INFORMATION RETRIEVAL SYSTEM USING FUZZY SET THEORY - THE BASIC CONCEPT

BHASKAR KARN
Assistant Professor
Department of MIS
Birla Institute of Technology
Mesra, Ranchi

ABSTRACT

The paper presents the basic idea involved in information retrieval using fuzzy set theory and fuzzy logic. It attempts to clarify the definition and some terms and discuss the possible sources of fuzziness within this sub-field. It also tries to prove that how this way retrieving information is more realistic and expressive than their crisp counterparts.

Key words:- Vagueness, imprecision, uncertainty, fuzzy, information retrieval

1. INTRODUCTION

The rapid production and circulation of information, makes the design of systems which automatically manage and retrieve the "right" information an important issue.

The primary purpose of establishing an information retrieval system lies in assisting the users to efficiently acquire desired information. Most commercial information retrieval systems currently still adopt the Boolean logic model. However, the information retrieval systems based on Boolean logic model are rather restricted in applications since these systems are unable to represent uncertain information. If there is uncertain information, the query processing of these systems are not handled properly, [1].

Fuzzy information retrieval methods based on fuzzy set theory has been proposed improving the disadvantage of Boolean logic model which can not handle uncertain information.

2. What Is Information Retrieval System [2]

Information retrieval system consists of two parts: the textual archive, which is a set of textual units (often called document), and a retrieval engine. A user of a retrieval system presents queries describing what kinds of documents are desired. The retrieval engine matches the queries against the documents in the textual archive. It then and returns the user a list of sub-collection of the documents which are deemed " best matches".

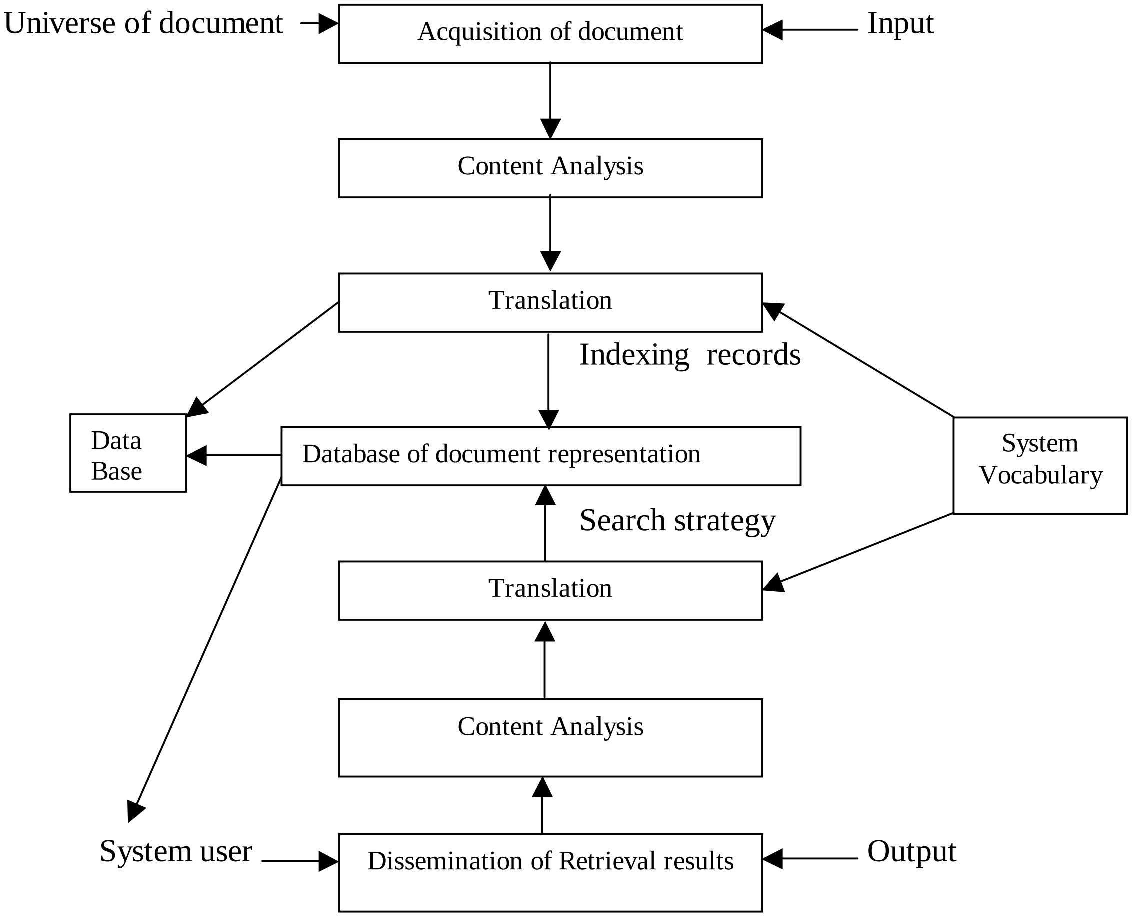

Fig. Functions of an IRS. [3]

Its aim is to evaluate user queries for information based on the content analysis of the document stored in an archive at minimal cost and documents that are of “maximum relevance” to the user.

Thus we can define information retrieval, in general, as the problem of the selection of documentary information from storage in response to search questions, i.e. to match the words or other symbols of the inquiry with those characterizing the individual document and make “appropriate selection”.

3. TERMINOLOGIES

3.1 Uncertainty:

Uncertainty refers to the truth of the information stated in a proposition, thus affecting the credibility of the information content (or value). Uncertain information is generally expressed by the terms such as likely, probable, possible, plausible, or credible. As a consequence of uncertainty, one is not able to establish the truth of the information. It is important to distinguish between two main types of uncertainty, one related to randomness, the other one to fuzziness. The classical approach to model uncertainty related to randomness is probability theory. A well established formal frame work for dealing with uncertainty derived by fuzziness is possibility theory. Within possibility theory a fuzzy set has a disjunctive interpretation. This means that just one of the values of the reference domain of a variable is the actual value of the variable. The fuzzy set membership is then interpreted as a possibility distribution: in this case the degree of membership of each domain value expresses the degree of possibility that this value in the real value.

3.2 Imprecision:

Imprecision is related to the content of the information stated in a propositional form with respect to the granularity of a variable. Precise information occurs when the value of the considered variable is atomic with respect to the reference domain of its possible values. Conversely, imprecise information is characterized by a value that is a subset of the reference domain. For example, John is between 20 and 30 years old, given the reference domain of the implicit variable age being equal to {0,...., 100}.

Maximum imprecision occurs when the value cannot be hypothesized (the whole reference domain must be considered). Imprecise information identifies a clear cut partition of a set of elements: those which can be the actual value and those which cannot.

3.3 Vagueness:

Vagueness occurs when the value of a variable in a proposition has a higher granularity than the reference domain. For example, the statement. John is young contain vagueness when the reference domain is the set {0,....,100} while it is precise when the reference domain is the set of labels { very young, young, middle aged, old, very old}. A vague proposition identifies an elastic constraint on a set elements. The elastic constraint is defined by a fuzzy set on the domain of a variable; it does not define a clear cut partition , but a “fuzzy” subset whose element satisfy the elastic constraint to a different degree. For example, in the previous statement young is an elastic constraint defined on the domain {0,150} of the implicit variable Age. Imprecision can be regarded as a particular case of vagueness, for any imprecise if information can be represented by a fuzzy subset in which all the elements in the support have the maximum membership value 1.

Both imprecise/vague and uncertain information can coexist and balance each other. Often, to cope with in complete knowledge, vagueness is introduced in a proportion to describe a system state so as to be more confident on its truth.

3.4 Inconsistency :

Inconsistency is derived by the coexistence of contradictory information about the value of a variable and may depend upon the existence of different sources of information.

4. WHERE IS THE VAGUENESS AND UNCERTAINTY IN IRS

4.1 Document indexing:

The indexing process is a procedure aimed at assigning a set of descriptive words (index terms) or phrases to synthesize the topic (s) a document is “about”. This activity can either be carried out manually by an expert in the field or automatically performed.

One can often employ a controlled vocabulary, generated by subject experts and/or librarians, which specify given words and phrases. Often, there are also rules for the assignment of terms, along with term relationships . Term relationships include synonyms, related terms, narrower terms, and broader terms. The downside of this is that users may not use the specific terms as given by the experts; moreover, the set of terms can often become outdated.

An alternative of the free vocabulary approach, which involves the use of keywords, terms that are provided by authors, by professional indexes, or are found in the document. Even in the case where a thesaurus is a available, sometimes authors and/or indexes will augment the document representation with terms not found in the controlled vocabulary.

4.2 Queries:

When expressing needs for information, users often formulate natural language queries, at least at first, in describing those needs. If a system is expected to respond, it will need to parse those natural language statements. For example, in an archive of recipes a natural language query could be:

“Retrieve the most relevant recipes dealing with at least one among (rice (most desired), pasta (desired), connelloni (appreciated) and possibly with (at least a few among (vegetable (a lot), spices (a few), wine (moderate), cheese (a little), not meat) and point out the main ingredients and cooking time.

This introduces a sense of ambiguity in that natural language is often vague.

Vagueness can be present at different levels in a query. In order to specify different levels of importance of the search terms, important labels can be associated with the search terms; these labels can have different semantics. In the example above, the labels `most desired’, `desired’, `appreciated’, specify the importance of the search terms `rice’, `pasta’, `cannelloni; giving the differing significance to the user of the recipes about the specified foods. The labels ` a lot’, ` a few’, ` moderate’, ` a little’ specify the desired degree of significance (amount) of ` vegetable’, ` spices ‘, ` wine’, `cheese’ in the recipes. Thus these labels have the semantics of ideal significance values.

Vagueness can also appear in specifying the aggregation criteria of the search terms. In the example above `at least a few’ expresses the need for a soft aggregation criterion, while `and possibly’ is used to specify primary selection criteria and optional topics of interest that may influence positively the relevance of selected recipes.

4.3 Relevance :

Another ill-defined concept in IR, is the concept of relevance: only the user, or perhaps experts on the topic(s) acting in the user’s stead, can determine the true relevance , i.e. the usefulness, pertinence, appropriateness , or utility of the documents with respect to the given query.

First, and foremost, relevance is in the mind and eye of the beholder, i.e., the user. Thus, relevance is time and user specific. Moreover, users are influenced by many factors that go beyond whether or not given document is “about “ the topics covered in the user’s queries. For example, the order of a document as presented in a ranked list can affect the user’s opinion. Moreover, sometimes relevance is dictated by the number of documents needed; for examples one world wide web site that tells the user what the weather is going to be tomorrow in Ranchi suffices, so that additional sites that repeat the information become non relevant. These are all factors influencing a user’s judgement of the relevance of a given document. This is why studies using the judgement of the subject experts who can determine, and imprecisely at that, what topics are covered by which documents are seen as limited. What is worse , it is sometimes the case that users can accidentally, i.e. with serendipity, find a good document that is seemingly not on topic, was somehow retrieved, and which the user finds useful. However, it is also possible that documents which the user should find useful, on topic , and good, are seen as non relevant; serendipity in reverse. Thus, the retrieval system is not able to retrieve all the relevant documents and nothing but the relevant document.

5. Use of Fuzzy Set Theory in Information Retrieval System.[4].

Information retrieval involves two finite crisp sets, a set of recognized index terms,
$X = \{x_1,x_2,.........x_n\}$
and a set of relevant documents,
$Y = \{y_1, y_2, ..., y_n\}$

In fuzzy information retrieval, the relevance of index terms to individual documents is expressed by a fuzzy relation,
R = X x Y ------> [0,1],
such that membership value R (x_i, y_j) specifies for each $x_i \in X$ and $y_j \in Y$ the grade of relevance of index term x_i to document y_j.

5.1 Methods for determining grades for documents:

As we have seen vagueness and uncertainly occur in information retrieval system at different levels and sub levels, therefore determining the grades for documents is trivial. In general grades for documents are determined either subjectively by authors of the documents, or objectively, by some algorithmic procedure. One way of determing the grades objectively is to define them in an appropriate way in terms of the numbers of occurrences of individual index terms in titles and/or abstract of the document involved. This can be combined with other criteria. One possible criteria is to discount the grade of relevance involving old documents or old index terms by some rate. Another possible criteria is defining grades of relevance are to discriminate among different types of document such as journal article, papers in conference proceedings unpublished reports etc to rank relevant journals and so on. These and other criteria may be specified by the user.

5.2 Relationship of document and inquiry:

Fuzzy thesaurus plays an important role in establishing relationship between document and inquiry. Fuzzy thesaurus is a reflexive relation, T, defined on X^2. For each pair of index term $<x_i, x_k> \in x^2$, T (x_i, x_k) expresses the association of x_i with x_k; that is the degree to which the meaning of the index term X_k is compatible with meaning of the given index term x_i. The role of this relation is to deal with the problem of synonyms among index terms. The relationship helps to identify relevant documents for a given inquiry that otherwise would not be identified. This happens whenever a document is characterized by an index term that is synonymous with an index term contained in the inquiry.

To construct fuzzy thesauri , experts in a given field are asked to identify, in a given set of index terms, pairs whose meaning they consider associated (or possibly, to give degree of association for each pair). Grades of membership in T are then determined by averaging the scores of each pair. Another way of obtaining these grades is to use statistical data obtained from the document or such as frequencies of associations based on citations. When a fuzzy thesaurus is updated by introducing new index terms, old index terms are usually discounted at some rate.

5.3 Setting up an inquiry:

In fuzzy information retrieval, an inquiry can be expressed by any fuzzy set defined on the set of index terms X. Let A denote the fuzzy set representing a particular inquiry then, by comparing A with fuzzy thesaurus T, we obtain new fuzzy set in X (say, set B), which represents an augmented inquiry (i.e. augmented by associated index terms). That is,

$A.o\ T \;=\; B$ --------------- (1)

where , o is the max-min composition, so that

$$B(x_j) = \underset{x_i \in X}{\text{max.min}} \, [A(x_i), T\,(x_i,\, x_j)]$$

for all $x_j \in X$. The retrieved documents expressed by a fuzzy set D defined on Y, are then obtained by composing the augmented inquiry, expressed by fuzzy set B, with the relevance relation R. That is

$B\ o\ R \;=\; D$ ----------------- (2)

5. EXAMPLE

We consider an inquiry which involves only the following the index terms:

x_1 = Fuzzy logic
x_2 = Fuzzy relation equations
x_3 = Fuzzy modus ponens

That is, $^{o+}A = \{x_1,x_2,x_3\}$ is a support of fuzzy set A expressing the inquiry.
Assume that the vector representation of A is

$$A \;=\; \begin{array}{ccc} x_1 & x_2 & x_3 \\ [1 & .4 & .1] \end{array}$$

Assume that the relevant part of fuzzy thesaurus, restricted to the support of A, and non zero columns, is given by the matrix

$$T = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \end{array} \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \left(\begin{array}{cccccc} 1 & .2 & 1 & 1 & .5 & 1 \\ .2 & 1 & .1 & .7 & .9 & 0 \\ 1 & .4 & 1 & .9 & .3 & 1 \end{array} \right) \end{array}$$

Where ,

x_4 = approximate reasoning
x_5 = max-min composition
x_6 = fuzzy implication

$$B = \begin{array}{c} \begin{array}{cccccc} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{array} \\ \left(\begin{array}{cccccc} 1 & .4 & 1 & 1 & .5 & 1 \end{array} \right) \end{array}$$

Assume now that the relevant part of the relevance relation, restricted to support of B and non zero columns is given by the matrix

$$R = \begin{array}{c} \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{array} \begin{array}{c} \begin{array}{cccccccccc} y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 & y_8 & y_9 & y_{10} \end{array} \\ \left(\begin{array}{cccccccccc} .2 & 0 & 1 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & .3 & 0 & .4 & 0 & 0 & 1 & 0 \\ 0 & 0 & .8 & 0 & .4 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & .9 & .7 & .5 \\ 1 & 0 & .5 & 0 & 0 & .6 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & .2 & 0 & 1 & 0 & 0 & .5 \end{array} \right) \end{array}$$

where y_1 y_2 .... y_{10} are the only documents related to index terms x_1, x_2 .. x_6. By eq (2) the composition BoR results in fuzzy set D, which characterizes the retrieved documents; its vector form is

$$D= \begin{array}{c} \begin{array}{cccccccccc} y_1 & y_2 & y_3 & y_4 & y_5 & y_6 & y_7 & y_8 & y_9 & y_{10} \end{array} \\ \left(\begin{array}{cccccccccc} .5 & 1 & 1 & .3 & .4 & .5 & 1 & .9 & .7 & .5 \end{array} \right) \end{array}$$

The user can now decide whether to inspect all documents capture by the support of D or to consider only documents captured by some α- cuts of D.

6. CONCLUSION

The use of fuzzy set theory in information retrieval has at least the following advantages in comparison with classical methods.

i) Fuzzy relevance relations and fuzzy thesauri are more expressive than their crisp counterparts, and their construction is more realistic;

ii) The fuzzy set characterizing the retrieved documents establishes by its α -cuts, layers of retrieved documents distinguished by their relevance (the value of α) and thus provide the user with a guideline regarding the order in which the documents should be inspected or which document to neglect when total number of retrieved document is too large; and

iii) Fuzzy inquiry provides the user with greater flexibility in expressing the subject area of interest.

7. REFERENCES

1. Bordogra, Glaria & Pasi Gabriella' "Handling vagueness in Information Retrieval System", 0-8186-7174-2195 $ 04.00 (C) 1995 IEEE. PP 110-112.

2. Kroaft, Donald H; D Bordogra Glaria & Pasi , Gabriella; "Information Retrieval System : Where is the Fuzz" 0-7803-4863 – x 198 $ 10.00 (C) 1998 IEEE, PP, 1367- 1372.

3. Riaz Mahamad, "Advanced Indexing and Anstracting Practices", Atlantic Publisher and Distributors, 1989, PP 19-21.

4. Klir, G.J. and Yuan, 60, "Fuzzy Sets and Fuzzy Logic Theory and Application" , PHI 1994, Pp. 379-387.