

## Soft Computing for Information Retrieval in the WEB

E. Herrera-Viedma. M<sup>a</sup>. J. Martín-B.  
 Dep of Comp. Science and Artificial  
 Intelligence  
 University of Granada  
 viedma@decsai.ugr.es  
 mbautis@decsai.ugr.es

S. Guadarrama .  
 Dep. de Informática  
 Univ. Rey Juan Carlos  
 sergio.guadarrama@urjc.es

A. Sobrino.  
 Dep. de Lógica y  
 Filosofía Moral  
 Univ. de Santiago de  
 Compostela  
 lflgalex@usc.es

J. A. Olivás.  
 Dep. de Informática  
 Univ. de Castilla la  
 Mancha  
 joseangel.olivas@uclm.es

### **Abstract:**

#### *1. Soft Computing*

The term SC refers to a family of computing techniques that, when L.A. Zadeh -the father of fuzzy logic- introduced the topic, originally comprised four different partners: fuzzy logic, evolutionary computation, neural networks and probabilistic reasoning. The term SC distinguishes these techniques from hard computing that is considered less flexible and computationally demanding.

The key point of the transition from hard to SC is the observation that the computational effort required by conventional computing techniques sometimes not only makes a problem intractable, but is also unnecessary as in many applications precision can be sacrificed in order to accomplish more economical, less complex and more feasible solutions. Imprecision results from our limited capability to resolve detail and encompasses the notions of partial, vague, noisy and incomplete information about the real world.

In other words, it becomes not only difficult or even impossible, but also inappropriate to apply hard computing techniques when dealing with situations in which uncertainty and imprecision are involved. The guiding principle of SC is “to exploit the tolerance for imprecision, uncertainty, partial truth, and approximation to achieve tractability, robustness, low solution cost and better rapport with reality”.

All the methodologies that constitute the realm of SC (the four abovementioned and some others that have been incorporated in the last few years such as rough sets or chaotic computing) are considered complementary as desirable features lacking in one approach are present in another. Hence, the SC framework is put into effect by hybrid systems combining two or more of the constituent technologies with complementary characteristics.

#### *2. Textual Information Retrieval*

IR may be defined, in general, as the problem of the selection of documentary information from storage in response to search questions provided by a user. IR systems (IRSs) are a kind of information system that deal with data bases composed of information items -documents that usually consist of textual information- and process user queries trying to allow the user to access to relevant information in an appropriate time interval. An IRS is basically constituted by three main components:

- (1) A documentary base, which stores the documents and the representation of their information contents. It is associated with the indexer module, which automatically generates a representation for each document by extracting the document contents. Textual document representation is typically based on index terms (that can be either single terms or sequences) which are the content identifiers of the documents.
- (2) A query subsystem, which allows the users to formulate their queries and presents the relevant documents retrieved by the system to them. To do so, it includes a query language that collects the rules to generate legitimate queries and procedures to select the relevant documents.
- (3) A matching or evaluation mechanism, which evaluates the degree to which the document representations satisfy the requirements expressed in the query, the so called retrieval status value, and retrieves those documents that are judged to be relevant to it.

The underlying retrieval model of most of the commercial IRSs is the Boolean one, which is a robust and well formulated model although presents some limitations. For example, it does not consider partial relevance and is not able to rank the retrieved documents by relevance. Due to this fact, some paradigms have been designed to extend this retrieval model and overcome these problems, with the vector space model being the most representative.

#### *3. Web Retrieval*

Although the textual IR techniques reviewed in the previous subsection are sometimes more than thirty years old, they still constitute the base of modern Web search engines. The popularity of the Web has transformed traditional IRSs into newer and more powerful search tools for locating content on the Internet.

However, there are several differences due to the special characteristics of the World Wide Web environment. As Zadeh enunciated in his foreword for F. Crestani and G. Pasi's edited book on “Soft Computing in Information Retrieval”, the problem of searching the Web has become far more complex that it was in the past mainly due to the increase on the size of the search space by several orders of magnitude and to the multimedia nature of Web documents, being composed of more information kinds than simple plain text. The main existing differences between Web retrieval and traditional IR, highlighting the following ones:

- (1) The HTML-based nature of Web documents, that make them present a structure defined by the HTML tags.

(2) The diversity of Web documents in terms of: i) length, structure, writing style and existence of grammatical and spelling errors; ii) language and domains; and iii) existing information formats, that Web applications have to appropriately deal with.

(3) The dynamic nature of many Web pages, that makes their retrieval difficult.

The previous aspects clearly show how Web retrieval have to extend traditional IR in order to deal with the special nature of Web documents. However, this usually makes Web engines focus more on the efficiency of the response than on the retrieval efficacy. Hence, as we shall see in the following section, SC can be a useful tool to build this gap obtaining textual IRSs and Web retrieval engines modelling better the retrieval activity.

#### 4. *Soft Computing in Information Retrieval*

So, what can actually do SC for IR?. Crestani and Pasi gave their view on the answer to this question in the preface of their previously mentioned edited book: “we think that a promising direction to improve IRSs’ effectiveness is to model the subjectivity and partiality intrinsic in the IR process, and to make IRSs adaptative, i.e., able to ‘learn the users concept of relevance’ ”. In a few words, they believe that SC can incorporate a greater flexibility to IRSs and, in view of the characteristics of this research area, it actually seems that this could be the case.

On the one hand, the modelling of the subjectiveness and uncertainty existing in the IR activity can be performed by the knowledge representation components of SC such as fuzzy logic, probabilistic reasoning, and rough sets. It is clear that uncertainty and imprecision are involved in the IR activity as, for example, the estimation of the relevance of a document to a user query or the own formulation of a query representing his information needs are pervaded with these characteristics. Concretely, fuzzy logic is a suitable tool to manage the retrieval activity as it is a formal tool designed to deal with imprecision and vagueness and as it facilitates the definition of a superstructure of the Boolean model, so that existing Boolean IRSs can be modified without completely redesigning them. Besides, probabilistic models are powerful and mathematically well formulated techniques to express and handle uncertainty since some decades ago.

On the other hand, the IRS adaptativeness mentioned by Crestani and Pasi is related to the machine learning perspective of SC, put into effect by evolutionary algorithms, neural networks and Bayesian networks, among others. These techniques and their hybridizations with IRSs based on the previous knowledge representation approaches can be applied to textual and Web retrieval tasks such as, for example, information extraction and Web mining, inductive query by example and relevance feedback, textual and Web document classification and clustering, and information filtering and recommendation systems.

#### **Contents:**

- 1.- Problems of the information retrieval and access in the web. (Enrique Herrera-Viedma)
- 2.- Techniques to solve the problems. (Enrique Herrera-Viedma)
- 3.- Classic models of Information Retrieval. (Enrique Herrera-Viedma)
- 4.- Soft Computing and Information Retrieval in the web: Fuzzy Model, Applications with Genetic Algorithms. (María J. Martín-Bautista)
- 5.- Fuzzy Logic tools and problems. (Sergio Guadarrama)
- 6.- Applications and Examples. (Alejandro Sobrino, José A. Olivas)

#### **About the speakers:**

**José A. Olivas.** Born in 1964 in Lugo (Spain), received his M.S. degree in Philosophy in 1990 (University of Santiago de Compostela), Master on Knowledge Engineering of the Department of Artificial Intelligence, Polytechnic University of Madrid in 1992, and his Ph.D. in Computer Science in 2000 (University of Castilla–La Mancha). In 2001 was Postdoc Visiting Scholar at Lotfi Zadeh’s BISC (Berkeley Initiative in Soft Computing), University of California-Berkeley, USA. His current main research interests are in the field of Soft Computing for Information Retrieval and Knowledge Engineering applications. He received the Environment Research Award 2002 from the Madrid Council (Spain) for his PhD. Thesis. **PRINCIPAL EMPLOYMENT AND AFFILIATIONS:** From 1997: Associate Professor of the Department of Computer Science, University of Castilla–La Mancha, Ciudad Real, Spain. From 1997: Professor of the Department of Computer Science, ICAI – Universidad Pontificia Comillas, Madrid, Spain. 1995-97: Head of the Department of Artificial Intelligence and Computer Science, University Antonio de Nebrija–UNNE, Madrid, Spain. From 1995: Collaborations with INSA (Aero Spatial Engineering and Services, NASA - Spain), Processing of forest fires data from satellites. 1992-1996: Head of the Computer Science Department of PPM studies center (Tres Cantos, MADRID): Consulting on Intelligent Systems to Enterprises such as SOUTHCO or ATT.

**María J. Martín-Bautista** is an Associate Professor of Computer Science at the University of Granada, Spain, where she received her Ph.D in Computer Science in 2000. Her current research interests include Data, Text and Web Mining, Intelligent Information Systems, and Information Retrieval with Fuzzy Logic and Genetic Algorithms, and she has served as a program committee member for several international conferences. She is a member of the European Society for Fuzzy Logic and Technology.

**Enrique Herrera-Viedma** was born in 1969. He received the M.S. degree in Computer Sciences in 1993 and the Ph.D. degree in Computer Sciences in 1996, both from the University of Granada, Spain. Currently, Dr. Herrera-Viedma is Senior Lecturer of Computer Science in the Dept. of Computer Science and Artificial Intelligence at the University of Granada. He has published more than 100 papers, 40 of them in international journals. He has coedited various journal special issues on Computing with Words and Preference Modeling and Soft Computing in Information Retrieval. His last special issue (with Gabriella Pasi) is titled "Soft Approaches to Information Retrieval and Information Access on the Web " and it will be published nextly in the Journal of American Society for Information Science and Technology (JASIST). His research interests include multicriteria decision making, decision support systems, aggregation of information, information retrieval, genetic algorithms, Web quality evaluation and recommendation systems.