정보검색에서 부울연산자를 연산하는 식의 수학적 특성

Mathematical Properties of the Formulas Evaluating Boolean Operators in Information Retrieval

이 준호 (Joon-Ho Lee)*, 이 기호 (Kyi-Ho Lee)**, 조 영화 (Young-Hwa Cho)***

□목 차□

초 록

부울 검색 시스템은 구현이 용이하고 빠른 검색 시간을 제공하기 때문에, 오늘날 정보 검색 분야에서 가장 널리 사용되고 있다. 그러나 순수한 부울 검색 시스템은 문서값을 계산할 수 없기 때문에, 검색된 문서들을 질의를 만족하는 정도에 따라 정렬할 수 없다. 부울 검색 시스템에 순위 결정 기능을 부여하기 위하여 퍼지 집합, Waller-Kraft, Paice, P-Norm, Infinite-One과 같은 확장된 부울 모델들이 개발되어 왔다. 이들 모델에서 부울 연산자 AND와 OR에 대한 계산식은 순위 결정의 성능을 결정하는 중요한 요소이다. 본 논문에서는 부울 연산자 계산식의 수학적 특성을 제시하고, 이들이 검색 효과에 미치는 영향을 분석한다. 분석 결과는 P-Norm 모델이 높은 검색 효과를 얻기에 가장 적합함을 보여준다.

ABSTRACT

Boolean retrieval systems have been most widely used in the area of information retrieval due to easy implementation and efficient retrieval. Conventional Boolean retrieval systems, however, cannot rank retrieved documents in decreasing order of query-document similarities because they cannot compute similarity coefficients between queries and documents. Extended Boolean models such as fuzzy set, Waller-Kraft, Paice, P-Norm and Infinite-One have been developed to provide the document ranking facility. In extended Boolean models, the formulas evaluating Boolean operators AND and OR are an important component to affect the quality of document ranking. In this paper we present mathematical properties of the formulas, and analyse their effect on retrieval effectiveness. Our analyses show that P-Norm is the most suitable for achieving high retrieval effectiveness.

*연구개발정보센터 정보시스템개발실 선임연구원
**연구개발정보센터 정보유통실 실장
***연구개발정보센터 정보시스템부 부장

■논문접수일 : 1995년 4월 24일

I. 서 론

부울 검색 시스템(Boolean Retrieval System)은 구현이 용이하고 빠른 검색 시간을 제공하기 때문에, 상용화된 정보 검색 시스템의 대부분을 차지하고 있다. 부울 검색 시스템에서 문서는 색인어들의 집합으로 표현되며, 질의는 색인어와 논리연산자 AND, OR, NOT으로 이루어진 부울 수식이다. 시스템은 질의로 주어진 부울 수식을 만족하는 문서들을 사용자에게 제공한다.

기존의 부울 검색 시스템은 질의와 문서 사이의 유사도를 의미하는 문서값(Document Value)을 계산할 수 없기 때문에, 질의를 만족하는 정도에 따라 검색된 문서들에 순위를 부여할 수 없다. 이러한 순위 결정 기능은 검색된 문서들의 양에 대한 통제를 가능하게 하며, 적합성 피드백(Relevance Feedback)에 의한 질의 수정을 용이하게 한다. 부울 검색 시스템에 순위 결정 기능을 부여하기 위하여, 퍼지 집합(Buell 1981; Radecki 1979; Sachs 1976), Waller-Kraft(Waller & Kraft 1979), Paice(Paice 1984), P-Norm (Salton 1983), Infinite-One(Smith 1990)과 같은 확장된 부울 모델들이 개발되어 왔다. 이러한 모델들은 문서값을 계산하기 위하여, 문서내에서 색인어의 중요성을 반영하는 색인어 가중치를 이용하는 공통된 특성을 지니고 있다.

퍼지 집합 모델이 문서값을 계산함으로써 부울 검색 시스템의 단점을 극복하였을 지라도, 많은 경우에 부정확한 문서값을 생성하는 단점을 지니고 있다. 이것은 퍼지 집합 모델이 AND와 OR 연산을 위하여 사용하는 MIN과 MAX 연산자가 문서값 계산에 대한 인간의 사고 방식과 부합하지 못하기 때문이다(Bookstein 1980; Lee et al. 1994). Waller- Kraft, Paice, P-Norm, Infinite-One 모델은 MIN과 MAX 연산자의 문제점을 발생시키지 않는 연산자를 사용함으로써 퍼지 집합 모델의 단점을 극복하였다. 이들 모델들의 연산자는 다항 연산자로서, 이는 다항 연산자의 이항 형태가 결합 법칙을 만족하지 못하기 때문이다.

퍼지 집합 이론에서 다양한 퍼지 연산자들이 AND와 OR 연산을 위해 제안되어 왔다(Zimmermann 1991). 이들 다양한 퍼지 연산자들의 특성을 분석함으로써 높은 검색 효과(Retrieval Effectiveness)를 제공하는 긍정적 보상 연산자(Positively Compensatory Operator)라 불리는 이항 연산자 집합이 정의되었다(Kim et al. 1993; Lee et al. 1992; Lee et al. 1993). 본 논문에서는 확장된 부울 모델에서 AND와 OR 연산자를 위해 사용된 연산자들의 수학적 특성을 분석하고, Waller-Kraft, Paice, P-Norm, Infinite-One 모델로부터 유도된 이항 연산자들이 긍정적 보상 연산자에 포함됨을 입증한다. 또한 다항 연산자들이 검색 효과에 미치는 영향을 분석하고, 불균등 중요성(Unequal Impor-tance)이라 불리는 Waller-Kraft, Paice, Infinite-One 모델의 문제점을 기술한다. 연산자들에 대한 분석을 기초로 하여 높은 검색 효과를 제공하는 다항 유연한 부울 연산자(N-ary Soft Boolean Operator)라 불리는 연산자 집합을 정의하고, P-Norm 모델의 연산자가 다항 유연한 부울 연산자에 포함됨을 보인다.

본 논문의 구성은 다음과 같다. 2장에서 확

장된 부울 모델에 대하여 설명한다. 3장과 4장에서 기존의 확장된 부울 모델의 이항 연산자와 다항 연산자를 분석하고, 높은 검색 효과를 얻기 위하여 연산자들이 지녀야할 수학적 특성들을 기술한다. 마지막으로 5장에서 결론을 맺는다.

II. 확장된 부울 모델

퍼지 집합, Waller-Kraft, Paice, P-Norm, Infinite-One과 같은 확장된 부울 모델들은 기존의 부울 검색 시스템에 순위 결정 기능을 부여하기 위하여 개발되어 왔다(Sachs, 1976; Smith 1990; Fox et al. 1992). 이들은 문서 내에서 색인어의 중요성을 반영하는 색인어 가중치를 이용하는 공통된 특성을 지니고 있다. 확장된 부울 모델을 기반으로 하는 정보 검색 시스템은 다음의 $<T, Q, D, F>$에 의해 정의된다.

(1) T는 질의와 문서를 표현하기 위해 사용되는 색인어들의 집합이다.

(2) Q는 시스템이 인식할 수 있는 질의들의 집합이다. Q에 속하는 각각의 질의 q는 색인어들과 논리 연산자 AND, OR, NOT으로 구성된 부울 수식이다.

(3) D는 문서들의 집합이다. D에 속하는 각각의 문서 d는 w_i가 색인어 t_i의 가중치일 때, $\{(t_1, w_1), \cdots, (t_n, w_n)\}$와 같이 표현된다. 색인어 가중치 w_i는 0부터 1사이의 값을 갖는다.

$$F(d, t_1 \ AND \ t_2)=MIN(w_1, w_2)$$
$$F(d, t_1 \ OR \ t_2)=MAX(w_1, w_2)$$

(a) The fuzzy set model

$$F(d, t_2 \ AND \cdots AND \ t_2)=(1-r) \cdot MIN(w_1, \cdots, w_n)+r \cdot MAX(w_1, \cdots, w_n), \quad 0 \le r \le 0.5$$
$$F(d, t1 \ OR \cdots OR \ t_n)=(1-r) \cdot MIN(w_1, \cdots, w_n)+r \cdot MAX(w_1, \cdots, w_n), \quad 0.5 \le r \le 1$$

(b) The Waller-Kraft model

$$F(d, t_1 \ AND \cdots AND \ t_n)=\frac{\sum_{i=1}^{n}(r^{i-1} \cdot w_i)}{\sum_{i=1}^{n} r^{i-1}} \quad 0 \le r \le 1 \text{ and wi 's are considered in ascending order}$$
$$F(d, t_1 \ OR \cdots OR \ t_n) =\frac{\sum_{i=1}^{n}(r^{i-1} \cdot w_i)}{\sum_{i=1}^{n} r^{i-1}} \quad 0 \le r \le 1 \text{ and wi 's are considered in descending order}$$

(c) The Paice model

$$F(d, t_1 \ AND \cdots AND \ t_n)=1-\left(\frac{\sum_{i=1}^{n}(1-w_i)^p}{n}\right)^{1/p} \quad 1 \le p \le \infty$$
$$F(d, t_1 \ OR \cdots OR \ t_n)= \left(\frac{\sum_{i=1}^{n} w_i^p}{n}\right)^{1/p} 1 \le p \le \infty$$

(d) The P-Norm model

$$F(d, t_1 \ AND \cdots AND \ t_n)=r \cdot (1-MAX(1-w_1, \cdots, 1-w_n))+(1-r) \frac{\sum_{i=1}^{n} w_i}{n} \quad 0 \le r \le 1$$
$$F(d, t_1 \ OR \cdots OR \ t_n)=r \cdot MAX(1-w_1, \cdots, 1-w_n)+(1-r) \cdot \frac{\sum_{i=1}^{n} w_i}{n} \quad 0 \le r \le 1$$

(e) The Infinite-One model

〈그림 1〉 AND와 OR 연산에 대한 연산식

(4) F는 문서값을 계산하는 순위 결정 함수(Ranking Function)로서 다음과 같이 정의된다.

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

함수 F는 각 쌍의 (d, q)에 0부터 1사이의 값을 부여한다. 이 값은 문서 d와 질의 q사이의 유사도이며,질의 q에 대한 문서 d의 문서값이라고 일컬어진다. *AND*와 *OR*에 대한 연산자 계산식은 순위 결정 함수의 성능을 결정하는 가장 중요한 요소이다. 〈그림 1〉 은 확장된 부울 모델에서 *AND*와 *OR* 연산을 위해 사용된 연산자들을 보여준다.

퍼지 집합 모델의 연산자 계산식은 단지 2개의 피연산자를 갖는 이항 연산자이고, Waller-Kraft, Paice, P-Norm, Infinite-One 모델의 연산자 계산식은 2개 이상의 피연산자를 갖는 다항 연산자이다. 이것은 *MIN*과 *MAX*연산자가 결합법칙을 만족하는데 비하여, Waller-Kraft, Paice, P-Norm, Infinite-One 모델의 이항 연산자는 결합법칙을 만족하지 못하기 때문이다. 결합법칙을 만족하지 못하는 이항 연산자의 사용은 2개의 논리적으로 동등한 질의 $(t_1$ *AND* $t_2)$ *AND* t_3와 t_1 *AND* $(t_2$ *AND* $t_3)$에 대해 서로 다른 문서값을 생성한다. 이러한 문제는 4장에서 자세히 설명될 것이다.

III. 이항 연산자의 분석

3.1 검색 효과를 저하시키는 특성들

퍼지 집합 모델은 문서들의 순위를 결정하는 문서값을 계산함으로써 부울 검색 시스템의 단점을 극복하였을 지라도, 부정확한 문서값을 생성하는 요인을 지니고 있기 때문에 정보 검색 모델로서 부적합하다고 비판되어 왔다(Bookstein 1980; Lee et al. 1994). 이것은 퍼지 집합 모델이 *AND*와 *OR* 연산을 위하여 사용하는 *MIN*과 *MAX* 연산자가 검색효과를 저하시키는 특성을 지니고 있기 때문이다. 퍼지 집합 이론이 개발된 이후로 *MIN*과 *MAX* 연산자를 대신할 수 있는 다양한 퍼지 연산자들이 제안되어 왔다. 이들 퍼지 연산자들이 확장된 부울 모델에서 *AND*와 *OR* 연산을 위해 사용될 때, 단일 피연산자 의존 특성(Single Operand Dependency Property)과 부정적 보상 특성(Negative Compensa-tion Property)을 지니는 연산자는 검색 효과를 저하시킴이 입증되었다(Kim et al. 1993; Lee et al. 1992; Lee et al. 1993).

단일 피연산자 의존 문제 : 임의의 연산자 θ가 단일 피연산자 의존 특성을 갖는다면, $\theta(x, y)=x$ 또는 y $(x, y \in [0, 1], x \neq y)$이다. *AND*와 *OR* 연산을 위하여 단일 피연산자 의존 특성을 갖는 연산자를 사용하는 확장된 부울 모델은 단일 피연산자 의존 문제를 발생시킨다. 예를 들면, 두개의 문서 d_1, d_2와 질의 q_1이 다음과 같이 주어졌다고 가정하자.

d_1 = {(Information, 0), (Retrieval, 0)}
d_2 = {(Information, 1), (Retrieval, 0)}
q_1 = Information AND Retrieval

곱하기 연산자가 *AND* 연산을 위해 사용될 때, 질의 q_1에 대한 문서 d_1, d_2의 문서값은 모두 0으로 동일하다. 그러나 대부분의 사람들은 d_1보다 d_2가 질의 q_1에 유사하다고 결정할 것이

다. 이러한 부정적 결과는 곱하기 연산자가 단일 피연산자 의존 특성을 지니고 있기 때문이다. 즉, $x \cdot 0 = 0$ ($x \in (0, 1]$).

부정적 보상 문제 : 임의의 연산자 θ가 부정적 보상 특성을 갖는다면, $\theta(x, y) < MIN(x, y)$ 또는 $\theta(x, y) > MAX(x, y)$ ($x, y \in [0,1]$)이다. *AND*와 *OR* 연산을 위하여 부정적 보상 특성을 갖는 연산자를 사용하는 확장된 부울 모델은 부정적 보상 문제를 발생시킨다. 예를 들면, 문서 d_3와 두 개의 질의 q_2, q_3가 다음과 같이 주어졌다고 가정하자.

d_3 = {(Information, 0.70), (Retrieval, 0.70)}

q_2 = Information *AND* Retrieval

q_3 = Information

곱하기 연산자가 *AND* 연산을 위해 사용될 때, 질의 q_2와 q_3에 대한 d_3의 문서값은 각각 0.49와 0.70이다. 즉, q_2와 d_3사이의 유사도가 q_3와 d_3사이의 유사도보다 작다. 그러나 이것은 사람들의 순위 결정에 대한 행동 방식과 상반되는 결과로서, 곱하기 연산자가 단일 피연산자 의존 특성뿐 아니라 부정적 보상 특성도 지니고 있기 때문이다. 즉, $x \cdot y < MIN(x, y)$ ($x \in (0,1)$).

3.2 긍정적 보상 연산자

퍼지 연산자들이 검색 효과에 미치는 영향을 분석함으로써, 높은 검색 효과를 제공할 수 있는 긍정적 보상 연산자라 불리는 이항 연산자 집합이 정의되었다(Kim et al. 1993; Lee et al. 1992; Lee et al. 1993). 높은 검색 효과를 제공하는 긍정적 보상 연산자는 다음과 같이 정의된다.

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

임의의 긍정적 보상 연산자 p는 다음과 같은 특성을 갖는다.

특성 p_1 : $p(x, x) = x$; 즉, p는 idempotent 이다.

특성 p_2 : $MIN(x, y) < p(x, y) < MAX(x, y)$, $x \neq y$

긍정적 보상 연산자에 대한 위의 정의로부터 긍정적 보상 연산자는 단일 피연산자 의존 특성과 부정적 보상 특성 모두를 지니고 있지 않음을 알 수 있다. 따라서 긍정적 보상 연산자를 사용하는 확장된 부울 모델은 단일 피연산자 의존 문제와 부정적 보상 문제를 발생시키지 않는다.

퍼지 집합 이론에서 개발된 퍼지 연산자들로부터 다음과 같은 2개의 긍정적 보상 연산자를 발견하였다.

$$(A_2) \quad (1-r) \cdot MIN(x, y) + r \cdot MIN(x, y), \quad 0 \leq r \leq 1$$

$$(A_{4,\,AND}) \quad r \cdot MIN(x, y) + (1-r) \cdot \frac{x+y}{2} \quad 0 \leq r \leq 1$$

$$(A_{4,\,OR}) \quad r \cdot MIN(x, y) + (1-r) \cdot \frac{x+y}{2} \quad 0 \leq r \leq 1$$

A_2와 A_4 연산자는 각각 다른 사람에 의해 다른 시점에 개발되었을 지라도, 수학적으로 동일한 식임이 증명되었다(Lee et al. 1993).

3.3 이항 유연한 부울 연산자

〈그림 2〉는 Waller-Kraft, Paice, P-Norm, Infinite-One 모델로부터 유도된 이항 연산자를 보여준다. 이들은 다음과 같은 형태

의 함수이다.

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

이항 연산자 b는 다음과 같은 특성을 갖는다.

특성 b_1 : $b(x, x) = x$; 즉, b는 idempotent이다.

특성 b_2 : $x < x'$이면 $b(x, y) < b(x', y)$이고, $y < y'$이면 $b(x, y) < b(x, y')$이다. 즉, b는 각각의 피연산자에 대하여 절대 단조성(Strict Monotonicity)을 갖는다.

특성 b_3 : b는 연속함수이다.

특성 b_4 : $b(x, y) = b(y, x)$; 즉, b는 교환법칙을 만족한다.

특성 b_1은 3개의 질의 $q_1 = t_1$ *AND* t_1, $q_2 = t_1$ *OR* t_1, $q_3 = t_1$이 임의의 문서에 대해 동일한 문서값을 지정함을 의미한다. 특성 b_2에 의해 색인어 가중치의 증가로써 문서값의 증가를 보장할 수 있다. 특성 b_3의 연속성의 특성은 색인어 가중치의 미세한 증가가 문서값의 큰 변화를 발생시키는 상황을 막는다. 교환법칙의 특성에 의해 질의 t_1 *AND* t_2와 t_2 *AND* t_1은 임의의 문서에 대해 동일한 문서값을 지정한다. 특성 b_1부터 b_4를 만족하는 연산자는 이항 유연한 부울 연산자(Binary Soft Boolean Operator)로 정의된다.

정리 1: 임의의 연산자 θ가 특성 b_1과 b_2를 만족한다면, 서로 다른 $x, y \in [0,1]$에 대해 $MIN(x, y) < \theta(x, y) < MAX(x, y)$이다.

증명 : 절대 단조성에 대해 다음의 식이 성립한다.

$\theta(MIN(x, y), MIN(x, y)) < \theta(x, y) < \theta(MAX(x, y), MAX(x, y))$

$$F(d, t_1 \text{ AND } t_2) = (1-r) \cdot MIN(w_1, w_2) + r \cdot MAX(w_1, w_2), \qquad 0 \le r \le 0.5$$
$$F(d, t_1 \text{ OR } t_2) = (1-r) \cdot MIN(w_1, w_2) + r \cdot MAX(w_1, w_2), \qquad 0.5 \le r \le 1$$

(a) The Waller-Kraft model

$$F(d, t_1 \text{ AND } t_2) = \frac{1}{1+r} \cdot MIN(w_1, w_2) + \frac{r}{1+r} \cdot MAX(w_1, w_2),$$
$0 \le r \le 1$ and wi's are considered in ascending order

$$F(d, t_1 \text{ OR } t_2) = \frac{1}{1+r} \cdot MIN(w_1, w_2) + \frac{r}{1+r} \cdot MAX(w_1, w_2),$$
$0 \le r \le 1$ and wi's are considered in descending order

(b) The Paice model

$$F(d, t_1 \text{ AND } t_2) = 1 - \left(\frac{(1-w_1)^p + (1-w_2)^p}{2}\right)^{1/p}, \qquad 1 \le p \le \infty$$
$$F(d, t_1 \text{ OR } t_2) = \left(\frac{w_1^p + w_2^p}{2}\right)^{1/p}, \qquad 1 \le p \le \infty$$

(c) The P-Norm model

$$F(d, t_1 \text{ AND } t_2) = r \cdot (1 - MAX(1-w_1, 1-w_2)) + (1-r) \cdot \frac{w_1 + w_2}{2}, \qquad 0 \le r \le 1$$
$$F(d, t_1 \text{ OR } t_2) = r \cdot MAX(1-w_1, 1-w_2) + (1-r) \cdot \frac{w_1 + w_2}{2}, \qquad 0 \le r \le 1$$

(d) The Infinite-One model

〈그림 2〉 AND와 OR 연산에 대한 이항 연산식

또한 idempotency에 의한 다음의 두 식이 성립한다.

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

$(MAX(x, y), MAX(x, y)) = MAX(x, y)$

따라서, $MIN(x, y) \langle \theta(x, y) \langle MAX(x, y)$가 성립한다.

정리 1로부터 이항 유연한 부울 연산자가 긍정적 보상 연산자에 속함을 알 수 있다. 따라서 Waller-Kraft, Paice, P-Norm, Infinite-One 모델의 이항 연산자들은 높은 검색 효과를 제공하는 긍정적 보상 연산자이다. A_2와 A_4 연산자도 특성 b_1부터 b_4를 만족하는 이항 유연한 부울 연산자이다. A_2, A_4와 Waller-Kraft, Paice, Infinite-One 모델의 이항 연산자는 수학적으로 동일한 식임을 쉽게 입증할 수 있다.

IV. 다항 연산자의 분석

4.1 다항 연산자의 필요성

AND 연산은 단어들을 연결하여 구를 생성하고, *OR* 연산은 두개의 색인어들을 동의어로 간주한다. 따라서 fuzzy *AND* set *AND* theory와 같이 연속적인 *AND* 연산이나, human *OR* people *OR* man과 같이 연속적인 *OR* 연산이 자주 발생한다. 그러나 이항 유연한 부울 연산자는 결합 법칙을 만족하지 못한다.

정리 2: 임의의 연산자 θ가 특성 b_1과 b_2를 만족한다면, θ는 결합 법칙을 만족하지 못한다.

증명 : $x \langle y$를 만족하는 x, y를 가정하자. 정리 1에 의해 다음의 식이 성립한다.

$$\theta(x, y) \rangle x,$$

첫번째 파라미터에 절대 단조성을 적용한다면 다음의 식을 얻을 수 있다.

$$\theta(\theta(x, y), y) \rangle \theta(x, y),$$

그러나, θ가 결합 법칙을 만족한다면 다음의 식이 성립한다.

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

따라서 θ가 결합 법칙을 만족한다는 가정은 모순을 발생시킨다.

결합 법칙을 만족하지 않는 이항 연산자는 의미적으로 동일한 질의에 대하여 서로 다른 문서값을 생성한다. 예를 들면, 문서 d가 $\{(t_1, 1), (t_2, 0.7), (t_3, 0.5)\}$로 표현되어 있고, 두 개의 질의 $q_1 = t_1\ AND\ (t_2\ AND\ t_3)$와 $q_2 = (t_1\ AND\ t_2)\ AND\ t_3$가 주어졌다고 가정하자. $t_{4.AND}$ (r=0.3) 연산자를 사용하는 확장된 부울 모델은 q_1에 대한 d의 문서값으로 0.721을 생성하고, q_2에 대한 문서값으로 0.607을 생성한다.

질의 $t_1\ OR\ t_2\ OR\ t_3$에 나타난 색인어 t_1, t_2, t_3는 같은 중요도로 사용자가 요구하는 정보를 표현하고 있다. 그러나 이항 유연한 부울 연산자를 사용한다면, 질의 q를 연산하는데 이러한 기본적인 가정을 유지할 수 없다. 질의 q를 연산하는 2가지 방법 $(t_1\ OR\ t_2)\ OR\ t_3$와 $t_1\ OR\ (t_2\ OR\ t_3)$를 고려하자. $(t_1\ OR\ t_2)\ OR\ t_3$의 문서값은 색인어 t_1, t_2보다 t_3에 의존적이고, $t_1\ OR\ (t_2\ OR\ t_3)$의 문서값은 색인어 t_2, t_3 보다 t_1에 의존적이다.

문서 $d = \{(t_1, w_2), (t_2, w_2), \cdots, (t_n, w_n)\}$와 질의 $q = t_1\ OR\ t_2\ OR \cdots OR\ t_n$를 가정하자. 질의를 왼쪽에서 오른쪽으로 $A_{4,\ OR}$, $(\gamma = 0)$를 적용하면, 다음과 같은 문서 d의 문서값을 얻을 수 있다.

$$F(d, q) = \frac{w_1}{2^{n-1}} + \frac{w_2}{2^{n-1}} + \frac{w_3}{2^{n-2}} + \frac{w_4}{2^{n-3}} + \cdots + \frac{w_n}{2}$$
$$= \frac{1}{2}\left(\frac{w_1}{2^{n-2}} + \frac{w_2}{2^{n-2}} + \frac{w_3}{2^{n-3}} + \frac{w_4}{2^{n-4}} + \cdots + w_n\right)$$

위의 결과로부터 먼저 연산에 참여한 색인어와 나중에 연산에 참여한 색인어의 중요도가 다르게 취급되고 있음을 알 수 있다. 색인어 t_1과 t_n을 살펴보면, 질의에서는 같은 중요도로 질의를 표현하고 있다. 그러나 연산 결과에 있어서 색인어 t_1의 중요도를 1이라고 할 때 t_n의 중요도는 $2^{n-2}(n \geq 2)$ 이다. 결론적으로 먼저 연산에 참여한 색인어는 나중에 연산에 참여한 색인어보다 훨씬 작은 중요도로 취급되며, n의 값이 크면 클수록 이러한 왜곡의 정도는 더욱 커진다.

4.2 다항 유연한 부울 연산자

Waller-Kraft, Paice, P-Norm, Infinite-One 모델은 다항 연산을 가능하게 함으로써 결합 법칙을 만족하지 못하는 문제를 회피하였다. 그러나, Waller-Kraft, Paice, P-Norm, Infinite-One 모델은 어떤 경우에 사람이 생각하는 것과 다르게 문서의 순위를 결정한다. 이 모델들은 질의에 주어진 모든 색인어들이 문서값의 계산에 있어서 동등하게 고려되어야 한다는 일반적 가정을 위반하며, 이것은 다음과 같은 불균등 중요성 문제(Unequal Importance Problem)를 발생시킨다.

불균등 중요성 문제 (유형 1) : 문서값 계산에 있어서 Waller-Kraft 모델은 단지 2개의 피연산자, 최소값과 최대값만을 고려한다. 예를 들면, 두개의 문서 d_1, d_2와 질의 q_1이 다음과 같이 주어졌다고 가정하다.

$d_1 = \{(t_1, 0), (t_2, 0.9), (t_3, 0.9), \cdots, (t_{99}, 0.9), (t_{100}, 1)\}$

$d_2 = \{(t_1, 0), (t_2, 0.1), (t_3, 0.1), \cdots, (t_{99}, 0.1), (t_{100}, 1)\}$

$q_1 = t_1\ AND\ t_2\ AND \cdots AND\ t_{100}$

Waller-Kraft 모델은 질의 q_1에 대한 d_1, d_2의 문서값으로 동일한 값을 지정한다. 그러나 대부분의 사람들은 질의 q_1에 대하여 문서 d_1의 만족도가 문서 d_2의 만족도보다 큰 것으로 결정할 것이다.

불균등 중요성 문제 (유형 2) : Paice 모델은 모든 피연산자들을 문서값에 반영함으로써, 유형 1의 불균등 중요성 문제를 회피한다. 그러나 Paice 모델은 질의에 주어진 색인어들에 서로 다른 중요도를 부여한다. 예를 들면, 두개의 문서 d_3, d_4와 질의 q_2가 다음과 같이 주어졌다고 가정하자.

$d_3 = \{(t_1, 0.1), (t_2, 0.3), (t_3, 0.3), (t_4, 0.3), (t_5, 0.3), (t_6, 0.3)\}$

$d_4 = \{(t_1, 0.1), (t_2, 0.7), (t_3, 0.3), (t_4, 0.3), (t_5, 0.3), (t_6, 0.2)\}$

$q_2 = t_1\ AND\ t_2\ AND\ t_3\ AND\ t_4\ AND\ t_5\ AND\ t_6$

Paice 모델(r=0.7)은 질의 q_2에 대한 문서 d_3의 문서값으로 0.2119를 지정하고, 문서 d_4의

문서값으로 0.2310을 지정한다. 즉, t_6의 색인어 가중치의 적은 감소가 t_2의 색인어 가중치의 많은 증가에 대한 효과를 상쇄시키며, 이러한 특성은 문서값 계산에 대한 사람들의 행동 방식과 일치하지 않는다.

불균등 중요성 문제 (유형 2) : Infinite-One 모델도 모든 피연산자들을 문서값에 반영함으로써, 유형 1의 불균등 중요성 문제를 회피한다. 그러나 Infinite-One 모델이 생성한 문서값은 *AND* 연산에 대하여 최소값의 피연산자, *OR* 연산에 대하여 최대값의 피연산자에 의해 보다 많은 영향을 받는다. 예를 들면, 두개의 문서 d_5, d_6와 질의 q_3가 다음과 같이 주어졌다고 가정하자.

$d_5 = \{(t_1, 0), (t_2, 1), (t_3, 1), \cdots, (t_{100}, 1)\}$

$d_6 = \{(t_1, 0.4), (t_2, 0.6), (t_3, 0.6), \cdots, (t_{100}, 0.6)\}$

$q_3 = t_1\ AND\ t_2\ AND \cdots AND\ t_{100}$

Infinite-One 모델($r=0.5$)은 질의 q_3에 대한 문서 d_5와 d_6의 문서값이 동일한 것으로 결정한다. 즉, t_1의 색인어 가중치에 대한 0.4의 증가가 t_2부터 t_{100}까지의 99개 색인어 가중치에 대한 0.4의 감소와 동일한 효과를 갖는다. 그러나 사람들은 d_5의 문서값이 보다 큰 것으로 결정할 것이다.

P-Norm 모델은 문서값 계산에 있어서 모든 피연산자들을 동일한 중요도로 고려하기 때문에, 불균등 중요성 문제를 발생시키지 않는다. P-Norm 모델의 다항 연산자는 다음과 같은 형태의 함수이다.

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

다항 연산자 n은 다음과 같은 특성을 지니며, 특성 n_1부터 n_5를 만족하는 연산자는 다항 유연한 부울 연산자(N-ary Soft Boolean Operator)로 정의된다.

특성 n_1 : $n(x, x, \cdots, x) = x$; 즉, n은 idempotent이다.

특성 n_2 : n은 각각의 피연산자에 대하여 절대 단조성(Strict Monotonicity)을 갖는다.

특성 n_3 : n은 연속 함수이다.

특성 n_4 : $y_1, y_2, \cdots, y_n$이 $x_1, x_2, \cdots, x_n$에 대한 임의의 조합일 때, $n(x_1, x_2, \cdots, x_n) = n(y_1, y_2, \cdots, y_n)$; 즉 n은 symmetric 함수이다.

특성 n_5 : n은 모든 피연산자를 동일한 중요도로 고려한다.

긍정적 보상 연산자 A_2와 A_4는 이항 유연한 부울 연산자에 속하기 때문에, 결합 법칙을 만족하지 않는다. 이들 연산자의 다항 연산식이 퍼지 집합 이론에 대한 연구 분야에서 다음과 같이 제시되었다(Zimmermann, 1987).

$$(A_{2.n}) \quad r\cdot MAX(w_1, \cdots, w_n)+(1-r)\cdot MIN(w_1, \cdots, w_n), \quad 0\le r\le 1$$

$$(A_{4.N.AND}) \quad r\cdot MIN(w_1, \cdots, w_n)+(1-r)\cdot \frac{w_1+\cdots+w_n}{n} \quad 0\le r\le 1$$

$$(A_{4.N.OR}) \quad r\cdot MAX(w_1, \cdots, w_n)+(1-r)\cdot \frac{w_1+\cdots+w_n}{n} \quad 0\le r\le 1$$

이항 연산자 A_2와 A_4는 수학적으로 동일한 식일 지라도, 다항 연산자 $A_{2.N}$과 $A_{4.N}$은 서로 다른 식이다. $A_{2.N}$은 Waller-Kraft 모델의 연산자 계산식과 동일하며, $A_{2.N}$은 Infinite-One 모델의 연산자 계산식과 동일하다.

V. 결 론

정보 검색 시스템의 중요한 역할 중의 하나는 검색된 문서들에 대하여 순위 결정 방법을 적용함으로써 문서가 질의를 만족하는 정도를 나타내는 문서값을 계산하고, 계산된 문서값에 따라 문서들에 순위를 부여하는 것이다. 높은 순위를 갖는 문서일수록 질의에 대한 만족도가 크며, 사용자는 높은 순위를 갖는 문서를 우선적으로 검토함으로써 필요한 정보를 얻는데 소모되는 시간을 최소화할 수 있다. 역화일을 기반으로 하는 기존의 부울 검색 시스템은 빠른 검색 시간을 제공하지만, 검색된 문서들에 대하여 문서값을 계산할 수 없는 단점을 지니고 있다. 이러한 부울 검색 시스템의 단점을 보완하기 위하여, 퍼지 집합, Waller-Kraft, Paice, P-Norm, Infinite-One과 같은 확장된 부울 모델들이 개발되어 왔다.

본 논문에서는 기존의 확장된 부울 모델에서 *AND*와 *OR* 연산을 위해 사용된 연산자들의 수학적 특성을 분석하여, 검색 효과에 영향을 미치는 중요한 쟁점들에 대하여 기술하였다. Waller-Kraft, Paice, P-Norm, Infinite-One 모델로부터 유도된 이항 연산자들은 결합법칙을 만족하지 못할 지라도, 높은 검색 효과를 제공하는 것으로 알려진 긍정적 보상 연산자에 속함을 증명하였다. 그러나 Waller-Kraft, Paice, Infinite-One 모델이 지니고 있는 불균등 중요성 문제는 검색 효과를 저하시킬 수 있다. 본 논문에서는 P-Norm 모델의 연산자가 불균등 중요성 문제를 발생시키지 않으며, 높은 검색 효과를 얻기에 적합한 다항 유연한 부울 연산자라는 연산자 집합에 속함을 입증하였다.

참 고 문 헌

Bookstein, A. (1980). Fuzzy requests: an approach to weighted Boolean searches. Journal of the American Society for Information Science, 31(4), 240-247.

Buell, D.A. (1981). A general model of query processing in information retrieval system. Information Processing & Management, 17(5), 249-262.

Fox, E.A., Betrabet, S., Koushik, M., & Lee, W. (1992). Extended Boolean models. In: Frakes, W.B., Yates, R.B. (ed) Information Retrieval Data Structures & Algorithms. Prentice Hall, 393-418.

Kim, M.H., Lee, J.H., & Lee, Y.J. (1993). Analysis of fuzzy operators for high quality information retrieval. Information Processing Letters, 46(5), 251-256.

Lee, J.H., Kim, M.H., & Lee, Y.J. (1992). Enhancing the fuzzy set model for high quality document rankings. In: Proceedings of the 19th Euromicro Conference, 337-344.

Lee, J.H., Kim, W.Y., Kim, M.H., and Lee, Y.J. (1993). On the evaluation of Boolean operators in the extended Boolean retrieval framework. In: Proceedings of the 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 291-297.

Lee, J.H., Kim, M.H., & Lee, Y.J. (1994). Ranking documents in thesaurus-based Boolean retrieval systems. Information Processing & Management, 30(1), 79-91.

Paice, C.P. (1984). Soft evaluation of Boolean search queries in information retrieval systems. Information Technology: Research and Development, 3(1), 33-42.

Radecki, T. (1979). Fuzzy set theoretical approach to document retrieval. Information Processing & Management, 15(5), 247-259.

Sachs, W.M. (1976). An approach to associative retrieval through the theory of fuzzy sets. Journal of the American Society for Information Science, 27, 85-87.

Salton, G, Fox, E.A., & Wu, H. (1983). Extended Boolean information retrieval. Communications of the ACM, 26(11), 1022-1036.

Smith, M.E. (1990). Aspects of the p-norm model of information retrieval: syntactic query generation, efficiency, and theoretical properties. PhD thesis, Cornell University.

Waller, W.G., & Kraft, D.H. (1979). A mathematical model of a weighted Boolean retrieval system. Information Processing & Management, 15, 235-245.

Zimmermann, H.J. (1987). Fuzzy sets, decision making, and expert systems. Kluwer Academic Publishers.

Zimmermann, H.J. (1991). Fuzzy set theory and its applications, 2nd edition. Kluwer Academic Publishers.