産業科學技術硏究所 論文集 第十卷 第二號
(J. of Ind. Sci. and Tech. Institute)
Vol. 10, No. 2, 1996. 12.

효율적인 정보검색을 위한 새로운 퍼지집합모델

유 재 수*, 이 석 희*, 최 길 성*, 조 기 형*, 안 성 윤**
* 충북대학교 정보통신공학과
** 목포대학교 전산통계학과

A New Fuzzy Set Model for Efficient Information Retrieval

Jae Soo Yoo*, Seok Hee Lee*, Kel Seong Choi*, Ki Hyung Cho*, Sung Yoon An**
* Dept. of Computer & Communication, Chungbuk National Univ.
** Dept. of Computer Science & Statistics, Mokpo National Univ.

Abstract

After the first introduction of fuzzy set theory with MIN and MAX operators, a variety of fuzzy operators have been developed for the AND and OR operations. They are classified into three groups such as averaging operators, T-norms and T-conorms. T-operators(T-norms and T-conorms) and the averaging operators have been used to model human decisions in many cases. However, it has been noted that neither T-operators nor some of the averaging operators are appropriate to model managerial decisions. In this paper, we propose a new averaging operator that overcomes the problem and improves the retrieval effectiveness of the fuzzy set model. We show through performance evaluation that the proposed new averaging operator provides better retrieval effectiveness and efficiency than the existing operators.

Ⅰ. 서론

정보 검색 시스템(Information Retrieval System)은 질의를 만족하는 문서들을 사용자에게 제공한다. 그러나 정보 검색 시스템의 주요한 목적은 단순히 질의를 만족하는 문서들을 검색하는 것이 아니라, 검색된 각각의 문서에 대하여 순위 결정 방법(Ranking Method)을 적용함으로써 문서가 질의를 만족하는 정도를 나타내는 문서값(Document Value)을 계산하고, 계산된 문서값에 따라 문서들에 순위를 부여하는 것이다. 높은 순위를 갖는 문서일수록 질의에 대한 만족도

가 크며, 사용자는 높은 순위를 갖는 문서를 우선적으로 검토함으로써 필요한 정보를 얻는데 소모되는 시간을 최소화할 수 있다.

불리안 검색 시스템(Boolean Retrieval System)은 역화일(Inverted File)을 기반으로 하여 짧은 검색 시간을 제공하고, 불리안 연산자들을 사용함으로써 비교적 쉽고 정확하게 질의를 표현할 수 있기 때문에 오늘날 가장 널리 사용되는 시스템이다. 그러나 불리안 검색 시스템은 순위 결정 기능을 제공하지 않기 때문에 검색된 문서들을 질의에 만족하는 정도에 따라 정렬할 수 있다. 불리안 검색 시스템의 이러한 단점을 개선하기 위해 색인어 가중치를 이용하는 퍼지 집합 모델(Fuzzy Set Model)[1]과 확장된 불리안 모델(Extended Boolean Model)[2]이 제안되었다. 색인어 가중치는 문서에서 색인어가 가지는 중요도, 즉, 색인어가 문서에 관련된 정도를 표현한다.

본 논문에서는 MIN과 MAX 연산자를 사용하는 기존의 퍼지 집합 모델의 문제점들이 MIN과 MAX 연산자 대신에 T-연산자를 사용하더라도 해결될 수 없음을 설명한다. 또한 긍정적 보상 연산자(Positively Compensation Operator)라 불리는 일부의 평균 연산자들은 높은 검색 효율(Retrieval Effectiveness)을 제공할 수 있는 특성들을 지니고 있음을 기술하고, 긍정적 보상 연산자에 포함되는 평균 연산자를 불리안 연산자 계산식으로 사용할 것을 제안한다. 지금까지 개발된 긍정적 보상 연산자의 이항 연산식은 탐색어의 중요도를 왜곡하여 문서값을 계산하기 때문에 긍정적 보상 연산자의 이항 연산식을 다항 연산이 가능하도록 확장함으로써 문서 값의 왜곡을 감소시킬 수 있음을 설명한다. 마지막으로 실험을 통하여 긍정적 보상 연산자를 사용하는 검색 모델은 다른 연산자를 사용하는 검색 모델보다 좋은 검색 효율을 나타냄을 입증한다. 또한 본 논문에서 제안한 불리안 연산자 계산식은 좋은 검색 효율을 제공하는 것으로 알려진 확장된 불리안 모델의 불리안 연산자 계산식보다 문서값 계산 시간이 짧고 확장된 불리안 모델의 불리안 연산자 계산식과 유사한 검색 효율을 제공함을 보인다.

본 논문의 구성은 다음과 같다. 제Ⅱ장에서는 불리안 검색 모델을 개선하기 위한 기존의 퍼지 집합 모델을 기술한다. 제Ⅲ장에서는 퍼지 집합 이론에서 AND와 OR 연산을 위하여 지금까지 개발된 다양한 퍼지 연산자들이 정보 검색 시스템의 검색 효율에 미치는 영향을 설명하고, 이를 기반으로 새로운 퍼지연산자를 제안한다. 제Ⅳ장에서는 실험을 통하여 검색 효율을 평가하며, 끝으로 제Ⅴ장에서는 결론과 앞으로의 연구방향을 제시한다.

Ⅱ. 기존의 퍼지 집합모델

1. 퍼지 집합 모델

불리안 검색 모델은 불리안 연산자를 이용하여 사용자가 원하는 문서를 비교적 정확하게 표현할 수 있다. 불리안 검색 모델은 문서를 색인어의 집합으로 표현하고, 검색된 모든 문서들의 문서 값은 1이다. 또한 이 모델은 역화일(Inverted File)을 이용하여 빠른 검색 시간을 제공하지만 다음과 같은 단점이 있다. (1) 불리안 검색 시스템은 문서값을 0과 1의 두 값으로만 평가하므로 어떤 문서가 사용자가 가장 필요로 하는 문서인지를 결정할 수 없다. 또한 주어진 질의에 대하여 검색되는 문서의 양을 조절하기 어렵다. (2) 문서는 색인어의 집합으로 표현되며 문서를 표현하는 색인어는 각각 서로 다른 중요도를 가지고 문서를 표현하게 된다. 그런데 불리안 검색 시스템은 문서를 표현하는 모든 색인어를 동일한 중요도로 평가한다. 즉, 문서의 색인에 이용된 색인어는 가중치가 1이고 이용되지 않은 색인어는 가중치가 0이다. 불리안 검색 시스템의 이러한 단점을 보완하기 위하여 퍼지 집합 모델과 확장된 불리안 모델이 제안되었다.

퍼지 집합 모델은 색인어가 문서에서 갖는 중요도를 반영하는 색인어 가중치를 이용한다. 퍼지 집합 모델을 기반으로 하는 정보 검색 시스템은 다음에 설명되는 <T, Q, D, F>로 정의될 수 있으며 각각의 요소는 다음과 같은 의미를 가진다.

. T는 질의와 문서를 표현하기 위해 사용되는 색인어들의 집합이다.
. Q는 시스템이 인식할 수 있는 질의들의 집합이다. Q에 속하는 질의 q는 색인어들과 불리안 연산자 AND, OR 그리고 NOT로 구성된 불리안 수식이다.
. D는 문서들의 집합이다. D에 속하는 각각의 문서 d는 w_i가 색인어 t_i의 가중치일 때, $((t_1,w_1),.. .,(t_n,w_n))$와 같이 표현된다. 색인어 가중치 w_i는 0부터 1사이의 값을 갖는다.
. F는 문서값을 계산하는 검색 함수(Retrieval Function)으로서 다음과 같이 정의된다.

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

검색 함수 F는 각 쌍의 (d,q)에 0부터 1사이의 값을 지정한다. 이 값은 문서 d와 질의 q사이의 유사성을 의미하며 질의 q에 대한 문서 d의 문서 값이다. 검색 함수 F(d,q)는 다음과 같은 2단계 과정을 거쳐서 계산된다.

1. 질의에 나타난 각각의 색인어 t_i에 대하여, F(d,q)는 문서 d에서 색인어 t_i의 가중치 w_i로 정의된다.
2. 불리안 연산자 AND, OR 그리고 NOT은 다음의 식을 이용하여 계산된다. 두 개 이상의 불리안 연산자를 포함하는 불리안 질의는 가장 안쪽에 위치하는 절부터 순환적으로 계산된다.

Boolean Formulation	Evaluation Formula
F(d, t_1 AND t_2)	MIN(F(d,t_1), F(d,t_2))
F(d, t_1 OR t_2)	MAX(F(d,t_1), F(d,t_2))
F(d, NOT t_1)	1 - F(d,t_1)

예를 들어, F(d,t_1) = 0.7, F(d,t_2) = 0.2, F(d,t_3) = 0.1이고, 질의를 ((t_1 OR t_2) AND NOT t_3)라 하자. 이때 F(t_1 OR t_2)는 MAX(F(d,t_1), F(d,t_2)) = 0.7이고, F(d, NOT t_3) = 0.9이며, 질의 최종 결과는 MIN(0.7, 0.9) = 0.7이 된다.

문서의 순위는 질의에 대하여 계산된 유사도의 크기에 따라 결정된다. 이 때, 문서가 색인어 가중치를 갖지 않으면 검색 결과는 불리안 모델과 일치한다. 또한 퍼지 집합 모델에서 논리적으로 같은 의미를 갖는 질의에 대한 탐색 결과는 탐색 순서에 상관없이 항상 일치한다. 그러나 퍼지 집합 모델이 문서들의 순위를 결정하는 문서값을 계산함으로써 불리안 검색 시스템의 단점을 극복하였지만 많은 경우에 부정확한 문서 값을 생성하기 때문에 정보 검색 모델로서 부적합하다고 알려져 왔다.[1][3]

III. 정보 검색을 위한 퍼지 연산자의 분석

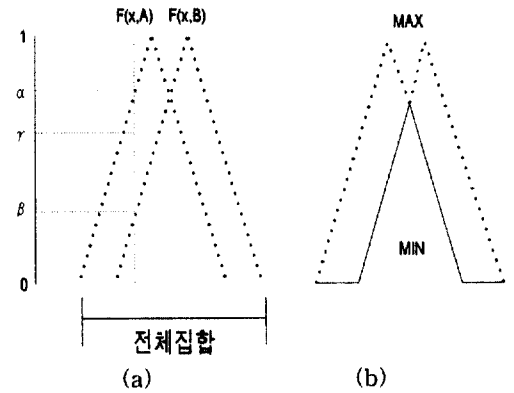
연산자 그래프는 퍼지 연산자의 특성을 시각적으로 편리하게 표현하는 방법이다. 그림 3.1(a)는 연산자 그래프를 작성하는 방법을 보여준다. 세로축은 0부터 1사이의 소속값을 나타내고 가로축은 전체 집합(Universal Set) X를 의미한다. 전체 집합 X 내에서 정의된 두 개의 퍼지 집합 A와 B의 소속 함수가 F(x, A)와 F(x, B)이 때, 임의의 퍼지 연산자에 대한 연산자 그래프는 다음과 같이 작성된다. 전체 집합에 속하는 임의의 원소 k의 퍼지 집합 A와 B에 대한 소속값 F(k, A)와 F(k, B)에 퍼지 연산자를 적용함으로써 얻은 값들은 그래프 상에 표기한다. 예를 들면 F(k, A) = α이고 F(k, B) = β일 때, 연산자 그래프의 작성을 위한 하나의 값 γ는 α와 β에 해당 연산자를 적용함으로써 계산된다. 그림 3.1(b)는 MIN과 MAX 연산자들에 대한 연산자 그래프이다. 굵은 선은 MIN 연산자에 대한 연산자 그래프이고, 점선은 MAX 연산자에 대한 연산자 그래프이다.

1. 퍼지 연산자의 분류

퍼지 집합 이론은 0부터 1사이의 소속값(Membership Value)을 갖는 원소들의 집합에, 집합 이론에서 정의된 집합 연산자에 대응하는 새로운 퍼지 연산자를 정의함으로써 개발되어

왔다. 일반적으로 하나의 집합 연산자에 대하여 이에 대응하는 다수의 퍼지 연산자가 개발되고 있으며, 서로 다른 퍼지 연산자는 서로 다른 특성을 지닌다. 이러한 퍼지 연산자들은 T-norm, T-conorm 그리고 평균연산자로 구분될 수 있다.[4] 그림 3.2는 이러한 세가지 부류의 연산자들 사이의 관계를 보여준다. 세로축은 0부터 1사이의 소속값을 나타내고, 가로축은 전체 집합(Universal Set) X를 의미한다. 전체 집합 X내에서 정의된 두 개의 퍼지 집합 A와 B의 소속함수를 각각 F(x, A)와 F(x, B)라 하고, 특히 원소 k의 소속값을 각각 F(k, A) = α, F(k, B) = β, $\alpha \leq \beta$라고 하자. 퍼지 집합 A와 B를 결합할 때, T-norm, T-conorm 그리고 평균연산자는 다음과 같은 특성을 갖는다. T, T^c, A는 각각 T-norm, T-conorm 그리고 평균연산자를 의미한다.

$$\alpha \leq A(F(k, A), F(k, B)) \leq \beta \quad (1)$$
$$0 \leq T(F(k, A), F(k, B)) \leq \alpha \quad (2)$$
$$\beta \leq T^c(F(k, A), F(k, B)) \leq 1 \quad (3)$$



(a) 생성 (b) MIN과 MAX연산자에 대한 그래프

그림 3.1 연산자 그래프

가. T-연산자

T-연산자는 AND 연산을 위한 T-norm과 OR 연산을 위한 T-conorm으로 구분된다. 지금까지 퍼지 집합 이론에서 다양한 T-연산자들이 개발되어 왔으며 AND와 OR 연산을 위해 가장 널리 사용된 MIN과 MAX 연산자는 각각 T-norm, T-conorm에 속한다. 표 3.1은 지금까지 개발된 다양한 T-연산자들을 보여준다.[5] x와 y는 결합되는 두 개의 퍼지 집합에서의 소속값을 의미한다.

그림 3.2 T-norm, T-conorm, 평균연산자의 특성 비교

표 3.1 T-연산자

	T(x,y)	Tc(x,y)	Comment
1	MIN(x,y)	MAX(x,y)	
2	x · y	x+y-xy	
3	MAX(x+y-1, 0)	MIN(x+y,1)	
4	xy/(x+y-xy) x if y = 1 y if x = 1 0 otherwise	(x+y-2xy)/(1-xy) x if y = 0 y if x = 0 1 otherwise	
5			
6	λxy/(1-(1-λ)(x+y-xy)	(λ(x+y)+xy(1-2λ))/(λ+xy(1-λ))	$0 \leq \lambda \leq \infty$
7	MAX(1-((1-x)p+(1-y)p)1/p, 0)	MIN((xp+yp)1/p, 1)	$1 \leq p \leq \infty$
8	1/(1+((1/x-1)λ+(1/y-1)λ)1/λ)	1/(1+((1/x-1)-λ+(1/y-1)-λ)-1/λ)	$0 \leq \lambda \leq \infty$
9	xy/MAX(x,y,λ)	1-((1-x)(1-y))/MAX((1-x),(1-y),(1-λ))	$0 \leq \lambda \leq 1$
10	MAX((x+y-1+λxy)/(1+λ),0)	MIN((x+y+λxy,1)	$-1 \leq \lambda \leq \infty$
11	MAX((1+λ)(x+y-1)-λxy,0)	MIN(x+y+λxy,1)	$-1 \leq \lambda \leq \infty$

표 3.1에서 제시된 11개의 T-연산자들은 매개변수를 갖지 않는 연산자들 T_1-T_5, T_1^c-T_5^c와 매개변수를 갖는 연산자들 T_6-T_{11}, T_6^c-T_{11}^c로 구분될 수 있다. 매개 변수를 갖는 않는 T-연산자들은 임의의 두값 x,y에 대하여 다음과 같은 관계를 갖는다.

$$T_1(x,y) \geq T_4(x,y) \geq T_2(x,y) \geq T_3(x,y) \geq T_5(x,y)$$
$$T_1^c(x,y) \leq T_4^c(x,y) \leq T_2^c(x,y) \leq T_3^c(x,y) \leq T_5^c(x,y) \quad (4)$$

매개변수를 갖는 T-연산자는 매개변수의 값을 변경하므로써 연산 결과로 생성되는 소속값을 조절할 수 있다. 매개변수를 갖지 않는 T-연산자는 매개변수를 갖는 T-연산자의 특별한 경우로 취급될 수 있으며 두 부류의 연산자는 다음과 같은 관계를 갖는다.

$$T_4(x,y) \geq T_6(x,y) \geq T_5(x,y), T_4^c(x,y) \leq T_6^c(x,y) \leq T_5^c(x,y)$$
$$T_1(x,y) \geq T_7(x,y) \geq T_3(x,y), T_1^c(x,y) \leq T_7^c(x,y) \leq T_3^c(x,y)$$
$$T_1(x,y) \geq T_8(x,y) \geq T_5(x,y), T_1^c(x,y) \leq T_8^c(x,y) \leq T_5^c(x,y)$$
$$T_1(x,y) \geq T_9(x,y) \geq T_2(x,y), T_1^c(x,y) \leq T_9^c(x,y) \leq T_2^c(x,y)$$
$$T_2(x,y) \geq T_{10}(x,y) \geq T_5(x,y), T_2^c(x,y) \leq T_{10}^c(x,y) \leq T_5^c(x,y)$$
$$T_2(x,y) \geq T_{11}(x,y) \geq T_4(x,y), T_2^c(x,y) \leq T_{11}^c(x,y) \leq T_5^c(x,y) \quad (5)$$

나. 평균연산자

퍼지 결정 이론(Fuzzy Decision Theory)에서 의사결정(Decision-Making)은 퍼지 집합의 교집합 또는 합집합을 생성함으로써 이루어지며, T-연산자는 많은 경우에 인간의 의사결정을 표현하기 위하여 사용되어 왔다. 그러나 T-연산자는 경영 결정(Managerial Decision)을 표현하기에 부적합한 것으로 판명되었으며, 이러한 문제점을 극복하기 위하여 평균연산자가 개발되었다[9]. 본 절에서는 지금까지 개발된 네가지 평균연산자를 소개한다.

Zimmermann[4]은 "compensatory and"라고 불리는 평균연산자를 제안하였다. "compensatory and"연산자는 대수합과 대수곱을 결합하여 생성되며 매개변수 γ의 값을 변경함으로써 연산의 결과로 생성되는 소속값을 조절할 수 있다.

$$(A_1) \quad (x \cdot y)^{1-\gamma} \cdot (x+y-x \cdot y)^{\gamma}, \quad 0 \leq \gamma \leq 1 \quad (6)$$

또한 AND와 OR 연산을 위해 개발된 퍼지 연산자들에 선형 블록 결합(Linear Convex Combination) 방법을 사용함으로써 매개변수 γ를 사용하는 평균연산자들을 정의할 수 있다[4][6]. 예를 들어, 선형 블록 결합 방법을 MIN과 MAX 연산자에 사용하면 평균연산자 A_2가 정의되고 대수곱과 대수합을 적용하면 평균연산자 A_3가 정의된다.

$$(A_2) \quad (1-\gamma) \cdot MIN(x,y) + MAX(x,y), \quad 0 \leq \gamma \leq 1 \quad (7)$$
$$(A_3) \quad (1-\gamma) \cdot (x \cdot y) + \gamma \cdot (x+y-x \cdot y), \quad 0 \leq \gamma \leq 1 \quad (8)$$

지금까지 기술된 평균연산자들은 교집합을 위한 연산자와 합집합을 위한 연산자를 구분하지 않는데 합집합과 교집합을 위한 연산자를 별도로 구분하는 평균연산자인 "fuzzy and"와 "fuzzy or"가 제안되었다. "fuzzy and" 연산자는 MIN 연산자와 산술평균(Arithmetic Mean) 연산자를 결합함으로써 생성되었고, "fuzzy or"연산자는 MAX 연산자와 산술평균 연산자를 결합함으로써 생성되었다.

$$(A_4) \quad \gamma \cdot MIN(x,y) + (1-\gamma)(x+y)/2, \quad 0 \leq \gamma \leq 1 \quad (9)$$
$$(A_5) \quad \gamma \cdot MAX(x,y) + (1-\gamma)(x+y)/2, \quad 0 \leq \gamma \leq 1 \quad (10)$$

2. 기존의 퍼지 집합 모델의 문제점

MIN과 MAX 연산자는 검색 효율을 감소시키는 특성을 지니고 있기 때문에 기존의 퍼지 집합 모델은 정보 검색 시스템의 검색 모델로서 부적합하다고 알려졌다.[1][4] 보기 1과 2는 AND 연산을 위해 MIN 연산자를 사용하는 기존의 퍼지 집합 모델이 사람이 생각하는 것과 다르게 문서의 순위를 결정함을 보여준다. MAX 연산자도 MIN 연산자와 유사한 문제점들을 발생시킨다.

보기 1 : 색인어와 가중치의 쌍으로 표현된 두

개의 문서 d_1, d_2와 불리안 질의 q_1이 다음과 같이 주어졌다고 가정하자.

d_1 = {(Fuzzy, 0.50), (Retrieval, 0.50)}
d_2 = {(Fuzzy, 0.99), (Retrieval, 0.49)}
q_1 = Fuzzy AND Retrieval

MIN 연산자가 AND연산을 위해 사용되었을 때, 질의 q_1에 대한 d_1과 d_2의 문서값은 각각 0.50과 0.49이다. 따라서 기존의 퍼지 집합 모델은 d_1이 d_2보다 높은 순위를 갖는 것으로 결정한다. 그러나 대부분의 사람들은 d의 질의에 대한 만족도가 d보다 높은 것으로 결정할 것이다.

보기 2 : 두 개의 문서 d_3와 d_4, 그리고 질의 q_2가 다음과 같이 주어졌다고 가정하자.

d_3 = {(t_1,0), (t_2,1), (t_3,1), . . .,(t_{100},1)}
d_4 = {(t_1,0), (t_2,0), (t_3,0), . . .,(t_{100},0)}
q_2 = t_1 AND t_2 AND . . . AND t_{100}

이때 기존의 퍼지 집합 모델은 질의 q에 대한 d의 문서값과 d의 문서값이 동일한 것으로 결정한다. 또한 문서 d가 질의에서 지정된 99개의 색인어를 포함하고 있음에도 불구하고 d의 문서값을 0으로 계산한다.

보기 1과 2에서 설명된 기존의 퍼지 집합 모델의 문제점은 MIN과 MAX 연산자가 두 개의 피연산자를 모두 고려하지 않고 단지 하나의 피연산자에 전적으로 의존하는 결과값을 생성하기 때문에 발생한다. MIN과 MAX 연산자가 발생시키는 이러한 문제를 "단일 피연산자 의존 문제(Single Operand Dependency Problem)"라 한다.

3. 정보검색시스템에서 T-연산자의 사용에 대한 문제점

MIN과 MAX 이외의 T-연산자들은 다음과 같은 두가지 공통적 특성을 지닌다. 첫째, 하나의 피연산자가 0또는 1일 경우 MIN, MAX와 마찬가지로 하나의 피연산자에 전적으로 의존하는 결과값을 생성한다. 둘째, 피연산자의 값들이 0과 1이 아닐 경우 두 개의 피연산자를 모두 고려하여 결과값을 계산하며, 계산된 결과값은 피연산자 값들의 최소값보다 작거나 최대값보다 크다.

퍼지 집합 모델에서 MIN과 MAX 이외의 T-연산자들의 사용은 위에서 언급된 첫 번째 공통적 특성으로 인하여 보기 2에서 기술된 문제점을 여전히 발생시키지만 두 개의 피연산자를 모두 고려하여 결과값을 계산하기 때문에 보기 1에서 기술된 문제점을 완화시킨다. 예를 들어, 보기 1에서 MIN 연산자 대신에 곱하기 연산자를 사용한다고 가정하자. 이때 d_1과 d_2에 대한 문서값은 각각 0.16과 0.39로 계산되고 따라서, d_2가 d_1보다 높은 순위를 부여받고 검색된다. 그러나 계산된 결과값이 피연산자 값들의 최소값보다 작거나 최대값보다 크다는 특성은 다음의 보기와 같은 새로운 문제점을 발생시킨다.

보기 3 : 문서 d_5와 두 개의 질의 q_1과 q_3가 다음과 같이 주어졌다고 가정하자.

d_5 = {(Information, 0.70),
(Retrieval, 0.70),
(System, 0.70)}
q_1 = Information AND Retrieval
q_3 = System

MIN과 MAX를 제외한 T-연산자들은 AND연산에 있어서 두 개의 피연산자들의 최소값보다 작은 값을 생성한다. 따라서 이러한 연산자들은 불리안 연산자 계산식으로 사용하는 퍼지 집합 모델은 문서 d_5와 질의 q_1사이의 유사성이 d_5와 q_2사이의 유사성보다 적은 것으로 결정할 것이다. 예를 들어, 곱하기 연산자 T_2가 사용되었을 때, 질의 q_1에 대한 문서 d_5의 문서값은 0.49이고 q_3에 대한 d_5의 문서값은 0.70이다. 이러한 결정은 대부분 사람들의 의사와 상반되는 것으로 정보 검색 시스템의 검색 효율을 저하시킨다.

MIN과 MAX 이외의 T-연산자들을 불리안 연산자 계산식으로 사용하는 정보 검색 시스템

의 문제점들은 다음과 같이 요약할 수 있다. 첫째, 연산자들의 첫 번째 공통적 특성은 보기 2와 같은 형태의 단일 피연산자 의존 문제를 발생시킨다. 둘째, MIM과 MAX 이외의 T-연산자들은 결과값을 계산하기 위해 두 개의 피연산자 값들의 상호 보상(Compensation)을 허용한다. 그러나 이러한 보상이 검색 효율에 부정적인 방향으로 이루어 지기 때문에 "부정적 보상 문제(Negative Compensation Problem)"라고 하는 보기 3과 같은 새로운 문제를 발생시킨다.[3]

IV. 새로운 평균연산자 기반의 향상된 퍼지 집합 모델

1. 새로운 평균연산자 기반의 퍼지 집합 모델

퍼지 결정 이론(Fuzzy Decision Theory)는 퍼지 집합의 합집합 또는 교집합으로 보여진다. T-연산자들과 평균연산자들은 많은 경우에 있어서 인간결정을 모델화하는데 사용되어 왔다. 그러나, T-연산자들이나 평균연산자들의 어느 것에 있어서 경영결정을 모델화하는데 어느 것도 적절치 않다고 알려져 왔다. 본 논문에서는 이전 절에에서 보인 문제를 극복하기 위해 새로운 평균연산자를 제안한다. 이 새로운 연산자는 퍼지 집합 모델의 검색 효율을 향상시킬 수 있다.

그림 4.1은 제안된 평균연산자와 연산자 그래프를 보여준다. 연산자 그래프는 연산자의 특성을 나타내는 한 방법이다. 본 논문에서는 퍼지연산자들의 행동특성을 분석하기 위해 연산자 그래프를 사용한다.

연산자그래프는 주어진 두 개의 연산자에 대하여 구성된다. 연산자와 연산자 그래프는 각각 선과 점선으로 표현된다. 그래프에서 수직적 교차점은 소속값(membership)의 정도를 나타낸다. 연산자 그래프는 연산자를 적용함으로써 계산된 점들의 집합이다.

새로운 평균연산자 AND

$\gamma \cdot (x+y-x \cdot y)+(1-\gamma) \cdot (x+y)/2, \quad 0 \le \gamma \le 0.5$

새로운 평균연산자 OR

$\gamma \cdot (x+y-x \cdot y)+(1-\gamma) \cdot (x+y)/2, \quad 0.5 \le \gamma \le 1$

$$\alpha \le A(\alpha,\beta) \le \beta$$

$$0 \le T(\alpha,\beta) \le \alpha$$

$$\beta \le T^c(\alpha,\beta) \le 1$$

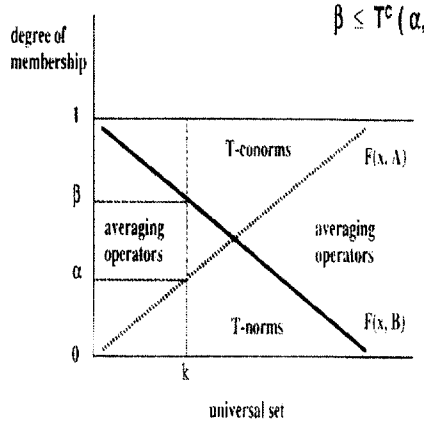


그림 4.1 평균연산자와 연산자 그래프

새로운 평균연산자의 연산자그래프는 결과값이 항상 두 개의 값들의 적은 값보다 더 크고, 큰 값보다 더 작다. 결과적으로, 본 논문에서는 퍼지 집합 모델의 평가 공식으로서 새로운 평균연산자를 사용할 것을 제안한다. 이 퍼지 집합 모델의 공식은 보기 1, 2와 3에서 기술된 모든 문제들을 해결할 수 있다. 또한, 다음절에서 기술되는 제안된 퍼지 집합 모델은 좀 더 높은 검색 효율을 제공한다.

2. 실험 및 결과 분석

가. 실험 설계

실험에 사용된 컴퓨터는 Sun Sparc이고 사용된 데이터 집합은 CACM 3204와 ISI 1460 두 종류를 사용하였다. 데이터 집합 CACM 3204는 1959년과 1979년 사이에 출판된 "Communications of the ACM"의 논문표지을 모은 것으로써

3,000개의 문서와 60개의 질의로 구성되어 있으며, ISI 1460은 "Social Science Citation Index"에서 추출한 것으로 1,500개의 문서와 40개의 질의로 구성되어 있다. 각 문서는 제목, 요약, 저자, 키워드 등으로 구성되어 있다. 각 문서 집합의 질의에 대한 문서의 관련성 평가는 목포대학교에서 수행되었다. CACM 3204의 경우에 어떤 질의는 문서의 관련성 평가가 존재하지 않고 2개의 질의는 특정 저자가 쓴 문서를 검색하는 질의이다. 본 실험에서는 관련된 문서의 평가가 없는 질의와 특정 저자에 대한 질의를 제외한 CACM 3204의 50개와 ISI 1460의 35개의 질의를 사용하여 실험하였다.

자연어로 작성된 질의를 목포대학교에서 불리안 질의로 작성하였다. 자연어 질의와 그에 대한 불리안 질의의 예는 다음과 같다.

자연어 질의 :
I'm interested in mechanisms for communicating between disjoint processes, possibly, but not exclusively, in a distributed environment. I would rather see descriptions of complete mechanisms, with or without implementations, as opposed to theoretical work on the abstract problem. Remote procedure calls and message-passing are examples of my interests.

불리안 질의:
#or (#and ('communicating', 'processes'),
#and ('processes', 'remote', 'procedure'),
#and('message', 'passing'));

문서는 색인어의 집합으로 표현되고 문서 표현을 위한 색인어는 문서의 제목과 요약에서 추출하였다. 문서 집합에서 색인어의 추출은 다음과 같은 방법으로 하였다.

1. 문서의 제목과 요약에 나타나는 각각의 단어를 색인어로 간주한다.
2. 문서 집합에 있는 모든 문서의 색인어 역할을 하는 단어를 색인어 집합에서 제거한다. 이러한 단어는 문서들을 서로 구별하는데 아무런 도움이 되지 않기 때문이다.
3. "retrieve", "retrieves", "retrieval"과 같은 색인어들에서 접미사를 제거하여 하나의 색인어로 인식한다. 이렇게 함으로써 검색 시스템의 재현 능력도를 향상 시키게 된다.
4. 색인어가 문서 내에서 갖는 중요성을 반영하기 위해서 색인어가 가중치를 부여한다.

어떤 문서를 표현하는 모든 색인어들은 서로 다른 중요도를 가지고 문서를 표현한다. 각각의 색인어가 갖는 중요도에 대한 특별한 정보가 없을 경우 색인어가 한 문서내에서 갖는 중요도는 그 색인어가 출현하는 빈도에 따라 가중치를 부여할 수 있다. 출현 빈도에 의한 색인어 가중치는 역 문서 빈도 IDF(Inverse Document Frequency)와 문서내에서 그 색인어가 나타나는 빈도 TF(Term Frequency)의 곱하기로 정의될 수 있다. 문서 i에 출현하는 색인어 k에 대한 가중치 W_{ik}는 다음과 같다. TF_{ik}는 문서 i에서 색인어 k가 출현한 빈도, IDF_k는 색인어 k에 대한 역 문서 빈도, N은 문서 집합에 있는 전체 문서의 수, n_k는 색인어 k를 포함하는 문서의 수 이다.

본 실험에서는 어떤 문서의 저자가 그 문서의 키워드를 선정해 두었을 경우 출현 빈도에 의해서 그 문서의 가장 중요한 색인어로 결정된 색인어와 동일한 가중치를 부여하고 색인어의 가중치가 0에서 1사이의 값을 갖게하기 위하여 다음과 같은 정규화된 가중치를 사용한다. TF_{ih}는 문서 i에서 색인어 h가 출현한 빈도이고 IDF_h는 색인어 h의 역 문서 빈도이다.

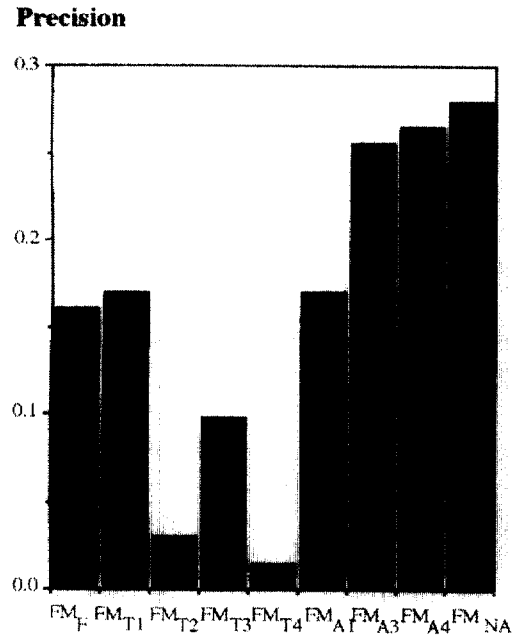
나. 실험 결과 및 분석

본 실험에서는 AND와 OR에 대한 연산자 계산식으로 새로운 평균연산자를 사용하는 것이 다른 연산자를 사용하는 것보다 우수한 검색 효율을 제공함을 입증한다. 또한 본 논문에서 제안한 새로운 평균연산자와 확장된 불리안 모델의 불리안 연산자 계산식의 검색 효율과 문서값 계

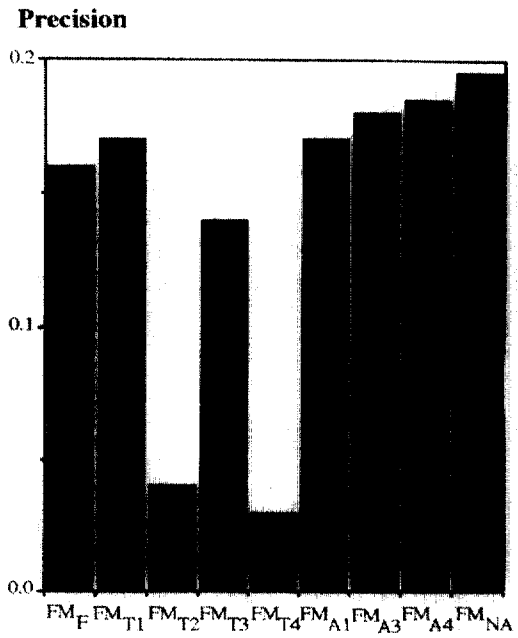
산 속도를 비교한다. 본 실험에서는 많은 비교를 하여야 하므로 여러 재현 능력도에 대한 각각의 평균 검색 정밀도를 비교하는 것은 매우 어려운 일이다. 따라서 본 실험에서는 비교를 쉽게하기 위하여 각 매개변수에 대하여 낮은 재현 능력도인 0.25, 중간 재현 능력도인 0.5, 높은 재현 능력도인 0.75에 대한 검색 정밀도를 산술 평균하여 하나의 값으로 비교한다. 0.25, 0.5 그리고 0.75에 대한 검색 정밀도는 질의들에 대한 평균 검색 정밀도이다.

본 논문에서는 MIN, MAX, T-연산자, 기존의 평균연산자 그리고 새로운 평균연산자와 같은 다양한 연산자들의 검색효율을 비교한다. 그림 4.4은 다양한 퍼지 연산자들에 근거한 퍼지 집합 모델의 검색효율을 보여준다. 여기서 FM_{op}는 퍼지 집합 모델에서 op연산자 가 AND와 OR연산을 평가하기 위해 사용하는 퍼지 검색 모델을 나타낸다. FM_f, FM_{Ti}, FM_{Ai}, FM_{NA}들은 각각 MIN, MAX 연산자들을 가진 퍼지 집합 모델, T-연산자를 가진 퍼지 집합 모델, 기존의 평균연산자를 가진 퍼지 집합 모델, 본 논문에서 제안한 평균연산자를 가진 퍼지 집합 모델과 등가이다. 본 논문에서는 세 개의 전형적인 재호출 레벨(0.25, 0.50, 0.75)에서 평균정도를 표현하는 하나의 정확한 값을 계산한다. 어떤 파라미터화된 연산자들에 대하여 본 논문에서는 그것의 적당한 범위내에서 해당되는 파라미터를 바꿈으로서 정도(precision) 값을 계산하고, 최대값을 선택한다. 그림 4.2에서 보이는 것과 같이 본 논문에서 제안한 평균연산자를 가진 퍼지 집합 모델은 각 집합에 대하여 기존의 퍼지 연산자들 보다 높은 검색 효율을 제공한다.

확장된 불리안 모델의 불리안 연산자 계산식과 본 논문에서 제안한 불리안 연산자 계산식의 연산 수행 시간을 비교하였다. 두 식의 연산 수행 시간을 비교하기 위하여 "A OR B"와 "A AND B"의 계산을 10,000회 수행하였을 때 소요된 시간을 측정하였다. 본 논문에서 제안한 불리안 연산자 계산식의 소요 시간은 0.04초이고 확장된 불리안 모델의 경우 1.54초가 소요되었다. 이 결과는 본 논문에서 제안한 불리안 연산자



(a) The CACM 3204 Collection



(b) The ISI 1460 Collection

그림 4.2 다양한 퍼지 연산자들의 검색 효율

계산식은 확장된 불리안 모델의 단점으로 지적된 느린 검색 시간의 문제점을 해결할 수 있음을 입증한다.

V. 결론

본 논문에서는 기존의 퍼지 연산자들의 문제를 극복하고 퍼지 집합 모델의 검색 효율을 향상시키는 새로운 평균연산자를 제안하였다. 본 논문에서 새로운 평균연산자가 기존의 T-연산자와 평균연산자들보다 좀더 효율적인 새로운 평균연산자라는 것을 실험을 통해 보였다. 결과적으로 새로운 평균연산자를 사용하는 퍼지 집합 모델이 좀더 높은 검색효율을 제공하고 확장된 불안 검색모델의 연산자들과 비교에서 유사한 효율을 보인다는 것을 실험결과를 통해 알 수 있다.

본 논문에서는 또한 FM_{na}와 확장된 불리안 모델의 검색효율을 비교한다. 확장된 불리안 모델은 기존의 퍼지 집합 모델의 문제점을 극복하기 위하여 제안되었고, 좀더 효율적인 IR 모델중의 하나로 알려졌다[2]. FM_{na}와 확장된 불리안 검색 모델을 비교하기 위하여 본 논문에서는 t_1 AND t_2와 t_1 OR t_2를 10,000번 평가하는데 소비하는 계산 시간을 측정하였다. 결과적으로 FM_{NA}와 확장된 불리안 검색 모델에 대하여 각각 0.035와 1.55초가 각각 걸렸다. 결과적으로 제안한 평균연산자는 확장된 불리안 모델의 연산자보다 좀더 효율적이다.

참고 문헌

1. A. Bookstein, "Fuzzy Request : An Approach to Weighted Boolean Searches", Journal of the American Society for Information Science, Vol. 31, No. 4, pp. 240-247, 1980
2. G. Salton, E.A. Fox, and H. Wu, "Extended Boolean Information Retrieval", Communications of the ACM, Vol. 26, No. 11, pp. 1022-1036, 1983
3. S. E. Robertson, "On the Nature of Fuzzy: A Diatrive", Journal of the American Society for Information Science, Vol. 29, No. 6, pp. 304-307, 1978
4. H. J. Zimmermann, "Fuzzy Set Theory and Its Applications", 2nd ed., Kluwer Academic Publishers, 1991.
5. M. M. Gupta and J. Oi, "Theory of T-Norms and Fuzzy Interface Methods", Fuzzy Sets and Systems, Vol. 40, No. 3, pp. 431-450, 1991
6. J. H. Lee, M. H. Kim and Y. J. Lee, "Enhancing the Fuzzy Set Model for High Quality Document Rankings", Proceedings of the 19th Euromicro Conference, Paris, France, pp. 337-344, 1992
7. 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
8. H. J. Zimmermann and P. Zysno, "Latent Connectives in Human Decision Making", Fuzzy Sets and Systems, Vol. 4, No. 1, pp. 37-51, 1980.
9. V. Tahani, "A Fuzzy Model of Document Retrieval Systems", Information Processing & Management, Vol. 12, No. 3, pp. 177-188, 1976
10. U. Thole, H. J. Zimmermann and P. Zysno, "On the Suitability of Minimum and Product Operators for the Intersection of Fuzzy Sets", Fuzzy Sets and Systems, Vol. 2, No. 2, pp. 167-180, 1979