# 퍼지개념을 적용한 질의식의 분석과 문헌정보 검색에 관한 연구

이 승 채*

〈목 차〉



## 1. 서 론

오늘날 우리 인간이 접하는 모든 자연, 사회현상은 갈수록 복잡해지고 있다. 컴퓨터 기술의 발달에 힘입어 이 복잡한 환경에서 발생하는 많은 문제를 해결하고 있지만, 이는 아직도 매우 제한되어 있다. 따라서 복잡한 현상의 문제를 인간이 해결한다는 것은, 대부분이 복잡한 문제를 단순화시켜 이것을 우리가 이해할 수 있는 정도의 간단한 문제로 만든 다음 그 문제를 해결하는 것이다. 그러나, 단순화하는 과정에서는 필연적으로 문제와 관련된 정보가 손실되기 마련이다.

오늘날 컴퓨터는 우리 주위에 없어서는 안될 중요한 도구가 되어 있으며 우리가 원하는 많은 일을 대신해 주고 있다. 컴퓨터가 일을 하기 위해서는,

* 전남대학교 문헌정보학과 강사.

제기된 문제를 수치로 바꾸어 주어야 하고 컴퓨터는 이 수치를 처리함으로써 우리가 원하는 바를 수행하는 것이다. 이때 컴퓨터에 인지시키는 수치는 정확한 것이어야 한다. 즉 사과 두개 또는 10℃ 등 정확한 수치 또는 개념으로 바꾸어 주어야 처리할 수 있다. 그러나 우리는 적합한 수치로 나타내기 어려워 애매한 표현을 하는 경우가 많이 있음을 알고 있다.

최근 인간과 비슷하게 생각하고 일하는 컴퓨터를 만들고자 하는 인공지능 연구가 활발하게 진행되고 있다. 컴퓨터가 인공지능을 가지고 인간이 원하는 바를 제대로 수행하기 위해서는 인간이 사용하는 숫자는 물론이고 애매한 표현을 처리할 수 있어야 한다. 이러한 인간의 애매한 표현을 처리할 수 있는 이론적인 바탕을 제공하는 것이 바로 퍼지이론이다. 퍼지이론은 현상의 불확실한 상태를 그대로 표현해 주는 방법으로서 1965년 미국 버클리대학의 자데(Lofti A. Zadeh)교수에 의해서 처음 소개되었다[Lofti A. Jadeh]. 퍼지이론은 애매하게 표현된 자료를 우리에게 유용한 자료로 만들기 위하여, 퍼지집합(fuzzy set), 퍼지논리(fuzzy logic), 퍼지숫자(fuzzy number) 등의 개념을 포함하고 있으며 수학적인 계산방법도 잘 개발되어 있다[오길록:1-5].

본 연구에서 퍼지집합이론을 적용하여 구현하고자 하는 정보검색시스템은 검색기법을 크게 분류하여 완전매치기법과 부분매치 기법으로 나누는 관점에서[Belkin & Croft; 김영귀] 볼 때 부분매치기법을 적용한 것이라고 할 수 있다. 내용면에서 본 정보검색시스템은 대상문헌의 저자의 의도를 정확히 추정하는 주제분석과 이용자의 정보요구를 확실하게 파악하는 요구분석을 통하여 양자간의 커뮤니케이션을 증진시키는 기관이라고 할 수 있다. 이러한 중요한 두 가지의 요소중 전자는 흔히 색인이나 초록의 형태로 나타나고, 후자는 이용자의 질문을 특징화시켜 탐색모형을 수립함으로써 수행된다. 그

러나 주제분석과 요구분석은 그 과정상 본질적으로 불확실성과 애매성·모호성을 내포하게 된다. 이상적인 정보검색시스템이란 이러한 요인들을 극소화시킴으로써 보다 적합한 정보를 제공할 수 있어야 한다[이순재:202].

기존에 개발된 정보검색기법들은 논리상 또는 실험적으로 수행된 결과들에서 의미있는 가능성을 제시해 준다. 물론 대부분이 실제의 방대한 시스템에서 적용된 것이 아니고 비교적 소규모의 실험집단을 대상으로 응용한 것이다. 이러한 검색기법들은 모두 그 수에 있어서 차이는 있으나 나름대로의 문제점과 한계성을 지니고 있다. 정보검색기법에 대한 연구는 선행연구들에서 고려하지 못한 새로운 관점, 상이한 기법들의 합병, 기법 자체가 가지고 있는 모순의 극복을 통한 보다 나은 모델을 제시하는 것이다[이순재:202].

또한 보다 효율적인 정보검색시스템은 첫째, 이용자가 표현한 정보요구 내용에 관련된 용어가 문헌에 포함되어 있지 않더라도 관련문헌(적합문헌)을 찾아낼 수 있어야 하며, 둘째, 역으로 이용자가 표현한 정보요구 내용에 있는 용어가 문헌에 나타나 있지 않더라도 관련문헌을 찾아낼 수 있어야 한다고 볼 때[Fox & Koll:259], 퍼지집합이론은 이러한 시스템의 개발에 있어서 하나의 대안이 될 수 있다.

퍼지집합은 소속함수의 경계가 뚜렷하지 않은 대상류로, 다시 말하면 고려된 대상들은 꼭 그러한 류에 속하거나 속하지 않거나 할 필요성이 없고 그 구성원 등급은 [0,1]의 중간에 조정될 수가 있다. 즉, 주어진 개념은 퍼지집합에서 전체집합의 특정요소들의 구성원등급이 특성함수의 일반화인 소속함수로 결정되는 것으로 보통 집합이론과 대비하여 소속함수와 비소속함수의 변이가 연속적인 것을 말한다.

문헌생산량의 증가에 기인하여 문헌의 축적과 검색업무는 현대의

정보관리 환경에서 매우 중요한 기능이 되었다. 비록 정보공학의 발전이 신속히 이루어지고 있기는 하나 과거의 불리안논리에 기초를 둔 검색방법은 급증하는 문헌들을 처리하기에 충분한 능력을 발휘하지 못하고 있는 것이 사실이며, 시스템의 구축과 탐색문 형성에 있어서의 용이함에도 불구하고 몇가지 결점을 지니고 있으며[Cooper,1983 & 1988;정영미:262], 그중에서도 가장 취약한 점으로는 탐색어로 표현되는 각 개념의 상대적 중요도나 관계의 정도(유사도)를 표현하지 못한다는 점을 지적할 수 있다. 현재 널리 이용되고 있는 정보검색시스템에서는 단순히 질문과 문헌간의 공통된 용어의 수만을 고려하므로 공통용어의 수가 많으면 질문과 문헌의 관련성을 높게 나타낼 뿐, 용어간의 관계는 반영하지 않고 있는 실정이다[강일중:1-2]. 따라서 일반적인 불리안 검색모델을 개선하고자 하는 여러 연구가 수행되었으며 그 방법으로는 가중치값을 도입하거나 서열값을 부여하는 방법이 모색되어 왔다.

퍼지집합이론은 경계가 불분명하여 잘못 정의될 수 있는 애매정보를 처리하는데 있어서 적합하기 때문에, 많은 연구자들이 가중치나 서열값을 부여하는데 적용해 왔다[Y. Ogawa et al:163]. 예를 들면, 퍼지색인시스템[Buell & Kraft; Radecki, 1981 & 1983], 일반 색인어에 기초한 퍼지 디소러스를 이용한 퍼지검색시스템[T. Murai et al, 1989], 퍼지색인에 기초한 퍼지디소러스를 이용한 퍼지검색시스템[S. Miyamoto et al, 1983; T. Murai et al, 1988], 인용문헌을 이용한 퍼지검색시스템[K. Nomoto et al, 1987 & 1990] 등이 발표된 바 있다.

한편, 오가와 등[Y. Ogawa et al, 1991]은 색인어 연계행렬(keyword connection metrix)을 이용한 퍼지 문헌검색시스템을 제안한 바 있다. 이 방법에서는 일반적인 불리안색인을 전제로 하고 있는 바, 이는 대부분의 데이터베이스들이 일반 색인방법에 기초를 두

고 있을 뿐만 아니라 통상 색인어의 숫자가 문헌집단의 숫자보다 적기 때문에 퍼지색인보다는 퍼지디소러스를 유지하는 것이 더 용이하다[Ibid:164]는 점에 주목한 것이다.

본 연구에서는 문헌의 표제들로부터 추출된 색인어들의 퍼지관계 행렬을 이용하여 문헌검색을 수행하기 위한 시스템을 구현하여 보았다. 퍼지관계행렬을 이용함으로써 일반 색인어들을 통한 퍼지색인의 작성이 가능하며, 문헌들은 이용자의 질의식과의 관계에 있어서의 적합도(relevance value)에 따라서 순위를 매길 수 있다. 또한 색인어 관계행렬을 통해, 일반 검색시스템에 있어서 검색용 디소러스를 활용하는 작업과 같은 방법으로 이용자가 검색하고자 하는 주제에 적합한 질의식을 작성할 수 있도록 하였다.

한편, 본 연구를 수행함에 있어서 실험대상 문헌집단으로는 우리나라에서 생산된 유전공학분야 연구논문들을 택하였으며, 실험적 시스템이라는 점을 감안하여 120개의 문헌들의 서지정보로 DB를 구축하였다.

본 연구 수행상의 제한점은 실험실 환경에서 시스템을 구축한 때문에 대형 DB를 구축하여 운용할 현장상황에서 어느 정도 효율적일 것인가에 대해서 까지는 접근을 하지 못하였으며, 최근의 정보검색 연구의 양상이 지능형 시스템의 개발에 있다고 할 때[Croft,1987; 이영자:2], 시스템에 높은 수준의 지능을 부여한다는 점에서는 초기적인 수준을 벗어나지 못하였다는 점이다.

## 2. 퍼지관계이론의 기본 개념

### 2.1. 퍼지관계(fuzzy relation) 개념의 소개

자데(L.A. Zadeh)의 퍼지집합이론은 수학에 있어서의 보통집합(crisp set) 이론을 일반화한 것이다[김성혁:69].. 논의되는 전체집

합 U에서 U의 퍼지부분집합(fuzzy subset) A는 다음과 같은 소속함수(membership function) μA에 의해 정의된다.

$$\mu A : U \quad [0,1]$$

여기서 소속함수 $\mu A(X_i)$는 퍼지부분집합 A에 있는 각 원소 $X_i$의 소속의 정도(degree)를 표현하는 것이다. 즉, $\mu A(X_i) = 0$은 소속관계가 없음을 나타내는 것이고, $\mu A(X_i) = 1$은 '완전한 멤버쉽'을, $0 < \mu A(X_i) < 1$은 '부분멤버쉽'을 표시하는 것이다.

일반적인 관계(crisp relation)를 소속함수를 이용하여 표시해 보자[오길록:4-27-31]. 관계 R이 집합 A에서 B로의 관계를 나타낼 때, $x \in A$와 $y \in B$에 대하여,

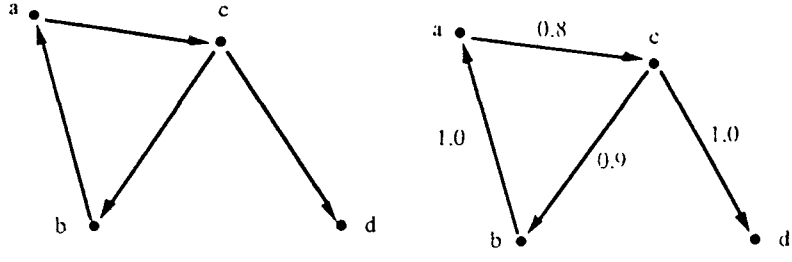
소속함수 $\mu_R(x,y) = 1$ iff (if and only if) $(x,y) \in R$

$0$ iff $(x,y) \in R$

즉, 소속함수는 A x B를 집합 {0,1}으로 대응시킨다.

$$\mu_R : A \times B \rightarrow [0,1]$$

지금까지 살펴본 관계(crisp relation) R은 쌍(x,y)가 R에 소속여부가 명확한 경우이다. 즉, $\mu_R(x,y) = 0$ 또는 1이다. 앞에서 살펴본 퍼지집합의 개념을 이용하면, 이 관계에서도 애매한 관계를 생각




(a) 일반적인 관계 (b) 퍼지관계 $\tilde{R}$

그림 1. 일반적인 관계와 퍼지관계

할 수 있다. 이 애매한 관계는 쌍(x,y)의 퍼지집합 R에 대한 소속여부가 애매하여 소속정도가 [0,1]의 값을 가진다. 즉

$$\mu_R : AxB \rightarrow [0,1]$$

여기에서 R(x,y)는 소속의 정도(membership degree)라기 보다는 관계의 강도(strength)로 해석되기도 한다. 즉 $\mu_R(x,y) \geq R(x',y')$ 일 때 (x,y)는(x′ ,y′ )보다 강하게 연결되어 있다고 해석한다.

예를 들어서 그림 1. (a)를 보자. 그림 (a)에 있는 일반적인 관계 R은 AxA의 관계를 보이고 있다. 이것을 소속함수를 이용하여 표현하면 다음과 같다.

$$\mu_R(a,c) = 1$$
$$\mu_R(b,a) = 1$$
$$\mu_R(c,b) = 1$$
$$\mu_R(c,d) = 1$$

이러한 관계에 소속정도를 0과 1사이의 값으로 준다고 하면, 이 관계는 퍼지관계가 되고 이를 소속함수를 이용하여 표현한 예와 여기에 대응되는 퍼지관계 행렬은 다음과 같다.(일반적으로 AxA 행렬에서는 대칭형관계를 갖게 되지만 여기에서는 관계의 예를 든 것이기 때문에 이를 무시하였다)

| | A/A | a | b | c | d |
|---|---|---|---|---|---|
| R(a,c) = 0.8 | a | 0 | 0 | 0.8 | 0 |
| R(b,a) = 1.0 | b | 1 | 0 | 0 | 0 |
| R(c,b) = 0.9 | c | 0 | 0.9 | 0 | 1 |
| R(c,d) = 1.0 | d | 0 | 0 | 0 | 0 |

이와 같이 퍼지관계로 나타내면 a와 c사이에는 0.8 정도로 관계가 있고, b와 a 사이에는 1.0, c와 b 사이에는 0.9, c와 d 사이에는 1.0의 관계가 있다고 해석할 수 있다. 이를 문헌집합 D = { d1, d2, d3, d4, d5, d6, d7}와 용어집합 T = {t1, t2, t3, t4, t5}의 구성원소들 사이의 퍼지관계행렬로 나타낸 예는 다음과 같다.

| | t1 | t2 | t3 | t4 | t5 |
|---|---|---|---|---|---|
| d1 | 0.6 | 0.9 | 0.8 | 0.14 | 0.9 |
| d2 | 0.9 | 0.15 | 0.4 | 0.4 | 0.8 |
| d3 | 0.2 | 0.15 | 0.4 | 0.4 | 0.8 |
| d4 | 0.1 | 0.3 | 1.0 | 1.0 | 0.1 |
| d5 | 0.9 | 0.4 | 0.2 | 0.0 | 0.9 |
| d6 | 0.4 | 0.2 | 0.8 | 0.4 | 0.1 |
| d7 | 0.3 | 0.1 | 0.3 | 0.3 | 1.0 |

표 1. 색인어와 문헌간의 관계행렬

표 1.에서 보는 바와 같이 행과 열의 교집합 F(di,tj)는 문헌 di와 용어 tj사이의 관계도를 나타낸다.

## 2.2. 퍼지그래프와 퍼지관계

퍼지그래프는 퍼지관계를 표현한 것으로서 결국 동일한 자료구조를 표현한 것이라 할 수 있으므로, 퍼지그래프는 그래프 외에도 퍼지관계처럼 퍼지행렬에 의해서 표현될 수 있다. 다음 퍼지관계 행렬 $M_G$로 나타난 퍼지그래프를 그림으로 표현하면 다음과 같다. 그림으로 표현할 때는 일반적으로 연결선 위에 소속함수값(관계의 강도)을 표시한다.

| | a | b | c | d | e |
|---|---|---|---|---|---|
| a | 1 | 0.9 | 0 | 0 | 0.8 |
| b | 0.4 | 1 | 0.5 | 1 | 0.3 |
| c | 0.9 | 0.8 | 1 | 0 | 0 |
| d | 0 | 0 | 0 | 1 | 0 |
| e | 0.2 | 0 | 0 | 0 | 1 |

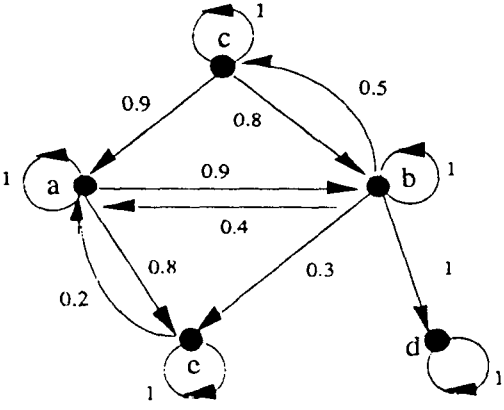
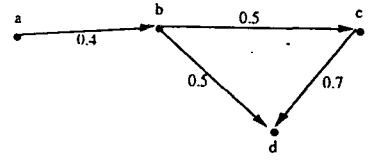


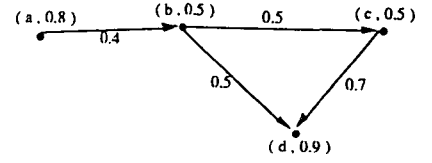

그림 2. 퍼지관계 R과 퍼지그래프

## 2.3. 퍼지망(fuzzy network)

앞에서 퍼지관계와 퍼지그래프는 사실상 동일한 것이라고 했다. 퍼지관계 R은 정의역(domain) A에서 치역(range) B로의 관계를 나타내며, 이때 집합 A와 B는 보통집합(crisp set)이다.일반적으로 연결된(connected) 유향그래 프(directed graph)를 망(network)이라고 한다. 이 망도 자연스럽게 퍼지화 될 수 있다. 이 퍼지화 역시 퍼지그래프에서처럼 첫째, 연결선(edge or flow)의 퍼지화, 둘째, 정점(node)들의 퍼지화가 가능하다.




(a) 퍼지망(퍼지 연결선)




(b) 퍼지망(퍼지정점, 퍼지 연결선)

그림 3. 퍼지망

1) 퍼지망의 정점의 보통집합을 V, 연결선(관계)의 퍼지집합을 R이라 해보자. 그리고 퍼지망에서 정점 xi1에서 xir사이의 경로를 Ci라 해보자.

$$C_i = (x_{i1}, x_{i2}, \ldots, x_{ir}), \quad x_{ik} \in V,\ k = 1, 2, \ldots, r$$

이때

$$\forall(x_{ik}, x_{ik+1}),\ \mu_R(x_{ik}, x_{ik+1}) > 0,\ k = 1, 2, \ldots, r-1$$

이와 같은 경로 Ci에 대하여 퍼지값 을 다음과 같이 부여할 수 있다.

$f(x_{i1},x_{i2},\ldots,x_{ir}) = \mu_R(x_{i1},x_{i2})\quad R(x_{i2},x_{i3}) \wedge \ldots \wedge \mu_R(x_{ir}-1,x_{ir})$

$\mu_R(x_{ir}-1,x_{ir})$

이 값은 $x_{i1}$에서 $x_{ir}$까지 갈 수 있는 최소한의 가능성을 나타낸다. 만약 $x_{i1}$에서 $x_{ir}$ 사이에 여러개의 가능한 경로가 있다고 해보자. 그러면 이 가능한 경로의 집합을 다음과 같이 나타낼 수 있다.

$C(x_i,x_j) = \{c(x_i,x_j)\ c(x_i,x_j) = (x_{i1} = x_i,\ x_{i2},\ \ldots,\ x_{ir} = x_j)\}$

이 가능한 모든 경로중에서 연결이 가장 강한 경로의 값(최대값의 경로) $f^*$은 다음과 같이 얻을 수 있다.

$$f^*(x_i,x_j) = \bigvee_{C(x_i,x_j)} (x_{i1} = x_i,\ x_{i2},\ \ldots,\ x_{ir} = x_j)$$

2) 퍼지망에서 정점들의 퍼지집합 V, 연결선의 퍼지집합 R을 생각해 보자. 이제 $x_{i1}$에서 $x_{ir}$까지의 경로 $C_i$는 다음과 같다.

$C_i = (x_{i1},\ x_{i2},\ \ldots,\ x_{ir}),\quad x_{ik}\quad V,\ k = 1,\ 2,\ \ldots,\ r$

이때

$\forall(x_{ik},\ x_{ik+1}),\ \mu_R(x_{ik},\ x_{ik+1}) > 0,\ k = 1,\ 2,\ \ldots,\ r-1$

$\forall x_{ik},\ \mu v(xik) > 0,\ k = 1,\ 2,\ \ldots,\ r-1$

이 경로의 값은 다음과 같다.

$(x_{i1},x_{i2},\ldots,x_{ir}) = \mu_R(x_{i1},x_{i2})\ \mu_R(x_{i2},x_{i3})\ \ldots\mu_R(xir_{-1},x_{ir})$

$\wedge\mu v(x_{i1}) \wedge\mu v(x_{i2}) \wedge \ldots\mu v(x_{ir})$

이제 앞에서와 같이 $f^*$을 구할 수 있다.

## 3. 퍼지관계 측면에서의 영역지식

대부분의 정보검색시스템이 채택하고 있는 검색추론과정은 일반적으로 용어의 통계적 속성에 기초한 것이다. 질문 및 문헌에서 발췌한 색인어에 의해 문헌집합이 주어지며, 확률적인 모델에 기초를 둔 검색알고리즘은 질문에 대한 관련성 확률을 추론하여 순위를 부여한다. 이러한 관련성 확률은 관련 및 비관련 문헌에서의 색인어의

각 용어의 발생빈도나 용어간에 존재하는 통계적 의존성을 주제영역에 있어서의 지식으로 이용하였다.

본 연구에서는 영역지식으로서 추출된 색인어들 사이의 퍼지관계를 활용한 것으로서, 색인어 퍼지관계행렬은 색인어들과 이들 사이의 퍼지관계들로

| | Virus | 유산균 | Acryl-amide | Biorea-ctor | 콜레스테롤 | Escherich-iacoli | ... |
|---|---|---|---|---|---|---|---|
| | $t_1$ | $t_2$ | $t_3$ | $t_4$ | $t_5$ | $t_6$ | ... |
| t1 | 1.0 | 0.0 | 0.5 | 0.3 | 0.3 | 0.5 | |
| t2 | | 1.0 | 0.5 | 0.8 | 0.6 | 0.5 | |
| t3 | | | 1.0 | 0.2 | 0.2 | 0.3 | |
| t4 | | | | 1.0 | 0.2 | 0.8 | |
| t5 | | | | | 1.0 | 0.4 | |
| t6 | | | | | | 1.0 | |
| . | | | | | | | |
| . | | | | | | | |
| . | | | | | | | |

표 2. 유전공학분야 색인어간 관계도

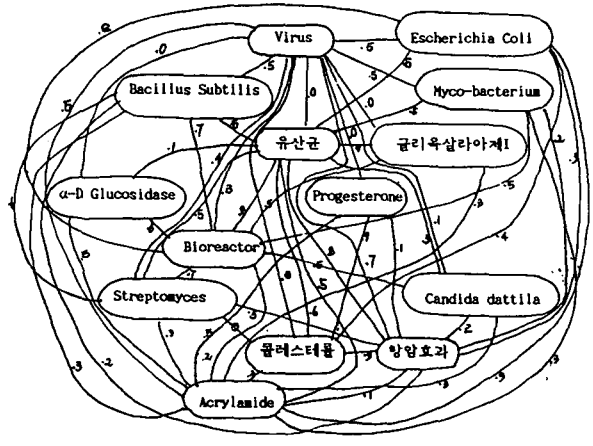



그림 4. 그래프와 망으로 표현한 영역지식

표현된다. 여기서 관계값은 색인어들 사이의 개념적 유사도 즉 관련도를 나타내 주고 있다. 이러한 의미에서 이 행렬은 색인어들 사이의 관계를 나타내는 퍼지디소러스의 일종이라고 할 수 있을 것이다.

실제로 관계값들은 관련 용어들에 있어서의 관련도에 따라서 주어진다. 색인어 관계행렬은 표 2.에서 보는 바와 같이 K x K 행렬 W로 표현할 수 있으며, 여기서 K는 색인어의 숫자와 같은 크기를 갖는다. 그림 4.는 유전공학분야의 연구논문들로부터 추출한 색인어들 사이의 관계를 그래프와 네트웍으로 표현한 것이다. 관계값들은 [0,1] 사이에서 주어지는데, 여기서 0은 색인어 사이에 아무런 관계가 없음을 의미하고 1은 논리적으로 가능한 최대값을 의미한다.

한편, 문헌들에 있어서 두개의 색인어가 동시출현하는 빈도수가 많을 수록 이들 색인어 사이의 관계가 높다는 가정하에 자동으로 색인어들 사이의 관계값을 부여할 수 있는 환경에서 초기관계값들을 부여한 연구결과들도 다수 발표된 바 있으나[Doyle; S. Miyamoto et al; T. Murai et al; Y. Ogawa et al], 본 연구에서는 실험환경을 여기까지 확장하지는 못하였다.

(참고): i번째 색인어와 j번째 색인어 사이의 초기관계값 $W^*_{ij}$

$$W^*_{ij} = \begin{cases} \dfrac{N_{ij}}{N_i + N_j - N_{ij}}, & i \neq j, \\ 1, & i = j, \end{cases} \qquad (1)$$

여기서, $N_{ij}$는 i번째와 j번째 색인어들을 모두 포함하는 문헌들의 갯수이며, $N_i$와 $N_j$는 각각 i번째 색인어와 j번째 색인어만을 포함하고 있는 문헌들의 갯수이다. 이 공식 (1)에 의해서 두개의 색인어가 동시에 포함된 문헌들의 정규빈도수가 결정된다[Y. Ogawa:165].

본 연구를 통해 수행하고자 하는 정보검색실험은 이러한 원칙들 즉, 색인어 사이의 관계, 색인어와 문헌 사이의 관계, 그리고 검색된 문헌들의 적합도 서열 등을 퍼지관계와 퍼지논리의 측면에서 부여함으로써 정보검색의 과정을 보다 일반화하고자 하는 궁극적 목표

를 염두에 둔 실험적 시도이며, 본 연구에서는 우선적으로 용어들간의 관계를 퍼지관계로 규정함으로써 정보검색에 있어서 또 하나의 접근방법을 모색하고 그 가능성을 모색하였다.

## 4. 퍼지관계행렬을 이용한 검색

### 4.1. 퍼지검색방법의 개관

퍼지문헌검색시스템은 다음과 같이 정의할 수 있다[Buell:398].

$D = \{D_i\}$ : 문헌집합

$T = \{T_i\}$ : 용어집합

퍼지소속함수 $f:D \times T \quad [0,1]$.

문헌과 색인어 사이의 관계함수 $f(d_i, t_i)$를 통해서는 문헌 $d_i$와 용어 $t_i$ 사이의 관계도를 측정한다. 또한 소속함수(membership function:MF)는 다음 규칙과 함께 모든 불리안 수식으로 확장될 수 있다[Buell & Kraft:127].

$F(d, t \text{ AND } t') = \text{Min}\,(F(d,t), F(d,t'))$,

$F(d, t \text{ OR } t') = \text{Max}\,(F(d,t), F(d,t'))$,

$F(d, \text{NOT } t) = 1 - F(d,t)$.

이용자는 그가 검색하고자 하는 주제어를 제시한다. 이 질의식은 색인어들과 AND, OR, NOT 등의 논리연산자들로 구성된다. 질의식은 불리안 연산의 기본 규칙들을 반복적으로 적용함으로써 연결형 공식으로 변환될 수 있는데, 이 공식은 OR와 NOT 만을 포함하는 부질의식(subquery)들로 구성된다. 연결형 공식으로 된 질의식은 다음과 같이 기술할 수 있다.

$\text{Query} = \text{SubQuery}(1) \wedge \cdots \wedge \text{SubQuery}(N)$.

$\text{SubQuery}(h) = K_1 \vee \cdots \wedge Kn_h \vee \neg Kn_{h+1} \vee \cdots \wedge \neg Kn_{h+m_h}$,

여기서 $\wedge, \vee, \neg$ 은 각각 불리안 연산자인 AND, OR, NOT을 가리킨다.

$K_i$는 질의식 내의 i번째 용어를 나타낸 것이다. 질의식에 있어서는 $N \geq 1$이며, h번째 부질의식은 $n_h \geq 0$, $mh \geq 0$ 그리고 $nh + mh_1$의 조건을 가진다. h번째 부질의식은 두개의 집합 Q(h)+ 와 Q(h)-에 의해 표현할 수 있는데, 여기서 Q(h)+는 연산자 NOT을 갖지 않은 용어집합을 의미하고 Q(h)-는 연산자 NOT을 포함하고 있는 용어집합을 의미한다. 따라서 어떤 용어도 Q(h)+ 와 Q(h)-에 동시에 포함되지 못한다.

기존의 일반적인 검색에 있어서의 검색결과는 다음과 같이 표현할 수 있는데, h번째 부질의식에 대한 검색결과는 다음과 같다.

$$SubResult(h) = D(K_1) \cup \ldots \cup D(Kn_h) \cup \overline{D(Kn_h+_1)} \cup \ldots \cup \overline{D(Kn_h+m_h)}$$

여기서, D(K)는 색인어 K를 갖는 문헌집합을 의미하고, $D_1 \cup D_2$는 두개의 집합 $D_1$과 $D_2$의 합집합이며, D는 집합 D의 여집합을 나타낸다. 질의식의 OR연산자는 해당 문헌집합들의 합집합을 구하는 연산자이며, NOT연산자는 여집합을 구하는 연산자이다. 따라서 검색결과는 다음과 같은 식으로 표현할 수 있다.

$$Result = SubResult(1) \cap \ldots \cap SubResult(N)$$

여기서 D1 D2는 두개의 집합 $D_1$과 $D_2$의 교집합이다. 질의식에 들어 있는AND 연산자는 해당 문헌집합들의 교집합을 구하는 연산자이다. 그리고 검색결과는 집합의 구성요소들이 정확히 질의식과 일치하는 일반집합(crisp set)으로 나타난다.

퍼지검색시스템에 있어서의 방법론은 검색결과가 퍼지집합이라는 사실만 제외한다면 일반집합이론에 따른 검색방법과 동일하다. 일반적으로 문헌집합 D(K)는 퍼지집합이며, 검색결과는 퍼지합집합, 퍼지교집합, 퍼지여집합 계산방법을 적용하여 퍼지적으로 추출하는 방법을 사용하고 있다.

그러나 본 연구에서는 문헌집합 D(K)를 일반집합으로 하고 퍼지디

소러스에 있어서와 유사한 방법으로 색인어간 퍼지관계행렬을 이용하는 방법을 써서 검색하는 방법을 취하였다[T. Murai et al; Y. Ogawa et al]. 각 문헌의 소속함수값(membership value)은 검색문헌으로서의 적합도(relevance as a retrieved document)를 나타내게 된다. 따라서 i번째 문헌 di의 소속함수값은 간단히 $r_i$로 표시한다 ($r_i = \mu_{Result}(d_i)$).

검색결과로서 각 문헌들은 소속함수값(적합도)에 따라서 내림차순으로 정렬되게끔 하였다. 그리고 검색자는 검색된 문헌들의 소속함수값에 기준치를 부여하는 방법($\alpha$-cut)과 검색대상문헌의 갯수를 통제하는 방법을 통해서 요구되는 적합문헌들을 검색할 수 있게 하였다[Y. Ogawa et al].

4.1.1. 퍼지질의식의 처리

불리안 검색시스템이 지니고 있는 중요한 특징중의 하나는 질의내용을 다른 편리한 형태로 변환할 수 있는 능력이 뛰어나다는 것이다. 이용자의 입장에서는 질의식을 통해 중요한 개념들 사이의 논리적 관계를 표현한다. 이용자는 자신의 입장에서 논리적이고 자연스러워 보이는 개념으로 질의내용을 표현함과 아울러, 정확성의 관점에서 자신이 표현하는 개념의 중요성을 표현할 수 있어야 한다. 또한 시스템도 입력된 질의어를 원래의 의도를 변형시키지 않고 유지하면서, 처리하기 편리한 다른 형태로 조직적으로 변환시킬 수 있어야 한다.

북스타인은 퍼지질의어를 동등한 다른 형태로 변환시킬 수 있는 규칙들을 퍼지집합이론에서의 연산방법에 준거하여 작성한 바 있다[Bookstein, 1980: 245-246; 1981: 276].

(1) 교환법칙(commutativity): "E AND F"와 "F AND E"는 동일하다. E 와 F가 가중치를 갖고 있는지의 여부와는 상관이 없다.

(2) 결합결합법칙(associativity): E AND (F AND G) = (E AND F)

AND G E OR (F OR G) = (E OR F) OR G.

(3) 분배법칙(distributivity):가중치를 가진 퍼지집합들도 퍼지집합이라는 전제하에 $E^a \cap (F^b \cup G^c) = (E^a \cap F^b) \cup (E^a \cap G^c)$ 이다. 여기서 $E^a \cap F_b$는 멤버쉽함수 $\min(fE^a, bfF) = b.\min(b^{-1}fE^a, fF)$와 같다. 그리고 이것은 다시 $b^{-1}fE^a(x) \leq 1$일 때, 점 x에 대하여 $b\ \min(fE^{ab}, fF)$와 동일하다. 그러나 $b^{-1}fE^a(x) > 1$이면, $b^{-1}fE^a(x) \geq fF(x)$이며, 이 경우 $b\ \min[b^{-1}fE^a(x), fF(x)] = bfF(x) = b\ \min[1, fF(x)] = b\ \min\ [fE^{ab}(x), fF(x)]$이다. 어떤 경우이건, $fE^a \cap Fb = b\ \min(fE^{ab}, fF)$임을 증명할 수 있으며, 후자는 집합$(E^{ab}, F)_b$와 일치한다.

따라서 $E^a \cap (F_b \cup G_c) = (E^{ab} \cap F)_b \cup (E^{ac} \cap G)_c$ 를 정의할 수 있으며, 마찬가지로 다음과 같은 정의를 내릴 수 있다.

$E_a \cup (F^b \cap G^c) = (E_{ab} \cap F)^b \quad (E_{ac} \cup G)^c$.

또한 이를 검색에 필요한 질의식으로 고친다면,

$E^a$ AND $(F_b$ OR $G_c)$ = $(E^{ab}$ AND $F)_b$ OR $(E^{ac}$ AND $G)_c$ 과

$E_a$ OR $(F^b$ AND $G^c)$ = $(E_{ab}$ OR $F)^b$ AND $(E_{ac}$ OR $G)^c$ 으로 정할 수 있다.

이러한 논의는 예를 들면, E AND (F OR G)과 같은 질의식의 경우에 (E AND F) OR (E AND G)과 같은 결과를 내게 된다는 것을 의미한다.

(4) 드 모르강의 법칙(duality; De Morgan's law): $\overline{E \cap F} = \bar{E} \cup \bar{F}$. 이 논리를 검색식에 적용시킬 경우에 가중치를 가진 경우라도 NOT $(E^a$ AND $F^b)$ = $(NOT_a$ E) OR $(NOT_b$ F)과 NOT $(E_a$ OR $F_b)$ = $NOT^a$ E AND $NOT^b$ F이 성립한다.

(5) Idempotency: $E \cap E = E \cup E = E$. 이를 가중치를 고려한 보다 일반적인 예로 만들어 보자면,

$E^a$ AND $E^b = E^{max(a,b)}$ 와

$E_a$ OR $E_b = E_{max(a,b)}$.

NOT E의 형식을 가진 표현식의 경우에도 같은 관계가 성립된다.

$(NOT_a\ E)\ OR\ (NOT_b\ E) = NOT_{max(a,b)}\ E,$

$(NOT_a\ E)\ AND\ (NOT_b\ E) = NOT^{max}{}_{(a,b)}\ E.$

이들 관계들을 통해서 볼 때, 중복된 용어들을 제거함으로써 불리안 표현식을 단순화시킬 수 있다.

(6) 가중치 분배법칙(weight distributivity): 어렵지 않게 다음 내용을 확인할 수 있을 것이다. 즉,

$(E\ OR\ F)_a = E_a\ OR\ F_a$

$(E\ AND\ F)_a = E^a\ AND\ F^a.$

그러나 이들 관계는 (E OR F)a와 (E AND F)b에서는 성립하지 않는다. 그러나 불리안 표현식을 단순화할 수 있는 방법으로서 다음과 같은 실례를 들 수 있다.

$(E\ OR\ F)^a\ AND\ G = (E^a\ AND\ G)\ OR\ (F^a\ AND\ G).$

$(E\ AND\ F)_a\ OR\ G = (E_a\ OR\ G)\ AND\ (F_a\ OR\ G).$

(7) Involution: NOT (NOT E) = E. 이를 일반화하면,

$NOT\ NOT_a\ E = E^a$와

$NOT\ NOT^a\ E = E_a$가 된다.

(8) 가중치의 처리

a. $(E_a)_b = E_{ab} = (E_b)_a$

b. $(E^a)^b = (E^b)^a = E^{ab}$

c. $(E) = E_a/_b$ if $a \le b$,

$E^b/^a$ if $a \ge b$.

## 4.2. 퍼지관계행렬을 이용한 적합도값(소속함수값) 산정

적합도값 즉, 소속함수값은 다음과 같은 절차를 통해서 계산한다. 먼저 퍼지 관계행렬을 통해 각 부질의식(subquery)에 대한 소속함수값을 계산한 다음, 전체적인 소속함수값을 계산한다.

값을 계산한 다음, 전체적인 소속함수값을 계산한다.

4.2.1. 퍼지색인의 작성

$A_i$를 i번째 문헌에 대해 색인된 색인어집합이라고 하자. 본 연구에서는 문헌의 표제들로부터 불용어들을[Rijsbergen:18-21;안현수:10-14] 제거한 후, 남은 용어들을 색인어 관계행렬상의 용어들과 비교한 다음 색인어 관계행렬을 이용해 해당 문헌의 적합도를 계산하는 방법을 택하였다. 여기서 불용어(stop word)란 검색에 불필요하면서 대체적으로 빈도가 높은 단어들로서 우리말의 경우 그 유형을 구분하면 다음과 같다[안현수:10].

첫째, '우리', '이', '이들', '그들', '등', '것', '상', '수'와 같은 대명사나 불완전명사 따위의 기능어들

둘째, '연구', '고찰', '문제', '큰', '높은' 등과 같이 내용규정에 관계가 없는 명사, 형용사, 부사류

세째, '본', '나타난', '되는' 등과 같은 동사와 그의 변화형들.

한편, i번째 문헌과 j번째 색인어사이의 관계값 $R_{ij}$는 다음과 같이 정의된다[Y. Ogawa et al:167].

$$R_{ij} \quad \underset{K_k \; A_i}{} \quad W_{jk} \tag{2}$$

여기서 $W_{jk}$는 색인어 관계행렬 내에서 j번째와 k번째 색인어사이의 관계값을 의미하며 는 식 ${}_iX_i = 1 - \Pi_i(1-X_i)$로 정의되는 대수합을 나타낸다. 따라서 공식 (2)는 다음과 같이 정리된다.

$$R_{ij} = 1 - \underset{K_k \in A_i}{\Pi_j} ( 1 - W_{jk})$$

따라서 $R_{ij}$는 i번째 문헌과 j번째 색인어사이의 관계값으로서 퍼지적 소속함수값이 된다.

4.2.2. 부질의식의 관계값 계산

퍼지집합이론에서 두 집합 A와 B의 합집합은 A U B로 표시하고, 이때 어느 원소 X의 소속함수값은 X가 A와 B에 포함될 가능성 중에

큰 것을 취한다. 즉,

$$\mu_{AUB}(X) = Max\,[\mu_A(X),\ \mu_B(X)],\ \forall X \in X.$$

이와 같이 하면 당연히 A와 B는 A U B의 부분집합이 된다. 본 연구에서는 퍼지집합에 있어서 대수합 계산방법을 적용하여 다음과 같은 식에 의해 합집합값을 구하였다. 대수합 연산자( )는 퍼지 합집합을 위한 Max연산자에 비해 A와 B사이의 교환성이 크다고 할 수 있다[오길록:3-16].

$$\mu_{A\ B}(X) = \mu_A(X) \quad \mu_B(X) = 1 - \mu A(X) \cdot \mu_B(X)$$

여기서 $\mu A(X)$와 $\mu B(X)$는 퍼지집합 A와 B에 있어서 원소의 소속함수값을 나타낸다. 퍼지집합 A의 여집합은 다음과 같은 식에 의해서 정의될 수 있다.

$$\mu_{\bar{A}}(X) = 1 - \mu_A(X)$$

부질의식에 대한 관계값은 다음 식을 통해서 계산된다.

$$r_i(h) = (\,{}_{K_j \in Q(h)^+} R_{ij})_{K_j}\ \{\,{}_{Q(h)^-}\ (1 - R_{ij})\}$$

$$= 1 - (\prod_{K_j \in Q(h)^+} Sij\ )(\prod_{K_j \in Q(h)^-} Rij) \qquad (3)$$

여기서 Sij는 다음 식으로 정의된다.

$$S_{ij}\quad 1 - R_{ij} = \prod_{K_k \in A_i} (1 - W_{jk})$$

이 결과는 W가 단위행렬일 때 일반 집합론의 계산방법을 사용했을 때의 값과 동일하다. $Q(h)_+$나 $Q(h)_-$가 공집합( )인 경우에는 관계값은 다음 식에 의해서 계산된다.

$$r_i(h) = 1 - \prod_{K_j \in Q(h)^+} Sij : \ Q(h)^- = \phi$$

$$ri(h) = 1 - \prod_{K_j\ Q(h)_-} R_{ij} : \ Q(h)^+ = \phi \qquad (4)$$

4.2.3. 질의식 전체의 관계값 계산

부질의식들에 대한 관계값이 결정되면 전체 관계값을 계산한다. 퍼지 교집합을 구할 때 소속함수값에 단순곱(product)을 적용할 수 있다. 이렇게 하여 얻은 교집합을 A · B라 하면 퍼지 교집합은 다음과 같이 정의된다.

$$\mu_{A \cdot B}(X) = \mu_A(X) \cdot \mu_B(X), \quad \forall X \in X.$$

여기서 연산자 · 는 교환법칙, 결합법칙, 항등법칙, 드모르강법칙 등을 만족하고 또한 $A \cdot {}^{2} = {}^{2}$ 을 만족한다. 이 연산자 역시 Min 연산자에 비해 A와 B 사이의 교환성이 크다고 할 수 있다[오길록:3-23]. 따라서 Min연산자 대신에 교집합연산으로서 대수적(확률적)을 사용하였다. 따라서 i번째 문헌의 관계값은 다음 식에 의해서 계산이 된다.

$$r_i = \prod_{h=1}^{n} R_i(h)$$

W가 단위행렬일 때, 위 식에 의한 퍼지결과값은 일반집합론의 계산방법을 사용했을 때의 값과 동일하다.

## 4.3. 질의식의 수정

정보검색과정은 통제된 불확실성의 범주에서 질의어를 포함하는 문헌집합을 찾는[Chiaramella & Defude:286], 그리고 상호작용적이면서 비결정론적인 과정이며, 따라서 이용자로 하여금 더 양호한 질의식을 작성할 수 있도록 지원하는 것이 매우 중요한 시스템요소이다[Y.Ogawa:168]. 색인어 관계행렬은 두 용어 사이의 유사도를 나타낸 것으로서 질의식을 수정하는 데에도 활용할 수 있다. 또한 색인어와 질의식 사이의 관계의 정도도 이 행렬을 통해서 계산할 수 있다.

질의식을 형성하는 과정에서 이용자는 시스템으로 하여금 색인어

와 질의식과의 관계도 순서에 따라서 색인어들을 열거하도록 요청할 수 있다. 관계도는 색인어로서의 적합도라고 할 수 있는 것으로서, 한 문헌의 적합도를 계산하는 방법과 같은 방법으로 계산할 수 있다. i번째 색인어의 적합도값 $T_i$는 다음 식과 같이 계산된다.

$$T_i(h) = 1 - \left( \prod_{K_j \in Q(h)} W_{ij} \right) \left\{ \prod_{K_j \in Q(h)} \left( \frac{1}{} - W_{ij} \right) \right\},$$

$$Ti = \sum_{h=1}^{N} T_i(h).$$

이용자가 관련된 색인어들을 제시하면 모든 색인어들의 적합도값이 계산되고, 색인어들은 그 값에 따라서 내림차순으로 배열된다.

## 5. 시스템의 구현

정보검색시스템이 기본적으로 갖추어야 할 요소들로는 문장구조분석프로그램, 문헌정보DB, 조직화된 검색알고리즘, 이용자와 시스템 사이의 상호작용을 가능케 하는 장치 등을 들 수 있다[Belkin et al:160]. 그리고 여기에 추가할 수 있다면, 이들 요소들이 상호 유기적으로 연결되고 각 하부기능들이 기능을 수행하는 과정에서 지능적인 작동원리에 따라서 운용될 수 있게 하는 것이 필요할 것이다.

본 연구에서는 실험적 시스템으로 다음과 같은 환경을 갖추도록 하였다. 워크스테이션으로는 IBM 호환기종인 386DX급 개인용컴퓨터를 사용하였는데, 본 시스템의 주기억장치 용량은 2 메가바이트, 보조기억장치 용량은 40메가바이트이다. 문헌데이터는 유전공학분야에서 생산된 연구논문 120건을 입력하여 DB로 활용하였으며, 이 논문들의 제목으로부터 불용어들을 제거한 다음, 추출된 용어 187개를 리스트로 작성하여 해당분야 전문가의 자문을 받아 퍼지 관계행렬로 작성하였다. 한글 색인어와 영문 색인어는 퍼지관계행렬을 통해 자연스럽게 관계로서 연결되므로 역어사전을 별도로 작성할 필요는 없

었다.

본 시스템은 주요 기능으로서 퍼지문헌검색과 관련색인어검색 두 가지 기능을 수행하도록 되어 있다. 퍼지문헌검색기능을 통해서는 적절한 수의 적합문헌을 검색하도록 하였으며, 관련색인어검색을 통해서는 이용자가 문헌을 검색함에 있어서 관련 용어들을 제시해 줌으로써 질의식을 수정할 수 있도록 하였다.

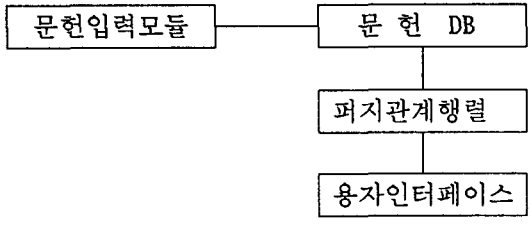



그림 5. 시스템 구성도

## 6. 검색결과의 평가

### 6.1. 평가방법

검색시스템의 효율은 이용자의 정보요구에 적합한 문헌을 검색해 내는 검색시스템의 능력을 의미하는 것으로 검색된 적합문헌과 부적합문헌, 검색되지 않은 적합문헌과 부적합문헌 사이의 비율로서 측정된다[사공철 등:213].

정보검색시스템에 있어서 검색성능 평가의 일반적인 방법은 정확률과 재현율을 측정하는 방법을 취한다. 재현율(Recall Ratio)과 정확률(Precision Ratio)은 각각 다음과 같이 구해진다.

$$RR = \frac{\text{검색된 적합문헌수}}{\text{전체 적합문헌수}}$$

$$PR = \frac{\text{검색된 적합문헌수}}{\text{검색된 전체문헌수}}$$

여기서 재현율(RR)은 얼마나 많은 적합문헌이 검색되었는가를 나타내는 것이고, 정확률(PR)은 검색된 문헌중 적합문헌이 얼마나 되는가를 의미한다.

한편, 코헨과 켈젠(P.R. Cohen R. Kjeldsen)은 검색결과의 평가에 있어서 누락률(fallout ratio)을 적용한 결과를 제시하고 있는데 [Cohen & Kjeldsen:261], 그 식은 다음과 같다. 그러나 본고에서는 그 내용이 해당분

$$\text{fallout} = \frac{\text{number of documents judged good by the system, bad by expert}}{\text{number of documents judged good by the system}}$$

야 전문가의 판단에 의해 재현율에 포함되는 것으로 보고 그 결과를 평가대상에 포함시키지는 않았다.

이들 결과들을 계산하기 위해서 실험용으로 작성한 세개의 질의식들에 대한 적합문헌들을 해당분야 전문가들이 수작업으로 추출하였다. 그리고 시스템상에서 해당 질의식들에 대해 검색을 수행하였다. 그리고 나서 재현율과 정확률을 계산하였다. 여기서 재현율과 정확률은 일반집합이론에 따라서 정의된 것이기 때문에 퍼지검색시스템의 검색결과를 일반집합이론적인 것으로 변환시키는 작업이 필요하다. 본 연구에서는 성능평가를 위해 3장에서 설명한 바와 같은 적합값 산출방법에 따라서 기준값을 적용하는 방법을 취하였다.

본 실험에서는 120개의 문헌을 대상으로 문헌 DB를 구축하고 검색을 수행한 결과를 평가하였다. 문헌표제를 대상으로 하여 불용어처리를 한 다음 추출된 187개의 색인어 중에서 해당분야 연구자들로 하여금 색인어적 가치가 있다고 판단된 용어들을 선별하여 퍼지 색

인어 관계행렬을 작성하였으며, 그 작성방법은 3장에서 설명한 바와 같이 전문가들의 협조를 얻어 수작업으로 수행하였다. 따라서 색인어들 사이의 퍼지관계를 자동으로 부여함으로써 시스템에 지능을 부여한다는 측면에서 본 연구의 차기 연구과제가 남아 있다고 할 수 있다.

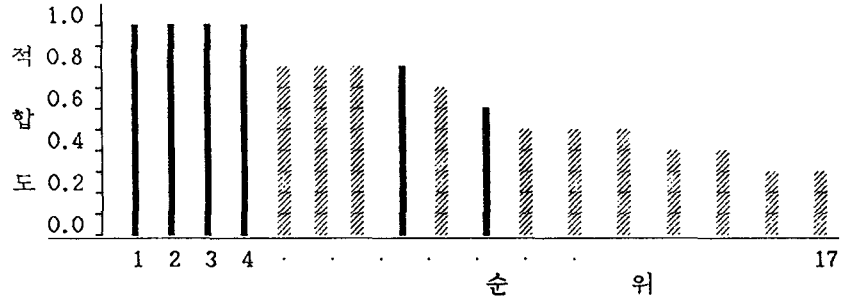



그림 6. 색인어 1에 대한 검색결과(적합도:순위)

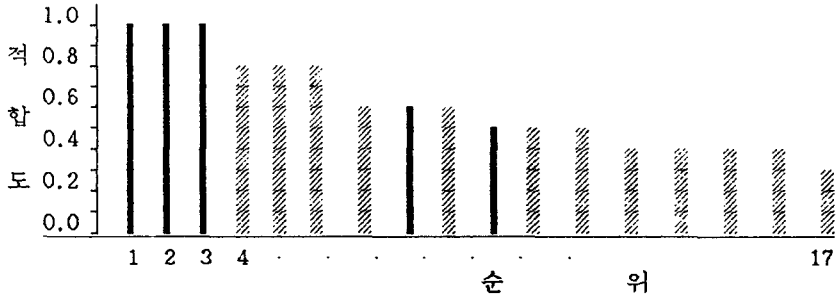



그림 7. 색인어 2에 대한 검색결과(적합도:순위)

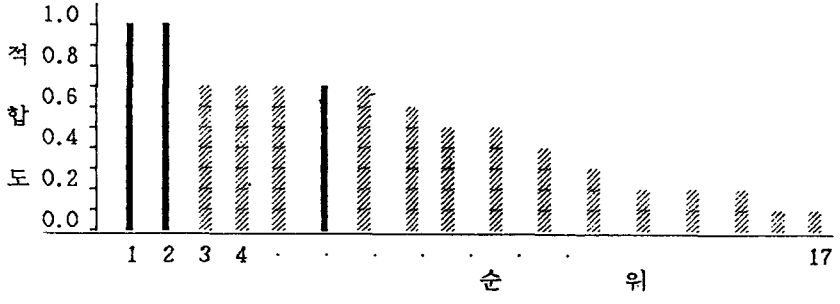



그림 8. 색인어 3에 대한 검색결과(적합도:순위)

## 6.2. 검색성능의 평가

그림 6,7,8은 각각 색인어 1,2,3에 대한 검색결과를 나타낸 것이다. 검은색의 막대는 적합문헌을 나타내는 것이고 사선막대는 부적합문헌을 각기 나타낸 것이다.

그림 9.는 질의식(색인어 1)에 대한 기준값과 재현율, 정확률 사이의 관계를 나타낸 것이다. 재현율과 정확률의 정의방법에 기술된 바와 같이 이들은 서로 반비례관계에 있다.

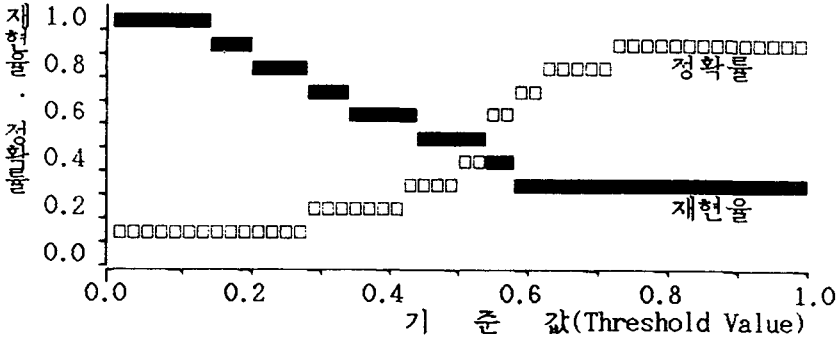



그림9. 질의식(색인어 1)에서 기준값과 재현율, 정확률 사이의 관계

다음으로, 본 연구에서는 퍼지관계행렬을 이용한 검색과 일반 집합이론에 따른 검색결과를 비교하였다. 일반적으로 재현율과 정확률을 동시에 증가시켜주는 최적기준값(optimal threshold value)은 그림 9.와 10.에서볼 수 있는 바와 같이 질의식에 따라서 달리 나타난다. 그림 10.은 질의식 (색인어1 색인어2)에 대한 검색결과인데, 정확률은 기준값 0.7 이상의 범위에서는 정확률이 계산되지 않으며, 이는 최대적합도가 0.7이기 때문에 그 이상의 기준값을 만족하는 문헌들이 없기 때문이다. 따라서, 기준값을 고정시키는 것은 부적합하며 이 값은 문헌과 적합도에 따라서 역동적으로 변동하

는 것이어야 한다[Y.Ogawa:176]. 기준값 는 다음과 같이 계산되도록 하였다.

$$\alpha = \mu x \frac{\text{전체 적합도 총계}}{\text{0 이상의 적합도를 갖는 문헌의 수}}$$

여기서 는 기준계수(threshold coefficient)로서 1.6의 값을 주었다[Y. Ogawa:176]. 본 연구에서는 색인어1, 2, 3을 개별적인 질의식으로 한 검색과 이들 색인어들을 AND, OR 를 통해 조합한 질의식과 AND와 NOT을 동시에 사용하여 두개의 색인어를 조합한 질의식에 의해서 실험을 수행하였다. 그 결과 이들 질의식에 대한 재현율과 정확률 평균치는 각각 평균재현율이 0.53이었고 평균정확률은 0.42이었다. 이 결과는 일반 집합이론에 의한 방법에 있어서 평균재현율

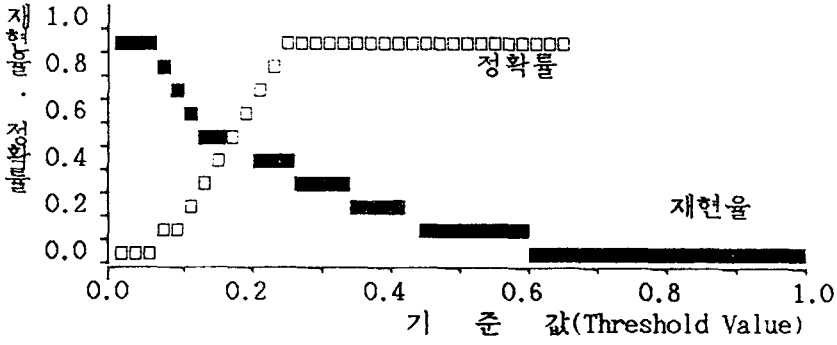



그림 10. 기준값과 재현율, 정확률 사이의 관계

(질의식 = 색인어1∧¬ 색인어2)

0.43과 평균정확률 0.45와 비교해 볼 때, 정확률에 있어서는 거의 비슷하다고 할 수 있으나 재현율에 있어서는 상당히 큰 폭으로 증가하였음을 확인할 수 있었다.

## 7. 결론 및 과제

정보검색에 있어서 이루어진 이론적인 발전사항들의 혜택을 받아 전통적인 불리안검색에 대한 다수의 대안들이 제시되었다. 그 중에서도 확률이론과 퍼지집합 이론은 기중 돋보이는 역할을 수행했다고 할 수 있을 것이다[Bookstein, 1985:117]. 퍼지집합이론은 전통적인 집합이론을 일반화하여 집합내에 부분적인 소속함수를 도입한 것이다. 이는 검색환경에서 색인용어와 문헌, 문헌과 문헌 사이의 관계를 정할 때 이치관계(0 또는 1)가 아닌 부분적 소속관계(0-1)를 적용할 수 있음을 의미한다. 퍼지집합이론을 적용한 검색시스템에서는 전통적인 불리안시스템들이 적용하고 있는 검색방법들을 대부분 활용하고 있다.

현재 운용되고 있는 범용 검색시스템들은 주로 불리안구조를 채택하고 있고, 시스템들의 기초가 되는 이론적인 모델은 문헌과 이용자 정보요구가 색인어집합과 부울탐색을 위한 요구공식으로 각각 정확히 그리고 완전하게 특징지울 수 있다는 불분명한 가설에 근거하고 있다. 그러나 이러한 가설에서와 같이 문헌검색과정에 있어서 불확실성이란 아주 본질적인 것이므로 정확성이 결여되어 있는 것이 사실이다. 검색과정이 가지고 있는 본질적인 오류성을 효과적으로 처리할 수 있는 표준 불리안모델이 없다는 사실이 현재 운용되고 있는 검색시스템들이 지니고 있는 여러가지 심각한 결함들의 주요 원인이 되고 있으며[Radecki,1988], 실제로 불리안 시스템에 있어서의 결함사항들로는 첫째, 시스템에 익숙치 않은 이용자에게는 복잡하다는 점, 둘째, 불리안 질의식에 대한 검색결과는 지나치게 많을 수도 있고 또는 지나치게 적을 수도 있다는 점, 세째, 검색결과에 있어서 서열이 매겨지지 않는 점, 네째, 불리안 언어의 표현력이 제한되어 있는 점 등이 지적되고 있다[Cooper,1983:32].

한편, 이러한 문헌검색시스템을 일반화하는데 있어서의 네가지 수

준은 다음과 같다[Kraft & Buell, 1983:48-49].

1) 불리안 색인과 불리안 질의
2) 퍼지 색인과 불리안 질의
3) 불리안 색인과 퍼지 질의
4) 퍼지 색인과 퍼지 질의

본 연구에서는 퍼지집합이론을 적용한 문헌검색시스템을 구현함으로써 용어와 이를 통해 표현된 질의식과 문헌간의 관계를 일반화하고자 하였으며, 이것은 색인용어와 문헌에 관계값으로서 이치관계가 아닌 퍼지관계를 채택함으로써 수행하였다.

이러한 퍼지집합이론 및 퍼지논리에 근거한 문헌검색시스템의 장점은[Cooper, 1983:48-49] 문헌을 표현하는 색인어의 다양한 중요도를 고려할 수 있으며, 문헌은 가중치가 부여된 색인어로 색인된 문헌검색시스템에 있어서 색인용어들을 통해 이론적으로 정당화된 부울 질의식을 대표질의어로서 사용할 수 있고, 집합이론과 2치 논리에 기반을 둔 문헌검색방법은 퍼지집합이론과 퍼지논리에 근거한 방법들의 특수한 경우이므로 더 일반적인 문헌검색이론을 개발하는 것이 가능하다는 점 등을 들 수 있다. 검색의 기법이라고 하는 측면에서 이러한 연구들이 주로 공헌한 점은 부울 질의어와 서열기법을 병합한 것이라고 할 수 있으나, 벡터공간모델이나 확률모델에 있어서 용어의존도를 이용한 확대 부울검색방법과 비교할 때 제한점은 있다[Cooper, 1983].

본 연구에서는 색인어 퍼지관계행렬을 이용한 문헌검색시스템에 관한 실험을 수행하고 그 결과를 분석하였다. 각 색인어들과 불리안 연산자인 AND, OR, NOT으로 이들 색인어들을 조합한 질의식들을 통해 실험을 수행한 결과 PC환경에서의 실험적 시스템에서 기존의 일반집합이론에 의한 검색실험에서보다 상당히 우수한 성능을 보였다. 특히 재현율과 정확률을 측정한 성능평가 결과는 퍼지 문헌검색

시스템이 효율적이라는 사실을 입증하였다고 할 수 있다.

| 구 분 | 재현율 | 정확률 |
|---|---|---|
| 일반적 방법 | 43 % | 45 % |
| 퍼지관계에 의한 방법 | 53 % | 42 % |

표 3. 퍼지관계행렬을 이용한 검색과 일반 집합이론에 의한 검색결과의 비교

표 3.은 성능평가 결과를 표로 작성한 것으로서 표에서 알 수 있는 바와 같이, 일반 집합이론에 의한 방법보다 정확률에 있어서는 소폭 감소하였지만 재현율에 있어서는 10 %가 증가하였음을 확인할 수 있다.

한편, 본 시스템을 설계함에 있어서 아직 해결하지 못한 부분은 자동입력과 자동색인에 관한 모듈이다. 수작업을 통해서 DB를 구축하고 색인하는 과정은 실험용이나 소규모 시스템에 있어서는 용이하나, 실제로 현장에서 활용하기 위해서는 입수문헌의 대량성과 인건비 압박 때문에 대량으로 입수되는 문헌들과 잡지논문들을 자동으로 입력하고 색인하는 일이 매우 필수적이라고 할 수 있으며, 그러한 모듈들이 갖추어 져야만 실용성과 경쟁력을 갖춘 시스템이 될 수 있을 것이다.

또 하나의 미해결 문제는 추출된 색인어들에 대해 퍼지관계 또는 가중치를 자동으로 부여할 수 있는 체계적인 방법을 개발하는 것이다. 이 문제를 해결하기 위한 접근방법으로서 통계적 으로 빈도수를 계산한다든지, 문헌과 용어와의 관계를 계산하여 퍼지관계 함수를 적용한다든지 하는 방법들이 제시된 바 있으나[Y. Ogawa et al; Buell], 이들 방법이 지니고 있는 논리적 타당성이 아직은 입증되지 못하고 있는 실정이다.

그러나 앞으로도 이러한 미해결된 부분을 보완하고 이론적인 가설들을 실험적으로 입증할 수 있는 계속연구가 필요하며, 아울러서 이를 현장기술로 적용하기 위한 연구를 계속하는 것이 요망된다고 하겠다.

감사의 말씀

본 연구를 수행하는 과정에서 조언과 협조를 아끼지 않으신 전남대학교 문헌정보학과 정준민 선생님, 전자계산소의 김철 선생님, 환경연구소의 고화석 선생님께 감사를 드립니다.

# 〈참 고 문 헌〉

강일중. 1990. 용어간 관계를 이용한 검색문헌의 순위부여에 관한 연구. 서울: 연세대학교대학원 문헌정보학과; 1990. (석사학위논문).

김성혁. 1990. 전문가 대체시스템에서의 퍼지 추론에 관한 연구. 정보관리학회지. 1990; 7(1): 68-78.

김영귀. 1990. 완전매치와 부분매치검색기법에 관한 연구. 정보관리학회지. 1990; 7(1): 79-95.

사공철 등. 1990. 최신정보검색론. 서울: 구미무역 출판부; 1990.

오길록; 이광형. 1991. 퍼지 이론 및 응용 I, II. 서울: 홍릉과학출판사; 1991.

안현수. 1986. 한글문헌의 자동색인에 관한 실험적 연구. 서울: 연세대학교대학원 문헌정보학과; 1986. (석사학위논문).

이순재. 1989. 정보검색 시스템에 Fuzzy Set이론의 적용. 도서관.정보학연구(경북대학교대학원도서관.정보학과). 1989; 1: 201-236.

이영자. 1989. 지능형 정보검색 시스템에 관한 고찰. 도서관.정보학연구(경북대학교대학원도서관.정보학과). 1989; 1: 1-36.

정영미. 1988. 정보검색론. 서울: 정음사; 1988.

Belkin, N.J.; Croft, W.B. 1987. Retrieval Techniques. Annual Review of Information Science and Technology. 1987; 22: 109-146.

Belkin, N.J. et al. 1982. ASK for Information Retrieval:Part I & II. Journal of Documentation. 1982;38(2) :61-71;38(3) :145-163.

Bookstein, Abraham. 1980. Fuzzy Requests: An Approach to

Weighted Boolean Searches. Journal of the American Society for Information Science. 1980; 31(4): 240-247.

Bookstein, Abraham. 1981. A Comparison of Two Systems of Weighted Boolean Retrieval. Journal of the American Society for Information Science. 1981; 32(4): 275-279.

Bookstein, Abraham. 1983. Outline of a General Probabilistic Retrieval Model. Journal of Documentation. 1983; 39(2): 63-72.

Bookstein, Abraham. 1985. Probability and Fuzzy Set Applications to Information Retrieval. Annual Review of Information Science and Technology. 1985; 20: 117-152.

Brooks, R.M. 1987. Expert Systems and Intelligent Information Retrieval. Information Processing and Management. 1987; 23(4): 367-382.

Bruandet, Marie-France. 1989. Outline of a Knowledge-Base Model for an Intelligent Information Retrieval System. Information Processing and Management. 1989; 25(1): 89-115.

Buell, D.A. 1985. A Problem in Information Retrieval with Fuzzy Sets. Journal of the American Society for Information Science. 1985; 36(6): 398-401.

Buell, D.A.; Kraft, D.H. 1981. Threshold Values and Boolean Retrieval Systems. Information Processing and Management. 1981; 17(3): 127-136.

Chiaramella, Y.; Defude, B. 1987. A Prototype of an Intelligent System for Information Retrieval: IOTA. Information Processing and Management. 1987; 23(4): 285-303.

Cohen, P.R.; Kjeldsen, R. 1987. Information Retrieval by Constrained Spreading Activation in Semantic Networks. Information Processing and Management. 1987; 23(4): 255-268.

Cooper, W.S. 1983. Exploiting the Maximum Entropy Principle to Increase Retrieval Effectiveness. Journal of the American Society for Information Science. 1983; 34(1): 31-39.

Cooper, W.S. 1988. Getting Beyond Boole. Information Proccessing and Management. 1988; 24(3): 243-248.

Croft, W.B. 1987. Approaches to Intelligent Information Retrieval. Information Processing and Management. 1987; 23(4): 249-254.

Croft, W.B.; Thompson, R.H. 1987. I3R: A New Approach to the Design of Document Retrieval Systems. Journal of the American Society for Information Science. 1987; 38(6): 389-404.

Doszkocs, T.E. et al. 1990. Connectionist Models and Information Retrieval. Annual Review of Information Science and Technology. 1990; 25: 209-260.

Fox, E.A. 1989. Research and Development of Information Retrieval Models and Their Application. Information Processing and Management. 1989; 25(1): 1-5.

Fox, E.A.; Koll, M.B. 1988. Practical Enhanced Boolean Retrieval: Experiments with the SMART and SIRE Systems. Information Processing and Management. 1988; 24(3): 257-267.

Fox, E.A.; Winett, S.G. 1990. Using Vector and Extended Boolean Matching in an Expert System for Selecting Foster Homes. Journal of the American Society for Information Science. 1990; 41(1): 10-26.

Fuhr, Norbert. 1989. Models for Retrieval with Probabilistic Indexing. Information Processing and Management. 1989; 25(1): 55-72.

K. Nakamura; S. Iwai. 1982. Topological fuzzy sets as a quantative description of analogical inference and its application to question-answering system for information retrieval. IEEE Trans. Systems Man Cybernet. 1982;12(2):193-204.

Kerre, E.E. et al. 1986. The Use of Fuzzy Set Theory in Information Retrieval and Databases: A Survey. Journal of the American Society for Information Science. 1986; 37(5): 341-345.

Koll, Mathew; Srinivasan, Padmini. 1990. Fuzzy versus Probabilistic Models for User Relevance Judgments. Journal of the American Society for Information Science. 1990; 41(4): 264-271.

Kraft, D.H.; Buell, D.A. 1983. Fuzzy Sets and Gerneralized Boolean retrirval systems. International Journal of Man-Machine Studies. 1983;19(1):45-56.

Mantaras, R.L.de. et al. 1990. Knowledge Engineering for a Document Retrieval System. Fuzzy Sets and Systems. 1990; 38: 223-240.

Markey, Karen. 1981. Levels of question Formulation in

Negotiation of Information Need during the Online Presearch Interview: a Proposed Model. Information Processing and Management. 1981; 17(5); 215-225.

McCarn, D.B.; Lewis, C.M. 1990. A Mathematical Model of Retrieval System Performance. Journal of the American Society for Information Science. 1990; 41(7): 495-500.

Miyamoto, S. 1990. Information Retrieval Based on Fuzzy Assocfiations. Fuzzy Sets and Systems. 1990; 38: 191-205.

Murai, T. et al. 1989. A Fuzzy Document Retrieval Method Based on Two-Valued Indexing. Fuzzy Sets and Systems. 1989; 30: 103-120.

Nakkouzi, Z.S.; Eastman, C.M. 1990. Query Formulation for Handling Negation in Information Retrieval Systems. Journal of the American Society for Information Science. 1990; 41(3): 171-182.

Nomoto, K. et al. 1990. A Document Retrieval System Based on Citations Using Fuzzy Graphs. Fuzzy Sets and Systems. 1990; 38: 207-222.

Ogawa, Y. et al. 1991. A Fuzzy Document Retrieval System Using the Keyword Connection Matrix and a Learning Method. Fuzzy Sets and Systems. 1991; 39: 163-179.

Prade, H.; Testemale, C. 1987. Fuzzy Relational Databases: Representational Issues and Reduction Using Similarity Measures. Journal of the American Society for Information Science. 1987; 38(2): 118-126.

Rada, Roy; Bicknell, Ellen. 1989. Ranking Documents with a

Thesaurus. Journal of the American Society for Information Science. 1989; 40(5): 304-310.

Radecki, T. 1983. Gerneralized Boolean methods of information retrieval. International Journal of Man-Machine Studies. 1983;18:407-439.

Radecki, Tadeusz. 1977. Mathematical Model of Time-Effective Information Retrieval System Based on the Theory of Fuzzy Sets. Information Processing and Management. 1977; 13(2): 109-116.

Radecki, Tadeusz. 1988. Trends in research on Information Retrieval-The Potential for Improvements in Conventional Boolean Retrieval Systems. Information Processing and Management. 1988; 24(3): 219-227.

Radecki, T. 1981. Outline of a fuzzy logic approach to information retrieval, International Journal of Man-Machine Studies. 1981;14:169-178.

Radecki, Tadeusz. 1976a. Mathematical Model of Information Retrieval System Based on the Concept of Fuzzy Thesaurus. Information Processing and Management. 1976; 12(5): 313-318.

Radecki, Tadeusz. 1976b. New Approach to the Problem of Information System Effectiveness Evaluation. Information Processing and Management. 1976; 12(5): 319-326.

Radecki, Tadeusz. 1979. Fuzzy Set Theoretical Approach to Document Retrieval. Information Processing and Management. 1979; 15(5): 247-259.

Radecki, Tadeusz. 1983. A Theoretical Background for Applying Fuzzy Set Theory in Information Retrieval. Fuzzy Sets and Systems. 1983; 10: 169-183.

Rijsbergen, C.J. van. 1979. Information Retrieval. 2nd ed. London: Butterworths; 1979.

S. Miyamoto et al. 1983. Generation of a pseudothesaurus for information retrieval based on co-occurences and fuzzy set operations. IEEE Trans. Systems Man Cybernet. 1983;13(1):62-70.

S. Miyamoto; K. Nakayama. 1986. Fuzzy information retrieval based on a fuzzy pseudothesaurus. IEEE Trans. Systems Man Cybernet. 1986;16(2):278-282.

Salton, G.; McGill, M.J. 1983. Introduction to Information Retrieval. NY: McGraw-Hill, 1983.

Schank, R.C. et al. 1981. Conceptual Information Retrieval. In Information Retrieval Research. ed by R.N. Oddy. London: Butterworth; 1981. pp. 94-116.

Smith, Linda C. 1987. Artificial Intelligence and Information Retrieval. Annual Review of Information Science and Technology. 1987; 22: 41-47.

Srinivasan, Padamini. 1991. The Importance of Rough Approximations for Information Retrieval. International Journal of Man-Machine Studies. 1991; 34(5): 657-671.

Sullivan, Michael et al. 1990. End-Users, Mediated Searches, and Front-End Assistance Programs on Dialog: A Comparison of Learning, Performance, and Satisfaction. Journal of the American Society for Information

Science. 1990; 41(1): 27-42.

T. Murai et al. 1988. A modeling of search oriented thesaurus use based on multivalued logical inference. Information Science. 1988;43:185-212.

Tahani, Valiollah. 1976. A Fuzzy Model of Document Retrieval Systems. Information Processing and Management. 1976; 12(3): 177-187.

Tahani, Valiollah. 1977. A Conceptual Framework for Fuzzy Query Processing: a Step Toward Very Intelligent Database Systems. 1977; 13(5): 289-303.

Teskey, F.N. 1989. User Models and World Models for Data, Information, and Knowledge. Information Processing and Management. 1989; 25(1): 7-14.

Thompson, Paul. 1990. A Sensitivity Analysis of a Probabilistic Information Retrieval System. Journal of the American Society for Information Science. 1990; 41(5): 348-358.

Tong, R.M. 1987. Conceptual Information Retrieval using RUBRIC. Proceedings of the Tenth Annual International ACMSIGIR Conference on Research & Development in Information Retrieval. June 1987:247-253.

Vickery, A.; Brooks, H.M. 1987. PLEXUS - The Expert System for Referral. Information Processing and Management. 1987; 23(2): 99-117.

Wong, S.K.M.; Yao, Y.Y. 1989. A Probability Distribution Model for Information Retrieval. Information Processing and Management. 1989; 25(1): 39-53.

Yager, R.R. 1980. A logical on-line bibliographic searcher: an application of fuzzy sets. IEEE Trans. Systems Man Cybernet. 1980;10(1):51-53.

Zadeh, L.A. 1965. Fuzzy Sets. Information and Control. 1965;8:338-353.

Zenner, R.B. et al. 1985. A New Approach to Information Retrieval Systems Using Fuzzy Expressions. Fuzzy Sets and Systems. 1985; 17: 9-22.

# An Experimental Study on Fuzzy Document Retrieval System

## Abstract

Seung Chai Lee*

Theoretical developments in the information retrieval have offered a number of alternatives to traditional Boolean retrieval. Probability theory and fuzzy set theory have played prominent roles here. Fuzzy set theory is an attempt to generalize traditional set theory by permitting partial membership in a set and this means recognizing different degrees to which a document can match a request.

In this study, an experimentation of a document retrieval system using the fuzzy relation matrix of the keywords is described and the results are offered. The queries composed of keywords and Boolean operaters AND, OR, NOT were processed in the retrieval method, and the method was implemented on the PC of 32bit level(30 MHz) in an experimental system.

The measurement of the recall ratio and precision ratio verified the effectiveness of the proposed fuzzy relation matrix of keywords and retrieval method. Compared to traditional crisp method in the same document database, the recall ratio increased 10 % high although the precision ratio decreased slightly.

The problems, in this experiment, to be resolved are first,

* Instructor, Chonnam National University

the design of the automatic data input and fuzzy indexing modules, through which the system can have the ability of competition and usefulness. Second, devising a systematic procedure for assigning fuzzy weights to keywords in documents and in queries.