

Fuzzy Kim. Un sistema de Minería de Datos con ayuda inteligente basado en técnicas de Soft-Computing TIC2002-04021-C02-02

Maria-Amparo Vila Miranda *
Departamento de Ciencias de la Computación de I.A.
E.T.S.I. Informática
Universidad de Granada

Jose Manuel Cadenas Figueredo †
Departamento de Ingeniería de la Información y de las Comunicaciones
Facultad de Informática
Universidad de Murcia

Abstract

The main aim of this project is the development of an intelligent system for Data Mining (DM) with the following features:

- It must be able to response to the different environments that any user may propose. Specially, it have to deal with imprecise and uncertain data and requirements.
- It must be able to provide information about the "quality" of the obtained knowledge, by giving alternative techniques to solve the arise problems
- It must be capable to learn, by recording the obtained knowledge and using it in new queries. Therefore, it should include deductive abilities.

Such a system should include a platform where the different "classical" techniques of DM are integrated together, namely those which developed by the project research groups, with a common representation schema and a unified DM language (dmlanguage). Both representation schema, dmlanguage and some DM techniques should be found in the "Soft Computing" paradigm.

Keywords:Data Mining, Soft Computing, Fuzzy Logic, Knowledge based systems, Intelligent Agents

*Email: vila@decsai.ugr.es

†Email: jcadenas@dif.um.es

1 Project objectives

1.1 General objectives, basic principles and starting points

According to the title the general objective of the project is developing a Data Mining system that provides to any user additional capacities, being different of a standard commercial system in the following points:

- Knowledge and data will be integrated, with the same representation schema which is able to deal with imprecision and uncertainty. Moreover, the could include in the data schema his/herself view of such a imprecision and/or uncertainty
- Queries about data or knowledge will be expressed in the a unified language, and/or in a unique user interface
- It is possible to reuse the acquired knowledge by querying jointly data and knowledge. To do it the systems must to have deductive capacities with imprecise data.
- in order to achieve a dynamic and adaptive user interconnection, the overall system architecture should be based in an agent philosophy.

To get a such a system we were based in the following methodology principles:

Using theoretical results developed by the project research groups

The project research groups have been doing researches about Data Mining issues since more than six years. Specially they have been interested in developing techniques which are better than the classical ones in those cases where imprecision and/or uncertainty appear. In this sense, the project uses previous results and obtains other new ones. Other results concerning to common representation and querying of imprecise data and knowledge are also to be used. Finally the project should be found the previous theoretical results concerning to agent development for Data Mining which were the origin of the METALA platform.

Reusing software products previously developed by the project research groups

Regarding this point, it should be remarked that each group had been working in software products, whose extensions and integration were the project basis:

- The Granada's group had builded a database management system which allow to deal with imprecise and uncertain data and knowledge. This system have a relational database as internal level. The database was implemented in Oracle and it was managed by PL-SQL and JAVA programs. The system DBL was FSQL, an SQL-based language which allows defining and querying imprecise data. The FSQL extension to deal with knowledge and DM issues was one of the project objectives.

- The Murcia's group had developed a software architecture, called METALA whose objective is to give a global structure to support inductive learning processes such as DM. This architecture were agent based and it should be the structure were the all of the possible extensions will be integrated.

As it can be seen both system were complementary, the first one was focused in the imprecise data and knowledge representation and querying, the second provided the needed flexible and active software architecture for DM.

Distribution, modularity and accessibility

One of our main our objectives is that the system acts in the most friendly and flexible way. Therefore, the user access environment have to be not very different of those of a modern

2 Summary of the project results

2.1 Summary of the scientific and technic results obtained by the Granada's group

2.1.1 New Data Mining theoretical results

Results concerning to attribute association

We are being focused on studying association rules issues, by extending this concept and by applying it to develop new DM techniques and/or to improve others ones. From this concept and using Soft Computing based ideas we are obtained the following results:

Fuzzy association rule: definition, properties and applications: we have extended the association rule definition to deal with imprecise data. To do it, we consider association rules where the itemsets may be fuzzy, and we compute the quality of such a rule by using fuzzy set cardinality concepts and the evaluation of linguistic quantified sentences. These results were mainly published in [18]. We have also used our definition in text mining and Web mining problems [23]

Fuzzy approximate dependencies: definition, properties and applications: these results were a part of the Ph degree thesis of the Dr. Serrano [24], and they extends the concept of approximate dependency in a relational database, previously developed by our group. It is possible to define approximate dependencies in a relational database through association rules by considering a transactional database derived from the original one. So, by means of the concept of fuzzy association rule we have defined fuzzy approximate dependencies which can be applied in the case of imprecise relational databases. An algorithm to obtain these dependencies has been also developed. These results mainly appears in [3, 4].

Fuzzy correspondence analysis: definition and applications: an important problem in DM is the knowledge fusion and a particular case is that of matching two or more classifications over the same set of objects. Special interest has the problem in the case of these classifications are fuzzy. We have defined different ways of correspondences and fuzzy correspondences, on the basis of the existence of certain approximated or fuzzy approximated dependencies. These results appears as a part of [24] and published in [15]

Results concerning to decision trees

Decision tree are probably the most popular and used classification model. They are recursively built following a top-down approach by repeated splits of the training data set. We have done researches about these methods in two main ways:

Problems of splitting rules: we have propose two new splitting rules which obtain similar result to other well-know criteria while their simplicity makes them ideal for non-expert users. These results appears in [5]. We have studied the case where continuous numerical attributes appear as variables of the training set. Usually binary split are performed by choosing a threshold values, we propose a multi-way split by clustering the domain of the considered attribute. The results appears in [1]

A new classification model: we have developed a new family of decision list algorithms based on the ideas from the association mining concepts. ART which stands for "Association Rule Tree" builds decision lists that can be viewed as degenerate decision tree. The basic idea is to generate association rules whose consequents are the class attribute values and whose

antecedent are set of values of different attributes. Quality criteria of the association rules gives the split rule and attribute values to be include in the node, the instances in the input data set which are not covered by the selected association rules are then grouped together in the "else" branches to be further processed following the same algorithm. The whole ART description appears in [2]

Many of the theoretical results above described have been implemented in the experimental platform TMINER (<http://idbis.ugr.es>)

All of the results presented above cover the work of the Granada's group at this moment in the developing of the task P

2.1.2 Researches concerning to a fuzzy data mining systems and languages

An SQL-based language for fuzzy data mining

This subject is mainly covered by the results of the Dr. Carrasco Ph degree thesis [16, 17] which were finished by the end of 2003. The most significant results of this study were:

A theoretical model GEFRED* is defined. This model is an extension of GEFRED, previously developed by our research group to represent fuzzy data in a relational database. GREFRED* extends it by dealing with any kind of attribute domain if it involves a complex fuzzy comparison operator. This enables it to be used in DM problem solving

The definition and implementation of the model dmFIRST which is the GEFRED* adaptation for implementation use. A client/server architecture has been designed for this model, it can be seen in the figure 1. The most interesting parts of this architecture are:

- The Fuzzy Metaknowledge Base for Data Mining (dmFMB) which includes all of necessary information about fuzzy data and DM issues. It appears as an extension of the database catalog, and consequently it has a relational eschema.

- The dmFSQL server which manages the dmFSQL language. It is an extension of the FSQL language defined by our research group for dealing with imprecise/uncertain data and queries. This extension is focused on solving DM queries and to store their results as imprecise/uncertain data. Consequently, dmFSQL has a closure property , by allowing an iterative DM proceses. The dmFSQL implementation includes sentences for clustering, classifying and discovering graded fuzzy functional dependencies, most of them based in our previous researches. But at this moment it does not cover all of the DM processes developed by our group

Fuzzy Queries 2+, a friendly environment for imprecise, deductive and data mining querying

Initially Fuzzy Queries (FQ) was designed as a querying client of an architecture very similar than that which appears in the figure 1. In fact it was a particular of this, where the dmFSQL server is restricted to be the FSQL server. Next, FQ was extended in several ways appearing FQ2+ whose main tools are:

- **FQbuilder:** an assistant for query builder which also provides information about the fuzzy issues on the database

- **Fdeductor:** a module for dealing with fuzzy knowledge included in the database as fuzzy rules, this implied to change the FSQL server to another new extended server (DFSQL), designed to deal with this kind of information. The bases of this extension appears en[8]

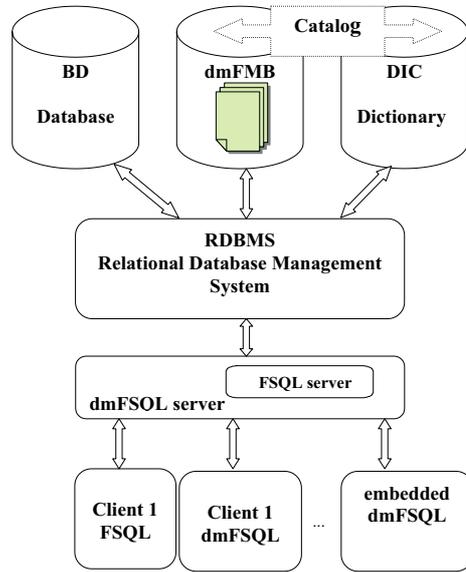


Figure 1: dmFSQL Client/server architecture

- **FDminer**: a module which allows to extract association rules, fuzzy association rules, approximating dependencies, and fuzzy approximate dependencies from the result on a fuzzy query

FQ2+ has also other important functionalities which are described in [24]. Their main advantages are the modularity and that it is very easy to use.

All of the result concerning to dmFSQL and FQ2+ cover the work of the Granada's group a this moment, in the developing of the task A.

2.2 Summary of scientific and technic results obtained by the Murcia's group

2.2.1 Evolution of the DM platform METALA

The previous METALA version (METALA-RMI) was defined on the basis of four different abstraction layers, these are (down-top): (1) the oriented object layer (2) The middleware layer (3) the agent layer and (4) the METALA application itself. One of the major drawbacks of METALA-RMI was its high complexity. For this reason we adopted a new architecture. We were concerned both in maintaining the initial functionalities and in to simplify the structure. Furthermore we wanted to have the possibility of adding new facilities by means of a new technology. We can compare the previous structure with the new architecture (METALA-J2EE) which appears in the figure 2, in this one only three layers appear: (1) the data layer, (2) the logic process layer (3) and the METALA application layer.

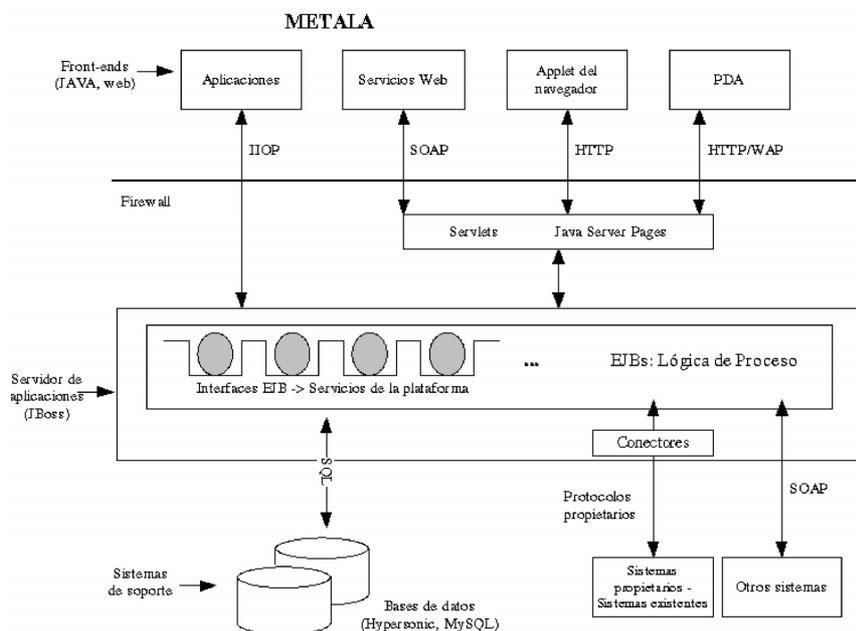


Figure 2: METALA-J2EE architecture

In the new architecture, we can see the possibilities to access to the METALA functionalities, we can do it either by means of a set of Web services which are presented in a Web gateway, or other by developing front-ends based on JAVA clients, JAVA applet or PDA access clients. All the internal processes logic is managed by the Enterprise JAVA Beans (EJBs) included in the application server J2EE. Finally the necessary data support is given by a relational DMBS (mySQL at this moment). The application server also sees about all of general aspect such as: system distributing, charge balancing, services placing, database accessing and transaction control.

The new METALA architecture has been widely disseminated in different congresses. In the presentation [20] and their later extended version as book chapter, we explain its structure and how it may be used in Web mining problems. In the paper [19] we describe the platform from the point of view of the distributed computing. METALA had been also successfully used in several applications such as the case of modelling the quality perception in a video conference session, this has been presented in [13] and its extended version [14]. However the more important results related with the project appears in the Ph degree thesis of Dra. Valdes [26], where the data induction processes are studied from formal point view. Concretely we are focused in the fuzzy modelling process, as a typical example hybrid DM task. The obtained results gives us the way for automatically combining different DM techniques in a hybrid one. This will provides us a complete autonomy in designing DM experiments. A part of these results has been published in [25]

2.2.2 The Web services and Portal

After describing the general METALA structure we will deal with other important part of the project results which is that of concern with the METALA use and connectivity. These functionalities are modelled as Web services. The Apache tool Axis has been used to design these Web services, we have chosen this tool since it allows us to use the EJBs and describe the particular methods of each bean as the methods as a concrete Web service. Axis can be used in two ways: either by providing a library set which we can invoke from our system for accessing to the Web services (SOAP engine), or as an independent server where our services can be registered to be called for external users. The later one is the option we have chosen for our system, where the libraries for the Web services are previously implemented and Axis acts as a independent server

At this moment the DM Web services are in a definition phase, however some of them are described. Concretely the starting learning service is already designed. A correct definition of such a services will be essential to establish the way of using the system functionalities, to share the knowledge with other platforms and to reuse the results obtained by the system users. The initial ideas for the implementation of the Web services using this technology appears in the [9]

Once the Web services are studied it is necessary to open them to the user. The most convenient way to do it is by means of a Web portal. Such a portal is being developing by using Tapestry, a tool for generating Web applications based on the view-control model. The user portal made by Tapestry consists of reusable components and this allows us a kick prototyping. At this moment we have available a first portal version which is used for checking both the services and the METALA basic components, because of this it is in permanent evolution.

2.2.3 System Agents

The supervisor agent

The implementation of this agent is almost finished. Its development is based in the tool ACLAnalyser, developed in this project context. This tool is designed for listening and interpreting the inter-agent communications, by storing all of the information which is relevant to the system and by advising of any possible fail. ACLAnalyser has been developed inside the agent platform JADE as plug-in of this platform, and can be obtained for the Web site of such a platform. The references concerning to this result are [10, 12, 11]

The evaluator and assistant agents

All these agents are in a design stage. Concretely, the assistant agent should have to be designed in accordance with the user interface of imprecise knowledge since it is the back-end of it

Several technological alternatives are being evaluated for developing these two agents. For example we could to create an special agent-box which interacts the EJB platform or directly include these agents in the EJB. In any case both two agents are being designed according to FIPA standards, to do it we are using JADE as platform for developing and evaluation. Until the end of the project the agent subsystem should be independent only when we are sure of a correct integration we will evaluate the possibility of including it in the EJB, by means of solution as BlueJADE.

2.3 Problems found and future activities

According to attained the results , at this moment the situation of the project is the following:

The task P has been successfully completed

The task A is in developing

We need to complete it:

(a) To integrate dmFSQL and FQ2+ by replacing the FSQL server with the dmFSQL one. We will call DMFQ this new tool.

(b) To finish the evaluator agent implementation and the assistant agent design.

In this point it should be remarked that, regarding some small the difficulties in the the project development, no major problems have appeared. We have only had some drawbacks in performing the integration (a), of FQ2+ and dmFSQL. Since both systems have based in a similar architecture, it could be expected that this integration will be straightforward, however the extension of data and knowledge representation formalisms is not so easy. In fact, at the user level, the data complexity makes necessary to use an structured representation, it will provide to the user with a complete and understandable view of all of information he/she manages. In short we had the following problems:

(1) To integrate at internal level the imprecise data and knowledge representation structures underlying in FQ2+ and dmFSQL.

(2) To formalize a conceptual model for imprecise knowledge and data which involves: initial data corresponding to facts, initial knowledge represented by means fuzzy intensive predicated (fuzzy rules) and derived knowledge represented by different ways: rules, trees, tables etc..

(3) To translate the conceptual structure to the internal one.

A proposal for solving (1) had been presented in the reference [7]. Regarding the problem (2) we have done a first approximation on the basis of the fuzzy object concept [22] . However the problem (3) as well as the necessity of include derived knowledge have suggested us to formulate a global representation, based on ontologies. This representation is in design phase.

The task G (integration) is to be starting

We have planned to devote the final year of the project in developing it, in accordance with the project timing.

This task implies two integration phases:

(1) The first one has no mind, it only consists of including inside the METALA services those algorithms (ART, approximate dependencies etc..) which does not involve imprecise information.

(2) The second one consists of integrating both systems. To do it, several alternatives are being considered:

- It is possible to configure DMFQ as a METALA client which directly interacts with the assistant agent. It should be a weak integration since the internal databases are managed each of them for their own system. Such a integration should be quite straightforward

- Since the both internal databases are relational model based, it is possible to think about a deeper integration, at internal level.

All these questions are in a discussion phase by the group teams and they will be analyzed and decided in the next coordination meeting.

3 Summary of data about results

Summary of Publications

<i>Type</i>	<i>Amount</i>	<i>References</i>
Phd Thesis	3	[16, 24, 26]
Papers in SCI journals	10	[1, 2, 3, 5, 6, 8, 11, 18, 22, 23]
Papers in others international journals	2	[7, 19]
Book chapters	5	[11, 14, 15, 17, 20]
International congress presentations	5	[4, 10, 12, 20, 26]
Spanish congress presentations	3	[7, 9, 13]

International projects

- The Granada´s group participates the following international projects related with this one:
 - . FIT-70000-2003-09061 PROFIT project associated to the EUREKA project IFK
 - . CIDAPA-CIACAP CYTED project.
- Also, the Murcia´s group participates in the following international project
 - . DAIDALOS (Designing Advanced network Interfaces for the Delivery and Administration of Location independent, Optimised personal Services) IST FP 6 project.

International and national contacts

- The Granada's group belongs to european excellence network KDNET (Knowledge Discovery Network).
- Both groups belong to the spanish network on Data Mining
- The Granada's group belongs the spanish network Reddb devoted to advanced information systems
- The Murcia's group belongs to the spanish network on agent technologies, Agentcities.es

Technological transfers

We have not any direct formal transfer but the PROFIT and CYTED projects above mentioned are based in the direct transfer of some DM results. In a informal way, we are starting a collaboration with the gateway "PULEVA SALUD" to apply some web mining results.

Teaching activities

- A this moment the project have produced three Phd thesis
- We hope to have finished the Phd thesis of Sr. Hernansaez and Sra. Martinez-Cruz, grant holders in charge of the project, by the end of the year 2005.

References

- [1] BERZAL F. CUBERO J.C., MARIN N SANCHEZ D. Building multi-way decision trees with numerical attributes *Information Sciences* Volume 165, pages 73-90, 2004
- [2] BERZAL F. CUBERO J.C., SANCHEZ D.,SERRANO J.M ART: A hybrid classification model *Machine Learning* Volume 54, Number 1, Pages 67-92,January 2004
- [3] BERZAL F., BLANCO I., SANCHEZ D., SERRANO J.M., VILA M.A. A definition for fuzzy approximate dependencies.*to appear in Fuzzy Sets and Systems* 2004
- [4] BERZAL F., CUBERO J.C., SANCHEZ D., SERRANO J.M., VILA M.A., Finding fuzzy approximate dependencies within STULONG data. *Proceedings of the ECML/PKDD 2003 workshop on discovery challenge*. Pp 34-46. 22-26 September 2003
- [5] BERZAL F., CUBERO J.C., CUENCA F. MARTIN-BAUTISTA M.J. On the quest for easy-to-understand splitting rules *Data and Knowledge Engineering* V 44 pp 31-48 2003
- [6] BERZAL F., MARIN N., PONS O., VILA M.A. Development of applications with fuzzy objects in modern programming platforms *to appear in International Journal on Intelligent System* 2004
- [7] BLANCO I.J., MARTINEZ-CRUZ C., VILA M.A., SERRANO J.M., Servidor de Bases de Datos Relacionales Difusas para Deduccion y Minería de Datos. *Actas del XII Congreso Español sobre Tecnologías y Logica Fuzzy ESTYLF* 2004, Jaen (España), 15-17 de Septiembre del 2004, pp. 135-140 (an extended version to appear in *Mathware and Soft Computing*)
- [8] BLANCO, I.; MARTIN BAUTISTA, M.J.; PONS, O.; VILA, M. A. A tuple-Oriented algorithm for deduction in a fuzzy relational database. *International Journal of Uncertainty and Fuzzy Knowledge Based Systems* 2003, 11: 47-66.
- [9] BOTÍA J.A., CABALLERO A., GÓMEZ-SKARMETA A.F. El papel de los agentes fipa en aplicaciones basadas en servicios web. In *Conferencia de la Asociación Española de Inteligencia Artificial (CAEPIA'2003)*, San Sebastián, Spain, November 2003. short paper.
- [10] BOTÍA J.A., HERNANSAEZ J.M., GÓMEZ-SKARMETA A.F.. Towards an approach for debugging mas through the analysis of acl messages. In *Proceedings of the German conference on Multi-agent system TechnologieS* 2004.
- [11] BOTÍA J.A., RUIZ P., SÁNCHEZ J.A., GÓMEZ-SKARMETA A.F.Towards an approach for debugging mas through the analysis of acl messages (to appear). CRL Publishing Ltd, 2005.
- [12] BOTÍA J.A., LOPEZ A., GÓMEZ-SKARMETA A.F... Aclanalyser: a tool for debugging multi-agent systems. In *European Conference on Artificial Intelligence (ECAI 2004)*, 2004.
- [13] BOTÍA J.A., RUIZ P., SÁNCHEZ J.A., GÓMEZ-SKARMETA A.F. Comunicación multimedia p2p adaptativa mediante aprendizaje híbrido. In *Conferencia de la Asociación Española de Inteligencia Artificial (CAEPIA'2003)*, San Sebastián, Spain, November 2003.

- [14] BOTÍA J.A., RUIZ P., SÁNCHEZ J.A., GÓMEZ-SKARMETA A.F. Adaptive p2p multimedia communication using hybrid learning. In R. Conejo, eds., *LNAI 3040*. Springer, 2004.
- [15] CALERO J., DELGADO G., SANCHEZ D., SERRANO J.M., VILA M.A., A Proposal of Fuzzy Correspondence Analysis Based On Flexible Data Mining Techniques. Eds. Lopez-Diaz M., Gil M.A., Grzegorzewski P., Hryniewicz O., Lawry J. *Soft Methodology And Random Information Systems*. Pp. 447-454. Springer 2004.
- [16] CARRASCO R, Lenguajes e interfaces de alto nivel para "Data Mining" con aplicacion practica a entornos financieros. TESIS DOCTORAL DIRECTOR Vila M. A. 2003
- [17] CARRASCO, R.; VILA, M. A.; GALINDO, J. Fsql: a flexible query language for data mining. *Enterprise Information Systems IV* Dordrech, Holanda: Kluwer Academic Publisher, 2003, p.68-74.
- [18] DELGADO, M.; MARIN, N.; SANCHEZ, D.; VILA, M. A.. Fuzzy association rules: general model and applications. *IEEE Transactions on Fuzzy Systems* 2003, 11: 214-225.
- [19] HERNANSAEZ J.M., BOTÍA J.A., GÓMEZ-SKARMETA A.F.. A j2ee technology based framework for web mining. *Revista Colombiana de Computación*, 2004.
- [20] HERNANSAEZ J.M., BOTÍA J.A., GÓMEZ-SKARMETA A.F. A j2ee technology based distributed software architecture for web usage mining. In *International Conference on Internet Computing 2004 (IC'04)*. Special Session on Web Mining., Las Vegas, June 2004. (an extended version). *Idea Group Publishing* August 2004.
- [21] JIMENEZ F. , CADENAS J.M., VERDEGAY J.L., SANCHEZ G. Solving fuzzy optimization problems by evolutionary algorithms. *Information Science* 152., pp. 303-311. 2003.
- [22] MARIN, N., MEDINA, J.M., PONS, O., SANCHEZ, D., VILA, M. A.. Complex object comparison in a fuzzy context. *Information and Software Technology*, 2003, 45: 431-444
- [23] MARTIN-BAUTISTA, M J, SANCHEZ, D., CHAMORRO-MARTINEZ, J, SERRANO, JM, VILA, MA." Mining Web Documents To Find Additional Query Terms Using Fuzzy Association Rules" *Fuzzy Sets And Systems* Volume/Issue 148/1 Pp. 85-104, 2004
- [24] SERRANO J.M. Fusion del conocimiento en bases de datos relacionales: medidas de agregacion y resumen. TESIS DOCTORAL.DIRECTORES: Vila, M. A.; Sanchez, D. 2003
- [25] VALDES M., BOTÍA J.A., GÓMEZ-SKARMETA A.F Towards a formal framework for the specification of fuzzy modeling. In *Proceedings of the International Conference on Fuzzy Systems*, St. Louis, MO. USA., May 2003.
- [26] VALDES M. Definición de un Marco para la Especificación de Técnicas Híbridas de Modelado Difuso (METHOD). TESIS DOCTORAL. Directores: A.F. Gómez-Skarmeta and J. A. Botía. 2003.