Fuzzy Web-based Information Retrieval Systems

DULCE MAGALHÃES DE SÁ
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa
Campus de Campolide, 1070-312 Lisboa
PORTUGAL
dulce@isegi.unl.pt

Abstract: The main problem on the Web is to locate information. Some reasons for this arise from the inefficiency of interfaces, incorrect information organization and data structure issues. A fuzzy approach to information retrieval systems can mitigate these problems. Advantages of fuzzy logic over sharp ones are the retrieval of information items which partial match the query and their ranking of importance, because in fuzzy logic an information element can reside in more than one set of different degrees of similarity. This provides a tool for natural language interfaces in retrieval systems and a solution for some Web-based information retrieval problems.

Key-Words: Fuzzy Logic, Information Organization, Information Retrieval Systems, Retrieval Techniques, Web-based Information Systems, Web Search

1 Introduction

The evolution of the Internet, particularly its Word Wide Web component, has created new opportunities and ways of application and analysis of information systems.

One of the components of information systems is information itself that, in certain ways, has become a social and economical product.

One of the ways of organizing information is by creating databases, developing applications and implementing database management systems.

Interactive databases in Web-based information systems are a critical component of the success of the system and, eventually, of the organization that owns it, because they constitute a decision support tool.

The inclusion of fuzzy logic models that generate filtered information, integrated in user–friendly interfaces, contribute for the competitiveness of the Web-based information retrieval system, ensuring the efficiency in getting results when queries to the database are made.

On the one hand, it guarantees a unique and tailored response to each search made by any user and, on the other hand, it makes possible the reduction of the necessary user knowledge to obtain information.

2 Information as Resource

The information, the way that information contents are treated by information systems and the way that it is presented as an answer to a query or request, are important factors to take a decision. Business processes depend upon the quality of information and of what it can be.

The information is a resource, not only for the information systems but also for the taking of decisions within different sorts of aspects of the economic life or other situations.

The Internet has brought a considerable development to the understanding of information as a resource [2].

Information can be understood as a resource at several levels, such as: the understanding of several phenomena and therefore as an input of information systems; the answer or output of those systems and by that way as a support to the taking of decisions; a knowledge base or an economic product.

The information can also have mixed characteristics of any possible kind of set of the referred levels. At the economical level it can be understood as a product, a good or a service and take different formats. It can be a Web page, a newspaper article, a set o bits on a CD or on a digital map, a database or many other things.

As a resource the information has costs and, frequently, its price is not easy to determine. At present it is burdensome to produce information but not to reproduce it, thanks to the technologies that allow us to do it each time with bigger easiness.

Therefore its price is more associated with its value than with its costs. That value is set up by who needs it and by the relevance that is given to it.

3 Retrieval Systems

Information retrieval systems are developed to provide efficient ways of search. Information technology gives some performance aspects to these systems. Information retrieval systems can operate in local mode or remote mode by a private network or Internet.

Salton and McGill, define an information retrieval system as a system used to store items of information that need to be processed, searched, retrieved and disseminated to various user population [9].

3.1 Components

The basic components of retrieval information systems are search formulation, search software, information storage environment and queries process [5]. This involves hardware for information storage and persons (searchers) that executes queries.

3.1.1 Searcher

Searcher is the person who has information or information contents needs and then begins the search process. To do this, the searcher executes a direct or indirect query to the system. A direct query occurs when the system provides access in a particular topic by user choice, for example such as word-key.

Indirect query is the option of navigation to a particular information topic on a variety of possibilities. Like a thematic index, with sub-thematic indexes. This provides the user (searcher) navigation control in a sub-world of information.

In a particular case, the searcher of World Wide Web can be a mixed entity between any person who elaborates part of the initial search process and a software agent who limits the scope of the search. The human search is mostly oriented by subjective processes.

The software agent searcher transmits to system the subjective aspects of the human user through rules previous fixed. Software agents can interfere in other phases of search, such as structures of search formulation [8].

3.1.2 Search Formulation

Search formulation is a complex process that requires some decisions. These decisions concern to topics of search, information fonts, information contents, design of search formulation and which resources should be used in search process.

3.1.3 Resources

Resources considered in a retrieval information system can be software to search in local mode or private networks or Internet [5]. On the World Wide Web can be used some mechanisms with interactive databases such as a search engine like Yahoo.

A search engine is a tool for specific localization of sites or information on the Word Wide Web [6]. Search engines provide search by hierarchic way of contents organization, direct way or mixed ways between hierarchical and direct modes.

3.1.4 Storage

Information storage is an important aspect of retrieval information systems because it allows search across the data structures. This is possible with resource to file systems or database management systems. Storage can be distributed (geographical or physical) or localized in a storage information support like a computer disk.

3.1.5 Retrieval Items

Answers to information queries or retrieval items depend on all elements referred before. It also depends of the design of retrieval information system. The way that the system answers to queries depends of its design and conception.

Particular types of agents are those that support search engines like Yahoo or Google. They work like search engines of search engines and can be activated by a query search formulation.

For this query, agents filter one of the possible answers between thousands of answers provided by some search engines.

4 Retrieval Techniques

Retrieval techniques are the methods or processes used by systems to extract information topics in a particular way.

A classification of retrieval techniques has been proposed by [1]. On the highest level one can distinguish between exact or partial match techniques.

Partial match can be divided in individual or network techniques. The first one searches single items nodes without considering the data collection as a whole. While the second one considers the set of all data items and their relationships are used to find the most relevant data with respect to a query.

Individual techniques can be separated on structure-based (logic and graph) and feature-based (formal and ad-hoc techniques).

Formal can be divided in probabilistic or vector-space techniques. And finally, network can be divided in cluster, browsing and spreading activation.

5 Web Search

The main problem on the Web is locating the information. Searching is a critical activity because they are millions of sites. There are half million of new Web-pages for week [6].

Almost other problems on the Web search are a consequence of this problem. For example, some problems are information organization, efficiency of links, performance of navigation interfaces, legal items of authoring and reproduction of information.

To Lennon, search of information on the World Wide Web has the following problems [7]:

- Break links, generated by inadequate update of addresses, by error on link creation and its pages or sites has removed.
- Lost orientation on navigation, due to inefficiently design of interfaces or site structure.
- Inappropriate interfaces for the sites and/or information systems context.
- Difficulty to obtain answers to specific queries or navigation trace route.
- Access difficulties to found information contents or valid references to information needed.

Some reasons for these problems are data structures issues, efficiency of interfaces, information organization, database management systems absence, quality of information and information systems not suitable for specific business process use.

6 Fuzzy Approach

Fuzzy logic is basically a multivalued logic that allows intermediate values to be defined between usual valuations like 1/0, yes/no or true/false.It has been used initially in system theory to describe and implement uncertain notions and general concepts [4].

In fuzzy logic an information element can reside in more than one set of different degrees of similarity. Fuzzy relations represent a degree of presence or absence of association, interaction or interconnectedness between the information elements of two or more fuzzy sets.

One of the components of a fuzzy logic system is rules. These rules will be expressed as logic restrictions, in the forms of IF-THEN statements [3]. They are usually of a form similar to the following: *if x is low and y is high then z=medium.*

The area of information systems and information retrieval and database management has also benefited from fuzzy logic methodology.

Some fuzzy operations like union, intersection or complement can filter information elements with values between absolute true (1) and absolute false (0).

This provides a tool for natural language interfaces in retrieval systems and solution for some Web-based information retrieval problems.

The software agent searcher that transmits to system the subjective aspects of the human user through rules previous fixed can be based in a fuzzy logic structure.

Fuzzy approaches to retrieval information systems can consist in establish rules linking information elements, for example between words and themes by way user interaction or first word and second word within an expression.

There already exist query systems that provide an ordering among the information items that more or less satisfy the request. The systems may allow for the presence or imprecise, uncertain or vague information in the system database.

7 Conclusion

Information can be understood as a resource at several levels, such as: the understanding of several phenomena and therefore as an input of information systems; the answer or output of those systems and by that way as a support to the taking of decisions; a knowledge base or an economic product.

Some reasons for information retrieval problems are data structures issues, efficiency of interfaces, information organization, database management systems absence, quality of information and information systems not suitable for specific business process use.

The advantages of fuzzy logic over sharp ones are the retrieval of information items which partial match the query and their ranking of importance, because in fuzzy logic an information element can reside in more than one set of different degrees of similarity.

Retrieval techniques are the methods or processes used by systems to extract information topics in a particular way.

The advantages of fuzzy logic queries over sharp ones are the retrieval of information items which partial match the query and their ranking of importance to user needs articulate to the search formulation.

Fuzzy logic can be used to detect break links, identify the orientation of user navigation, discover partial needs of user queries and access information elements with some efficiency by way to reduce the number of problems in information retrieval.

References:

[1] Belkin, N. and Croft, W., Retrieval Techniques, *Annual Review of Information Sciences and Techniques*, Vol. 22, 1987, pp. 109-145.

[2] Burke, M. and Hall, H., *Navigating Business Information Sources, a practical guide for information managers*, Library Association Publishing, 1998

[3] Cloete, I. And Zurada, J., *Knowledge-Based Neurocomputing*, MIT Press, 2000

[4] Ignazio, J., *Introduction to Expert Systems: The Development and Implementation of Rule-Based Expert Systems*, McGraw-Hill, 1991

[5] Large, A. et al., *Information Seeking in the Online Age: Principles and Practice*, Bowker-Saur, 1999

[6] Laudon, K. and Laudon, J., *Essentials of Management Information Systems*. Prentice Hall, 1999

[7] Lennon, J., Aspects of Large World Wide Web Systems, *Proceedings of WebNet International Conference*, 1996

[8] Pitkow, J., In Search of Reliable Usage Data on the WWW, *Proceedings of the 6th International World Wide Web Conference*, 1997

[9] Salton, G. and McGILL, M., *Introduction to Modern Information Retrieval*, McGraw-Hill, 1983