Jornada de Seguimiento de Proyectos, 2005
Programa Nacional de Tecnologías Informáticas

GLIRS-II: An information retrieval system based on fuzzy linguistic information and genetic algorithms TIC2003-07977

Enrique Herrera Viedma*
Dept. Of Computer Science and A.I
University of Granada

Abstract

In most existing Information Retrieval Systems (IRSs) users do not participate in the retrieval activity and systems have not appropriate tools to express correctly user queries. The main consequence is that the system performance and users' satisfaction degree are disminished. The Fuzzy Linguistic Modeling and Genetic Algorithms (GA) are Soft Computing tools useful for the design of IRS. The first one to manage the qualitative information in the formulation of queries by faciliting users the expression of their information needs, and GA for the optimization of queries as a tool that improves the relevance feedback (tool to define queries that allows the users participation in the retrieval processes). Our idea is to study the application of both tools in the design of IRSs to combine their advantages and to improve the effectivity of IRSs and the users' satisfaction degree.

Keywords: information retrieval, relevance feedback, query language, soft computing, fuzzy linguistic modelling, genetic algorithms.

1 Goals of the Project

This research project is a continuation of the project TIC-2002-03276 (GLIRS: Information retrieval systems based on fuzzy linguistic information and genetic algorithms). In that project a basic study on the development of IRS based on fuzzy linguistic information and GA is being carried out. In this occasion, we propose the development of a IRS based on fuzzy linguistic information and GA, called GLIRS-II (Genetic Linguistic Information Retrieval System), that incorporates more tools to help users to express better their information needs and better tools of relevance feedback based on GA. To do so, we focus on the developing the following specific goals:

- The design of a more expressive fuzzy linguistic multi-level weighted query subsystem that represents better the information needs by means of three elements: i) a scheme of multi-level weighting to associate weights with any element of a query(terms, sub-expressions, Boolean connectives and whole query), ii) the possibility of managing multi-granular and unbalanced fuzzy linguistic information to assess weights, and iii) an appropriate interface.

* Email: viedma@decsai.ugr.es

- The design of an evaluation subsystem of multi-level linguistic weighted queries that allows to manage multi-granular and unbalanced linguistic information.

- The design of relevance feedback tools based on GA for the optimization of multi-level linguistic weighted queries. We study two research areas: i) the use of new GA based strategies to model the relevance feedback tools, and ii) the design of new fitness functions, selection mechanisms and crossover operators more appropriate for relevance feedback.

- Implementation and evaluation of GLIRS-II. GLIRS-II will be a prototype of an IRS that integrates the above proposals. In such a way, GLIRS-II will be the issue of the project that stores all our proposals.

On the other hand, to develop our project we have the tools necessary to achieve the above goals. In particular, we have a good library to study the different advances about the development of information retrieval based on fuzzy linguistic modelling and genetic algorithms, we also have some computers to implement the different models of information retrieval systems, and finally, we also have some collections of documents to validate such models (CRANFIELD,TREC y CACM).

Finally, the approximate schedule of the project is the following:

Activities	First year (*)	Second year (*)	Third year (*)
	M1 M2		
Study of the art state	x| x||||||||	|||||||||||	|||||||||||
	M3 M7		
Design of the basic prototype GLIRS-II	||| x |x| x|x||||	|||||||||||	|||||||||||
	M8	M16	
Implementation of basic prototype GLIRS-II	|||||| |x|x|x |x|x	x|x|x|x| |||||||	|||||||||||
		M13 M20	
Development of multi-level linguistic weighted query subsystem	|||||||||||	x| x| x|x |x|x | x| x| |||	|||||||||||
		M16 M22	
Development of multi-level linguistic weighted query evaluation subsystem	|||||||||||	||| x |x | x| x|x| x| |	|||||||||||
		M23	M29
Development of relevance feedback tools	|||||||||||	|||||||||||	x|x|x|x|x ||||||
Based on GA			
			M30 M36
Experimentation, evaluation of GLIRS-II	|||||||||||	|||||||||||	|||| |x| x |x|x|x |x |x

2 Level of Success

The level of success achieved through the carrying out project has been important and very high. We have solved the nearly all the goals by implementing several prototypes of the different information retrieval

system models and an basic prototype of GLIRS-II. In the last phase of the project, we are integrating all the advances of those models in GLIRS-II.

In the following subsections, we analyze briefly how we have solved each goal, the main problems that we have detected, and the relevant results generated in the project.

2.1 The Design of a More Expressive Fuzzy Linguistic Multi-Level Weighted Query Subsystem

We have developed a fuzzy linguistic multi-level weighted query subsystem by means of design of different models of information retrieval systems (IRSs).

Firstly, we have developed an IRS model that allows the use of linguistic multi-level weighted queries, in such a way, that a user could use any element of a query (atoms, operators, sub-expressions, whole query) to express his information needs. In this IRS model we use always the same linguistic term set. Then, we have designed some IRS models with multi-level weighted queries but with multi-granular linguistic information, that is assuming different term sets to assess the different concepts of the activity of a IRS, and with unbalanced linguistic information, that is assuming linguistic term sets whose distribution is not uniform and symmetrical with respect to the mid linguistic term. In both cases, the main problem is how to represent in a uniform way the different linguistic information. We have designed two methodologies to manage the multi-granular and unbalanced linguistic information using the 2-tuple linguistic representation model [8].

We have observed that these fuzzy linguistic query languages can be easily applied in the search engines on the Web, and then, we have proposed some models of information gathering distributed systems on the Web that use such query languages as a way to increase the expression possibilities of information needs.

Studying the information retrieval problem on the Web we find out that there exist two different approaches to focus the information access processes on the Web [1, 5, 9]: i) approaches based on the Classical Information Retrieval which retrieve information from user queries and provide relevance scores, and ii) approaches based on the Information Filtering tools which retrieve information from user profiles and provide recommendations. On the Web community it is known that information access models combining both approaches are a promising tool to improve the users' satisfaction degrees and the performance of information access systems. Then, in order to improve the performance of the Web search systems we have studied how to combine both, that is, IRS models together with filtering tools using linguistic modelling to represent the users' information needs and profiles. We also have studied how to generate recommendations in the filtering tools facilitating the user participation in the quality evaluation procedures of the objects/documents to recommend by means of the fuzzy linguistic modelling.

In all above models, the big problem to solve is to design adequate interfaces that allow users to provide their information needs by means of the linguistic information. We are designing a graphic interface that allows users to express their preferences by means of simple clicks on the mouse.

2.2 The Design of an Evaluation Subsystem of Multi-Level Linguistic Weighted Queries

For each one of the above fuzzy linguistic query languages we have defined its respective evaluation subsystem. To do that, we have defined in each IRS model three elements:

1. A strategy of evaluation of multi-level weighted queries to apply the different semantics associated with the query weights in a consistent way.
2. Some linguistic matching functions to interpret each one of the semantics of the query weights, and
3. appropriate soft aggregation operators of multi-granular and unbalanced linguistic information. We have defined such aggregation operators using 2-tuple linguistic hierarchies [3] and the family of the OWA operators [6, 7, 10, 11].

2.3 The Design of Relevance Feedback Tools Based on GA

We have studied how to improve the learning of the Boolean user queries and multi-weighted user queries by means of the multi-objetive genetic algorithms [2, 4]. In such a way, we have provided new tools to improve the relevance feedback in the formulation of the queries. In particular we have used the mutli-objetive genetic algorithm called SPEA [12]. With these tools we improve the GA based procedures to learn queries existing in the literature.

We have observed that in the case of learning of multi-weighted user queries, sometimes we obtain a poor performance. Now, we are studying how to tune such algorithms by improving some of the usual operations of the GAs, as for example, the mutation, the crossover, the generation of the initial population, etc.

2.4 Implementation and Evaluation of GLIRS-II

Initially, we designed the basic architecture of system GLIRS-II, by incorporating the different parameters and variables (for example, different linguistic term sets with different cardinality, different fuzzy linguistic quantifier to compute the weighting vector of the OWA operators, different databases to validate our developments, etc) necessary to implement the IRS models that we have developed in the project. When we define a new IRS model we implement it and then, it is incorporated in the system GLIRS-II. As aforementioned, now we are designing an appropriate interface to facilitate users the use of the different query languages proposed and also to facilitate the use of the system GLIRS-II.

2.5 Relevant Results

Here, we are going to show the more relevant results that we have provided until this moment:

- We have introduced the concept of multi-weighted query language. We have defined different fuzzy linguistic query languages that allow to weigh the different levels of a query, to use linguistic weights, and to use multi-granular and unbalanced linguistic information. They can be consistently used in the search engines on the Web to improve the specification of the user information needs.
- We have defined for each multi-weighted fuzzy linguistic query language its respective evaluation method of queries that allows to evaluate coherently the different types of multi-weighted user queries.
- We have defined new semantic linguistic interpretation functions for the query weights to soften their evaluation.
- We have defined a method to help users in the formulation of their queries based on multi-objective GA, which improves the learning of the concept of user relevance with respect to other genetic proposals existing in the literature.

- We have defined a method based on multi-objetive genetic algorithms to learn simple Boolean user queries and user queries with multiple weighting levels.

3 Indicators of Results

The main indicators that we can show about the results of the basic Research Project GLIRS-II are classified in the following categories: formation of students, publications, spanish projects, international and spanish collaborations.

3.1 Formation of Students

The results generated in this category have been important. A doctoral thesis has been defended and three PhD students have presented their research projects as a previous step to present the next year three doctoral thesis. In particular:

1. Dª. María Luque Rodríguez, member of Project and associate professor in Cordoba University, has finished her doctoral thesis titled "*Information Retrieval Models Based on Fuzzy Linguistic Information and Evolutionary Algorithms. Improving the Representation of Information Needs*" with Drs. Oscar Cordon and Enrique Herrera-Viedma as advisors. She presented it by March 2005.
2. D. Antonio Gabriel Lopez Herrera, PhD student in the Department of Computer Science and Artificial Intelligence (DECSAI), has presented his research Project titled "*Information Retrieval Models Based on Fuzzy Linguistic Information*" with the Dr. Enrique Herrera Viedma as advisor. He presented it by July 2005. He will finish his doctoral thesis the next year 2006.
3. D. Carlos Porcel Gallego, programmer in the UGR, has presented his research Project titled "*Information Access Models Based on Fuzzy Linguistic Information and Filtering Tools*" with the Dr. Enrique Herrera Viedma as advisor. He presented it by July 2004. He will finish his doctoral thesis at the end of this year 2005.
4. D. Jose Manuel Morales del Castillo, PhD student in the Department of Library Science, will present by September 2005 his research project titled "*Improvement of the Information Access on the Web Using Semantic Web and Fuzzy Linguistic Tools*" with the Drs. Eduardo Peis (member of the project) and Enrique Herrera Viedma as advisors. He will present his doctoral thesis the next year 2006.

3.2 Publications

The main publications correspond to the following categories:

- **Articles in international journals**

1. *E. Herrera-Viedma, O. Cordon, M. Luque, A.G. Lopez, A.M. Muñoz.* **An Information Retrieval System Based on Multi-Granular Linguistic Information.** International Journal of Approximate Reasoning 34 (3) (2003) 221-239.
2. *E. Herrera-Viedma, F. Herrera, L. Martínez, J.C. Herrera, A.G. Lopez.* **Incorporating Filtering Techniques in a Fuzzy Linguistic Multi-Agent Model for Gathering of Information on the Web.** Fuzzy Sets and Systems 148 (1) (2004) 61-83.
3. *E. Herrera-Viedma, A.G. Lopez-Herrera C. Porcel.* **Tuning the Matching Function for a Threshold Weighting Semantics in a Linguistic Information Retrieval System.** International Journal of Intelligent Systems 20 (2005) 921-937.

4. *O. Cordon, E. Herrera-Viedma, M. Luque.* **Improving the Learning of Boolean Queries by Means of a Multiobjective IQBE Evolutionary Algorithm.** Information Processing & Management. To appear, 2005.
5. E. Herrera-Viedma, G. Pasi, A.G. Lopez-Herrera, C. Porcel. **Evaluating the Information Quality of Web Sites: A Methodology Based on Fuzzy Computing with Words.** Journal of American Society for Information Science and Technology (JASIST). To appear, 2005.
6. E. Herrera-Viedma, A.G. Lopez-Herrera, M. Luque, C. Porcel**. A Fuzzy Linguistic IRS Based on 2-Tuple Modelling.** International Journal of Uncertainty, Fuzziness and Knowlegde-Based Systems. Submitted, **2005**.
7. E. Herrera-Viedma, E. Peis, J.M. Morales-del-Castillo, D. Gómez. **On the quality of Web sites based on XML documents: An evaluation model using fuzzy computing with words.** Journal of Information Science. Submitted, 2005.
8. E. Herrera-Viedma, A.G. Lopez-Herrera. **A Model of Information Retrieval System with Unbalanced Fuzzy Linguistic Information**. IEEE Trans. On Systems Man and Cybernetics, Part B. Submitted, 2005.
9. E. Herrera-Viedma. **Controlling the Retrieval of an Ordinal Fuzzy Linguistic Information Retrieval System Using a Multi-Level Weighting Scheme in the Formulation of Queries**. Journal of American Society for Information Science and Technology (JASIST). Submitted, 2005.

• Book chapters

1. O. Cordón, E. Herrera-Viedma, M. Luque. **A Realistic Information Retrieval Enviroment to Validate a Multiobjective GA-P Algorithm for Learning Fuzzy Queries.** In Soft Computing: Methodologies and Applications, Springer Verlag, 2005, pp. 299-310.

2. E. Herrera-Viedma, E. Peis. J.M. Morales-del-Castillo. **A Fuzzy linguistic multi-agent model based on Semantic Web technologies and user profiles**. In Soft Computing for Information Retrieval on the Web, Springer. Accepted, 2005.

3. E. Herrera-Viedma, C. Porcel, F. Herrera, L.Martínez, A.G. Lopez-Herrera. **Techniques to Improve Multi-Agent Systems for Searching and Mining the Web.** In Intelligent Data Mining: Techniques and Applications, Springer Verlag, 2005, pp. 463-486.

4. F. Herrera, E. Herrera-Viedma, L. Martínez, C. Porcel. **Information Gathering on the Internet Using a Distributed Intelligent Agent Model with Multi-Granular Linguistic Information**. In Fuzzy Logic and the Internet, Springer Verlag, 2005, pp. 95-116.

• Articles in international and national conferences

1. E. Herrera-Viedma, C. Porcel, A.G. Lopez, M.D. Olvera, K. Anaya. **A Fuzzy Linguistic Multi-Agent Model for Information Gathering on the Web Based on Collaborative Filtering Techniques.** Atlantic Web Intelligence Conference, AWIC'04. Cancún(México), 2004. Lect. Notes in Artificial Intelligence 3034, pp.3-12, 2004.
2. E. Herrera-Viedma, O. Cordon, M. Luque, A.G. Lopez. **A Model of Fuzzy Multi-Granular Linguistic IRS.** IPMU 2004, Vol II pp. 1365-1362, Perugia (Italia).
3. E. Herrera-Viedma, A.G. Lopez-Herrera, L. Hidalgo. **Tuning a Linguistic Information Retrieval System.** RASC 2004, Nottingham (UK). Proc. of the 5th Int. Conf. on Recent Advances in Soft Computing, pp. 506-511, 2004.
4. E. Herrera-Viedma, O. Cordón, M. Luque. **Improving the performance of ordinal fuzzy linguistic IRSs.** Proceedings of the XII Congreso Español sobre Tecnología y Lógica Fuzzy (ESTYLF 2004), Jaén (Spain), 509-514, 2004.

5. E. Herrera-Viedma, A.G. Lopez-Herrera, L. Hidalgo. **A New Linguistic Modelling of the Symmetrical Threshold Semantics.** Proceedings of the XII Congreso Español sobre Tecnología y Lógica Fuzzy (ESTYLF 2004), Jaén (Spain), 309-314, 2004.
6. E. Herrera-Viedma. **ON AGGREGATION OPERATORS FOR INFORMATION ACCESS ON THE WEB.** Proceeding of the 3rd International Summer School on Aggregation Operators (AGOP 2005), Lugano (Suiza), 141-146, 2005.
7. E. Herrera-Viedma, A. G. Lopez-Herrera, L. Hidalgo, C. Porcel. **Improving the Computation of Relevance Degrees in a Linguistic Information Retrieval System.** Proceeding of CEDI, Granada, 2005**.**
8. E. Herrera-Viedma, A. G. Lopez-Herrera, C. Porcel. **A New Model of Linguistic Weighted Information Retrieval System,** Proceeding of EUSFLAT, Barcelona, 2005.

3.3 Spanish Projects

We have participated with other spanish research groups and enterprises in two national projects:

1. **Network about Decisión Making, Modelling and Aggregation of Preferences (REDEMAP).** Our research group is the coordinator of this spanish network in which we are collaborating with 20 or 22 research groups from different spanish universities and two enterprises (Telefonica I+D, Puleva) to join our proposals in the development of systems to help users in decision making, e.g, decision support systems, information retrieval systems, web multi-agent systems, recommendation systems, etc. We are studying with Telefonica I+D the possibility to ask an european project together with others international parnets about information filtering systems to help in decision making.
2. **Network about Soft Computing for information retrieval on the Web.** Our research group is an important research group of this spanish network. In this network, similarly, we collaborate with 15 or 20 research groups and an important enterprise, SOLUCIONA, in the development of information access systems based on Soft Computing tools. We are studying with Soluciona the possibility to ask an european project together with others international partners about information access systems on the Web based on Soft Computing tools.

On the other hand, at this moment we are organizing a regional regional project together with Dr. Jose Angel Olivas's research group, Soluciona enterprise about the use of Soft Computing tools in the design of information access systems. We shall ask it in September 2005 in the research program of the Castilla-La Mancha Community.

3.4 International and Spanish Collaborations

With respect to the international collaborations we have to point out two collaborations with two important european research groups, Dr. Gabriella Pasi's group from University of Biccoca (Milan) and Dr. Fabio Crestani's group from Univ. of the Strathclyde, Glasgow (UK). Both research groups are participating in several european projects about development of technologies to improve the information access tools on the Information and Knowledge Society. We are collaborating with them to organize in the future some european project. Until this moment our collaboration has produced the following results:

1. **Edition of a special issue** in the *Journal of American Society for Information Science and Technology (JASIST)*: E. Herrera-Viedma and G. Pasi (Eds.). Special issue on Soft approaches in information retrieval and information access on the Web. JASIST, to appear in 2006.
2. **Organization of two invited sessions together** with Dr. Gabriella Pasi about Soft Computing approaches in two International Conferences EUSFLAT 2003 (Zitau, Germany) and IPMU 2004 (Perugia, Italy).

3. **Edition of a book in Springer-Verlag**: E. Herrera-Viedma, G. Pasi and F. Crestani (Eds.). Soft approaches to information access on the Web. Springer Verlag. To appear in 2006.
4. **Edition of a special issue** in the *International Journal of Intelligent Systems (IJIS)*: G. Pasi and E. Herrera-Viedma (Eds). Special issue on Aggregation operators in the information management and access processes. IJIS, in preparation.

With respect to the national collaborations we have collaborated mainly with the Dr. Jose Angel Olivas's research group, in particular organizing two invited sessions about the use of Soft Computing Tools in Information Retrieval on the Web in the conferences:

1. National Conference on Fuzzy Logic and Technology (Estylft 2004) celebrated in Jaen in September 2004, and
2. European Conference on Fuzzy Logic and Technology (EUSFLAT 2005) to be celebrated in Barcelona in September 2005.

4 References

[1] N.J. Belkin and W.B Croft. Information filtering and information retrieval: Two sides of the same coin?. *Communications of the ACM*, 35(12) (1992) 29-38.

[2] C.A. Coello, D.A. Van Veldhuizen, G. B. Lamant. *Evolutionary Algorithms for Solving Multi-Objective Problems*. Kluwer Academic Publishers, (2002).

[3] O.Cordón, F.Herrera, I.Zwir. Linguistic modelling by hierarchical systems of linguistic rules. *IEEE Transactions on Fuzzy Systems*, 10(1) (2001) 2-20.

[4] K. Deb. *Multi-objective Optimization using Evolutionary Algorithms*. Wiley, (2001).

[5] U. Hanani, B. Shapira, P. Shoval, Information filtering: Overview of issues, research and systems. *User Modeling and User-Adapted Interaction* 11 (2001) 203-259.

[6] F. Herrera, E. Herrera-Viedma, Aggregation operators for linguistic weighted information. *IEEE Trans. on Systems, Man and Cybernetics, Part A: Systems and Humans*, 27 (5) (1997) 646-656.

[7] F. Herrera, E. Herrera-Viedma, J.L. Verdegay, Direct approach processes in group decision making using linguistic OWA operators. *Fuzzy Sets and Systems*, 79 (1996) 175-190.

[8] F. Herrera and L. Martínez, A 2-tuple fuzzy linguistic representation model for computing with words, *IEEE Transactions on Fuzzy Systems* 8:6 (2000) 746-752.

[9] P. Resnick, H.R. Varian, Guest Editors. Recommender systems. *Communications of the ACM*, 40 (3) (1997), 56-89.

[10] R.R. Yager. On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Transactions on Systems, Man, and Cybernetic*, 18 (1988) 183-190.

[11] R.R. Yager. Quantifier Guided Aggregation Using OWA Operators. *Int. J. of Intelligent Systems*, 11 (1996) 49-73.

[12] E. Zitzler, K. Deb, L. Thiele. Comparison of Multiobjective Evolutionary Algorithms: Empirical Results. *Evolutionary Computation*, 8(2) (2000)173-195.