Analyzing the Effectiveness of Extended Boolean Models in Information Retrieval

Joon Ho Lee*
Department of Computer Science
Cornell University
Ithaca, NY 14853-7501

Abstract

Many extended Boolean models such as fuzzy set, p-norm, et al. have been proposed in the information retrieval literature to support ranking facility for the Boolean retrieval system. They can be explained within the same framework, and each extended Boolean model is characterized by evaluation formulas for AND and OR operations. A variety of operators have been also developed in the area of fuzzy set theory for AND and OR operations, and can be used in extended Boolean models. In this paper we analyze the behavioral aspects of various operators for AND and OR operations, and address important properties in terms of retrieval effectiveness. Our analyses show that the four properties, namely single operand dependency, negative compensation, double operand dependency and unequal importance decrease retrieval effectiveness in some circumstances. This suggests that the two properties, namely positive compensation and equal importance might help retrieval effectiveness. We also provide the experimental results that coincide with our analyses.

1 Introduction

Information Retrieval (IR) systems represent, store and retrieve the input information, which is likely to include the natural language text of documents or of document excerpts and abstracts [1]. The output of IR systems in response to a search request consists of references. These references are intended to provide the system users with the information about items of potential interest. A major role of IR systems, however, is not just to generate a set of relevant references, but to help determine which documents are most likely to be relevant to the given requirements. IR systems should present to users a sequence of documents ranked in decreasing order of query-document similarities. Users are able to minimize their time spent to find useful information by reading the top-ranked documents first.

The Boolean model has been most widely used among commercially available IR systems because it is easy to implement and is very efficient in terms of query processing time. In the standard

*On leave from Korea Research and Development Information Center, Korea Institute of Science and Technology, P.O. Box 1, Yusung, Taejon 305-600, Korea

Boolean model, each document is indexed with a set of keywords or terms, and each query is represented by a legitimate Boolean expression. The Boolean expression consists of a set of index terms connected by the Boolean operators AND, OR and NOT. The documents retrieved by a given query are those satisfying the expression of the query. The standard Boolean model, however, does not provide document ranking facility because it cannot compute similarity coefficients between queries and documents.

The fuzzy set model, which uses the original fuzzy set operators, i.e. MIN and MAX, has been developed in the past to support the ranking facility for the Boolean retrieval system [2, 3, 4]. The fuzzy set model, however, has been criticized to generate incorrect ranked output in some circumstances because MIN and MAX operators do not well correspond with human behavior for the calculation of query-document similarities [5, 6]. The Waller-Kraft [7], Paice [8], p-norm [9, 10], infinite-one [11], network Boolean [12] and INQUERY Boolean [13] models have been proposed in the information retrieval literature. They use different operators for AND and OR operations, which can avoid the problem of MIN and MAX operators. All of the aforementioned models can be explained within the same framework, namely extended Boolean model.

The theory of fuzzy sets was founded by Zadeh, primarily to analyze complex systems [14]. In the original theory of fuzzy sets the specific operators MIN and MAX were used for AND and OR operations, respectively. When the range of operand values is restricted to the set $\{0, 1\}$, the MIN and MAX operators perform precisely as the corresponding operators for classical sets, thus establishing fuzzy sets as clear generalization of classical sets. It is now understood, however, that these operators are not the only possible generalization of the classical set operators [15]. A variety of fuzzy operators have been developed in fuzzy set theory since late 1970, and different operators have different operational characteristics. They can be incorporated in the extended Boolean models.

A few operators, which might give the higher effectiveness of information retrieval, have been evaluated in the information retrieval literature. Theoretically, however, there are an infinite number of operators for AND and OR operations. This makes it impossible to evaluate all of the operators that can evaluate AND and OR operations in extended Boolean models. Furthermore, even if we evaluate all the operators known so far and use the operators providing the highest retrieval effectiveness, we cannot exclude a possibility of finding better operators. In this paper, by analyzing some well-known operators for AND and OR operations, some fundamental properties of the operators are investigated which may increase or decrease retrieval effectiveness. Even though this paper does not present optimal operators, the properties given in this paper can be used as a high base line to approach optimal operators.

First of all, we introduce a variety of operators from the areas of information retrieval and fuzzy set theory that can be used for AND and OR operations in extended Boolean models, and indicate that some operators are mathematically equivalent. We then address important properties of the operators to affect retrieval effectiveness. We explain through counter-intuitive examples that the properties of single operand dependency, negative compensation, double operand dependency and unequal importance could decrease retrieval effectiveness in some circumstances. This suggests that the properties of positive compensation and equal importance might help retrieval effectiveness. As a result, we define an operator class called n-ary soft Boolean operators that is suitable for achieving high retrieval effectiveness.

The remainder of this paper is organized as follows. Section 2 gives the description of extended Boolean models, and introduce a variety of operators that can be used for AND and OR operations in extended Boolean models. We analyze the binary forms of the operators in section 3, and analyze their n-ary forms in section 4. Section 5 describe experimental results to show the effectiveness of the various operators. Finally, conclusion and future works are given in section 6.

2 Extended Boolean Models

Extended Boolean models can support document ranking facility for the conventional Boolean retrieval system by calculating the similarities between documents and Boolean queries. An IR system based on extended Boolean models can be defined by the quadruple $< T, D, Q, F >$, where

- T is a set of index terms used to represent queries and documents.
- 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 document d and w_i may take any value between zero and one, i.e. $0 \leq w_i \leq 1$.
- Q is a set of queries that can be recognized by the system. Each query $q \in Q$ is a legitimate Boolean expression composed of index terms and Boolean operators AND, OR and NOT. In some extended Boolean models, queries can be also formulated with term and clause weights. Even though it has been known that weighting query terms and clauses provides better retrieval effectiveness, considering its effect on retrieval effectiveness is beyond the scope of the present study.
- F is a ranking function

  $$F : D \times Q \rightarrow [0, 1]$$

  which assigns to each pair (d, q) a number in the closed interval [0, 1]. This number is a measure of similarity between document d and query q, and is called the *document value* for document d with respect to query q. The retrieval function F is defined as follows:

  1. For each term t_i in query q, the function $F(d, t_i)$ is defined as the weight of term t_i in document d, i.e. w_i.
  2. Boolean operators, i.e. AND, OR and NOT are then evaluated by applying the corresponding formulas. The evaluation formulas of the operators are an important factor to determine the quality of ranked output. $F(d, \text{NOT } t_i)$ is evaluated as $1 - w_i$, and a variety of operators for AND and OR operations will be reviewed in the following sections. For Boolean queries containing more than one Boolean operator, the evaluation proceeds recursively from the innermost clause.

Table 1: T-operators

	AND	OR	
T_1	$\text{MIN}(x,y)$	$\text{MAX}(x,y)$	
T_2	$x \cdot y$	$x+y-xy$	
T_3	$\text{MAX}(x+y-1,0)$	$\text{MIN}(x+y,1)$	
T_4	$\frac{xy}{x+y-xy}$	$\frac{x+y-2xy}{1-xy}$	
T_5	$\begin{cases} x & \text{if } y=1 \\ y & \text{if } x=1 \\ 0 & \text{otherwise} \end{cases}$	$\begin{cases} x & \text{if } y=0 \\ y & \text{if } x=0 \\ 1 & \text{otherwise} \end{cases}$	
T_6	$\frac{\lambda xy}{1-(1-\lambda)(x+y-xy)}$	$\frac{\lambda(x+y)+xy(1-2\lambda)}{\lambda+xy(1-\lambda)}$	$0 \le \lambda \le \infty$
T_7	$\text{MAX}(1-((1-x)^p+(1-y)^p)^{1/p},0)$	$\text{MIN}((x^p+y^p)^{1/p},1)$	$1 \le p \le \infty$
T_8	$\frac{1}{1+\left((\frac{1}{x}-1)^\lambda+\left(\frac{1}{y}-1\right)^\lambda\right)^{1/\lambda}}$	$\frac{1}{1+\left((\frac{1}{x}-1)^{-\lambda}+\left(\frac{1}{y}-1\right)^{-\lambda}\right)^{-1/\lambda}}$	$0 \le \lambda \le \infty$
T_9	$\frac{xy}{\text{MAX}(x,y,\lambda)}$	$1-\frac{(1-x)(1-y)}{\text{MAX}(1-x,1-y,\lambda)}$	$0 \le \lambda \le 1$
T_{10}	$\text{MAX}((1+\lambda)(x+y-1)-\lambda xy,0)$	$\text{MIN}(x+y+\lambda xy,1)$	$-1 \le \lambda \le \infty$

Table 2: Averaging operators

A_1	$(1-(1-w_1)\cdot\ldots\cdot(1-w_n))^\gamma \cdot (w_1 \cdot w_2 \cdot \ldots \cdot w_n)^{1-\gamma}$	$0 \le \gamma < 1$
A_2	$\gamma \cdot \text{MAX}(w_1,w_2,\ldots,w_n)+(1-\gamma)\cdot\text{MIN}(w_1,w_2,\ldots,w_n)$	$0 \le \gamma \le 1$
A_3	$\gamma\cdot(1-(1-w_1)\cdot\ldots\cdot(1-w_n))+(1-\gamma)\cdot(w_1\cdot w_2\cdot\ldots\cdot w_n)$	$0 \le \gamma \le 1$
$A_{4.\text{AND}}$	$\gamma\cdot\text{MIN}(w_1,\ldots,w_n)+\frac{(1-\gamma)(w_1+\ldots+w_n)}{n}$	$0 \le \gamma \le 1$
$A_{4.\text{OR}}$	$\gamma\cdot\text{MAX}(w_1,\ldots,w_n)+\frac{(1-\gamma)(w_1+\ldots+w_n)}{n}$	$0 \le \gamma \le 1$

Table 3: The operators from information retrieval

fuzzy set	$T_{1.\text{AND}}$	$\text{MIN}(w_1,w_2)$	
	$T_{1.\text{OR}}$	$\text{MAX}(w_1,w_2)$	
Waller-Kraft	A_2	$\gamma\cdot\text{MAX}(w_1,w_2,\ldots,w_n)+(1-\gamma)\cdot\text{MIN}(w_1,w_2,\ldots,w_n)$	$0 \le \gamma \le 1$
Paice	P_{AND}	$\frac{\sum_{i=1}^n (r^{i-1}\cdot w_i)}{\sum_{i=1}^n r^{i-1}}$, w_i's are considered in ascending order	$0 \le r \le 1$
	P_{OR}	$\frac{\sum_{i=1}^n (r^{i-1}\cdot w_i)}{\sum_{i=1}^n r^{i-1}}$, w_i's are considered in descending order	$0 \le r \le 1$
p-norm	PN_{AND}	$1-\left(\frac{(1-w_1)^p+\ldots+(1-w_n)^p}{n}\right)^{1/p}$	$1 \le p \le \infty$
	PN_{OR}	$\left(\frac{w_1^p+\ldots+w_n^p}{n}\right)^{1/p}$	$1 \le p \le \infty$
infinite-one	$A_{4.\text{AND}}$	$\gamma\cdot\text{MIN}(w_1,\ldots,w_n)+\frac{(1-\gamma)(w_1+\ldots+w_n)}{n}$	$0 \le \gamma \le 1$
	$A_{4.\text{OR}}$	$\gamma\cdot\text{MAX}(w_1,\ldots,w_n)+\frac{(1-\gamma)(w_1+\ldots+w_n)}{n}$	$0 \le \gamma \le 1$
network Boolean	A_3	$\gamma\cdot(1-(1-w_1)\cdot\ldots\cdot(1-w_n))+(1-\gamma)\cdot(w_1\cdot w_2\cdot\ldots\cdot w_n)$	$0 \le \gamma \le 1$
INQUERY Boolean	$T_{2.\text{AND}}$	$w_1 \cdot w_2$	
	$T_{2.\text{OR}}$	$w_1+w_2-w_1\cdot w_2$	

2.1 Evaluation formulas from fuzzy set theory

The development of fuzzy set theory has allowed researchers to apply set theoretic concepts to sets of objects whose membership values vary in the interval [0, 1]. This applicability has been achieved by defining new operators for classical set theory operators. Often there is more than one fuzzy operator corresponding to a given classical operator, and different operators have different operational characteristics. A variety of fuzzy operators have been developed for AND and OR operations in fuzzy set theory. They can be classified into two groups such as T-operators and averaging operators [16], which will be denoted by T and A, respectively, for simplicity.

The T-operators consisting of T-norms and T-conorms originated from the studies of probabilistic metric spaces. Later, it was proposed that T-norms and T-conorms could be used for AND and OR operations, respectively. Table 1 shows some well-known T-operators [17]. The ten T-operators are classified into non-parameterized T-operators, i.e. $T_1 - T_5$ and parameterized T-operators, i.e. $T_6 - T_{10}$. The non-parameterized T-operators have the following relationships for any two values x and y.

$$T_{1.AND}(x,y) \geq T_{4.AND}(x,y) \geq T_{2.AND}(x,y) \geq T_{3.AND}(x,y) \geq T_{5.AND}(x,y)$$
$$T_{1.OR}(x,y) \leq T_{4.OR}(x,y) \leq T_{2.AND}(x,y) \leq T_{3.OR}(x,y) \leq T_{5.OR}(x,y)$$

In the parameterized T-operators the resulting value is controlled by the given parameter. The non-parameterized T-operators can be considered as special cases of the parameterized T-operators that have the following relationships.

$$\begin{array}{ll}
T_{4.AND}(x,y) \geq T_{6.AND}(x,y) \geq T_{5.AND}(x,y), & T_{4.OR}(x,y) \leq T_{6.OR}(x,y) \leq T_{5.OR}(x,y) \\
T_{1.AND}(x,y) \geq T_{7.AND}(x,y) \geq T_{3.AND}(x,y), & T_{1.OR}(x,y) \leq T_{7.OR}(x,y) \leq T_{3.OR}(x,y) \\
T_{1.AND}(x,y) \geq T_{8.AND}(x,y) \geq T_{5.AND}(x,y), & T_{1.OR}(x,y) \leq T_{8.OR}(x,y) \leq T_{5.OR}(x,y) \\
T_{1.AND}(x,y) \geq T_{9.AND}(x,y) \geq T_{2.AND}(x,y), & T_{1.OR}(x,y) \leq T_{9.OR}(x,y) \leq T_{2.OR}(x,y) \\
T_{2.AND}(x,y) \geq T_{10.AND}(x,y) \geq T_{5.AND}(x,y), & T_{2.OR}(x,y) \leq T_{10.OR}(x,y) \leq T_{5.OR}(x,y)
\end{array}$$

In fuzzy decision theory the "decision" has been viewed as the intersection or the union of fuzzy sets, and T-operators have been used to model human decisions in many cases. However, it has been noted that neither T-norms nor T-conorms are appropriate to model "managerial decisions". Averaging operators have been developed to overcome this problem, and Table 2 shows some well-known averaging operators, i.e. $A_1 - A_4$ [18, 19].

2.2 Evaluation formulas from information retrieval

The fuzzy set model has been proposed by studying the use of fuzzy set theory in information retrieval [2, 3, 4]. It uses MIN and MAX operators to evaluate AND and OR operations, respectively. Though the fuzzy set model can provide ranked retrieval output, it has been criticized to generate incorrect ranked output in some circumstances [5, 6]. Many retrieval models such as Waller-Kraft, Paice, p-norm, infinite-one, network Boolean and INQUERY Boolean have been proposed as alternatives to the fuzzy set model. Different retrieval models provide different evaluation formulas for AND and OR operations, which are shown in Table 3.

We have reviewed the operators that can evaluate AND and OR operations in extended Boolean models. Even though the operators are independently proposed in many cases by different researchers for different purposes such as information retrieval and fuzzy set theory, closer investigation shows that some operators from information retrieval are equivalent to those from fuzzy set theory as follows:

$$\text{fuzzy set} = T_1$$

$$\text{Waller-Kraft} = A_2$$

$$\text{infinite-one} = A_4$$

$$\text{network Boolean} = A_3$$

$$\text{INQUERY Boolean} = T_2$$

The five operators of fuzzy set, Waller-Kraft, infinite-one, network Boolean and INQUERY Boolean can be found in the area of fuzzy set theory. This phenomenon might happen because the research on fuzzy set theoretic approaches in the information retrieval literature faded out since late 1970, whereas many fuzzy operators, which can replace MIN and MAX, have been developed after that time. In the remainder of this paper, the notations such as T_1, A_2, A_4, A_3 and T_2 will be used for the operators of fuzzy set, Waller-Kraft, infinite-one, network Boolean and INQUERY Boolean, respectively, for notational convenience.

The evaluation formulas of the T-operators are binary functions having only two operands whereas those of the averaging operators, Paice and p-norm are n-ary functions having two or more operands. This is due to the fact that the binary forms of the n-ary functions do not satisfy associativity. The use of binary functions violating associativity results in the problem of generating different document values for two logically equivalent queries "(t_1 AND t_2) AND t_3" and "t_1 AND (t_2 AND t_3)". This problem will be detailed in section 4.1. In the following two sections, we will analyze the sixteen distinct operators out of twenty-one in terms of retrieval effectiveness — investigate the binary and n-ary functions in section 3 and section 4, respectively.

3 Analyzing Binary Operators

We have presented the following sixteen distinct operators that can evaluate AND and OR operations in extended Boolean models.

$$T_1 - T_{10}, \quad A_1 - A_4, \quad P, \quad PN$$

Some of the operators, which do not satisfy associativity, were given in n-ary forms. Their binary forms are shown in Table 4. Closer investigation shows that some binary forms of the n-ary operators are mathematically equivalent as follows [20].

$$A_{2.B} = A_{4.B} = P_B$$

Table 4: Binary forms of the n-ary operators

$A_{1.B}$	$(w_1 + w_2 - w_1 \cdot w_2)^{\gamma} \cdot (w_1 \cdot w_2)^{1-\gamma}$	$0 \leq \gamma < 1$
$A_{2.B}$	$\gamma \cdot \mathrm{MAX}(w_1, w_2) + (1-\gamma) \cdot \mathrm{MIN}(w_1, w_2)$	$0 \leq \gamma \leq 1$
$A_{3.B}$	$\gamma \cdot (w_1 + w_2 - w_1 \cdot w_2) + (1-\gamma) \cdot (w_1 \cdot w_2)$	$0 \leq \gamma \leq 1$
$A_{4.B.AND}$	$\gamma \cdot \mathrm{MIN}(w_1, w_2) + \frac{(1-\gamma)(w_1+w_2)}{2}$	$0 \leq \gamma \leq 1$
$A_{4.B.OR}$	$\gamma \cdot \mathrm{MAX}(w_1, w_2) + \frac{(1-\gamma)(w_1+w_2)}{2}$	$0 \leq \gamma \leq 1$
$P_{B.AND}$	$\frac{1}{1+r} \cdot \mathrm{MIN}(w_1, w_2) + \frac{r}{1+r} \cdot \mathrm{MAX}(w_1, w_w)$	$0 \leq r \leq 1$
$P_{B.OR}$	$\frac{1}{1+r} \cdot \mathrm{MAX}(w_1, w_2) + \frac{r}{1+r} \cdot \mathrm{MIN}(w_1, w_w)$	$0 \leq r \leq 1$
$PN_{B.AND}$	$1 - \left(\frac{(1-w_1)^p + +(1-w_2)^p}{2}\right)^{1/p}$	$1 \leq p \leq \infty$
$PN_{B.OR}$	$\left(\frac{w_1^p + w_2^p}{2}\right)^{1/p}$	$1 \leq p \leq \infty$

Therefore, the following fourteen binary operators out of sixteen are distinct, and their properties are analyzed in the next sections.

$$T_1 - T_{10}, \quad A_{1.B} - A_{3.B}, \quad PN_B$$

3.1 Properties adverse to retrieval effectiveness

We have analyzed the behavioral aspects of various fuzzy operators, and have addressed important issues to affect retrieval effectiveness [21, 22]. As a result, we have described the two problems, namely single operand dependency and negative compensation to decrease retrieval effectiveness.

Single Operand Dependency Problem: If an operator θ has a single operand dependent property, then $\theta(x, y) = x$ or y for some $x, y \in [0, 1]$ such that $x \neq y$. When an extended Boolean model uses for AND or OR operations the operators having the single operand dependent property, it causes the single operand dependency problem. For example, suppose that we have two documents d_1, d_2 and a query q_1 shown below. The documents are represented by two pairs of an index term and its weight.

$$d_1 = \{(\text{Information}, 0.4), (\text{Retrieval}, 0.4)\}$$
$$d_2 = \{(\text{Information}, 1), (\text{Retrieval}, 0.4)\}$$
$$q_1 = \text{Information AND Retrieval}$$

When $T_{1.AND}$, i.e. MIN is used for the AND operation, the document values of d_1 and d_2 for query q_1 are the same, i.e. 0.4. Most people, however, will obviously decide that d_2 rather than d_1 is more similar to q_1. This undesirable result is due to the fact that the MIN operator has the single operand dependent property — gives the resulting value that depends on only one operand without considering the other.

The compensatory property reflecting both operand values on the document value can alleviate the single operand dependency problem in some circumstances. For example, when $T_{2.AND}$, i.e.

product operator replaces the MIN operator in the above example, the document values of d_1 and d_2 are evaluated as 0.16 and 0.4 respectively, and hence d_2 is retrieved with a higher rank than d_1. However, if the compensated value is less than the lower operand value or greater than the higher operand value, it may result in another problem called negative compensation.

Negative Compensation Problem: If an operator θ has a negatively compensatory property, then $\theta(x, y) < \text{MIN}(x, y)$ or $\theta(x, y) > \text{MAX}(x, y)$ for some $x, y \in [0, 1]$. When an extended Boolean model uses for AND or OR operations the operators having the negatively compensatory property, it causes the negative compensation problem. For example, suppose that two documents d_3, d_4 and a query q_2 are given as follows:

$$d_3 = \{(\text{Information}, 0.9), (\text{Retrieval}, 0.01)\}$$
$$d_4 = \{(\text{Information}, 0.1), (\text{Retrieval}, 0.1)\}$$
$$q_2 = \text{Information AND Retrieval}$$

When $T_{2.AND}$, i.e. product operator is used for the AND operation, the document values of d_3 and d_4 with respect to query q_2 are evaluated as 0.009 and 0.01, respectively. The similarity between q_2 and d_3 is less than that between q_2 and d_4, which may not agree with most people's decision. This is because the product operator has the negatively compensatory property, that is, $x \cdot y < \text{MIN}(x, y)$ for all $x, y \in (0, 1)$.

3.2 Positively compensatory operators

If an operator θ has a positively compensatory property, then $\text{MIN}(x, y) < \theta(x, y) < \text{MAX}(x, y)$ for some $x, y \in [0, 1]$. The problems of single operand dependency and negative compensation can be overcome if the operator incorporated in the extended Boolean has the positively compensatory property for any $x, y \in [0, 1]$ — generates a resulting value between the lower operand value and the higher operand value for any operand values. The positively compensatory operator is defined as a function of the form:

$$p : [0, 1] \times [0, 1] \rightarrow [0, 1]$$

It satisfies at least the following properties:

Property p1 $p(x, x) = x$; that is, p is idempotent.
Property p2 $\text{MIN}(x, y) < p(x, y) < MAX(x, y)$ where $x \neq y$.

The operator graph is useful to analyze the behavioral properties of binary operators. The operator graph of an operator is a set of points that are computed by applying the operator to some pairs of operand values. The pairs of operand values are given by any arbitrary operand graphs. Figure 1(a) shows how the operator graph can be constructed where the vertical coordinate denotes the values ranging from zero to one. The shapes of operand graphs A and B was presented in [17]. Given an operator, its operator graph can be drawn as follows: For an element in the set of objects, we can have a pair of operand values, e.g. α and β. Then, we can get a point on the operator graph, e.g. γ by applying the operator to the operand values. The operator graph can be constructed by repeating the same procedure for every element in the set of objects. The operator graphs for T_1, i.e. MIN and MAX, for instance, are depicted in Figure 1(b), where the bold line represents the values of MIN and the other line represents the values of MAX.

We can easily understand the behavioral properties of an binary operator from its operator graph. For example, Figure 1(b) shows that MIN and MAX operators are single operand dependent because all points of the operator graphs are on either one operand graph or the other. For each operator graph in Figure 2, the corresponding operator has the following properties: (a) single operand dependent and negatively compensatory, (b) single operand dependent, negatively compensatory and positively compensatory, (c) negatively compensatory and positively compensatory, (d) positively compensatory.

Figure 3 shows the operator graphs of the fourteen distinct binary operators. Though we do not provide the operator graphs for the parameterized T-operators, i.e. $T_6 - T_{10}$, the shapes of the operator graphs can be derived from those relationships given section 2.1. We can easily understand the following properties of the operators through their operator graphs.

T_1	single operand dependent
$T_2 - T_{10}$	single operand dependent, negatively compensatory
$A_{1.B}$	single operand dependent, negatively compensatory, positively compensatory
$A_{2.B}$	positively compensatory
$A_{3.B}$	negatively compensatory, positively compensatory
PN_B	positively compensatory

3.3 Binary soft Boolean operators

The operator graphs of the binary operators $A_{2.B}$ and PN_B suggest that the operators may not have the properties of single operand dependency and negative compensation, and may belong to positively compensatory operators that can avoid the problems of single operand dependency and negative compensation. Such suggestion is proved in the following. $A_{2.B}$ and PN_B are functions of the form

$$b : [0,1] \times [0,1] \rightarrow [0,1]$$

They satisfy at least the following properties:

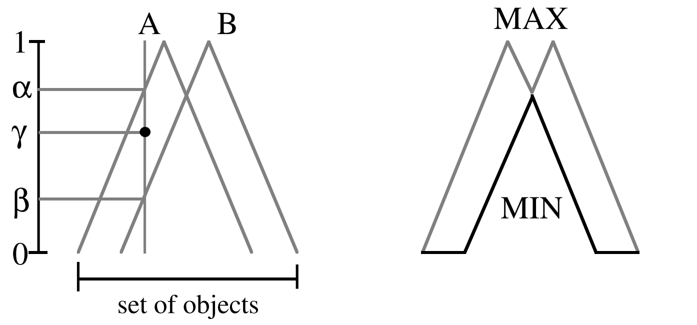
Property b1 $b(x,x) = x$; that is, b is *idempotent.*

Property b2 If $x < x'$ then $b(x,y) < b(x',y)$, and if $y < y'$ then $b(x,y) < b(x,y')$; that is, b is *strictly monotonic* on each operand.

Property b3 b is a *continuous* function.

Property b4 $b(x,y) = b(y,x)$; that is, b is *commutative.*

The first property indicates that a document is retrieved with the same document values with respect to the queries "t_1", "t_1 AND t_1", and "t_1 OR t_1". The second property is the natural requirement that an increase in an document term weight produces an increase in the document



(a) Creating the operator graph

(b) The operator graphs for MIN and MAX operators

Figure 1: The operator graph

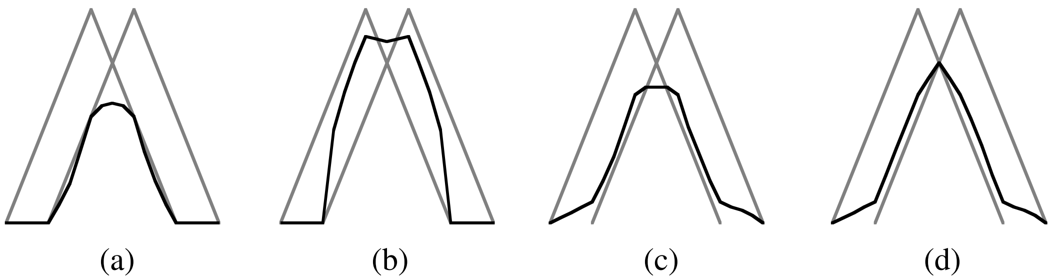


Figure 2: The behavioral properties of various operators

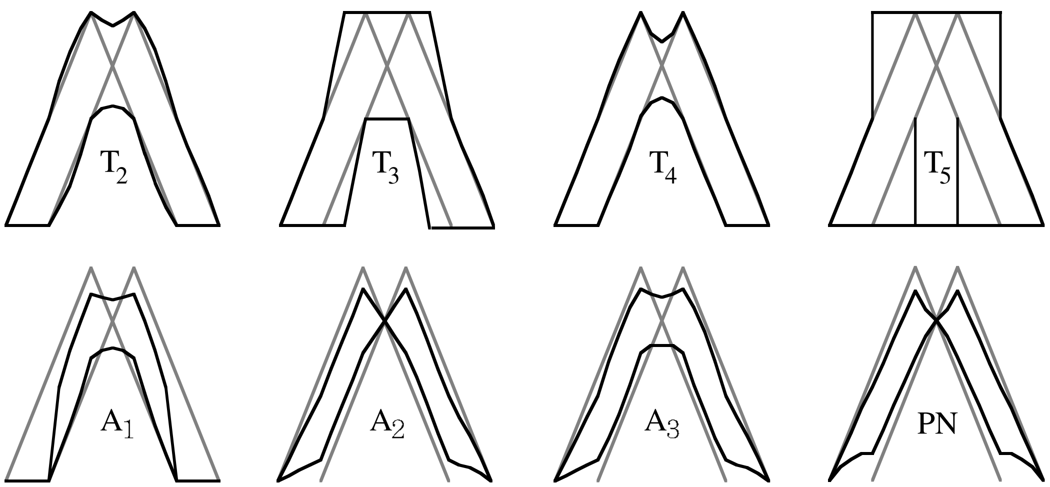


Figure 3: Operator graphs for various operators

value. The property of continuity prevents a situation in which a very small increase in an document term weight produces a large change in the document value. The property of commutativity insures that a document is retrieved with the same document value for the request "t_1 AND t_2" as for the request "t_2 AND t_1", and similarly for "t_1 OR t_2" and "t_2 OR t_1". In the remainder of this paper an operator satisfying the properties $b1$ through $b4$ will be called the *binary soft Boolean operator*. The following theorem shows that a binary soft Boolean operator belongs to positively compensatory operators.

Theorem 1 *If an operator θ satisfies the properties $b1$ and $b2$, then* $\mathrm{MIN}(x,y) < \theta(x,y) < \mathrm{MAX}(x,y)$ *for all $x, y \in [0,1]$ such that $x \neq y$.*

Proof: By the property of strict monotonicity we have that

$$\theta(\mathrm{MIN}(x,y), \mathrm{MIN}(x,y)) < \theta(x,y) < \theta(\mathrm{MAX}(x,y), \mathrm{MAX}(x,y)),$$

and by the property of idempotency we have that

$$\theta(\mathrm{MIN}(x,y), \mathrm{MIN}(x,y)) = \mathrm{MIN}(x,y),$$

$$\theta(\mathrm{MAX}(x,y), \mathrm{MAX}(x,y)) = \mathrm{MAX}(x,y).$$

Hence, $\mathrm{MIN}(x,y) < \theta(x,y) < \mathrm{MAX}(x,y)$, which concludes the proof.

4 Analyzing N-ary Operators

We have presented the sixteen distinct operators such as $T_1 - T_{10}$, $A_1 - A_4$, P and PN that can evaluate AND and OR operations in extended Boolean models. We have also shown that the binary forms of the operators A_2, A_4, P and PN belong to positively compensatory operators that can avoid the problems of single operand dependency and negative compensation. In what follows, we analyze the properties of the four distinct n-ary operators A_2, A_4, P and PN, which are the operators of the Waller-Kraft, infinite-one, Paice and p-norm models, respectively.

4.1 Necessity of n-ary operators

AND operations combine terms into term phrases and OR operations treat two terms as effectively synonymous. Thus, we often use consecutive AND operations such as "fuzzy AND set AND theory" and consecutive OR operations such as "human OR people OR person OR man". Unfortunately, a binary soft Boolean operator cannot satisfy associativity.

Theorem 2 *If an operator θ satisfies the properties $b1$ and $b2$, θ cannot be associative.*

Proof: Let x and y be such that $x < y$. From the proof of Theorem 1, we know that

$$\theta(x,y) > x.$$

Hence, through strict monotonicity on the first component we have that

$$\theta(\theta(x,y),y) > \theta(x,y).$$

However, if θ is associative

$$\theta(\theta(x,y),y) = \theta(x,\theta(y,y)) = \theta(x,y).$$

Therefore, the associativity of θ entails a contradiction, which completes the proof.

If an operator violating associativity is used in extended Boolean models, different document values are obtainable for logically equivalent queries such as $q_1 = t_1$ AND (t_2 AND t_3) and $q_2 =$ (t_1 AND t_2) AND t_3. For example, suppose that a document d is represented by three pairs of an index term and its weight $\{(t_1,1),(t_2,0.7),(t_3,0.5)\}$. An extended Boolean model using $A_{4.AND}$ ($\gamma = 0.3$) gives 0.721 and 0.607 as document values of document d with respect to queries q_1 and q_2, respectively.

Given a query $q = t_1$ OR t_2 OR t_3, users assume that query terms t_1, t_2 and t_3 have equal importance in query q. However, we cannot preserve this basic assumption in evaluating query q if we use binary soft Boolean operators. Consider two ways of evaluating query q such as "(t_1 OR t_2) OR t_3" and "t_1 OR (t_2 OR t_3)". The document value for "(t_1 OR t_2) OR t_3" is more dependent on term t_3 than terms t_1 and t_2 while that for "t_1 OR (t_2 OR t_3)" is more dependent on term t_1 than terms t_2 and t_3.

Suppose that we have a document $d = \{(t_1,w_1),(t_2,w_2),\ldots,(t_n,w_n)\}$ and a query $q = t_1$ OR t_2 OR ... OR t_n. When we apply the operator $A_{4.B.OR}$ ($\gamma = 0$) from the left of query q to the right, the similarity between query q and document d is calculated as follows:

$$\begin{aligned} F(d,q) &= \tfrac{w_1}{2^{n-1}} + \tfrac{w_2}{2^{n-1}} + \tfrac{w_3}{2^{n-2}} + \tfrac{w_4}{2^{n-3}} + \ldots + \tfrac{w_n}{2} \\ &= \tfrac{1}{2}\left(\tfrac{w_1}{2^{n-2}} + \tfrac{w_2}{2^{n-2}} + \tfrac{w_3}{2^{n-3}} + \tfrac{w_4}{2^{n-4}} + \ldots + w_n\right) \end{aligned}$$

The above expression shows that the document value of document d is more affected by the weight of term t_i than the weight of term t_j where i is greater than j.

4.2 N-ary soft Boolean operators

The Waller-Kraft, infinite-one, Paice and p-norm models have avoided the problem of violating associativity by expanding their operators into n-ary forms. Waller-Kraft, infinite-one and Paice, however, have undesirable properties of generating incorrect ranked outputs in certain cases. They do not satisfy the usual assumption that all the terms given in a query are equally important, which results in the following problems.

Double Operand Dependency Problem: The Waller-Kraft model considers only two operand values, i.e. the minimum and maximum operand values in calculating query-document similarities. For example, suppose that two documents d_5, d_6 and a query q_3 are given as follows.

$$d_5 = \{(t_1, 0), (t_2, 0.9), (t_3, 0.9), \ldots, (t_{99}, 0.9), (t_{100}, 1)\}$$
$$d_6 = \{(t_1, 0), (t_2, 0.1), (t_3, 0.1), \ldots, (t_{99}, 0.1), (t_{100}, 1)\}$$
$$q_3 = t_1 \text{ AND } t_2 \text{ AND} \ldots \text{AND } t_{100}$$

Waller-Kraft decides that document values of d_5 and d_6 for query q_3 are the same. Most people, however, will obviously decide that d_5 is more similar to q_3 than d_6.

Unequal Importance Problem: The infinite-one model avoids the double operand dependency problem by reflecting all operand values on document values. It, however, generates the document value that is more affected by minimum and maximum values for AND and OR operations, respectively. For example, suppose that we have two documents d_7, d_8 and a query q_4 shown below.

$$d_7 = \{(t_1, 0), (t_2, 1), (t_3, 1), \ldots, (t_{100}, 1)\}$$
$$d_8 = \{(t_1, 0.4), (t_2, 0.6), (t_3, 0.6), \ldots, (t_{100}, 0.6)\}$$
$$q_4 = t_1 \text{ AND } t_2 \text{ AND} \ldots \text{AND } t_{100}$$

infinite-one ($\gamma = 0.5$) decides that the similarity between d_7 and q_4 is equal to that between d_8 and q_4. Note that the increase in the weight of term t_1 corresponds to the reduction in the weights of all the remaining terms t_2 through t_{100}. Most people, however, will obviously decide that d_7 is more similar to q_4 than d_8.

Unequal Importance Problem: The Paice model also avoids the double operand dependency problem by reflecting all operand values on document values. It, however, gives different importance to each term in a query in that document values are more dependent on smaller term weights for AND operations, and are more dependent on larger term weights for OR operations. For example, suppose that two documents d_9, d_{10} and a query q_5 are given as follows:

$$d_9 = \{(t_1, 0.1), (t_2, 0.3), (t_3, 0.3), (t_4, 0.3), (t_5, 0.3), (t_6, 0.3)\}$$
$$d_{10} = \{(t_1, 0.1), (t_2, 0.7), (t_3, 0.3), (t_4, 0.3), (t_5, 0.3), (t_6, 0.2)\}$$
$$q_5 = t_1 \text{ AND } t_2 \text{ AND } t_3 \text{ AND } t_4 \text{ AND } t_5 \text{ AND } t_6$$

Since Paice ($r = 0.7$) evaluates document values of d_9 and d_{10} for q_5 as 0.2319 and 0.2310, respectively, d_9 is retrieved with a higher rank than d_{10}. This is because the small reduction in the weight of term t_6 dominates the large increase in the weight of term t_2, which may not agree with human behavior.

Since the p-norm model takes into account all operand values with equal importance, it does not cause the problems of double operand dependency and unequal importance. Its n-ary operators are functions of the form

$$n : [0,1] \times \ldots \times [0,1] \rightarrow [0,1]$$

They satisfy at least the following properties:

Property n1	$n(x, x, \ldots, x) = x$; that is, n is *idempotent*.
Property n2	n is *strictly monotonic* on each operand.
Property n3	n is a *continuous* function.
Property n4	$n(x_1, x_2, \ldots, x_n) = n(y_1, y_2, \ldots, y_n)$ where $y_1, y_2, \ldots, y_n$ is any permutation of $x_1, x_2, \ldots, x_n$; that is, n is a *symmetric* function.
Property n5	n considers all operands *equally important*.

The first property indicates that the queries "t_1", "t_1 AND ... AND t_1", and "t_1 OR ... OR t_1" generate the same document values. The second property implies that an increase in an document term weight produces an increase in the document value. The property of continuity prevents a situation in which a very small increase in an document term weight produces a large change in the document value. The symmetric property insures that the request "t_1 AND t_2 AND t_3 AND t_4" generates the same document value as the requests made by the permutation of terms t_1 through t_4, e.g. "t_2 AND t_1 AND t_4 AND t_3". The fifth property of equal importance reflects the usual assumption that all operands are equally important. Note that Waller-Kraft does not satisfy the properties $n2$ and $n5$, and infinite-one and Paice do not satisfy the property $n5$. In the remainder of this paper an operator satisfying the properties $n1$ through $n5$ will be called the *n-ary soft Boolean operator*. The operators of p-norm have the properties of positive compensation and equal importance, which prevent the problems of single operand dependency, negative compensation, double operand dependency and unequal importance. Hence, the p-norm model may be more effective than the others in terms of retrieval effectiveness.

5 Performance Evaluation

A new large data collection is used to evaluate the effectiveness of the various operators in extended Boolean models. The large data collection consists of one of the TREC (Text REtrieval Conference) subcollections, namely Wall Street Journal Disk 2 (WSJ.D2) and 100 Boolean queries. The WSJ.D2 document collection contains 74,520 articles published in the Wall Street Journal between 1990 and 1992 [23]. The 100 Boolean queries were formulated from TREC topics 51 – 150 at Virginia Tech [24].

5.1 Document term weighting

Extended Boolean models exploit document term weights to calculate document values. The experimental collections do not have specific information about the actual importance of document terms. In this case, the weights of document terms can be derived from their occurrence frequency [27]. The effectiveness of a information retrieval system is much dependent on the weighting used in the system [28]. Thus, we use two well-known weighting schemes, namely *Fox-weights* and *INQUERY-weights*.

The following weighting scheme referred to as Fox-weights was introduced by E.A. Fox in his Ph.D. thesis [29]. The weight of term t_k in document d_i is calculated as

$$w_{ik.Fox} = \begin{cases} (0.5 + 0.5 \cdot \frac{tf_{ik}}{maxtf_i}) \cdot \frac{\log(N/n_k)}{\log(N)} & \text{if } tf_{ik} > 0 \\ 0 & \text{if } tf_{ik} = 0 \end{cases}$$

where tf_{ik} is the occurrence frequency of term t_k in document d_i, $maxtf_i$ is the maximum frequency of any term (excluding function words) in document d_i, N is the number of documents in the collection, and n_k is the number of documents to which term t_k is assigned. The two extended Boolean models such as INQUERY Boolean and network Boolean were developed based on the following specific weighting schemes referred to as INQUERY-weights [12].

$$w_{ik.INQUERY} = \begin{cases} (0.4 + 0.6 \cdot (0.4 + 0.6 \cdot \frac{\log(tf_{ik}+0.5)}{\log(maxtf_i+1.0)})) \cdot \frac{\log(N/n_k)}{\log(N)} & \text{if } tf_{ik} > 0 \\ 0.4 & \text{if } tf_{ik} = 0 \end{cases}$$

The two weighting schemes such as Fox-weights and INQUERY-weights generate different range of document term weights as follows:

$$0 \leq w_{ik.Fox} \leq 1$$

$$0.4 \leq w_{ik.INQUERY} < 1$$

The document term weight is minimum in the both weighting schemes when tf_{ik} is equal to zero, and is maximum when tf_{ik} is equal to $maxtf_i$ and n_k is equal to one. The properties of an operator depend on the range of document term weights. Table 5 and Table 6 show the positive and negative properties of the sixteen distinct operators in terms of retrieval effectiveness when document terms are weighted with the two weighting schemes.

The properties of single operand dependency, negative compensation, double operand dependency can be considered as special cases of the unequal important property. For example, the single and double operand dependent properties reflect only one and two operand values, respectively, on document values, and the negatively compensatory property makes document values more affected by minimum and maximum operand values for AND and OR operations, respectively. However, the terminology of the unequal importance property will be used to refer to the above unequal importance problem caused by the infinite-one and Paice models. The operators T_2 – T_{10} and A_1 have the single operand dependent property when document terms are weighted with Fox-weights. They, however, do not have the single operand dependent property when document terms are weighted with INQUERY-weights.

5.2 Experimental results

We performed the experiments to evaluate the effectiveness of the sixteen distinct operators that can evaluate AND and OR operations in extended Boolean models. Before we describe experimental results, the following relationships should be noted again.

$$T_1 = \text{fuzzy set}$$

$$T_2 = \text{INQUERY Boolean}$$

Table 5: Properties of various operators (Fox-weights)

	Positive	Negative
p-norm	positively compensatory equally important	
T_1		single operand dependent
$T_2 - T_{10}$		single operand dependent negatively compensatory
A_1	positively compensatory	single operand dependent negatively compensatory
A_2	positively compensatory	double operand dependent
A_3	positively compensatory	negatively compensatory
A_4	positively compensatory	unequally important
Paice	positively compensatory	unequally important

Table 6: Properties of various operators (INQUERY-weights)

	Positive	Negative
p-norm	positively compensatory equally important	
T_1		single operand dependent
$T_2 - T_{10}$		negatively compensatory
A_1	positively compensatory	negatively compensatory
A_2	positively compensatory	double operand dependent
A_3	positively compensatory	negatively compensatory
A_4	positively compensatory	unequally important
Paice	positively compensatory	unequally important

$$A_2 = \text{Waller-Kraft}$$

$$A_3 = \text{network Boolean}$$

$$A_4 = \text{infinite-one}$$

We evaluate the various operators with the experimental collection. We retrieve top-ranked 200 documents for the 100 Boolean queries, and evaluate the performance using the total number of documents retrieved, the total number of relevant documents retrieved and the 11-point average precision. % change of the 11-point average precision is also given with respect to the effectiveness of the p-norm model. The experimental results are shown in Table 7 and Table 8 where document term weights are calculated with fox-weights and INQUERY-weights, respectively. The following results might be drawn from the experimental evidence.

- Each run should retrieve at least 200 documents for each query — obtain at least 200 documents whose document values are greater than zero for each query. However, Table 7 shows that the operators T_1 – T_{10} and A_1 retrieve those less than 200 documents for some specific queries when document terms are weighted with Fox-weights. The major reason is that Fox-weights gives the weight of zero to the term that is not included in the document, that is, $w_{ik} = 0$ if $tf_{ik} = 0$. The operators $T_{1.AND}$ – $T_{10.AND}$, and A_1 give zero as a resulting value when one of operand values is zero. INQUERY-weights can alleviate this problem because it assigns the weight of 0.4 to the term that is not included in the document, that is, $w_{ik} = 0.4$ if $tf_{ik} = 0$. Table 8 shows that the operators except T_3 and T_5 retrieve at least 200 documents for each query. The poor retrieval effectiveness of T_3 and T_5 is due to the fact that $T_{3.AND}(x, y) = 0$ if $x + y \leq 0$, and $T_{5.AND}(x, y) = 0$ if $x, y \neq 1$

- The tables show that extended Boolean models using T_2, T_9 and T_{10} provides the same retrieval effectiveness. This is because T_9 is equivalent to T_2 when the parameter values of AND and OR are 1.0, and T_{10} is also equivalent to T_2 when the parameter values are -1.0. The tables also show that T_2 and T_4 generating similar operator graphs provide slightly different retrieval effectiveness.

- When document terms are weighted with INQUERY-weights, the fuzzy set model has only the single operand dependency property and T_2 – T_{10} have only the negatively compensatory property. The latter provides significantly better retrieval effectiveness that the former except T_3 and T_5. This suggests that the single operand dependency problem may be more adverse to retrieval effectiveness than the negative compensation problem.

- While both Waller-Kraft and network Boolean have the positively compensatory property, the former causes the double operand dependency problem and the latter results in the negative compensation problem. Waller-Kraft provides 0.2286 (0.2469) when Fox-weights (INQUERY-weights) is used for document term weighting while network Boolean gives 0.2763 (0.2500). This suggests that the double operand dependency problem as well as the single operand dependency problem may be more adverse to retrieval effectiveness than the negative compensation problem.

Table 7: The effectiveness of various operators in extended Boolean models (WSJ.D2, Fox-weights)

	Parameter Val (AND : OR)	Total Retrieved	Total Rel Retrieved	11-pt Avg Precision
p-norm	2.0 : 2.0	20000	2609	0.2929
T_1, fuzzy set		16816	1744	0.1489 (-49.0%)
T_2, INQUERY Boolean		16816	2152	0.2265 (-22.4%)
T_3		1626	305	0.0475 (-83.7%)
T_4		16816	2124	0.2188 (-25.0%)
T_5		9350	1212	0.1035 (-64.5%)
T_6	1.5 : 1.5	16816	2162	0.2275 (-22.1%)
T_7	13.0 : 13.0	16076	1961	0.1776 (-39.2%)
T_8	0.8 : 0.8	16816	2129	0.2220 (-24.0%)
T_9	1.0 : 1.0	16816	2152	0.2265 (-22.4%)
T_{10}	-1.0 : -1.0	16816	2152	0.2265 (-22.4%)
A_1	0.5 : 0.5	294	37	0.0139 (-95.3%)
A_2, Waller-Kraft	0.4 : 0.6	20000	2255	0.2286 (-21.7%)
A_3, network Boolean	0.5 : 0.5	20000	2482	0.2763 (-5.3%)
A_4, infinite-one	0.1 : 0.1	20000	2559	0.2851 (-2.3%)
Paice	1.0 : 1.0	20000	2528	0.2741 (-6.1%)

Table 8: The effectiveness of various operators in extended Boolean models (WSJ.D2, INQUERY-weights)

	Parameter Val (AND : OR)	Total Retrieved	Total Rel Retrieved	11-pt Avg Precision
p-norm	2.0 : 2.0	20000	2570	0.2893
T_1, fuzzy set		20000	1799	0.1562 (-46.0%)
T_2, INQUERY Boolean		20000	2266	0.2461 (-15.0%)
T_3		16916	1289	0.1276 (-55.9%)
T_4		20000	2354	0.2605 (-9.9%)
T_5		15600	494	0.0448 (-84.5%)
T_6	15.0 : 15.0	20000	2365	0.2610 (-9.8%)
T_7	6.0 : 6.0	20000	2487	0.2712 (-6.3%)
T_8	2.5 : 2.5	20000	2435	0.2709 (-6.4%)
T_9	1.0 : 1.0	20000	2266	0.2461 (-15.0%)
T_{10}	-1.0 : -1.0	20000	2266	0.2461 (-15.0%)
A_1	0.15 : 0.85	20000	2651	0.2962 (+2.4%)
A_2, Waller-Kraft	0.35 : 0.65	20000	2363	0.2479 (-14.3%)
A_3, network Boolean	0.15 : 0.85	20000	2247	0.2500 (-13.6%)
A_4, infinite-one	0.2 : 0.2	20000	2569	0.2948 (+1.9%)
Paice	1.0 : 1.0	20000	2540	0.2832 (-2.1%)

- INQUERY-weights generates smaller range of document term weights than Fox-weights. The smaller range can slightly alleviate the problems of single and double operand dependency. This may be the reason why INQUERY-weights gives better retrieval effectiveness to the fuzzy set and Waller-Kraft models than Fox-weights.

- The network Boolean model is part of the inference network-based retrieval model that is the basis of the INQUERY system. However, the INQUERY system has used the INQUERY Boolean operators, i.e. product and algebraic sum rather than the network Boolean operators. When Fox-weights is used for document term weighting, the effectiveness of network Boolean is much better than that of INQUERY Boolean. This is because INQUERY Boolean causes the single operand dependency problem that is a bit critical to retrieval effectiveness. On the other hand, if we use INQUERY-weights for document term weighting, the INQUERY Boolean operators do not have the single operand dependency property. Table 8 shows that INQUERY Boolean provides slightly worse retrieval effectiveness than network Boolean. This may be due to the fact that network Boolean has the positively compensatory property in some operand values.

- The infinite-one and Paice models provide a little comparable retrieval effectiveness to the p-norm model in the both weighting schemes. It should be noticed that infinite-one, Paice and p-norm can be reduced to the vector space model in that they provide the same results that the vector space model when the parameter values for AND and OR operators are as follows: 0 and 0 for infinite-one, 1.0 and 1.0 for Paice, and 1.0 and 1.0 for p-norm. It has been known that the effectiveness of the p-norm model is slightly better than the vector space model for the well-formulated Boolean query.

6 Conclusion

A major role of IR systems present to users a sequence of documents ranked in decreasing order of query-document similarities. Users are able to minimize their time spent to find useful information by reading the top-ranked documents first. Although conventional Boolean retrieval systems based on inverted files have the advantages of easy implementation and efficient retrieval, they cannot rank the documents retrieved. Extended Boolean models such as fuzzy set, p-norm et al. have been developed to provide ranked output in the Boolean retrieval systems.

In this paper, we have reviewed a variety of operators that can evaluate AND and OR operations in extended Boolean models, and have analyzed their properties to affect retrieval effectiveness. The following conclusions may be drawn from the analyses and experimental evidence described in this study.

- The partial operand dependent property, which exploits some specific operands for the similarity calculation, decrease retrieval effectiveness in some circumstances. For instance, the fuzzy set and Waller-Kraft models have the single operand dependent property and double operand dependent property, respectively.

- The negatively compensatory property, which generates a resulting value less than the minimum operand value or greater than the maximum, decrease retrieval effectiveness in some circumstances. For instance, the INQUERY Boolean and network Boolean models have this property in some range of operand values.

- The unequal importance property, which reflects all operand values on document values but gives different importance to each operand, may decrease retrieval effectiveness. For instance, the Paice and infinite-one models have the unequal importance property.

- The properties of positive compensation and equal importance may help retrieval effectiveness by preventing the problems of partial operand dependency, negative compensation and unequal importance.

- The operators of the p-norm model have the positive properties of positive compensation and equal importance, but do not have the negative properties of partial operand dependency, negative compensation and unequal importance. Hence, the p-norm model may provide better retrieval effectiveness than the others.

In developing extended Boolean models, AND operations have been used to identify query phrases as in "fuzzy AND set AND theory", and similarly, OR operations have been used to relate synonymous terms as in "human OR people OR person OR man". This simple interpretation of AND and OR operations may make sense in the linguistic viewpoint, and many users may also use AND and OR operations to identify phrases and synonyms, respectively, in some circumstances. However, many conventional Boolean retrieval systems allow user queries to be formulated with proximity and synonym operators. For example, in the STAIRS system, ADJ specifies that the two terms must be adjacent to one another, WITH indicates that the two terms must appear in the same sentence, SAME specifies that the two terms must appear in the same paragraph, and SYN specifies that the two terms are to be considered as synonyms. ADJ, WITH and SAME look more appropriate to specify phrases than AND, and SYN is an operator itself to specify synonyms. The meaning of the aforementioned operators should be investigated in more detail, and their evaluation formulas should also be developed to provide sophisticated ranking facility for the conventional Boolean retrieval systems.

Acknowledgments

I would like to thank Gerard Salton for reading earlier drafts of this paper. I would also like to give special thanks to Chris Buckley for his help in setting up experimental environment.

References

[1] M. Bartschi, "An overview of information retrieval subject," IEEE Computer, pp. 67-84, 1985.

[2] D.A. Buell, "A general model of query processing in information retrieval system," Information Processing & Management, Vol. 17, No. 5, pp. 249-262, 1981.

[3] T. Radecki, "Fuzzy set theoretical approach to document retrieval," Information Processing & Management, Vol. 15, No. 5, pp. 247-259, 1979.

[4] W.M. Sachs, "An approach to associative retrieval through the theory of fuzzy sets," Journal of the American Society for Information Science, Vol. 27, pp. 85-87, 1976.

[5] A. Bookstein, "Fuzzy requests: An approach to weighted Boolean searches," Journal of the American Society for Information Science, Vol. 31, No. 4, pp. 240-247, 1980.

[6] 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.

[7] 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.

[8] C.P. Paice, "Soft evaluation of Boolean search queries in information retrieval systems," Information Technology: Research and Development, Vol. 3, No. 1, pp. 33-42, 1984.

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

[10] J.H. Lee, M.H. Kim, and Y.J. Lee, "Information retrieval based on conceptual distance in is-a hierarchies," Journal of Documentation, Vol. 49, No. 2, pp. 188-207, 1993.

[11] M.E. Smith, "Aspects of the p-norm model of information retrieval: Syntactic query generation, efficiency, and theoretical properties," Ph.D. Thesis, Department of Computer Science, Cornell University, 1990.

[12] H. Turtle and W.B. Croft, "Evaluation of an inference network-based retrieval model," ACM Transactions on Information Systems, Vol. 9, No. 3, pp. 187-222, 1991.

[13] J.P. Callan, W.B. Croft and S.M. Harding, "The INQUERY retrieval system," Proceedings of the 3rd International Conference on Database and Expert Systems Applications, pp. 78-83, Springer-Verlag, 1992.

[14] L.A. Zadeh, "Fuzzy sets," Information and Control, Vol. 8, pp. 338-353, 1965.

[15] G.J. Klir and T.A. Folger, Fuzzy Sets, Uncertainty, and Information, Prentice-Hall, 1992.

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

[17] M.M. Gupta and J. Oi, "Theory of T-norms and fuzzy inference methods," Fuzzy Sets and Systems, Vol. 40, No. 3, pp.431-450, 1991.

[18] B. Werners, "Aggregation model in mathematical programming," in Mathematical Models for Decision Support, G. Mitra ed., pp. 295-319, 1988.

[19] H.J. Zimmermann and P. Zyso, "Latent connectives in human decision making," Fuzzy Sets and Systems, Vol. 4, No. 1, pp. 37-51, 1980.

[20] J.H. Lee, "Properties of extended Boolean models in information retrieval," Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 182-190, 1994.

[21] M.H. Kim and J.H. Lee and Y.J Lee, "Analysis of fuzzy operators for high quality information retrieval," Information Processing Letters, Vol. 46, No. 5, pp. 251-256, 1993.

[22] J.H. Lee, W.Y. Kim, M.H. Kim and Y.J. Lee, "On the evaluation of Boolean operators in the extended Boolean retrieval framework," Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Pittsburgh, PA USA, pp. 291-297, 1993.

[23] D. Harman, "Overview of the second text retrieval conference," Proceedings of the 2nd Text REtrieval Conference (TREC-2), National Institute of Standards and Technology Special Publication 500-215, pp. 243-252, 1994.

[24] E.A. Fox and J.A. Shaw, "Combination of multiple searches," Proceedings of the 2nd Text REtrieval Conference (TREC-2), National Institute of Standards and Technology Special Publication 500-215, pp. 243-252, 1994.

[25] E.A. Fox, S. Betrabet, M. Koushik and W. Lee, "Extended Boolean Models," in Information Retrieval: Data Structures & Algorithms, W.B. Frakes and R.B. Yates ed., Prentice Hall, pp. 393-418, 1992.

[26] E.A. Fox, "Characterization of two new experimental collections in computer and information science containing textual and bibliographic concepts," Technical Report 83-561, Dept. of Computer Science, Cornell Univ., Ithaca, N.Y., Sept. 1983.

[27] G. Salton and C. Buckley, "Term weighting approaches in automatic text retrieval," Information Processing and Management, Vol. 24, No. 5, pp. 513-523, 1988.

[28] J.H. Lee, "Combining multiple evidence from different properties of weighting schemes," submitted to the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1995.

[29] E.A. Fox, "Extending the Boolean and vector space models of information retrieval with p-norm queries and multiple concept types," Ph.D. Thesis, Department of Computer Science, Cornell Univ., 1983.