퍼지情報檢索시스템의 검색효율에 관한 연구

A Study on Evaluating Effectiveness of Fuzzy Information Retrieval System

김현희(Hyun-Hee Kim)*
배금표(Gum-pyo Bae)*

□목 차□

초 록

본 연구에서는 이진색인체계를 유지하면서 퍼지디소러스를 통해 퍼지정보검색을 구현할 수 있는 시스템을 구축하고 그 검색결과를 불리언 검색결과와 비교해 보았다. 실험결과는 재현률의 경우 퍼지집합검색이 75%로 불리언 검색의 60% 보다 15% 높았으며, 정확률의 경우 불리언검색이 73%로 퍼지집합검색의 69% 보다 4% 정도 높았다.

ABSTRACT

This study compared the retrieval effectiveness of conventional Boolean and fuzzy set retrieval strategies through a retrieval experiment. The recall ratio of fuzzy set retrieval is 75%, higher than the 60% of Boolean retrieval. On the other hand, the precision ratio of fuzzy set retrieval is 69%, a little bit lower than the 73% of Boolean retrieval.

*명지대학교 문헌정보학과

1. 서 론

1.1 연구의 목적

현재 많이 이용되고 있는 정보검색기법인 불리언 검색은 이용자들에게 익숙하고 컴퓨터처리가 용이하다는 장점이 있는 반면, 단어와 단어사이의 중요도를 표시하지 못하고 완전히 일치되는 문헌만이 검색되는 등의 몇가지 단점들이 있다.

퍼지집합검색은 이러한 불리언검색이 안고 있는 문제점들을 어느 정도 해결할 수 있을 것으로 본다. 그러나 퍼지집합검색을 하기 위해서는 퍼지색인체재가 필요한데 대부분의 상업용 데이타베이스가 이진색인체계에 기초하고 있으므로 이 기법의 실용화를 위해서는 좀 더 효율적인 방법이 요망되고 있다.

따라서, 본 논문에서는 이진색인체제를 유지하면서 퍼지집합검색을 실현할 수 있는 퍼지정보검색시스템을 구축하여 그 검색 결과를 불리언집합 검색결과와 비교해 보았다.

1.2 연구의 방법 및 범위

본 연구에서 구축한 퍼지정보검색시스템은 문헌데이타베이스는 이진색인체제를 유지하면서 퍼지디소러스화일을 이용하여 퍼지정보검색을 구현하는 시스템이다.

구축된 퍼지정보검색시스템은 시스템 내부에 자동색인기능과 퍼지검색기능을 갖추고 있다. 색인어는 실험문헌집단을 분석하여 언어학적 분석기법인 형태소 해석방법에 의해 자동으로 추출하고, 실험문헌집단내에서의 키워드들의 분포패턴과 퍼지집합연산을 이용하여 퍼지디소러스화일을 구성한다. 그 다음 평균-연산자를 이용하 여 퍼지정보검색을 수행해 보고, 그 결과를 불리언검색 결과와 비교해 본다.

실험문헌은 「정보관리학회지」와 「정보관리연구」에 수록된 135편의 논문을 선정하여 구성하였다. 선정된 135편의 논문을 대상으로 논문의 제목과 초록에서 색인어를 추출하였는데, 한글과 영어는 그대로 입력하였고 한자는 한글로 변환하여 표기하였다. 화일조직과 검색실험을 위하여 사용한 프로그램밍 언어는 Turbo C와 Fox Base+ 이며, 하드웨어는 IBM-PC 호환기종을 이용하여 실험하였다.

1.3 선행연구

다음은 정보검색에 퍼지이론을 응용한 국내외 연구들과, 정보센터에서 실제 운용되고 있는 퍼지정보검색시스템에 대해 살펴 보고자 한다.

외국에서의 연구는 퍼지색인시스템, 퍼지디소러스를 이용한 정보검색시스템, 그리고, 키워드관계행렬을 이용한 퍼지정보검색시스템에 관한 연구 등으로 구분된다.

라데키(Radecki, 1981)는 퍼지논리에 기초한 정보검색시스템의 이론적인 기초를 처음으로 확립하였다. 그는 문헌의 디스크립터에 가중치를 부여하고 검색기법으로 퍼지집합검색을 이용한 퍼지색인시스템을 제안하였다. 그러나, 문헌에 할당된 디스크립터에 정확한 가중치를 부여하는 것은 매우 어려운 작업이다. 또한 가중치 부여에 대한 분명한 기준이 없어서 색인자의 주관에 따라 많이 달라질 수 있으므로 객관성 유지에도 문제가 있을 수 있다.

미야모토(Miyamoto, 1990a ; 1990b) 등은 단어의 동시출현빈도와 퍼지집합연산에 기초하여 퍼지디소러스를 생성하고, 이 퍼지디소러스를 이용한 정보검색시스템을 제안하였다. 이들이 제안한 시스템에

서는 디스크립터쌍의 동시출현이 퍼지화의 기준이 되고 있다.

오가와(Ogawa, 1991) 등은 통계정보를 이용해 퍼지디소러스와 비슷한 키워드관계행렬을 생성하였다. 그런 다음, 키워드관계행렬을 이용해 문헌집단의 이진색인으로부터 퍼지색인을 생성하고 문헌들을 적합도에 따라 등급지우는 퍼지정보검색시스템을 설계하고 그 검색 결과를 불리언검색 결과와 비교해 보았다. 비교 결과, 퍼지정보검색의 정확률은 불리언검색에 비해 2% 정도 낮았지만 재현률은 16% 정도 향상되었다. 더욱이, 통계정보를 이용해 구축한 초기키워드관계값을 주제전문가들의 지적 판단을 반영하여 수정보완한 후에는 불리언검색에 비해 정확률과 재현률이 7%, 35%로 각각 증가하였다.

미야모토(Miyamoto, 1990a)가 개발한 퍼지디소러스와 퍼지색인을 이용한 퍼지정보검색시스템은 Tsukuba 대학의 과학정보처리센터에 있는 일반 정보검색시스템 FACOM FAIRS-I의 기능을 확장하여 구축하였다. 이 시스템에서는 퍼지디소러스를 통해서 정보검색을 하라는 명령어를 입력하면 탐색어와 관련된 용어들이 가중치와 함께 검색되며 또한 질문에 대한 적합도(소속함수)에 따라 그룹지워진 문헌들이 검색된다.

국내의 연구를 살펴보면 이순재(1989)는 확률 및 퍼지집합이론에 의한 정보검색기법을 고찰하고 퍼지검색 모델의 특징과 문제점을 제시하였다.

이승채(1991)는 퍼지키워드관계행렬을 전문가들의 협조를 얻어 수작업으로 구성하고 이 퍼지키워드관계행렬을 이용한 퍼지정보검색과 불리언검색에 따른 검색 결과를 비교해 본 결과 퍼지정보검색이 정확률은 3% 정도 낮았으나, 재현률은 10% 정도 높았다.

조혜민(1990)은 기존의 불리언 정보검색시스템의 단점을 보완하고자 불리언 검색시스템에 가중치를 포함시킨 가중치 불리언 검색시스템을 설계하였다. 이 시스템에서는 색인어와 탐색어에 모두 가중치를 주는 방법을 채택하였으며 문헌의 적합도를 계산하기 위해 퍼지집합개념을 이용하였다. 또한, 이 시스템은 퍼지디소러스를 사용하여 탐색문에 있는 탐색어는 물론 탐색어의 상위어, 하위어 등의 관련어까지 검색될 수 있도록 설계하였다.

배금표(1993)는 퍼지키워드관계 행렬을 이용한 퍼지정보검색의 결과와 동일한 관계행렬을 통해 탐색문을 확장하여 수행한 불리언검색의 결과를 비교해 보았다. 비교 결과 이 두 기법의 재현률은 거의 비슷했으나 정확률은 퍼지정보검색이 11% 정도 더 높게 나왔다.

1.4 가 설

앞의 선행연구에서의 연구 결과와 이론을 기초로 하여 수립한 가설은 다음과 같다.

1) 퍼지집합검색이 불리언검색 보다 재현률은 더 높을 것이다.

2) 퍼지집합검색과 불리언검색의 정확률은 같을 것이다.

2. 이론적 배경

본 연구에서 구축한 퍼지정보검색시스템은 시스템 내부에 자동색인기능, 퍼지디소러스구축기능 및 퍼지검색기능을 갖추고 있다. 여기에서는 시스템을 구축하기 위해서 고찰한 기법들에 대하여 설명하고자 한다.

2.1 자동색인모형

본 연구에서는 언어학적기법의 형태소해석방법을 중심으로 색인어를 추출하였기 때문에 한글문헌에 적용된 형태소해석방법에 의한 자동색인모형들을 살펴보기로 하겠다.

형태소해석방법이란 대상문헌에서 단어를 형태소 단위로 해석한 후, 각 형태소에 대해 통계적방법을 이용하여 색인어를 선정하거나 또는 특정품사를 색인어로 선정하는 방식 등이 있다(김현회, 김용호 1993).

형태소해석방법을 활용하여 색인어를 추출한 연구에는 국내의 경우 김영환(1983)의 논문에서 워드프로세싱 시스템을 이용하여 한글과 한자, 영어 혼용문의 전 색인과정을 자동처리하고 색인어리스트를 출력하는 방법이 제시되었다.

안현수(1986)는 어휘론과 구문론을 적용해 한글문헌의 자동색인 기법에서 한글문헌의 주요어는 체언이 될 수 있으며 체언뒤에는 대체적으로 조사가 붙는다는 한글의 특성을 가정하여 먼저 복합어구를 추출한 후 문헌의 처음으로 돌아가 단일 색인어 추출 알고리듬을 수행하도록 하였다. 여기서 문제점은 조사가 생략된 체언은 색인어로 추출되지 못한 점이다.

서경주 사공철(1991)은 체언과 한자어로 된 어간을 갖는 용언의 어간이 키워드로 적합하다고 가정하였다. 조사 용언어미 불용어 화일들과 키워드 추출 알고리듬을 이용하여 비주제어를 제거하는 방식으로 신문기사들을 대상으로 단어들을 추출하였으며 추출된 단어는 인위적인 어휘조절 과정을 거친 다음 확정된 키워드는 키워드 화일에 첨가시켰다.

홍순철(1990)의 논문에서는 한글 문헌에서 체언과 용언의 원형을 색인어로 선정하였는데 체언 다음에는 대체적으로 조사가 붙고 용언에는 어미가 붙어서 활용한다는 한국어의 형태론적인 특성을 이용하였다. 조사표와 용언어미표를 이용하여 체언과 용언을 식별하는데 이때 문제가 되는것은 체언의 어미가 조사나 용언의 어미형태와 같은 경우가 발생하게 된다. 이를 방지하기 위해서 체언의 어미가 조사나 용언의 어미형태와 일치하는 체언들을 따로 모아서 「조사형명사사전」[1]을 구축하여 활용하였다. 이 색인방법은 한 어절만을 분석함으로써 복합어구를 처리하지 못했다.

2.2 키워드관계값 구축모형

키워드 상호간의 관계값을 측정하는 방법으로 오가와 등(1991)의 키워드간의 관계값 측정공식과 미야모토(1990a)의 퍼지디소러스 생성알고리듬 등이 있다.

2.2.1 퍼지키워드관계행렬

오가와 등(1991)은 다음의 공식을 이용하여 문헌집단에서 키워드간의 초기 관계값을 측정하여 퍼지 키워드 관계행렬을 구성한다.

$$Wij = \frac{Nij}{Ni + Nj - Nij} \quad \cdots\cdots\cdots\cdots\cdots\cdots (1)$$

Wij=색인어 i,j의 초기관계값

Nij=i번째와 j번째 색인어들을 모두 포함하는 문헌수

1) '강의, 개인…'처럼 체언으로 사용되나 그 형태 또는 어절의 어미가 조사의 형태를 지닌 것.

Ni, Nj=i번째 색인어와 j번째 색인어만을 포함하고 있는 문헌수

i=j 인 경우는 Wij=1 이 된다.

공식(1)에 의해서 두개의 색인어가 동시에 포함된 문헌들의 정규빈도수가 결정된다. 그러나 이 초기관계값은 통계정보에 의해서 할당된 값이기 때문에 이용자의 유사측정도와 항상 일치하는 것은 아니다. 따라서 통계정보에 기초한 관계값과 이용자 평가간의 차이를 감소시키기 위해 관계값을 수정하는 알고리듬을 첨가하였다.

2.2.2 퍼지디소러스 생성알고리듬

미야모토(1990a)는 단어의 동시출현빈도와 퍼지집합연산에 기초하여 퍼지디소러스를 생성한다. 색인어간의 관련도(s(W_i, W_j))와 포함도(t(W_i, W_j))를 다음의 공식을 이용하여 구한다.

$$s(W_i, W_j)=\frac{\sum_k \min[h_{ik}, h_{jk}]}{\sum_k \max[h_{ik}, h_{jk}]} \quad \cdots\cdots\cdots\cdots\cdots\cdots (2)$$

$s(W_i, W_j)$=용어 W_i와 W_j의 관련도 (RT, Related Term)를 나타냄

h_{ik}=문헌 d_k에서 용어 i의 출현빈도

h_{jk}=문헌 d_k에서 용어 j의 출현빈도

$$t(W_i, W_j)=\frac{\sum_k \min[h_{ik}, h_{jk}]}{\sum_k h_{ik}} \quad \cdots\cdots\cdots\cdots\cdots\cdots (3)$$

$t(W_i, W_j)$=용어 W_j와 W_i의 포함도 (NT, Narrower Term)를 나타냄

h_{ik}=문헌 d_k에서 용어 i의 출현빈도

h_{jk}=문헌 d_k에서 용어 j의 출현빈도

만약 $s(W_i, W_j) \geq \alpha$ 이면, (W_i RT W_j)으로 표현한다. 또한 $t(W_i, W_j) \geq \beta$ 이면 (W_i NT W_j)와 (W_j BT W_i)으로 표현한다.

다음은 구체적 예를 들어 위의 공식 (2)(3)을 이용하여 퍼지디소러스를 생성하는 과정을 설명한다. 다음 〈표 1〉과 같이 문헌 3개 d1, d2, d3와, 색인어 6개 w1, w2, w3, w4, w5, w6이 있다고 가정해 보자. 괄호안에 있는 숫자는 색인어의 출현빈도를 나타내는 것이다. 문헌 d1에서의 w1(2)는 색인어 w1이 문헌 d1에서 두번 출현한다는 것을 뜻한다.

다음 s와 t에 공식(2)와 (3)을 응용하면, 다음과 같은 값들을 얻는다.

$$s(w1, w2)=\frac{2+0+0}{2+0+1}=\frac{2}{3}$$

$$t(w1, w2)=\frac{0}{2+0+1}=\frac{2}{3}$$

$$t(w2, w1)=\frac{2}{2}=1$$

색인어의 모든 쌍의 s와 t의 계산 결과가 다음의 〈표 2〉, 〈표 3〉와 같다.

〈표 1〉 색인어 출현빈도

d1 : w1(2), w2(2), w3(1), w4(1), w5(1)
d2 : w3(2), w4(1), w6(2)
d3 : w1(1), w4(3), w6(1)

〈표 2〉 s의 계산 결과

	w1	w2	w3	w4	w5	w6
w1	1	2/3	1/5	1/3	1/3	1/5
w2	2/3	1	1/4	1/6	1/2	0
s=w3	1/5	1/4	1	1/3	1/3	1/2
w4	1/3	1/6	1/3	1	1/5	1/3
w5	1/3	1/2	1/3	1/5	1	0
w6	1/5	0	1/2	1/3	0	1

〈표 3〉 t의 계산 결과

	w1	w2	w3	w4	w5	w6
w1	1	2/3	1/3	2/3	1/3	1/3
w2	1	1	1/2	1/2	1/2	0
t=w3	1/3	1/3	1	2/3	1/3	2/3
w4	2/5	1/5	2/5	1	1/5	2/5
w5	1	1	1	1	1	0
w6	1/3	0	1/3	2/3	0	1

위의 행렬에서 항목 (i,j)는 s(wi, wj) 또는 t (wi, wj)를 나타낸다. 만약 $\alpha=1/2$, $\beta=1$로 잡으면 다음과 같은 결과를 얻을 수 있다.

$\alpha=1/2$: w1 RT w2, w2 RT w5, w3 RT w6

$\beta=1$: w2 NT w1 (w1 BT w2), w5 NT w1 (w1 BT w5), w5 NT w2 (w2 BT w5), w5 NT w3 (w3 BT w5), w5 NT w4 (w4 BT w5)

2.3 퍼지집합검색모형

초기의 가장 간단한 퍼지집합검색은 문헌에 할당된 색인어에 그 문헌에 대한 색인어의 0-1 사이의 소속함수(가중치)를 부여하거나 또는 탐색문의 탐색어에 해당 질문에 대한 탐색어의 소속함수를 부여한 후 min-max 연산 즉 'and'인 경우는 소속함수의 최소값을 취하고 'or'인 경우는 소속함수의 최대값을 취하는 계산방식을 이용하여 관련 문헌들을 해당 질문에 대한 소속함수순으로 검색

〈표 4〉 전통적 퍼지집합모형에서 유사성평가규칙

불리언 공식화	평가공식
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)

하는 것이었다.

그러나, 이와 같은 min-max 연산자에 기초한 전통적인 퍼지집합모형은 정보검색시스템에 적합하지 않는 것으로 판명되었다. 왜냐하면, 이 모형은 하나의 피연산자값에 의존하여 문헌의 적합도(소속함수)를 측정하므로 많은 경우에 주제 전문가의 지적 판단과 일치하지 않는 것으로 판명되었기 때문이다.

min-max 연산자이외에 많은 연산자들이 제안되었는데 이러한 연산자들은 크게 T-연산자들과 평균-연산자들로 구분할 수 있다. 다음은 먼저 T-연산자의 한 종류인 min-max 연산자를 응용한 전통적인 퍼지집합모형에 대해 설명한 다음 다른 T-연산자들과 평균-연산자들을 응용한 퍼지집합모형에 대해 각각 설명한다(Kim et al. 1993 ; Lee et al. 1992 ; Lee et al. 1993a ; Lee et al. 1993b).

2.3.1 MIN-MAX 연산자 모형

min-max연산자를 응용한 전통적인 퍼지집합모형에서 문헌의 소속도 계산은 다음과 같이 간단하게 기술되어진다. 먼저 질문 Q가 주어지면, 질문 Q와 문헌 D간의 유사성 Sim(Q, D)는 함수 F(D, Q)로 정의된다. 함수 F(D, Q)는 주어진 질문 Q에 의해 검색된 문헌 집합에서 문헌 D의 소속함수이다. 전통적인 퍼지집합모형을 채택한 정보검색시스템에서, 다음 두 단계는 문헌과 질문간의 유사성을 계산하기 위해서 일반적으로 이용된다.

1) 탐색문에 있는 각 탐색어 Ti에 대하여, F(D, Ti)는 문헌 D에서 색인어 Ti의 가중치로 정의되어진다.

2)불리언 연산자 즉 AND, OR, NOT는 〈표 4〉에 주어진 규칙에 기초하여 평가되어진다. 두개 이상의 불리언 연산자를 포함한 불리언 탐색문의 경우, 계산순서는 가장 안 쪽의 절부터 수행해 나간다.

예를 들어서, 탐색문이 Q=((T1 OR T2) AND T3)이고 문헌 D=((T1, 0.7), (T2, 0.5), (T3, 0.8))이라고 가정해 보자. 전통적인 퍼지집합모형은 다음과 같은 절차에 따라 탐색문 Q와 문헌 D간의 유사성을 계산한다. 먼저, 탐색문에서 각 탐색어 Ti에 대해 함수 F(D, Ti)가 계산된다. 즉 F(D, T1)=0.7, F(D, T2)=0.5, F(D, T3)=0.8이 된다. 그런 다음, 첫절 F(D, T1 OR T2)=MAX(F(D, T1), F(D, T2))=0.7이 된다. 탐색문 전체에 대한 최종 문헌값은 MIN(F(D, T1 OR T2), F(D, T3))=0.7이 된다.

그러나, 이러한 전통적인 퍼지집합모형은 AND (또는 OR) 연산을 위한 MIN(또는 MAX) 연산자가 소속도가 가장 낮은 (또는 높은) 하나의 피연

산자에 의존하는 단일 피연산자의존문제를 야기시켜 주제 전문가의 지적 판단과 맞지 않는 많은 검색 결과들을 산출하고 있다.

다음은 두 가지 예를 들어 이러한 문제점들을 살펴 보고자 한다. 아래 (예 1)에서는 'AND' 연산자의 경우만 예를 들었는데 'OR'연산자의 이용도 비슷한 문제점들을 야기시키고 있다.

(예 1) 아래와 같은 문헌 D_1, D_2가 있고 질문 Q_1이 있다고 가정해 보자.

D_1={(디소러스, 0.40), (클러스터링, 0.40)}
D_2={(디소러스, 0.99), (클러스터링, 0.39)}
Q_1=디소러스 AND 클러스터링

〈표 3〉에서 기술한 유사성평가규칙에 따라, 문헌 D_1, D_2의 소속함수가 0.40, 0.39로 각각 산출되었다. 이에 따라, 문헌 D_1이 문헌 D_2 보다 등급이 더 높게 나왔는데, 대다수의 사람들은 문헌 D_2가 문헌 D_1 보다 질문에 더 적합하다고 결정할 것이다. 이러한 문제는 AND 연산을 위한 MIN 연산자가 소속도가 가장 낮은 하나의 피연산자에만 의존하기 때문이다. 다음의 예는 이러한 문제를 좀 더 분명하게 나타낸다.

(예 2) 아래와 같은 문헌 D_3, D_4가 있고 질문 Q_2가 있다고 가정해 보자.

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}

여기서, 전통적인 퍼지집합모형은 질문 Q_2에 대한 문헌 D_3, D_4의 소속함수가 0으로 동일하게 결정될 것이다. 더욱이, D_3의 99개의 색인어들이 탐색문 Q_2에 있는 탐색어들을 만족시킨다 하더라도 문헌 D_3의 소속함수는 0이 되는 결과를 만들어낸다.

2.3.2 T-연산자 모형

T-연산자는 T-norms와 T-conorms로 구분되며, 퍼지집합검색에서 T-norms와 T-conorms은 각각

〈표 5〉 T-연산자

	T(x, y) (T-norms)	T^c(x, y) (T-conorms)
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+y−2xy/(1−xy)
5	x if y=1 y if x=1 0 otherwise	x if y=0 y if x=0 0 otherwise

AND와 OR의 연산에 응용될 수 있다. 따라서 min 연산자는 T-norms에 속하며 max 연산자는 T-conorms에 속한다. 〈표 5〉는 T-연산자들을 나타낸 것이다.

min-max연산자들(T_1과 T^c_1)을 제외한 T-연산자들(T_2-T_5와 T^c_2-T^c_5)은 다음과 같은 두가지 공통된 특징을 갖는다. 첫째는 하나의 피연산자가 0 또는 1일 경우 MIN, MAX와 마찬가지로 하나의 피연산자에 전적으로 의존하는 결과값을 생성하며, 두번째는 피연산자의 값들이 0과 1이 아닐 경우 두개의 피연산자를 모두 고려하여 결과값을 계산하며, 계산된 결과값들은 피연산자값들의 최소값 보다 작거나 최대값 보다 크다.

T-연산자들(T_2-T_5와 T^c_2-T^c_5)의 사용은 앞에서 언급한 첫번째 공통적 특성으로 인하여 두번째 예에서 기술한 문제점은 여전히 해결되지 않은 채 남아 있지만 두개의 피연산자를 모두 고려하여 결과값을 계산하기 때문에 첫번째 예(예 1)에서 나타난 문제점이 완화될 수 있다. 첫번째 예에, 곱연산자, $T_2(x,y)$을 응용하면 질문 Q_1에 대한 문헌 D_1, D_2의 소속함수는 각각 0.16, 0.39가 된다. 따라서, 문헌 D_2가 D_1 보다 더 높은 등급으로 검색된다.

그러나, 계산된 결과값이 피연산자값들의 최소값 보다 작거나 최대값 보다 크다는 두번째 특성은 다음과 같은 또다른 부정적 보상문제(negative compensation problem)를 야기시킨다.

(예 3) 아래와 같은 질문 Q_1, Q_3에 대하여 문헌 D_5가 있다고 가정해 보자.

D_5={(디소러스, 0.70), (클러스터링, 0.90), (시스템, 0.70)}

Q_1=디소러스 AND 클러스터링 0.63

Q_3=시스템 0.70

위의 예에 곱연산자, $T_2(x,y)$을 응용하면 질문 Q_1에 대한 문헌 D_5의 소속도는 0.63이며, 질문 Q_3에 대한 문헌 D_5의 소속도는 0.7이 된다. 그러나, 대부분의 사람들은 질문 Q_1과 문헌 D_5간의 유사성이 질문 Q_3과 D_5간의 유사성 보다 더 높다고 판단할 것이다.

결론적으로, T-연산자들은 min-max 연산자의 단일 피연산자의존문제를 완전히 해결해 주지도 못하면서, 세번째 예에서 나타난 부정적 보상문제(negative compensation problem)라는 새로운 문제를 야기시키고 있다고 할 수 있다.

2.3.3 평균-연산자 모형

위에서 기술했듯이 T-연산자들은 전통적인 min-max 연산자들의 문제점들을 완전히 해결해 주지 못하였는데 평균-연산자들이 이러한 문제점들을 해결하기 위해서 제안되었다. 다음 〈표 6〉의 공식들은 평균-연산자들을 나타낸 것이다.

불리언 연산자 계산식이 항상 피연산자들의 최소값과 최대값 사이의 값을 생성하는 특성을 지니고 있다면, 앞의 예 1, 2, 3을 통하여 설명된 단일 피연산자의존문제와 부정적 보상문제를 해결할 수 있을 것이다. 평균-연산자들 중 A_2, A_4 연산자들만이 언제나 피연산자들의 최소값과 최대값 사이의 값을 생성하는 특성을 지니고 있기 때문에 이 두 문제를 해결해 주고 있는 것으로 증명되었다(Kim et al. 1993). 평균연산자 A_2와 A_4는 이러한 특성때문에 '긍정적 보상연산자(positively com pensatory operator)'로 지칭된다.

다음은 평균연산자 A_4를 응용한 예들을 들어 어떻게 단일 피연산자의존문제와 부정적 보상문제를 해결하는지를 살펴 보고자 한다.

(예 1) 평균연산자 A_4를 응용하여 질문 Q_1에 대

〈표 6〉 평균-연산자

$$(A_1)\ (X \cdot Y)^{(1-\gamma)}(X+Y-X \cdot Y)^{\gamma},\ 0 \leqq \gamma \leqq 1$$

$$(A_2)\ (1-\gamma) \cdot MIN(x, y)+\gamma \cdot MAX(x, y),\ 0 \leqq \gamma \leqq 1$$

$$(A_3)\ (1-\gamma) \cdot (x \cdot y)+\gamma \cdot (x+y-x \cdot y),\ 0 \leqq \gamma \leqq 1$$

$$(A_{4.and})\ \gamma \cdot MIN(x, y)+\frac{(1-\gamma)(x+y)}{2},\ 0 \leqq \gamma \leqq 1$$

$$(A_{4.or})\ \gamma \cdot MAX(x, y)+\frac{(1-\gamma)(x+y)}{2},\ 0 \leqq \gamma \leqq 1$$

〈표 7〉 평균-연산자를 응용한 D_1, D_2 소속도

	$A_{4.AND}$	
	D_1	D_2
γ 0.1	0.40[2]	0.66
0.2	0.40	0.63
0.3	0.40	0.60
0.4	0.40	0.57
0.5	0.40	0.54
0.6	0.40	0.51
0.7	0.40	0.48
0.8	0.40	0.45
0.9	0.40	0.42

한, 문헌 D_1과 D_2의 소속도를 예시한 〈표 7〉를 보면 문헌 D_2의 소속도가 D_1의 소속도 보다 항상 더 큰 것으로 나타나 대부분의 주제 전문가의 판단과 일치함을 알 수 있다. 이것은 평균연산자 A_4가 두 피연산자간의 적절한 보상을 허용하기 때문이다.

(예 2) 만약 더 작은 피연산자값이 더 큰 피연산자값에 의해 항상 보상되어진다면 앞의 두번째 예에서 예시한 문제점을 피할 수 있는데 평균-연산자 A_4는 그러한 문제점을 해결하고 있다. 평균연산자 A_4를 응용하여 질문 Q_2에 대한, 문헌 D_3과 D_4의 소속도를 예시한 〈표 8〉를 보면 주제 전문가의 판단과 같이 문헌 D_3의 소속도가 D_4의 소속도 보다 항상 더 큰 것으로 나타나고 있다.

---

2) $0.1 \cdot MIN(0.40, 0.40)+(1-0)(0.4+0.4)/2 = 0.4$

〈표 8〉 평균-연산자를 응용한 D_3, D_4 소속도

	$A_{4.AND}$	
	D_3	D_4
γ 0.1	0.90	0
0.2	0.80	0
0.3	0.70	0
0.4	0.60	0
0.5	0.50	0
0.6	0.40	0
0.7	0.30	0
0.8	0.20	0
0.9	0.10	0

〈표 9〉 Q_1과 Q_3에 대한 D_5 소속도

	$A_{4.AND}$	
	Q_1	Q_3
γ 0.1	0.79	0.70
0.2	0.78	0.70
0.3	0.77	0.70
0.4	0.76	0.70
0.5	0.75	0.70
0.6	0.74	0.70
0.7	0.73	0.70
0.8	0.72	0.70
0.9	0.71	0.70

(예 3) 평균-연산자 A_4는 가장 낮은 피연산자값과 가장 높은 피연산자값 사이에 있는 문헌의 소속도를 항상 산출하기 때문에 세번째 예에서 기술한 문제를 피할 수 있다. 평균연산자 A_4를 응용하여 질문 Q_1, Q_3에 대한 D_5의 소속도를 예시한 〈표 9〉를 보면, 대부분의 주제 전문가들의 판단 처럼 질문 Q_1과 문헌 D_5간의 유사성이 질문 Q_3과 D_5간의 유사성 보다 더 높게 나오고 있다.

3. 퍼지정보검색시스템구축

3.1 시스템구축

본 연구에서 구축한 퍼지정보검색시스템은 앞에서 논의된 오가와 등(1991)이 개발한 퍼지정보검색시스템과 미야모토(1990a ; 1990b)가 개발한 퍼지정보검색시스템을 기본틀로 하여 구축하였으며 시스템 내부에 자동색인기능, 퍼지디소러스화일 구축기능 및 퍼지검색·불리언검색기능을 갖고 있다.

이 시스템는 이진색인체계를 유지하면서 퍼지디소러스화일을 통해 퍼지정보검색을 실현하고 있다. 이와같이 퍼지색인을 이용하지 않은 이유는 대부분의 상업용 데이타베이스가 이진색인에 기초하고 있으며 또한 키워드의 수가 문헌의 수보다 보통 작기 때문에 퍼지색인보다 퍼지디소러스를 유지하는 것이 훨씬 용이하기 때문이다.

시스템이 갖는 화일들에는 문헌화일(마스터화일), 도치색인화일, 퍼지디소러스화일이 있다.

3.2.1 자동색인기능

통계정보와 형태소해석방법에 의해 체언을 색인어로 추출하였는데 실험문헌 135개를 분석하여 207개의 색인어가 추출되었다. 그런 다음 각 문헌에 가중치가 없는 색인어를 할당하여 문헌화일(마스터화일)과 도치색인화일을 구축하였다.

1) 자동색인절차

다음의 자동색인 절차는 초록에서 색인어를 추출할 때 적용한 방법이다. 제목에서 색인어를 추

〈표 10〉 문헌화일의 레코드 구조

문헌번호	1
제 목	컴퓨터와 연관된 지적소유권 보호책의 현황과 문제점
저 자	이순자
색 인 어	미디어, 선진국, 지적소유권, 컴퓨터
초 록	이 논문은 컴퓨터와 연관된 지적소유권에 대한 기존 법적보호의 적용문제를 다룬 것이다. 선진국의 현황과 그 문제점을 조사함으로써 우리나라의 저작권법 개정안에 보완되어야 할 사항, 또는 특허권법의 적용가능성등을 제시하였고 어떤 법적보호책도 새로운 미디어를 포괄적으로 다루어야 할 것과 국제적인 협의 안에서 서로 동등하게 인정되어야 할 것을 결론으로 내 놓았다.
서지사항	한국정보관리학회, 「정보관리학회지」 1 : 1 (1984), 9-24.

〈표 11〉 도치색인화일의 레코드 구조

색 인 어	네트워크
문헌번호	34, 41, 97, 129

출할 때는 제목에서는 체언의 조사가 생략된 형태가 많기 때문에 불용어를 제외한 모든 어절들중에서 총 출현빈도가 3 이상인 어절을 색인어로 추출하였다.

(a) 어절분석 : 공란을 어절의 구분자로 하여 어절을 분석하였으며, 실험문헌집단에서 총 출현빈도가 2 이하인 어절들은 분석 대상에서 제외시켰다.

(b) 불용어제거 : 불용어사전을 구축하고 불용어사전에 나오는 어절들은 제외시켰다. 불용어사전에는 304개의 단어가 등록되어 있다.

(c) 조사형명사사전체크 : 체언에 붙는 조사가 생략된 경우 체언의 어미가 조사와 같은 형태의 체언들(예, 전문가(전문가 시스템)) 중 주제어로 생각되는 키워드들은 조사형명사로 분류하여 조사형명사사전을 구축하여 색인어로 뽑아주었다. 조사형명사사전에는 16개의 단어가 등록되어 있다.

(d) 조사사전체크 : 어절의 어미를 조사사전과 비교하여 일치하면 체언으로 간주하고 색인어로 추출한다.

(e) 용언제거 : 체언 또는 용언의 어미가 조사와 일치하는 어절들이 있기 때문에 바로 앞 단계(d)에서 체언으로 뽑힌 단어들 중 '이용하는', '되어지는', '생략되는' 처럼 용언이면서도 체언으로 인식되어 '이용하', '되어지', '생략되'가 체언으로 최종적으로 선택된다. 이런 오류를 방지하기 위해서 어절의 어미가 '는'으로 인식된 어절은 바로 그 앞 글자를 식별하여 만약 그 글자가 '하', '지', '되'의 세 자 중 한 글자에 해당되면 용언으로 인식하여 제외시켰다. 이 과정에 의해서 모든 용언들을 완전히 제외시키지 못하였지만 출현빈도가 높은 용언들은 대부분 제거된다.

(f)단수형으로 변환 : 선택된 체언중에서 어절의 끝 단어가 접미사 '들'인 경우는 '들'을 제거하여 단수형으로 변환시켰다.

2) 데이타베이스의 구축

자동색인 절차에서 선택된 색인어들을 이용하여 문헌화일(마스터화일)[〈표 10〉 참조]과 도치색인화일[〈표 11〉 참조]을 구축하였다. 문헌화일은 문헌번호, 제목, 저자, 색인어, 초록, 서지사항의 필드로 구성되는데 색인어 필드에 수록한 색인어들은 제목과 초록을 분석하여 앞에서 설명한 자동색인 절차에 따라 추출한 체언들이다. 다음의 도치색인화일은 색인어, 문헌번호의 필드로 구성된다.

3.2.2 퍼지디소러스화일 구축기능

앞에서 구축한 데이타베이스를 이용하여 다음의 미야모토공식으로 키워드간의 관계값을 측정하여, 각 키워드의 관련어들을 관련도와 함께 수록한 퍼지디소러스화일을 구성하였다. 다음 〈표 12〉은 퍼지디소러스화일의 일부이다.

$$s(W_i, W_j) = \frac{\sum_k \min[h_{ik}, h_{jk}]}{\sum_k \max[h_{ik}, h_{jk}]}$$

〈표 12〉 퍼지디소러스화일

대 출 업 무	(대학도서관 0.14), (분석 0.11), (분석기법 0.29)
대학도서관	(대출업무 0.14), (설계 0.12), (수준, 0.20) (정보사서 0.5), (정보서비스 0.33), (주제전문사서 0.50), (토탈시스템 0.13)

3.2.3 퍼지정보검색기능

퍼지검색과정은 먼저 이용자의 질문이 들어 오면 탐색문을 작성한 후 이를 시스템에 입력한다. 이 때 필요에 따라 탐색문의 탐색어에 가중치를 부여하거나 또는 가중치를 부여하지 않을 수도 있다. 시스템에 입력된 탐색문은 퍼지디소러스화일을 이용하여 탐색문이 확장되며 확장된 탐색문에 기초하여 퍼지집합연산을 이용하여 적합문헌들을 등급순으로 검색한다.

1) 탐색문확장

시스템내부에서는 퍼지디소러스화일을 체크하여 탐색어의 관련어(들)과 각 탐색어의 가중치를 첨가한 확장된 탐색문을 생성한다.

탐색문을 작성하는 형식에는 CNF(Conjunctive Normal Form)(예, (w_1 or ⋯ or w_3) and (w_4 or ⋯ or w_{10}))와 DNF(Disjunctive Normal Form)(예, (w_1 and ⋯ and w_4) or (w_3 and ⋯ and w_{10}))이 있는데 여기서는 CNF를 채택하였다.

탐색문의 각 탐색어에 가중치가 부여되는데 이때 이용자의 질문에 있는 용어에는 가중치가 1.0이 부여되고 퍼지디소러스화일을 통해 연결된 용어에는 1.0 이하의 가중치가 부여된다. 다음은 하나의 예를 들어 탐색문을 확장하고 퍼지집합연산을 통해 적합문헌들을 검색하는 과정을 설명한다. 예로 든 탐색문에서 'AND NOT'연산자 탐색어만 가중치 0.5를 부여하였다.

(예) 탐색문(q1) :

(자동색인 AND 언어학적 기법(의미분석) AND (NOT 통계적기법 0.5))

확장된 탐색문(q1') :

((자동색인 1) OR (색인어 0.56) OR (의미분석 0.33)) AND ((언어(의미분석 1)) OR (자동색인 0.33))) AND NOT ((통계적기법 0.5) OR (관련성 0.17) OR (색인어 0.19))

탐색문 q1'는 퍼지집합연산을 용이하기 위해서 불리언연산자 'AND'를 기준으로, 다음과 같이 q1'(1), q1'(2), q1'(3)로 구분하여 각각 계산한 다음 나중에 'AND'연산자에 의해 결합한다.

q1'(1)=((자동색인 1) OR 색인어 0.56) OR (의미분석 0.33))

q1'(2)=((의미분석 1) OR (자동색인 0.33))

q1'(3)=NOT((통계적기법 0.5) OR (관련성 0.17) OR (색인어 0.19))

2) 검색함수에 의한 문헌의 적합도계산

검색함수 R은 주어진 질문(탐색문) q에 대한 각 문헌의 적합도(Relevance Status Value, RSV)를 계산하여 정해진 기준치(α=0.44)를 넘으면 검색 결과로 출력해 주는 함수이다. 이 검색함수 R은 검색함수 R1과 R2로 구성된다.

(a)검색함수 R1

좀 더 효율적인 검색을 하기 위해서 먼저 검색함수 R1에 의해서 문헌집단에서 문헌의 적합도(소속함수)를 계산할 대상문헌들을 가려낸다. 검색함수 R1은 4가지 경우로 정의하면 다음과 같다.

· 단일어로 된 탐색문의 경우 : 탐색문 q가 하나의 용어 t1으로 구성될 때 용어 t1을 포

함하고 있는 문헌들이 검색된다.

· AND로 연결된 탐색문의 경우 : 탐색문이 q=t1 AND t2 일때
R1(q)=R1(t1) ∩ R1(t2) 이다.

· OR로 연결된 탐색문의 경우 : 탐색문이 q=t1 OR t2 일때
R1(q)=R1(t1) ∪ R2(t2) 이다.

· NOT이 포함된 탐색문의 경우 : 탐색문에 NOT이 포함된 경우는 대개 q t1 AND (NOT t2)의 형태인데 여기서는 R1(q)=R1(t1)로 하였다.

위의 확장된 탐색문(q1')에 검색함수(R1)를 적용하면 다음과 같이 5개의 문헌이 검색된다. 1차적으로 검색된 5개의 문헌들은 색인어 집합, s1={자동색인, 색인어, 의미분석}에서 최소한 하나의 단어를 색인어로 갖고 있으면서, 색인어 집합, s2={의미분석, 자동색인}에서 최소한 하나의 단어를 색인어로 갖는 문헌들이 검색된다.

R1(q1')={d29, d64, d68, d104, d110}

(b) 검색함수 R2

검색함수 R2는 검색함수 R1에 의해 검색된 문헌들을 대상으로 주어진 질문어 q∈Q에 대한 문헌의 적합도(RSV)를 평균연산자(A_4)에 의해 0과 1사이의 값으로 계산해 낸다. 이때 평균연산자(A_4)의 매개변수, γ값으로 0.7이 가장 효율적인 정보검색결과를 가져다 주는 것으로 확인되어 여기서는 0.7을 사용한다(Lee et al. 1992). 함수 R2는 R2 : Q*R1(Q)→[0, 1]로서 아래와 같이 정의된다.

· 단일 용어로 된 탐색문의 경우 : 탐색어 q=(t1,w1)일 때
R2(q,d)=w1*fd(t1)

· AND로 연결된 탐색문의 경우 : 탐색문이 q=(t1,w1) AND (t2,w2) 일때
R2(q,d)=γ · min(w1*fd(t1), w2*fd(t2))+((1−γ)(w1*fd(t1)+w2*fd(t2))/2

· OR로 연결된 탐색문의 경우 : 탐색문이 q=(t1,w1) OR (t2, w2) 일때
R2(q, d)=γ · max(w1*fd(t1), w2*fd(t2))+((1−γ)(w1*fd(t1)+w2*fd(t2))/2

· NOT이 있는 탐색문의 경우 : 탐색문 q=NOT(t1, w1) 일때
R2(q, d)=1−w1*fd(t1)

다음은 검색함수 R2를 이용하여 검색함수 R1에 의해 검색된 문헌들의 소속함수를 구하는 과정을 설명한다. 먼저, 검색함수 R1에 의해 검색된 문헌들에 할당된 색인어들은 다음과 같다.

d29=(데이터, 문헌, 색인, 색인어, 자동색인, 정보, 조사, 초록, 표제, 한글 문헌)

d64=(기법, 분석, 분석기법, 색인, 색인어, 설계, 성능, 시스템, 신문기사, 의미분석, 자동색인, 컴퓨터, 키워드, 특성, 한글, 화일)

d68=(검색, 관련성, 기술, 모형, 색인어, 시스템, 의미분석, 인공지능, 자동색인, 정보, 정보검색, 통계적기법, 한글)

d104=(분석, 색인어, 색인원, 자동색인, 정보량, 초록, 표제)

d110=(시스템, 신문기사, 자동색인, 전망, 컴퓨터)

다음은 평균연산자(A_4)를 이용하여 서브탐색문, 즉 q1'(1), q1'(2), q1'(3)의 결과를 계산한 다음, 'AND'연산자를 이용하여 이 결과값들을 결합하는 과정을 설명한다.

q1'(1)=((자동색인 1) OR (색인어 0.56) OR (의미분석 0.33))

R2(q1'(1))

d29 ={(색인어 0.56), (자동색인 1)}
d64 ={(색인어,0.56), (의미분석 0.33), (자동색인 1)}
d68 ={(색인어 0.56), (의미분석 0.33), (자동색인 1)}
d104={(색인어 0.56), (자동색인 1)}
d110={(자동색인 1)}

d29 : $0.7 \cdot MAX(0.56, 1)+((1-0.7)(0.56+1))/2=0.93$
d64 : $0.7 \cdot MAX(0.56, 0.33, 1)+((1-0.7)(0.56+0.33+1))/3=0.89$
d68 : $0.7 \cdot MAX(0.56, 0.33, 1)+((1-0.7)(0.56+0.33+1))/3=0.89$
d104 : $0.7 \cdot MAX(0.56, 1)+((1-0.7)(0.56+1))/2=0.93$
d110 : $0.7 \cdot MAX(1)+(1-0.7)(1)=1$

q1'(2)=((언어(의미분석 1)OR(자동색인 0.33))

R2(q1'(2))

d29 ={(자동색인 0.33)}
d64 ={(의미분석 1), (자동색인 0.33)}
d68 ={(의미분석 1), (자동색인 0.33)}
d104={(자동색인 0.33)}
d110={(자동색인 0.33)}

d29 $=0.7 \cdot MAX(0.33)+(1-0.7)(0.33)=0.33$
d64 $=0.7 \cdot MAX(1, 0.33)+((1-0.7)(1+0.33))/2=0.90$

d68 =0.7 · MAX(1, 0.33)+((1−0.7)(1+0.33)) /2=0.90
d104=0.7 · MAX(0.33)+(1−0.7)(0.33)=0.33
d110=0.7 · MAX(0.33)+(1−0.7)(0.33)=0.33

q1'(3)=NOT((통계적기법 0.5)OR(관련성 0.17) OR(색인어 0.19))

R2(q1'(3))

d29 ={(색인어 0.19)}
d64 ={(색인어 0.19)}
d68 ={(통계적기법 0.5), (관련성 0.17), (색인어 0.19)}
d104={(색인어 0.19)}

q1'(3)는 'NOT'연산자로 연결되어 있기 때문에 1에서 평균−연산자(A_4)를 이용하여 구한 값을 뺀 나머지를 취한다.

d29 =1−(0.7 · MAX(0.19)+(1−0.7)(0.19))=0.81
d64 =1−(0.7 · MAX(0.19)+(1−0.7)(0.19))=0.81
d68 =1−(0.7 · MAX(0.5, 0.17, 0.19)+((1−0.7)(1+0.17+0.19)) /3=0.51
d104=1−(0.7 · MAX(0.19)+(1−0.7)(0.19))=0.81

끝으로, 위에서 계산한 결과들을 'AND'연산자로 결합하여 문헌의 최종 소속도를 구한다.

R2(q1')=q1'(1) AND q1'(2) AND q1'(3)

d29 =0.7 · MIN(0.93, 0.33, 0.81)+((1−0.7)(0.93+0.33+0.81)) /3=0.44
d64 =0.7 · MIN(0.89, 0.90, 0.81)+((1−0.7)(0.89+0.90+0.81)) /3=0.83
d68 =0.7 · MIN(0.89, 0.90, 0.51)+((1−0.7)(0.89+0.90+0.51)) /3=0.59
d104=0.7 · MIN(0.93, 0.33, 0.81)+((1−0.7)(0.93+0.33+0.81)) /3=0.44
d110=0.7 · MIN(1, 0.33)+((1−0.7)(1+0.33)) /2=0.43

따라서, 소속함수가 0.44 이상인 다음의 문헌들이 최종적으로 검색된다.

R2(q1')={(d64, 0.83), (d68, 0.59), (d29, 0.44), (d104, 0.44)}

3.3 시스템구현

〈표 13〉 퍼지집합검색의 결과

)FSEARCH "자동색인" and "의미분석" and not "통계적기법 0.5" 1. 자 동 색 인 : 색인어(0.56) 2. 자 동 색 인 : 의미분석(0.33) 3. 의 미 분 석 : 자동색인(0.33) 4. 통계적기법 : 관련성(0.17) 5. 통계적기법 : 색인어(0.19)
***결과 : 용어(들)=5 : 등급 0 0 문헌(들) : 등급 1 1 문헌(들) : 등급 2 0 문헌(들) : 등급 3 3 문헌(들) : 합계 4 문헌(들)

검색시스템은 불리언검색과 퍼지집합검색을 함께 수행할 수 있도록 설계하였다. 동일한 문헌데이타베이스를 사용하면서 불리언검색일 때는 명령어를 SEARCH로 입력하고 퍼지집합검색일 때는 FSEARCH를 사용하도록 하였다. 다음 〈표 13〉는 위에서 예로 든 퍼지집합검색의 결과이다.

FSEARCH는 퍼지디소러스화일을 통해서 정보검색을 하라는 명령어로 탐색어와 관련된 5개의 용어가 가중치와 함께 검색되었다. 등급 0은 소속함수가 1인 것을 나타내며 나머지는 등급 1(소속함수 0.80−0.99), 등급 2(소속함수 0.60−0.79), 등급 3(소속함수 0.44−0.59)을 나타낸다. 따라서, 소속도가 0.44 미만인 문헌들은 부적합 문헌들이 많아 검색되지 않게 하였다.

3.4 검색실험

3.4.1 질문의 구성

검색실험은 불리언검색과 퍼지집합검색 모두 동일한 탐색문 8건을 가지고 수행하였다. 탐색문은 도치색인화일을 참조로 하여 작성하였는데 다음 〈표 14〉는 이용자 질문과 시스템에 입력하기 위해 작성한 탐색문을 나타내는 것이다.

3.4.2 검색실험

불리언 검색시에 탐색문을 작성하여 시스템에 입력하면 퍼지집합검색과는 달리 탐색문을 확장하지 않고 바로 불리언검색을 수행하였다.

퍼지집합검색시에는 탐색문을 시스템에 입력하면 퍼지디소러스화일을 통해 가중치가 부여된 확장된 탐색문을 생성한다. 그런 다음 이 탐색문을 기초로 하여 이진색인체계가 퍼지색인체계로 변환되며 퍼지집합연산자에 의해 각 문헌의 소속도가 계산되어 내림차순으로 정렬되어 진다. 이때 최종적으로 적합문헌을 검색하는 방법에는 검색된 문헌들의 소속도에 기준치를 부여하는 방법과 검색대상문헌의 갯수를 통제하는 방법을 통해서 적합문헌을 검색하는 방법이 있다. 이 시스템에서는 기준치를 0.44로 정하여 문헌의 소속함수가 0.44 이상인 문헌만을 검색하도록 하였다. 그 이유로는 소속도가 낮은 문헌들 중에 상당수의 문헌들이 비적합문헌으로 조사되었기 때문이다.

〈표 14〉검색질문과 탐색문

1	언어학적 기법에 의한 자동문헌에 관한 문헌을 검색하고 싶다. 단 이때 통계적 기법에 의한 자동색인을 다룬 문헌은 검색에서 제외시킨다. (자동색인 AND 의미분석 AND NOT 통계적기법)
2	언어학적 기법에 의한 자동색인에 관한 문헌을 검색하고 싶다. (자동색인 AND 의미분석)
3	전문가시스템 또는 지식기반시스템에 관한 문헌을 검색하고 싶다. (전문가 OR 지식기반시스템)
4	통계적기법에 의한 자동색인에 관한 문헌을 검색하고 싶다. (자동색인 AND 통계적기법)
5	문헌정보학에 관한 문헌을 검색하고 싶다. (문헌정보학)
6	디소러스에 관한 문헌을 검색하고 싶다. (디소러스)
7	계량서지학에 관한 문헌을 검색하고 싶다. (계량서지학)
8	여러 나라의 MARC 포멧을 비교연구한 논문을 검색하고 싶다. MARC AND (비교 OR 비교연구)

〈표 15〉 재현률과 정확률

탐색문 번 호	불리언 검색		퍼지집합검색	
	재현률	정확률	재현률	정확률
1	0.25	1.00	0.75	0.75
2	0.50	1.00	0.75	0.75
3	0.80	0.67	0.80	0.67
4	0.00	0.00	0.50	0.20
5	1.00	0.40	1.00	0.40
6	1.00	0.75	1.00	0.75
7	0.22	1.00	0.22	1.00
8	1.00	1.00	1.00	1.00
평 균	0.60	0.73	0.75	0.69

3.4.3 검색결과 비교분석

검색효율을 평가하는 방법들에는 여러가지가 있는데 본 연구에서는 재현률과 정확률을 이용하였다. 다음 〈표 15〉는 실험결과 각 질문에 대해 정확률과 재현률을 계산한 것이다.

〈표 15〉에서 보는 바와 같이 다른 연구들의 결과와 비슷하게 재현률의 경우 퍼지집합검색이 75%로 불리언 검색의 60% 보다 15% 높았으며, 정확률의 경우 불리언검색이 73%로 퍼지집합검색 보다 4% 정도 높았다. 이 두 정보검색기법의 재현률과 정확률이 유의한 차이가 있는지를 검증하기 위해서, Wilcoxon의 부호순위검정(signed rank-test)을 해 보았다. 그 결과, 유의도 수준 α=0.05에서 정확률과 재현률 모두 유의한 차이가 없는 것으로 나타났다.

4. 결 론

정보검색에 관한 연구가 시작된지 40여년이 지났지만 현재 운용되고 있는 대부분의 정보검색시스템들은 아직도 많은 문제점들을 갖고 있다. 특히 부정확한 색인, 비효율적인 검색기법 등이 해결해야 할 구체적인 과제들로 지적되고 있다. 최근에는 좀 더 효율적인 색인·검색기법들을 위해 많은 수리모형들이 활용되고 있으며 계량언어학까지 이 분야에 적용되고 있다.

본 논문에서는 형태소해석방법에 의한 자동색인기능, 통계정보와 퍼지집합연산을 응용한 퍼지디소러스화일 구축기능, 그리고 퍼지집합 검색기능을 갖는 퍼지정보검색시스템을 구현해 보았다. 그런 다음 이 시스템의 효율성을 평가하기 위해서 재현률과 정확률을 이용하여 그 검색 결과를 불리언검색 결과와 비교해 보았다.

이 연구를 통하여 얻은 결과는 다음과 같다.

1) 실험결과는 재현률의 경우 퍼지집합검색이 75%로 불리언 검색의 60% 보다 15% 높았으며, 정확률의 경우 불리언검색이 73%로 퍼지집합검

색의 69% 보다 4% 정도 높았다. 이 두 정보검색 기법의 재현률과 정확률이 유의한 차이가 있는지를 검증하기 위해서, Wilcoxon의 부호순위검정(signed rank-test)을 해 본 결과, 유의도 수준 α=0.05에서 정확률과 재현률 모두 유의한 차이가 없는 것으로 나타났다.

2) 퍼지정보검색기법은 해당 질문에 대한 각 문헌의 소속함수에 따라 관련문헌을 검색하기 때문에 탐색문에 'AND NOT'이라는 불리언 연산자를 사용할 경우, 'AND NOT' 불리언연산자 탐색어가 색인어로 포함된 적합문헌들도 검색될 수 있는 가능성은 있다. 그러나, 불리언검색에 의해서 이런 유형의 적합문헌들은 항상 누락되었다.

본 시스템 구현시 발생된 문제점과 앞으로 연구되어야 할 과제는 다음과 같다.

1) 본 시스템에서는 퍼지디소러스의 키워드간의 관계값이 재현률과 정확률 향상의 가장 중요한 요인으로 판단되기 때문에 정보검색과정에서 주제전문가들의 의견을 반영하여 퍼지디소러스화일의 색인어간의 초기관계값을 수정 보완하는 알고리듬을 설계하는 것이 바람직하다.

2) 본 시스템에서는 퍼지디소러스화일 구축시 각 색인어의 관련어화일만을 생성하였으나, 마야모토(1990a)의 퍼지디소러스 알고리듬에 의해 각 색인어의 광의어화일(또는 협의어화일)을 구성하여 탐색의 폭을 확장(또는 축소)시킬 수 있다.

3) 한글문헌에서 형태소해석방법에 의해 단일어 형태의 체언을 색인어로 추출하는 시스템에서는 한글띄어쓰기의 부정확성과 조사생략의 체언들이 문제가 되었다. 조사생략의 체언들은 조사형명사사전의 이용이나 조사를 많이 생략하는 제목의 경우는 불용어를 제외한 모든 어절을 색인어로 선정하는 등의 방법들을 통하여 부분적으로 해결할 수 있다. 그러나, 한글 띄어쓰기의 부정확성과 일관성 결여는 여전히 해결해야 할 문제로 남았다.

4) 대규모 문헌데이타베이스에서 퍼지디소러스 화일을 응용할 경우 자연히 키워드의 수가 많아지게 된다. 이 경우에는 키워드들을 통제하여 의미가 거의 비슷한 키워드들을 모아서 하나의 키워드 클래스로 처리해야 할 것이다.

참고문헌

1. 김영환. 1983. 한글 한자 영어 혼용문의 자동색인 시스템. 석사학위논문, 한국과학기술원대학원.
2. 김현회, 김용호. 1993. 계량정보학. 서울 : 구미무역(주) 출판부.
3. 배금표, 1993. "퍼지정보검색시스템의 검색효율에 관한 실험적 연구." 석사학위논문, 명지대학교 대학원.
4. 서경주, 사공철. 1991. "언어학 분석기법에 의한 신문기사 자동시스템설계에 관한 연구." 정보관리학회지 8(1) : 78-99.
5. 안현수. 1986. "한글문헌의 자동색인에 관한 실험적 연구." 석사학위논문, 연세대학교 대학원.
6. 이순재. 1989. "정보검색시스템에 Fuzzy Set이론의 적용," 도서관·정보학연구 (경북대학교대학원 도서관·정보학과) 제1집 : 201-236.
7. 이승채. 1991. "퍼지개념을 적용한 질의식의 분석과 문헌정보검색에 관한 연구." 도서

관학 제 21집 : 249－289.

8. 정영미. 1993. 정보검색론. 서울 : 구미무역(주) 출판부.
9. 정영미, 노영희. 1992. "정보검색기법의 검색효율 비교연구." 연세논총 제28집 : 107－129.
10. 조혜민. 1990. "퍼지논리를 이용한 가중치 부울정보검색시스템." 석사학위논문, 서강대학교 공공정책대학원.
11. 홍순철. 1990. "형태소분석을 이용한 한글 자동색인 시스템 구현에 관한 연구." 석사학위논문, 숭실대학교 산업대학원.
12. Bookstein, A. 1980. "Fuzzy Requests : An Approach to Weighted Boolean Searches." *JASIS* 16(July) : 240－247.
13. Kim, M. H., et al. 1993. "Analysis of Fuzzy Operators for High Quality Information Retrieval." *Information Processing Letters*(to appear).
14. Lee, J. H., et al. 1992. "Enhancing the Fuzzy Set Model for High Quality Document Rankings." *Microprocessing and Microprogramming 35 : 337－344.*
15. _____. 1993a. "Ranking Documents in Thesaurus－Based Boolean Retrieval Systems." *Information Processing & Management*(to appear).
16. _____. 1993b. "On the Evaluation of Boolean Operators in the Extended Boolean Retrieval Framework," *ACM SIGER'93* (to appear).
17. Miyamoto, S. 1990a. "Fuzzy Sets in Information Retrieval and Cluster Analysis." Dordrecht : Kluwer Academic Publishers.
18. _____. 1990b. "Information Retrieval Based on Fuzzy Associations." *Fuzzy sets and Systems 38 : 191－205.*
19. Ogawa, Y., et al. 1991. "A Fuzzy Document Retrieval System Using the Keyword Connection Matries and a Learning Method." *Fuzzy Sets and Systems 39 : 163－179.*
20. Radechi, T. 1981. "Outline of a Fuzzy Logic Approach to Information Retrieval." *International Journal of Man－Machine Studies 14 : 169－178.*
21. Zadeh, L. A. 1965. "Fuzzy Sets." *Information and Control 8(3) : 338－353.*