Fuzzy approaches to access information on the Web: recent developments and research trends

E. Herrera-Viedma
School of Library Sciences
University of Granada, Spain
viedma@decsai.ugr.es

G. Pasi
ITC-CNR
Via Bassini 15 Milano Italy
gabriella.pasi@itc.cnr.it

Abstract

In this paper some research trends are presented concerning the application of fuzzy techniques to model flexible systems for the access to information on the WWW. The focus is on the definition of flexible systems, i.e. systems that can represent and manage the vagueness and uncertainty which is characteristic of the process of information searching and retrieval. We also identify some research directions which are worth to explore by the fuzzy community.

Keywords: Information Access, Recommender Systems, Information Retrieval, Semantic Web.

1 Introduction

The huge production of multimedia information on the World Wide Web stimulates the development of fast and effective systems that support an easy access to the information items relevant to specific users' needs. The activity aimed at locating on the WWW the information relevant to specific information needs is extremely hard: a user who accesses the network looking for something relevant to her/his needs can be seen as a traveller opening a door on a wild forest, which has to be traversed to the aim of reaching a (more or less) known destination. The users approaching the wild world of the Web have different ways to access the big amount of available and mostly unknown information, which are also related to the purposes of their search. The most immediate way is to directly navigate through the Web sites by means of a chain of links found in the pages; in this way a formal expression of information needs is not necessary. However, when some specific information is searched, this point and click access paradigm is unpractical, and the effectiveness of the results strongly depends on the starting page. The definition of systems that help users to automatically access information relevant to their needs is a very important research field. What the users expect from a system that provides an automatic support to information access is a list of the information items ordered according to their preferences. Some important research efforts are aimed at defining systems tolerant to imprecision and uncertainty in the elicitation of users' preferences and able to learn them through an interactive and adaptive behavior. We are interested in the fuzzy approaches to define flexible systems which support an automatic access to information on the Web. The most known systems belonging to this category are Information Retrieval Systems (IRSs) (on the Web, the search engines) and systems for the electronic commerce.

The aim of this paper is to synthetically present an overview of the main applications of Fuzzy Set Theory to the definition of flexible systems for locating and accessing information on the Web. In section 2 the main fuzzy applications to access information on the Web are reviewed. In section 3 some possible research directions which could benefit of the application of fuzzy techniques are suggested and commented.

2 Main applications of fuzzy techniques to Web search

In this section we briefly present the main applications of Fuzzy Set Theory to model systems which help users to access information on the Web. In particular, we analyze the following applications: fuzzy information retrieval models (indexing mechanisms and query languages, fuzzy document clustering, fuzzy data mining, fuzzy approaches to distributed IR, and fuzzy recommender systems.

2.1 Fuzzy information retrieval models

Most of the existing IRSs offer a very simple modeling of IR, which privileges the efficiency at the expenses of the effectiveness. A crucial aspect affecting the effectiveness of the system is related to the characteristics of the query language, which should represent in the more accurate and faithful way the user's information needs. The available query languages are based on keywords specification, and do not allow to express any uncertainty and vagueness in the expression of constraints that the relevant information items must satisfy.

Another important aspect which affects the effectiveness of IRSs is related to the way in which the information items are formally represented; the documents' representations are extremely simple, based on keywords extraction and weighting; moreover the IRSs generally produce a unique representation of documents for all users, not taking into account the each user looks at a document content in a personalized way, by emphasizing some subparts with respect to others. This adaptive view of the document is not modelled. Another important aspect is related to the fact that on the WWW some standard for the representation of semi-structured information are becoming more and more employed (such as XML); for this reason it is important to exploit their structure in order to represent the information they contain.

A promising direction to improve IRSs is to model the concept of partiality intrinsic in the IR process and to make the systems adaptive, i.e.able to "learn" the users' concept of relevance. In recent years big efforts have been devoted to the attempt to improve the performance of IR systems, and the research has explored many different directions to the aim of modelling the vagueness and uncertainty that invariably characterize the management of information.

To the aim of defining flexible IRS, Fuzzy Set Theory has been successfully employed for dealing with the imprecision and subjectivity that characterize the indexing process and for managing the user's vagueness in query formulation.

A survey of fuzzy extensions of IRSs and of fuzzy generalizations of the Boolean IR model can be found in [4]. Fuzzy generalizations of the Boolean model have been defined to the aim of defining IRSs able to produce discriminated answers in response to users' queries. In fact, Boolean IRSs apply an exact matching between a Boolean query and the representation of each document, defined as a set of index terms. They partition the archive of items into two sets: the relevant documents and the irrelevant ones. As a consequence of this crisp behavior, they are liable to reject relevant items as a result of too restrictive queries, and to retrieve useless material in reply to general queries. To the aim of softening the Boolean IR model, Fuzzy Set Theory has been applied at distinct levels. In documents' indexing some fuzzy techniques have been applied to the aim of providing more specific and personalized representations of documents' information content than those generated by the existing indexing procedures.

In [5] a fuzzy indexing model of documents structured in logical sections (such as XML documents) has been defined. The main idea is to explicitly model an indexing strategy that adapts the formal document representation to the user personalized view of documents' information contents. By the proposed model, the document representations can be tuned by users on the basis of their personal criteria for interpreting the content of documents.

Fuzzy Set Theory has also been successfully employed for defining flexible query languages, able to capture the vagueness of user needs as well as to simplify the user system interaction. This aim has been pursued at two levels: through the definition of soft selection criteria (soft constraints), which are more expressive selection criteria that allow

the specification of the distinct importance of the search terms by means of weights with different semantics [3, 10]. Another level of flexibility concerns the definition of soft aggregation operators specified by means of linguistic quantifiers such as at least k or most of [6].

As it happens with search engines, the incorporation of a weighted document representation in a Boolean IRS is a sufficient condition to improve the system with a document ranking ability. As a consequence of this extension the exact matching applied by a Boolean system is softened to a partial matching mechanism, evaluating the degree of satisfaction of the user's query for each document retrieved. This value is called the Retrieval Status Value (RSV), and is used to rank the retrieved documents [3].

In [18] a formulation has been proposed of the Boolean model and of some weighted Boolean models in terms of first order logic and fuzzy logic respectively.

2.2 Fuzzy associative mechanisms

Fuzzy associative mechanisms based on thesauri or clustering techniques have been defined in order to cope with the incompleteness characterizing either the representation of documents or the users' queries. In [16] a wide range of methods for generating fuzzy associative mechanisms is presented. Fuzzy thesauri and pseudo-thesauri can be used to expand the set of index terms of documents with new terms by taking into account their varying significance in representing the topics dealt with in the documents; the degree of significance of the associated terms depends on the strength of the associations with the documents' descriptors. An alternative use of fuzzy thesauri and pseudo-thesauri is to expand each of the search terms in the query with associated terms, by taking into account their distinct importance in representing the concepts of interest; the varying importance is dependent on the associations' strength with the search terms. Fuzzy clustering can be used to expand the set of the documents retrieved by a query with associated documents; their degrees of association with respect to the documents originally retrieved influence their RSVs.

2.3 Fuzzy data mining

As outlined in [15] data mining is an interdisciplinary field with a general goal of predicting outcomes and uncovering relationships in data. It is based on algorithms aimed at discovering hidden patterns and associations from large amounts of data stored in information repositories. In [15] the authors outline that "data mining tasks can be descriptive, i.e., discovering interesting patterns describing the data, and predictive, i.e., predicting the behavior of the model based on available data". Data mining can be either based on fitting models to or to determining patterns from observed data. A fitted model plays the role of inferred knowledge. A decisional activity in data mining is to establish whether the model reflects useful knowledge or not. For this kind of activity a subjective human judgment is usually required.

Recently various soft computing methodologies have been proposed to solve some aspects of the data mining problem. In [15] and in [17] a survey of the main applications of Fuzzy Set Theory to data mining and Web mining can be found.

2.4 Fuzzy approaches to distributed IR

In this section, the issue of searching for information distributed on a wide area network is considered. As outlined in [7], there are two main models describing the problem of retrieving such information: in the first model the information is considered as belonging to a unique, huge database which is centrally indexed for retrieval purposes. This is the model adopted by search engines on the WWW. A second model is based on the distribution of the information on distinct databases, independently indexed, and thus constituting distinct sources of information. This last model gives rise to the so called distributed or multi-source information retrieval problem. In this second case the databases reside on distinct servers each of which can be provided with its own search engine (IRS). A common problem to both models is list fusion. In the case in which we have a unique, huge and distributed information repository (like in the WWW), and distinct IRSs (search engines) the metasearch engines have been used to improve the effectiveness of the individual search engines.

The main aim of a metasearch engine is to submit the same query to distinct search engines and to fuse the individual resulting lists into an overall ranked list of documents that is presented to the user. In this case we typically have overlapping individual lists since a document may be retrieved by more than a single search engine. The fusion method thus has to be able to handle situations in which a document may appear in more than one list and in different positions within them. In the case of multi-source information retrieval the problem is to merge the lists resulting from the processing of the same query by (generally distinct) search engines on the distinct databases residing on distinct servers. However, in this second case we generally do not have overlapping lists as a result of the same query evaluation. Typically a document will be retrieved by just one single search engine, and thus the fusion problem is simplified with respect to the previous case.

In [19] some fuzzy methods for distributed IR have been proposed. These methods address the problem of deciding how many retrieved documents to select from each information source located on a different server.

In [7] some approaches to a soft fusion of the list of documents retrieved are proposed.

2.5 Fuzzy recommender systems

To filter the great amount of information available across the Web can improve the information access. Information filtering is a name used to describe a variety of processes involving the delivery of information to people who need it. Operating in textual domains, *filtering systems* or *recommender systems* evaluate and filter the great amount of information available on the Web (usually, stored in HTML or XML documents) to assist people in their search processes [22]. Traditionally, these systems have fallen into two main categories [21]. *Content-based filtering systems* filter and recommend the information by matching user query terms with the index terms used in the representation of documents, ignoring data from other users. These recommender systems tend to fail when little is known about user information needs, e.g. as happens when the query language is poor. In [24] was shown how to build content-based filtering systems (so-called reclusive recommender systems) based on fuzzy logic.

Collaborative filtering systems use explicit or implicit preferences from many users to filter and recommend documents to a given user, ignoring the representation of documents. These recommender systems tend to fail when little is known about a user, or when he/she has uncommon interests [21]. In [14] a collaborative filtering system based on fuzzy associative memory for automatically recommending high-quality information to users with similar interests is proposed. In [11] a fuzzy linguistic method to generate recommendations on the quality of Web sites in collaborative filtering systems is proposed. In [12] a fuzzy linguistic multi-agent model that incorporates content-based and collaborative filtering techniques to gather high-quality information on the Web is also defined. And, in [20] a fuzzy preference-based filtering technique is presented, which integrates content-based and collaborative filtering principles.

3 Some promising research directions

In this section we point out some new research paradigms appeared on the Web, and outline some possible applications of fuzzy techniques.

In last years the Web has witnessed an exponential growth of both documents and services. Today we can say that the Web is the largest available repository of data with the largest number of visitors searching information. Additionally, the Web is an infrastructure on which many different applications or services (such as e-commerce or search engines) are available. In this context, accessing relevant documents and developing useful services on the Web requires human intelligence. These Web challenges generate new research directions which could benefit of the application of fuzzy techniques. Some of the most important research directions are [1, 9, 23]: improving the query language of search engines, identifying Web content of good quality, and developing the Semantic Web. The first problems are related to the improvement of search processes of documents on the Web and the latter with the quantitative and qualitative improvement of services that the Web

provides or can provide. We briefly analyze these research directions in the following subsections.

3.1 Improving the query language of search engines

A fundamental problem on the Web is how to specify a search for information. Most Internet search engines propose query language based on keywords to express users' information needs. However, the user system communication requires a more complex interaction than the specification of keywords. For example, query languages that allow users to add the context of the information needed (as gender or time) or to distinguish the kind of information need (informational, navigational, or transactional) are necessary. The use of fuzzy linguistic modelling can be very useful to help users in the expression of their information needs. Another potentially useful development concerns the application of soft aggregation techniques to calculate the relevance with respect to all criteria expressed in a query. The problem is to make users able to use such query languages in a simple and satisfactory way. Then, the development of adequate interfaces that allow a fuzzy query formulation is necessary.

3.2 Identifying Web content of good quality

The Web is dense of noisy, low-quality and unreliable content. It would be extremely helpful for the Web search engines to be able to identify the quality of web documents independently of a given user request. Some proposals use link analysis for estimating the quality of Web documents. However, to assess the quality of a Web document requires additional sources of information. User judgments can help to evaluate the quality of accessed Web documents. The problem is that users typically do not make the effort to give explicit feedback. Web search engines can collect implicit user feedback using log files. However, this data is still incomplete. To achieve better issues of evaluation the direct participation is necessary. The use of fuzzy linguistic modelling to facilitate users in the expression of their judgements can be a good start to increase the participation of users in the evaluation models of the quality of Web documents. Additionally, to develop mechanisms to store such judgements in the structure of personal Web documents would facilitate the quality evaluation. This is possible by developing fuzzy linguistic information representations based on XML.

3.3 Developing the Semantic Web

A fundamental problem with the existing Web is that the data is machine-readable but not machine-understandable. The Semantic Web appears to create a new form of Web content meaningful to computers as well as humans [2]. The development of the Semantic Web implies two main research addresses which can benefit of the use of fuzzy tools:

1. Creation of new technologies to formalize the knowledge on the Web, i.e., to formalize the Semantic Web: They must allow that Web resources can be accessed by machines in a semantic fashion.

2. Creation of new applications: Like the Web, the Semantic Web is not an application; it is an infrastructure on which different applications (intelligent personal assistant, semantics-based Web search engines) may be developed.

On the one hand, to develop a Web with semantics, resources on the Web need to be represented with structured machine-understandable descriptions of their contents and relationships, using vocabularies and constructs that have been explicitly and formally defined with a domain ontology. An ontology is usually conceived as a hierarchical description of a set of concepts, a set of properties and their relationships, and a set of inference rules. The concept of ontology is central to the development of the Semantic Web. In this context, the Semantic Web is a web of distributed knowledge bases, and intelligent agents can read and reason about published knowledge with the guidance of the ontology [13]. In an ontology many aspects appear which require flexible knowledge representation, learning and reasoning: notions of approximate equality in data, semantic equivalence of syntactically different structures,

robustness against inconsistent or partial data, etc. Then, the fuzzy techniques can be used to avoid rigid definitions and to manage uncertainty in hierarchical representation of concepts and in inference or matching processes.

On the other hand, the Semantic Web is a collection of Web-applications described by ontologies. Semantic Web applications seem to fall into two types [8]:

1. Applications for organizations: For example, ontology-based marketplace development for business-to-business e-commerce.
2. Applications for users: For example, an intelligent travel assistant that gathers information, filters the relevant information, and composes a travel itinerary for a user.

In developing Semantic Web applications oriented to users one should consider that humans use vague terms in their interactions. Thus, the Semantic Web must provide definitions for linguistic terms used by humans with the aim of enabling machines to provide better solutions. In this context the fuzzy linguistic modelling can have an important role.

References

[1] R. Baeza (2003). Information retrieval in the Web: Beyond current search engines, Int. J. of Approximate Reasoning. To appear.

[2] T. Berners-Lee, J. Hendler, and O. Lassila (2001). The Semantic Web, Scientific American 284(5), 34-43.

[3] G. Bordogna and G. Pasi (1993). A fuzzy linguistic approach generalizing Boolean information retrieval: A model and its evaluation, J. of the American Society for Information Science 44, 70-82.

[4] G. Bordogna and G. Pasi (2001). Modelling vagueness in information retrieval. In Lectures in Information Retrieval, M. Agosti, F. Crestani and G. Pasi eds., Springer Verlag.

[5] G.Bordogna and G. Pasi (1995). Controlling retrieval trough a user-adaptive representation of documents. Int. J. of Approximate Reasoning, 12, pages 317-339.

[6] G. Bordogna and G. Pasi (1995). Linguistic aggregation operators of selection criteria in fuzzy information retrieval, Int. J. of Intelligent Systems, 10, 233-248.

[7] G.Bordogna, G. Pasi and R.R. Yager (2003). Soft approaches to distributed information retrieval. Int. J. of Approximate Reasoning, to appear.

[8] J. Euzenat (2002). Research challenges and perspectives of the Semantic Web, IEEE Intelligent Systems 17(5), 86-88.

[9] M.R. Henzinger, R. Motwani and C. Silverstein (2002). Chanllenges in Web search engines, SIGIR Forum 36(2).

[10] E. Herrera-Viedma (2001). Modeling the retrieval process of an information retrieval system using an ordinal fuzzy linguistic approach, J. of the American Society for Information Science and Technology 52(6), 460-475.

[11] E. Herrera-Viedma, E. Peis, M.D. Olvera, Y.H. Montero and J.C. Herrera (2003). Evaluating the informative quality of Web sites by fuzzy computing with words, Lectures Notes in Artificial Intelligence 2663, 62-72.

[12] E. Herrera-Viedma, F. Herrera, L. Martinez, J.C. Herrera, and A.G. Lopez (2003). Incorporating filtering techniques in a fuzzy Linguistic multi-agent model for information gathering on the Web, Fuzzy Sets and Systems, to appear.

[13] S. Lu, M. Dong and F. Fotouhi (2002). The Semantic Web: Opportunities and Challenges for Next-Generation Web Applications. Int. J. of Information Research 7(4).

[14] D.S. Lee, G.Y. Kim and H. Choi (2003). A Web-based collaborative filtering system, Pattern Recognition 36, 519-526.

[15] S. Mitra, S. K. Pal and P. Mitra (2002). Data mining in soft computing framework: A Survey (2001) IEEE Transactions on Neural Networks 13(1), 3-14.

[16] S. Miyamoto (1990). Fuzzy Sets in Information Retrieval and Cluster Analysis. Kluwer Academic Publishers.

[17] S.K. Pal, V. Talwar and P. Mitra (2002). Web mining in soft computing framework: Relevance, state of the art and future directions, IEEE Transactions on Neural Networks 13(5), 1163-1177.

[18] G. Pasi (1999). A logical formulation of the Boolean model and of weighted Boolean models. Workshop on Logical and Uncertainty Models for Information Systems (LUMIS 99), University College London, England, 5-6 July 1999.

[19] G. Pasi and R.R. Yager (2000). Document retrieval from multiple sources of information. In "Uncertainty in Intelligent and Information Systems", B. Bouchon-Meunier, R.R. Yager and L. Zadeh eds., World Scientific.

[20] P. Perny and J.D. Zucker (2001). Preference-based search and machine learning for collaborative filtering: The film conseil movie recommender System, Information - Interaction - Intelligence 1(1),9-48.

[21] A. Popescul, L.H. Ungar, D.M. Pennock and S. Lawrence (2001). Probabilistic models for unified collaborative and content-based recommendation in sparce-data environments. In Proc. of the Seventeenth Conference on Uncertainty in Artificial Intelligence (UAI), San Francisco, 437-444.

[22] P. Reisnick and H.R. Varian (1997). Special issue on recommender systems, Communication of the ACM 40(3).

[23] H. Tirri (2003). Search in Vain: Challenges for Internet search, IEEE Computer 36(1), 115-116.

[24] R.R. Yager (2003). Fuzzy logic methods in recommender systems, Fuzzy Sets and Systems 136(2), 133-149.