Available at
www.ElsevierComputerScience.com
POWERED BY SCIENCE @ DIRECT®

ELSEVIER International Journal of Approximate Reasoning 34 (2003) 89–95

INTERNATIONAL JOURNAL OF APPROXIMATE REASONING

www.elsevier.com/locate/ijar

Editorial

Special issue on soft computing applications to intelligent information retrieval on the Internet

This special issue encompasses eleven papers devoted to the recent developments in the applications of soft computing (SC) techniques to information retrieval (IR), both in the text and Web retrieval areas. The seed of the current issue were some of the presentations made in two special sessions organized by the guest editors in two different conferences: the First Spanish Conference on Evolutionary and Bioinspired Algorithms (AEB'02), that was held in Mérida, Spain, February 2002, and the Seventh International ISKO Conference (ISKO'02), held in Granada, Spain, July 2002. The scope of both special sessions was pretty related. In the former conference, the session topic was "Applications of Evolutionary Computation to Information Retrieval" while in the latter the session was entitled "Artificial Intelligence Applications to Information Retrieval".

Five contributions were selected from these two special sessions in order to cover as much as possible the range of the different branches of SC in their application to IR. This way, these five works were instances of the four main SC constituent techniques: fuzzy logic, evolutionary computation, neural networks and probabilistic reasoning (Bayesian networks). The five original contributions were thoroughly revised and expanded to become the papers currently presented in this issue.

On the other hand, with the aim of giving a wider outline of the current state of the topic, we decided to incorporate a group of papers written by recognized people in the field to the issue. Again, we tried to follow the idea of including a contribution on each of the SC techniques that had been applied to IR. This way, several representative researchers on the area were invited by the guest editors to write a paper on their specific expertise subject, thus resulting in four different areas: fuzzy logic, fuzzy clustering, Bayesian networks, and rough sets. Besides, Baeza-Yates was also invited by us to write an introductory paper exploring the existing challenges on Web retrieval and suggesting how SC can help to achieve them.

0888-613X/$ - see front matter © 2003 Elsevier Inc. All rights reserved.
doi:10.1016/j.ijar.2003.07.001

First, we briefly introduce SC and textual and Web-based information retrieval. Then, we try to give a picture of the current state of the art of the applications of SC techniques to IR. Afterwards, we give an overview on the contents of the papers in the issue.

1. Soft computing, textual information retrieval and Web retrieval

1.1. Soft computing

The term SC refers to a family of computing techniques that, when Zadeh—the father of fuzzy logic—introduced the topic [7], 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. According to [8], 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 [2]. Hence, the SC framework is put into effect by hybrid systems combining two or more of the constituent technologies with complementary characteristics.

1.2. 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 [1,6]. 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 [6] being the most representative.

1.3. Web retrieval

Although the textual IR techniques reviewed in the previous section 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 Crestani and Pasi's edited book on "Soft Computing in Information Retrieval" [4], 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 are listed as follows [3]:

(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 it difficult to Retrieve them.

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 modeling better the retrieval activity.

2. 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 [4]: "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 modeling 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 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 [5] 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.

3. The papers in the special issue

As said in the beginning of this editorial, the papers in this issue are divided into two main groups. The first block, composed of six papers, corresponds to relevant researchers on the topic and outlines a general framework of the application of different SC branches to textual and Web-based IR. Then, the second group, composed of five contributions, includes the selected papers from the two special sessions, providing an additional insight of other proposals on the topic.

In the first paper, entitled "Information retrieval on the Web: beyond current search engines", Baeza-Yates analyzes the latest challenges of Web retrieval and provides some guidelines on how SC can be used to reach them.

In the second paper, entitled "Soft approaches to distributed information retrieval", Bordogna, Pasi and Yager tackle the problem of IR in a distributed environment, both from the situation where different documentary bases stored in several servers with different IRSs are available and from the case when several search engines look for information in the same large document collection. They propose some fuzzy logic-based approaches to merge the documents retrieved from the different IRSs in order to provide a single, ranked document list to the user.

The contribution "P-FCM: a proximity-based fuzzy clustering for user-centered Web applications" by Loia, Pedrycz and Senatore introduces an interesting application to Web retrieval based on a browsing process where the user provides a Web page and asks the search engine to locate other pages similar to it. To generate the document categorization needed for the system to put into effect the retrieval activity, they consider a novel fuzzy clustering algorithm guided by the user's feedback on the sensibility of the Web document cluster composition.

Kraft, Martín-Bautista, Chen and Sánchez present an exhaustive review on the use of classical logic rules and fuzzy rules in textual information analysis in the contribution "Rules and fuzzy rules in text: concept, extraction and usage". This way, they deal with topics such as text categorization, indexing, retrieval and classification, and analyze their relation with other disciplines such as data mining and machine learning when working with the rule knowledge representation structure.

In the paper "Bayesian belief networks for IR", Pinehiro de Cristo, Pereira Calado, de Lourdes da Silveira, Silva, Muntz and Ribeiro-Neto develop another review, in this case of the applications of Bayesian networks to the IR field, both for text and Web retrieval. The authors specially focus on the powerful aid that the Bayesian network can provide by incorporating new evidence, coming from different information sources such as past queries, thesauri and so on, to the IRS usual activity.

The last contribution in the first group is written by Miyamoto and is entitled "Proximity measures for terms based on fuzzy neighborhoods in document sets". It also deals with the document categorization problem, as Loia et al.'s work, but in this case the SC tools considered are fuzzy logic and rough sets applied on term-document co-occurrence matrices. In this case, fuzzy logic is considered to define new measures establishing the similarity between the documents, taking the previous information as a base.

The second group of papers starts with the contribution entitled "An application of the FIS-CRM model to the FISS metasearcher: using fuzzy synonymy and fuzzy generality for representing concepts in documents", coauthored by Olivas, Garcés and Romero. It presents the composition of a new Internet search engine that extends the usual ones based on the vector space model by considering fuzzy logic-techniques to expand the queries performed by the users considering synonyms and hierarchy relations of the terms involved in them. The engine also considers a soft clustering algorithm to build the hierarchy of related documents.

Herrera-Viedma, Cordón, Luque, López and Muñoz propose a new model for a linguistic IRS in the contribution "A model of fuzzy linguistic IRS based on multi-granular linguistic information". The main aim of this IRS is to facilitate the user-system interaction by allowing the use of different linguistic term sets (multi-granular linguistic information) to express both the user weighted queries and the relevance of retrieved documents.

The paper "A review on the application of evolutionary computation to information retrieval", by Cordón, Herrera-Viedma, López-Pujalte, Luque and Zarco, constitutes the third review in the issue. It provides a wide and complete view on the different areas of application of evolutionary algorithms to solve different IR problems, both in traditional text retrieval and in modern Web retrieval, as well as an exhaustive list of references on the topic. The different existing application areas are related to the machine learning activity in the most of the cases, such as in automatic document indexing, document and term clustering, query definition, user profiles, Web page classification or Internet search agents.

de Campos, Fernández-Luna and Huete introduce a retrieval model based on the use of Bayesian networks in the work entitled "The BNR model: foundations and performance of a Bayesian network-based retrieval model". The document collection is represented in the form of a Bayesian network, automatically learned by machine learning techniques, and new algorithms are proposed to estimate the probability distributions and infer on the network in order to solve the usual drawbacks of previous proposals.

The last contribution to the issue, "Comparison of neural models for document clustering", has been developed by Guerrero-Bote, López-Pujalte, De Moya-Anegón and Herrero-Solana, and is devoted to introduce an experimental study on the application of different neural networks and fuzzy neural networks to the document clustering task. The authors determine that Kohonen's neural network model was the best performing one in their experiment and analyze the interesting characteristics of the algorithm for the clustering activity.

Finally, as guest editors of this special issue, we should like to thank all the authors for their high quality contributions and the referees for their outstanding cooperation, and interesting comments and suggestions that helped to improve the final versions of the papers. Besides, we sincerely thank E. Alba, F. Fernández, J.A. Gómez, F. Herrera, I. Hidalgo, J. Lanchares, J.J. Merelo and J.M. Sánchez, General Chairpersons Commitee of AEB'02, M.J. López-Huertas, General Chairperson of ISKO'02, and P. Bonissone, Editor of the International Journal of Approximate Reasoning journal, for providing us with the opportunity to edit this issue.

References

[1] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, Addison-Wesley, 1999.

[2] P.P. Bonissone, Soft computing: the convergence of emerging reasoning technologies, Soft Computing 1 (1) (1997) 6–18.

[3] H. Chen, Introduction to the special issue on "Web retrieval and mining: a machine learning perspective", Journal of the American Society for Information Science and Technology 54 (7) (2003) 621–624.

[4] F. Crestani, G. Pasi (Eds.), Soft computing in information retrieval. Studies in Fuzziness and Soft Computing Series, vol. 50, Physica-Verlag, 2000.

[5] M. Nikravesh, V. Loia, B. Azvine, Fuzzy logic and the Internet (FLINT): Internet, World Wide Web and search engines, Soft Computing 6 (5) (2002) 287–299 (special issue on "Fuzzy Logic and the Internet").

[6] G. Salton, M.J. McGill, Introduction to Modern Information Retrieval, McGraw-Hill, 1989.

[7] L.A. Zadeh, Fuzzy logic, neural networks and soft computing, Communications of the ACM 37 (3) (1994) 77–84.

[8] L.A. Zadeh, What is soft computing? Soft Computing 1 (1) (1997) 1.

O. Cordón, E. Herrera-Viedma
Department of Computer Science and Artificial Intelligence, E.T.S. de Ingeniería
Informática, University of Granada, C/Daniel Saucedo Aranda,
s/n, 18071 Granada, Spain
E-mail addresses: ocordon@decsai.ugr.es (O. Cordón),
viedma@decsai.ugr.es (E. Herrera-Viedma).