Fifth IFSA World Congress (1993), 489-492

The Effect of Positive Compensation on Retrieval Effectiveness

Won Yong Kim, Joon Ho Lee, Yoon Joon Lee and Myoung Ho Kim

Department of Computer Science
Korea Advanced Institute of Science and Technology
373-1, Kusung-dong, Yusung-gu, Taejon, 305-701, Korea

ABSTRACT

Due to adverse properties of MIN and MAX operators, the fuzzy set model generates incorrect document rankings in certain cases. In the area of fuzzy set theory a variety of fuzzy operators have been developed, which can replace the MIN and MAX operators. In this paper we analyze how the fuzzy operators affect document rankings. We describe that the fuzzy operators with positive compensation properties, i.e. positive compensatory operators provid better retrieval effectiveness than the others. It is also shown through performance evaluation that the fuzzy set model based on positively compensatory operators gives high quality document rankings.

1. Introduction

The ranked output facility is an important component of Information Retrieval(IR) systems because it minimizes users' efforts spent to find relevant information. Boolean retrieval systems have been most widely used among commercially available IR systems. Conventional boolean retrieval systems, however, do not support document ranking.

The fuzzy set model overcomes an inablity to rank documents by using document term weights[1, 2, 3]. Though the fuzzy set model is an elegant approach, it generates in certain cases incorrect document rankings not to agree with humans' intuition[4, 5]. This is because the MIN and MAX operators give the resulting value that depends on only one operand without considering the other.

Since the first introduction of fuzzy set theory a variety of fuzzy operators have been proposed for the AND and OR operations. They can be classified into three groups such as averaging operators, T-norms and T-conorms depending on their operational characteristics[6]. In this paper we first describe that the problems of the fuzzy set model cannot be overcome even though the MIN and MAX operators are replaced with any types of T-operators, i.e. T-norms and T-conorms. We then present a class of averaging operators called positively compensatory operators is suitable for achieving high retrieval effectiveness, which is shown through performance experiments.

The remainder of this paper is organized as follows. Section 2 describes the conventional fuzzy set model and its problems. In section 3 we present the effect of various fuzzy operators on retrieval effectiveness. The performance evaluation and concluding remarks are given in sections 4 and 5, respectively.

2. Fuzzy Set Model

An IR system based on the fuzzy set model can be defined as a quadruple $< T, Q, D, F >$.

T is a set of index terms used to represent queries and documents.

Q is a set of queries that can be recognized by the system. A query $q \in Q$ is constructed from the terms in T and logical operators AND, OR and NOT.

D is a set of documents. Each document $d \in D$ is represented by $((t_1, w_1), \ldots, (t_n, w_n))$ where w_i designates the weight of term t_i in d and w_i takes a value between zero and one, i.e. $0 \leq w_i \leq 1$.

F is a retrieval function taking a pair (d,q) in $D \times Q$ to a *document value* in the closed interval [0,1]. The document value means the similarity between the document d and the query q. The retrieval function F(d,q) is defined as follows :

1. When a query q is composed of an index term t_i, the function F(d,q) is defined as the weight of t_i in document d, i.e. w_i.
2. Given a complex query with logical operators, it is evaluated by applying the following formulas. The evaluation proceeds recursively from the innermost clause.

$$F(d, q_1 \text{ AND } q_2) = MIN(F(d, q_1), F(d, q_2))$$
$$F(d, q_1 \text{ OR } q_2) = MAX(F(d, q_1), F(d, q_2))$$
$$F(d, \text{NOT } q_1) = 1 - F(d, q_1)$$

A document value for a query is a measure to rank the document. The fuzzy set model, however, has been criticized to generate inappropriate document values in certain cases[4, 5]. Example 1 illustrates that the ranked ouput in the fuzzy set model does not agree with humans' intuition. This is because the MIN and MAX operators have the *single operand dependency problem* – they generate the resulting value which depends on only one operand without considering the other. Although we explain only problems incurred by the AND operation, it should be noted that the use of the OR operation causes similar problems.

Example 1: Suppose that we have two documents d_1 and d_2 shown below. The documents are represented by two pairs of an index term and its weight.

$$d_1 = \{(\text{Thesaurus}, 0.40), (\text{Clustering}, 0.40)\}$$
$$d_2 = \{(\text{Thesaurus}, 0.99), (\text{Clustering}, 0.39)\}$$
$$q_1 = \text{Thesaurus AND Clustering}$$

When the MIN operator is used for the AND operatioin, the documents values of d_1 and d_2 for the query q_1 are evaluated as 0.40 and 0.39, repectively. Hence, d_1 is retrieved with a higher rank than d_2. Most people, however, will obviously decide that d_2 rather than d_1 is more similar to q_1.

3. Effect of Fuzzy Operators on Retrieval Effectiveness

3.1 Classification of Fuzzy Operators

There are a variety of fuzzy operators corresponding to a given classical operators, and the different operators have different properties. These operators can be classified into three groups such as averaging operatos, T-norms and T-conrms[6]. Fig 1 shows the T-norms and T-conorms called T-operators, Fig 2 gives the averaging operators.

	T(x, y)	Tc(x, y)	Comment
1	MIN (x, y)	MAX (x, y)	
2	$x \bullet y$	$x + y - xy$	
3	MAX (x + y - 1, 0)	MIN (x + y, 1)	
4	$\frac{xy}{x + y - xy}$	$\frac{x + y - 2xy}{1 - xy}$	
5	$\begin{cases} x & \text{if } y = 1 \\ y & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} x & \text{if } y = 0 \\ y & \text{if } x = 0 \\ 1 & \text{otherwise} \end{cases}$	
6	$\frac{\lambda xy}{1 - (1 - \lambda)(x + y - xy)}$	$\frac{\lambda(x + y) + xy(1 - 2\lambda)}{\lambda + xy(1 - \lambda)}$	$0 \leq \lambda \leq \infty$
7	$MAX\ (1 - ((1 - x)^P + (1 - y)^P)^{1/P}, 0)$	$MIN\ ((x^P + y^P)^{1/P}, 1)$	$1 \leq P \leq \infty$
8	$\frac{1}{1 + \left(\left(\frac{1}{x} - 1\right)^{\lambda} + \left(\frac{1}{y} - 1\right)^{\lambda}\right)^{1/\lambda}}$	$\frac{1}{1 + \left(\left(\frac{1}{x} - 1\right)^{-\lambda} + \left(\frac{1}{y} - 1\right)^{-\lambda}\right)^{-1/\lambda}}$	$0 \leq \lambda \leq \infty$
9	$\frac{xy}{MAX\ (x, y, \lambda)}$	$1 - \frac{(1 - x)(1 - y)}{MAX\ (1 - x, 1 - y, \lambda)}$	$0 \leq \lambda \leq 1$
10	$MAX\left(\frac{x + y - 1 + \lambda xy}{1 + \lambda}, 0\right)$	$MIN\ (x + y + \lambda xy, 1)$	$-1 \leq \lambda \leq \infty$
11	$MAX\ ((1 + \lambda)(x + y - 1) - \lambda xy, 0)$	$MIN\ (x + y + \lambda xy, 1)$	$-1 \leq \lambda \leq \infty$

Fig 1. The T-Operators

$$\begin{array}{lll}
(A_1) & (x \bullet y)^{1-\gamma} \bullet (x + y - x \bullet y)^{\gamma}, & 0 \leq \gamma \leq 1 \\
(A_2) & (1 - \gamma) \bullet MIN(x, y) + \gamma \bullet MAX(x, y), & 0 \leq \gamma \leq 1 \\
(A_3) & (1 - \gamma) \bullet (x \bullet y) + \gamma \bullet (x + y - x \bullet y), & 0 \leq \gamma \leq 1 \\
(A_{4.AND}) & \gamma \bullet MIN(x, y) + \frac{(1-\gamma)(x+y)}{2}, & 0 \leq \gamma \leq 1 \\
(A_{4.OR}) & \gamma \bullet MAX(x, y) + \frac{(1-\gamma)(x+y)}{2}, & 0 \leq \gamma \leq 1
\end{array}$$

Fig 2. The Averaging Opertors

3.2 Problems of the T-Operators

The T-operators except MIN and MAX operators have the following two common properties. First, when one operand value of the two is 0 or 1, they generate the same resulting value as MIN and MAX operator(i.e. they can not overcome the single operand dependency problem). Second, they allow some compensation between two operand values in other cases, and the resulting value is less than the lower operand value, or greater than the higher.

When the fuzzy set model uses one of the T-operators except MIN and MAX operators as evaluation formulas for the AND and OR operations, the second property can alleviate the problem illustrated in Example 1. For example, suppose that the product operator, i.e. $x \bullet y$ is used instead of the MIN operator in Example 1. Then the document values of d_1 and d_2 are evaluated as 0.16 and 0.39 respectively, and hence d_2 is retrieved with a higher rank than d_1. The second property, however, brings out a new problem called *negative compensation problem*, which is shown in the next example.

Example 2: Suppose a document d_3 and two queries q_1 and q_2 are given as follows:

$d_3 = \{(\text{Thesaurus}, 0.70), (\text{Clustering}, 0.70), (\text{System}, 0.70)\}$

$q_1 = \text{Thesaurus AND Clustering}$

$q_2 = \text{System}$

Though the fuzzy set model uses any operators, the similarity between q_2 and d_3 is evaluated as 0.70 that is the weight of term 'System' in d_3. The T-operators except MIN and MAX operators always decide that the similarity between q_1 and d_3 is less than 0.70. Note that the similarity between q_1 and d_3 is less than that between q_2 and d_3, which clearly does not agree with most people's decision.

3.3 Positively Compensatory Operators

The single operand dependency and negative compensation problems can be overcome if fuzzy operators have the property generating a resulting value between the lower operand and the higher operand. The Operators A_2 and A_4, which will be called *positively compensatory operators*, have the aforementioned property. We propose to use them as evaluation formulas of the fuzzy set model.

Averaging operator A_1 and A_3 have the single operand dependency or negative compensation problem in some cases. When one operand value of the two is 0, A_1 constantly generates 0, i.e. it can not overcome the single operand dependency problem. In some cases the resulting value generated by A_1 or A_3 is lower or higher than the two operand values, i.e. they have the negative compensation problem.

Though the averaging operators A_2 and A_4 are independently developed by different researchers at different time, they are mathematically equivalent. The distinction of the AND and OR operation separates the averaging operator A_2 into two parts as follows:

$$(A_{2.AND})\ (1-\gamma) \bullet MIN(x,y) + \gamma \bullet MAX(x,y), 0 \le \gamma \le 0.5$$
$$(A_{2.OR})\ \ (1-\gamma) \bullet MIN(x,y) + \gamma \bullet MAX(x,y), 0.5 \le \gamma \le 1$$

In order to coincide the value range of the parameter of $A_{2.AND}$ with that of the parameter of $A_{2.OR}$, we change the operator $A_{2.AND}$ to a different form having the same value. By replacing γ with $1-\gamma$, we obtain the following expression.

$$(A'_{2.AND})(1-\gamma) \bullet MAX(x,y) + \gamma \bullet MIN(x,y), 0.5 \le \gamma \le 1$$

Then we can transform $A_{2.AND}$ and $A_{2.OR}$ into $A_{4.AND}$ and $A_{4.OR}$, respectively, by replacing γ with $(\gamma+1)/2$.

The extended boolean model has been known as an effective retrieval model in the area of IR[7]. The difference between the extended boolean and the fuzzy set model is in the evaluation formulas. The following evaluation formulas of the extended boolean model are also positively compensatory operators.

$$(E_{AND})\quad 1 - \left[\frac{(1-x)^p + (1-y)^p}{2}\right]^{1/p} \quad 1 \le p \le \infty$$
$$(E_{OR})\quad \left[\frac{x^p + y^p}{2}\right]^{1/p} \quad 1 \le p \le \infty$$

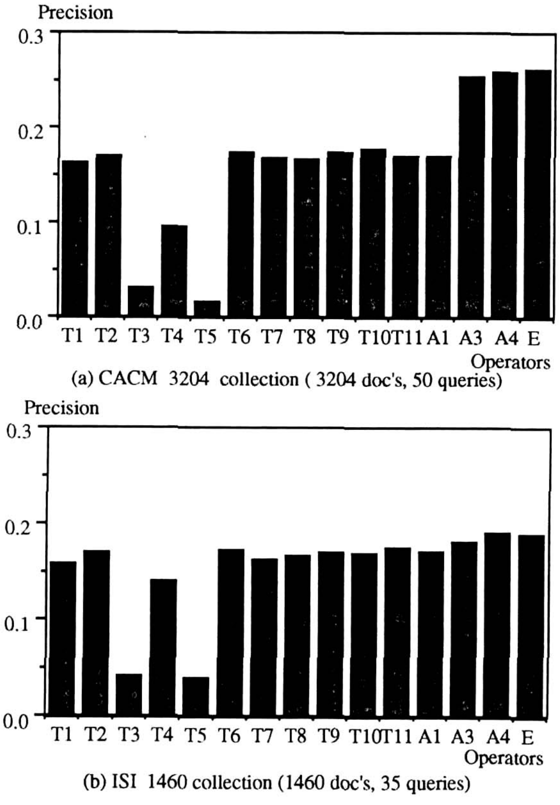
4. Performance Evaluation

Two different document collections, ISI 1460 and CACM 3204[7] are used to compare retrieval effectiveness of the positively compensatory operators with others. ISI 1460 consists of 1460 documents and 35 queries. CACM 3204 consists of 3204 documents and 50 queries. Both collections also contain relevance assessment of each document with respect to each query.

To evaluate the effectiveness of an IR system, it is customary to compute values of the recall and precision. The *recall* represents the proportion of relevant items retrieved out of the total number of relevant in the whole collection and the *precision* represents the proportion of relevant items retrieved out of the total number retrieved.

Some retrieval system providing ranking facility make it possible to compute a recall and a precision value after the retrieval of each item. By interpolation the precision values can be calculated for fixed values of the recall, say, for a recall of 0.1, 0.2 and so on up to a recall of 1.0[8]. By averaging the precision values at a fixed recall level for a number of user queries, one finally obtains the precision at the recall level.

For easy comparison we use a single precision value that represents the average precision at three typical recall levels, including a low recall level of 0.25, a medium recall of 0.50, and a high recall of 0.75. When an operator has a parameter, we find the parameter value providing the best precision, and then empoly the precision to compare retrieval effectiveness. Fig 3 shows the evlauation results.



(a) CACM 3204 collection (3204 doc's, 50 queries)

(b) ISI 1460 collection (1460 doc's, 35 queries)

Fig 3. The retrieval effectiveness of fuzzy operators

5. Concluding Remarks

It has been argued that the conventional fuzzy set model based on the MIN and MAX operators is not appropriate as a model of IR systems. This is because the MIN and MAX operators have properties adverse to effective document ranking. In rescent years a variety of fuzzy operators have been developed, which are classified into T-operators and averaging operators. We have shown that the problems of the conventional fuzzy set model are not overcome with any types of T-operators. We have then proposed to use the positively compensatory operators as the boolean evaluation formulas for the fuzzy set model. Performance evaluation has shown that the positively compensatory operators provide better retrieval effectiveness than any other operators.

References

[1] A.Bookstein, "A Comparison of Two Weighting Schemes for Boolean Retrieval," *Journal of the Americal Society for Information Science*, Vol. 32, No. 4, pp. 275-279, 1981.

[2] T. Radecki, "Fuzzy Set Theoretical Approach to Document Retrieval," *Information Processin & Management*, Vol. 15, No. 5, pp 247-259, 1979.

[3] W.G. Waller and D.H. Kraft, "A Mathematical Model of a Weighted Boolean Retrieval System," *Information Processing & Management*, Vol. 15, pp. 235-245, 1979.

[4] A. Bookstein, "Fuzzy Requests: An Approach to Weighted Boolean Seraches," *Journal of the American Sociery for Information Science*, Vol. 31, No. 4, pp. 240-247, 1980.

[5] S.E. Robertson, "On the Nature of Fuzz: A Diatribe," *Journal of the American Society for Information Science*, Vol. 29, No. 6, pp. 304-307, 1978.

[6] H.J. Zimmermann, "Fuzzy Set Theory and Its Applications," *2nd ed., Kluwer Academic Publishers*, 1991.

[7] G. Salton, E.A. Fox, and H. Wu, "Extended Boolean Information Retrieval," *Communications of the ACM*, Vol. 26, No. 11, pp. 1022-1036, 1983.

[8] G. Salton, Ed. "The Smart Retrieval System-Experiments in Automatic Document Processing", *Prentice Hall. Inc., Englewood Cliffs, New Jersey*, 1971.