Fuzzy Ontology and Information Access on the Web

Stefania GALLOVA

Abstract—**Web is the largest available repository of data. In this contribution a solved application of Fuzzy set theory technique to the definition of flexible systems for locating and accessing information on the Web is presented. A purpose of our research is also a fact, that there are various ways to access the big amount of available and mostly unknown information for users. Clustering methods are also appropriate for helping on the process of document retrieval. We introduce a fuzzy clustering methodology for solved system. Interface provides the users with tools for navigating through hierarchies of documents and visualise selected documents with using of fuzzy clustering for indexing in the HSC interface.**

***Index Terms*—Information Retrieval, Clustering, Semantic Web, Fuzzy ontology**

I. INTRODUCTION

The definition of systems that helps users to automatically access information important to their own needs is a very relevant domain of research. An important research works are aimed at defining tolerant to imprecision, uncertainty and vagueness in the elicitation of users' preferences and able to learn them through an interactive and adaptive behaviour. We are interested in the fuzzy approach to define solved flexible system which support an automatic access to information on the *Web*. There are some known relevant systems such as *Information Retrieval Systems* on the Web with the search engines and systems for the electronic commerce.

There is also a relevant aspect related to the way in which the information items are formally represented. Generally, documents produce a unique documents representation for users. On the „*www*" some standard for the representation of semi-structured information are becoming more and more employed, such as *XML*. This fact leads to exploit their structure in order to represent information they contain.

Manuscript received March 22, 2007. This work was supported by the Slovak Republic's Ministry of Education under Scientific Grant Projects:

1. *Diagnostic System Building of Machinery Equipment with Knowledge-based System and Chaos Theory Applying* (Number: VEGA 1/1084/04).
2. *Intelligent Approach to Automated Diagnostic Problem Solving of Machinery Equipment* (Number: VEGA 1/4133/07).

Stefania GALLOVA. Author is with the Technical University of Kosice, Letna 9, SK-042 00 Kosice, Slovak Republic; phone: +421-904-584-426; e-mail: stefania.gallova@zoznam.sk

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. There are many methods of soft computing, which have been proposed to solve some aspects of the problems about data mining tasks.

We solve two approaches to model, which describe the problems of information retrieving. In model with search engines on the *www* the information is considered as belonging to a unique and huge database. This database is centrally indexed for retrieval purposes. A type of model, which is based on the distribution of information on distinct databases, independently indexed, gives rise to the *distributed or multi-source* information retrieval problem. In this case the solved model constitutes distinct sources of information and the databases reside on distinct servers each of which can be provided with its own search engine, i.e. information retrieval system. Fig.1 illustrates a solved fuzzy ontology approach in information retrieval process.

A common solved problem mentioned above models is list fusion. In the case in which we have a unique, huge and distributed information repository, like in the *www*, and distinct information retrieval systems with search engines, the metasearch engines have been used to improve the effectiveness of the individual search engines.

The main aim of 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 users.

The fusion methodology has to be able to handle situations in which a document may appear in more than one list and in various positions within them.

In the case of multi-source information retrieval the task 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.

II. PROBLEM SOLVING

There is an information filtering, i.e. a variety of processes involving delivery of information to users who need it.

The great amount of information available across the *Web* can improve the information access [1], [2]. Operating in textual domains, filtering systems or recommenders systems evaluate and filter the great amount of information available on the *Web* usually stored in *HTML* or *XML* to users in their

search processes. There are two used principles: „Content-based filtering system“ and „Collaborative filtering system“.

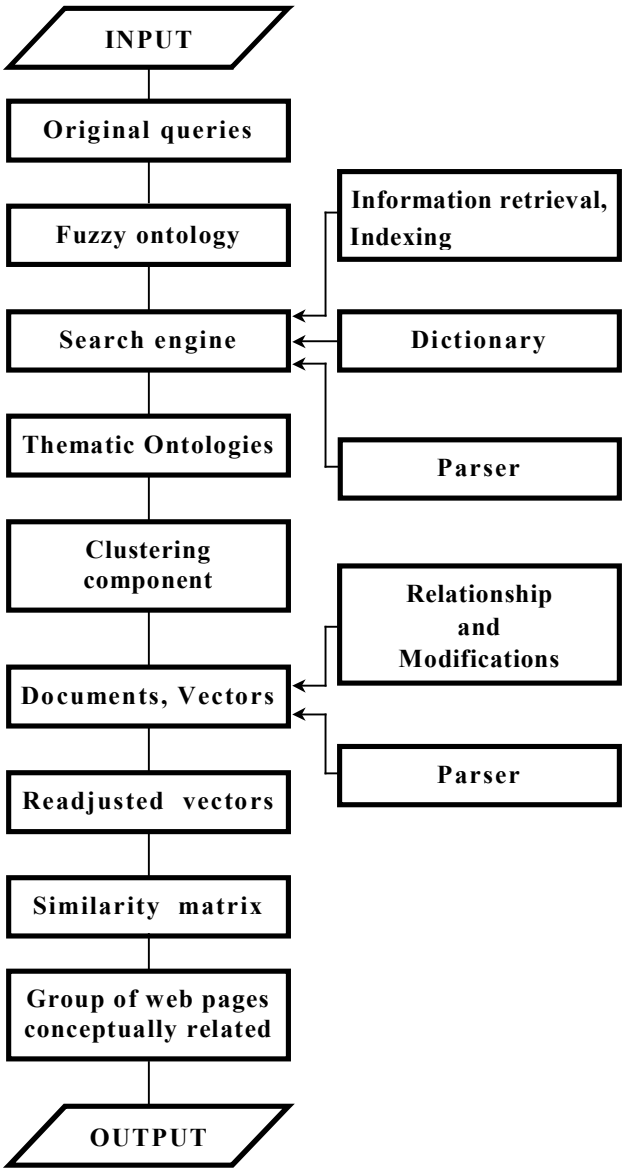

Fig.1 Fuzzy ontology in information retrieval process.

Content-based filtering system filters and recommends information by matching user query terms with the index terms used in the representation of documents, ignoring data from other users. This system tends to fail when little is known about user information needs (when the query language is poor and so on).

Collaborative filtering system use explicit or implicit preference from many users to filter and recommend documents to a given user, ignoring the representation of documents. This system tends to fail when little is known about the user, or when the user has uncommon interests.

There is fuzzy clustering, which is applied to the set of domain documents. Then, we use a heuristic approach methodology for selecting and appropriate number of clusters. Then we apply fuzzy clustering for selected cluster data. Elements with a smaller membership should have less influence on the final partition than elements, which have larger membership. Visualisation of the clusters is not so easy. There is considered application of Shannon's map to the whole data set and initial position of data selected at random. The fuzzy clustering algorithm does not only compute the fuzzy partition but also permits to compute membership functions values for any arbitrary data point. There is applied also entropy based fuzzy means, which permits this.

In experiments with usable membership function μ – one of the usable forms is as follows:

$$\mu_{act} = \mu_{prev} \pm \frac{\sqrt{(\mu_i - \mu_{prev})^2}}{Q_{hist}} \tag{1}$$

where μ_i belongs to particular document. The previous value of membership function μ_{prev} determines the actual value of membership function μ_{act} and μ_i provide a membership value of each document D_i (i=1, , n). Q_{hist} is the number of queries that have confirmed the intended meaning of the term Q, which is used as a denominator in order to reduce the effect of later learning in order to stabilise the values.

There is a fundamental question on the Web – „how to specify a search for information?“. Classical Internet search is based on specification of „keywords“, without more complex interactions. The use of fuzzy linguistic modelling is seemed to be very useful to help users in the expression of their information needs.

Fuzzy labels can be attached to the recovered documents according to a membership function with understanding of the meaning and acceptance of vagueness. The potential advantages of fuzzy ontologies include not only improved satisfaction with query results, but also the potential to represent the domain knowledge. This approach should remove artefacts caused by sharp transitions between ontologies. The set of elements in Fig. 2 with the assignment function μ_{E0} is dissolved over the elements of the bottom sets. The composed objects function is the set with the assignment function illustrated in Fig. 3. The assignment coefficients of the singular objects (E_1, E_2, E_3, ...) to the given concrete element (E5) is equal to the discrete points of $\mu_{E0,5}$. Fig. 2,3 illustrate the capability of the organisational memory to determine the overall assignment or membership of a single (composed) object to the overall knowledge stored in the organisational memory, the capability to identify the degree of membership of a single (composed) object (i.e. word) to the overall knowledge, i.e. local similarity/global similarity or local membership/global membership.

The new query is obtained as follows [5]:

$$Q_{new} = c_1 Q + \frac{c_2}{n_r} \sum D_r - \frac{c_3}{n_{nr}} \sum D_{nr} \tag{2}$$

where:
- Q is the initial query,
- Q_{new} is the new query,
- n_r is the number of relevant documents,
- $n_{n,r}$ is the number of nonrelevant documents,
- D_r is the relevant document representation,

D_{nr} is the nonrelevant document representation,
c_1, c_2, c_3 are the constants.

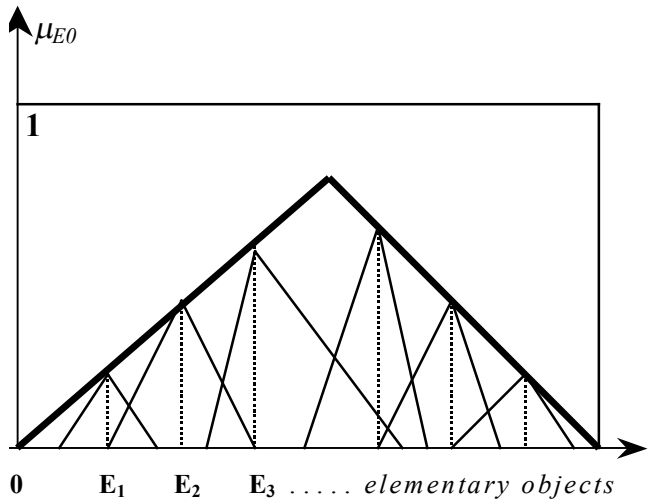

Fig. 2 Clustering of concept sets

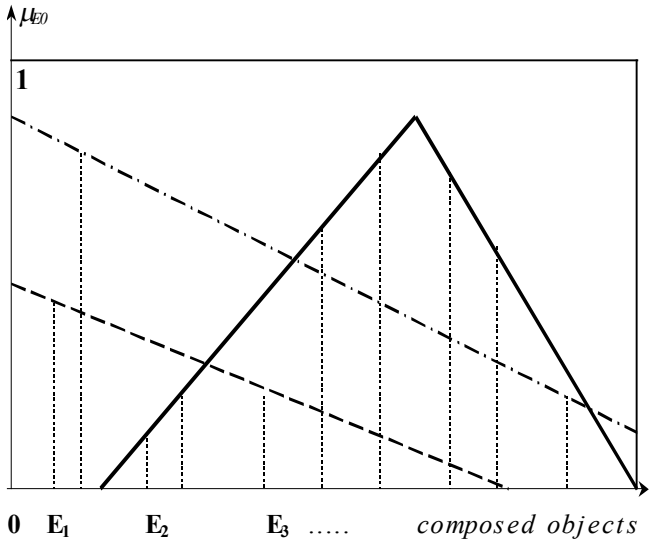

Fig. 3 Clustering of concept sets

We realise an algorithm for query expansion based on the partial judgements of retrieval documents. There are introduced one classical approach and two fuzzy approach methodologies. Fuzzy principles are applied with constant four relevant levels and then with changeable amount of judgements from two to nine relevant levels in accordance with their real needs. The scale values are from *„very relevant state“* to *„non relevant state“*. The degree of a membership of judged document to a relevance level is classically on the binary base.

The first phase is to formalise the document judgements. The second phase is to transform them onto the terms occurring in these documents. The objective is to build new query with the best couples (term, weight). The new query is built by considering the terms extracted from the judged document and the judgements of these documents.

We use a modified term weight expression such as „Mercure Information Retrieval System approach“ in this form:

$$w_{ij} = \frac{k_1 * ft_{ij} * \log \frac{N}{n_i}}{k_2 * ft_{ij} + k_3 \frac{h_j}{\Delta d}} = \mu_d(t) \qquad (3)$$

where: w_{ij} is link weight between term t_i and document j,
ft_{ij} is the frequency of occurrence of the term i in document j,
n_i is the number of document containing the term t_i,
h_j is the width of document d_j,
N is the number of documents of the collection,
k_1, k_2, k_3 are the constant parameters,
$\mu_d(t)$ is the document representation as a membership function.

A term occurs in the document d at a certain degree of membership. The final weight of this term is mostly aggregated by *Maximum/Minimum* pair operators. There are judged documents produced. Mostly one feedback iteration is realised. For each intended relevance feedback methodology and for each query we submit the query to the Information retrieval system.

Fuzzy feedback approach is compared to the classical binary approach. The top twenty documents are judged according to the considered method and also the new query is built according to the used methodology approach. This query is submitted to the *Information retrieval system* which returns a ranked list of 800 documents. At the end of experiment three lists of 800 documents are selected for each query. One list is the result of classical binary feedback, and the others two lists are the results of two fuzzy feedbacks approaches. Fuzzy feedback search with changeable judgements from two to nine levels was realised also by special modified chaos theory principles combined with entropy measure. A classical measure precision developed in *Information Retrieval domain* and used in *TREC* programme evaluates created lists. Realised experiment involves nine precision measures. „P_n“.with $n=10,...,800$ is the precision at n documents. It is measured by following way:

$$P_n = \frac{\text{"number of relevant documents"}}{\text{"number of retrieval documents which are kept"}} \qquad (4)$$

System is functioning this way that the user can decrease the number of documents retrieved by the first query or can query the web again with a different vocabulary to retrieve a new set of documents. Each one of the new queries has a fuzzy compatibility degree with the original query that is determined from the synonymy degree between the words included in it and the words specified in the original query. It

must be mentioned that some results are counterintuitive when we compare the values of confidence and certainty, which discloses that where the rules relate two items very frequent, the confidence is quite high by the certainty is not.

Table I. Values for precission curves

Approach	P10	P20	P30	P50	P80
classical	0,515	0,45	0,29	0,26	0,22
constant fuzzy	0,58	0,51	0,33	0,28	0,25
changeable fuzzy	0,59	0,57	0,405	0,37	0,35

Approach	P150	P300	P500	P800
classical	0,19	0,16	0,14	0,11
constant fuzzy	0,21	0,19	0,17	0,15
changeable fuzzy	0,318	0,26	0,24	0,23

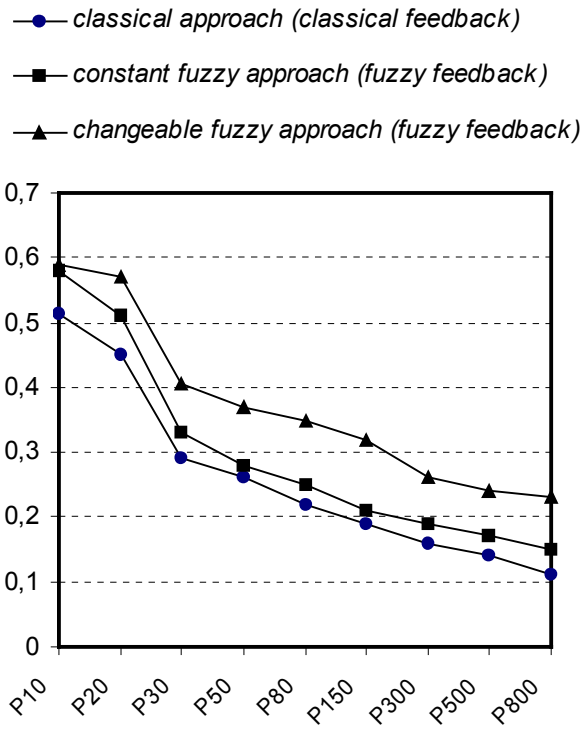

Fig. 4 Courses of precision curves for 3 solved methodology approaches

Fig. 4 illustrates the precision curves of realised experiments at different points P_n (see table I.). The solved methods improve the classical methodology approach and the best documents are in the top of the retrieved-document list. Results and experiments undertaken on TREC collection have relevant conclusions and show the effectiveness of the solved approach.

Fuzzy logic elements give an adequate basis to handle these partial judgements. When users have more levels, they can easily make choices to assign documents to important levels which best match their idea and perceptions of the documents.

A flexible and effective information retrieval is the goal of our effort. The main goal of information retrieval is to retrieve relevant documents in response to a user need [5].

DEVELOPING THE SEMANTIC WEB

The Web is dense of noisy, low-quality and unreliable content [1], [2]. It is known that if we have obtained information about Web document quality, we need to realise additional sources of specific information. 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. It is also useful 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*. A communication modul is illustrated in Fig. 5.

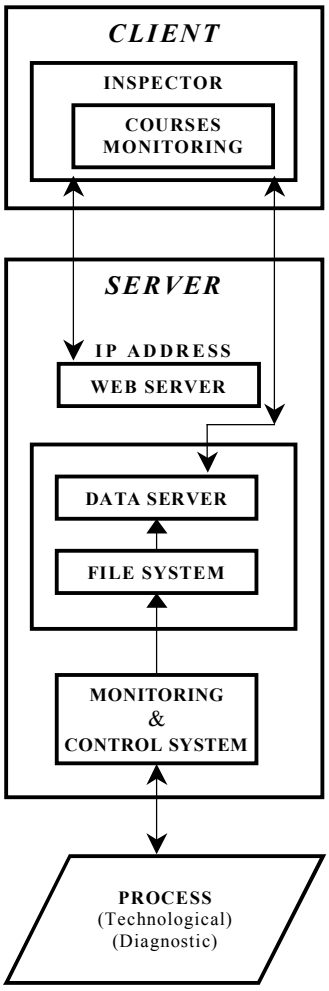

Fig. 5 A communication modul

Existing Web has fundamental problem. This problem 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, and its development involve a creation of new technologies to formalise the knowledge on the Web, and creation of new applications like the Web.

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 public knowledge with the guidance of the ontology [3]. 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, and so on. Then the fuzzy techniques can be used to avoid rigid definitions and to manage uncertainty in hierarchical representations of concepts and in inference or matching processes. The Semantic Web is also a collection of Web applications described by ontologies.

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. [4].

III. Conclusion

The flexible nature of the fuzzy ontology may support a wide range of approaches to the problems of retrieving relevant, accurate, appropriate and most of all useful information which is a relevant key aspiration of research of semantic web.

Future work includes working with larger data sets. In this case, a hierarchical clustering methodology seems to be appropriate tools. Tools for document clustering and structuring permit to have a better understanding of the documents available.

The interfaces provide the user with tools for navigating through hierarchies of documents and visualise selected documents and similar ones. Similarity is based on Wordnet 1.7 and Latent Semantics Analysis.

Further research is directed to task of improving the query language of search engines, identifying Web content of demanded quality and also developing the Semantic Web problems.

The research reported on this paper is supported by following scientific projects of Slovak Grant Agency VEGA (Ministry of Education):

1. *Diagnostic System Building of Machinery Equipment with Knowledge-based System and Chaos Theory Applying*. (Number: VEGA 1/1084/04)
2. *Intelligent Approach to Automated Diagnostic Problem Solving of Machinery Equipment.* (Number: VEGA 1/4133/07).

REFERENCES

[1] E. Herrera-Viedma, *Modelling the retrieval process of an information retrieval system using an ordinal linguistic approach.* American Society for Information Science and Technology 6/2001, pp.460-475

[2] G. Pasi, *A logical formulation of the Boolean model and weighted Boolean models.* Lumis'99, University College London, England, 1990

[3] M. Lu, F. Dong, F. Fotouhi, *The Semantic Web. Opportunities and Challenges for next Generation Web Applications.* Information Research (7), 2002

[4] U. Kruschwitz, *An Adaptable Search System for Collections of Partially Structured Documents.* „Intelligent Systems, IEEE vol.18, pp.44-52, 2003.

[5] I. J. Rochio, *Relevance feedback of Information Retrieval.* The Smart System Experiments in Automatic Document of Processing. Prentice-Hall, pp.313-323, 1971.

[6] A. Brini, *Introduction de la Gradulaite dans le Jugement Utilisateur.* Dea Report. Toulouse, France, 2002