F-META: Soft-Computing Techniques for Fuzzy Meta Search Engines Development TIN2007-67494

José A. Olivas [*]

Dpto. de Tecnologías y Sistemas de Información
Universidad de Castilla-La Mancha

Abstract

The aim of this Project is to develop, to implement and to test a prototype of the F-Meta Fuzzy Meta search engine, with an agent-based structure that allows to test in an autonomic way but in a global environment different soft computing based techniques. The goal would improve Web access and Web search. The final objective is to favour a Technology Transference process, for making possible the use of these kinds of systems in real conditions.

To accomplish this goal, in this project the tasks of reviewing existing techniques and State of Art, developing of new soft computing based mechanisms for information access on the Web, developing an agent-based meta-search engine, developing and programming the F-Meta Fuzzy meta search engine and testing the prototype, will be accomplished.

Keywords: Web search, metasearch engines, soft-computing, fuzzy logic.

1 Objectives of the project

This project is been developed by the SMILe (Soft Management of Internet and Learning) research group of the Department of Information Technologies and Systems of the University of Castilla-La Mancha. The main initial hypothesis is that it would be possible to have really fuzzy searchers, or what is the same thing, searchers with the aptitude to accomplish searches in terms of approximate meanings [1]. The main point of reference to the searchers must be the Web, not so much for elements of general searches but rather for metasearchers, because those ones use as base general search engines in the Web.

The fact of having fuzzy searchers would offer the possibility of doing interesting tests and experiments. The Artificial Intelligence is an area where different logics are mixed, because the approximations in the formal analysis of a phrase can be different. Then, the logical form of the following phrase, a bit long but not surprising: "I suppose that you believe that I will call you a bit later", it implies the use of different logics: belief, non monotonic, fuzzy, temporal… But the problem is more complex, because for the words with vague meaning there can be several modalities of fuzzy logics. The election is not a factor that has been studied too much. The metasearchers might provide a very useful tool to investigate in the variety of formalisms that fuzzy logic provides. The use of user profiles in web metasearchers might provide some advantages to

---

[*] Email: joseangel.olivas@uclm.es

improve the search. The user profiles can be another parameter to keep in mind to expand the queries (with concepts related to the profile: synonymous, etc.), to select the search engines and to adapt the queries to them and to choose and to establish a ranking among the search results. Soft-Computing techniques can help in learning and representation tasks. Therefore, this project tries to describe what a fuzzy web metasearcher is, which is the role that fuzzy logic can play as Soft-Computing technique to improve the search in this type of tools, and to develop the prototype of the F-Meta metasearch engine.

Fuzzy Logic can provide tools for the extraction and use of knowledge from thesaurus and ontologies, allows to formalize sentences and to implement deduction capabilities in Question-Answering Systems, to combine fuzzy values and different logics, to design clustering algorithms and to handle the different architectures of a metasearcher.

Meta Question-Answering Systems?, probably the next aim might be to develop Meta Web Question-Answering Systems, which analyze the question of the user and generate a set of precise queries (expanded queries) that send to the directory or the most suitable search engine, to achieve the response adapted to the question of the user. Soft-Computing technologies and principally Fuzzy Logic, as tool closer to human expressions, can play an essential role to detect the intention and the correct meaning of the questions of the user to the system.

The proposed chronogram of F-Meta is shown in table I.

Activities/Tasks	First year	Second year	Third year
1. Review of the technologies developed till now and of the state of knowledge on the topic.			
1.1 Criteria unification.	▓░░		
1.2 Review of the formalisms.	▓▓▓		
1.3 Bibliographical review.	▓▓▓		
2. Development of new mechanisms of access to the information in the Web based on Soft Computing			
2.1 Development of new mechanisms of access to the information in the Web based on Soft Computing	░▓▓	▓▓▓	░░
3. Design, integration and implantation of the prototype of the platform based on agents F-Meta			
3.1 Components integration.		░▓▓	▓░
3.2 Implantation.		░▓▓	▓░
4. Evaluation of the prototype of F-Meta			
4.1 Verifications of the formalisms and the architecture.			▓░
4.2 Comparison with other tools using TREC.			▓░
4.3 Usability, usefulness (utility) and scalability tests.			▓░

Table I: F-Meta Chronogram.

F-Meta is a research prototype, but we said that one of the main goals of this project is to favour the transference of technology to the real market. For this reason, all the F-Meta developments have been simultaneously implemented in the BUDI (BÚsqueda Difusa de la Información), as a metasearch engine shell. All the results achieved in the F-Meta project are implemented in BUDI. BUDI is the result of the R+D work carried out by the University of Castilla La Mancha and the CSD Company for Software and Development (Valencia), with the financial support of the Spanish Ministry of Industry, Tourism and Commerce under the auspices of the PROFIT program (FIT-340001-2007-4).

2 Level of success achieved in the project

The need of any user when they consult a documentary database, whether departmental or general such as the web, is to obtain relevant results that provide them with the knowledge necessary to cover their requirements. To this end, the system must be able to carry out a series of previous and differentiated processes for the correct extraction of information. BUDI arises in response to this urgent need of today's world with the aim of improving the results of current search processes in document repositories, through the use of tools for the extraction and use of knowledge from thesauruses (lists of words or terms used to represent concepts) and ontologies (common and shared understanding of a knowledge domain), fuzzy linguistic modelling and clustering (vector grouping) algorithms.

- What is BUDI?
  - o BUDI is a software component developed in Java that enables us to improve, by means of semantic mechanisms, the search in all types of documentary repositories.
  - o BUDI provides developers with an API (Application Programming Interface) to integrate semantic search into their applications.
  - o BUDI lets us include new adaptors to be able to carry out searches through any engine.

- What are the techniques used in BUDI?
  - o Massive extraction of results.
  - o Fuzzy clustering based on thematic meaning groups.
  - o Fuzzy association rules (subsequently defined) to search for new terms.
  - o Document representation using fuzzy extensions of the standard vector space model.
  - o Metasearch engines: Merging of results provided by other browsers and/or specific search engines.
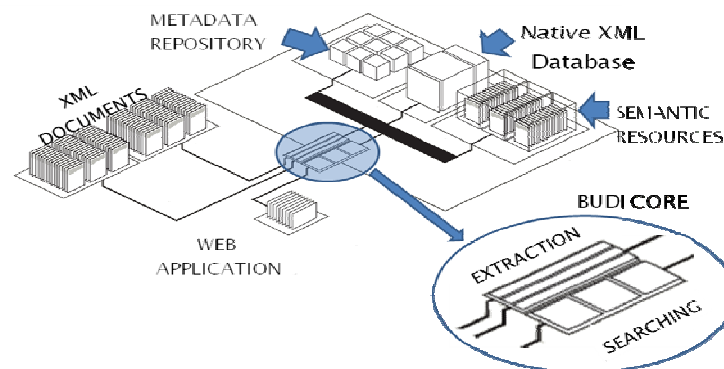

Figure 1. BUDI Architecture.

BUDI is designed to include an add value to the search in different documentary frameworks through the use of several fuzzy logic techniques. It is focused on global applicability, not limited to its interaction with Web browsers or documentary databases. Any of the components that make up BUDI can easily be interchanged to adapt the search engine to different scenarios, modifying a component for specific applications for efficiency purposes or adding new functionalities to the

engine by means of adaptors. BUDI has a series of operative modules that carry out key system tasks, such as the transformation of queries and assessment of results. These modules interact with key elements such as the search engine or the end client application which displays the results to the user. There are also management or support modules to deal with the system's internal tasks, such as handling the semantic resources (term sets, synonymies, ontologies, etc.) and configuration.

- F-Meta/BUDI keeps evolving

The BUDI search engine continues to evolve in functionality and enhancement, through new R+D contributions from the University of Castilla La Mancha and CSD (Company for Software and Development). Recently (October 2009), the Spanish Ministry of Industry, Tourism and Trade (MITYC) approved the FOG (Fuzzy Ontology Generation, TSI-020302-2009-55) project for the Generation of Fuzzy Ontologies [2] in the frame of the AVANZA R+D sub-program, to be carried out in collaboration by *Tea Cegos S.A.* and *Colegio territorial de arquitectos de Valencia* as tester entities. FOG is a new system that lets us harness the capabilities of thematic search engines or web meta-browsers by providing them the availability of ontologies perfectly adapted to their search objectives. This way, the search services necessary for the so-called Internet of the Future are provided. Within the FOG framework, steps will be taken so that the provision of semantic features becomes an automated process that lets us model the implicit knowledge stored in the databases or document repositories in which the searches are run, with no need for human intervention by an expert. The ontologies, as well as representing knowledge, also allow us to include deductive rules that can enable the company to take intelligent or guided decisions. So, thanks to the tasks to be performed in the frame of the FOG project, a search engine like BUDI would be equipped not only with structured data with which to interact, but also rules enabling it to make deductions in order to take decisions coherently.

- FIDELIO and FIS Proposals

To continue this research line, we proposed in February 2010 to the official announcement the project FIDELIO: Fuzzy Information Filtering Techniques applied to Metasearch Engines, as continuation of F-Meta. The main objective of this project is the application of fuzzy information filtering techniques on metasearch engines in order to retrieve more relevant results for the user, especially in within environments where the underlying contents change dynamically. Examples of this kind of applications are the Web, blogs, e-mail, etc. For this purpose a review of the main techniques and strategies in the field of Soft Computing is planned. Strategies such as clustering, information filtering based on user profiles, knowledge representation based on prototypes, automatic ontology generation or pseudo/relevance feedback can be some examples. Once the review has been made, a platform will be developed whose main goal is the implementation of the main methods used to organize dynamic document collections in an automatic way by means of fuzzy content-based structures. The evaluation of the system performance will be tested by the application of the built structures to the search process carried out by metasearch engines. Once the application is developed and the methods are tested then the last step will be the technology transfer in order to transmit the acquired knowledge to the society.

We are also preparing a proposal to the FIS (*Fondo de Investigaciones Sanitarias*) with Mapfre, to apply the F-Meta research to medical documents.

3 Results Indicators

- Relevancy and originality of the obtained results.

The work carried out is also corroborated by the investigations on the topic developed by different groups of international relevancy, such as the SCI2S (Soft Computing and Intelligent Information Systems) of the University of Granada (a narrow relation is kept). Our research group SMILe has an extensive and prolific experience in the application of Soft Computing's technologies for the solution of problems in the field of Information Retrieval in Internet, field in which SMILe has turned into a reference group with the organization of thematic special sessions in relevant conferences, such as for example:

- o *Soft approaches to information access on the Web*, 12th Int. Conference IPMU'08: Information Processing and Management of Uncertainty in Knowledge-based Systems, Torremolinos, Málaga, june 2008. Organizers: Enrique Herrera-Viedma, University of Granada, Gabriela Pasi, Università degli Studi di Milano Bicocca, Italy, Jose A. Olivas, University of Castilla La Mancha.
- o *Soft Computing y Recuperacion de Informacion en Internet*, ESTYLF'08: XIV Congreso Español sobre Tecnologías y Lógica Fuzzy, September 2008. Organizers: Jose A. Olivas, University of Castilla La Mancha, Enrique Herrera-Viedma, University of Granada, María José Martín Bautista, University of Granada, Alejandro Sobrino, University of Santiago de Compostela.
- o *Workshop: Soft approaches to information access on the Web (SAIAW 09)*, 2009 IEEE/WIC/ACM International Conference on Web Intelligence (WI-09), Milan, Italy, September 2009. Organizers: Guy De Tré, Ghent University, Belgium, Enrique Herrera-Viedma, Granada University, Jose A. Olivas, University of Castilla La Mancha, and Slawomir Zadrozny, Polish Academy of Sciences, Poland.
- o *Soft Computing y Recuperacion de Informacion en Internet*, ESTYLF'10: XV Congreso Español sobre Tecnologías y Lógica Fuzzy, Punta Umbría, Huelva, February 2010. Organizers: Jose A. Olivas, University of Castilla La Mancha, Enrique Herrera-Viedma, University of Granada, María José Martín Bautista, University of Granada, Alejandro Sobrino, University of Santiago de Compostela.
- o *II Workshop on Soft approaches to information access on the Web (SAIAW 10)*, 2010 IEEE/WIC/ACM International Conference on Web Intelligence (WI-09), Toronto, Canada, September 2010. Organizers: Guy De Tré, Ghent University, Belgium, Enrique Herrera-Viedma, Granada University, Spain, Jose A. Olivas, University of Castilla La Mancha, and Slawomir Zadrozny, Polish Academy of Sciences, Poland.
- o *Soft Computing y Recuperación de Información en Internet*, CEDI 2010, September 2010, Valencia. Organizers: Carlos Porcel, University of Jaen, Antonio Gabriel López Herrera, University of Granada, Francisco P. Romero, University of Castilla La Mancha, Jesús Serrano-Guerrero, University of Castilla La Mancha.

- Scientific and Technological Production.

These achieved results are clearly presented in three doctoral theses advised by J. A. Olivas in the framework of the F-Meta project:

- o "Metodología basada en técnicas de soft-computing para la organización y gestión de documentos de naturaleza heterogénea (Methodology based on soft-computing technologies for the organization and management of documents of heterogeneous nature)", Francisco Pascual Romero Chicharro. Departamento de Tecnologías y Sistemas

de Información, Universidad de Castilla-La Mancha. April 2008. Qualification: Distinction *cum laude* for unanimity.

o "Fuzzy Approach to Conceptual Meaning Processing in Natural Language Documents", Andrés Soto Villaverde. Departamento de Tecnologías y Sistemas de Información, Universidad de Castilla-La Mancha. December 2008. Qualification: Distinction *cum laude* for unanimity.

o "Characterization of Soft Computing-based Semantic Distances for Internet Search", Jesús Serrano Guerrero. Departamento de Tecnologías y Sistemas de Información, Universidad de Castilla-La Mancha. September 2009. Qualification: Distinction *cum laude* for unanimity. Accreditation of "European Doctorate".

And in the publications obtained in the framework of this project, in scientific magazines and contributions to conferences, where references [3 to 21], can be emphasized.

- Utility of the results and relations with the socio-economic environment.

As results before this research, the development, implementation and evaluation of several platforms of fuzzy search and metasearch of information (FISS, GUMSE…) have been developed and presented in the conferences and magazines of the area. During the years 2005 and 2006 we led the "Red Temática Nacional sobre Sistemas de Acceso a la Información en la Web basados en Soft Computing, TIN2004-20503-E (Thematic National Network on Soft Computing based Web Information Access Systems).

Due to this scientific activity in the years of development of this project, one has met(competed) successfully to diverse national research projects official announcements, receiving funding in order to develop a prototype of fuzzy metasearcher (F-META, TIN2007-67494), to apply this metasearcher in a commercial environment with the company CSD (Company for Software and Development) of Valencia (BUDI –Búsqueda Difusa-, FIT-340001-2007-4) as well as in order to apply fuzzy technologies inside specific environments, as the medicine, with Indra (HERMES, TSI-020100-2008-155) and to use semantic technologies in order to improve the performance of the technologies of search, again with CSD after the good results obtained with BUDI (FOG-Fuzzy Ontologies Generator-, TSI-020302-2009-55).

- Formation of human resources.

In the framework of F-Meta project, three research visits were done:

o Cristina Puente: Laboratoire d'Informatique Fondamentale, Université Aix-Marseille, France, with prof. Elie Sánchez, 14 weeks, 2009, Soft-Computing and Information Retrieval.

o Jesús Serrano-Guerrero: Centre for Computational Intelligence, De Montfort University, UK, with prof. Robert John, 16 weeks, 2008, Type-2 fuzzy sets applied to Information Retrieval.

o Francisco P. Romero: Centre for Computational Intelligence, De Montfort University, UK, with prof. Robert John, 8 weeks, 2008, Representing Fuzzy Prototypes using type-2 fuzzy sets.

A researcher was taken on (Jesús Serrano-Guerrero) in the last part of the project to accomplish tests leadership.

- Collaborations with other European or international research groups.

As it has been presented, several collaborations with other international research groups has been done. Those ones can be emphasized:

- o University of Granada, profs. Enrique Herrera-Viedma, María José Martín Bautista, Antonio Gabriel López Herrera.
- o University of Jaen, prof. Carlos Porcel.
- o University of Santiago de Compostela, prof. Alejandro Sobrino [19].
- o Pontificia Comillas University, profs. Cristina Puente, Israel Alonso and Pedro López [11][19].
- o European Centre for Soft Computing, Mieres, Asturias, profs. Sergio Guadarrama, Eloy Renedo, Gracián Triviño and Enric Trillas.
- o Laboratoire d'Informatique Fondamentale, Université Aix-Marseille, France, with prof. Elie Sánchez.
- o Centre for Computational Intelligence, De Montfort University, UK, with prof. Robert John.
- o Università degli Studi di Milano Bicocca, Italy, prof. Gabriela Pasi.
- o Ghent University, Belgium, prof. Guy De Tré.
- o Polish Academy of Sciences, Poland, prof. Slawomir Zadrozny.
- o Universidad Nacional del Carmen, Mexico, prof. Andrés Soto [2][4][8][9][18].
- o Universidad Autónoma de Yucatán, Mexico, profs. Victor Hugo Domínguez and Alfredo Zapata [21].
- o Universidade Federal de Viçosa, Brazil, profs. José Luis Braga and Mateus Ferreira-Satler [16][21].

- Development and management of the project.

There were no problems in the management and development of the project.

4 References

[1] Olivas, J. A.: *Fuzzy Sets and Web Meta-search Engines.* In H. Bustince, F. Herrera, J. Montero (eds), Fuzzy Sets and Their Extensions: Representation, Aggregation and Models, Studies in Fuzziness and Soft Computing vol. 220-2008 Springer Verlag, pp. 537 – 552, 2008.

[2] Viñas, E. F., Serrano-Guerrero, J., Olivas, J. A., de la Mata, J., Soto, A.: "FOG: Arquitectura flexible para la generación automática de ontologías". Actas del XV Congreso Español sobre Tecnologías y Lógica Fuzzy. ESTYLF 2010, pp. 525-530, 2010.

[3] Romero, F. P.; Olivas, J. A.; Garcés, P.: *FzMail: Using FIS-CRM for e-mail classification.* Journal of Advanced Computational Intelligence and Intelligent Informatics (JACIII), Vol.11, No.1, pp. 40 - 50, 2007,

[4] Romero, F. P.; Peralta, A.; Soto, A.; Olivas, J. A.; Serrano-Guerrero, J.: *Fuzzy Optimized Self-Organizing Maps and their Application to Document Clustering,* Soft Computing, 2009. (Online).

[5] Serrano-Guerrero, J.; Romero, F. P.; Olivas, J. A. J. de la Mata, A. Soto: BUDI: *Architecture for Fuzzy Search in documental Repositories,* Mathware & Soft Computing 16, pp.71-85, 2009.

[6] Soto, A.; Olivas, J. A.; Prieto, M.: *Using Generalized Constraints and Protoforms to deal with adverbs.* Proc. of EUSFLAT-07, 5th Conference of the European Society for Fuzzy Logic and Technology, 2007.

[7] Romero, F. P.; Olivas, J. A.; Garcés, P. J.: *Inference based on Fuzzy Deformable Prototypes for Information Filtering in Dynamic Web Repositories.* Proc. of the FUZZ-IEEE 2007, IEEE International Conference on Fuzzy Systems, 2007.

[8] Soto, A.; Olivas, J. A.; Prieto, M. E.: *Using Generalized Constraints and Protoforms to Deal with Adjectives*. Proc. of the FUZZ-IEEE 2007, IEEE International Conference on Fuzzy Systems, 2007.

[9] Romero, F. P.; Soto, A.; Olivas, J. A.: *A Hybrid Model for Document Clustering based on a Fuzzy Approach of Synonymy and Polysemy*. Proc. of the 12th International Fuzzy Systems Association World Congress IFSA 2007, 2007.

[10] Carretero, A.; Romero, F. P.; Olivas, J. A.: *A Representation Model based on Wordnet Meanings for Internet Search*. Proc. of the Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU'08), pp. 1076-1083, 2008.

[11] Puente, C., Olivas, J. A.: *Analysis, detection and classification of certain conditional sentences in text documents*. Proc. of the Information Processing and Management of Uncertainty in Knowledge-based Systems (IPMU'08), pp. 1097-1104, 2008.

[12] Romero, F. P.; Genero, M.; Olivas, J. A.; Piattini, M.: *Automatic Extraction of the Main Terminology used in Empirical Software Engineering through Text Mining Techniques*, Proc. of the 2nd International Symposium on Empirical Software Engineering and Measurement ESEM 2008, pp. 357-358, 2008.

[13] Serrano-Guerrero, J.; Romero, F. P.; Olivas, J. A.: *Filtering Short Queries by means of Fuzzy Semantic-Lexical Relations for Meta-Searchers using WordNet*. Proc. of the the IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology (WI-IAT 2008). ACM/IEEE, pp. 269-272, 2008.

[14] Romero, F. P.; Serrano-Guerrero, J.; Olivas, J. A.: *Bucefalo: A Tool for Intelligent Search and Filtering for Web-based Personal Health Records*. Proc. of the 18th international conference on World Wide Web WWW 2009, pp. 1077-1078, 2009.

[15] Romero, F. P.; Caballero, I.; Olivas, J. A; Verbo, E.; Serrano-Guerrero, J.: *Web-based Personal Health Records Filtering using Fuzzy Prototypes and Data Quality Criteria*. Proc. of the 2009 IEEE/WIC/ACM International Joint Conference on Web Intelligence (WI'09) and Intelligent Agent Technology (IAT'09), pp. 159-162, 2009.

[16] Romero, F. P.; Olivas, J. A; Ferreira-Satler, M.; Serrano-Guerrero, J.: *Using Fuzzy Deformable Prototypes to represent User Profiles in Information Retrieval Context*. Proc. of the EUROFUSE WORKSHOP Preference Modelling and Decision Analysis, pp. 90-96, 2009.

[17] Serrano-Guerrero, J.; Olivas, J. A; Gallego, J. A.; Romero, F. P.; Soto, A.: *CASTALIA: Architecture of a Fuzzy Metasearch Engine for Question Answering Systems*. Proc. of the ISDA 09: 9th International Conference on Intelligent Systems Design and Applications, pp. 73-78, 2009.

[18] Soto, A; Romero, F. P.; Olivas, J.A.; Serrano-Guerrero, J.: *An experiment about using copulative and comparative sentences as constraining relations*. Proc. of the ISDA 09: 9th International Conference on Intelligent Systems Design and Applications, pp. 79-84, 2009.

[19] Puente, C., Sobrino, A., Olivas, J. A.: *Representación de relaciones causales a través de un proceso semi-automático de recuperación*. Actas del XV Congreso Español sobre Tecnologías y Lógica Fuzzy. ESTYLF 2010, pp. 519-524, 2010.

[20] Ruiz-Morilla, J.; Serrano-Guerrero J.; Olivas, J. A.; Viñas, E. F.: *Representación Múltiple de Consultas: Una alternativa a la Expansión de Consultas en Sistemas de Recuperación de Información*. Actas del XV Congreso Español sobre Tecnologías y Lógica Fuzzy. ESTYLF 2010, pp. 531-536, 2010.

[21] Ferreira-Satler, M.; Menéndez, V. H.; Romero, F. P.; Zapata A.; Prieto, M. E.: *Ontologías borrosas para representar perfiles de usuario en una herramienta de gestión de objetos de aprendizaje*. Actas del XV Congreso Español sobre Tecnologías y Lógica Fuzzy. ESTYLF 2010, pp. 537-542, 2010.