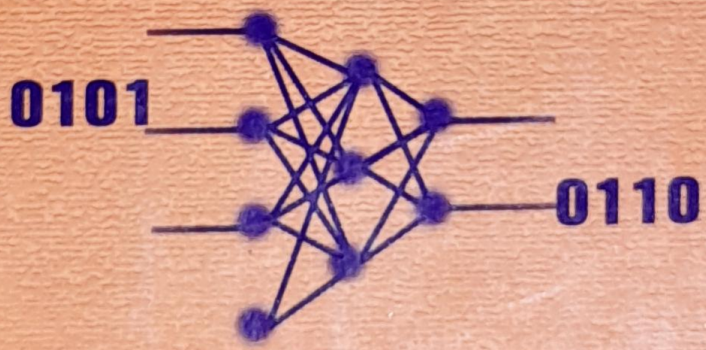
Algorithms And Artificial Systems

CONTROL & COMMUNICATION SYSTEMS
SPEECH & IMAGE PROCESSING SYSTEMS
DATABASE SYSTEMS
ARTIFICIAL NEURAL NETWORKS



PARALLEL & DISTRIBUTED ALGORITHMS
GENETIC ALGORITHMS
AUTOMATA &
THEORY OF COMPUTATION

EDITOR
P. THANGAVEL

Algorithms and Artificial Systems,
2003, pp. 297-307.

Fuzzy logic-based information retrieval

K.R. CHOWDHARY[a] AND V.S. BANSAL[b]

[a]*Dept. of Computer Science & Engg.*
[b]*Dept. of Electrical Engg.,*
M.B.M. Engg. College,
J.N.V. University, Jodhpur.

Abstract

We propose a method for Information Retrieval(IR) using fuzzy logic approach using thesaurus in the form of index. A comparison with IR method using classical logic concludes the fuzzy logic technique as more realistic approach. A typical IR strategy and system structure has also been proposed.

Keywords: Classical Logic, Fuzzy Logic, Information Retrieval.

1 Introduction

The Problem of Information Retrieval in the present context is highly relevant when the volume of Information generated is much more than the individual can digest easily. It is overly difficult to search, locate, and disseminate the precisely desired information from the storage media, irrespective of whether it is local or globally distributed over the electronic media. More and more information, the newly generated as well as old, are finding their place in the digital storage. This process has been fuelled first by arrival of Internet and the

World-Wide-Web, and subsequently by digital libraries, research publications repositories, electronic editions of news papers, journals, and magazines.

The aim of IR is to locate the document or file which holds the desired information. Some times, it is also required to locate the actual position of required information in the document selected when the document is of large size.

The recent work on IR is largely based on binary IR [1], structured language-based IR [7], frequency-based retrieval [4], automatic summarization [6], probabilistic retrieval [2], and document analysis based on the verbs [5], are the major approaches for IR.

IR is a process of matching the patterns in the user inquiry with the patterns in the prospective text documents. If the inquiry words are taken as set of words X, and text document words are taken as set of words Y, then IR is nothing but finding the binary relation $X \times Y$. However, since words may appear in different morphological forms but their meaning remains unchanged before the matching is performed, the words in the inquiry as well as the words in the text document must be reduced to their basic form called *stem* words, by a process called *stemming*. The following section discusses the concept of binary IR using Crisp sets (i.e., when an element is a member of the set a property of the element, called characteristic function, has value 1, and when the element is not a member, its characteristic function value is zero).

This paper presents the fuzzy retrieval technique, which is based on fuzzy set theory and fuzzy logic an extension of the classical set theory. This fuzzy retrieval technique is based on thatthe word matching between the inquiry word set and the text and the text word set should not be limited to the matching of stem words. But, since the words in inquiry also match with their the text documents, with relative degree of matchingthe the matching should not be simply considered as Boolean with values '1' or '0'. In fact the matching should be graded, depending on the degree or level of matching, which can range between 0 and 1. The extremes of this range of 0

to 1, a special case in fuzzy logic, corresponds to the Boolean matching. This logic is more realistic than the Boolean logic, simply due to the fact that it considers the exact matching as in Boolean logic as well as vague matching, the later being more frequently encountered in the real world.

2 Binary Relations in Crisp and Fuzzy Sets

Considering the key words in the inquiry as a set X, and the key words in the document as set Y, a binary relation from X to Y can be represented as ,

$$R : X \times Y \in \{0, 1\} \tag{2.1}$$

If for every $x \in X$, there is a corresponding $y \in Y$ and $R(x, y)$ such that $x = y$, the relation is Reflexive relation, i.e, $R(x, y) = 1$.

In relations over fuzzy sets the elements of two sets have a degree of association as a form of relation rather than simply related (binary 1) or not related (binary 0) as in the case of crisp sets. The degree of association ranges from 0 to 1, where 0 indicates a total absence of relation and a 1 indicates a total presence of the relation; therefore,

$$R : X \times Y \in [0, 1] \tag{2.2}$$

and the range of $R(x, y)$ varies from 0 to 1 depending on how close the y, $y \in Y$, is associated with x, $x \in X$. If for every $x \in X$, there is a corresponding $y \in Y$ such that $R(x, y)$ ranges between 0 and 1, this is fuzzy reflexive relation whose degree is between 0 and 1.

3 Crisp-Set based IR

Consider a set of text documents Y consisting of documents $y_1, y_2, ..., y_n$, as potential documents to be searched for the key-words $x_1, x_2, ...x_m$ in the inquiry set X. Assume that $n = 2$, and y_1, y_2 are as given below: (quoted from, "Einstein - The Life and Times", by Ronald W. Clark).

y_1 : "Thus the new concept of the subatomic world was even by 1920 beginning to produce a gulf. Bohr, Born, and a number of Einstein's other contemporaries, as well as the many of younger men who were in great part responsible for the new idea readily jumped the gap. Einstein stayed where he was. Therefore, the scene in many ways paralleled that into which he has launched his theory of relativity two decades earlier. But then he had been in the iconoclastic vanguard; now he took up station with the small conservative rearguard."

y_2 : "Plank, the man of honor who had yet signed the manifesto of 93, had in fact for the first time done as much to keep Einstein in Berlin as he had done to bring him there in 1914. His letter, which, in Einstein's words, had induced him ...".

Assuming that the inquiry be : *Einstein's Scientific Theory of Relativity*, and therefore the corresponding set of keywords in the inquiry is, $X = \{x_1$="Einstein", x_2 ="Scientific", x_3 = "Theory of Relativity"$\}$, and document set Y consists documents y_1, y_2, therefore, $Y = \{y_1, y_2\}$.

In binary IR it is required to find R - a subset of $X \times Y$, i.e., R("Einstein", y_1), R("Scientific", y_1), R("Theory of Relativity", y_1), and R("Einstein", y_2), R("Scientific", y_2), R("Theory of Relativity", y_2). Where, the first, third and fourth terms results to 1 as perfect match have been found for these terms in the documents, where as the remaining three terms (i.e. 2, 5, 6) results to zero as no match have been found for these. Based on above it can be concluded that the document y_1 has match of 2, out of 3 terms in the inquiry, (a relevancy of the document as 66%), and document y_2 has match of 1, out of 3 terms of the inquiry (a relevancy of 33%), for this particular inquiry. (It is a mere coincident that the sum of relevance from the two documents is 100%). However, if it were only one term in the inquiry, the relevance of the entire document would come out either 100 % or 0 %. For larger number of documents $y_1, y_2, ..., y_n$, the relevance can be computed in the similar manner for a specified inquiry. Once the relevance is computed, the IR system lists the documents in the order of relevance or grading.

In the classical set theory, and the corresponding logic which may be called as binary logic or classical logic or crisp logic; there is either 100% match for an index term in the inquiry and corresponding term in the document, so $R(x_i, y_i) = 1$; or there is no match at all with $R(x_i, y_i) = 0$. However, the reality is not this. It may happen that the two terms from index and document which are being matched are two forms of the same word, e.g., *real* and *reality, phrases and phrases, exact* and *exactness* etc. In all these cases the crisp logic returns zero percent value of relevance, but factually there is almost a complete relevance. Similarly, in the words *good* and *better, bad* and *worse, amazon* and *wonder*, there is high level of relevance, but the crisp-logic will return a relevance of zero for the corresponding relation $R(x_i, y_i)$ for each of the above cases. Number of times the words which are to be matched are synonymous, therefore a simple pattern matching will show no match.

Different morphological form, like *total, totaled, totaling,* can be first reduced to their root form, called stem value, before matching is carried out, and then their stem form will be used for the matching. While crisp logic is being used, however, the problem of matching still remains to be resolved, for those text document words which are either partly or exact synonym of a word(s) in the inquiry. Because, in this particular situation it should not be concluded that there is no match, as degree of matching exists in the meaning of these words. The measure of this degree of match can be taken as ranging from 0 to 1. The fuzzy logic and fuzzy IR, described in the next section takes care of this problem.

4 Fuzzy Information Retrieval

The term Fuzzy Information Retrieval (FIR) refers methods of IR that are based on the fuzzy set theory. These methods are more realistic than those based on classical set theory. In FIR, the relevance of index terms to individual documents

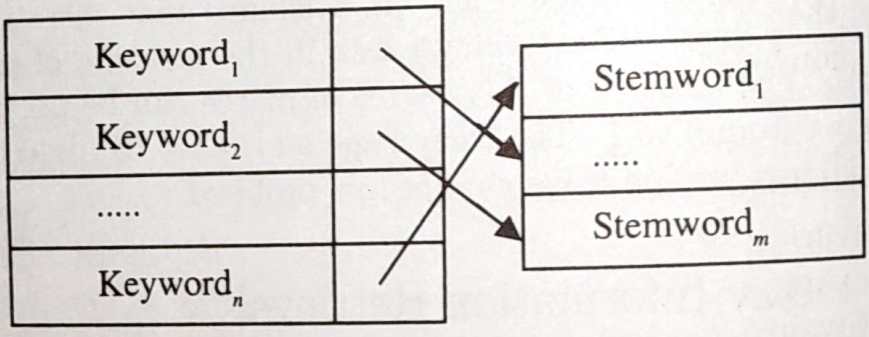
is expressed by the relation [3],

$$R : X \times Y \rightarrow [0, 1] \tag{4.1}$$

such that the membership value $R(x_i, y_i)$ specifies for each $x_i \in X$, $y_i \in Y$, the grade of relevance of index term x_i with the document y_i. The value of these grades range from 0 (no match) to 1 (full match). The grade of relevance depends on many factors (i) position of term y_j in the text document. If the document is a research article, and y_i appears in the list of keywords, in abstract, or in the conclusion part, the relevance is higher; (ii) frequency of occurrence of y_i in the docu-ment, (iii) x_i and y_i are terms formed from the same basic stem word and, (iv) y_i is synonym of x_i the proximity of meaning of x_i and y_i decide value synonym of x_i the proximity of meaning of x_i and y_i decide value of $R(x_i, y_i)$'s closeness with 1.

The criteria (i) and (ii) above are user defined and they can be programmed in the implementation according to the user needs. The stem word criteria (iii) requires a data structure similar to the one shown below in the Figure. 1, which helps to locate the stem word for a given word, and then stem word is substituted in the original text before the retrieval technique is applied on it. Another important relation for IR

Figure 1: Data structure for finding stem words.



based on criteria (iv) above is fuzzy thesaurus, which is basis for FIR. The fuzzy thesaurus shows the relationship between the pairs of words based on their centrality or degree of relevance. The structure [8] of a fuzzy thesaurus is,

<WC1> <WC2> <RD>

where WC stands for word category and RD is relationship degree between the words WC1 and WC2. For a typical case it can be,

attraction, love, .8
studious, hardworking, .9
war, crime, .7

arranged in a matrix form. When relationship degree is zero, no entry is recorded in the fuzzy thesaurus. When the relationship is degree is 1, there is actual synonym, as in,

violence, conflict, 1.0
pain, harm, 1.0

etc. The fuzzy thesaurus can be manually constructed, or can be generated from the lexicons. Transitivity relationship can be applied by computing the missing relationship degrees from the existing ones. The thesaurus is a reflexive fuzzy relation, say T, defined over X^2. For each pair of index terms $(x_i, x_j) \in X^2$, $T(x_i, x_j)$ expresses the degree of association of x_j with x_i, such that the degree to which the meaning of the index term x_j is compatible with the meaning of the index term x_i. The role of this relation is to deal with the problem of synonyms among the index terms. The relation helps to identify the relevant documents which otherwise would not be identified in the absence of perfect match between the keywords in the user inquiry and those in the text document. This happens when the index term characterizing the document happens to be synonym of the index term present in the inquiry.

5 Illustrative Example of FIR

Various methods exist for construction of fuzzy thesaurus. For example, experts in the domain of text can be asked to identify, in a given set of index terms, the pairs of words whose meaning they consider are associated, and provide the degree of association for each pair. In FIR an inquiry can be expressed in the form of a fuzzy set (say Q) based on the index term X. Then, by composing Q with the fuzzy thesaurus T, we obtain a new fuzzy set on X, say A which represents the

augmented inquiry, i.e.,

$$A = Q \circ T \tag{5.1}$$

where $\circ$ is called *max-min* composition operator, so that,

$$A(x_j) = max - min[Q(x_i), T(x_i, x_j)] \tag{5.2}$$

$x_i \in X$, for all $x_j \in X$. The retrieved documents, expressed by a set F defined over Y, are then obtained by composing the augmented inquiry, expressed by the fuzzy set A, with the relevance relation R, i.e.,

$$F = A \circ R \tag{5.3}$$

Therefore, (5.1) and (5.3) above represent the process of FIR. Let us illustrate the above mentioned process using a simple example in which the inquiry is same as it was for crisp set logic in the beginning of this article, i.e., Q=*Einstein's Scientific Theory of Relativity*, and let the index terms are: x_1 = "Einstein", x_2="Scientific", x_3="Theory of Relativity". Therefore, $Q = \{x_1, x_2, x_3\}$ is fuzzy set expressing the inquiry. Assuming that vector representation of Q is,

$$Q = \begin{matrix} x_1 & x_2 & x_3 \\ [1 & .6 & .8] \end{matrix} \tag{5.4}$$

where 1, .6, and .8 are called the centralities of x_1, x_2, x_3 respectively. The centrality indicates the presence of certain qualities, whose computations are modeled as computation of fuzzy membership degree. The relevant part of the fuzzy thesaurus T, restricted to the support of Q is given by the matrix:

$$T = \begin{matrix} \\ x_1 \\ x_2 \\ x_3 \end{matrix} \begin{matrix} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{matrix} \\ \begin{bmatrix} 1 & .6 & .9 & 0 & 0 & 0 \\ .6 & 1 & .8 & .6 & .5 & .9 \\ .9 & .8 & 1 & 0 & .1 & .7 \end{bmatrix} \end{matrix} \tag{5.5}$$

Where x_4 = "Bohr", x_5 = "subatomic", x_6 = "New idea". Then, the composition of $Q \circ T$ results in fuzzy set A, which represents the augmented inquiry. The vector form of A is:

$$A = max - min[Q \circ T]$$

$$= \begin{matrix} & x_1 & x_2 & x_3 \end{matrix} \begin{bmatrix} 1 & .6 & .8 \end{bmatrix} \circ \begin{matrix} \\ x_1 \\ x_2 \\ x_3 \end{matrix} \begin{matrix} x_1 & x_2 & x_3 & x_4 & x_5 & x_6 \end{matrix} \begin{bmatrix} 1 & .6 & .9 & 0 & 0 & 0 \\ .6 & 1 & .8 & .6 & .5 & .9 \\ .9 & .8 & 1 & 0 & .1 & .7 \end{bmatrix}$$

$$= [1 \quad .8 \quad .9 \quad .6 \quad .5 \quad .7]$$

Assume that part of the relevance relation R, which is restricted to support A, is given by the matrix,

$$R = \begin{matrix} & y_1 & y_2 \\ x_1 \\ x_2 \\ x_3 \\ x_4 \\ x_5 \\ x_6 \end{matrix} \begin{bmatrix} 1 & 1 \\ .2 & .4 \\ 1 & 0 \\ 1 & 0 \\ 1 & .1 \\ .5 & .2 \end{bmatrix}$$

where y_1, y_2 are the documents related to index terms $x_1, x_2, \ldots, x_6$. By composing $A \circ R$ as per (5.3) above, it results in a fuzzy set F, which characterizes the retrieval documents. The vector form of F, for documents y_1, and y_2 is,

$$F = max - min(A \circ R)$$

$$= max - min([1 \quad .8 \quad .9 \quad .6 \quad .5 \quad .7] \circ \begin{bmatrix} 1 & 1 \\ 2 & .4 \\ 1 & 0 \\ 1 & 0 \\ 1 & .1 \\ .5 & .2 \end{bmatrix}) \tag{5.6}$$

$$= [1 \quad 1]$$

The fuzzy retrieval matrix F in the above shows that y_1, and y_2 are relevant documents, where as crisp logic earlier shown these to be partially relevant. Once the FIR system lists the documents with their relevance value, the user can now decide whether to inspect all the retrieved documents supported by the fuzzy set F or to inspect only some of the documents depending on the degree of association of the document with the index terms.

6 Conclusion and Future Developments

When fuzzy set theory is used for IR, the main advantage is that the fuzzy relevance relation and fuzzy thesaurus are more expressive than their crisp set counter parts. Also, since the degree of association is returned along with the retrieved documents, it helps the user to decide the order in which the documents can be viewed, particularly when the documents are in large number.

The FIR promises higher potential for cross-language text processing and IR. Every language and its semantics have close association with the culture in which it has its roots, and therefore, exact matching terms for any language are not possible in other languages. In fact a degree of relevance or, only fuzzy relation exists between the matching words of the two or more languages. We have planned to work on the cross language areas, which includes English, Hindi, and Sanskrit languages. The Sanskrit being the oldest, developed, and a rich language, having the oldest literature which is yet unexplored by many, and not accessible to most the FIR would help in accelerating research in this language, and subsequently its possible adoption in modern times. The work in this direction has been planned at this place to be carried out in a phased manner, though presently at the initial stage.

References

[1] Salton G., Fox E.A., and Wu Harry, "Extended Boolean Information Retrieval", Communications of the ACM, Vol. 26, No.12, pp. 1022-1036, Dec. 1983.

[2] Fung R. and Favero B.D., "Applying Bayesian Networks to Information Retrieval", Communications of ACM, March 1995/Vol.38, No. 3, pp. 42-57.

[3] Klir G.J. and Yuan B.; Fuzzy Sets and Fuzzy Logic Theory and Applications, Prentice-Hall, 1997.

[4] Niwa Y., Nishioka S., Iwayama M. and Takano A., "Topic Graph Generation for Query Navigation : Use of Frequency Classes for Topic Extraction", In Proceedings of NLPRS'97, Natural Language Processing Pacific Rim Symposium, pp. 95-100, Phuket, Thailand, 1997.

[5] Klavans J., Kan M., "Role of Verbs in Document Analysis", In Proceedings of the 17th international Conference on Computa-tional linguistics(COLING-ACL '98) Montreal, Canada: Aug. 1998, pp.680-686.

[6] Hahn U. and Mani I., "The Challenges of Automatic Summarization", In IEEE Computer, Nov. 2000, Vol. 33, No.11, pp. 29-36.

[7] Chelba C. and Mahajan M.; "Information Extraction using Structured language Model", EMNLP / NAACL Conference, 2001.

[8] Subasic P. and Huettner A., "Affect Analysis of Text using Fuzzy Semantic Typing", In IEEE Trans. on Fuzzy Systems, Vol. 9, No.4, Aug. 2001, pp. 483-496.