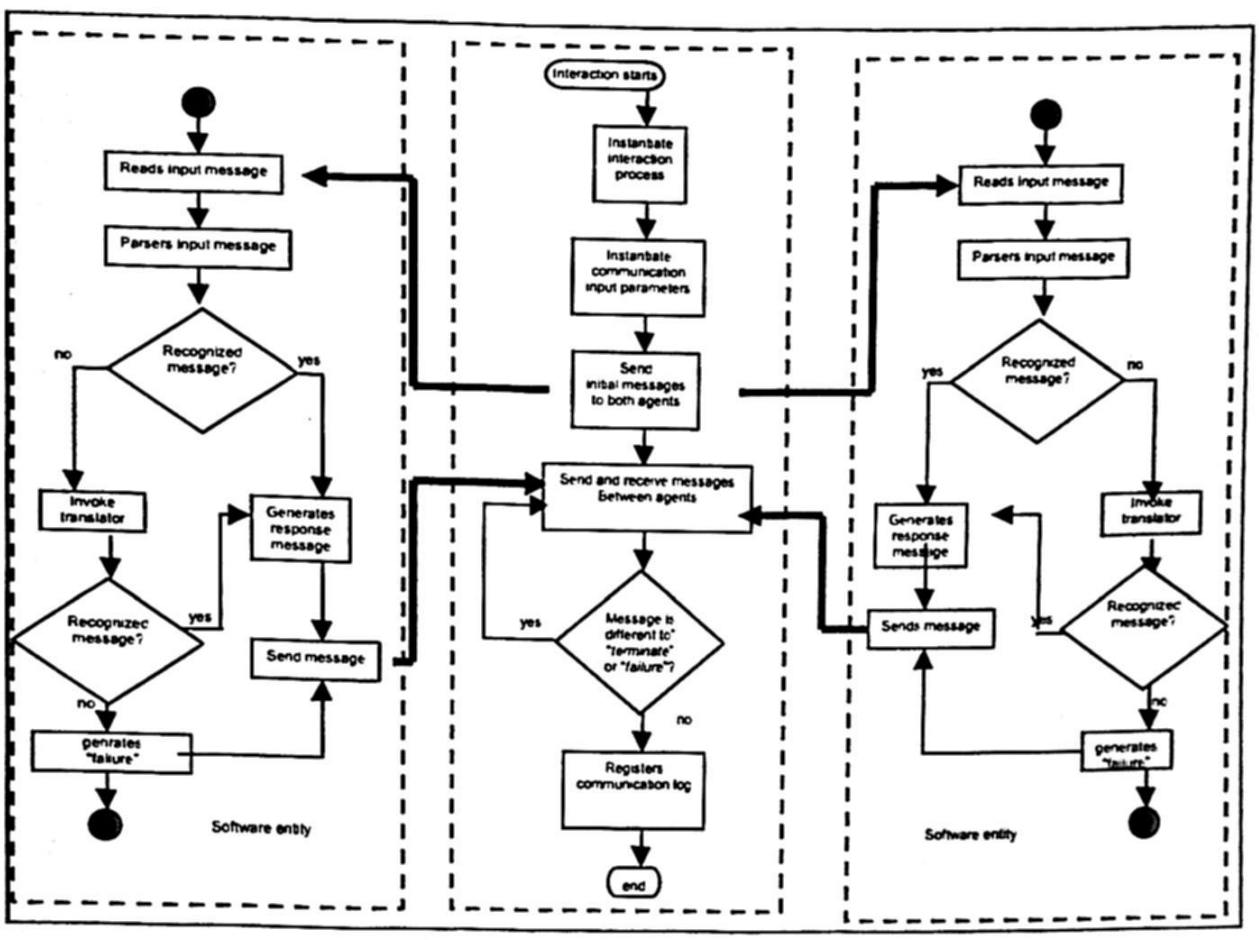


Fig. 4. Interaction Web-based architecture

The software entities are programmed by their respective owners to define their preferences and interaction strategies. For example, a seller software entity will be programmed to maximize his profit, establishing the lowest acceptable price and the

desired price for selling. In contrast, a buyer software entity is seeking to minimize his payment.

The translator module is invoked whenever the software does not recognize a message from the other software. The translator module was implemented using Jena[1], a framework for building Semantic Web applications. It provides a programmatic environment for OWL, including a rule-based inference engine. For the description and execution of communication processes, we used BPEL4WS. BPEL4WS defines a model and a grammar for describing the behavior of a business process based on interactions between the process and its partners. BPEL4WS represents a convergence of the ideas in the XLANG and WSFL specifications. Both XLANG and WSFL are superseded by the BPEL4WS specification. The interaction with each partner occurs through Web service interfaces, and the structure of the relationship at the interface level is encapsulated in what we call a partner link. The BPEL4WS process defines how multiple service interactions with these partners are coordinated to achieve a business goal, as well as the state and the logic necessary for this coordination.

5.2 Translator Functionality

The translator acts as an interpreter of different software entities. The translator module first reads the input parameters, and then opens a connection to the Ontology to make queries. When the connection has been established it narrows the search by selecting only instances of the same *Type* of the incoming "message". It then searches for and retrieves all primitives that belong to the "receiver" software entity in the same *Type*. And finally it makes a comparison to find a similar primitive, when this process ends it returns the equivalent primitive; in other case it returns a failure message.

A misunderstanding event occurs when a software entity receiving a message, compares it to its own message knowledge base, and acknowledges that the received message is not in his base, and then invokes the translator to find the relation with its own messages.

6 Experimentation Results

We executed a series of tests with these software entities. In this section we present one experiment of a set of 30 communication executions between software entities A and B. The result of this experiment is shown in Table 8.

Table 8 shows that for the first set of executions many ended because of misunderstandings. For the second set of executions we can appreciate a reduction in the number of misunderstandings, although the problem remains, some communications are still ending due to this problem. For this case we suggest using a learning approach, in order to fully eliminate the problem. However, the result is good enough to evaluate the main contribution of this paper. The set of discovered

[1] http://jena.sourceforge.net

pragmatic relations were well defined by our semi-automatic approach. The Ontology was populated with these primitives and relations among them, and when integrated to the general environment it solved the main problem of our work: interoperability.

Table 8. Experimental results

	Result	Result
Test number	**No translation**	**Using translation**
1	notUnderstood	Accept
2	notUnderstood	Reject
3	notUnderstood	Accept
4	notUnderstood	Accept
5	notUnderstood	Accept
6	notUnderstood	Accept
7	notUnderstood	Reject
8	notUnderstood	Reject
9	notUnderstood	Accept
10	notUnderstood	notUnderstood
11	notUnderstood	Accept
12	notUnderstood	Accept
13	notUnderstood	Accept
14	notUnderstood	Accept
15	notUnderstood	Accept
16	notUnderstood	Accept
17	notUnderstood	Reject
18	notUnderstood	Accept
19	notUnderstood	Accept
20	notUnderstood	Accept
21	notUnderstood	Reject
22	notUnderstood	Accept
23	notUnderstood	Accept
24	notUnderstood	notUnderstood
25	notUnderstood	Reject
26	notUnderstood	Reject
27	notUnderstood	Reject
28	notUnderstood	Accept
29	notUnderstood	Reject
30	notUnderstood	notUnderstood

Figure 5 shows the graphical results of this experiment. The first set of bars represents results of communications without translations, and the second set of bars represent results of communications invoking the translator.

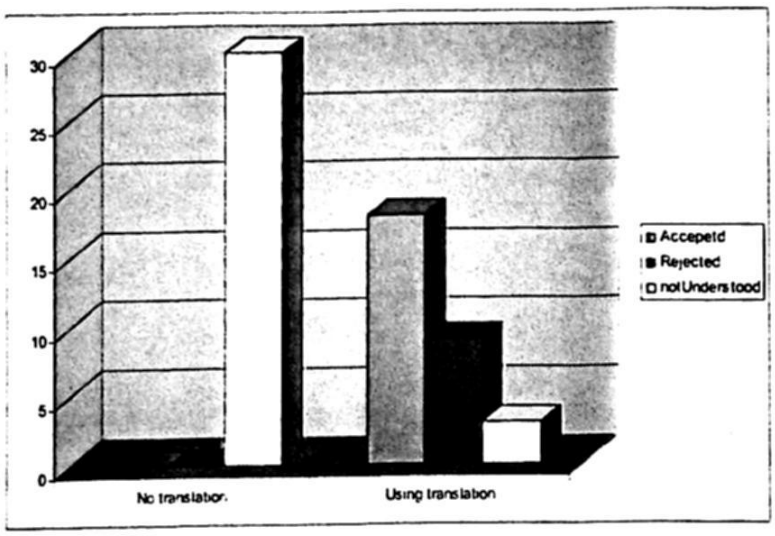


Fig. 5. Experimental results

7 Related Work

There is currently a research effort dealing with the problem of modeling Web service processes; in particular, some of them have proposed conformance and interoperability measures between two composed Web service processes. In [8] authors present the behavior equivalence concept, which states that two executable processes are equivalent if they can be transformed by a set of transformation rules.

Baldoni et al. [9], present a calculus to allow the automatic construction and update of compatible services. They define formally the compatibility, interoperability, conformance and substitutability between Web services concepts. They present a set of derivation rules to verify the conformance level and to construct new conformant and interoperable services.

Van der Aalst et al. [10] present the concept of process equivalence between two process models based on their observed behavior. They provide a measure of equivalence with a Lumber ranking from 0 to 1. This represents a more precise measure.

In this work we are dealing with agent communication protocols, which have similar functionality compared to composite Web service processes. However, communication protocols follow specific interaction rules; for example, an agent following a conversation protocol never issues a completer communicative act before receiving a starter communication. While Web service processes may follow different flows depending on the problem or objective they are following. Therefore, even

though we consider important these works, they are not directly transferable to our domain.

8 Conclusions

In this paper we have presented how to compute pragmatic similarity in DIS. Our approach is novel, because it presents an algorithm based on the analysis of FSM transition functions, which in turn are capable of being implemented and compared to obtain pragmatic relations and pragmatic similarity measures. In contrast to the semantic approach, which has proven good results for data, vocabularies and information sources, we propose a semi-automatic approach for comparing software pragmatics. The result of this analysis helps the developer to measure pragmatic similarity and identify pragmatic relations. The set of defined relations were implemented in an ontology-based translator which showed better results in communication environments when the translator was invoked.

The pragmatic similarity measure will help developers to decide and propose a good solution, for example, translation, learning or reprogramming all software entities to be fully interoperable among them.

As the Semantic Web has been evolving we are facing new requirements, such as discovering semantic and pragmatic relations from heterogeneous sources. This is a promising research area, because nowadays there is a tremendous amount of legacy software which in turn will require to be incorporated transparently giving the idea of an integral solution independently of the inherent heterogeneity inside their logics or protocols.

FSM is a good formalism for representing communication scenarios between software entities, in particular they are suitable to compare interaction protocols and identify the state of a conversation in a DIS environment.

Finally, we are sure that our pragmatic computing approach can be applied in other application areas such as Web service discovery, process engineering, and program comparison or application integration, which have in common that they provide functionality information similar to protocols in communication scenarios.

References

1. Petri, Carl A. Kommunikation mit Automaten. Ph. D. Thesis. University of Bonn, 1962.
2. R. Scott Cost, Ye Chen, Tim Finin, Yannis Labrou, and You Peng. Modeling agent conversations with colored petri nets. In working notes of the Workshop on Specifying and Implementing Conversation Policies, pp. 59-66, Seattle, Washington, 1999.
3. R. Milner. Communicating and Mobile Systems: The Pi-Calculus, Cambridge Univ. Press, 1999.
4. Odell J. Van Dyke Parunak, H., and Buer, B. Representing agent interaction protocol UML. In the first international workshop, AOSE 2000 on Agent Oriented Soft Engineering, pp. 121-140, Springer Verlag, Berlin, 2001.BPEL, Business process Execution Language for Web Services. Version 1.1, May, 2003.

5. DAML-S: Web service description for the semantic Web. In Proceedings of the first International Web Conference, ISWC, 2002.
6. J.E. Hopcroft, R. Motwani, and J. D. Ullman. Introduction to Automata Theory, Languages and Computation, Addison Wesley, 2001.
7. Müller, H. J., 1996. Negotiation Principles, Foundations of Distributed Artificial Intelligence. *In G.M.P. O'Hare, and N.R. Jennings*, New York: John Wiley & Sons.
8. Köning D., Lohmann, N., Moser, S., Stahl, C., Wolf, K. Extending the Compatibility Notion for Abstract WS-BPEL Processes. En las memorias de la 17 Conferencia Internacional de la World Wide Web, (WWW 2008), Beijing, China, 2008.
9. Baldoni, M., Baroglio, C.,Patti, V., Chopra, A., Desai, N., Munindar, S. Calculi for Service Conformance and Interoperability. Draft paper, 2008.
10. Van der Aalst, W.M.P., de Medeiros, A.K. Alves, Weijters, A.J.M.M., Process Equivalence: Comparing Two Process Models Based on Observed Bahvior. LNCS, Business Process Management, Vol. 4102, 2006.

Effects of Cheaters on Altruistic Signaling

Grecia C. Lapizco-Encinas

Department of Computer Science
University of Maryland
College Park, MD 20742

Abstract. This report explores some conditions that influence the evolution of cheating individuals in an altruistic signaling society. It explores specifically three kinds of signaling: advertising of food, warning of nearby predators and begging for food. The experimental method considers food distribution, predator density and population size as factors that could influence the evolution of cheating behaviors on these signals. The results of these simulations show that predator density seems to be the most influential factor.

1 Introduction

Altruism is defined as a situation in which an individual acts to promote or enhance the fitness of other members of a group while at the same time reducing its own fitness [1]. When all individuals of the group engage (by providing help and/or benefiting from it) on these behaviors we have an altruistic society. This work explores the effects of cheaters (individuals who perform deceiving activities) on an altruistic society.

The first part of this work explores an altruistic signaling society. In an altruistic signaling society individuals emit signals to advertise the presence of food, or to warn about nearby predators, and there can be two main kinds of deceiving activities: one is to emit a false signal (e.g. a bluff), and the second one is to withhold a signal (e.g. an alarm call). This work focus on the latter. There are biological studies [2] that suggest that withholding information (e.g. food calls) can be beneficial to the individual, and can result in a higher fitness. But there is no evidence about which factors constrain or expedite this behavior. We explore some conditions that influence the evolution of cheating behaviors (withholding information) on alarm and food calls.

The second part of this work extends the idea of taking advantage of altruism one step further, in this part we explore the effect of individuals who take advantage not from the emission of signals but from even more explicit altruistic behaviors. An example of these explicit altruistic behaviors is reported in [3], where vampire bats share food (at a cost) with recipient bats with no directly apparent benefits. This idea inspires the next part of this study. In here we introduce a beg-for-food call, and the cheating behavior consists of taking advantage of this signal (using it) but not responding to it. Conditions that can influence the evolution of this cheating behavior are explored.

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 171-182

Received 01/07/08
Accepted 01/08/08
Final version 13/09/08

The final sections consist of the experimental results and discussion of these results.

2 Methods

For this work, a predator-prey simulation framework for evolving a shared communication system was adopted and extended. In here we just provide a brief summary of this framework, a detailed description can be found in [4].

The world consists in a two-dimensional grid with three kinds of objects: preys, predators and food.

- Food: static food (e.g. plants) is distributed in a pre-determined number of sites in the world.
- Predators: agents that move around the world looking for preys. they are not able to communicate among them, but they are able to hear prey's communication.
- Preys: agents that move around the world, looking for food and avoiding predators. They are able to emit/hear signals.

2.1 Preys

Preys are the object of interest of the present work; these are the initially altruistic individuals on which the evolution of deceiving behaviors will be studied.

Signals Preys are able to emit two distinctive signals:

- Food call: when they arrive to a food site, they can advertise this food site to other preys by emitting this signal.
- Alarm call: when they see a predator, they can advertise its presence to warn nearby preys by emitting this signal.

States Preys have six possible behavioral states; these behavioral states are activated according to the world perceptions.

- Wander: the prey is just moving around.
- Search: the prey has heard a food call, so it is trying to find it.
- Forage: the prey saw a food site, so it is moving toward it.
- Consume: the prey is at a food site, and consuming the food.
- Avoid: the prey has heard a alarm call, so it is moving away from the signal source.
- Flee: the prey saw a predator, so it is moving away from it.

Kind The kind of prey is defined by its communicative abilities. These abilities of communicating signals are determined by the genome of the prey, and do not change during its lifetime. The genome of the prey is represented as a two-bit string which is interpreted as follows:

- 00 indicates a NC prey (not able to communicate about food or predators).
- 10 indicates a FO prey (able to communicate only about food).
- 01 indicates a OP prey (able to communicate only about predators).
- 11 indicates a FP prey (able to communicate about food and predators).

2.2 Evolving cheating in food/alarm calls

The existent framework was extended with the purpose of studying the influence of multiple factors in the evolution of cheating (deceiving) behaviors in the advertising of food and the warning of predator call.

Added Cheater Gene We modify the genome to include a new gene. This is called the cheater gene. Preys that have this gene activated do not contribute to communication but take advantage of it (if possible). In this case we have two kinds of individual:

- Altruist: advertises food and warn about predators (if able to communicate about them).
- Cheater: hears the advertisements and warnings, and takes advantage of them, but never advertise or warn (similar to altruistic individuals, the ability to hear the signals is determined by the genome).

This gene has no influence with the other two genes described in the previous section; this means that the prey can have this gene activated regardless of the value of the other two, which can give as a result a NC cheater (which will not take advantage of communication since it is not able to communicate in the first place).

Modified behaviors The behaviors of the prey were modified accordingly to consider the new genome by adding the constraint of advertising food or warn about predators if and only if the cheater gene is not activated.

2.3 Evolving cheating in beg-for-food calls

The original framework was extended to consider a new type of signal, a *beg-for-food call* When an altruistic individual hears this signal, it approaches the emitter of the signal and donates some of its own food to him. A cheating individual is an individual who can emit this signal when in need, but ignores the signal when it hears it. In this task we do not consider the food or alarm calls to limit our focus to the relevant information.

Added signal Beg-for-food call: This call is emitted to elicit a donate behavior from fellow preys.

Added behaviors

- Begging: when the prey founds itself with an internal storage of food that is below of a predetermined critical amount then it changes its state to Begging and emits a beg-for food-call.
- Approach: when a prey hears a *beg-for-food call*, the it tries to approach to the source of the message to be able to donate him food.
- Donate: when the prey founds the starving fellow, it donates him an amount of its internal storage of food.

Some constraints apply to these behaviors state transitions:

- The prey can only change to a Begging state if it is currently on the wander state. If the prey is on Forage or Consume state it has more opportunity to get food directly from the food site than from begging, and if the prey is on Avoid or Flee, then these behavioral states get precedence (what is the point of getting food if you are going to be eaten by a predator?).
- The prey can only try to approach to source of the *beg-for-food call* if its internal storage of food is higher than a pre-determined amount. (what is the point of approaching to a starving fellow if you are starving yourself or if donating food can put you in a dangerous situation?). So in a sense, this limits the altruism, the prey can only be altruistic if it does not jeopardize its situation.

In this case we have two kinds of individual:

- Altruistic individual: Donates food as a result of *beg-for-food call.*
- Cheater individual: Ignores the *beg-for-food calls* it hears, but it could emit this signal if in need.

3 Experiments

Since an approach of systematically varying parameters is not feasible (because of computing-time constrains) we took the parameters that were shown to evolve the shared communication system (as described by the study in [5]) as our reference frame. We considered that parameters that did not evolve the communication system in the first case, were unlikely to have relevance on this study.

Unless explicitly noted otherwise, all the simulations reported below where done using the following parameters: the environment was 60x60 in size, predators and preys had a vision of three cells, could hear signals at a distance of six cells, and could move in any direction. Preys had a maximum storage capacity of 30 food units and a newborn had an initial storage of 25 units. Spatial constrained selection and placement, and tournament size of 2 was used. Simulations were run under 100,000 iterations.

The results reported below for each simulation are the averaged values over 10 different runs of the simulation.

4 Results

4.1 Evolving withholding of food/alarm calls

Baseline simulation For the first part of this work, the focus is in the conditions that influence the evolution of cheating in food/alarm calls. To have a baseline to which compare the results of varying several parameters, the first simulation was made set with the default parameters, no food and no predators. The initial population was 200 FO preys, and mutation of 0.003 was set for mutating the cheater-gene, other mutations were not allowed. So, our population consists initially of altruistic FO preys, and could evolve to cheater FO preys. The long-term expected fraction of the population in the baseline simulation that is altruistic or cheater is about 0.5, because of the fixed mutation rate. As expected, in Figure 1 we can see this result. This baseline simulation is also valid when considering an initial population of 200 altruistic OP preys, which can evolve to cheater OP preys.

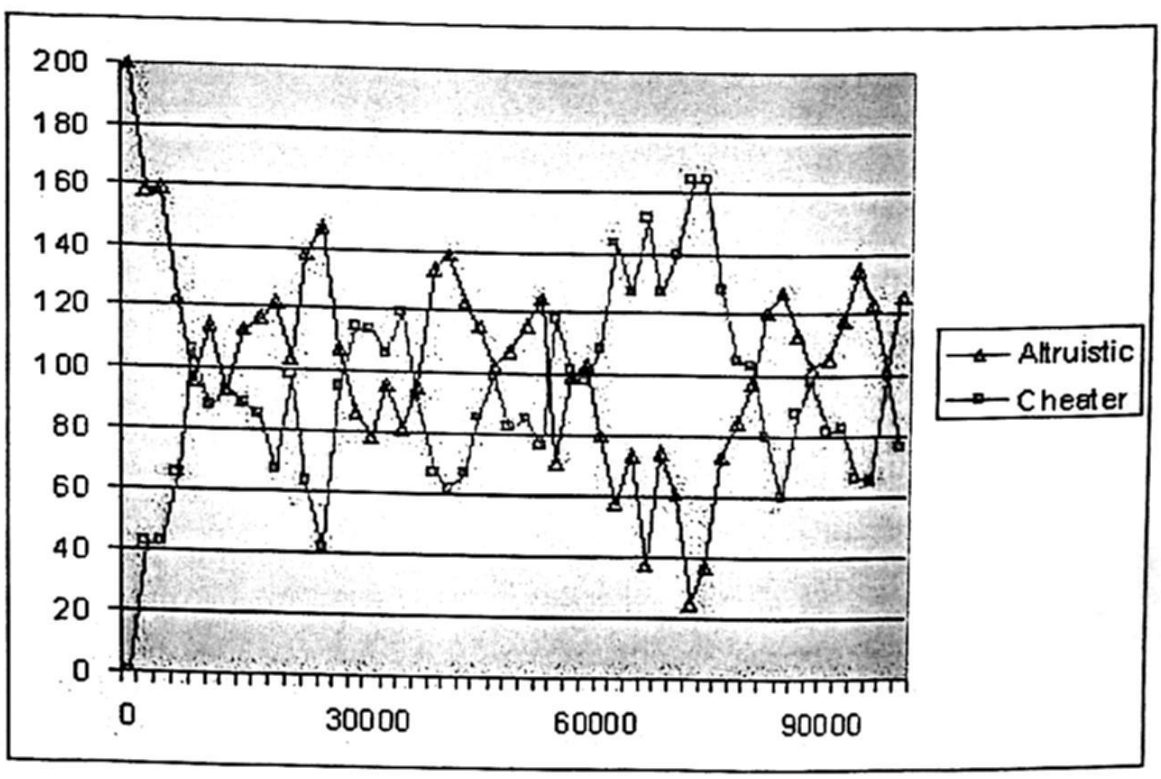


Fig. 1. Baseline simulation with no food / no predators. (Time vs Number of Agents)

Evolution of cheating in food calls As mentioned before, a cheating individual will hear and take advantage of food calls, but will withhold advertising of food. Two factors that could influence the evolution of cheating individuals for the advertising of food are studied: the distribution of food and the population size.

Varying food distribution In this experiment the amount of food was fixed (1600 units) and the number of food sites was varied. A hypothesis for this experiment is that cheaters would seem more possible to evolve in situations where the resources are scarcer, and advertising of a scarce resource would limit its own fitness due to competition. In this case, cheaters would be more likely to evolve in environments where the amount of the food site is smaller. The results obtained in this simulation do not agree with the hypothesis. In the simulations cheaters were unlike to evolve independently of the distribution of food. An explanation could be that, even though the distribution of food varies, this does not put enough pressure over resource competition. Results are shown in Figure 2.

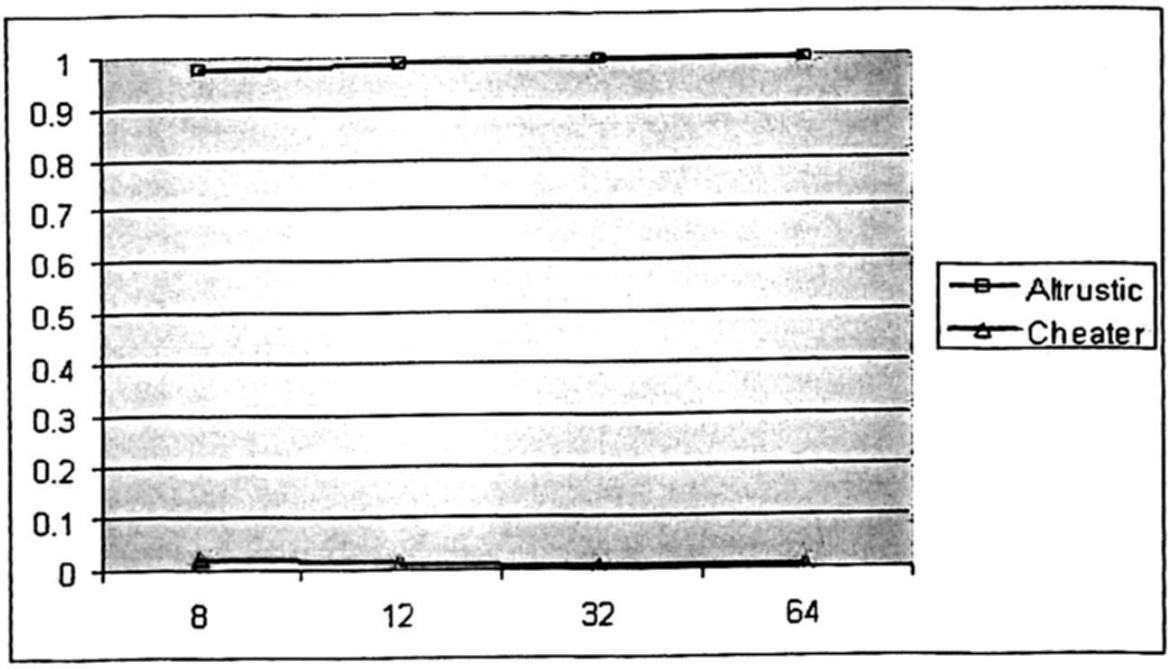


Fig. 2. Evolution of cheating in food calls. Fraction of population of each type averaged in the last 30,000 iterations. Axis X shows the number of food sites.

Varying population size In this experiment the population size varies from 75 to 200 agents. A hypothesis for this experiment is that cheaters would seem more possible to evolve in situations where there is more resource competition. In this case, cheaters would be more likely to evolve in environments where the population size is bigger. Again, the results obtained in this simulation do not agree with the hypothesis. In the simulations cheaters were unlike to evolve independently of the population size. Results are shown in Figure 3.

Evolution of cheating in alarm calls A cheating individual is defined here as an individual that is able to hear and take advantage of alarm calls, but withhold warning. Two factors that could influence the evolution of cheating individuals

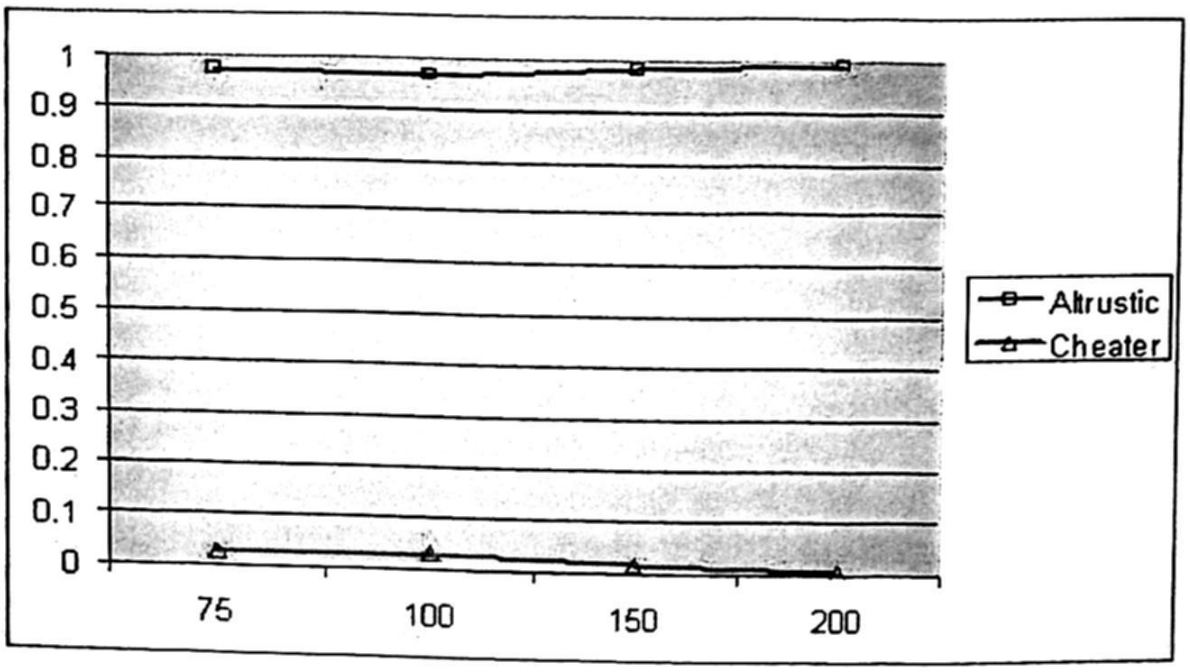


Fig. 3. Evolution of cheating in food calls. Fraction of population of each type averaged in the last 30,000 iterations. Axis X shows the population size.

in the warning of nearby predators are studied: the density of predators and the population size.

Varying predator density In this experiment the predator density varies from 4 to 24 agents. Warning of predators can be costly to the sender, since predators are able to hear these messages and locate a prey that they would have missed before. So, a hypothesis for this experiment is that cheaters would seem more possible to evolve in situations where there the risk of predation is higher. The results of the simulation agree with this hypothesis. We can observe that cheaters get a higher fraction of the population when the number of predators increases. The maximum percentage of the population is 60% , and then it starts to drop. An explanation of these could be the cheaters can not dominate the population because they have a parasite-relationship with altruistic agents (they need to be warned about nearby predators). Results are shown in Figure 4.

Varying population size In this experiment the population size was varies from 75 to 200 agents. A hypothesis for this experiment is that cheaters would seem more possible to evolve in situations where there are more warning agents to take advantage from. So, in this case, cheaters would be more likely to evolve in environments where the population size is bigger. The results obtained in this simulation do not agree with the hypothesis. Increasing the population above 100 seems to have a negative effect on cheaters. Results are shown in Figure 5.

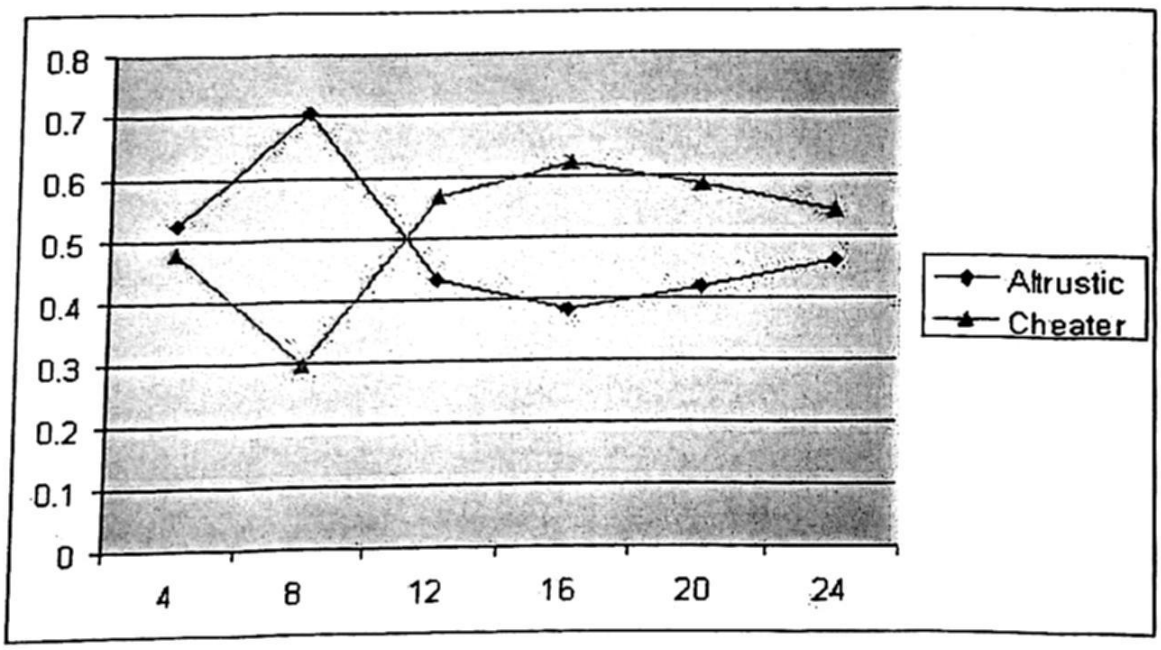


Fig. 4. Evolution of cheating in alarm calls. Fraction of population of each type averaged in the last 30,000 iterations. Axis X shows the number of predators.

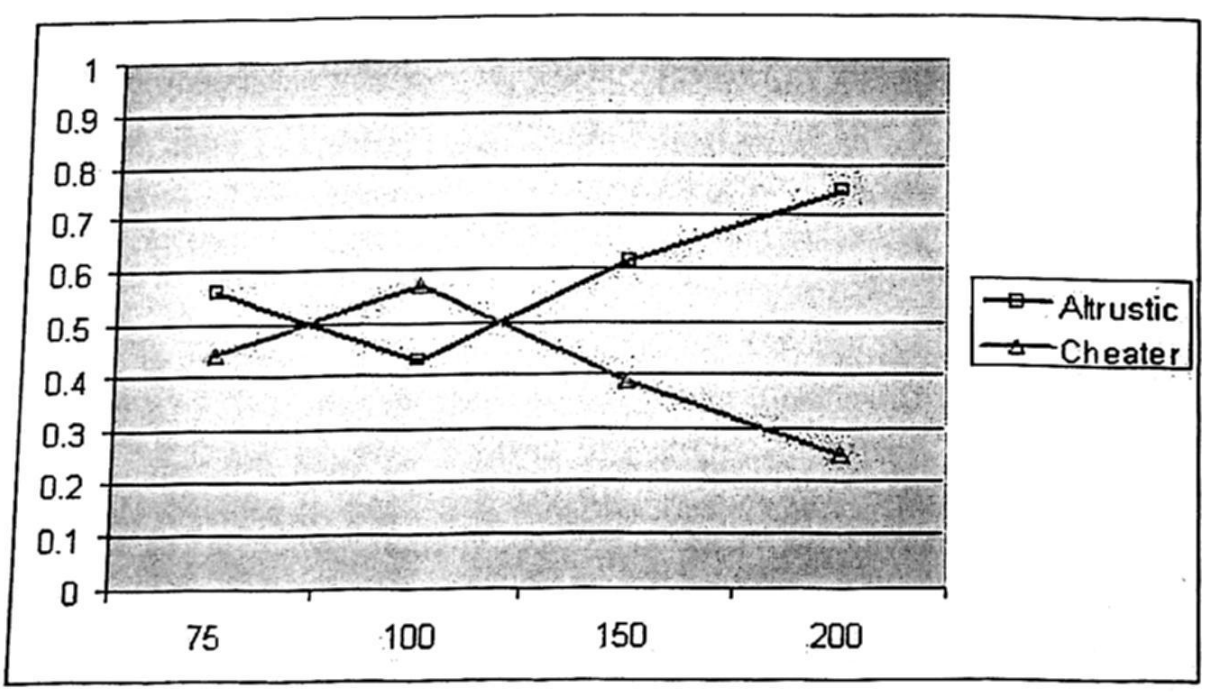


Fig. 5. Evolution of cheating in alarm calls. Fraction of population of each type averaged in the last 30,000 iterations. Axis X shows the population size.

4.2 Evolving cheating in beg-for-food calls

Altruistic preys respond to the *beg-for-food call* by approaching the message source and donating food, while cheaters ignore these calls (but they use them if they need them). Two factors that could be of influence in the evolution of cheating individuals in this donating-food behavior are explored: food distribution and predator density.

Baseline simulation In here the simulation was set with the default parameters, no food and no predators. We are interested in conditions that influence the evolution of cheating in *beg-for-food calls*. In order to explore these we started with a population of 200 altruistic preys, and set a mutation rate of 0.003 for mutating the cheater-gene, other mutations were not allowed. In here we are not considering food or alarm calls. The long-term expected fraction of the population in the baseline simulation that is altruistic or cheater is about 0.5, because of the fixed mutation rate. In Figure 6 we see a mixed result; the altruistic and cheaters agents fluctuate and neither of them can dominate the population and reach an equilibrium point.

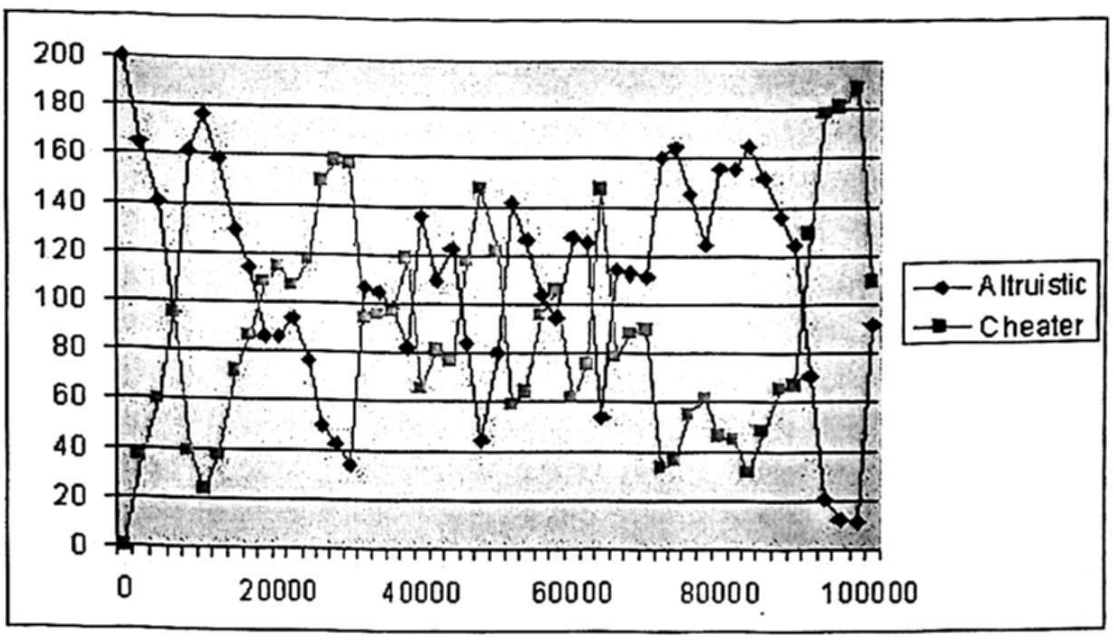


Fig. 6. Baseline simulation with no food / no predators. (Time vs Number of Agents)

Varying food distribution In this experiment the amount of food was fixed (1600 units) and the number of food sites was varied. A hypothesis for this experiment is that cheaters would seem more possible to evolve in situations where the resources are distributed through the environment, so they have a

bigger opportunity of being approached for an altruistic agent that has access to food (otherwise even if there are altruistic agents, they will not be able to donate food). The results obtained in this simulation agree with the hypothesis. In the simulations the percentage of cheaters increases when the number of food sites is higher than 16. See Figure 7.

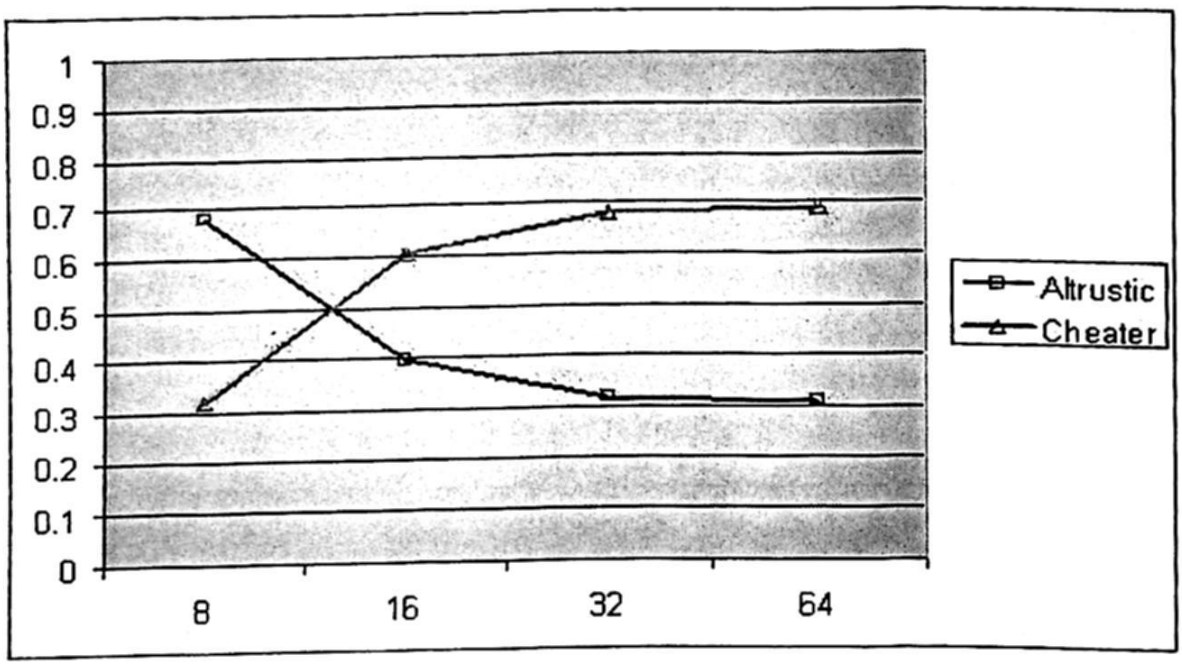


Fig. 7. Evolving cheating in beg-for-food calls. Fraction of population of each type averaged in the last 30,000 iterations. Axis X shows the number of food sites.

Varying predator density In this experiment the predator density varies from 4 to 24 agents. Begging for food can attract predators, which put in danger not only the starving agent, but also any altruistic agent that approaches to the message source to donate food. So, the hypothesis here would be that cheaters would be likely to evolve when the number of predators increases. In here the results agree with the hypothesis, we can see an increase of the cheater population as the number of predators increase, but when the predators reaches 20, the number of cheater individuals starts decreasing. An explanation for that could be that they reach the critical point at 16 predators, when the percentage of cheating individuals is higher than the altruistic individuals reaching a 60% percentage, and cheater individuals can not dominate the population since they need the altruistic individuals to take advantage of.

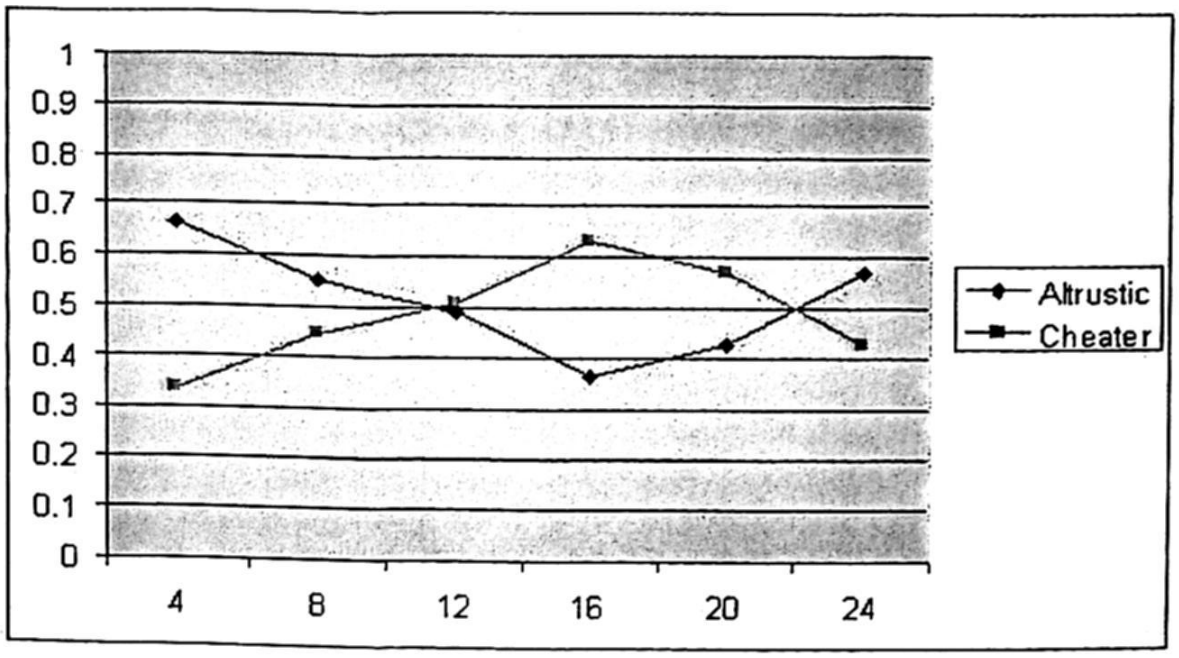


Fig. 8. Evolving cheating in beg-for-food calls. Fraction of population of each type averaged in the last 30,000 iterations. Axis X shows the numbers of predators.

5 Discussion

The results of the simulations were mixed, since it did not always agreed with the hypothesis. Nevertheless, this study gives an insight of possible conditions affecting the evolution of cheating behaviors in initially altruistic agents.

In the case of the evolution of cheating behaviors for the food call it is still unclear which factors can truly influence its evolution, since neither the food distribution nor the population size seem to have affected it.

In both alarm calls and beg-for-food calls, predator density has a strong influence over the evolution of cheating individuals. In both cases the number of cheating individuals reach maximum percentage at 16 predators when they have approximately 60% of the population, and the y start decreasing. This decrease could be due to the fact that they have a parasite-relationship with the altruistic individuals (they need altruistic individuals who would emit the alarm calls or who will donate food when they hear the *beg-for-food call* so they can not take over the entire population.

Another factor that influence the evolution of cheating behavior in the *beg-for-food call* is the food distribution. If the food is distributed throughout the environment, then the cheating individual has a bigger opportunity to be approached by an altruistic individual who has access to food (remember that altruistic individuals donate food only if they have enough reserves).

6 Future work

One extension is to enable preys to discriminate between them, so individuals (and not member of an anonymous society) can interact with each other. As pointed out in [6] agents should have histories; they should perceive and interpret the world in terms of their own experiences. This ability to discriminate between individuals would enable an arms-race between cheaters and discriminating receivers, so donors can recognize and expel cheaters (as mentioned in [7]).

Acknowledgment

The author would like to thank James Reggia for reviewing this manuscript.

References

1. Lincoln, R., Boxshall, G., Clark, P.: A Dictionary of Ecology, Evolution and Systematics. Cambridge University Press, Cambridge, UK (1998)
2. Lachmann, M., Bergstrom, C.T.: Signalling among relatives - ii. beyond the tower of babel. Theoretical Population Biology **54** (October 1998) 146–160(15)
3. Wilkinson, G.S.: Food sharing in vampire bats. Scientific American **262**(2) (February 1990) 76–82
4. Reggia, J.A., Schultz, R., Uriagereka, J., Wilkinson, J.: A simulation environment for evolving multiagent communication. UMIACS Technical Reports CS-TR-4182, UMIACS, University of Maryland, College Park, Maryland 20742 (2000)
5. Reggia, J.A., Schulz, R., Wilkinson, G.S., Uriagereka, J.: Conditions enabling the evolution of inter-agent signaling in an artificial world. Artif. Life **7**(1) (2000) 3–32
6. Dautenhahn, K.: The art of designing socially intelligent agents. Applied Artificial Intelligence Journal, Special Issue on Socially Intelligent Agents **1**(7) (1998) 573–617
7. Hauser, M.D.: Costs of deception: Cheaters are punished in rhesus monkeys (macaca mulatta). Proceedings of the National Academy of Sciences **89**(24) (December 1992) 12137–12139

Natural Language Processing and Information Retrieval

Automatic Generation of Document Summaries in Spanish Language

Rodolfo Rodríguez, Darnes Vilariño, Beatriz Beltrán, and Mireya Tovar

Benemérita Universidad Autónoma de Puebla, Puebla, Puebla, México.
Facultad Ciencias de la Computación
Puebla, Puebla, 72570 México, www.cs.buap.mx

Abstract. Without a doubt, Internet has become the biggest source of available information. Nowadays one of the biggest sources of information is the WEB, the information grows in a chaotic and not controlled way, originating certain limitations for its handling, organization and recovery. The Spanish language is very complicated for its study, and exist few tools that make an analysis to obtain an abstract and so the few available versions unfortunately are not freeware. In the present work we developed a learning technique and a set of metrics that applied at the original document return us the most representative sentences of the document to construct the automatic abstract by extraction, including a study of anaphoras.

Keywords. Retrieval, information, abstract, automatic, extraction, sentence, learning.

1 Introduction

A summary is to reduce to brief terms the essential of a document. The use of the summary of a document helps to reduce the storage space, facilitates the access to the most important information and accelerates the time of reading to locate the required information of a particular document, in at a certain moment [11].

Nowadays it is necessary to count on suitable computer science tools for the recovery of the important information in an efficient and fast way. The manual techniques have demonstrated to be inefficient, because basically they consist in manual elaboration. This work is very expensive in time, in addition it suffers from abundant inconsistencies, like orthography and coherence [3,4]. Now identified the problem, the challenge of the investigators is achieve that the computers be an efficient tool in the process of information retrieval.

By automatic generation of text summaries is understood the process by which the originating substantial information of a source (or several) is identified to produce a brief version destined to a particular user (or user group) and a task (or tasks) [22].

A Phrase and an sentence are different in which the first lacks of a express preaching (does not say anything of a subject, type: Buenos Días) and in the second

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 185-197

Received 31/05/08
Accepted 04/08/08
Final version 10/09/08

there is explicit preaching (one affirms or it denies, something is said from somebody or something: Buenos Días nos de Dios (Subject: Dios; Predicate: de).

Sciences and the philosophy tend to establish truth, but the concept of true - or false - is not possible apply it to smaller units of the language than the sentences: nor to the sound. Therefore the minimum unit of the susceptible language of true is the proposal, since all proposal or sentence implies a judgment (establishment of the truth or false principle of the knowledge) [21].

There are a lot of articles that propose methods to extract the most important orations of a determined document, using the frequency of the words within the same document, the frequency of words within the title, considering most important the first five and last five sentences of a document, as well as other statistical parameters [4,7]. Nevertheless the great majority of these articles does not work over the Spanish language, the most important results are on the English language, that is substantially different with the Spanish language.

In the present article we obtained automatic summaries by extraction in the Spanish language are, where some classic techniques like eliminate stopwords and special symbols are used, stemm the words, make a general study of the most well-known abbreviations, as well as a brief study of proper names in a pre processed, also we used some other metrics for the selection of the important sentences, as they are: eliminate very short sentences, important words, comparison with the title, etc., in addition we developed a technique that includes intelligence (learning) for the important information retrieval of each document and when the extract is obtained, a study of the most well know anaphora [1]is applied with the intention of have more coherency in the obtained extract. All used software has been made by our group of investigation and the techniques used for the English language have been adapted to work in the Spanish language.

The present article is divided in different sections. In the second section, we describe a general way all the different metrics that were used to evaluate each sentence in a document. In the third section, we present the used algorithm to obtain the automatic summary by extraction, where the system decides which metrics will be used to evaluate the document and obtain the summary. The fourth section is dedicated to explain how works the algorithm that include anaphoras. The fifht section of the article is all about to evaluate the system, where the obtained results are show in the first table. This table shows the similarity degree between the obtained summary by the system without anaphoras against the obtained summary for one of the three experts, and also shows the degree of similarity between the summary of the system without anaphoras and with anaphoras, and in the second table we present the obtained results of the threshold for each metric (that help us to decide which metric will be used or not to evaluate a document) after analyzed all the documents of our corpus. The last section are the conlusions.

---

[1] Anaphora. Word that is used to reference other words that were used before in the document.

2 Score of Sentences of the Document

Supported in the analysis made at the different reviewed articles, it had been collected and adapted to our problem a set of metric that they search to give certain score at each sentence of the document, and extract the five most important sentences (the five sentences with bigger score) to conform an automatic summary by extraction.

2.1 Selection of Important Sentences using the Transition Point (TP)

It is defined the TP like the frequency of a term of the text that divides in two the terms of a vocabulary (in terms of high and low frequency), that is to say, the terms nearest the TP as much of high and low frequency are going to determine of what it is all about the document [12].

From the transition point a determined threshold can be taken, works generally with 25% or with 40%, in this work, words above 25% of TP and under 25% the TP of the threshold are selected, with these terms is constructed a paragraph, it is denominated Virtual Paragraph (VP).

With the aid of the VP, it is given a score at each sentence according to its degree of similarity, applying the similarity formula of Jaccard (1) at each sentence (O_i) of the document, and the obtained VP.

$$Sim_i(PV, O_i) = \left|\frac{PV \cap O_i}{PV \cup O_i}\right| \quad (1)$$

2.2 Sentences Length

This metric is sustained in the idea that the sentences of very small length are not important, because generally they are not in the summaries generated by the humans, by this, it is assign a negative score to them that affects to be part of the automatic summary [10]. In the present work it is considered to be a small sentence, if it consists of less than 5 words, after doing corresponding pre processing.

2.3 Title Compare

The title of a document usually is strongly related to its content and often constitutes the best summary of itself [11].

It is constructed a VP in the same way that was in the TP, but in this case this VP is formed by the terms that conform the title of the document [15]. The VP created from the title words is compare with each one of the sentences of the document, applying the similarity of Jaccard (1), then it is give a weight to each sentence of the

original document, and those with greater weight are candidates to be part of the automatic summary by extraction.

2.4 Centroid

The centroid value is a measurement that indicates how frequent are the words of a document [10, 11, 13, 15]. Importance is attributed to this metric, because using it, is possible know how near are the sentences of a document to the main subject. The centroid is a very great vector; each component of the vector is one of the words that appear in the document. Be D the document and $/D/$ the number of sentences in the document, it is described by:

$$Centroide(D) = (v_{w0}, v_{w1}, \dots, v_{wn}) \tag{2}$$

Where:

$$v_{wi} = \frac{TF(w_i) * IDF(w_i)}{|D|} \tag{3}$$

$TF(w_i)$ Is define like the occurrence number of w_i in the document D.

$IDF(w_i)$ Is known like inverse frequency, and this is calculated by the following formula:

$$IDF(w_i) = \log\left(\frac{|D|}{Sentences\ in\ D\ that\ include\ w_i}\right) \tag{4}$$

IDF indicates how "weird" is the word w_i in the document D.

2.5 Position

This metric is considered important due to the fact that in different types of documents, the most important information occurs in the first sentences, is to say the main idea occurs in the first paragraphs and in the rest of the document this idea simply is developed [10, 11, 13, 15]. For each sentence O_i in the document D, the value of the position of each sentence it is calculate by:

$$P_i = \left(\frac{(|D| - i + 1)}{|D|}\right) * C_{max} \tag{5}$$

Where C_{max} is the maximum centroid value obtained by the previous formula (2).

This metric is combined with the previous one, thus finally the sentence O_i obtain a higher score in function of the position and the obtained centroid value, the sentences importance diminishes with the position.

In addition to the previous score, a greater weight is offers to the first 5 and last 5 sentences of the document.

2.6 Proper Names

For some authors the presence of proper names is important [11, 13, 15]. Is considered that the sentences that have proper names are more important, since it can refer a particular person or place, and this is very common in bibliographies. In this work it has been considered like proper name those words that begin with a capital letter and which are not stopwords.

2.7 Important Words

The important Words metric [10, 11, 14, 17, 18] are words and expressions that do not have relation with the central subject of the document, but they indicate that the sentence can contain important information and must be part of the summary, words like: "importante", "esencial", "en conclusion", "resumen", "fundamental", etc. [11]. As we would desire the most possible generic system, a set of certain words is used, that reveal importance independently of the type of document.

The implemented metric looks for those important words within the text to summarize, granting additional scores to the sentences that contain them.

2.8 Sentences Compare

With this metric we try to find the sentences with greater degree of similarity. Each sentence of the document is compared with all the rest, applying the similarity of Jaccard (1); to define that two sentences are similar, is taken a threshold of 0.5, this is a good measurement (if it be 1 would be the same statement the one that is being compared) for the comparison of two sentences, if the similarity is 0 means that no one of the words that contains the sentence are similar to which is compared. The sentences with a greater degree of similarity are given a bigger score to conform the automatic summary by extraction.

2.9 Bayes Probability

For each sentence s we are going to calculate the probability if it will be included in a summary S given k characteristic, $F_j; j = 1, \ldots, k$, that can be expressed using the Bayes Rule and assuming statistical independence of the metrics, we have:

$$P(s \in S|F_1, F_2, \ldots, F_K) = \frac{\prod_{j=1}^{k} P(F_j|s \in S)P(s \in S)}{\prod_{j=1}^{k} P(F_j)} \quad (6)$$

Where:

- $P(s \in S|F_1, F_2, \ldots, F_K)$ It is the probability that the sentence s is contained in the summary given by the metrics $F_j; j = 1, \ldots, k$.
- $P(s \in S)$ It is the probability that the sentence s is selected between all the others of the document.
- $P(F_1, F_2, \ldots, F_K|s \in S)$ It is a probability that can be considered counting the number of metrics that were used of the total in the sentence s.
- $P(F_j)$ This probability is calculated of the average of the number of times that is used a metric in the document (just it is taken once like maximum account by sentence).

3 Algorithm to Extract Important Sentences

We made and developed an algorithm in order to decide if the j metric will be used or it does not to give an additional score to a sentence s, taking the idea from the Bayes probability, so $P(F_j)$ must be greater to a given threshold to be used; This threshold will be calculated applying the following algorithm.:

1. Fort the first analyzed document we are going to consider that there is the same probability that a metric j is used or not, reason why the initial threshold will be equal to 0.5, i.e. $Threshold_1 = 0.5$.
2. The threshold is calculated for the following documents in this way:

$$Threshold_j = \frac{Threshold_1 + \sum_{l=1}^{k} P_l(F_j)}{k+1} \quad (7)$$

Where k is the number of analyzed documents of the experiment, and i is the number of the analyzed document $i: 1 \ldots k$.

When concluding the experiment of analyzed documents, the threshold will be fixed in the obtained value. In this way if the average of times that was used a metric in a document is smaller than the threshold obtained from that metric, then it will not be used to evaluate the sentences, so the metric will not influence in the decision of the sentences to be included or not in the summary given by the system.

Once all the metrics finished giving the corresponding scores to each one of the sentences, the scores are added and mediated by sentences. The 5 sentences with greater average score are selected and ordered by appearance to construct the automatic summary by extraction.

4 Algorithm to Obtain Anaphoras

Once we have the automatic summary by extraction, we made a study of anaphoras, the main objective is give coherence to the extract made by the system, we did this in the following way:

1. Search for anaphoras in the sentences of the extract obtained by the system.
2. If the system finds a sentence that contains an anaphora, the previous sentence of which was the anaphora will be included in the new extract, this new sentence is taken from the original document.
    a. If the new sentence is already in the original extract do nothing.
3. Include in the new extract the sentence in which was found the anaphora.
4. Do the same for each one of the sentences of the original extract obtained by the system.

There are not methods of resolution of anaphora for the Spanish language that not use much semantic information. In this article one of our main objectives were implement a tool for the resolution of anaphoras for the Spanish language using limited knowledge, it means, not using semantic information.

5 Evaluation

Until now to analyze the behavior of the method of obtaining automatic summaries by extraction, was decided work with a set of 100 heterogeneous documents in the Spanish language, each one with its respective summary of one of a set of three experts.

An automatic pre processing was made at each document, first the stopwords were eliminated, then the document was stemmed and a treatment of abbreviations was made (that consists of changing the most common abbreviations by the complete word), to conform the initial groups.

Between the 100 documents that conform our CORPUS, there are very small documents and greater others, to all documents the formula of similarity of Jaccard (1) was applied to them between the summary offered by the Expert, against the summary obtained by the system, and the summary obtained by the system without anaphoras and with anaphoras, with the objective to measure the proportion of similar sentences, the results are presented in Table 1.

In the Table 1 can be appreciate that the similarity degree given by the extract of the system without anaphoras and the extract by one of the three experts is approximately of a 60%, and the similarity degree between the extract of the system without anaphoras and with anaphoras is approximately of a 98%, this means that many of the sentences of the extract of the system without anaphoras did not have anaphoras or its anaphora was included already in the initial extract, reason why when the additional sentences were included in the new extract, the similarity degree decrement between the extracts.

Table 1. Obtained similarity results

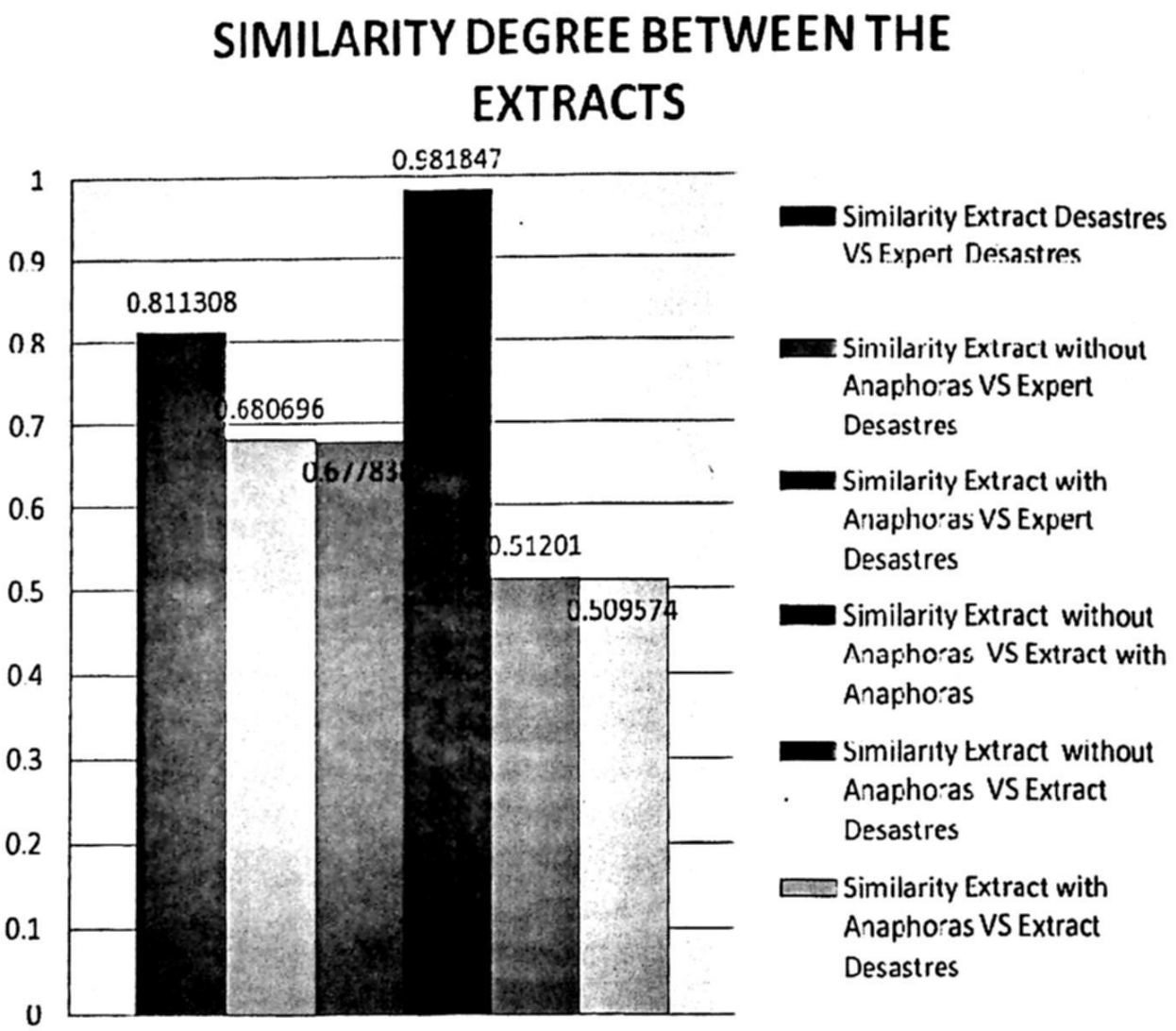
Document	Similarity Extract without Anaphoras VS Expert	Similarity Extract without Anaphoras VS Extract with Anaphoras
n1	0.534031	0.921052
n2	0.735135	1.0
n3	0.629107	1.0
n4	0.663414	1.0
...	...	...
n21	0.40566	0.916334
n22	0.62	1.0
n23	0.769230	0.917391
n24	0.721649	1.0
n25	0.591928	1.0
Antrax	0.593023	1.0
Arqueología	0.690476	1.0
Ciencia	1	1.0
Comercio electrónico	0.5	0.962358
Cristianismo	0.349999	1.0
...	...	...
Genoma humano	0.473282	0.960548
Guanajuato electrónico	0.661710	0.955453
Alimentación	0.581818	0.965442
General Average	**0.599416**	**0.981847**

Table 2. Similarity degree between the extracts

Document Corpus Desastres	Similarity Extract Desastres VS Expert Desastres	Similarity Extract without Anaphoras VS Expert Desastres	Similarity Extract with Anaphoras VS Expert Desastres	Similarity Extract without Anaphoras VS Extract with Anaphoras	Similarity Extract without Anaphoras VS Extract Desastres	Similarity Extract with Anaphoras VS Extract Desastres
Afecta Incendio	1.0	0.437956	0.437956	1.0	0.372262	0.372262
Alarma de Incendios	0.705882	0.462686	0.462686	1.0	0.283105	0.283105
Alerta por un incendio	1.0	0.550387	0.479729	0.907767	0.488372	0.425675
Alerta Roja	0.888482	0.789189	0.789189	1.0	0.607594	0.607594
Amaga incendio	0.607028	0.778688	0.778688	1.0	0.615803	0.615803
Amenazan a Arizona	0.927756	1.0	1.0	1.0	0.784482	0.784482
Amenazan Sydney	0.771929	0.833333	0.778098	0.934844	0.384615	0.350194
Arraso incendio	0.638888	0.769230	0.769230	1.0	0.741258	0.741258
Aumentan daños	0.857142	0.729805	0.729805	1.0	0.672268	0.672268
Ya son 205	0.478260	0.719008	0.719008	1.0	0.0625	0.0625
Visitara Fox zona	1.0	0.929752	0.929752	1.0	0.743801	0.743801
Tormenta Tropical	0.757201	0.585365	0.585365	1.0	0.525252	0.525252
Toneladas de ayuda	0.894977	0.633451	0.633451	1.0	0.528	0.528
Tifon Siembra	0.638115	0693121	0.693121	1.0	0.833638	0.833638
TERREMOTO	0.808510	0.714285	0.714285	1.0	0.561797	0.561797
Temporada de incendios	0.876190	0.946808	0.942105	0.987013	0.655172	0.653409
...	...	...	...	...	...	...
General Average	**0.811308**	**0.680696**	**0.677838**	**0.981847**	**0.512010**	**0.509574**

In Table 2 we made the comparison between the results obtained by our system against the results obtained by the system realized in the master thesis "Automatic Generation of Multiple Document Summaries" [24]. The Corpus with which was made the comparison called DISASTERS was realized in the INAOE (National Institute of Astrophysics Optical and Electronic) and is a corpus specialized in Natural Disasters of news of Mexican newspaper of national circulation; this corpus has 300 documents, where we took a sample of 100 documents to realize our tests.

With the aid of the author of the thesis [24] was possible compare the degree of similarity between the summaries of our system against his system, since he provided us the corpus disasters and the summaries obtained by his system. We have to mention that the system of the thesis [24] is a system dedicated to extract specialized information of natural disasters reason why the used metrics were adapted essentially to extract certain kind of information in this type of documents, and our system is not specialized, it means is of general propose, in addition, our system is designed to always extract 5 sentences, the same that the extract given by the experts. In the corpus disasters the extracts given by the experts did not have a minimum or maximum length at the same that the extracts given by the system of the thesis [24].



GRAPHIC 1. Similarity degree between the extracts

All the previous did that the performance of similarity of our system against the extract of the experts be less. In Table 2 at the same that in the Graph 1, we can observe the degree of similarity between the summaries of both systems against the extract of the experts, observing that the degree of similarity of the extract generated by the system of the thesis [24] against the expert of the corpus disasters has better results 0.811308 against the 0.680696 of our system and the degree of similarity between the extracts generated by both systems is 0.512010, also is show the similarity degree when the sentences with anaphoras are added to the extract, in all the cases the results were less.

Table 3 shows the results of the thresholds obtained for each metric after analyzed the 100 documents that conform ours Corpus, is to say, this is the threshold that is considered important for the decision if the metric is or not used to evaluate the sentences of the document. The most used metrics that were selected by the system were: Centroid, SentencesCompare, SmallSentences, follow by TransitionPoint. The less used metrics were: ImportantWords and FisrtsLastsentences.

Table 3. Results of the obtained threshold

Metrics	General Average of the Thresholds
Centroide	0.949328
FirstLastSentences	0.249236
ImportantWords	0.090041
ProperName	0.515761
SentenceCompare	0.647503
SentencePosition	0.797081
SmallSentences	0.900797
TitleCompare	0.407710
TransitionPoint	0.837108

6 Conclusion

In this article we present a way to obtain automatic summaries by extraction in the Spanish language. The obtained results were compared against the extracts of documents made by one of the three experts and also we compare the obtained extract of the system with and without anaphoras between them.

The results obtained by the automatic summaries by extraction using the techniques that were shown in the article, have obtained favorable results, even better than the shown in the article [19]. It is important say that the results obtained for the VP by the TP [19] are good, since that algorithm obtained around 50% of the sentences of the summaries given by the experts, and in spite of the previous the new system recovered approximately 60% of the sentences, improving the effectiveness in the automatic summaries by extraction, in addition in this new article is included a study of anaphoras, with this it is tried to give a major coherence to the obtained summaries, the degree of similarity between the extract without anaphoras and the extract with anaphoras is very high, 98%, this means that after including the anaphoras did not manage to recover a major amount of important sentences.

To validate again the quality of the obtained summaries, we obtained the original CORPUS and the extracts which were used to worked in [24]. The summaries obtained by that system and ours had a level of similarity of 50%, this was because the system developed in [24] did not necessarily gives back 5 sentences, as our system did, and his metrics were adapted to keywords within the area of natural disasters, aspect that our system does not consider, so we intended to do a more general system.

The obtained results are encouraging, although the similarity levels were not so high, this does not affect the obtained results, because the way of make the summaries depends so much in the dominion of Spanish language by the person that make them.

Nowadays our group of investigation still working in the improvement of the algorithms already proven as well as new algorithms to also continue obtaining still better results in the obtaining of automatic summaries non just by extraction but by abstraction for the sake of obtaining greater precision.

References

1. Diccionario Enciclopédico de la Lengua Castellana, Ed. Codex, Buenos Aires, 1968
2. Real Academia Española. http://www.rae.es
3. Baeza-Yates R., Ribeiro-Neto B., (eds.), Modern Information Retrieval. ACM Press, New York, 1999.
4. Bueno C., Tesis de Maestría: "Métodos para la Generación de Extractos mediante el uso del Párrafo Virtual", Benemérita Universidad Autónoma de Puebla Facultad de Ciencias de la Computación, Otoño 2005, pp. 8-13.
5. Guha Sudipto, Rastogi Rajaeev, Shim Kyuseok, "ROCK, A Robust Clustering Algorithm for Categorical Attributes", Information Systems, 2000, Vol. 25, No. 5, pp. 345-366.
6. Leal L., Vilariño D., López F., Jiménez H., "The Virtual Paragraph as a Retrieval Information Technique Implanted in Mobile Agents", Proceedings of the International Conference on Computational Intelligence for Modelling, Control and Automation CIMCA 2006, Sydney Australia, Nov. 2006, ISBN 0-7695-2731-0.
7. Reyes B., Moyotl E., Jiménez H., Tesis de Licenciatura "Reducción de términos índice usando el punto de transición, Facultad de Ciencias de la Computación, BUAP.
8. Urbizagástegui R., "Las posibilidades de la Ley de Zipf en la indización automática", Reporte de la Universidad de California Riverside, 1999.
9. MINITAB, http://www.minitab.com/
10. KUPIEC J., PEDERSEN J., CHEN F.. A Trainable Document Summarizer, 2005.

11. Mateo P. L., González J.C., Villena J., Martínez J.L., Un sistema para resumen automático de textos en castellano., DAEDALUS. S.A., Madrid, España.
12. ROJAS F., JIMENEZ H., PINTO D., LOPEZ A., Dimensionality Reduction for Information Retrieval, 2006.
13. GUERRA A.. Aprendizaje Automático: Clasificación, páginas 6-8, 2004.
14. Harabagiu S., MOLDOVAN D., CLARK Christine, BOWDEN M., HICKL A., WANG P.. Employing Two Question Answering Systems in TREC-2005, 2005.
15. ZAJIC D., DORR B., SCHWARTZ R.. Automatic Headline Generation for Newspaper Stories, 2002.
16. Karamuftuoglu M., Approach to Summarisation Based on Lexical Bonds, 2002.
17. TEUFEL S., MOENS M., Sentence extraction as a classification task. Workshop 'Intelligent and scalable Text summarization', 1997.
18. EDMUNSON. New Methods in Automatic Extracting. Journal of the Association for Computing Machinery, páginas 264-285, 1969.
19. MARQUEZ J. A., RENDON P. R., RODRIGUEZ R., VILARIÑO D., BELTRAN B., Comparación de dos métodos para la obtención de resúmenes automáticos, CIINDET 2007, ISBN-968-9152-00-9.
20. www.hipertexto.info/documentos/resumen.htm
21. Carcedo Elena F., "Los Géneros y su práctica", Ed. Textos UAP, 2003, pág. 57.
22. Moldovan D., Clark C., Harabagiu S. Temporal Context Representation and Reasoning, 2005.
23. Sidorov Grigori, Olivas Zazueta Omar, Resolución de anáfora pronominal para el español usando el método de conocimiento limitado, CIC, IPN, México D.F.
24. Villatoro Tello Esaú, "Generación Automática de Resúmenes de Múltiples Documentos", Tesis de Maestría del INAOE, Febrero 2007, Tonantzintla, Puebla.

Bilingual Information Retrieval using a Parallel Platform

Alberto Márquez, Darnes Vilariño, Erick Pinacho,
Mireya Tovar, and Beatriz Beltrán

Abstract. This paper discusses the problem of Bilingual Information Retrieval over bilingual documents using lexical resources and implementation over a parallel Kraken Platform designed by [1]. The system uses bilingual documents (English and Spanish), which are pre processed and post processed to group these documents using metric as the title, proper names and the first paragraph of the document. It was subsequently used the platform to send clusters obtained and the query to each slave lifted in the Platform Kraken and each slave showing the similarity to the query.

Keywords: Clustering algorithms, unsupervised learning, parallel system, incremental algorithms.

1 Introduction

In the last few years, natural language processing (NLP) techniques and tools have been incorporated into information retrieval (IR) systems with varying degrees of success, it has taken interesting among researchers, when they have to devise new techniques to assist user in the process of searching. Currently one of the richest in containers information is the Web, but its growth has generated problems at the time of information retrieval, mainly due to repetitive information, diversity of data, lack of standardization to represent data, and others. This become even more complex if they are to a query as provided in both English and Spanish will return documents relevant to it.

Since the 1980s, corpus linguistics has developed at an accelerated speed. While the construction and exploitation of English language corpora still dominate the research of corpus linguistics, corpora of other languages, particularly typologically related European languages such as French, German and Portuguese and Asian languages such as Chinese, Korean and Japanese, have also become available and have notably added to the diversity of corpus-based language studies. In addition to monolingual corpora, parallel and comparable corpora have been a key focus of non-English corpus linguistics, largely because corpora of these two types are important resources for translation and contrastive studies.

A parallel corpus can be defined as a corpus that contains source texts and their translations. Parallel corpora can be bilingual or multilingual. For a comparable corpus, the sampling frame is essential. The components representing the languages involved must match with each other in terms of proportion, genre, domain and sampling period.

This paper presents a parallel system to process bilingual text (English and Spanish) for cross language querying, the sets of metrics are dependent on lexical resources, and linguistic tools. Which cluster is send to Slave Node with query, which

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 199-209

Received 23/06/08
Accepted 01/08/08
Final version 10/09/08

Slave Node returns the comparison between the clusters of documents and query. It is verified in a practical way that a trading system of machine translation can employ approximately ten months to provide the translation of 250 000 documents, this is not feasible for a real system of IR. In this system only used a first paragraph of document.

2 Metrics

Several systems make use of independent modules, each assigned a score (weight) to each unit (words, sentences, paragraphs, etc.). Subsequently a module in particular is responsible for making the combination of different weights assigned to each unit getting a single value for each unit; eventually the system will return n the first units with the highest weight. Below are the most commonly used techniques.

a) **Proper Names**: It is a kind of attribute that refers to individuals and/or locations indicate that the sentence could contain important information. One example is the news of a natural disaster, the sentence containing the name of the place where the damage occurred could be important, which is a point of comparison with other documents. This paper takes its own name as one that begins with capital letters, which is not an empty word or closed and it is not possible to find its [12].

b) **Similar to title**: Several systems make use of keywords to identify relevant sentences [6]. But as we know it is impossible to define a set of keywords that apply to all kinds of documents because doing so would force the system to depend on the thematic domain and further language. With what is considered is with the title. Often the title of a document of great information on the contents. This is why it was decided to use this attribute. When there is no title, the first sentence of the document is taken as the title and thus prayers with greater similarity to the title are considered important [13] - [16].

c) **First paragraph**: It has been shown that the terms are both in the title as in the first paragraph largely reflect the theme of a document, these results have been exhibited in [17].

To compare two sentences are taken on a role of similarity, this is done by taking a value greater than a threshold defined β (between 0 and 1). The function $Sim\ (O_i, O_j) \geq \beta$ [18] to see (1).

$$sim(O_i, O_j) = \frac{|O_i \cap O_j|}{|O_i \cup O_j|} \quad (1)$$

Where Oi and Oj are a set of words. Another method is to use the Point of Transition (PT), which separates the words of a document in terms of high and low frequencies. The words about the point of transition are the most relevant. This method is ideas based on investigations by the formula for calculating the PT is presented in (2).

$$PT = \frac{\sqrt{1 + 8 \times I_1} - 1}{2} \quad (2)$$

Where I_1 represents the number of words frequently 1. According to the characterization of frequencies mean the PT can occur in the vocabulary of a text by identifying the lowest frequency of high, which is not repeated. For example, with the words around the PT by 25%, it generates a virtual paragraph. The results when applying the metrics are favorable in information retrieval.

3 Clustering Algorithms

Clustering algorithms heuristically build clusters from sets of objects which are characterized by several features and a similarity function. These algorithms try to maximize the similarity between objects in the same cluster and/or minimize the similarity between objects in different cluster. Clustering algorithms are used in many fields of non formalized sciences like information retrieval, data mining genetics, computer vision, biology, earth science, and others.

Essential elementals for to resolve the problem of clustering are: the representation of space objects, the measure of similarity between objects (not necessarily a distance) and clustering or heuristic approach to implement. In some of these methods is necessary to define also a measure of similarity between the clusters, which is defined in terms of similarities between objects that make up clusters [10].

We used the single-pass algorithm [11]. There are a small number of clustering algorithms which only require one pass of the file of object descriptions. Basically they operate as follows:

1. – The object descriptions are processed serially.
2. – The first object becomes the cluster representative of the first cluster.
3. – Each subsequent object is matched against all cluster representatives existing at its processing time.
4. – A given object is assigned to one cluster (or more if overlap is allowed) according to some condition on the matching function.
5. – When an object is assigned to a cluster the representative for that cluster is recomputed.
6. – If an object fails a certain test it becomes the cluster representative of a new cluster.

Once again the final classification is dependent on input parameters which can only be determined empirically (and which are likely to be different for different sets of objects) and must be specified in advance. For the clusters of documents were used metric described above.

4 Kraken Plataform

The search for solutions to real problems usually requires using a lot of calculations. This causes delivering results is too late. One solution to this problem is offered by parallelism, which allows several operations at once, favoring reducing the time required to execute a task.

Java provides mechanisms for competition, management functions and low-level technologies for developing distributed applications. These and other features that owns Java, they do an excellent candidate to develop software side, for details see [18]. The platform contains elements that are listed below:

a) **Node's Name**: The platform is shaped by a set of computers, which will be running the Demon Slave, Master Demon, or both. When you start the Demon in a particular computer, the Demon is responsible for announcing to the rest of the Platform that the node is added. The Platform is responsible for distributing classes between the nodes. For sending messages is necessary to know the addressee, and sometimes the source.

b) **Sokets**: To make communication across the platform is used in a socket UDP multicast. The use of UDP socket implies that the package delivery is not guaranteed, nor the order. However, the platform is targeting a cluster or a computer network LAN. IP addresses are among 244.0.0.0 and 239. Even used to multicast. Regardless of the direction given to any network interface can be used any direction in the range previously indicated in any application. The address in the range indicated, plus a number of standards UDP port, and can receive and send messages on multicast.

c) **Shared Memory**: The platform offers a building which serves as shared memory. With it, you may give the developer a region of memory that is shared among all the nodes of the *Platform*. The region of memory offered is seen by all the nodes of the platform, using replicas. Each node has a complete replica of shared memory. The Shared Memory is designed to perform a disco paging through pages of small size. With this, it gives the developer a memory size theoretically limited by the capacity of disk.

d) **Synchronization:** Several hardware platforms offer parallel mechanisms synchronization. Depending on the type of platform, processors can be synchronized by the control unit-level instruction, or you can synchronize at regular intervals at certain points of execution. The synchronization points are a common construction in the software tools for developing parallel and distributed applications. At software, a point of synchronization is to ensure that the different running processes reach a point in the execution in which processes one-one will stop until they all arrive there, when his execution continued.

All functionality accessible user classes are grouped within the class *PlatformBind*. One such instance is communicated to each user's task that is created within the Platform, and this is done through the interface *ClassStub*. All user tasks to be run in

the Platform must implement the interface *ClassStub*. With this class of user must implement methods setProperties() and *execute ()*.

Through the method *setProperties()* will have access to the user's task an instance of *PlatformBind* who used to use the features of the Platform. This method will be invoked before the implementation of the method *execute()*. The method *execute()* must meet the code that the user wishes to be executed. With the implementation of the interface *ClassStub*, will be achieved expose the functionality of the user platform.

5 System Design

This section presents the design for the construction of information retrieval systems in Parallel (RIP). The objective, as mentioned before, is to develop a system that given a query by the user, it can retrieve documents in both English and Spanish. The most important components for the development of the system are illustrated in the figure below:

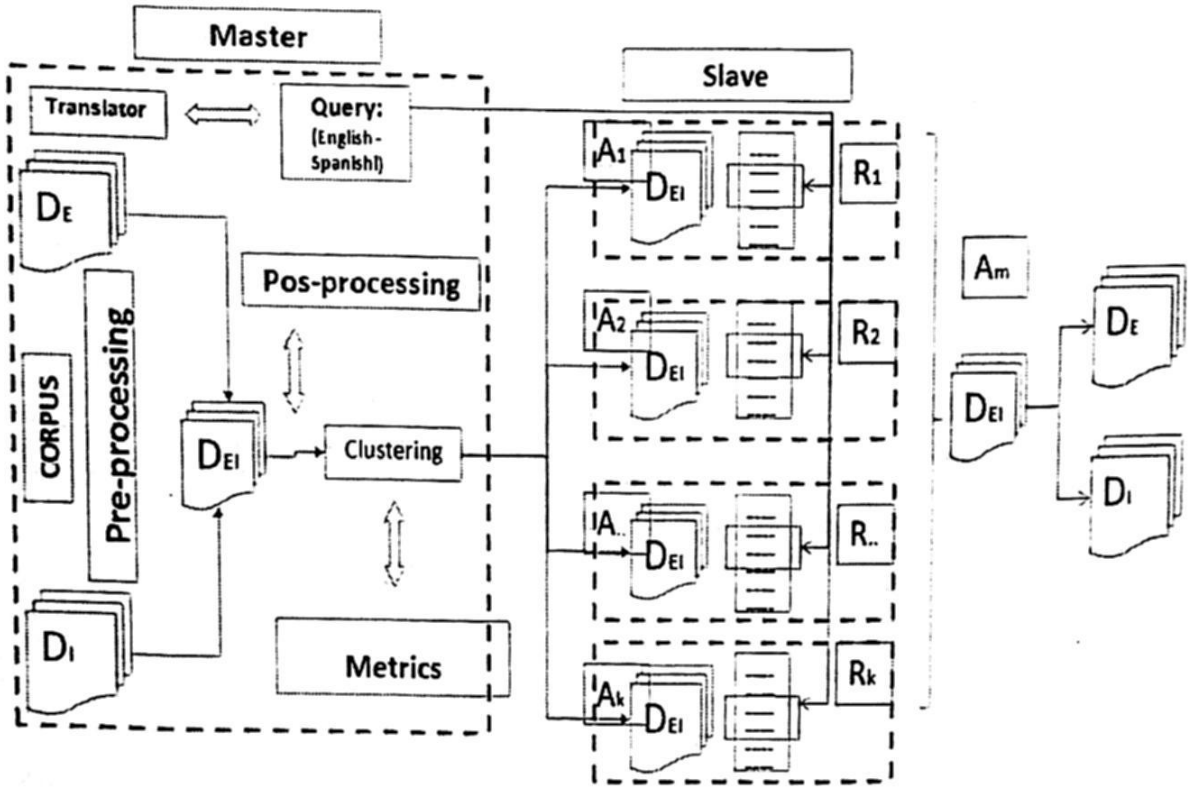


Fig.1. RIP System Scheme.

· Description:

a) **Corpus**: Our corpus procedures were initially developed to process D_I (English documents) and D_E (Spanish documents), but are extendible to other languages.

b) **Pre processing**: For each document (English - Spanish), it was divided to sentences, we get the titles, we get proper name, we get the language and to extract the text that will be translated.

c) **Post processing**: It removes stopwords (both languages), titles, proper name and text to translate, and to stemmer.
d) **Clustering** ($A_1..A_k$): The grouping of documents according to the characteristics obtained (proper names, paragraph original-translate, titles original-translate) by an algorithm *single pass*.
e) **Query**: We get the user's query and this is processed, to translate, to remove stopwords and to stemmer.
f) $R_{1..n}$: It is the representation of each set of documents previously grouped. The representation is divided into two parts (English-Spanish) each obtained through implementing Paragraph Virtual Point Transitional

g) D_{EI}: It is the representation of each set of documents previously grouped.

The communication scheme proposed for the parallel algorithm is the master-slave, and then defines the tasks which will make the master and slave respectively.

a) Master
   - ♣ Pre processing of corpus.
   - ♣ Post processing of corpus.
   - ♣ Generating groups (k).
   - ♣ Send *k-groups/n-tasks*.
   - ♣ Send query to each Slave.

b) Slave
   - ♣ Receive k groups.
   - ♣ Receive query.
   - ♣ 2 process (English - Spanish).
   - ♣ Each process:
     - Convert t documents to 1 document.
     - Generate Point Transition (PT).
     - Make the Virtual Paragraph (PV).
     - Evaluate query with PV.
     - Show the similarity.

6 System Implementation

The search for solutions to real problems usually requires using a lot of calculations. This causes delivering results is too late. One solution to this problem is offered by parallelism, which allows several operations at once, favoring reducing the time required to execute a task. The overall activities being carried out on the platform are:

a. **Starting the Master Daemon**: It is controls the platform.
b. **Starting the Slave Daemon**: It lifted the demons slaves for each node.
c. **Registration Demons Slaves**: Each slave demon informs to slave master and its existence is saved by the master slave.

d. **Receiving Task Definition**: According to the definition of user classes, the master demon communicates with the demons slaves to the distribution of classes.
e. **Execution of tasks:** Each node where there is a demon slave begins a task that is user instance of the class.
f. **Closure of Demons Slaves**: After performing the tasks, can be closed by the demons slaves.
g. **Closure of Master Demon**: Once closed demons slaves, we can close the master daemon.

Master: By creating the master should take into account the implementation of the platform and is declared as follows: private *PlatformBind name*; order to use the functionalities of the platform. The platform has been developed in Java 1.5, from this version has a specialized package for the development of competing sections, the *java.util.concurrent* package. Taken out this is due to declare the interface *ClassStub* is required to be implemented this method in order to pass the request of *PlatformBind*, but in the *execute()* contains the code to execute. It is stored consultation with the role *strQuery()*, and to process information is the role *LeerDatosPreProcessing()*; which obtains documents to carry out the consultation, the format is presented in text mode.

Slave: The slave receives the group and takes the appropriate consultations to carry out the comparison, as the documents are in English-Spanish language must verify the document, this will help to shape what is a single text information in Spanish and English, Formed this is done the PT (Point Transitional) with *getTransitionPoint (String strModified)* where *strModified* is a string that contains documents in English or Spanish. To assess the virtual consultation with paragraph was used according to Jaccard: *Jaccard (String cad1, String cad2)*, which returns a value as a comparison.

7 Results

The next result is the collection of 19 documents in English and Spanish with threshold .005, the results are shown in the following Table 1.

Table 1. Groups formed and assigned to each slave.

Groups	No. Documents	Slaves
0	4	Slave 1
1	4	Slave 2
2	8	Slave 3
3	2	Slave 1
4	1	Slave 2

In the able above were obtained with the 19 documents 5 groups, which were divided between 3 slaves, each processed this information to compare with the query.

The following Table 2 shows the results when comparing the consultation with each group formed and three slaves up.

As can be seen, each slave contains a number of groups, table [I], and each group generates its representation, in both English and Spanish, location of the transition points and then shaping the Virtual Paragraph, the comparison of paragraph Virtual in English and Spanish with the consultation is given in the Table 2.

Subsequently took place with more documents, 40 documents was conducted with both English and Spanish, and were divided into groups each slave to be able to process information and compare it with the consultation.

It is worth mentioning that each slave is generated the point of transition, and it conforms around what is called virtual paragraph will then be compared to the query by applying the role of similarity between Jaccard, the results are shown in the Table 3.

Table 2. It shows the results when comparing the query with each slave, it is worth to say that each slave its rightful certain groups.

Query: Spanish: ataud metal subsuel English: metal conffi subsoil			
PV	**Esclavo 1**	**Esclavo 2**	**Esclavo 3**
Spa	0.0	0.0	0.1176
Eng	0.0	0.0	0.1176
Spa	0.0	0.0	
Eng	0.0	0.0	

Table 3. Groups formed and assigned to each slave.

Groups	No. Documents	Slaves
0	4	Slave 1
1	8	Slave 2
2	12	Slave 3
3	8	Slave 1
4	3	Slave 2
5	4	Slave 3
6	1	Slave 1

As can be seen that the group has more elements should have consulted more widely accepted view, this provides a level of acceptance in the results.

Table 4. It shows the results when comparing the query with each slave, it is worth to say that each slave its rightful certain groups.

Query: Spanish: Campesinos realizan una marcha, ley agraria. English: Peasants perform a march, land law.			
PV	**Slave 1**	**Slave 2**	**Slave 3**
Spa	0.0	0.0	0.03846
Eng	0.0	0.0	0.03333
Spa	0.0	0.0	0.0
Eng	0.0540	0.0	0.0
Spa	0.0	0.0	
Eng	0.0	0.0	

It took time to recover, both linear and in parallel, The results of the linear system shown in Fig. 2 and in parallel in Fig. 3.

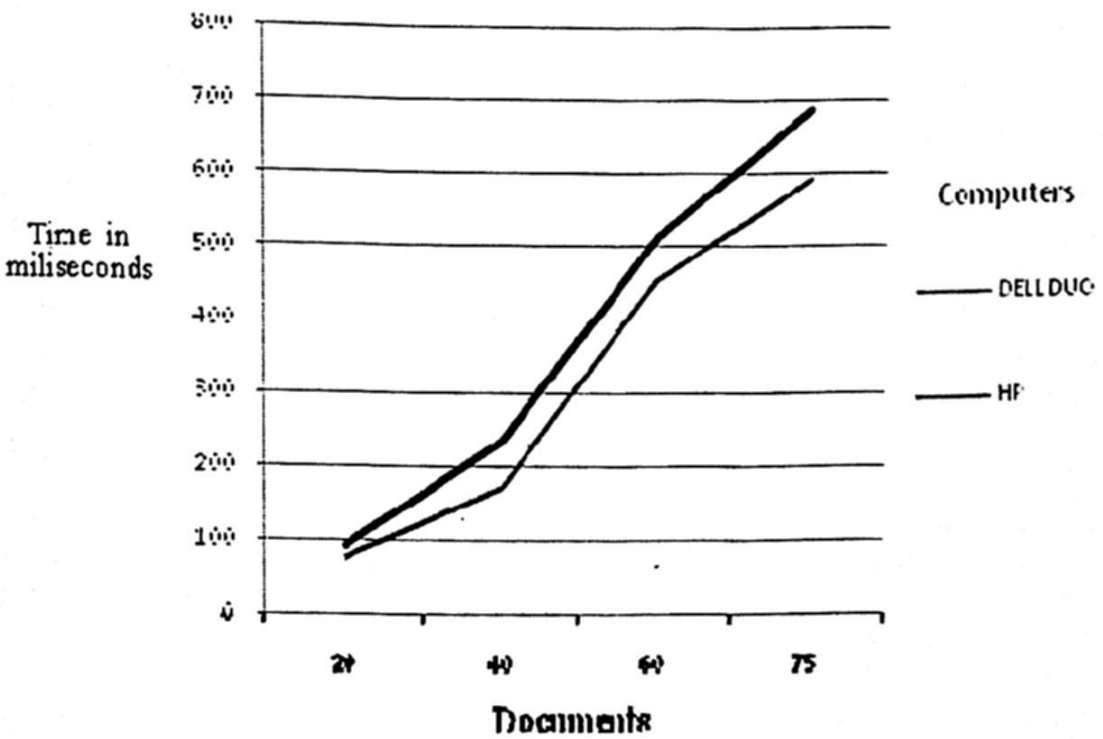


Fig.2. RIP sequential.

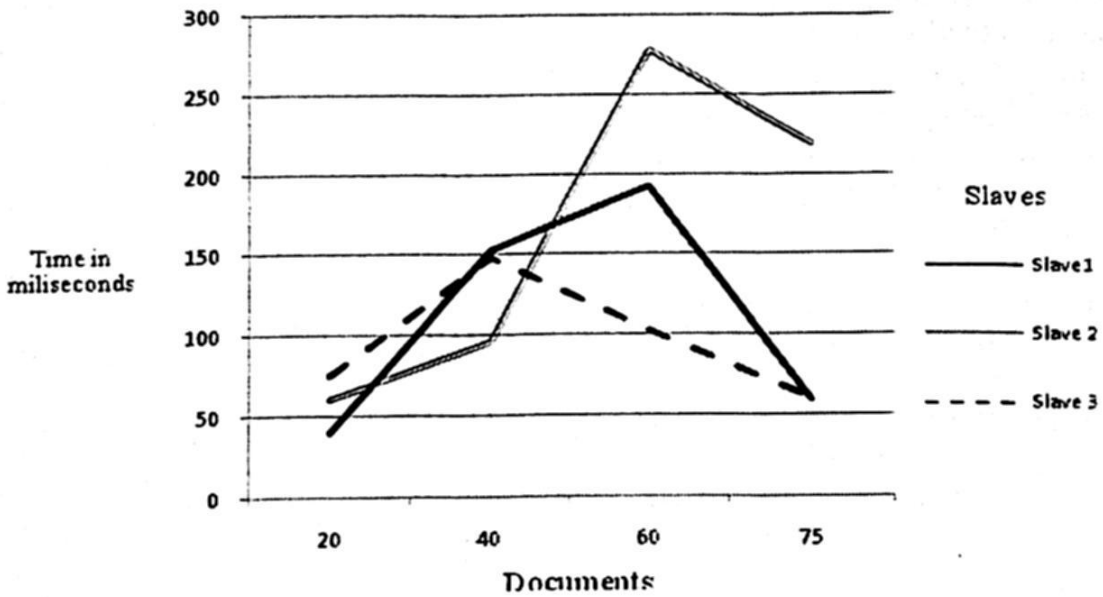


Fig.3. RIP parallel.

As expected, in parallel reduces the time and cost for computer information retrieval.

8 Conclusions

The results so far are satisfactory, work is under way on a platform developed in Java, applications for parallel and distributed [18], the functions of master node are programmed in its entirety. The platform allows for an appropriate way and user-friendly management of shared or distributed memory. The system detects the language of the text and analyze texts in both English and Spanish, also achieves these documents grouped according to criteria such as language and can be seen from the tables above yields a high degree of information retrieval in both languages.

In order to solve problems optimization algorithms using large-scale parallel, it was decided to create a platform itself, aimed directly to exploit the architecture of clusters. One of the most popular tools in science to develop solutions using the mechanism of passing messages is MPI.

The usefulness of Java for concurrence, include a queue asynchronous. With queues asynchronous, is no longer necessary to create critical regions or traffic lights for the addition or subtraction of elements, as they include the control of competition within the same class. The Platform asynchronous queues were used for receiving messages.

Another problem that occurs in the information retrieval recovery is bilingual in both languages, however, for this work was not necessary to translate the whole document but a single paragraph, which significantly reduces the computational cost. In addition the use of a Kraken platform parallel.

References

1. Ángel F. Carlos G. J. Luis. Recuperación de Información utilizando el modelo vectorial. Participación del taller CLEF-2101, Mayo 2002.
2. Van Rijsbergen, C.: "Information Retrieval". Butterworth, London, 1979.
3. Greengrass, E.: "Information Retrieval": A Survey. Tecnical Report, november, 2000.
4. Ruiz Shulcloper, J.; Lazo Cortés, M. ; Alba Cabrera, E.; Pico Peña R.; Sanches Gutierres, I.: "Workshop on Data Mining". Logical Combinatorial Pattern Recognition Group, ICIMAF, Cuba, december, 1999
5. Ester, M.; Kriegel, H.; Sander, J.; Wimmer, M.; Xu, X.:"Incremental Clustering for Mining in a Data Warehousing Enviroment". 1998.
6. Allan, J.; Carbonell, J.; Doddington, G.; Yamm, J; Yang, Y.: "Topic Detection and Tracking Pilot Study: Final Report". Proceeding of DARPA Broadcast News Transquiption and Understanding Workshop, pp. 204-228, 1998.
7. Nagesh, H.; Goil, S.; Choudhary, A.: "A Scalable Parallel Subspace Clustering Algorithm for Massive Data Set". 2000.
8. Forman, G.; Zhang, B.; "Linear Speedup for a parallel non aprocimate recasting of center based clustering algorithms; including K-Means, K-Harmonics Means and EM", 2000.
9. Gil-García, R. "Paralelización de algoritmos de agrupamientos jerárquicos para semejanzas reducibles en redes de difusión". Tesis de maestría, Departamento de Computación, Universidad Oriente, Cuba, 2000.
10. Sánchez, G.: "Desarrollo de Algoritmos para el agrupamiento de grandes volúmenes de datos mezclados". Tesis de Maestría CIC. IPN. México. 2001.
11. GUERRA A. Aprendizaje Automático: Clasificación, páginas 6-8, 2004.
12. J. L. Neto, A. A. Freitas, and C. A. A. Kaestner. Automatic text summarization using a machine learning approach. In Proceedings of the 16th Brazilian Symposium on Artificial Intelligence, pages 215-225, Porto de Galinhas/Recife, Brazil, 2002.
13. W. T. Chuang and J. Yang. Text summarization by sentence segment extraction using machine learning algorithms. In Proceedings of the 4th Paci¯c-Asia Conference on Knowledge Discovery and Data Mining. Current Issues and New Applications, pages 454-457, London, UK, 2000.
14. E. Hovy and C.-Y. Lin. Advances in Automatic Text Summarization, chapter Automated text summarization in SUMMARIST, pages 81-94. MIT Press, Cambridge. 1999.
15. S. Teufel and M. Moens. Sentence extraction as a classication task. In Proceedings of the ACL Workshop on Intelligent Text Summarization, pages 58-65, Madrid, España, 1997.
16. K. Rosales-López, M. Tovar-Vidal, D. Vilariño-Ayala, B. Beltrán-Martínez, H. Jiménez-Salazar. "Confección de resúmenes automáticos usando n-gramas" en IEEE 5° Congreso Internacional en Innovación y Desarrollo (Morelos, Cuernavaca). 2007.
17. Manning, D. C. y H. Schütze. Foundations of statistical natural language processing. MIT Press. 1999.
18. Pinacho, E.: "Una plataforma para el desarrollo de aplicaciones en paralelo usando Java" Tesis de Maestría BUAP, México, 2007.

Machine Learning and Data Mining

Academic Performance Model Through the Use of Data Mining

Claudio Gutiérrez-Soto, Patricio Oliva, and Angélica Paredes

Information Systems Department, Faculty of Business Sciences
Universidad del Bío-Bío, Concepción, Chile
cogutier@ubiobio.cl, patoliva@alumnos.ubiobio.cl,
anpade@ubiobio.cl

Abstract Data Mining is used in different disciplines for search of patterns and hidden models in databases. It is usually applied in business and marketing areas. This paper presents a Data Mining application in the superior education area. The main contributions of this paper are, at first, a set of standard variables with effect in the academic performance of students, secondly, obtaining predictive model based on Bayesian Network which determines with 96,55% the probability of successful semester for students from Department of Information Systems, Universidad del Bío-Bío, Chile.

Key Word: Data Mining, Survey, Data Base.

1 Introduction

Nowadays, of the superior education institutions generate a large amount of information related with their students. This information corresponds to data used in the administration of superior education organizations, as well as to academic backgrounds and, information of courses among other subjects. This information is relevant in the strategic decision taking of universities. However, there is a large uncertainty about depth of knowledge as well as factors which directly affect the academic performance of students [1]. The formerly mentioned makes complex to take policies intended to improve teaching/learning and so as reduce the student dropout rate. In fact, it's known that 40% of Chilean university students do not complete their studies, which is mainly attributed to lack of maturity, social and economical problems among other factors [2]. Even more, statistical data have been found in Chile and Latin America which demonstrate that careers such as engineering, architecture and laws present a graduation index under or equal to 30%. Careers like medicine, Dentistry, basic level education and special education show a graduation index over 69% [3].

On the other hand, the incorporation of Data Mining in the education area is recent. However, most of the Data Mining applications in the education don't consider all psycho-social variables which affect the student teaching/learning process. Through Data Mining application on Information Systems Department students from Universidad del Bío-Bío, factors affecting teaching/learning process are possible to be obtained.

The main contributions of this paper are: first, to obtain set of standard variables with impact in the academic success and, secondly, a predictive model by using

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 213-220

Received 25/06/08
Accepted 01/08/08
Final version 10/09/08

Bayesian Networks. Our work is based on CRISP-DM methodology, and SPSS Clementine and Weka are used to obtain predictive model.

This paper is organized as follow. Section 2 makes factorial analysis of surveys to students in order to establish the variables with impact on academic success of students. Section 3 presents predictive model through cluster analysis. Finally, section 4 presents conclusions and future works.

2 Obtain Variables with Impact in Academic Success

Data Mining is a tool applied in different areas, such as, DNA Analysis and biomedical applications [4][5], industrial sale and marketing [4][6], telecommunications [4][7], banking processes [4], financial industry [4] and medicine [8][9], among other areas. Recently, the education area has been favored with the use of Data Mining [1][10][11][12][13][14].

In Data Mining applications of education area, it is possible to find relationships between admission tests and the academic success [15]. In [12] a global classification model is obtained, which assigns the best tutor according to student profiles. In [1] taxonomy of processes where the students are involved and the specific tools supposed to support these processes is presented. On the other hand, in [5] university success predictive tool for students entering to the university is presented. This predictive tool is based on the use of neural networks.

However, varied factors have impact on academic success and desertion rate. In [16] a categorization composed by five significant factors affecting desertion is presented. These factors are classified as psychological, sociological, economical, organizational aspects being interactive between student and the institution. In [17] it is also possible to find four groups of factors that impact desertion rate of Caribbean and Latin America students. These are: external factors to the superior education system and own factors of system and institutions, academic performance and personal problems of students. This categorization includes all the factors with impact in the student desertion rate and their academic success.

Nevertheless, the works exposed in [1][10][11][12][13][14] do not consider all the factors affecting the academic performance and desertion rate of students formerly mentioned in [16] and [17].

One of the main contributions of this paper is to have a number of significant variables (factors), which impact the performance and academic success of Information System Department students of the Universidad del Bío-Bío, Chile. However, to reach this purpose, it was necessary to look for information from the University corporate databases.

2.1 Obtain Non-available Data

Surveys were carried out to obtain information non available in databases. Five dimensional groups of data were established which were later standardized through the analysis of main components (also well-known as factorial analysis). These dimensions included *student data* (sex, age, university entrance year, number of subject

matters renounced, number of failed subject matters and others); *Learning and study techniques of students* (data about their individual learning styles, number of hours dedicated to the study, use of the bibliographical material and relevance of group learning and others); socio-economic aspects (the student's socio-economic information), professor and used techniques (information regarding the styles and educational models applied by professors in their classes); and the use of technological tools for learning (information about impact of technological tools in the student's learning).

The sample corresponds to stratified random sample. The population size corresponded to 614 students from Computing Civil Engineering and Information Computing Engineering. The number of students interviewed corresponded to 87. Given the population sample, reliability level used corresponded to 5%, with a 10% of error.

In order to verify feasibility of factorial analysis, Kaiser-Meyer-Olkin (KMO) and Barlett tests were carried out. KMO test obtained value of 0,6145. On the other hand, Barlett test, specifically chi- square, the value obtained was 760,303; in freedom degrees the value obtained was 253 and, for significance the value obtained was 4308E-52. These values assure feasibility to carry out analysis [18].

Factorial analysis was made with SPSS Clementine software Eight components with correlated variables from surveys were found in this factorial analysis. These eight components represent 70% from total variables, being significant percentage. These components are in the Table 1 associated with the variables of the dimensional groups. The components founded are standardized according to the correlation of variables, that is, from highest to lowest. However, components Personal Information, and Works are available the University database. Therefore, these components won't be used to obtain models. However, Personal Information component is fundamental to carry out mapping between existing data in database and non-available data in databases in order to join both data sources.

Tabla 1. Name of the components related with the dimensional groups

Name of component	Dimensional Set
Personal Information	*Student data*
Time devoted to study	*Learning and study techniques of students*
Work	*socio-economic aspects*
Team study or Internet	*the student's learning and their study techniques*
	Use of technological learning tools
To understand the professor's classes (to learn in the classes)	*Professor and used techniques*
Interest in the subject matter	*Learning of Students and their study techniques*
	Professors and techniques used
Attendance	*the student's learning and their study techniques*
Study of guides and from works given by professor	*Professor and used techniques*

Other important item obtained from personal surveys was the definition of academic success. Here, a percentage of 95.34% determined that academic success corresponds to the non failure of subject matters in academic semester. Nevertheless, it is not possible to appreciate the incidence of the eight components in the definition of academic success since components only reflect correlations among variables. In order to standardize the incidence of the eight components on the academic success definition, a second personal survey was carried out. Where preference 1 (the most relevant component to achieve academic success) was to understand the professor's classes (P1), the second preference corresponds to the component *Attendance* (P2), the third preference corresponds to *time devoted to study* (P3), the fourth preference corresponds to *Study of guides (work sheets) and works given by professor* (P4), the fifth preference corresponds to *Interest in subject matter* (P5), and the last preference corresponds to *Study in group or by Internet* (P6).

2.2 Data Joining

One of the most effective classifiers, in the sense that its performance is competitive with state-of-the-art classifier, is so-called Naïve Bayes (NB)[21]. NB supposes that all the attributes are independent known the value of class variable value. Although this supposition is not very realistic the classifier NB is one of the most usedand competitive classifiers. An example of use of this classifier is for the use against the spam or mail garbage [22].

The information existing on database corresponds both to Computing Civil Engineering and Information Computing Engineering students. These data correspond to the years 1998 and 2006. These data correspond to Admit Year, Birth Year, Number of Applications for credit, Number of applications for failed subject matters, Accumulated transcript's average, Accumulated Credits, failed subject matters and Renounced subject matters.

In order to meet data from personal surveys with databases information, we proceeded to select a group of similar data among Personal Information, Admit Year, Birth Year and Failed Subject Matters. This allowed a subset of 180 registrations. Afterwards, 87 registrations were randomly selected to generate predictive model.

3 Obtain Predictive Model

3.1 Bayesian Networks

Classification is one o the basic tasks in the data analysis patterns recognition. Classification requires the construction of classifier, which is a function that assigns a class label to instances described by a set of attributes. The induction of classifiers from data sets of preclassified instances is a central problem in machine learning. Numerous approaches to this problem, which are based on various functional representations such as decision trees, decision lists, neural networks, decision graphs and decision rules [20].

The Bayesian Networks allow the NB performance improvement as well as manage the independency assumptions among variables [20]. Bayesian Networks represent the qualitative knowledge of a model by means acyclic graph. This knowledge is articulated in the dependence/independence definition among the variables that compose the model. Graphic representation for the model specification makes the Bayesian Networks to be a very attractive tool for the knowledge representation, where the representation of knowledge is important aspect of Data Mining [22]. Within the most popular classifiers based on Bayesian Networks, we find the Tree Augmented Naïve Bayes (TAN). TAN is a NB extension classifier, which seeks to maintain the NB computing simplicity but trying to improve the success rate during the classification. So, instead of supposing all the independent variables, certain dependences are admitted among attributes. Therefore it is supposed that attributes constitute a Bayesian Networks with tree form. The advantage of restricting the topology from the net to a tree, is that this structure can memorize easily [20][22]. Another popular classifier is the Bayesian Network Augmented Naïve Bayes (BAN). BAN possesses the same philosophy of the TAN, it proceeds learning a Bayesian Networks for the attributes excluding the class and later on by adding the C class variable and edges from C toward all the attributes is increased.

3.2 Predictive Model

In order to predict if a student will reach academic success, nominal variable disapproved subject matter has been selected. This variable allows us to predict the disapproval of a student per semester. To obtain this predictive model Bayesian Networks has been selected. Bayesian Networks possess several advantages, like the generation of a simple analysis model, even more, is one of the most solid theoretical focuses [19]. On the other hand, Bayesian Networks provide more exact model, that is; more classification tests more robust model in the time.

In order to generate tests for model, cross validation has been selected since the tuplas amount to validate model is not big. However, the amount is significant to total population. On the other hand, to validate model the parameters delivered by weka have been considered, these parameters are Correclty Classified Instantes, Mean Absolute Error, Root mean squared error, Relative absolute error and Root relative squared error.

Algorithm selected for model construction corresponds to the algorithm K2 with two parents. K2 is TAN algorithm extension which allows generation of classifier based on Bayesian Networks. On the other hand, K2 has been selected by the understanding that gives to the model, because TAN generates a classification tree where a node only has a predecessor, while K2 generates a graph, where a node possesses two predecessors.

Table 2 and Table 3 allow a view of results for K2 algorithm with one father, K2 with two parents, and the TAN algorithm. Here, we can observe that the K2 algorithm with a single father possesses higher percentage of classified instances. It also possesses the lowest absolute relative error and the smallest square relative error. Nevertheless, this algorithm has been discarded and we have selected the K2 with two parents since resulting model is more descriptive for analysts. Besides, it must be noticed

that in obtained model, nodes correspond to all the variables affecting the academic performance. These are the same variables in Table 1, by adding them the variable Admit Year, Age, Sex, N° of Disapprove Courses Last Semester, Preference 4, Preference 5, and Preference 6. The nodes number reaches the 19 variables.

By checking models, 84 cases were correctly classified from a total of 87 instances, and 3 cases were wrongly classified.

On the other hand, classifications summary with 87 cases is possible to be obtained from confusion matrix. Total of 34 classified instances obtained in this matrix were positive. It means a higher probability to fail in these 34 instances. On the other hand, a total of 50 instances classified obtained were negative, which indicates the probability of not failing.

Tabla 2. Weka Results for the Data Mining Process

Algorithm	Description	Correctly Classified Instances	Mean absolute error
Hill Climber	Maximum Parents 1	97.7011%	0.0547
Hill Climber	Maximum Parents 2	93.1034%	0.0828
K2	Maximum Parents 1	97.7011%	0.0547
K2	Maximum Parents 2	96.5517%	0.0547
TAN		93.1034%	0.0764
K2	Maximum Parents 2, simulate BAN	95.4023%	0.0606

Tabla 3. Weka Results for the Data Mining Process

Algorithm	Root mean squared error	Relative absolute error	Root relative squared error
Hill Climber	0.1485	11.3012 %	30.2492 %
Hill Climber	0.2083	17.1831 %	42.4263 %
K2	0.1363	11.3562 %	27.7728 %
K2	0.1645	11.9203 %	33.5125 %
TAN	0.2103	15.8646 %	42.8381 %
K2	0.1756	12.58 %	35.7676 %

4 Conclusions and Future Work

In this paper, we have presented the development of a Data Mining Application in the Information System Department of the Universidad del Bío-Bío, Chile. Here, we have tried to include most of the variables that have influence on the academic success and in the student desertion rates. For this purpose, we have appealed to non-existing data in the university database through personal surveys. Later on, this information has been joined to the University database data. Joining these data, a group of eight correlated variables denominated components has been obtained. It is important to delimit the existing error margin in this procedure for joining data from personal surveys and databases. However, this procedure is justified since there is no Data warehouse with every data which affect the student performance and the student retention rate

Nevertheless, we think it is a good approach of variables affecting the academic performance of students. Moreover, one of the main contributions of this work is to have a group of eight standardized variables that have impact in academic success of Information System Department Students.

On the other hand, predictive model has been obtained, based on Bayesian Networks which predicts 96,5517% of probability if a student will reprove some subject matter in the semester. Nevertheless, we think it is a good approach of variables affecting the academic performance of students. Moreover, one of the main contributions of this work is to have a group of eight standardized variables that have impact in academic success of Information System Department Students.

On the other hand, predictive model has been obtained, based on Bayesian Networks which predicts 96,5517% of probability if a student will reprove some subject matter in the semester .

References

1. Naeimeh, D.: Application of Enhanced Analysis Model for Data Mining Processes in Higher Educational System. In 6th Information Technology Based Higher Education and Training., pp. F4B/1 - F4B/6. (2005)
2. Díaz, J. P.: Los por qué de la deserción universitaria. consultado 08-09-2006. http://www.universia.cl/portada/actualidad/noticia_actualidad.jsp?noticia=58803
3. González, L., Uribe. D.: Estimaciones sobre la repitencia y deserción en la educación superior chilena. Consideraciones sobre sus implicaciones. consultado 21-01-2007. http://www.cse.cl/public/Secciones/seccionpublicaciones/publicaciones_revista_calidad_detalle.aspx?idPublicacion=35
4. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Simonv Fraser University, Morgan Kaufmann publishers. Second Edition. San Francisco, United States. Volume 1. pp. 5-15. (2001)
5. Han, J.: How can Data Mining Help Bio-Data Analysis. BIOKDD02 Workshop on data mining in Bioinformatics. Volume 2. pp.1-2. (2002)
6. Edelstein, H.: Building Profitable Customer Relationships with Data Mining. Two Crows Corporation, SPSS white paper-executive briefing. pp. 1-13. (2000)
7. Chang, W., Lee, H. Y.: Telecommunications Data Mining for Target Marketing. Journal of Computers, Volume 12 No. 4. pp.60-74. (2000)

8. Baylis, P.: Better Health Care with Data Mining. Two Crows Corporation, SPSS white paper-executive briefing. pp. 1-9. (1999)
9. Brossette, S., Sprague, E., Hardin, P., Waites, J. M., Jones, K. B., Moser, T.: Association Rules and Data Mining in Hospital Infection Control and Public Health Surveillance. Journal of the American Medical Informatics Association (JAMIA), Volume 5. pp.373-381. (1998)
10. Luan, J.: Data mining and Knowledge Management, A System Analysis for Establishing a Tiered Knowledge Management Model (TKMM). Proceedings of Air Forum, Volume 5. pp. 373-381. (2001)
11. Gabrilson, S.: Data Mining with CRCT Scores. Office of information technology, Geogia Department of Education. (2003)
12. Waiyamai, K.: Improving Quality of Graduate Students by Data Mining. consultado 11-02-2008. http://www.ku.ac.th/icted2003/document/kritsana.ppt#280,1,Improving quality of graduate students by data mining
13. Luan, J.: Data Mining Application in Higher Education. consultado 21-12-2007. http://www.pse.pt/Documentos/Data%20mining%20in%20higher%20education.pdf
14. Luan, J.: Data Mining and Knowledge Management in Higher Education- Potential Applications. consultado 12-12-2007. http://www.cabrillo.edu/services/pro/oir_reports/DM_KM2002AIR.pdf.
15. Erdoğan, Ş., Timor, M.: A Data Mining Application in a Student DataBase. Journal of Aeronautics and Space Technologies. Volume 2. pp. 53-57. (2005)
16. Braxton, J., Jonson, R.,. Shaw-Sullivan, A.: Appraising Tinto's theory of college student departure. In Smart, J. C.(Ed) Higher Education Handbook of theory and research, Agathon Press. Volume. 12. pp. 107-164. (1997)
17. Miel, E.: Modelos De análisis de la deserción estudiantil en la educación superior. consultado 21-01-2007. http://www.cse.cl/public/Secciones/seccionpublicaciones/publicaciones_revista_calidad_detalle.aspx?idPublicacion=35
18. ATS, UCLA. "Annotated SPSS output proncipal components análisis". consultado. 22-1-2007. http://www.ats.ucla.edu/stat/SPSS/output/principal_components.htm
19. J, Hernández. C, Ferri. JP, Ramírez. "Introducción a la minería de datos", Pearson Educación S.A . Primera Edición. Madrid, España. ISBN: 84-205-4091-9. pp. 257-278. 2004.
20. Friedman N., Gieger D., Goldszmidt M., "Bayesian Network Classifiers". *Machine Learning* 29. pp.131-163. (1997)
21. Duda, R.O., Hart P. E., "Pattern Classification and Scene Analysis. New York : John Wiley & Sons. (1973).
22. Hernández J, Ferri C. y Ramírez JP (2004). "Introducción a la minería de datos". Pearson Educación, Madrid. (2004).

Explorations of the BDI Multi-Agent support for the Knowledge Discovery in Databases Process

Alejandro Guerra-Hernández, Rosibelda Mondragón-Becerra, and Nicandro Cruz-Ramírez

Departamento de Inteligencia Artificial
Universidad Veracruzana
Facultad de Física e Inteligencia Artificial
Sebastián Camacho No. 5, Xalapa, Ver., México, 91000
{aguerra, rmondragon, ncruz}@uv.mx

Abstract. Knowledge Discovery in Databases (KDD) is the process of finding valid, novel, useful and understandable patterns in data, to verify hypothesis of the user or to describe/predict the future behavior of some event. The KDD process involves diverse techniques provided by tools like the Waikato Environment for Knowledge Analysis (WEKA), but usually without guidance. This work is an exploration of the use of Multi-Agent Systems (MAS) methodologies and tools to provide support in the KDD process while using such tools. The Belief-Desire-Intention (BDI) model of agency provides the right level of abstraction to approach this problem. First, the Prometheus methodology is used to analyse the KDD process in terms of MAS of BDI agents. Then, a MAS of decision trees inducers and Bayesian networks builders, that compete to generate the "best" hypothesis for a given KDD problem, is implemented. The main result of this exploration is a framework where it is possible to implement AgentSpeak(L) agents that use primitive actions of WEKA to form intentions for solving problems in the KDD process. Extensions in terms of the number of agents and their capabilities are easy to implement in this framework.

1 Introduction

The Knowledge Discovery in Databases (KDD) process [13] consists in selecting, preprocessing and transforming a data set obtained from several heterogeneous sources such as databases, plain files, data warehouses, etc., in order to facilitate the application of data mining algorithms that obtain the hidden patterns in this dataset. Subsequently, these patterns are interpreted and evaluated to select those that represent useful and novel knowledge [9].

There are several algorithms to carry out each one of the tasks of this process, implemented in tools as Clementine [34], DBMiner [20], Waikato Environment for Knowledge Analysis (WEKA) [36], among others. The difficulty arises when a neophyte has to choose the suitable algorithms for a given problem, since this decision depends on the nature of data, their representation, the mining task,

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 221-238

Received 21/06/08
Accepted 01/08/08
Final version 11/09/08

among other factors. Therefore, KDD is a complex process, that requires well trained users in a variety of disciplines including machine learning, statistics and domain knowledge. But even in this case, some parts of the KDD process can be simply tedious.

Multi-Agent Systems (MAS) have been proposed as a solution to the problems mentioned above. MAS are composed by agents specialized in data mining, which sometimes are distributed in a network [32, 23, 22, 3]. However, we observe that these approaches, often does not provide the level of abstraction required to reason about the KDD process and their participants. This is due to a weak notion of agency, which focuses on basic properties of agents as autonomy, reactivity, pro-activity and social ability [37]. The strong notion of agency conceptualizes and implements the agents as intentional systems. The Belief-Desire-Intention (BDI) model of agency [15, 33] is the best known of these approaches. In this model the agents perform practical reasoning using plans to form intentions to satisfy their desires. The folk-psychology language used in the model, is argued to offer a more effective communication among the participants in a KDD process in order to analyse it to implement support using BDI agents.

Our explorations for such a support are as follows: In order to design a MAS of rational agents that help the human experts in the KDD process, this one is analysed as performed by a MAS of human experts, following the *Prometheus* methodology [30]. The *Prometheus Design Tool* was used to get assistance for the design process and to generate a specification of the MAS easy to be implemented. As a result of the analysis we decided to implement a BDI MAS of six agents: a coordinator that receives requests to mine a given data set; a preprocessing specialist; and agents capable to execute ID3, C4.5, Naive Bayes, and TAN algorithms. These last four agents compete to produce the "best hypothesis" for a given problem, and cooperate with the preprocessing agent if necessary. The MAS was implemented in Jason [4], an interpreter of the BDI agent oriented programming language AgentSpeak(L) [33]. The idea was that these agents were capable of executing actions in WEKA [36], as part of their plans. We chose Jason because it is implemented in Java, as well as WEKA, so that the implementation of the intended agents seemed easier in this way.

Finally, the MAS was tested with different data sets, in order to know if such agent competency has sense. The results show that there is not such a thing as the "best method" for all the data sets, and that the exploration automatised by the MAS could be useful. But the main result is that we obtained a framework where it is possible to easily define new agents in the system and to improve the capabilities of the existent ones, i.e., extending their plan library or their beliefs.

The rest of this paper is organized as follows: Section two presents the antecedents, mainly the KDD process and the BDI agents as defined in AgentSpeak(L); section three introduces the tools and methods that were used to implement the MAS; section four comments some aspects of the design and implementation of the KDD process like a BDI MAS; section five reports the results of our experiments; and finally, section six presents the conclusions and discusses future work.

2 Antecedents

In this section, we describe the KDD process, as well as AgentSpeak(L), an agent oriented programming language under the BDI model of agency.

2.1 The KDD Process

KDD refers the non-trivial process of identifying valid, novel, potentially useful and understandable patterns in data [11, 12]. Figure 1 (adapted from Fayyad et al. [11]) shows that the KDD process has an iterative and interactive nature. Results obtained in the process are enhanced incrementally, possibly reconsidering previous decisions in the process [11, 12]. The steps of the KDD process [20, 11, 12] include:

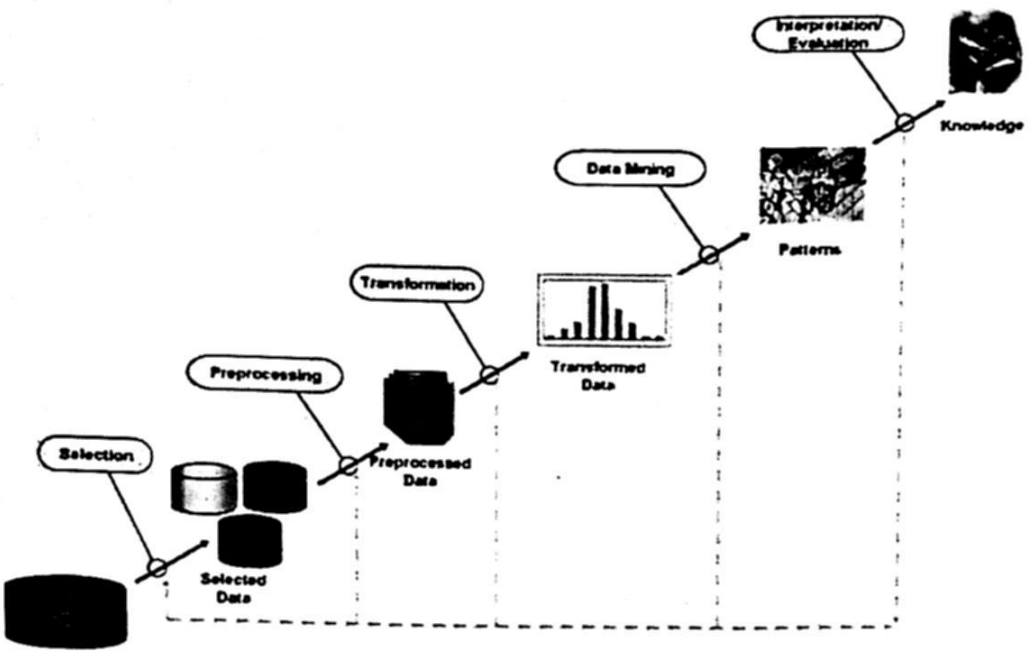


Fig. 1. Steps in the KDD process

1. Understanding the domain of the application and a background knowledge, as well as identifying the target of KDD process from the point of view of the user.
2. Selecting a subset of variables or a sample of data to create a set of target data, in which the discoveries will take place.
3. Cleaning and preprocessing the set of target data, e.g., removing noise if needed, dealing with missing data, etc.
4. Reducing and projecting data, through the selection of examples and attributes that are important for the target of the KDD process, or using dimensionality reduction or transformation methods to reduce the number

of variables under consideration or to find invariant representations for the data.
5. Selecting a data mining procedure, e.g., classification, regression, summarization, clustering.
6. Selecting some data mining algorithms and techniques to search patterns in the reduced data. The data mining expert decides which models or parameters are most appropriate.
7. Searching for patterns of interest (data mining) in a particular representational form as classification trees, rules, regression, clustering, etc.
8. Interpreting and evaluating the mined patterns, possibly reconsidering some of the steps $1 \ldots 7$.
9. Consolidating the discovered knowledge through its incorporation within another system or simply documenting and sending it to the interested participants in the process.

2.2 AgentSpeak(L) and BDI agents

BDI agents are intentional systems that continuously perceive their enviroment and take actions to modify it, based on their mental attitudes: beliefs, desires and intentions [15, 33, 8]. Beliefs represent the informational state of the agent. Desires, or goals, represent states that the agent would like to accomplish or bring about, considering its internal or external stimuli. Intentions represent the deliberative state of the agent, e.g., its commitment to some courses of action to accomplish its desires. These courses of action are built from plans in a plan library, e.g., the procedural knowledge of the agent. An event queue is usually used to process perception.

AgentSpeak(L) [33] is the language that was chosen to implement the BDI agents in this work, because it provides an abstract and elegant framework to program such agents [16]. The syntax and semantics of AgentSpeak(L) have been defined formally by means of a grammar and a operational semantics based on a transition system.

The grammar of AgentSpeak(L) as defined for its interpreter Jason [5], is shown in table 1. As usual, an agent ag is formed by a set of plans ps and beliefs bs. Each belief $b_i \in bs$ is a ground first-order term. Each plan $p \in ps$ has the form *trigger event* : *context* — *body*. A trigger event can be any update (addition or deletion) of beliefs (at) or goals (g).The context of a plan is an atom, a negation of an atom or a conjunction of them. A non empty plan body is a sequence of actions (a), goals, or belief updates. $\top$ denotes empty elements, e.g., plan bodies, contexts, intentions. Atoms (at) can be labelled with sources. Two kinds of goals are defined, achieve goals (!) and test goals (?).

The operational semantics [5] of the language, is given by a set of rules that define a transition system between configurations $\langle ag, C, M, T, s \rangle$, where:

- ag is an agent program formed by a set of beliefs bs and plans ps.
- An agent circumstance C is a tuple $\langle I, E, A \rangle$, where: I is a set of intentions $\{i, i', \ldots\}$, each $i \in I$ is a stack of partially instantiated plans $p \in ps$; E is a

Table 1. Grammar of *AgentSpeak(L)* [5]

ag	$::= bs \quad ps$		
bs	$::= b_1 \ldots b_n$	$(n \geq 0)$	
ps	$::= p_1 \ldots p_n$	$(n \geq 1)$	
p	$::= te : ct \leftarrow h$		
te	$::= +at \mid -at \mid +g \mid -g$		
ct	$::= ct_1 \mid \top$		
ct_1	$::= at \mid \neg at \mid ct_1 \wedge ct_1$		
h	$::= h_1; \top \mid \top$		
h_1	$::= a \mid g \mid u \mid h_1; h_1$		
at	$::= P(t_1, \ldots, t_n)$	$(n \geq 0)$	
	$\mid P(t_1, \ldots, t_n)[s_1, \ldots, s_m]$	$(n \geq 0, m \geq 0)$	
s	$::=$ `percept` $\mid$ `self` $\mid id$		
a	$::= A(t_1, \ldots, t_n)$	$(n \geq 0)$	
g	$::= !at \mid ?at$		
u	$::= +b \mid -b$		

set of events $\{(te, i), (te', i'), \ldots\}$, each te is a trigger event and each i is an intention (internal events) or the empty intention $\top$ (external events); and A is a set of actions to be performed in the environment.

- M is a tuple $\langle In, Out, SI \rangle$ working as a mailbox, where: In is the mailbox of the agent; Out is a list of messages to be delivered by the agent; SI is a register of suspended intentions (intentions that wait for an answer message).
- T is a tuple $\langle R, Ap, \iota, \epsilon, \rho \rangle$ that registers temporary information as follows: R is the set of relevant plans for a given event; Ap is the set of applicable plans (the subset of applicable plans which contexts are believed true); ι, ϵ, and ρ register the current intention, event and applicable plan along one cycle of execution.
- The label s indicates the current step in the reasoning cycle of the agent.

Figure 2 shows the interpreter for *AgentSpeak(L)* as a transition system. The operational semantics rules [5] define the transitions. Because of space limitations, table 2 shows only some these rules.

3 Methods

In this section, the methodology and tools used to design and implement the MAS in this work, are introduced. First, the Prometheus [30] methodology for designing MAS, and its associated software PDT, are introduced; then the interpreter of AgentSpeak(L), Jason [27, 6, 28, 1, 35] and the KDD tool, WEKA [36], are briefly described.

3.1 The Prometheus Methodology and its PDT Tool

Prometheus [30] is a methodology for developing intelligent agents and MAS. It covers all the phases of development of a system – specification, design, implementation and testing/debugging. Prometheus consists of three main phases:

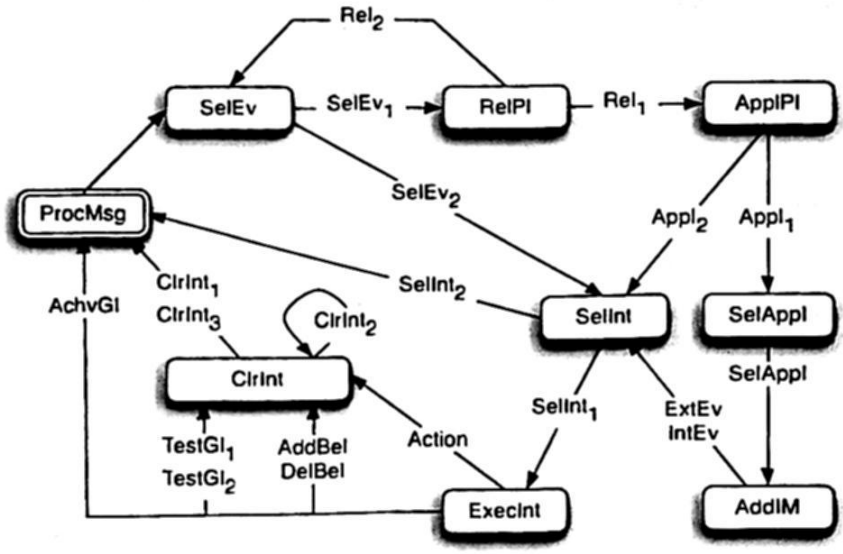


Fig. 2. The interpreter for *AgentSpeak*(L) as a transition system.

- **System Specification.** First, the goals and subgoals of the system are identified. Also, the actors (persons or roles that interact with the system) and their interactions with the system in form of perceptions and actions, are specified. Then, some scenarios are created for each actor. These scenarios show the operation of the system and consist of a series of steps including: goals, other scenarios, perceptions and actions. When the goals and scenarios are specified completely, the similar goals are grouped to form roles. Particular perceptions and actions are assigned to these roles. Finally, each step in the scenarios is assigned to their role and the data requirements for these scenarios are identified.
- **Architectural Design.** Based on the definitions generated in the previous phase, it is possible now to determine what kind of agents will be included in the system, as well as the interactions that will take place among them. To achieve this, several mechanisms are proposed, such as data coupling diagrams and agent acquaintance diagrams. In addition, in this phase is created the system overview diagram, that shows the structure of the system.
- **Detailed Design.** The internal details of each agent are designed and it is specified how the agents will carry out their jobs. Each agent is refined in terms of its capacities, internal events, plans, and data structures.

Additionally, the open software PDT [29] gives support to the Prometheus methodology. This tool provides: a graphic interface that allows to develop the definitions (diagrams) obtained through the Prometheus methodology and assists in the maintenance of a sound design because it provides verification between the diagrams; automatic spread of design elements when it is possible and appropriate; and assistance in the search for names.

Table 2. Some rules of the operational semantics of AgentSpeak(L).

(SelEv$_1$)	$\frac{S_\mathcal{E}(C_E)=\langle te,i\rangle}{\langle ag,C,M,T,SelEv\rangle \longrightarrow \langle ag,C',M,T',RelPl\rangle}$	s.t. $C'_E = C_E \setminus \{\langle te,i\rangle\}$ $T'_\epsilon = \langle te,i\rangle$
(Rel$_1$)	$\frac{T_\epsilon=\langle te,i\rangle, RelPlans(ag_{ps},te)\neq\{\}}{\langle ag,C,M,T,RelPl\rangle \longrightarrow \langle ag,C,M,T',AppPl\rangle}$	s.t. $T'_R = RelPlans(ag_{ps},te)$
(Rel$_2$)	$\frac{RelPlans(ps,te)=\{\}}{\langle ag,C,M,T,RelPl\rangle \longrightarrow \langle ag,C,M,T,SelEv\rangle}$	
(Appl$_1$)	$\frac{ApplPlans(ag_{bs},T_R)\neq\{\}}{\langle ag,C,M,T,ApplPl\rangle \longrightarrow \langle ag,C,M,T',SelAppl\rangle}$	s.t. $T'_{Ap} = AppPlans(ag_{bs},T_R)$
(SelAppl)	$\frac{S_\mathcal{O}(T_{Ap})=(p,\theta)}{\langle ag,C,M,T,SelAppl\rangle \longrightarrow \langle ag,C,M,T',AddIM\rangle}$	s.t. $T'_\rho = (p,\theta)$
(ExtEv)	$\frac{T_\epsilon=\langle te,\top\rangle, T_\rho=(p,\theta)}{\langle ag,C,M,T,AddIM\rangle \longrightarrow \langle ag,C',M,T,SelInt\rangle}$	s.t. $C'_I = C_I \cup \{[p\theta]\}$
(SelInt$_1$)	$\frac{C_I\neq\{\}, S_\mathcal{I}(C_I)=i}{\langle ag,C,M,T,SelInt\rangle \longrightarrow \langle ag,C,M,T',ExecInt\rangle}$	s.t. $T'_i = i$
(SelInt$_2$)	$\frac{C_I=\{\ \}}{\langle ag,C,M,T,SelInt\rangle \longrightarrow \langle ag,C,M,T,ProcMsg\rangle}$	
(AchvGl)	$\frac{T_i=i[head\leftarrow !at;h]}{\langle ag,C,M,T,ExecInt\rangle \longrightarrow \langle ag,C',M,T,ProcMsg\rangle}$	s.t. $C'_E = C_E \cup \{\langle +!at,T_i\rangle\}$ $C'_I = C_I \setminus \{T_i\}$
(ClrInt$_1$)	$\frac{T_i=[head\leftarrow \top]}{\langle ag,C,M,T,ClrInt\rangle \longrightarrow \langle ag,C',M,T,ProcMsg\rangle}$	s.t. $C'_I = C_I \setminus \{T_i\}$
(ClrInt$_2$)	$\frac{T_i=i[head\leftarrow \top]}{\langle ag,C,M,T,ClrInt\rangle \longrightarrow \langle ag,C',M,T,ClrInt\rangle}$	s.t. $C'_I = (C_I \setminus \{T_i\}) \cup$ $\{k[(head' \leftarrow h)\theta]\}$ if $i = k[head' \leftarrow g;h]$ and $g\theta = TrEv(head)$
(ClrInt$_3$)	$\frac{T_i\neq[head\leftarrow \top] \wedge T_i\neq i[head\leftarrow \top]}{\langle ag,C,M,T,ClrInt\rangle \longrightarrow \langle ag,C,M,T,ProcMsg\rangle}$	

3.2 Jason

Jason is an interpreter for an extended version of the AgentSpeak(L) agent programming language (see section 2.2). This interpreter is written in Java and implements the operational semantics of AgentSpeak(L) [27, 6] and its extensions [28, 1, 35]. The integrated develop enviroment of Jason provides a graphical interface, that allows to edit the configuration file of a MAS and the code of the agents written in AgentSpeak(L). Through of this environment is possible to control the execution of the MAS and distribute the agents over a network in a simple way. Another tool that comes with this environment is a "mind inspector" that allows observing the internal state of the agents in run time. Jason also includes: speech acts based on KQML for communication between agents; annotations of the plans for using selection functions based on decision theory;

selection functions configurable in Java; and mechanisms of extension and use of legacy code, by means of the "internal actions" defined by the user.

3.3 WEKA

WEKA [36] is a tool implemented in Java to perform experiments in a KDD process. It provides methods for preprocessing data, e.g., replacing the missing values and performing discretization on data sets; data mining algorithms, e.g., ID3, C4.5, NB, and TAN; and pattern evaluation algorithms as the stratified cross-validation method. Other algorithms to carry out each one of the tasks of data mining, but we have focused in the methods mentioned above due to our research lines. In addition, WEKA has tools for visualizing and preprocessing data [14]. The input of all algorithms takes the form of a relational table, that can be read from a file or generated through of a database query. The file can be in *csv* format or *arff* format, which is the native format of WEKA.

4 Explorations

This section describes our explorations for the BDI MAS support for the KDD process, from the design to the implementation of the system. Later, some results on the use of the system are reported.

4.1 Design

Figure 3 shows the overview diagram for the MAS implemented for this paper. The *Coordinator* agent perceives the requests for learning from the users of the system as well as the databases associated to the requests; and their format. If a database is in *xls* or *csv* formats, she converts it into *arff* format; in any another case, she asks the user for a database in one of the mentioned formats. Then, the Coordinator asks the *Preprocessing* agent to review the database. If it is not nominal, the former tells the user that the target must be of this kind, given the nature of the learning algorithms used by the agents; otherwise, the Coordinator prints "Class is nominal" and sends a preprocessing require to the Preprocessing agent. This agent replaces the missing values (with the mode when the attributes are nominal and the mean when these ones are numerical) and discretizes the database (supervised and non-supervised). Once Preprocessing informs to Coordinator that the database has been preprocessed, the latter sends learning requests to the *ID3*, *C4.5*, *NB* and *TAN* agents. These ones learn their respective models for both kinds of discretization of data, and inform the Coordinator of their results. The Coordinator selects and reports the winning model given some criteria, e.g., the most accurate model, the fastest result obtained, etc.

To obtain the system overview diagram is necessary to identify the goals and subgoals of the system in a goal overview diagram (System Specification phase). For example, the figure 4 shows the goal and its subgoals for preprocessing the

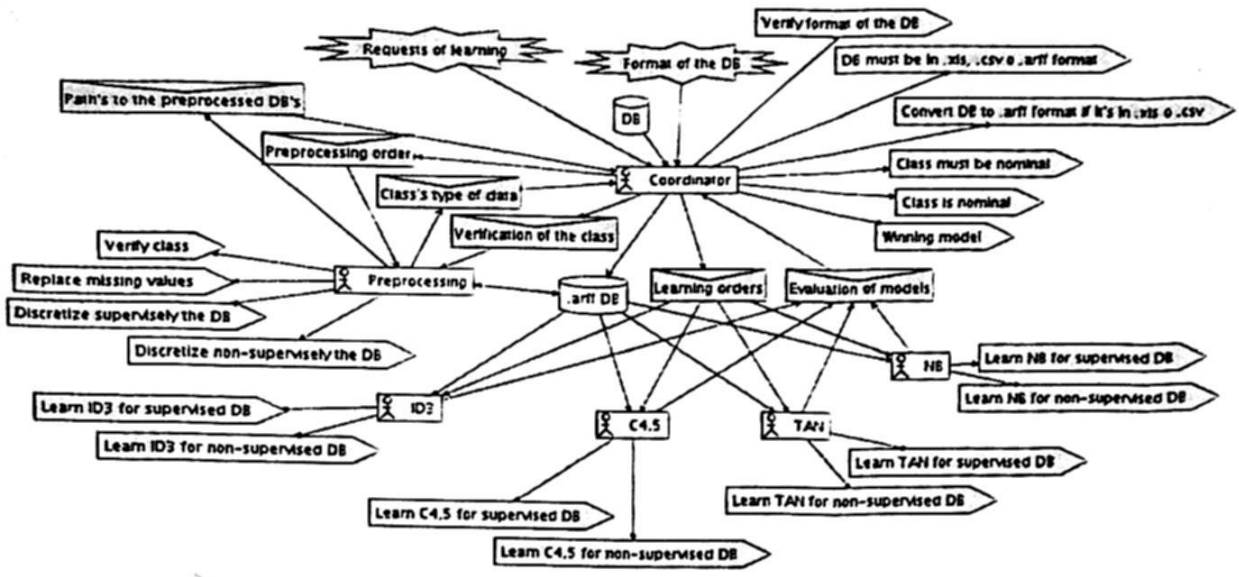


Fig. 3. Overview Diagram of our Multi-Agent System.

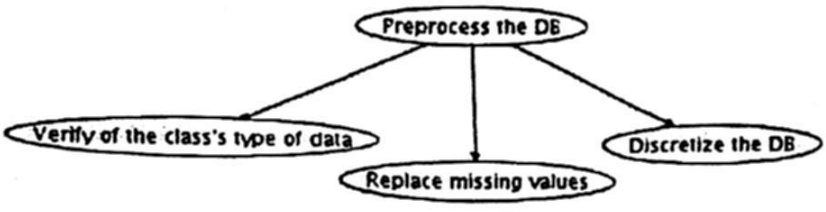


Fig. 4. Goal and its subgoals preprocess the database.

database: one of them verifies what kind the class is, another one replaces the missing values and the last one discretizes this database.

After that, the different possible scenarios for the system were defined in a scenario diagram (System Specification phase). For example, the figure 5 shows the scenario that will take place when the database has to be discretized supervisely, e.g., when the process of discretization takes into account the value of the class to create the different categories of data. It is worth mentioning that the discretization is based on the Minimum Description Length method (MDL) [10].

The roles of the system were created once the goals and scenarios were defined. These roles are formed by clustering similar goals, perceptions and actions in a system role diagram (System Specification phase). For example, the figure 6 shows the "Preprocessing of the DB" role. Observing the scenario shown in figure 5, we can see that both steps of the scenario are associated with this role.

The data coupling diagram and the agent-role grouping diagram (Architectural Design phase) were useful to identify the types of the agents in the MAS. Figure 7 shows the first of these diagrams, with the roles of the system and the data identified in the scenarios. Figure 8 shows the agent-role grouping di-

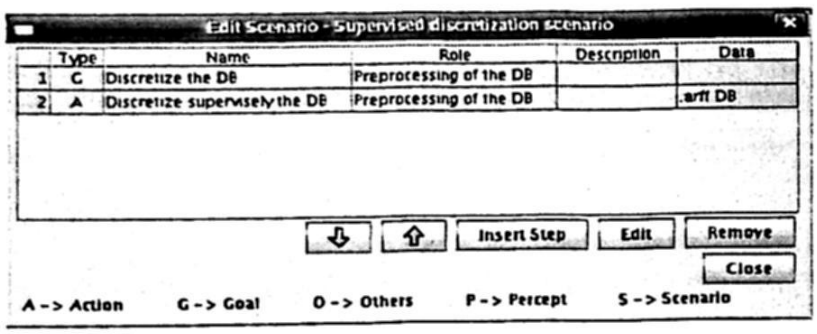


Fig. 5. Supervised discretization scenario.

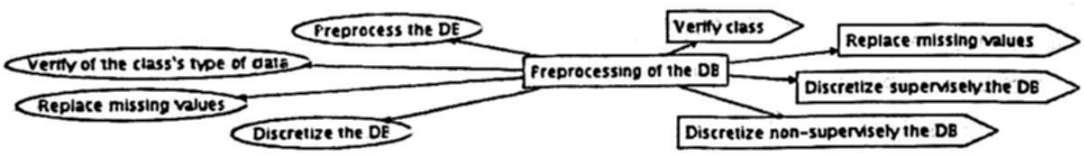


Fig. 6. Preprocessing of the DB role.

agram, which shows the roles assigned for each agent. Thus, the Coordinator is responsible for managing the input/output of the system and the database, etc.

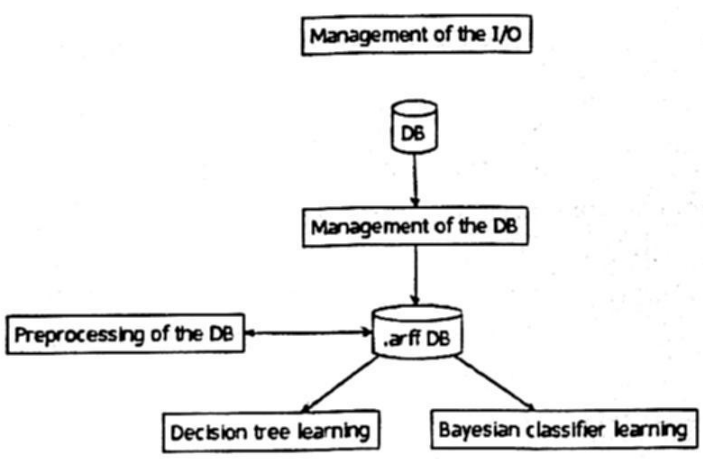


Fig. 7. Data Coupling Diagram.

In the last phase of this process (Detailed Design phase), each agent is refined in terms of its capacities, internal events, plans and structures of data. For example, the figure 9 shows the Preprocessing agent overview diagram, including two plans: "Verify the class" and "Preprocess the DB". The first plan is executed when a request to verify the target class is perceived, then it executes the action that verifies the class and finally it sends the type of data of this target class.

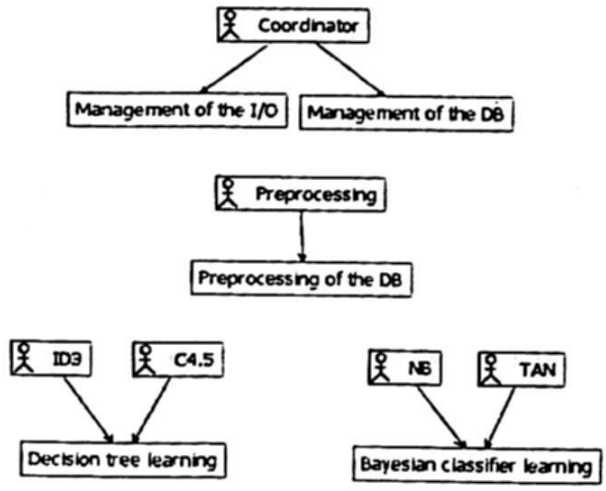


Fig. 8. Agent-Role Grouping Diagram.

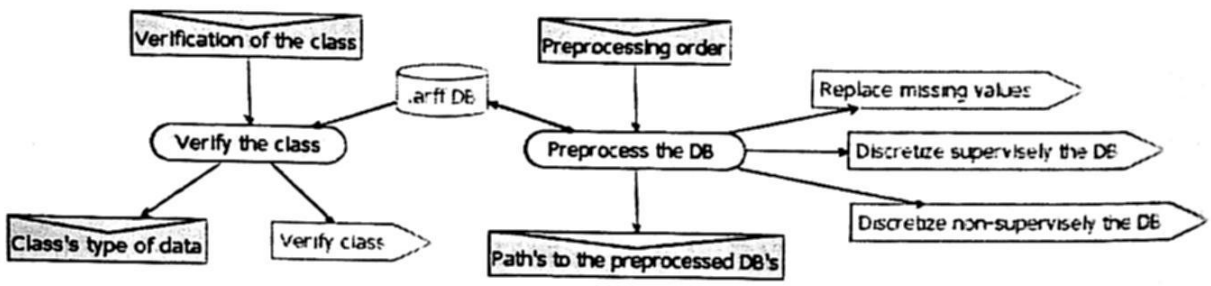


Fig. 9. Preprocessing-Agent Overview Diagram.

4.2 Implementation

As mentioned, the MAS was implemented in AgentSpeak(L)'s interpreter Jason v.1.0 (section 3.2). The development was carried out in a laptop with the following characteristics: Fedora Core 5 as operative system, Intel(R) Pentium(R) M 1.70 GHz processor, 1 Gb in RAM, 80 Gb in Hard Drive.

According to the design described in the previous section, the MAS consists of six agents: *Coordinator*, *Preprocessing*, *ID3*, *C4.5*, *NB* y *TAN*. The features of each one of these agents are shown in the table 3. The code in table 4 shows the initial beliefs and some plans of the Coordinator agent.

Table 3. Features of the agents.

Coordinator	Preprocessing	ID3, C4.5, NB and TAN
Manage I/O	Replace missing values	Learn data mining models
Manage the database	Discretize the database	

Table 4. A fragment of the code for the Coordinator agent

```
1  //Coordinator in project metaclasif.mas2j
2
3  //BELIEFS
4  start. //Start the execution of the MAS
5  file("./DATABASES/Iris.csv").
6
7  //PLANS
8  @pi
9  +start: true
10    <- .print("Coordinator agent, who manages the MAS...");
11       ?file(PathDB);
12       weka.verifyFormatDB(PathDB,Format);
13       .print("DB given in format ",Format);
14       !pVerifyFormatDB(PathDB,Format).
15
16 //Plans take actions depending on what type the database is
17 @pVFBD1
18 +!pVerifyFormatDB(PathDB,Format): not (Format == ".xls" | Format == ".csv" |
19                                        Format == ".arff")
20    <- .print("DB must be given in format .xls, .csv o .arff").
21
22 @pVFBD2
23 +!pVerifyFormatDB(PathDB,Format): Format == ".xls" | Format == ".csv" |
24                                        Format == ".arff"
25    <- !pConviertDB(PathDB,Format);
26       ?file(PathDBArff);
27       .send(preprocessing,achieve,verifyClass(PathDBArff));
28       .wait("+typeClass(TypeClass)");
29       !printTClass.
```

Table 5. Preprocessing Agent.

Plan	Internal Actions	WEKA Classes
@ppBD	replaceMissing	ArffLoader, ArffSaver Instances, Instance Filter, ReplaceMissingValues
	discretizeS	ArffLoader, ArffSaver Instances, Instance Filter, Discretize
	discretizeNS	ArffLoader ArffSaver, Instances Instance, Filter, Filter

The internal actions used to implement the plan library for each agent are built from WEKA classes. Table 5 illustrates this for the Preprocessing agent. For example, the *@ppBD* plan uses the *replaceMissing*, *discretizeS*, and *discretizeNS* internal actions, implemented using methods inherited from the *ArffLoader*, *ArffSaver* classes, etc., provided by WEKA.

5 Tests and Results

The MAS was tested with different databases from the repository at the University of California [2] to observe its viability offering support. Table 6 describes databases used in the tests. Performance, in terms of accuracy, was estimated using the *Stratified 10-fold Cross-Validation*, since it has been reported [24] as the best method to select one model from a set of them. Tables 7 and 8 show the percentage of correctly classified instances for each model learned by the agents, with the databases discretized supervised and non-supervised, respectively.

Table 6. Description of the Databases.

DB's	$\|Instances\|$	$\|Attributes\|$	$\|Classes\|$
Anneal	798	39	5
Balance_Scale	625	5	3
Credit_S	690	16	2
Diabetes	768	9	2
Ecoli	336	9	8
Hypothyroid	3163	26	2
Ionosphere	351	35	2
Iris	150	5	3
Lymphography	148	19	4
Segment	2310	20	7
Soybean	307	36	19
Vehicle	846	19	4
Zoo	101	18	7

The performance results show that there is no such a thing as the "best method" and the MAS automatizes the search for the best classifier using the selected algorithms. More importantly, the MAS avoids getting the "worst method" results, which in some cases (Zoo/ID3 and Ecoli/ID3) is quite relevant. The dynamics of the experiments suggests that the coordinator or another new agent, can be programmed to produce comparative reports, instead of the "the winner is..." behavior. Tables 9 and 10 provide the time taken to build the models learned by the agents, with the data discretized supervised and non-supervised, respectively.

The results of these tables indicate us that for both kinds of data discretization, NB always builds its models in the least time (because it only learns the parameters of the model, since this one is always the same) than the rest of the classifiers, whereas TAN is the one takes the most time to learn its models. In addition, we can see that when the data are discretized non-supervised, in general, the models are built in the least time.

Table 7. Percentage of correctly classified instances obtained with the data discretized supervised.

DB's	%ID3	%C4.5	%NB	%TAN
Anneal	**93.23**	92.98	92.1	92.1
Balance_Scale	70.72	71.04	71.68	**72.32**
Credit_S	77.97	86.96	86.23	**87.25**
Diabetes	75.26	77.73	77.99	**78.12**
Ecoli	0	80.95	**85.42**	85.12
Hypothyroid	98.86	**99.24**	98.58	98.83
Ionosphere	89.17	88.89	90.6	**91.17**
Iris	**95.33**	93.33	94.67	94
Lymphography	79.05	77.7	**83.78**	82.43
Segment	94.85	**95.41**	91.77	92.21
Soybean	91.21	**92.51**	85.99	88.6
Vehicle	**71.63**	70.45	62.53	69.5
Zoo	1.98	94.06	**96.04**	**96.04**
μ	**72.25**	**86.25**	**85.95**	**86.74**

Table 8. Percentage of correctly classified instances obtained with the data discretized non-supervised.

DB's	%ID3	%C4.5	%NB	%TAN
Anneal	89.6	**89.97**	88.47	87.72
Balance_Scale	39.52	65.92	**90.88**	86.72
Credit_S	74.64	**85.8**	84.35	84.64
Diabetes	59.89	72.66	75.52	**76.95**
Ecoli	0	73.21	**83.03**	78.27
Hypothyroid	95.92	**97.53**	96.9	97.03
Ionosphere	85.75	88.6	**90.6**	**90.6**
Iris	91.33	**96**	95.33	91.33
Lymphography	76.35	75	**83.11**	79.05
Segment	91.13	**93.16**	88.96	89.52
Soybean	84.69	**89.25**	80.78	83.71
Vehicle	58.51	71.63	59.93	**72.46**
Zoo	1.98	92.08	90.1	**97.03**
μ	**65.33**	**83.91**	**85.23**	**85.77**

6 Conclusions

This paper presents the design and implementation of a BDI MAS to support the KDD process. More precisely we present a framework that uses methodology and tools proposed in the MAS literature, to approach the automatizing of support for the use of common tools, as WEKA, in the KDD process. The MAS was designed according to the *Prometheus* methodology with the support of the

Table 9. Time taken to build the models learned with data discretized supervisely.

DB's	%ID3	%C4.5	%NB	%TAN
Anneal	0.96	0.91	**0.22**	2.48
Balance_Scale	0.24	0.41	**0.11**	0.16
Credit_S	0.6	0.78	**0.15**	0.53
Diabetes	0.66	0.81	**0.2**	0.42
Ecoli	1.62	1.28	**0.3**	1.65
Hypothyroid	1.57	1.15	**0.33**	3.25
Ionosphere	0.63	0.65	**0.21**	1.47
Iris	0.31	0.17	**0.1**	0.26
Lymphography	0.48	0.46	**0.21**	0.37
Segment	1.04	0.57	**0.16**	2.72
Soybean	0.48	0.2	**0.05**	3.48
Vehicle	0.51	0.44	**0.04**	1.3
Zoo	0.12	0.05	**0.01**	0.53
μ	**0.71**	**0.61**	**0.16**	**1.43**

Table 10. Time taken to build the models learned with data discretized non-supervised.

DB's	%ID3	%C4.5	%NB	%TAN
Anneal	0.34	0.14	**0.02**	0.33
Balance_Scale	0.17	0.11	**0**	0.02
Credit_S	0.15	0.08	**0.01**	0.07
Diabetes	0.2	0.13	**0**	0.06
Ecoli	0.31	0.08	**0**	0.28
Hypothyroid	0.51	0.31	**0.06**	0.56
Ionosphere	0.21	0.06	**0.01**	0.14
Iris	0.06	0.13	**0**	0.04
Lymphography	0.12	0.06	**0**	0.16
Segment	0.4	0.18	**0.02**	0.37
Soybean	0.14	0.13	**0.01**	2.73
Vehicle	0.34	0.11	**0.02**	0.28
Zoo	0.13	0.06	**0.02**	0.29
μ	**0.24**	**0.12**	**0.01**	**0.41**

PDT tool. Some advantages that were found when using Prometheus to model the KDD process like a BDI MAS are:

- The KDD process was clearly understood, because of the goal oriented analysis to define the MAS. The use of diagrams in the design was also helpful.
- The transition between the design and implementation of the BDI MAS was really easy. The diagrams obtained with Prometheus are expressed in terms of beliefs, goals, plans, events, etc., which are the natural constructors for the BDI agents and its AgentSpeak(L) programming language.

- Prometheus can be used to model any process within an organization, since its diagrams offer a high level of abstraction, in the sense that they can be referred in a similar language to ours (for example, through beliefs, goals, scenarios, roles, plans, events, actions, etc.).

The MAS was implemented in the *AgentSpeak(L)* agent programming language, through the *Jason* interpreter. The main point in favor of Jason, regarding other development platforms of MAS (as JACK [7], JAM [21], JADEX [31], among others), is its theoretical basis, which gets to implement the operational semantics of AgentSpeak(L).

Java helps since both WEKA and Jason are written in this language, so that inherit methods in both senses was natural. We have opted for taking methods in WEKA to built internal actions in Jason agents. Future work will explore another possibility: considering WEKA as the environment for the MAS implemented in Jason. This should enable a richer interaction between the agents and the users of WEKA, e.g., an agent can execute directly commands from WEKA, in the same way the user does. Another contribution due to Java is portability. The MAS has been executed successfully on Windows, Linux, Solaris, Mac OS X, all for free.

Our exploration provides the basis to built more elaborated MAS in this context, e.g, extending the number of agents (more learning algorithms adopted) or extending their competences (wiser agents using better the algorithms). A much more ambitious goal is to approach meta-learning by this way, e.g., the agents learn intentionally [17, 18] to become wiser. This kind of learning enables the agents to learn when a given plan is really useful, given the desires and beliefs of an agent, and its past experience, i.e., agents that learn to learn. Given that, such a system would be much more elaborated, we are currently developing formal tools for AgentSpeak(L) program verification [19].

Acknowledgments. The second author is supported by the CONACYT scholarship 197800 and DIP-UJAT (DAIS-02 UJAT-EGRESADA/2005) agreement.

References

1. Ancona, D., Mascardi, V., Hübner, J., Bordini, R.: Coo-AgentSpeak: Cooperation in AgentSpeak through plan exchange. In Nicholas R. Jennings, Carles Sierra, Liz Sonenberg, and Milind Tambe, editors, Proceedings of the Third International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS-2004), pages 698-705. New York, NY (2004)
2. Asuncion, A., Newman, D.: UCI Machine Learning Repository. Irvine, CA: University of California, Department of Information and Computer Science. http://www.ics.uci.edu/ mlearn/MLRepository.html [Consulted: friday, June 27, 2007].
3. Bailey, S., Grossman, R., Sivakumar, H., Turinsky, A.: Papyrus: A System for Data Mining Over Local and Wide Area Clusters and Super-clusters. In Proceedings of the 1999 ACM/IEEE conference on Supercomputing, page 63. ACM Press, Portland, OR (1999)

4. Bordini, R., Dastani, M., Dix, J., El Fallah Seghrouchni, A.: In Multi-Agent Programming: Languages, Platforms and Applications. Springer, New York (2005)
5. Bordini, R.H., Hübner, J.F., Wooldridge, M.: Programming Multi-Agent Systems in AgentSpeak using Jason. Wiley, England (2007)
6. Bordini, R., Moreira, A.: Proving BDI properties of agent-oriented programming languages: The asymmetry thesis principles in AgentSpeak(L). Annuals of Mathematics and Artificial Intelligence, 42(1-3):197-226. Special Issue on Computational Logic in Multi-Agent Systems, September (2004)
7. Busetta, P., Rönnquist, R., Hodgson, A., Lucas, A.: JACK Intelligent Agents - Components for Intelligent Agents in Java. Technical report, Agent Oriented Software Pty. Ltd, Melbourne, Australia (1998)
8. d'Inverno, M., Luck, M.: Engineering AgentSpeak(L): A formal computational model. Journal of Logic and Computation, 8(3). 233-260 (1998)
9. Dunham, M.: Data Mining: Introductory and Advanced Topics. Prentice Hall PTR (2002)
10. Fayyad, U., Irani, K.: Multi-interval discretization of continuous-valued attributes for classification learning. Proceedings of the 13th International Joint Conference on Artificial Intelligence, 1022-1027. Morgan Kaufmann, San Francisco, CA (1993)
11. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases. AI Magazine, 17(3):37–72 (1996)
12. Fayyad, U., Piatetsky-Shapiero, G., Smyth, P.: Knowledge Discovery and Data Mining: Towards a Unifying Framework. Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD-96), Portland, Oregon (1996)
13. Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From Data Mining to Knowledge Discovery: An Overview. In AKDDM, AAAI/MIT Press (1996)
14. Frank, E., Hall, M., Holmes, G., Kirkby, R., Pfahringer, B., Witten, I., Trigg, L.: WEKA - A Machine Learning Workbench for Data Mining. In O. Maimon and L. Rokach, editors, The Data Mining and Knowledge Discovery Handbook, pages 1305-1314. Springer (2005)
15. Georgeff, M. P. and Lansky, A. L.: 'Procedural knowledge. In Proceedings of the IEEE, (74)10:1383–1398 (1986)
16. Guerra-Hernández, A., Cruz-Ramírez, N., Mondragón-Becerra, R.: Exploraciones sobre el soporte Multi-Agente en Minería de Datos. Conferencia Internacional en Información, Comunicación y Diseño. Universidad Autónoma Metropolitana, Cuajimalpa, México (2006)
17. Guerra-Hernández, A., El-Fallah-Seghrouchni, A., Soldano, H.: Learning in BDI Multi-agent Systems. In: Dix, J., Leite, J. (eds.) CLIMA IV. LNCS, vol. 3259, pp. 218–233. Springer, Heidelberg (2004)
18. Guerra-Hernández, A., Ortíz-Hernández, G.: Toward BDI sapient agents: Learning intentionally. In: Mayorga, R.V., Perlovsky, L.I. (eds.) Toward Artificial Sapience: Principles and Methods for Wise Systems, pp. 77–91. Springer, London (2008)
19. Guerra-Hernández, A., Castro-Manzano, J.M., El-Fallah-Seghrouchni, A.: Toward an AgentSpeak(L) theory of commitment and intentional learning. In: LNAI 5317:848–858. Springer Verlag, Berlin Heidelberg (2008)
20. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers (2000)
21. Huber, M.: JAM: A BDI-Theoretic Mobile Agent Architecture. In Proceedings of the Third Annual Conference on Autonomous Agents, pp. 236-243, O. Etzioni, J. P. Müller, J. Bradshaw, editors, ACM Press (1999)
22. Kargupta, H., Hamzaoglu, I., Stafford, B.: Scalable, Distributed Data Mining Using

An Agent Based Architecture. In David Heckerman, Heikki Mannila, Daryl Pregibon, and Ramasamy Uthurusamy, editors, Proceedings of Knowledge Discovery And Data Mining, pages 211-214. AAAI Press, Menlo Park, CA (1997)
23. Kargupta, H., Park, B., Hershberger, D., Johnson, E.: Collective Data Mining: A New Perspective Towards Distributed Data Mining. In Hillol Kargupta and Philip Chan, editors, Advances in Distributed and Parallel Knowledge Discovery, pages 133-184. MIT/AAAI Press (2000)
24. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence. Morgan Kaufmann, San Mateo, CA (1995)
25. Mascardi, V., Demergasso, D., Ancona, D.: Languages for Programming BDI-style Agents: an Overview. In F. Corradini, F. De Paoli, E. Merelli, and A. Omicini, editors, WOA 2005 - Workshop From Objects to Agents, pages 9-15 (2005)
26. Mondragón-Becerra, R.: Exploraciones sobre el soporte Multi-Agente BDI en el Proceso de Descubrimiento de Conocimiento en Bases de Datos. Master's thesis submitted at the Universidad Veracruzana, December (2007)
27. Moreira, A., Bordini, R.: An operational semantics for a BDI agent-oriented programming language. In John-Jule Ch. Meyer and Michael J. Wooldridge, editors, Proceedings of the Workshop on Logics for Agent-Based Systems (LABS-02), held in conjunction with the Eighth International Conference on Principles of Knowledge Representation and Reasoning (KR2002), pages 45-59. Toulouse, France (2002)
28. Moreira, A., Vieira, R., Bordini, R.: Extending the operational semantics of a BDI agent-oriented programming language for introducing speech-act based communication. In Proceedings of the Workshop on Declarative Agent Languages and Technologies (DALT-03), held with AAMAS-03, Melbourne, Australia (2003)
29. Padgham, L., Thangarajah, J., Winikoff, M.: Tool support for agent development using the prometheus methodology. In Fifth International Conference on Quality Software (QSIC 2005), pages 383-388. IEEE Computer Society, Melbourne, Australia (2005)
30. Padgham, L., Winikoff, M.: Prometheus: A Methodology for Developing Intelligent Agents. Proceedings of the Third International Workshop on Agent-Oriented Software Engineering, at AAMAS, Bologna, Italy (2002)
31. Pokahr, A., Braubach, L., Lamersdorf W.: Jadex: A BDI reasoning engine. In R. Bordini, M. Dastani, A. Seghrouchni, and J. Dix, editors, Multi-Agent Programming. Kluwer (2005)
32. Prodromidis, A., Chan, P., Stolfo, S.: Meta-learning in Distributed Data Mining Systems: Issues and Approaches. In Hillol Kargupta and Philip Chan, editors, Advances of Distributed Data Mining. MIT/AAAI Press (2000)
33. Rao, A.S.: AgentSpeak(L): BDI agents speak out in a logical computable language. In: de Velde, W.V., Perram, J.W. (eds.) MAAMAW. LNCS, vol. 1038, pp. 42–55. Springer, Heidelberg (1996)
34. SPSS. *Clementine*. http://www.spss.com/clementine/ [Consulted: thursday, May 29, 2008].
35. Vieira, R., Moreira, A., Wooldridge, M., Bordini, R.: On the formal semantics of speech-act based communication in an agent-oriented programming language. Submitted article to appear (2005)
36. Witten, I., Frank, E.: Data Mining: Practical machine learning tools and techniques. 2nd Edition, Morgan Kaufmann, San Francisco (2005)
37. Wooldridge, M., Jennings, N.: Intelligent Agents: Theory and Practice. In Knowledge Engineering Review 10(2) (1995)

Neural Networks, Image Processing, and Scheduling

Enhancement Color Method by Luminance Modulation

Hayde Peregrina-Barreto[1], J. Gabriel Avina-Cervantes[1], Jose J. Rangel-Magdaleno[1], Sergio Ledesma-Orozco[1], and Mario Alberto Ibarra-Manzano[2]

[1] University of Guanajuato.
Facultad de Ingeniería Mecánica, Eléctrica y Electrónica.
Carr. Salamanca-Valle de Santiago km. 3.5 + 1.8km.
Comunidad de Palo Blanco. Salamanca, Guanajuato. 36730. México.
peregrina.barreto.hayde@gmail.com, avina@salamanca.ugto.mx
jjrangel@hspdigital.org, {selo, ibarram}@salamanca.ugto.mx
http://www.ugto.mx

[2] Laboratoire d'Analyse et d'Architecture des Systèmes
7, Avenue du Colonel Roche, 31077 Toulouse Cedex 4, FRANCE
maibarra@laas.fr
http://www.laas.fr/laas/

Abstract. Processing outdoor images with low saturated colors is a very complicated task for color based algorithms. Areas such as color segmentation and object classification require to distinguish clearly among regions with dissimilar chromatic features. Some problems with this chromatic discrimination in outdoors images, make it necessary to apply a method for color balancing and image enhancing. In this work, it is proposed an innovative color enhancement method which is based on color modulation; modulation is a function of the original image luminance and it is automatically computed from the same image statistics. This technique has two main advantages: no false colors are created nor oversaturation is presented, for poorly saturated images. In addition, experiments have shown the reliability of the proposed approach.

Keywords: Color enhancement, segmentation, outdoor scenes.

1 Introduction

Color is an important feature for some pattern recognition tasks, e.g., color segmentation, object classification and tracking algorithms. Sometimes images present faded colors due to variable conditions as lighting sources or acquisition devices. Results are not accurate when color segmentation is applied to uncalibrated or low satured color images; because chromatic differences among regions are small. Natural images exhibit rich and complex structure, their colored nature is determined by the physical and geometric properties depending on illumination, reflections and imaging on the scene; color permits to distinguish details that in the original image are not clearly defined. Image enhancement is basically a technology to improve image quality in terms of visual perception of an expert, i.e. human being [1].

Color enhancement algorithms solve these drawbacks and many of them are based on contrast improvement [2]. Therefore, it appears that color enhancement can become

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 241-250

Received 01/07/08
Accepted 01/08/08
Final version 11/09/08

an important tool for improving the acquisition of powerful color based descriptors for machine vision applications [3]. Image understanding requires segmenting (low level task) a scene into a set of meaningful regions, that is partitioning a natural color image into a set of perceptually color-uniform regions. In this case, if segmentations fails all the further stages could produce unfit results. Limited color constancy strongly affects algorithms, e.g. in outdoor robotics, color must be carefully used and reinforced with some other kind of features (texture, shape, and so on). By the way, in the case of assessing skin lesions in medicine, the expert has to provide a diagnosis whether skin lesion is a serious pathology. For instance he could trust a color analysis by image processing techniques to detect skin lesions such as melanomas. Therefore, color quality enhancement is a fundamental factor for many algorithms based on chromatic features [1]. Nowadays, many techniques for improving image appearance are available and some of them are focused on project details. However, they have important disadvantages: original color changes [4] and lose real appearance of the scene [5]. The main characteristic of the proposed approach [6] is that it permits enhance color to distinguish more easily the difference between regions; all of this without changing the original chromatic information [7].

This paper is organized as follows: a brief state of art in color enhancement is presented in section 2, our approach is presented in section 3, some preliminary results are presented in section 4 and finally the conclusions of this work are discussed in section 5.

2 State of the Art

Image details (e.g. textures) are more evident when the contrast is high (the difference between the smallest and biggest colored pixel in the image), but if contrast is modified, then generally the original color is also altered. Techniques of color enhancement available of improving image appearance (weakly affecting the original chromatic representation) are needed. Moreover, techniques based on histogram modification are frequently used for color enhancement [8]. This kind of techniques modifies the histogram distribution to get an improved image, and good results are obtained in clarity or contrast, but they are not adequate for color processing. Generally, these methods modify each RGB color component without taking into account the correlation among them: this substantially affects the results and the original colors [9].

A color calibration can also be taken further, to ensure the correct analysis (or reproduction) of color as well as intensity. Some other color enhancement techniques have been proposed: based on histogram equalization, contrast enhancement techniques, adaptive neighborhood histograms equalization method and 3D equalization methods in the RGB color space. Some other methods exploit the correlation of luminance and saturation color component of the image locally. Besides some authors exploit genetic algorithms to cope with color enhancement as an optimization problem, as well as multispectral approaches where adaptive strategies for wavelet based image enhancement have also been proposed [1].

Histogram equalization is frequently used for color improvement; in practice this method consists on creating an accumulated histogram, which determines the quantity to be affected for each pixel. In that way, the new histogram is distributed along the

desired dynamic rank of pixels. However, the new distribution modifies the value of original pixels assigning them a new color [10], this does not always give good results and object contours are not naturally modified [11]. Displacing histogram is another technique used to make a new image lighter or darker, and it is also based on a arbitrary change of values. Both techniques, equalization and histogram displacing, could be useful in many situations but not in color enhancement or critical color based applications, due to the fact they do not preserve original colors nor maintain color appearance [12] [13]. Retinex is another technique for improving image aspect. It is frequently used when it is not possible to distinguish important details on image. A noticeable improvement is obtained in oversaturated images (Fig. 1). However, it does not work so well with enhance color as it is shown further on.

Usually, some images may require some kind of color enhancement, even if they have been taken by special devices [14]. However, for some applications a short processing time is fundamental and complex algorithms for color enhancement (e.g. multispectral approaches) could be inadequate [15]. Taking into account real-time applications, a technique able to cope with low saturated color problem on images and improve color appearance is proposed in this work. This approach exploits luminance modulation to increase saturation on natural images.

(a) Original image (b) Retinex

Fig. 1. Enhance image using retinex method.

3 Proposed Algorithm

3.1 Image Evaluation

Before applying color enhancement, it is necessary to detect if an image needs to be enhanced, otherwise enhancing an image with good color quality (good saturation) will

turn out to be in a new oversaturated and unnatural colored image (Fig. 2). It was defined a parameter, based on the image statistics, to make a decision whether color enhancement must be applied or not. The parameter is the color definition of saturation; we observed that images with good color quality have a saturation factor above of 0.55. On the contrary, images requiring color enhancement have saturation below the factor. In this way, this threshold was established as condition to determine if the color enhancement algorithm must be applied.

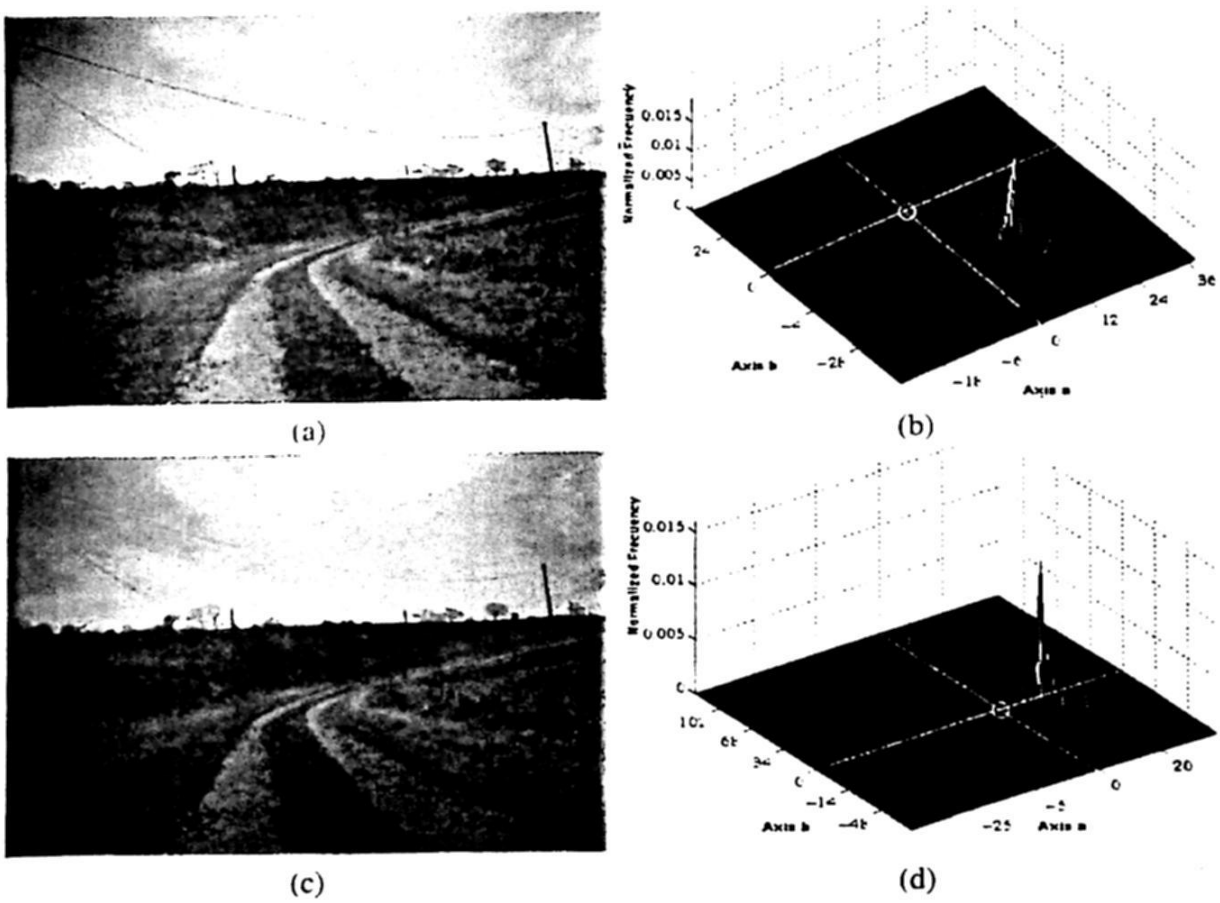


Fig. 2. Oversaturation effect (a)original image (b) output image

3.2 Enhancement Algorithm

The proposed color enhancement method is based on image luminance like modulator element, since luminance (L^3) is closely associated with color intensity. L is taking into account to determine how much the original pixel value must be increased or decreased. Three enhance coefficients ($Coef(R, G, B)$) are calculated, dividing the value of each color band ($BandIn(R, G, B)$) between the luminance of actual pixel ($L(i)$). The biggest coefficient is chosen as the enhance factor ($Efactor$) and dividing each

[3] luminance vector in CIELab space.

coefficient, obtaining enhance quantities. Finally, the original values for each band are multiplied by their corresponding enhance quantities (Ec. 1). The process is applied to each pixel of image.

$$\begin{aligned} Coef(R,B,G) &= BandIn(R,G,B)/L \\ Efactor &= Max(CoefR, CoefG, CoefB) \\ BandOut(R,G,B) &= BandIn(R,G,B) * (Coef(R,G,B)/Efactor) \end{aligned} \tag{1}$$

3.3 Luminance Adjust

Subsequently, an image is adjusted in luminance to enhance color, it means in color intensity, with the objective to get a better distribution in luminance. This process (Algorithm 1) consists on adjusting only the intensity, not the color; improving in this way the aspect of previously enhanced image. New RGB values must be switched to CIELab space, because it is necessary to work with luminance vector. The process consists on a single adjustment over each pixel, taking into account maximum ($Lmax$) and minimum ($Lmin$) values of L, obtaining the quotient between the differences since $L(i)$ to $Lmin$ and since $Lmax$ to $Lmin$. This means, the adjustment factor is determined based on distance between maximum and minimum, and current pixel value. Later, this result must be affected by the maximum desired value of L, less the minimum desired value of L. Involving all luminance rank, the maximum value is 100 and the minimum is 0. Finally, minimum desired is added at previous result and in this way, it is possible to make a luminance correction.

Algorithm 1: Luminance adjustment.

```
Switch the new RGB values to CIELab space.
Determine maximum value of luminance: Lmax
Determine minimum value of luminance: Lmin
for i=1 until i=TotalPix
  L(i) = (L(i)-Lmin) / (Lmax-Lmin))*100
  if L(i) > 100 then L(i)=100 end if
  if L(i) < 0   then L(i)=0 end if
end for
Switch from CIELab space to RGB space.
```

3.4 RGB Adjust

For luminance adjustment it is necessary to transform again the values from CIELab to RGB space. It must be taken into account that not all values CIELab can be represented on RGB space; it could happen that the values exceed the valid RGB rank (0, 255). This must be avoided because false colors are not desired nor changes on the original information. Again, value adjustment is necessary, but now in RGB space. This avoids producing new colors in the output image, taking into account that we pretend

to project actual colors, not new colors. In this way, to adjust existing RGB values it is necessary to calculate the maximum values ($Rmax$, $Gmax$, $Bmax$) and the minimum values ($Rmin$, $Gmin$, $Bmin$) for each band, because based on them, the new correction will be made. Maximums and minimums must be compared, and at the same time, they determine the total maximum and minimum ($Tmax$, $Tmin$). To obtain new desired maximums, it is calculated the quotient of maximum of each band and $Tmax$, and after it is multiplied by the maximum desired value ($MaxDes$). New minimums result from the differences between the minimum of each band and total minimum $Tmin$, added with minimum desired ($MinDes$) and multiplied by the quotient by the same elements. Later, it is made the RGB correction (Algorithm 3), which is similar to luminance adjustment, with the difference that it is applied to each band. The values must not be greater than 255 neither less 0, reason why a verification is necessary. The process described is shown next.

Algorithm 2: RGB adjustment.

```
Determine Rmax, Gmax and Bmax
Determine Rmin, Gmin and Bmin
Tmax=Max(Rmax, Gmax, Bmax)
Tmin=Min(Rmin, Gmin, Bmin)
R_Max=MaxDes*Rmax/Tmax
G_Max=MaxDes*Gmax/Tmax
B_Max=MaxDes*Bmax/Tmax
if Tmin >= 0 then
  R_Min = (Rmin-Tmin+MinDes)*Rmin/Tmin
  G_Min = (Gmin-Tmin+Mindes)*Gmin/Tmin
  B_Min = (Bmin-Tmin+Mindes)*Bmin/Tmin
end if
for i=1 until i=TotalPix
   N_R(i)=((R(i)-Rmin)/(Rmax-Rmin))*(R_Max-R_Min)+R_Min
   N_G(i)=((G(i)-Gmin)/(Gmax-Gmin))*(G_Max-G_Min)+G_Min
   N_B(i)=((B(i)-Bmin)/(Bmax-Bmin))*(B_Max-B_Min)+B_Min
   if N_R(i) > 100 then N_R(i)=100 end if
   if N_R(i) < 0   then N_R(i)=0   end if
   if N_G(i) > 100 then N_G(i)=100 end if
   if N_G(i) <  0  then N_G(i)=0   end if
   if N_B(i) > 100 then N_B(i)=100 end if
   if N_B(i) < 0   then N_B(i)=0   end if
end for
```

4 Experimental Results

Next, the results of color enhancement in outdoor images are shown. The image must be analyzed before any color improvement; this avoids applying enhance color process at an image with a good saturation level. For example, Fig. 2 (a) does not need color enhancement and its shows a blue dominant color visible in the lightest region; if analysis

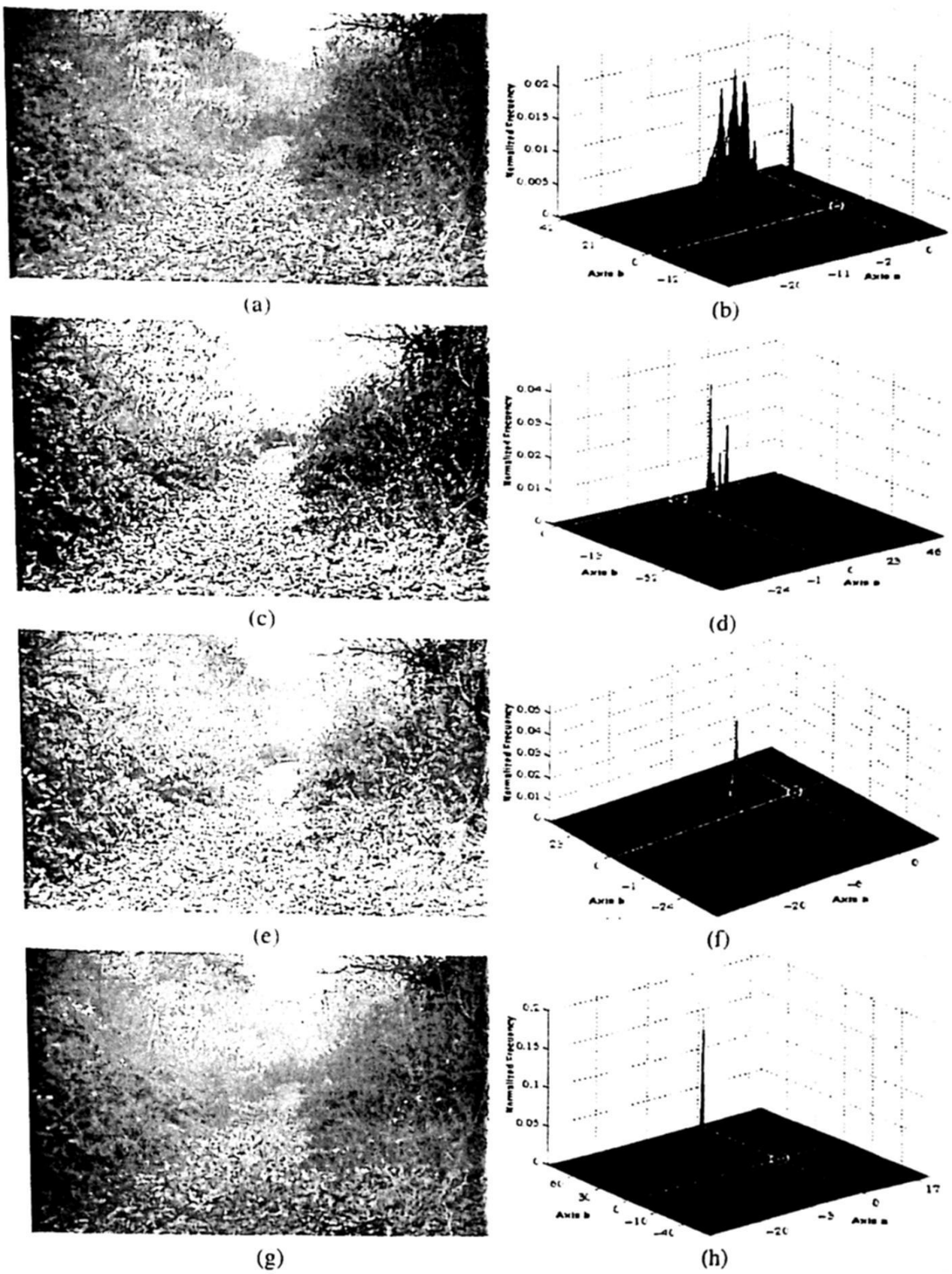


Fig. 3. Color enhancement (a) original image (c) by equalization (e) by retinex (g) by luminance

is omitted, the result is an oversaturated image (Fig. 2 (b)) with an unnatural and excessive color appearance. The histogram center of an image with good saturation is remote of coordinate (0, 0), the central point is the brightest point; it means that histograms very close to (0, 0) have faded colors and histograms very far from it are oversaturated. Neither previous conditions are desired, due to the fact the objective is to obtain a balanced color image with good saturation. As it can be observed histogram of Fig. 2 (b), it is remote to central point, it means image has good saturation level. In this way with the previous analysis this image does not need a color enhance. If the analysis is omitted, then resulting image is oversaturated (Fig. 2 (c)) for which, the histogram is farther from central point. It is not the expected result of a good enhance color method. Otherwise, as we can observe on Fig. 3 (a), the image has good clarity but colors are subdued and it is difficult to distinguish clearly between contiguous regions. Fig. 3 (b)) shows the histogram close to central point, and it is the reason of the grayish and subdued appearance of image.

(a) Outdoor image 3.

(b) Equalization result.

(c) Retinex result.

(d) Color enhancement by luminance

Fig. 4. Outdoor shadowed image

Previously, it was said that equalization is frequently used for improve the image contrast. However, when color improve is desired it is not convenient by reasons mentioned before. Fig. 3 (c) shows the result of equalization and, as it can be observed, the contrast is higher, since it is possible to distinguish more details that in the original image could not be observed. Color is more intense, but the tree zone at the end of road has been loss; it means lost of chromatic information and natural appearance.

Retinex method generally produced grayish colors and lighter image (Fig. 3 (e)), that is not the goal in this case. Previously, it was mentioned that retinex has good results with dark images; it enhances the appearance because it permits to observe details that in the original image are not noticeable. The main disadvantage is the result of low color image, so it is not a good technique for color enhancement.

Fig. 3 (g) shows the result of applying proposed color enhancement by the luminance method. Changes are more notorious and the image looks more attractive (hot colors). The road zone can be clearly distinguished between bushes, such as leaves color are projected among branches. The image was improved in color and false colors were not created; so color was enhanced and the original chromatic structure was maintained. The histogram of the new image (Fig. 3 (h)) shows clearly the difference obtained with the color enhancement method. As it can be observed, it exists a noticeable saturation enhance due to the histogram is now further of central point and this causes than colors are more intense. It implies a color difference reflected on the image. Thus, selection of R band on lighter region in original image, has been balanced with others bands G and B. Other results are shown for outdoor images in Fig. 4.

5 Conclusions

This work was intended to improve image quality with low saturated color, since this preprocessing makes it easier to apply a color based method as well as color segmentation or color classification. Image enhancing by the proposed method can help to improve the results of processes that use color like main feature. This is possible due to increase a better region discrimination with powerful descriptors. Proposed method is based on simple operations, thereby, avoiding an extensive processing and the result is a reliable method; it makes possible a noticeable chromatic enhancement, taking image luminance as the principal element to modulate original color. Keeping original colors appearance is one of the main advantages in the proposed method. Furthermore, the proposed algorithm determines when a color enhancement is required in the image based on the saturation parameter.

References

1. Ding Xiao and Jun Ohya. Contrast Enhancement of Color Images Based on Wavelet Transform and Human Visual System, *Proc. of the IASTED International Conference GRAPHICS AND VISUALIZATION IN ENGINEERING*, January 3-5, 2007.
2. N. W. Campbell, B. T. Thomas and T. Troscianko. Automatic segmentation and classification of outdoor images using neural networks. *International Journal of Neural Systems*, 8(1):137-144, February 1997.

3. C. Fernández - Maloigne, D. Laugier and C. Boscolo. Detection of apples with texture analyze for an apple picker robot. *Intelligent Vehicles '93 Symposium*, pages 323-328, July 1993.
4. J. Huang and D. Mumford. Statistics of natural images and models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(Fort Collins, CO):541-547, December 1999.
5. E. Salvador, A. Cavallaro and T. Ebrahimi. Shadow identification and classification using invariant color models. *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*, volume 3, pages 1545-1548, Salt Lake City, Utah, May 2001.
6. J. W. Hsien, W. F. Hu, C. J. Chang and Y. S. Chen. Shadow elimination for effective moving object detection by Gaussian shadow modeling. *Image and Vision Computing*, 21(6):505-516, June 2003.
7. A. Torralba, K. P. Murphy, W. T. Freeman and M. A. Rubin. Context based vision system for place and object recognition. *Ninth IEEE International Conference on Computer Vision*, volume 1, pages 273-280, Nice, France, October 2003.
8. T. M. Strat and M. A. Fischler. Context based vision: Recognizing objects using information from both 2-d and 3-d imagery. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(10):1050-1065, October 1991.
9. G. Finlayson, S. Hordley, G. Schaefer and G. Y. Tian. Illuminant and device invariant colour using histogram equalization. *Pattern recognition*, 30(2):179-190, February 2005.
10. B. Funt and F. Ciurea. Retinex in matlab. *Journal of Electronic Imaging*, 13(1):48-57, January 2004.
11. A. Rizzi, C. Gatta and D. Marini. A new algorithm for unsupervised global and local color correction. *Pattern Recognition Letters*, 24(11):1663-1677, July 2003.
12. G. M. Bianco and A. Rizzi. Chromatic adaptation for robust visual navigation. *Advanced Robotics*, 16(3):217-232, May 2002.
13. C. Poynton. The rehabilitation of gamma. *SPIE Conference on Human Vision and Electronic Imaging II 1*, 3299:232-249, 1998.
14. D. Merrill. The next generation digital camera. *Optics and Photonics News*, pages 26-34, January 2003.
15. T. Sakamoto, C. Nakanishi and T. Hase. Software pixel interpolation for digital still cameras suitable for a 32-bit mcu. *IEEE Transactions on Consumer Electronics*, 44(4):1342-1352, November 1998.

Artificial Neural Networks for Diagnosing Stator Induction Motor Faults

Pablo Serrano[1], Antonio Zamarrón[1], Arturo Hernandez[2], and Alberto Ochoa[3]

(1) Instituto Tecnológico de León; Guanajuato México
(2) Centro de Investigación en Matemáticas; Guanajuato, México
(3) Instituto de Ingeniería y Tecnología. Universidad Autónoma de Ciudad Juárez; México
E - mail: *jpsr9@hotmail.com*

Abstract. Stator Winding Fault can be detected by monitoring any abnormality of the Park's spectrum. In this paper is presented a fault-detection performance comparison between the Support Vector Machine (SVM) and backpropagation algorithm (BP) using experimental data for a healthy and faulty case. Support Vector Machine and Backpropagation Algorithm provide environments to develop fault-detection schemes because of their multi-input-processing and its good generalization capability. The training patterns are obtained using motor current signature analysis (MCSA) and using Spectral Park's Vector. The neural networks are evaluated by means of the cross-validation technique to determine easily the diagnosis and severity of turn-to-turn faults.

Keywords: Artificial Neural Networks. Faults Diagnosis. Induction Motor.

1 Introduction

At the present time the induction motor has a multiplicity of applications in the human life. Induction Motors applications are presented in different processes in the industry. However, the induction motors, as other machines. can fault during operation. One of the most important faults presented in induction motors is turn-to-turn short-circuits. Degradation of winding insulation can lead to these faults, starting a process that can progress to severe phase-to-phase or turn-to-ground faults. The investigation presented in this paper promotes induction motor preventive maintenance. This paper is organized as follows. Section II discusses about Artificial Neural Networks particularly backpropagation algorithm and support vector machine. Section III briefly describes the fundamental properties of the Park transformation complex vector. Section IV shows the motor-data specifications and the measurement and analysis data. Section V presents the fault-detection schemes and the experimental results. Conclusions are presented in section VI.

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 251-262

Received 23/06/08
Accepted 04/08/08
Final version 17/09/08

2 Artificial Neural Networks (Backpropagation Algorithm)

Stator winding fault diagnostic is essentially a classification problem in pattern space. The artificial neural networks (ANN) can be used to classify patterns of a motor in regular and fault condition. The ANN is a massively parallel processor made up of simple processing units, which has a natural propensity for storing experimental knowledge and making it available for use [1]. There are many neural networks models; however the Backpropagation (BP) networks are simple in structure and stable in operation [2]. Neural Networks based on Backpropagation algorithm have been successfully used for pattern recognition and nonlinear mapping. The BP is a supervised learning method, and is an implementation of the Delta Rule; in this algorithm are calculated desired outputs for any given input. The BP network is structured by hide layers which are capable of classifying an arbitrary region of multidimensional space. A three-layer BP networks is presented in the figure 1 where n is the input layer node number as a set of input data (x_1, x_2,..., x_n), m is the output layer node number, i is the hidden layer node number pattern number, and w_{ji} are the weights between input layer node and hidden layer node.

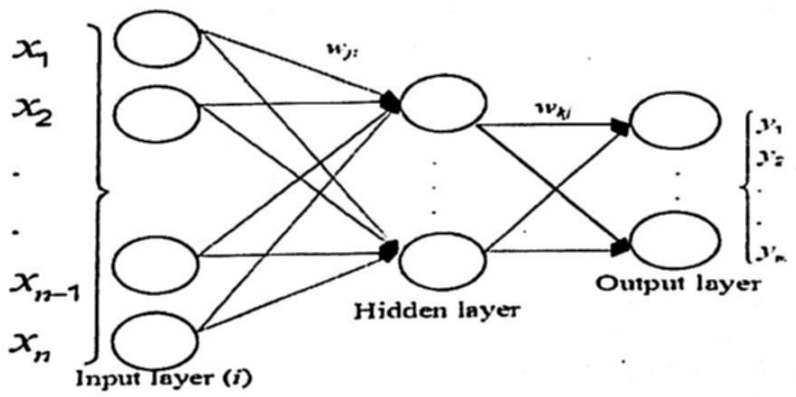


Fig. 1. A three-layer BP networks.

In mathematical terms, the input Net_j is obtained as following:

$$Net_j = \sum_{i=1}^{n} \left(x_i * w_{ji} \right) \tag{1}$$

The output $Output_j$ of layer node input is obtained by the activation function

$$Output_y = f\left(Net_j \right) \tag{2}$$

In this paper was used the bipolar sigmoid function

$$f(x) = \frac{2}{1+e^{-2*x}} - 1 \tag{3}$$

The Mean Square Error (MSE) is used to calculate the total error in training patterns.

$$MSE = \frac{1/2 \sum_{n=1}^{N} \sum_{k=1}^{K} (y_{kn} - \hat{y}_{kn})^2}{N \cdot K} \tag{4}$$

where y_{kn} is the target output of the pattern n, $\hat{y}_{kn}$ is the actual output of the neuron k at output layer for pattern n, N is the number of patterns, and K is the number of output neurons.

3 Support Vector Machine (SVM)

An SVM is a method for separating clouds of data in the feature space F using an optimal hyperplane [3]. Considering a training set $D = \{(x_1, y_1), (x_2, y_2), ..., (x_l, y_l)\}$ with input data $x_i \in R^N$ and corresponding binary class label $y_i \in \{+1, -1\}$, the data can be classified by means the SVM classifier. In the general case, the SVM classifier is [4]:

$$f(x) = w^T \Phi(x) + b \tag{5}$$

Where w^T is an m-dimensional vector, $\Phi(x)$ is a nonlinear function, and b is a scalar. Data points are mapped by means of a kernel with the purpose of searching a maximal separation between classes. The kernel $K(\cdot, \cdot)$ corresponds to an inner product of vector in the higher dimensional feature space if and only if Mercer's condition is met [3]. In this paper, we used a polynomial kernel, which is described as:

$$K(x, z) = (\langle x, z \rangle)^m \quad \text{With } m \in \mathbb{N} \tag{6}$$

An example of the SVM is presented in the figure 2.

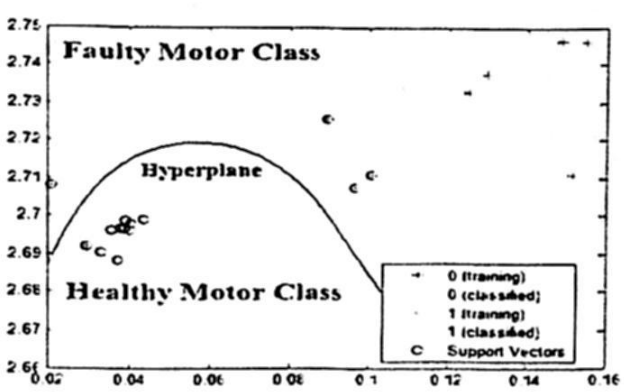


Fig. 2. Example of classification regions of a SVM for a healthy motor and faulty case.

4 Stator Current Complex Vector

Stator current complex vector can be represented as a Park's vector. The Park vector is generally used to carry out a simplified analysis of three-phase stator motor variables. It consists on a two dimensional representation that describes tree-phase induction motor phenomena [3]. In mathematical terms the current complex vector is [3], [4]:

$$\hat{I} = \frac{2}{3}\left\{ i_a + \left(-\frac{1}{2} + j\sqrt{\frac{3}{2}} \right) i_b + \left(-\frac{1}{2} - j\sqrt{\frac{3}{2}} \right) i_c \right\} \tag{7}$$

Where i_a, i_b, i_c are stator currents. Therefore Park's vector has two components which are:

$$i_d = \frac{2}{3} i_a - \frac{1}{3} i_b - \frac{1}{3} i_c \tag{8}$$

$$i_q = \frac{\sqrt{3}}{3}(i_b - i_c) \tag{9}$$

Under ideal conditions, for a healthy motor, Lissajou's curve $i_q = f(i_d)$ has a circular shape, centered at the origin and having a radio equal to the stator current complex vector corresponding to the state of operating of the motor [5]. In case of faulty motor, the Lissajou's curve changes in shape because of the harmonics presence generated by the fault. In this paper, Park's vector complex spectrum is used to detect induction motor faults.

5 Measurement and Analysis Data

We performed invasive experiments on an induction motor to obtain fault data or our analyses. The characteristics of the induction motor used in the experiment are listed in Table 1.

Table 1. Induction motor characteristics

Description	Value
Power.	0.75 kW (1Hp)
Voltage.	220 V
Current.	3.2 A.
Frequency.	60 Hz
Number of Poles.	4
Speed.	1745 rpm

In the figure 3 (a) and 3 (b) is presented the experiment setup. The induction motor was tested in healthy and fault conditions for different speeds and faults. A modified induction motor with shorted adjacent turns was used in the tests carried out. Figure 4 schematically shows the stator winding design, including how turn-to-turn faults can

be created. With this machine, a turn-to-turn fault ranging from 1 to 9 turns can be created.

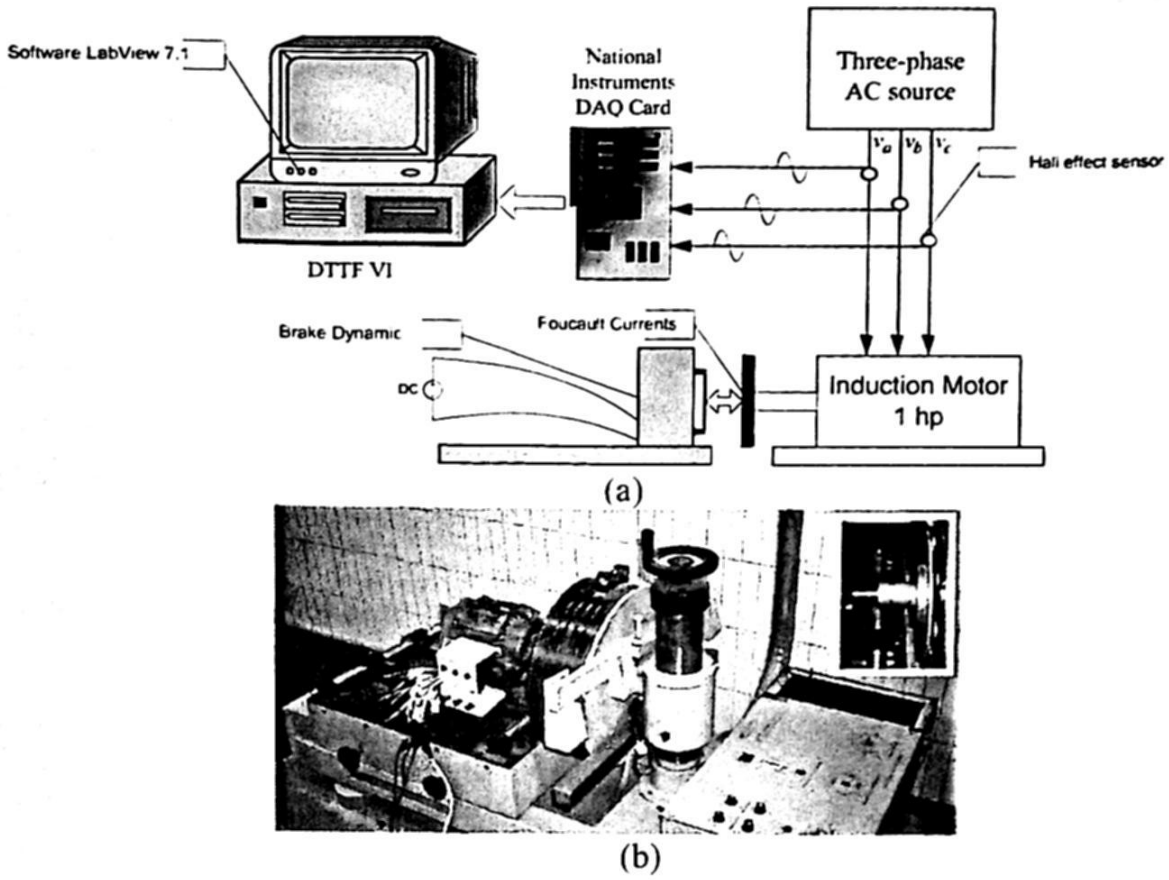


Fig. 3. a) Schematic of the experimental setup. b) Actual experiment setup to collect healthy and faulty motor data.

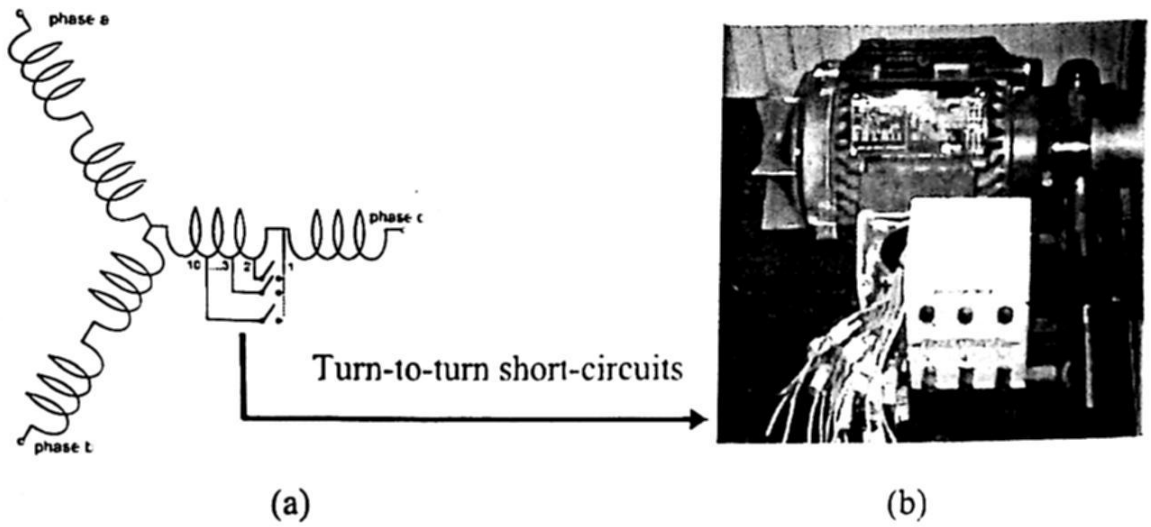


Fig. 4. a) Schematic of the stator winding design. b) Induction motor tested

Experiment consisted on collecting stator current data at different load motor conditions and faults. The induction machine was tested at 0%, 25%, 50%, 75% and full-load. Three Hall effect based sensors, Data Acquisition Card (National

Instruments DAQ-Card USB-9162), and a computer was used to obtain motor current data. Motor speeds were fixed by a Foucault Currents based break dynamic. Induction motor speed was measured by a tachometer. Figure 5 shows the stator current time evolution for a healthy and a faulty (nine- turn fault) motor respectively.

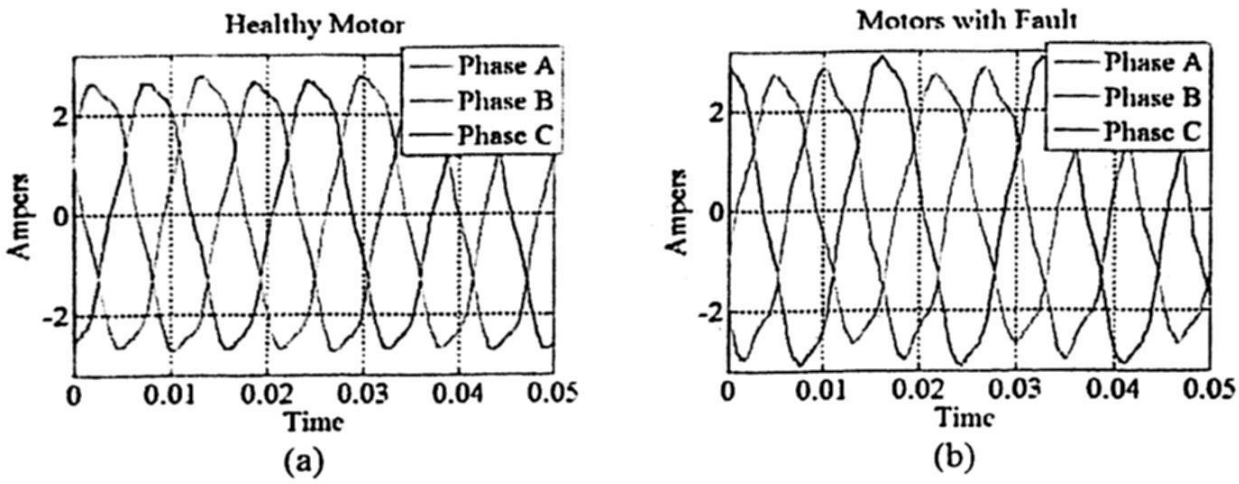


Fig. 5. Stator current time evolution under no load. (a) Healthy motor, (b) Motor with nine-turn short-circuit fault.

In figure 6 is presented the Park's vector trajectory and spectral for two situations: healthy and nine-turn fault. As can be seen in this figure, the magnitude of the harmonic at -60 Hz gives partial information about the fault. Harmonic at -180 Hz is also taken in account for fault detection. Park's vector in different practical situations presents drawbacks like was mentioned previously. For this reason the Park's vector modulus spectrum is proposed to detected abnormalities in the induction motor. The Park's Vector FFT is a power method to detected stator-winding faults. Stator current fault harmonics can be obtained as following [6]:

$$f_{stator} = \left\{ \frac{1}{p}(1-s) \pm k \right\} f_0 \tag{10}$$

Where p is the number of pole pairs of the motor, k = 1, 2, 3..., n is the index harmonic, and f_0 is the fundamental frequency. The equation 10 can be located whether we are doing reference to Motor Current Signature Analysis (MCSA). The slip s is defined as the relative mechanical speed of the motor, n_m, with respect to the motor synchronous speed n_s as [7]:

$$s = \frac{n_s - n_m}{n_s} \tag{11}$$

The motor synchronous speed n_s is related to the line frequency f_0 as:

$$n_s = \frac{120 f_0}{p} \tag{12}$$

Where 120 is a constant used to express the motor synchronous speed n_s in revolutions per minute (r/min) unit.

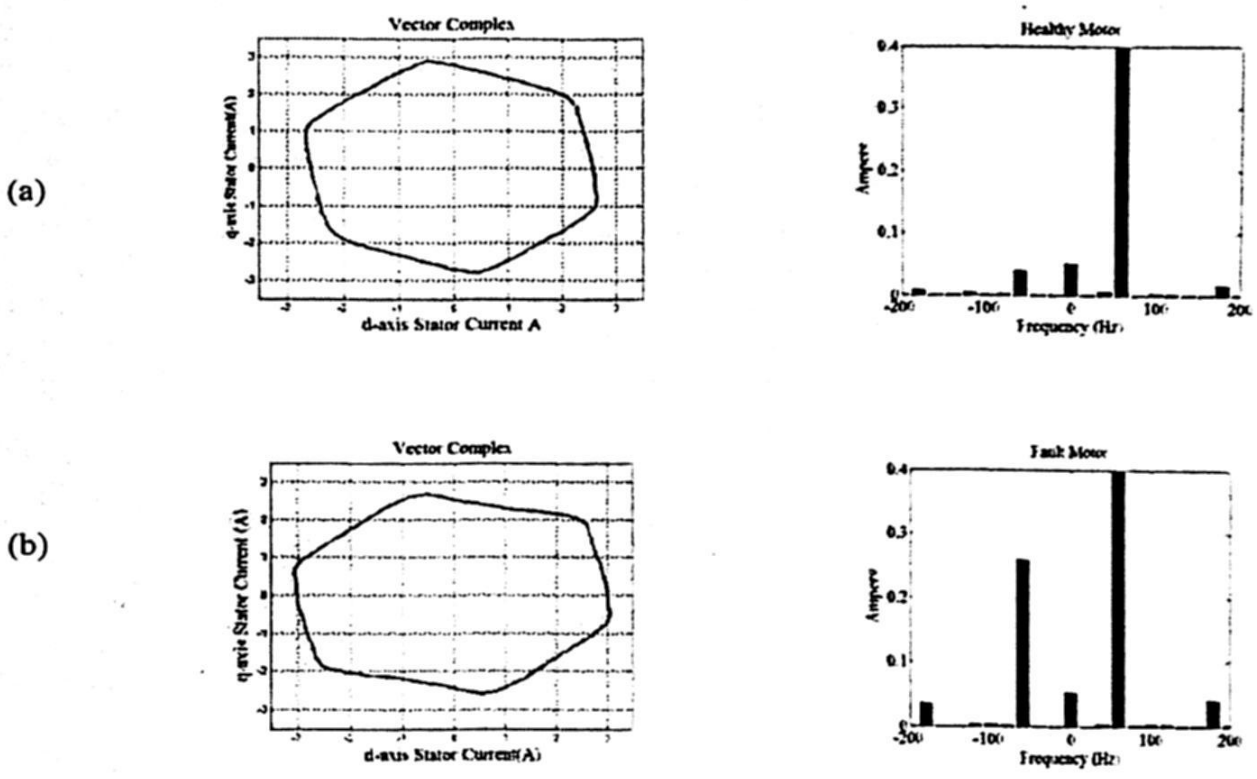


Fig. 6. Park's vector trajectory and spectral for a) Healthy, and b) Faulty condition.

Thus, we can do an analysis of the relation between the magnitude differences of current spectrum versus turn-to-turn fault severity under different load conditions. This relation is presented in the figure 7.

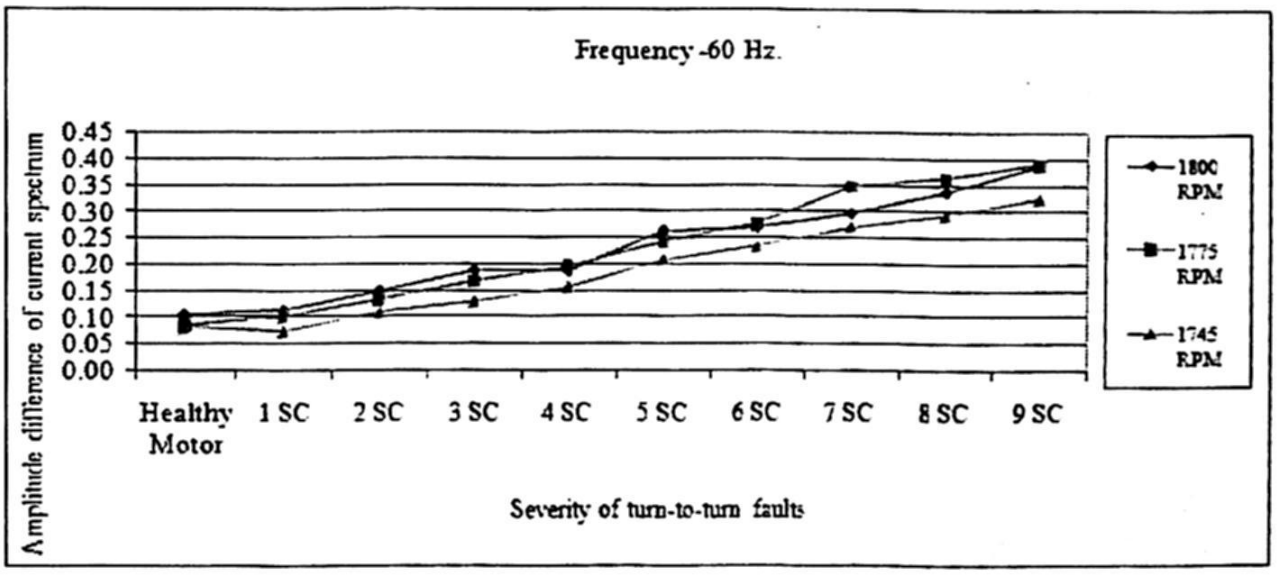


Fig. 7. Magnitude of the harmonic at -60 Hz for several speeds and fault severities.

6 Fault-Detection Schemes and Experimental Results

The experimental data were collected with a sampling frequency of 9.84 KHz for what each motor-current data set contains 9840 samples for duration of 1 s. There are a phase of pre-processing for that the current-data are processed by means of the spectral Park's Vector and using the equation 10 are obtained the training set according the frequency where is presented the stator winding fault. Each training set correspond to two types of condition: Healthy Motor and Faulty Motor under different load conditions. A training pattern is a vector of three columns whose data corresponding to the frequencies of -180 Hz and -60 Hz and 60 Hz of the Park's vector spectral. The experiments were based in two schemes: an artificial neural network with Backpropagation algorithm and a Support Vector Machine. The number of patterns in both schemes was obtained according to the rule of Baum and Haussler (1989) and is determinate for the next condition [11]:

$$P = \frac{W'}{e} \tag{13}$$

Where P is the number of training patterns, W is the number of weights in the neural network Backpropagation, and e is the percent of error in the classification on the validation set. Thus, with e = 0.1 and a maxim of 100 weights in the neural network, are obtained 1000 training patterns. These 1000 patterns are divided in two sets that correspond to the training patterns and validation patterns. The dimension of the training set is of 666 patterns and of the validation set is of 333 patterns.

6.1 ANN with Backpropagation Algorithm

The development neural network of this paper is based in the process of Rodvold 2001. This process is composed of five steps ("Network Requirements, Goals, and Constraints", "Data Gathering and Preprocessing", "Training and Testing Loops", "Network Deployment" and "Independent Testing and Verification") [12]. The values of the weights are initialized according to the method of Nguyen-Widrow with the purpose of improvement the learning ability of the hidden units. This method is based on geometrical analysis of the response of the hidden neurons to a single input. First is calculated the scale factor by means of the next equation [11]:

$$\beta = 0.7\,(p)^{\frac{1}{n}} = 0.7\sqrt[n]{p} \tag{14}$$

Where n (n = 3) is the number of input units, p is the number of hidden units. To initialize the weights is necessary to follow the next steps:

1) For each hidden unit, initialize its weights vector v_{ij} that has relation with the inputs units.
2) Firstly, set random number between -0.5 and 0.5 to the weights vector.

(15)

$v_{ij}(old)$ = random number between -0.5 and 0.5

To calculate:

$$\|v_j(old)\| = \sqrt{v_{1j}(old)^2 + v_{2j}(old)^2 + ... + v_{nj}(old)} \tag{16}$$

Reinitialize weights:

$$v_{ij} = \frac{\beta v_{ij}(old)}{\|v_j(old)\|} \tag{17}$$

3) Finally, set bias v_{oj} random number between $-\beta$ and $+\beta$.

We have two schemes of training for the neural network. The gradient descent is used in both schemes with the next parameters: learning rate is 0.001 and momentum is 0.8. In this paper is used the bipolar sigmoid function.

$$f(x) = \frac{2}{1+e^{-2*x}} - 1 \tag{18}$$

In case 1, we have stopped training if the network training error reaches a pre-set value, which in our case is set to 0.0005. The training and test results for this second case are shown in Table 2. In this case the ANN only recognizes a motor with fault and without fault. Two output units are codified in this model: 1 for faulty motor and -1 for healthy motor.

Table 2. Training and test results of the neural networks structures

ANN	Units Hidden	Validation Error	%	Test Error.	%
3	10	0.391097	92	0.494273	90
4	15	0.524772	87	0.424829	90
5	20	0.543908	87	0.627349	85
6	40	1.386552	83	1.083323	80
7	60	3.665848	74	3.723645	75
8	80	2.171371	78	3.828273	75
9	100	3.533251	74	4.049451	70

In case 2, we have taken into consideration that we have a large number of units in the hidden layer than the input layer. The gradient descent is used to train the ANN structure. In this case the neural network diagnoses the healthy and faulty condition motor. However, also, diagnose the severity fault since one to nine short-circuits in the stator winding. In the Table 3 are presented the outputs units.

Table 3. Outputs units of the artificial neural network.

Conditions Motor	Outputs Unit			
	1	2	3	4
Without Fault.	-1	-1	-1	-1
1 short-circuit.	1	-1	-1	-1
2 short-circuits.	-1	1	-1	-1
3 short-circuit.	1	1	-1	-1
4 short-circuit.	-1	-1	1	-1
5 short-circuit.	1	-1	1	-1
6 short-circuit.	-1	1	1	-1
7 short-circuit.	1	1	1	-1
8 short-circuit.	-1	-1	-1	1
9 short-circuit.	1	-1	-1	1

The training and test results are shown in Table 4.

Table 4. Training and test results of the neural networks structures .

ANN	Epochs	Training Error.	Validation Error	Test Error
1	310,000	0.034070	0.14150	0.24350
% Classification Validation/Test				86% / 80%
2	450,001	0.004098	0.05766	0.05966
% Classification Validation/Test				90% / 90%
3	620,001	0.011436	0.05853	0.06150
% Classification Validation/Test				90% / 88%

We applied training stop technique known as "Cross-validation". Cross-Validation is a technique to prevent overtraining which consist in to divide of the data in two disjoints sets. The first data set is the training set, which is used to train the ANN and the second set is used to validate the ANN structure. Thus, the validation error is checked throughout the training process. From Table 4, the absolute errors between the network outputs and the object outputs for all training patterns are less than 0.004098 after 450, 001 iterations. It is shown that the BP networks have the very high diagnosis accuracy and good generalized ability. The figure 9 illustrates the principle of the cross-validation. The figure shows that training stops at the 450,001 epochs. In this figure can be seen the increment of the validation error, the training is stopped to avoid overtraining and most recent weight and the biases are used as the neural networks parameters.

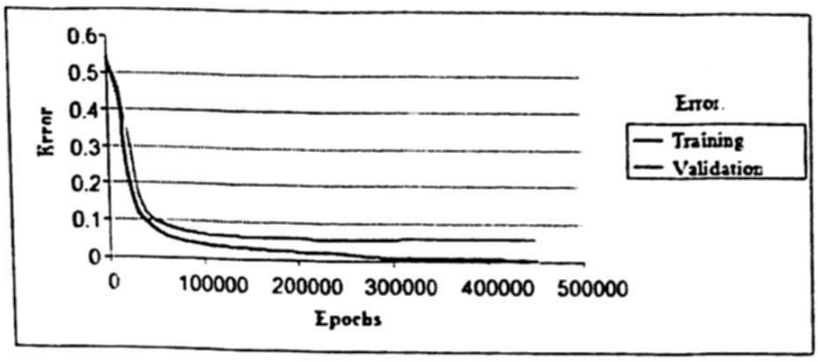


Fig. 9. Training and validation error curves with cross validation technique.

6.2 Support Vector Machine.

The learning samples and the test samples for the SVM are the same that we were used in the neural network with backpropagation algorithm. The polynomial kernel is used to train the SVM. The percent of accuracy rate in the classification, the accuracy rate in the generalization, and the number of support vectors in relation with the grade of the kernel polynomial is presented in the table 5. The percent of the rate of generalization and classification is obtained of the division of the correct number of patterns recognize and the total number of patterns in the set.

Table 5. Training and test results of the neural networks structures .

Polynomial Grade	Rate (%) Classification	Rate (%) Generalization	Number Support Vectors
3	91	87%	21
4	91	87%	15
5	97	90%	18
6	96	91%	15
7	98	91%	15
8	99	96%	21
9	98	96%	12
10	93	90%	9

7 Conclusion

Current spectrum analysis based on vector complex and a neuronal network to diagnose windings faults of an induction motor has been presented. Support Vector Machine and Backpropagation Algorithm were implemented in software to do the diagnostics on line and off line of an induction motor. The system is limited to the diagnostic of stator windings faults. The patterns are obtained by means of the spectral Park's Vector and using the Motor Current Signature Analysis. The MCSA does possible to identify the harmonics that describes the presence of a fault by mean the equation 10. A pattern of behavior is observed in the frequency of -60 Hz of the

spectral Park's Vector while the harmonic on 60 Hz can describe a feature about of velocity of the induction motor. A SVM and Backpropagation Algorithm are trained to diagnose faults in an induction motor; however the results suggested that the SVM could be used to develop fault-detection schemes because of their multinput-procesing and its good generalization capability. Cross-Validation technique was used to prevent overtraining and to stop the training of the neural network with backpropagation algorithm.

References

1. S. Haykin, "Neural Networks a Comprehensive Foundation", Prentice Hall, Second Edition, Upper Saddle River NJ, 1999. ISBN. 0-13-273350-1.
2. Q. He; D. Du, "Fault Diagnosis of Induction Motor Using Neural Networks", IEEE in Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007.
3. D.J. Sebald and J.A. Bucklew, "Support Vector Machine Techniques for Nonlinear Equalization", IEEE in Transactions On Signal Processing, Vol. 48, No.11, November 2000.
4. Z. Luo and Z. Shi, "On Electronic Equipment Fault Diagnosis Using Least Squares Wavelet Support Vector Machines", IEEE in Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21 – 23, 2006, Dalian, China.
5. M. Hachemi Benbouzid, "A Review of Induction Motors Signature Analysis as a Medium for Faults Detection", IEEE in Transactions on Industrial Electronics, Vol. 47, No. 5, October 2000.
6. B.K. Bose, "Neural Network Applications in Power Electronics and Motor Drives An Introduction Perspective", IEEE in Transactions on Industrial Electronics, Vol 54. No. 1, February 2007.
7. J. Zarei; J. Poshtan, "An Advance Park's Vectors Approach for Bearing Detection" IEEE, International Conference on Industrial Technology (ICIT) 2006, 15-17 December.
8. J. Jung, J. Lee and B. Kwon, "Online Diagnosis of Induction Motors Using MCSA", IEEE in Transactions on Industrial Electronics, Vol. 53, No. 6, December 2006.
9. W. T. Thompson, M. Fenger, "Analysis to Detected Induction Motor Faults", IEEE Industry Applications Magazine. July/August 2001.
10. B. Ayhan, M. Chow, M. Song, "Multiple Discriminant Analysis and Neural-Network-Based Monolith and Partition Fault-Detection Schemes for Broken Rotor Bar in Induction Motors", IEEE in Transactions on Industrial Electronics, Vol. 53, No. 4, August 2006.
11. V. Faussett, "Fundamentals of Neural Networks: Architectures, Algorithms and Applications", Addison Wesley, Florida Institute of Technology 1994, ISBN -10. 0133341860.
12. B. J. Taylor, "Methods and Procedures for the Verification and Validation of Artificial Neural Networks", Springer Science, Institute for Scientific Research, Inc., Fairmont, WV, USA 2006. ISBN-13: 978-0-37-28288-6.

Design of a Flexible Graphic Visualizer for Flowshops Scheduling

Rodolfo Ruiz Nangusé, Larysa Burtseva, and Gabriel A. López Morteo

Engineering Institute of Autonomous University of Baja California, Calle de la Normal S/N, 21280, Mexicali, BC, Mexico
{rodolfor, lpb, galopez}@iing.mxl.uabc.mx

Abstract. This paper describes the Flexible Graphic Visualizer, designed for the graphic interpretation of the results of flowshop scheduling algorithms obtained in a computational system. The visualizer is intended for the understanding of the data model generated by a scheduling algorithm. The visualizer interface integrates different visualization and interaction techniques for a deep and interactive exploration of the information represented in a visual manner. Due to the capacity of the visualizer to represent graphically data of any scheduling context, this system is useful to visualize complex data from flowshop scheduling algorithms both on production environments or research.

Keywords: visualization, scheduling, flowshop, Gantt chart, visualization techniques.

1 Introduction

This paper describes flexible visualization software for graphic interpretation of the results of the execution of algorithms for flowshop scheduling.

The first work dedicated to job scheduling, was published in 1950's by S. M. Johson [1]. Since those times, the problem attracted the attention of many researches because the theory complexity, and diverse of practical's applications [2, 3]. Due to this complexity, the analysis and execution of scheduling algorithms require the development of computational systems, such as visualization tools, that help in the interpretation of the algorithms' results. With the help of these tools, it is possible to make clear different schedules effects, such as "bottle necks", machines employment, homogeneity, etc.

When a specific scheduling algorithm is implemented in industrial environments, it is important for the operators to have adequate representations of the data, showing not just the schedules but related information, in a comprehensible and comfortable form, to support the organization of the productive process. The information related to the schedules are, allocation of jobs to the resources, monitoring of jobs' dates to machines, setup times in machines, WIP, interruptions of machines for blocking, etc.

The visualization systems or visualizers available at this time, both commercial and particular, were designed for certain groups of problems according to the interests of their users. Taking this in account, the properties and behaviors of visualizers allow

Advances in Computer Science and Artificial Intelligence.
Research in Computing Science 39, 2008, pp. 263-276

Received 01/07/08
Accepted 01/08/08
Final version 11/09/08

classifying them in three categories: *layout visualizers*, *automatic visualizers* and *enterprises solutions visualizers*.

Layout visualizer provides the user with a set of tools, such as lines and forms, to generate graphs from the results of its scheduling algorithms in a manual way [4, 5].

Automatics visualizers have an interface for interaction with the user; involve programming to generate graphics according to a wanted context. They provide diverse tools for the exploration of the data [6, 7, 8].

Enterprises solutions visualizers are integrated in packages of applications that offer to the companies' solutions to their problems in diverse areas, such as the management of departments, the planning of the production and projects, etc. The features for these visualizers are very general due to their tendency of embracing diverse areas. The specialization and robustness of this kind of visualizers are in proportion with the cost and the maker's prestige [9, 10].

Due to the complexity in the generation of calendars, standard free software doesn't exist for charting the results of specialized algorithms in scheduling. The commercial visualizers are oriented to enterprise companies at very high cost [7, 9, 10]. The visualizers developed for research is oriented to particular uses, for example, [11, 12].

The investigation of algorithms and visualizations of results for complex models of scheduling imply a variety of elements to be represented in the same chart. Otherwise, the implementation of visualizer in the factory requires diverse focuses in a schedule with the purpose of acquiring valid conclusions about the productive process.

This article describes the design of a flexible graphic visualizer (VGF) that achieves all requirements above mentioned. It is employed as an independent component for the system PLARETF, developed for execution and analysis of flowshops scheduling algorithms [13, 14].

The rest of the paper is organized as follows: Section 2 explains the definitions and intrinsic notations in a scheduling problem, shows up the visualization of problem results using a Gantt chart. The Section 3 describes the state of the art for the knowledge acquisition from a dataset that are studied in a specific context. The Section 4 provides the design of VGF. Presenting the scheduling knowledge model and the visualization techniques identifies according to the characteristics of the scheduling; later is detailed the elements of VGF interface in which are reflected the visualization techniques studied and the way we relate them to provide many perspectives and details of the information. In another section the characteristics of VGF flexible input format that allows represented graphically different models of shops scheduling is showed. Also, the VGF architecture showing the outline of its elements and the communication that exists among them to provide the required functionalities are presented. Lastly, section 5 contains the conclusions of the work.

2 Scheduling Problem

Different aspects of scheduling are examined in scientific literature [15, 16]. The problems are formulated in terms of "jobs" which are processed in "machines" with

various constraints. The jobs flow pattern of the machines defines the shop model. A conveyor with m successive operations represents a simple flowshop (FS). If for the execution of at least operation are several machines available, such shop called as flexible flowshop with m stages (FFS), if the machines in a group are identical, and hybrid flowshop (HFS), if these they are uniform (machines with different speeds) or not related (different machines). The Fig. 1 show the resources system in a FFS with m stages, each of which contains m_i parallel machines, $i = 1, \ldots, m$.

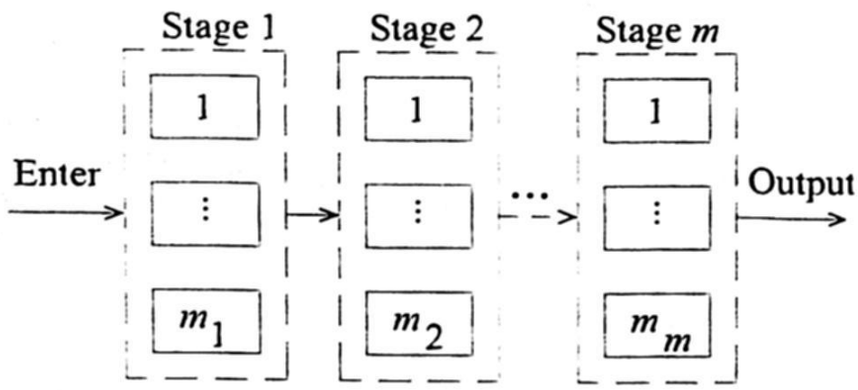


Fig. 1. Flexible (hybrid) flowshop with m stages.

With each one job j, $j = 1, \ldots, n$, several data are associated, as processing time p_{ilj} in the machine i of stage l, arrival instant r_j in shop, due day d_j that represents the dates the job j is promised to the customer, weight w_j which is basically a priority factor of job related to other jobs in the system, setup time s_{iljk} of the machine i to process the job k after the job j. the model specified other properties and restrictions as set M_j of eligible machines for job j when in the machine environment are several machines in parallel, machines blocking which may occur in flowshops when it has limited buffer between two successive machines for work in process (WIP) between two successive machines, etc.

The scheduling aim is to satisfy criteria. The most common optimization criterion is to minimize the jobs set processing time (makespan) since usually implies a high utilization of the machines. Others criteria exist, such as minimize the jobs' finalization time, considering jobs arrival at the system (release times), finalization compromise dates (due dates), priorities (weights), precedence relationships, the number of jobs retarded, among others.

Here an example is presented. It is necessary to process 6 jobs in a HFS of two stages with a machine in the first stage and two machines in the second. All the jobs consist of two successive operations. The first operation carries out by unique machine of the first stage. To execute the works 1, 2, 3 is eligible the machine 1 of the second stage and for the other three works the machine 2 is eligible. The jobs release times are $r_j = 0$, $j = 1,\ldots,6$. The Table 1 shows the jobs' processing time for the example. The non eligible machines for a definitive job are marked with the sign "-".

Table 1. Jobs processing time

Job j	Machina 0	Machina 1	Machina 2
1	1	4	-
2	5	9	-
3	3	6	-
4	5	-	5
5	4	-	10
6	8	-	7

The permutation $\pi^* = (5, 2, 6, 3, 4, 1)$ corresponds to optimal schedule with $C_{\max} = 30$. The Fig. 2 shows the problem solution in Gantt chart form for the permutation π^*.

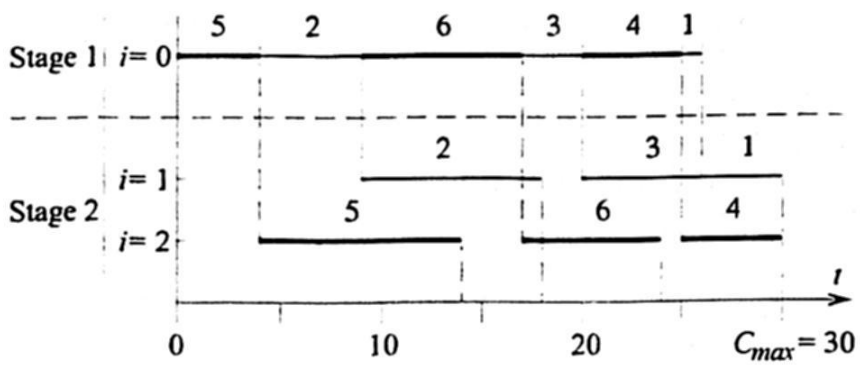


Fig. 2. Gantt chart for the optimal permutation $\pi^* = (5, 2, 6, 3, 4, 1)$

As shown in the example, the Gantt chart clarifies the visualization of the assignment and the movement of jobs on the machines, timing of each job, and idle times of each machine. In a similar way, other properties and restrictions of the shop are indicated in the chart. The Gantt chart is applicable for any kind of shop: flowshops, openshop, jobshop, therefore they are used in different scheduling visualizers [6, 7, 8, 9, 10].

The visualization of scheduling shops has its intrinsic complexity due to variations in:

- Length of Gantt chart that vary from short periods of time to long periods, until thousands time units:
- Models of resources as number of stages in shop, number of machines for stage, type of flow of jobs.
- Concepts and values to be displayed, such as makespan, setup times of the machines, release times, WIP, and blocking.
- Necessity of obtaining and visualizing the information for demand.

3 Acquisition of the Knowledge through the Visualization

3.1 Knowledge Acquisition Stages

The visualization is the creation of a mental image from an abstract concept [17], to perceive the characteristics of the data examined. Diverse approaches exist about the knowledge acquisition, locating the visualization as a layer between the human and the information [18]. In publications referred below, authors analyzed how the knowledge was represented to visual form from raw data. According to [19], the knowledge acquiring process is composed by four conceptual elements that evolve in the time: those elements are involved in an understanding context.

In the lowest part of the model, the raw data are stored as a characters set. In the following level metadata are added, to bind the raw data to particular context, resulting in what is denominated as information.

Until this point the raw data and the information are denominated as *producers* because they produce the information about a context. In a higher level, the information interacts with a human. The data are interpreted when being placed in a context. Based on the experience, the knowledge has achieved. In this point, the knowledge and the information are denominated as *consumers*. Finally become a concept called *wisdom*, which is achieved using and combining the knowledge in different forms and situations inside the individual's context.

Another approach is proposed in [20]. In the visualization process there are four basic phases that feedback between the last in the first. The first phase consists on the recollection and storage of the data, the second phase pre-processes and it transforms the data in information according to the individual's environment; the third phase deploys the graphically information that activates the visual and cognitive human systems. An analyst interacts with the information in three aspects. The first one is the data recollection from any physical environment where they are stored, and a social environment that determines their interpretation. Aspects as the exploration and manipulation form the second aspects where the data are been understanding for the human, and the third aspect is about the visual device, the manner which the information is represented to the user in visual form.

In [21] described the transformation phases of a dataset into visual forms for human collectors. In the first phase, the raw data are converting in tables of data through of metadata aggregation. The following phase is to add a visual structure to the data tables producing a visual representation. This is achieved through visual mappings. The last phase involves the visual mappings interpretation to show them to the user through a visual device. The user feedback the system through the change of parameters that control them the three phases of transformation above mentioned.

The visualization concept approach development by Card [21] was used as conceptual base for the VGF design, and is described in section 4.

3.2 Visual Exploration of Data

Three processes compose the exploration of data in a visual way: panorama, zoom

and filtrate [22]. First, the user needs to get an early panorama of the data, where patterns are identified and focused in one or more of them. To analyze patterns, it is needed to access data details. The user identifies a data subset, deepens the panorama through zoom and filtration process, decomposing the subset and exploring its elements to obtain details. It is important stress that the visualization technology not only provides the bases of the techniques for the three processes of the data exploration but it is also the bridge among them.

3.3 Visualization Techniques

Below the techniques applied for the interactive visualization charts are described.

Keim [22] classifies the visualization techniques according to three approaches:

- Characteristics of the data,
- Technique used for visualization,
- Technique of interaction with the data.

The data characteristics specified the data types represented: one-dimensional, bi-dimensional data, text, hypertext, hierarchy, graph, algorithm, and software.

In [23] a deep analysis is made about the visualization techniques and the context in what are appropriate: 2D and 3D graphics, geometric transformations, iconic representations, density diagrams of pixels (stacked diagrams are an example of them.) Also in the same paper authors analyzed how the data can be explored and manipulated.

4 VGF Design

The VGF described in this article, is an application developed for the necessity of representing graphically the results from PLARETF [13] system. The VGF system has a flexible input format, and allows the graphical representation of any shop type, including openshop and jobshop, with their characteristics and restrictions.

The agile understanding of the model is reached through the application of interaction technologies. Below some visualization aspects are described in the scheduling context.

4.1 Scheduling Results Knowledge Model through VGF

The model of understanding through visualization, presented in fig. 3, describes how the data generated by a scheduling algorithm, are understood by an analyst. Three transformation processes and a feedback process compose it.

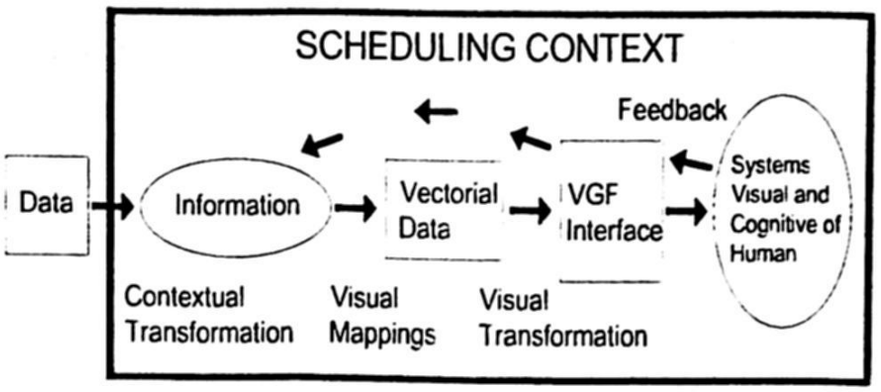


Fig. 3. VGF scheduling results knowledge model

In the process of *contextual transformation*, output data of an algorithm, stored in a text file in the input format of VGF are interpreted in the scheduling context. Thus, the raw data has become information. The VGF component, transforms the information from scalar to the vector adding *visual mappings*.

Later, the VGF component *transforms graphically* the vector information through the interface, showing the algorithms results in graphical form. In addition, the VGF interface offers a collection of tools that allows user interaction with the information. In this point, the *feedback* process facilitates the access to the calendars data in plain text, to manipulate the information and to explore it in detail, activating the cognitive and visual systems of the user. The transformation and processes feedback are inside the scheduling context.

4.2 Visualization Techniques

The VGF used the following visualization techniques [22].

– **Data representations:**

The system uses text to define labels and comments. The jobs processing dates are defined as two-dimensional data of text; it is an array whose each line describes the job timing. The data involved in the shop model representation are hierarchical, in tree form whose root is the shop, the machines occupy a second level, and the works are the leaves.

– **Visualization Techniques:**

As the main representation for the algorithms results provided by PLARETF, we use Gantt charts for the representation of scheduling characteristics and restrictions. The Gantt charts allow visualizing, in an agile way, the jobs processing order in the shop, indicating the machine used, and the beginning and end instants for each operation of the jobs.

The diagram radar view shows a scaled representation of the diagram of Gantt to give the user a total perspective of the image. This is usefully when an original schedule is very extensive and doesn't spread out entirety in the screen. This is important when, for example, a schedule has a big number of stages, machines, and

jobs; big values of times; and numerous restrictions and characteristics. This kind of diagrams offers the possibility to browse the data and to observe in detail the characteristics of a subset of data.

The third visualization technique implemented is a tree diagram, which takes advantage of the hierarchical characteristics of the data and it deploys to the tree a directory of elements whose three levels are the shop, the machines and the jobs. This kind of representation allows observing the complete structure of the shop.

- **Interaction techniques:**

The interaction techniques added to the VGF allow the interaction with the data through an interface [20].

Interactive Zooming provides a scale diagram handling in two senses: nearing to the data to show information in detail on certain areas on the image; reduction of the image to give a wide perspective. In dependence of the original size the Gantt chart offers 25, 50, 75, 100, 150, 200 % of zooming.

Interactive Filtering consists on the selection of a data subset to observe its details, for example, to show the work-load assigned to each machine.

Linking and brushing provide simultaneous visualization of the same information in related diagrams that provide several perspectives of the same information to the user [25, 26].

4.3 Interface

The VGF interface integrates the visualization techniques and the interaction techniques mentioned previously.

The screen is divided in five areas (Fig. 4a): 1) Menus and shortcuts, 2) Gantt chart, 3) Shop tree, 4) Radar view, 5) Schedule properties.

The menus and shortcuts area incorporate basic actions: open files, save files, print diagrams, export a graphic schedule, as well as the shortcuts to provide greater agility in the handling of the system.

The tree diagram corresponds to the hierarchical structure of the shop, having the shop as root, and in second level the machines are located with the jobs as leaves. This diagram is related with the Gantt chart, the radar area and the schedule properties in the following way. When the user selects a node of the tree, it is shadowed in the Gantt chart and in the diagram of radar view. Also the interface shows the information details of the selected node according to the context of the element. For example, if the user selects the root of the tree, the Gantt chart stand out and the characteristics of the shop are shown in the area of properties containing: kind of shop (simple flowshop, hybrid or flexible flowshop), characteristics of machines (parallel, identical, uniform or not related machines), among other data (see Fig. 4a). If a machine is specified, then in the area of properties the characteristics of the suitable machine are shown containing information such as: number of jobs assigned for processing in a machine, completion time of all the works, jobs for stage with their respective completion time (see Fig. 4b). In the event of selecting a specific job, in the properties area are shown the instants of their release times in the shop, beginning and end of the setup times, and completion times (see Fig. 4c).

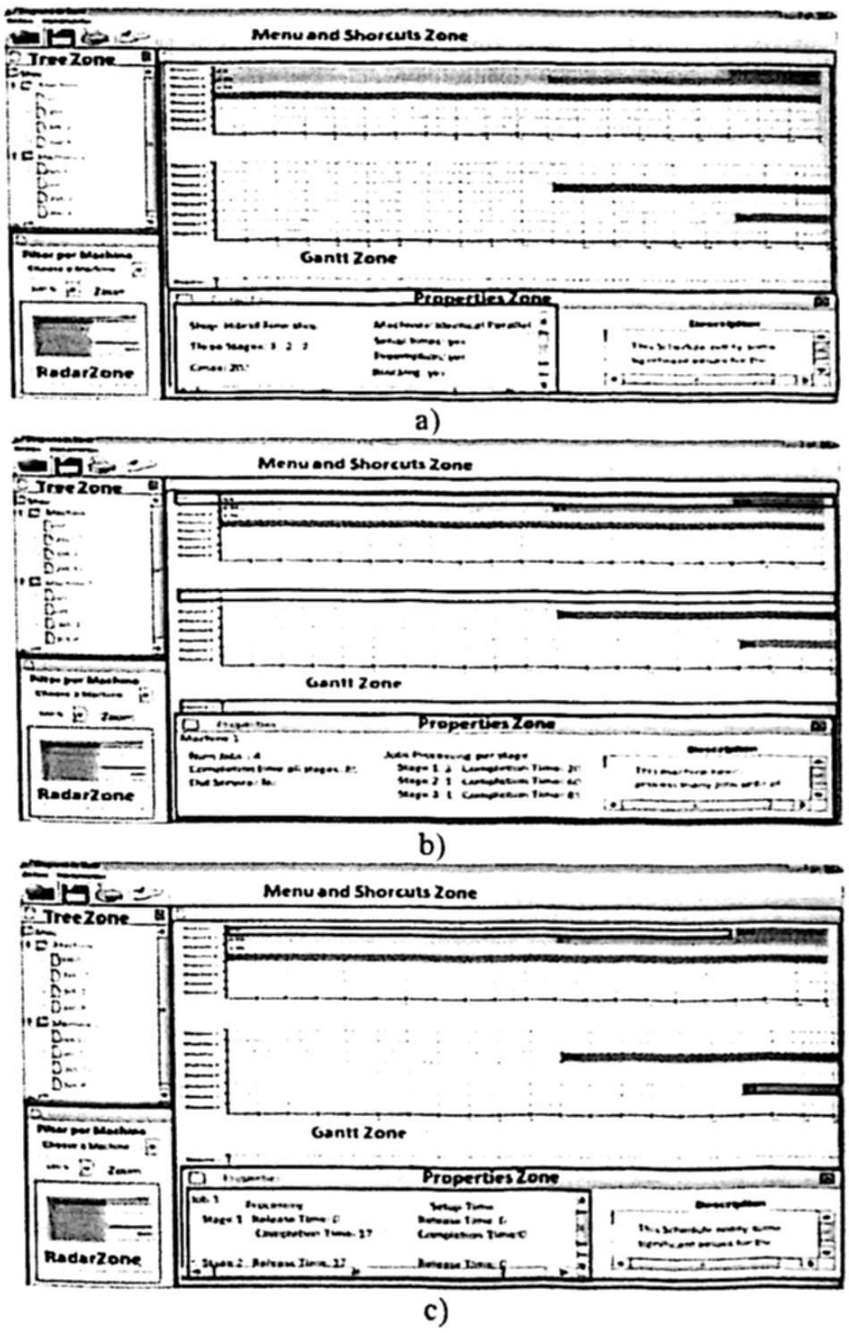


Fig. 4. VGF interface: a) selection of root tree; b) selection of machine; c) selection of job.

The Gantt chart spreads the schedule with all their characteristics defined in the input file of VGF: release time and finish time for each job, and their operations with regard to the machines. In this space it is possible to add the analyst's annotations about the graphic schedule.

Radar area shows the complete Gantt chart, adapted to the size of the area, using a reduced scale. It offers a wide perspective of the diagram, and allows the navigations to specified parts of the graphic representation in the area of Gantt chart.

In the area of schedule properties, some properties are shown according to the level selected in the area of shop tree. It also allows to the analyst to make comments about the schedule in description field.

All the diagrams that appear in the VGF interface are related to each other to offer

different views from the same information, as it prescribes the use of the multiple visualizations, in such a way to provide bigger detail of data [27, 28].

4.4 Input Format

The computational system PLARETF provides to the VGF system with scheduled data in a special input format. The solution of the scheduling of a workshop, provided by PLARETF is not agile for understanding due to the extension of involved information. The input format of VGF was designed in such a way, that can adapt it to any scheduling context.

Below the input format implemented in VGF is described. The data is divided in four sections (Fig. 5).

The first section (Fig. 5a) contains the schedule notations. It is a summary of declarations of the notations in the text file for their visual representation.

The second section (Fig. 5b) summarizes the shop characteristics, which contains Graham's triplet: shop characteristics (FS, FFS or HFS), and characteristics of machines (parallel, identical, standardize, not related). Also, all the jobs characteristic and optimization criteria are specified.

The third section (Fig. 5c) describes the shop resources, detailing the number of stages and the number of machines for each stage, as well as the total jobs quantity in the shop. All the machines, jobs characteristics and optimization criteria are specified.

The fourth section (Fig. 5d) is the area of specification for schedule times. An array contains the timing of the works. The items correspond to concepts related with each job: processing time, release time, setup time, WIP, and blocking.

The columns are organized by triads. For an Item 1, the first triad describes the timing of job in the machine that carries out the first stage (operation): the machine number, time of beginning in this machine and time of the end. The second triad in the same way describes timing of the second stage, etc., for all stages and for all items.

The form of representation of input data in an array gives the flexibility to VGF because it is independent of the scheduling model.

The system allows to add more elements than they are in a traditional Gantt chart, that represent another data besides of jobs processing time and sequence depends setup times: in order to accomplish this, it will be enough with adding another label called description in the section one, and add an array in the section four for the new description.

4.5 VGF Architecture

The architecture of VGF is organized in three layers (see Fig. 6). The first layer corresponds to the interface by means of which the user interacts with the information, represents visually the algorithms results and allows to manipulate and explore the data. The second layer works as VGF administrator that monitoring and responds to the user's actions. The third layer contains the VGF component, divided in three modules: interpretation engine, entity data and mapping visual.

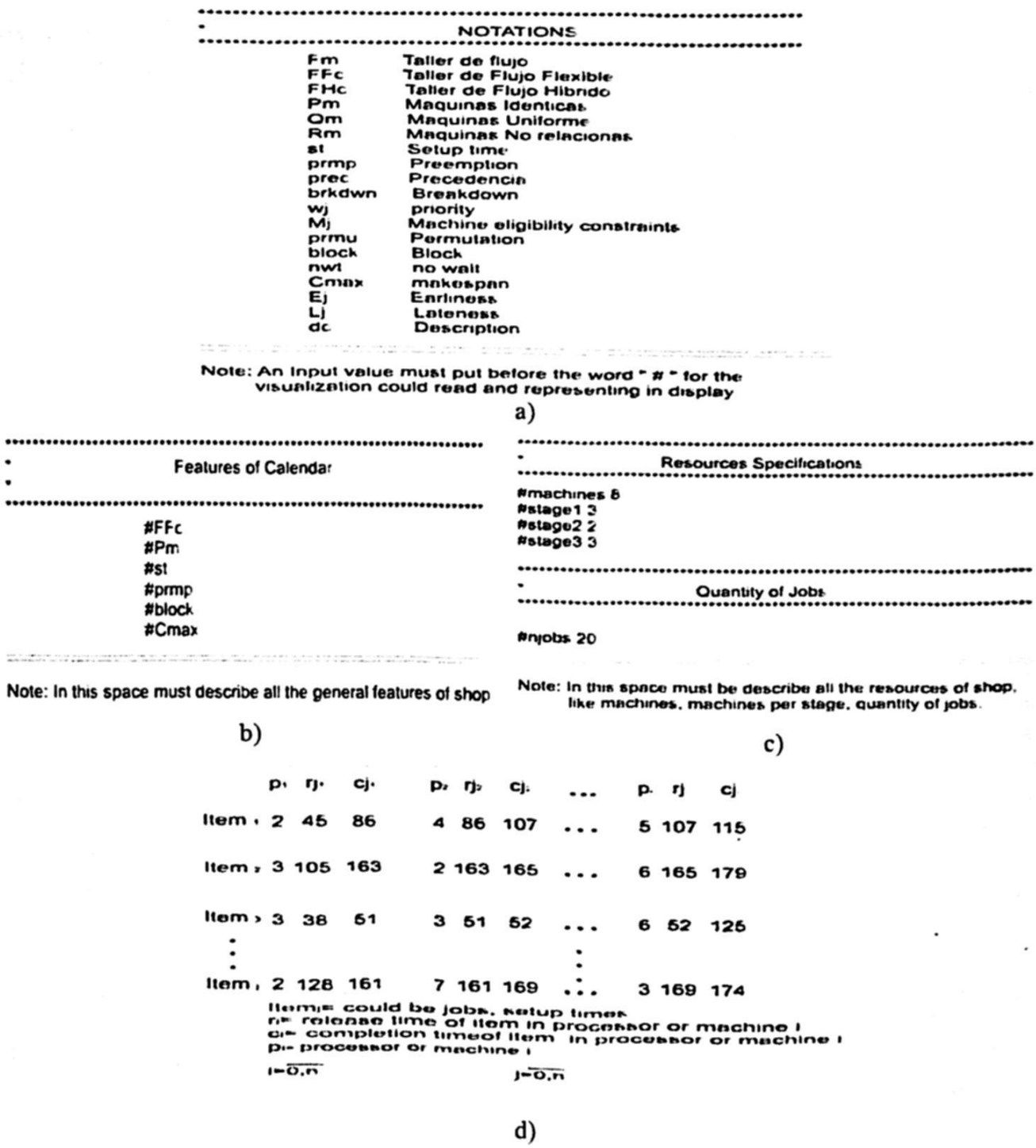


Fig. 5. The VGF input format: a) notations section; b) schedule features; c) resources specification; d) array of processing times.

The interpretation engine is the one in charge of reading all the information of the input text file, to incorporate the information into the system and to store it in the data entity module. The mapping visual module take the stored information from data entity and transforms the information from scalar to the vector. and generate a diagram that is sent to the VGF interface to present it to the user.

The VGF component represents the Gantt chart. Therefore, it has the property of been reusable in other projects. It requires as input, the schedule information generated by the jobs scheduling algorithm and creates a Gantt chart with all the schedule's characteristics. The VGF component is independent of the management

and visualization modules. It is designed to be re-used in any visual platform customizing its characteristics to respond to the necessities of future research where the visualization of job shops scheduling should be required.

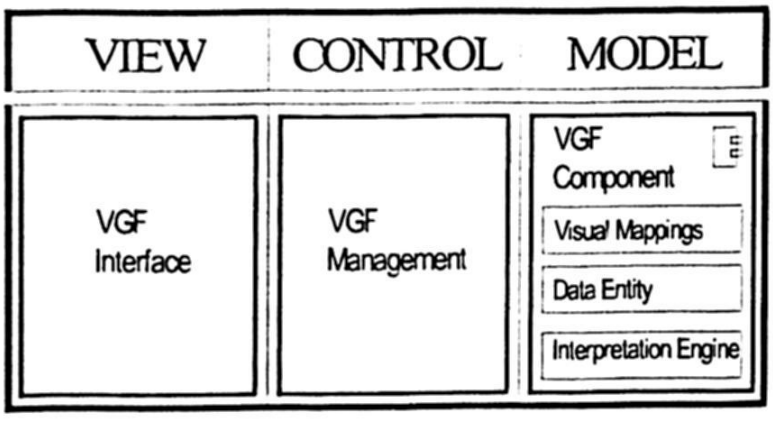


Fig. 6. VGF Architecture

5 Conclusion

In the development of visualization systems is important to take in account the specific context of the area in order to extract most information about the graphical representation of the data.

The real production implies necessities in the development and researching of scheduling algorithms with specific characteristic of the problem. The complexity in the visualization of calendars is due to the variety of elements to be represented in the same chart. At the moment doesn't exist software of free distribution that shows all the elements in a schedule. The commercial visualizers don't achieve with the investigation necessities and they have a high cost.

In this article we proposed a model of acquisition of the knowledge in the scheduling context through of the user's interaction with VGF. The design of the interface in form of four related areas gives to the user different perspectives of the information. VGF is a flexible visualization system due to his input data format, and visualizes diverse scheduling shops. This VGF could be reused in future researches that require the visualization of schedules generated for different scheduling models.

References

1. Johnson, S.M.: Optimal two- and three- stage production schedules with setup times included. Naval Research Logistics Quarterly 1(1), 61-68 (1954).
2. Garey, M.R., Johnson, D.S., Sethi, R.: The complexity of flowshop and jobshop scheduling. Mathematics of Operations Research 1(2), 117-129 (1976).
3. McKay, K. N., Pinedo, M., Webster, S.: Practice-focused research issues for scheduling systems. Production and Operations Management 11(2), 249-258 (2002).
4. Microsoft. Información General de Visio 2003.
5. http://www.microsoft.com/latam/office/visio/prodinfo/overview.mspx (October, 2007).

6. Mindjet. Mindget MindManager7.
7. http://www.mindjet.com/us/products/index.php (October, 2007).
8. Microsoft. Información General del Producto Microsoft Office Project Standard 2007 http://office.microsoft.com/es-es/project/HA101656383082.aspx (October, 2007)
9. Oracle. Oracle E-Business Suite. http://www.oracle.com/applications/e-business-suite.html (October, 2007)
10. Aqua eSolutions. Aqua EBS 2007. http://www.aquaesolutions.com/index.htm (October, 2007)
11. Deister. DEISTER ERP SUITE: Axional ERP. http://www.deister.es/es/products/e-erp/ (October, 2007)
12. Visual Consulting. Módulos Base de Visual Manufacturing.
13. http://www.visualconsulting.com.mx/visualmanufacturing.htm (October, 2007)
14. Ruiz, R., Şerifoğlu, F.S., Urlings, T.: Modeling realistic hybrid flexible flowshop scheduling problems. Computers & Operations Research 35(4) 1138-1150 (2008).
15. Pinedo, M., Chao, X., Leung, J. Michael Pinedo web site.
16. http://www.stern.nyu.edu/om/software/lekin/index.htm (March, 2008)
17. Romero, R.: Sistema computacional para la evaluación de algoritmos de planificación de trabajos en un taller de flujo híbrido. M.C. thesis, Engineering Institute, UABC, Mexicali, México (2007).
18. Yaurima, V., Burtseva, L., Tchernykh, A.: Hybrid Flowshop with Unrelated Machines, Sequence Dependent Setup Time and Availability Constraints: An Enhanced Crossover Operator for a Genetic Algorithm. In: Wyrzykowski et al. (Eds.) Parallel Processing and Applied Mathematics. LNCS, vol. 4967, pp. 609-617. Springer (2008).
19. Leung, J. Y-T.: Handbook of scheduling: algorithms, models and performance analysis. Chapman and Hall/CRC, Fl. (2004).
20. Pinedo, M.: Scheduling: theory algorithms and systems. Prentice-Hall, New Jersey (2002).
21. Dursteler, J. C.: Visualización de información. Gestión 2000, España (2001).
22. Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., Uthurusamy R.: Advances in Knowledge Discovery and Data Mining. American Association for Artificial Intelligence, LA (1996).
23. Wurman, S. R.: Information Architects. Watson-Guptill Pubns, NY (1997).
24. Colin, W.: Information Visualization: Perception for design. Morgan Kauffman, San Francisco (1999).
25. Card, S. K., Mackinlay, J. D., Shneiderman, B.; Readings in information Visualisation: Using Vision to Think. Morgan Kaufmann, San Francisco (1999).
26. Keim, A. D.: Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and computer Graphics 7(1), 100-107 (2002).
27. Keim, A. D.: Visual Techniques for Exploring Databases. Int. Conference on Knowwledge Discovery in Databases. http://www.dbs.informatik.uni-muenchen.de/~daniel/publication.html (1997)
28. Wainer H., Velleman, P.: Statistical graphics: Mapping the pathways of science. Annual Review of Psychology, 52, 305-335 (2001).
29. Swayne, D., Lang, D.T., Buja, A., Cook, D.: GGOBI: Evolving from XGobi into an extensible framework for interactive data visualization. Computational Statistics & Data Analysis, 43(4), 423-444 (2003).
30. Young, F.W.: ViSta: The Visual Statistics System. http://www.visualstats.org (March, 2008)
31. Ledesma, R., Molina, J. G., Young, F. W., Valero, P. M.: Desarrollo de Técnicas de visualización Múltiple en el programa Vista: Ejemplo de aplicación al análisis de componentes principales. Psicothema 19(3), 497-505 (2007).
32. Young, F. W., Faldowski, R. A., McFarlane, M. M.: Multivariate statistical visualization in computational statistics. In C.R. Rao (ed.): Handbook of Statistics, Amsterdam: Elsevier Science, pp. 959-998 (1993).

Author Index
Índice de autores

Editorial Board of the Volume
Comité editorial de volumen

Senior Editorial Board

Editorial Board Members

Florence Sèdes	UPS, France
Genaro Rebolledo Méndez	University of Coventry, UK
Georgina Flores Becerra	ITP, Mexico
Gero Decker	HPI, University of Potsdam, Germany
Guillermo Morales	CINVESTAV, Mexico
Héctor Duran Limón	University of Guadalajara, Mexico
Helena Graziottin-Ribeiro	University of Caxias do Sul, Brazil
Henry Muccini	University of L'Aquila, Italy
Iman Poernomo	King's College London, UK
Ingrid Kirschning	UDLAP, Mexico
Iván Olmos	BUAP, Mexico
Ivica Crnkovic	Mälardalen University, Sweden
Jaime Muñoz Arteaga	UAA, Mexico
Jair C. Leite	Federal U. of Río Grande del Norte, Brazil
Jesús Favela	CICESE, Mexico
José Ángel González Fraga	UABC, Mexico
José Arrazola	BUAP, Mexico
Juan Antonio Díaz	UDLA, Mexico
Juan Antonio Navarro	Max Planck Inst. for Software Systems, Germany
Juan-Manuel Ahuactzin	Probayes, SAS, France
Julio Garibay	IBM, USA
Julita Vassileva	University of Saskatchewan, Canada
Kenneth Regan	State University of New York at Buffalo, USA
Khalid Belhajjame	University of Manchester, UK
Laurence Duchien	INRIA - University of Lille, France
Laurence Nigay	UJF, France
Leonel Morales Díaz	University Rafael Landívar, Guatemala
Luis Francisco Revilla	University of Texas, USA
Manuel Coronado Arreaga	OPEN Group, Mexico
Marcela D. Rodríguez	UABC, Mexico
María del Pilar Villamil	Universidad de los Andes, Colombia
María Isabel Sánchez Segura	University Carlos III, Spain
Marina de Vos	University of Bath, UK
Mario A. Moreno Rocha	Technologic University of la Mixteca, Mexico
Marlon Dumas	University of Tartu, Estonia
Michel Chaudron	Eindhoven U. of Technology, Netherlands
Miguel A. Redondo	University of Castilla-La Mancha, Spain
Nicandro Cruz-Ramírez	UV, Mexico
Oscar Iván Lepe Aldama	UABC, Mexico
Pablo Romero	University of Sussex, UK
Pascal Molli	LORIA, Nancy University, France
Patricia Serrano Alvarado	University of Nantes, France
Patrick Jermann	EPFL, Switzerland
Patrick Ziegler	University of Zurich, Switzerland
Pedro Santana	University of Colima, Mexico
Petr Hnetynka	Charles University, Czech Republic
Pietro Baroni	Università degli Studi di Brescia, Italia

Pilar Pozos Parra	UJAT, Mexico
Rafael Lozano	ITESM-CCM, Mexico
Raúl Aquino Santos	University of Colima, Mexico
Remco Dijkman	Eindhoven U. of Technology, Netherlands
Ricardo Pérez	Universidad Tecnológica de la Mixteca, Mexico
Roberto López Herrejon	University of Oxford, UK
Saúl de los Santos	Producen, Mexico
Silvia B. Fajardo Flores	University of Colima, Mexico
Simone Diniz Junqueira	Pontific Catholic U. of Rio de Janeiro, Brazil
Stefan Jähnichen	Fraunhofer, FIRST, Germany
Steffen Becker	FZI Karlsruhe, Germany
Sylvie Doutre	IRIT - University of Toulouse 1, France
Tania Cerquitelli	Politécnico di Torino, Italy
Thierry Delot	University of Valenciennes, France
Victor González	University of Manchester, UK
Vladik Kreinovich	University of Texas at El Paso, USA
Yann Laurillau	UPMF, France
Yannis Dimitriadis	University of Valladolid, Spain

Additional Reviewers
Árbitros adicionales

Alberto Portilla-Flores
Alessandro Fiori
Antonio Benítez
Antonio de Aescua Seco
Antonio Menendez León de Cervantes
Ariadi Nugroho
Arturo Mora-Soto
Claudia-Lavinia Ignat
Fahima Cheikh
Fernando Zacarias
François Charoy
Fuensanta Medina-Domínguez
Gaincarlo Bigi
Gallo Giorgio
Gerald Oster
Giulia Bruno
Harold Castro
José Federico Ramírez-Cruz
José Luis Carballido
Juan Carlos Nieves
Miguel Angel Palomera Pérez
Salvador Venegas Andraca
Werner Heijstek
Yulia Ledeneva

Impreso en los Talleres Gráficos
de la Dirección de Publicaciones
del Instituto Politécnico Nacional
Tresguerras 27. Centro Histórico, México, D.F.
Octubre de 2008.
Printing 500 / Edición 500 ejemplares.

This volume contains 18 carefully selected papers by 56 authors from 6 countries: Chile, France, Mexico, Peru, Spain, and USA. These papers present the most recent developments in a range of areas related to computer science and artificial intelligence. The papers are arranged into 7 thematic fields:

Computer Science

- Software Technology and Human-Computer Interfaces
- Workflow and Collaboration
- Networking

Artificial Intelligence

- Logic and Multi-Agent Systems
- Natural Language Processing and Information Retrieval
- Machine Learning and Data Mining
- Neural Networks, Image Processing, and Scheduling

The volume will be useful for researchers, students, and general public interested in the corresponding areas of computer science and artificial intelligence.

50th Anniversary of Computer Science in Mexico

ISSN: 1870-4069

www.ipn.mx
www.cic.ipn.mx

INSTITUTO POLITÉCNICO NACIONAL
"La Técnica al Servicio de la Patria"