색인어 퍼지 관계와 서열기법을 이용한 정보 검색 방법론

김 철[†] · 이 승 채[††] · 김 병 기[†††]

요 약

본 연구에서는 색인어 퍼지 관계행렬을 이용한 정보검색 방법을 제안하고 간단한 문헌정보 검색시스템을 사용하여 실험을 수행하고 그 결과를 분석하였다. 불리안 연산자인 AND, OR, NOT으로 색인어들을 조합한 질의식을 통해 실험을 수행한 결과 일반 집합이론에 의한 검색실험에서보다 상당히 우수한 성능을 보였다. 특히 재현율과 정확률을 측정한 성능평가 결과는 퍼지 문헌검색 시스템이 가능한 검색 대안이라는 사실을 확인하였다고 할 수 있다.

한편, 검색의 기법 측면에서 고려하였을 때 본 실험은 먼저, 색인어 관계행렬에 따라서 검색결과에 서열을 부여하였고, 기준적합도값의 변동에 따라 검색결과가 유동적으로 대응하도록 하였으며, 관계값을 의미적 거리로 파악함으로써 검색과정과 검색 시멘틱스를 일치시키고자 새롭게 시도하였다.

A Methodology of the Information Retrieval System Using Fuzzy Connection Matrix and Document Connectivity Order

Chul Kim[†] · Seungchai Lee[††] · Byungki Kim[†††]

ABSTRACT

In this study, an experiment of information retrieval using fuzzy connection matrix of keywords was conducted. A query for retrieval was constructed from each keyword and Boolean operator such as AND, OR, NOT. In a workstation environment, the performance of the fuzzy retrieval system was proved to be considerably effective than that of the system using the crisp set theory. And both recall ratio and precision ratio showed that the proposed technique would be a possible alternative in future information retrieval. Some special features of this experimental system were; ranking the results in the order of connectivity, making the retrieval results correspond flexibly by changing the threshold value, trying to accord the retrieval process with the retrieval semantics by treating the averse-connectivity (fuzzy value) as a semantic approximation between keywords.

1. 서 론

1.1. 연구의 내용

컴퓨터는 인간의 지적 활동영역에서 없어서는 안될 중요한 도구가 되어 있으며 우리가 원하는 많은 일을 대신해 주고 있다. 컴퓨터 내부의 처리과정은 바로 수치처리 과정이라 할 수 있으며, 그렇게 하기 위해서는 제기된 문제를 수치로 바꿔 주어야 하고 컴퓨터는 이 수치를 처리함으로써 문제를 해결하는 것이다. 이 때 컴퓨터에 입력하는 수치는 정확한 것이

† 정 회 원:광주교육대학교 전산교육과 조교수
†† 정 회 원:광주정보컨설팅
††† 종신회원:전남대학교 전산학과 교수
논문접수:1996년 6월 17일, 심사완료:1996년 8월 16일

어야 한다. 즉, 사과 세개 또는 영하 10℃ 등, 정확한 수치나 개념으로 바꾸어 주어야 처리할 수 있다. 그러나 인간이 머리속에 담고 있는 개념이나 지식은 정확한 수치로 나타내기 어려운 모호한 표현을 포함하는 경우가 많이 있다.

컴퓨터가 인간의 문제를 제대로 해결하기 위해서는 인간이 사용하는 숫자는 물론이고 애매한 표현을 처리할 수 있어야 한다. 이러한 애매한 표현을 처리할 수 있는 이론적인 바탕을 제공하는 것이 바로 퍼지이론이다. 퍼지이론은 현상의 불확실한 상태를 그대로 표현해 주는 방법으로서 1965년 버클리대학의 자데(Lofti A. Zadeh)교수에 의해서 처음 소개되었다. 퍼지이론은 애매하게 표현된 데이터를 우리에게 유용한 자료로 만들기 위한 것이다[1].

정보의 폭발로 표현되고 있는 만큼 문헌생산량의 증가로 인하여 문헌정보의 축적과 검색은 정보처리 및 관리 환경에서 매우 중요한 기능이 되었다. 비록 정보공학의 발전이 신속히 이루어지고 있기는 하나 과거의 불리안식 일반집합 이론(crisp set theory)에 기초를 둔 검색방법은 이용자의 심도깊은 정보요구에 정확하게 부응하기에는 충분하지 못한 것이 사실이며, 몇가지 취약점을 지니고 있다[2, 3, 4]. 그중에서도 가장 취약한 점으로는 탐색어로 표현되는 각 개념간의 상대적 중요도나 관계의 정도를 표현하지 못한다는 점을 지적할 수 있다.

현재 널리 이용되고 있는 정보검색 시스템에서는 단순히 질문과 문헌간의 공통된 용어의 존재와 수만을 고려하므로 공통용어의 수가 많으면 검색질문과 문헌의 관련성을 높게 나타낼 뿐, 용어간의 관계도는 반영하지 않고 있는 실정이다[5]. 따라서 불리안식 검색모형을 개선하고자 하는 연구가 여러 각도에서 수행되었으며, 그 방법으로는 가중치 개념을 도입하거나 서열을 부여하는 방법등이 모색되어 왔다[6].

본 연구는 문헌의 초록에서 추출된 색인어들 사이의 퍼지 관계를 이용하여 문헌검색을 수행하기 위한 시스템을 실험적으로 구현하는 것을 내용으로 한다. 퍼지 관계행렬을 이용함으로써 퍼지색인의 작성이 가능하며, 이용자의 질의식과의 관계도(connectivity)에 따라서 검색된 문헌에 대해 적합관계도 서열을 부여할 수 있다. 또한 관계행렬을 통해, 일반 검색시스템에 있어서 검색용 시소러스(thesaurus)를 활용하는 것과 같은 방법으로 이용자가 검색하고자 하는 주제에 적합한 질의식을 작성할 수 있도록 하는 것을 내용으로 한다.

또한 본 실험에서 적용하고 있는 퍼지 관계는 관계값을 역으로 해석할 때 용어와 용어간의 거리개념으로 이해할 수 있다. 일반적으로 검색자가 적합문헌을 찾는 과정은 검색자의 질의식에 포함되어 있는 색인어들에 의미적으로 가장 가까운 거리에 있는 문헌집단을 찾는 것으로 해석이 가능하다. 본 실험에서는 검색 시맨틱스와 검색 매카니즘을 일치시키기 위해 검색서열을 적용하고자 하였다.

본 연구를 수행하는 과정에서 사용한 실험용 데이터베이스는 1992년 이후 발표된 유전공학 분야 영문 논문들의 초록으로 구성하였으며, 따라서 비교적 소규모의 영문 데이터베이스를 대상으로 한 실험 환경이 연구 수행상의 제한점이다.

1.2. 관련연구

문헌정보 검색기법은 완전매치 기법과 부분매치 기법으로 대별된다[7, 8]. 문헌정보 검색시스템의 기능은 대상문헌의 저자의 의도를 정확히 추정하는 주제분석과 이용자의 정보요구를 확실하게 파악하는 요구분석을 통하여 양자간의 커뮤니케이션을 증진시키는 것이다. 이러한 중요한 두가지의 요소중 주제분석은 흔히 색인이나 초록과 같은 정보가공의 결과로 만들어지고, 요구분석은 이용자의 질문내용을 공식화하여 탐색모형을 수립함으로써 수행된다. 그러나 주제분석과 요구분석은 과정상 본질적으로 불확실성과 애매성을 내포하게 된다. 이상적인 정보검색 시스템이란 이러한 요인들을 극소화시킴으로써 이용자에게 보다 적합한 정보를 제공할 수 있는 시스템을 말한다[9].

또한 보다 효율적인 정보검색 시스템은 첫째, 이용자가 표현한 정보요구 내용에 관련된 용어가 문헌에 포함되어 있지 않더라도 관련문헌(적합문헌)을 찾아낼 수 있어야 하며, 둘째, 역으로 이용자가 표현한 정보요구 내용에 관련 없는 용어라도 문헌에 이 용어가 나타나 있는 경우에도 관련문헌을 찾아낼 수 있어야 한다[10].

본 실험에서 적용하고자 하는 퍼지집합 이론은 경계가 불분명하여 잘못 정의될 수 있는 애매정보를 처

리하는데 있어서 적합하기 때문에, 많은 연구자들이 가중치나 서열값을 부여하는데 적용해 왔다[11]. 예를 들면, 퍼지 색인시스템[12, 13, 14], 일반색인어에 기초한 퍼지 시소러스를 이용한 퍼지 검색시스템[15], 퍼지색인에 기초한 퍼지시소러스를 이용한 퍼지 검색시스템[16], 인용문헌을 이용한 퍼지 검색시스템[17], 등이 발표된 바 있다.

한편, 오가와 등(Y. Ogawa, et al.)은 색인어 관계행렬(keyword connection matrix)를 이용한 퍼지 문헌검색시스템을 제안한 바 있다[11]. 이 시스템에서는 불리안 색인을 전제로 하고 있는데, 이는 대부분의 데이터베이스들이 일반 색인방법에 기초를 두고 있을 뿐만 아니라 통상 색인어의 숫자가 문헌집단의 숫자보다 적기 때문에 퍼지색인보다는 퍼지 시소러스를 유지하는 것이 더 용이하다는 점을 전제로 한 것이다.

최근에 들어서 엔라도(P. Enrado)는 정보 고속도로와 문헌검색의 문제와 관련하여 정보 검색과정이 더욱 어려워져 가고 있는 상황을 지적하면서 비구조적인 포맷을 지닌 문헌들을 분류, 저장, 검색하는데 있어서 퍼지논리를 응용한 ITMS(Intelligent Text Management System)을 소개하고 있다. ITMS의 특징은 문헌을 검색하고 분석하는데 있어서 인간의 판단과정을 시뮬레이션한 J-Space(판단공간) 기법이라 할 수 있는데, 이 기법의 개요는 질의식을 형성하는 이용자의 주제기준을 정해주는 탐색 매개변수 집합과 문헌사이의 관계에 따라서 문헌을 분류할 수 있도록 하고 있다[18]. 그리고 콕스(E. Cox)와 고우츠(M. A. Goetz)는 퍼지논리를 응용한 관계형 데이터베이스 질의식에 관한 연구를 발표하고 있음을 볼 수 있다[19, 20].

한편 국내에서는 퍼지개념을 적용한 질의식의 분석과 문헌정보 검색에 관한 실험적 논고를 통해서 문헌정보 검색시스템에 퍼지이론을 적용하고자 시도한 바 있으나[21], 당시의 연구는 영역지식에 해당하는 용어간 퍼지 관계행렬의 값을 해당분야 전문가의 지적 판단에 의존할 만큼 초기적인 수준의 것이었으며, 색인어 생성과정 또한 자동화 측면에서 부족한 점을 많이 안고 있었다.

2. 영역지식과 퍼지 관계행렬

기존의 정보검색 시스템들이 채택하고 있는 검색 추론과정은 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 문헌의 색인어에 있어서 각 용어의 발생빈도나 용어간에 존재하는 통계적 의존성을 주제영역에 있어서의 지식으로 이용하였다.

본 연구는 추출된 색인어들 사이의 퍼지관계를 영역지식으로 활용한 것으로서, 색인어 관계행렬은 다음과 같이 표현된다.

〈표 1〉 색인어 퍼지 관계행렬
〈Table 1〉 fuzzy connection matrix of keywords

t_i/t_j	Virus t_1	lactob－acillus t_2	Acryl－amide t_3	Biore－actor t_4	chole－strol t_5	Escher－ichiacoli t_6	…
t_1	1.0	0.50	0.35	0.33	0.43	0.45	
t_2		1.0	0.52	0.81	0.62	0.55	
t_3			1.0	0.23	0.22	0.32	
t_4				1.0	0.24	0.82	
t_5					1.0	0.44	
t_6						1.0	
:							

여기서 관계값은 색인어들 사이의 개념적 유사도 즉 의미적 거리를 나타내고 있다. 이러한 점에서 이 행렬은 색인어들 사이의 의미적 거리관계를 나타내는 퍼지시소러스라고도 할 수 있을 것이다. 실제로 관계값들은 관련 용어들 사이의 관련도에 따라서 주어진다. 색인어 관계행렬은 K x K 행렬 W로 표현할 수 있으며, 여기서 K는 색인어의 숫자와 같은 크기를 갖는다.

한편, 문헌들에 있어서 두개의 색인어가 동시에 출현하는 빈도수가 많을 수록 이들 색인어 사이의 관계가 높다는 가정하에 자동으로 색인어들 사이의 관계값을 부여할 수 있는 환경에서 초기 관계값들을 부여한 연구결과들도 다수 발표된 바 있다. 본 연구에서

는 다음과 같은 산출식에 의해 색인어 사이의 관계값을 산출하였다[11, 16, 22].

$$W_{ij} = \frac{N_{ij}}{N_i + N_j - N_{ij}}, \quad i \neq j$$

$$1, \qquad i = j$$

여기서, W_{ij}는 i번째와 j번째 색인어 사이의 초기 관계값이며, N_{ij}는 i번째와 j번째 색인어들을 모두 포함하는 문헌들의 갯수이다. N_i와 N_j는 각각 i번째 색인어와 j번째 색인어만을 포함하고 있는 문헌들의 갯수이다. 이 공식에 의해서 두개의 색인어가 동시에 포함된 문헌들의 정규빈도수가 결정된다.

본 연구에서는 이러한 원칙들 즉, 색인어 사이의 관계값, 색인어와 문헌 사이의 관계값을 퍼지 집합이론의 측면에서 자동 생성함으로써 정보의 검색과정을 보다 일반화하고자 하는 궁극적 목표를 염두에 두었으며, 용어(keyword)간의 관계를 의미적 거리관계로 규정함으로써 정보검색과 관련하여 새로운 접근방법과 가능성을 모색하고자 한 것이다.

3. 퍼지 관계행렬을 이용한 검색

3.1. 퍼지 문헌검색시스템의 정의

퍼지 문헌정보 검색시스템은 다음과 같이 정의된다[23].

$D = \{d_i\}$: 문헌집합
$T = \{t_i\}$: 용어집합
$f: D \times T \rightarrow [0, 1]$ 퍼지소속함수

문헌과 색인어 사이의 관계함수 $f(d_i, t_i)$를 통해서는 문헌 d_i와 용어 t_i 사이의 관계도를 측정한다.

먼저, 검색자는 검색하고자 하는 색인어들을 조합하여 질의식을 만든다. 이 질의식은 색인어들과 AND, OR, NOT 등의 논리연산자들로 구성된다. 질의식은 불리안 연산의 기본 규칙들을 반복적으로 적용함으로써 연결식으로 변환될 수 있는데, 이 식은 OR와 NOT만을 포함하는 부질의식(subquery)으로 나뉜다. 연결식의 예를 들면,

$$Query = SubQuery(1) \wedge \cdots \wedge SQ(N),$$
$$SQ(h) = K_1 \vee \cdots \vee Kn_h \vee Kn_{h+1} \vee \cdots \vee \lceil Kn_h + m_h,$$

여기서 $\vee$, $\wedge$, $\lceil$은 각각 불리안 연산자인 AND, OR, NOT을 가리킨다. 그리고 K_i는 질의식 내의 i번째 용어를 나타낸 것이다. 질의식 내에서는 $N \geq 1$이며, h번째 부질의식은 $n_h \geq 0$, $m_h \geq 0$ 그리고 $n_h + m_h < 1$의 조건을 가진다.

기존의 일반적인 검색에 있어서의 검색결과는 다음과 같다.

$$SubResult(h) = D(K_1) \cup \cdots \cup D(Kn_h) \cup \overline{D(Kn_{h+1})}$$
$$\cdots \cup \overline{D(Kn_h + m_h)}$$

여기서, D(K)는 색인어 K를 갖는 문헌집합을 의미하고, $D_1 \cup D_2$는 두개의 집합 D_1과 D_2의 합집합이며, $\overline{D}$는 D의 여집합을 나타낸다.

질의식의 OR 연산자는 여집합을 구하는 연산자이다.

따라서 검색결과는 다음과 같은 식으로 표현될 수 있다.

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

여기서 검색결과는 집합의 구성요소들이 정확히 질의식과 일치하는 일반집합(crisp set)으로 나타난다.

즉, 본 연구에서의 검색 방법은 검색결과가 퍼지집합이라는 사실만 제외하면 일반 집합이론에 따른 검색방법과 동일하다고 할 수 있다.

3.2. 퍼지색인의 작성

A_i는 i번째 문헌에 대해 색인된 색인어집합이라고 하자. 본 연구에서는 문헌의 초록들로부터 불용어들을 제거한 후, 남은 용어들을 색인어 관계행렬상의 용어들과 비교한 다음 관계행렬을 이용해 해당 문헌의 적합도를 계산하는 방법을 사용하였다.

한편, i번째 문헌과 j번째 색인어사이의 관계값 R_{ij}는 다음과 같이 정의된다[11].

$$R_{ij} \equiv \bigoplus_{K_k \in A_i} W_{jk}$$

여기서 W_{jk}는 색인어 관계행렬 내에서 j번째와 k번

째 색인어 사이의 관계값을 의미하며, ⊕는 식 $\oplus_i X_i = 1 - \prod_i (1 - X_i)$로 정의되는 대수합을 나타낸다. 이 공식은 다음과 같이 정리할 수 있다.

$$R_{ij} = 1 - \prod_{Kk \subset Ai} (1 - W_{jk})$$

즉, R_{ij}는 i번째 문헌과 j번째 색인어 사이의 관계값으로서 퍼지 소속함수값이 된다.

3.3. 질의식의 관계값

3.3.1. 부질의식의 관계값 계산

퍼지 집합이론에서 두 집합 A와 B의 합집합은 A ∪ B로 표시하고, 이때 어느 원소 x의 소속함수값은 x가 A와 B에 포함될 가능성 중에 큰 것을 취한다. 즉,

$$\mu_A \cup B(x) = max\,[\mu_A(x), \mu_B(x)], \forall x \in X.$$

본 연구에서는 퍼지집합에 있어서 대수합 계산방법을 적용하여 다음과 같은 식에 의해 합집합 값을 구하였다. 대수합 연산자 ⊕는 퍼지 합집합을 위한 Max연산자에 비해 A와 B사이의 교환성이 크다고 할 수 있다.

$$\begin{aligned}\mu_A \cup B(x) &= \mu_A(x) \oplus \mu_B(x) \\ &= 1 - \mu_A(x) \cdot \mu_B(x)\end{aligned}$$

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.

$$\mu_{\overline{A}}(x) = 1 - \mu_A(x)$$

부질의식에 대한 관계값은 다음 식을 통해서 계산된다.

$$\begin{aligned}r_i(h) &= \Big(\bigoplus_{K_j \in Q(h)^+} R_{ij}\Big) \oplus \Big\{\bigoplus_{K_j \in Q(h)^-} (1 - R_{ij})\Big\} \\ &= I - \Big(\prod_{K_j \in Q(h)^+} S_{ij}\Big)\Big(\prod_{K_j \in Q(h)^-} R_{ij}\Big)\end{aligned}$$

여기서 S_{ij}는 다음 식으로 정의된다.

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

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

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

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

3.3.2. 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 교집합이 A∩B라 하면 퍼지 교집합은 다음과 같이 정의된다.

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

따라서 i번째 문헌의 관계값은 다음 식에 의해서 계산된다.

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

W가 단위행렬일 때, 위의 식에 의한 퍼지 결과값은 일반 집합이론의 계산방법을 사용했을 때의 값과 동일하게 처리된다.

3.3.3. 질의식의 수정

정보검색과정은 통제된 불확실성의 범주에서 질의어를 포함하는 문헌집합을 찾는[24], 상호작용적이면서 미결정적인 과정이며, 이용자로 하여금 더 양호한 질의식을 작성할 수 있도록 지원하는 것이 중요한 요소이다. 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 시스템은 이용자로 하여금 색인어와 질의식과의 관계도 순서에 따라서 색인어들을 나열하도록 요청할 수 있다. 관계도는 색

인어의 적합도라고 할 수 있는 것으로서, 한 문헌의 적합도를 계산하는 방법과 같은 방법으로 계산할 수 있다. 예를 들어 i번째 색인어의 적합도 T_i는 다음 식과 같이 계산된다.

$$T_i(h) = 1 - \Big(\prod_{K_j \in Q(h)^+} W_{ij}\Big)\Big\{\prod_{K_j \in Q(h)^-} (1 - W_{ij})\Big\},$$

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

4. 문헌 검색서열

본 연구에서 적용하고 있는 색인어 퍼지 관계행렬은 그 관계값을 역으로 해석할 때 용어와 용어간의 거리개념으로 이해할 수 있다. 따라서 검색자가 적합 문헌을 찾는 과정은 질의식에 포함되어 있는 색인어에 가장 가까운 문헌집단을 찾는 것으로 해석이 가능하고, 이를 의미적 거리관계로 파악하고자 하였다. 일반적으로 검색시스템에 있어서 검색을 한다고 하는 것은 의미적으로 가장 근접해 있는 용어나 문헌집단을 추출해 내는 작업이기 때문에 적용한 서열기법과 개념은 검색의 시맨틱스와 검색과정을 일치시킨 것이라 할 수 있다.

문헌검색 서열을 이용자의 질의식에 포함된 색인어에 의해 검색된 각 문헌의 관련 색인어들의 관계값을 산출하여 그 값에 따라서 문헌들을 내림차순으로 배열되도록 하였다. 예를 들어 문헌과 색인어집합이 다음과 같이 표현될 때,

$D = \{d_1, d_2, d_3 \ldots\ldots d_n\}$: 문헌집합
$T = \{t_1, t_2, t_3 \ldots\ldots t_n\}$: 색인어집합

질의식에 포함된 색인어 t_i와 검색된 문헌 집단 d_j의 색인어들 간의 관계값은 다음 식에 의해 구할 수 있으며, 여기서 μ는 색인어들 간의 관계값이다.

$$d_m^{t_i} = \left\{\frac{\mu_k}{t_k}\right\} \quad \begin{Bmatrix} k = 1, 2, \ldots n \\ 0 \le \mu \le 1 \\ m = 1, 2, \ldots n \\ i = 1, 2, \ldots n \end{Bmatrix} \tag{1}$$

또한 t_i에 대한 검색된문헌 $d_{\ldots n}$ 각각의 관계값 V는 다음과 같이 구할 수 있다.

$$V_m = \frac{\sum_{i=0}^{k-1} \mu_i}{k-1} \quad \begin{Bmatrix} k = 1, 2, \ldots n \\ m = 1, 2, \ldots n \end{Bmatrix} \tag{2}$$

구해진 검색의 결과 $V_1, \ldots, V_n$는 관계값이 높은 순서(descending order)대로 정렬한다.

예를 들어 검색서열 계산과정을 살펴 보면,

먼저, 실험용 DB에 들어 있는 문헌들 d_4, d_{18}, d_{35}을 예로 들어 이들의 색인어 집합을 구한 결과가 다음과 같다.

d_4 { t_{24}(protein production), t_{20}(protein engineering), t_{520}(new process), t_{11}(genetic manipulation)}

d_{18} { t_{372}(cloning vector), t_{20}(protein engineering), t_{231}(lantibiotic), t_{310}(nisin), t_{81}(mutation), t_{634}(shuttle plasmid)}

d_{35} { t_{92}(enzyme activity), t_{35}(protein folding), t_{20}(protein engineering), t_{70}(detergent), t_{57}(agglomeration), t_{11}(genetic manipulation)}

또한 이들 문헌들에 포함된 색인어 관계행렬은 다음 〈표 2〉와 같이 나타난다.

여기서, 검색자의 질의식에 색인어 protein engineering이 들어 있는 상황을 가정하면, 먼저 t_{20}에 대한 d_4, d_{18}, d_{35}의 색인어들과의 관계값은 식 (1)에 의해서 다음과 같다.

$$d_4^{t_{20}} = \left(\frac{0.89}{t_{24}}\ \frac{0.2}{t_{520}}\ \frac{0.30}{t_{11}}\right),$$

$$d_{18}^{t_{20}} = \left(\frac{0.72}{t_{372}}\ \frac{0.38}{t_{231}}\ \frac{0.56}{t_{310}}\ \frac{0.56}{t_{281}}\ \frac{0.85}{t_{634}}\right),$$

$$d_{35}^{t_{20}} = \left(\frac{0.51}{t_{92}}\ \frac{0.31}{t_{35}}\ \frac{0.84}{t_{70}}\ \frac{0.85}{t_{57}}\ \frac{0.30}{t_{11}}\right),$$

다음 단계로 검색어 t_{20}에 대한 각 문헌들의 관계값은 식 (2)에 의해서 다음과 같이 구해진다.

$$V_4 = \frac{1.39}{3} = 0.46,$$

$$V_{18} = \frac{3.07}{5} = 0.614,$$

〈표 2〉 실험시스템의 색인어 관계행렬

〈Table 2〉 experimental keyword connection matrix

	t_{11}	t_{20}	t_{24}	t_{35}	t_{57}	t_{92}	t_{70}	t_{231}	t_{281}	t_{310}	t_{372}	t_{520}	t_{634}
t_{11}	1	0.30	0.58	0.62	0.46	0.59	0.45	0.37	0.31	0.68	0.65	0.31	0.76
t_{20}		1	0.89	0.91	0.85	0.51	0.84	0.38	0.56	0.56	0.72	0.20	0.85
t_{24}			1	0.84	0.80	0.47	0.83	0.72	0.71	0.70	0.87	0.30	0.43
t_{35}				1	0.58	0.71	0.79	0.47	0.35	0.45	0.67	0.40	0.35
t_{57}					1	0.73	0.69	0.33	0.75	0.67	0.53	0.42	0.56
t_{92}						1	0.65	0.32	0.63	0.51	0.54	0.45	0.74
t_{70}							1	0.56	0.53	0.37	0.56	0.51	0.82
t_{231}								1	0.25	0.73	0.79	0.29	0.52
t_{281}									1	0.45	0.81	0.35	0.33
t_{310}										1	0.76	0.26	0.45
t_{372}											1	0.19	0.71
t_{520}												1	0.50
t_{634}													1

$$V_{35} = \frac{3.41}{5} = 0.682,$$

위에서 산출된 V_4, V_{18}, V_{35}를 관계값이 높은 것 부터 순서대로 배열하면 (V_{35}, V_{18}, V_4)가 된다. 이는 V_{35}가 V_{18}, V_4에 비해 검색서열이 높은 것을 알 수 있다. 이는 검색결과로서의 d_{35}가 d_{18}, d_4에 비해 적합도가 높은 것을 의미하며, 동시에 이를 통해서 검색시스템의 정확률을 높일 수 있음을 의미한다.

따라서, 검색결과로서 각 문헌들은 소속함수값(관계값)에 따라서 내림차순으로 정렬되며, 검색자는 검색된 문헌들의 소속함수값에 기준값을 부여하는 방법과 검색대상 문헌의 갯수를 통제하는 방법을 통해서 요구되는 적합문헌들을 검색할 수 있도록 함으로써 시스템이 유동적으로 검색자의 요구에 대응할 수 있도록 하였다.

5. 실험 환경

정보검색 시스템이 기본적으로 갖추어야 할 요소들로는 구문분석 프로그램, 문헌정보 DB, 검색 알고리즘, 이용자와 시스템사이의 상호작용을 가능케 하는 커뮤니케이션 도구 등을 들 수 있다[25]. 그리고 이들 요소들이 상호 유기적으로 연동되고 각 하부기능들이 기능을 수행하는 과정에서 지능적인 작동원리에 따라서 운용될 수 있게 하는 것이 필요할 것이다.

본 실험은 다음과 같은 환경에서 수행되었다. 실험에 사용된 컴퓨터 시스템은 120 MHz 처리속도의 PA7200 프로세서를 탑재한 워크스테이션 SWS J720 시스템으로서, 이 시스템의 주기억장치 용량은 32 MB, HDD 용량은 2 GB이다. 문헌데이터는 유전공학 분야에서 발표된 연구논문 240건을 *Genetics Abstracts*로부터 다운받아 DB로 활용하였으며, 이 논문들의 초록으로부터 불용어들을 제거한 다음, 추출된 용어 727개를 색인어로 채택하였다.

실험에서 사용한 문헌데이터 초록정보의 예는 아래와 같다.

DIALOG(R)File 315:Chemlbg & Biotec Abs
(c)1996 RoySocChm,DECHEMA,FizChemie.
All rts. reserv.

383328 CEABA Accession No: 27-04-00732
DOCUMENT TYPE: Journal
Title: Expectations in molecular engineering.
Orig. Title: Erwartungen an das Molekul-Engineering.
AUTHOR: Kutschor, H.
CORPORATE SOURCE: ASTA Medica 60001 Frankfurt, Germany
JOURNAL: Bioforum, Volume:18, Issue:10,
Page(s): 383-388
CODEN: QQQQQQ
PUBLICATION DATE: 1995 (950000)
LANGUAGE: German
ABSTRACT: This article discusses the possibilities of molecular engineering. For example, modified E.coli could be used to synthesize muconic acid from D-glucose, and thus provide an alternative raw material for the synthesis of nylon, thus replacing benzene, whose production requires a lot of energy and leads to depletion of vital fossil resources. Other examples include the synthesis of plastics by bacteria, which would be a cheaper alternative to current procedures which also deplete mineral oil reserves, and the application of molecular engineering in the development of drugs. (Perez)
DESCRIPTORS: English: molecular biology; biotechnology; biopolymer; drug design
SECTION: Fermentation Technology (57)
SECTION CROSS REFERENCE: Pharmaceuticals (58)
DECHEMA CLASSIFICATION: Biology, Microbiology, Molecular Biology (classification, taxonomy, morphology, physiology, strain improvement, genetics, ecology, inoculum, maintenance, storage) (914); Culture, production and isolation of bacterial strains and biochemically active substances (144); High-molecular products (enzymes, hormones, peptides, etc.)(9435)

6. 검색성능 평가

검색시스템의 효율은 검색요구에 적합한 문헌을 검색해 내는 능력을 의미하는 것으로 검색된 적합문헌과 부적합문헌, 검색되지 않은 적합문헌과 부적합문헌 사이의 비율로서 측정된다.

정보검색 시스템에 있어서 검색성능을 평가하는 일반적인 방법은 정확률과 재현율을 측정함으로써 구하고 있다. 본 실험에서는 퍼지 관계행렬을 이용한 검색과 일반 집합이론에 따른 검색결과를 비교하였다. 그리고 재현율과 정확률을 동시에 증가시켜주는 최적기준값(optimal threshold value)을 고정시키지 않고 문헌과 색인어의 관계값에 따라서 유동적으로 변동하도록 하였다. 기준적합도값 α는 다음과 같은 식에 의해 계산되도록 하였다.

$$\alpha = \frac{\sum_{i=1}^{n} D_i}{n-1}$$

검색결과의 산출을 위해서는 세개의 색인어 t_{20}(protein engineering), t_{92}(enzyme activity), t_{11}(genetic manipulation)을 개별적인 질의식으로 한 검색과, 이들 색인어들을 AND, OR를 통해 조합한 질의식과 AND와 NOT을 동시에 사용하여 두 개의 색인어를 조합한 질의식에 의해서 실험을 수행하였다.

〈표 3〉에서 나타나는 것과 같이 검색어 t_{20}에 대한 결과는 예상했던 대로 적합도값이 높을 수록 적게 나타나고 적합도값이 낮을 수록 높게 나타나고 있다. 물론 적합도 0.0에서 240개의 검색결과는 실험용 DB 전체의 값이다.

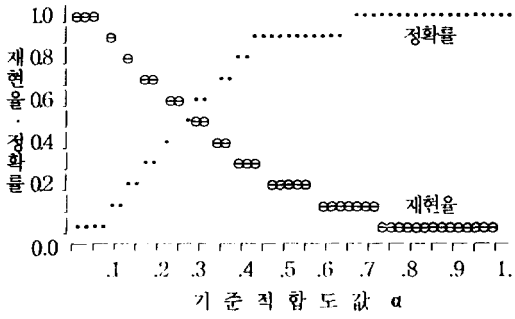
〈표 3〉 색인어20(protein engineering)의 검색결과
〈Table 3〉 retrieval results of t_{20}(protein engineering)

		문헌군서열										
		1	2	3	4	5	6	7	8	9	10	11
	1.0	2										
적	0.9		3									
	0.8			3								
합	0.7				5							
	0.6					6						
도	0.5						7					
	0.4							14				
값	0.3								27			
	0.2									31		
	0.1										36	
	0.0											240

〈표 3〉에서 나타나는 것과 같이 검색어 t_{20}에 대한 결과는 예상했던 대로 적합도값이 높을 수록 적게 나타나고 적합도값이 낮을 수록 높게 나타나고 있다. 물론 적합도 0.0에서 240개의 검색결과는 실험용 DB 전체의 값이다.

(그림 1)은 색인어20과 11을 AND 연산자와 NOT 연산자를 통해 조합한 질의식에 의한 실험결과이다.

이들 질의식에 대한 재현율과 정확률 평균치는 기준적합도값 α를 0.52로 했을 때 평균재현율 0.58이었고 평균정확률은 0.46이었다. 이 결과는 일반 집합이론에 의한 방법에 있어서 평균재현율 0.48과 평균정확률 0.43과 비교해 볼 때, 정확률에 있어서는 거의 비슷하다고 볼 수 있으나 재현율에 있어서는 상당히 높은 값을 보이고 있음을 확인할 수 있었다. 또한 선행 연구[21]에서 보인 평균재현율 0.53과 평균정확률 0.42보다도 상당히 개선된 결과가 나왔다.



(그림 1) 색인어20과 11에서 기준적합도값, 재현율, 정확률의 관계($t_{20} \wedge \neg t_{11}$)
(Fig. 1) relation between , RR, and PR in t_{20} and t_{11} ($t_{20} \wedge \neg t_{11}$)

7. 결 론

일반 집합이론에 기초한 부울 검색기법 이후에 개발된 정보검색 기법들은 논리적으로 또는 실험적으로 수행된 결과들을 볼 때 의미있는 가능성을 제시하고 있다. 물론 대부분이 실험적 시스템을 통해 소개된 이들 기법은 실제로 활용되는 대규모 시스템에서 적용된 것이 아니고 비교적 소규모의 실험집단을 대상으로 한 것이다.

이러한 검색기법들은 모두 그 수에 있어서 차이는 있으나 나름대로의 문제점과 한계성을 지니고 있다.

이는 정보검색이라고 하는 인간의 지적 행동과정을 시뮬레이션하는데 있어서의 어려운 점 때문이다. 정보검색 기법에 대한 연구의 의미는 선행연구들에서 고려하지 못한 새로운 관점, 상이한 기법들의 조합, 기법 자체가 가지고 있는 모순의 극복을 통한 보다 나은 모델을 제시하는데 있다.

정보검색에 있어 전통적인 불리안 검색에 대한 다수의 대안이 제시된 가운데서 퍼지집합 이론은 그중 돋보이는 역할을 수행했다고 할 수 있을 것이다. 이는 퍼지이론이 인간의 검색 시맨틱스에 가장 근접한 과정을 제시하고 있기 때문이다.

본 연구에서는 퍼지 집합이론을 적용한 문헌검색 시스템을 구현함으로써 용어와 이를 통해 표현된 문헌간의 관계를 일반화하고자 하였으며, 이는 색인용어와 문헌사이의 관계값으로서 퍼지관계를 채택함으로써 수행하였다.

이러한 퍼지 집합이론 및 퍼지논리에 근거한 문헌검색 시스템의 장점은 문헌을 표현하는 색인어의 다양한 관계도에 따른 다중 적합도와 의미적 관계값의 표현력에 있으며, 집합이론과 2치[0, 1] 논리에 기반을 둔 검색방법은 퍼지 집합이론과 퍼지 논리에 근거한 방법들의 특수한 경우이므로 더 일반적인 문헌검색 이론을 개발하는 것이 가능하다는 점 등을 들 수 있다.

각 색인어들과 불리안 연산자인 AND, OR, NOT으로 이들 색인어들을 조합한 질의식을 통해 실험을 수행한 결과 워크스테이션 환경에서의 실험적 시스템에서 일반 집합이론에 의한 검색실험에서보다 상당히 우수한 성능을 보였다. 특히 재현율과 정확률을 측정한 성능평가 결과는 퍼지 문헌검색 시스템이 가능한 검색 대안이라는 사실을 실험으로 입증할 수 있다. 즉, 두개의 색인어를 AND 연산자와 NOT 연산자를 통해 조합한 질의식에 의해서 재현율과 정확률을 구하는 실험을 한 결과, 이들 질의식에 대한 재현율과 정확률 평균치는 기준적합도값 α를 0.52로 했을 때 평균재현율이 0.58이었고 평균정확률은 0.46이었다. 이 결과는 선행 연구에서 보인 평균재현율 0.53과 평균정확률 0.42보다도 상당히 개선된 결과이다.

한편, 검색의 기법 측면에서 고려하였을 때, 본 연구는 먼저, 색인어 관계값을 통해 검색문헌에 서열을 부여하였고, 기준적합도값의 변동에 따라 검색결과가 유동적으로 대응하도록 하였으며, 관계값을 의미적 거리로 파악함으로써 검색과정과 검색사고를 일치시키고자 시도하였다.

한편, 본 연구를 수행하면서는 실험영역을 유전공학 단일영역으로 국한하였는데, 이는 이후의 계속 연구를 통하여 다른 영역의 DB를 확보하고 비교실험을 통해 보완하고자 한다.

참 고 문 헌

[1] 오길록, 이광형. 퍼지이론 및 응용 I, II. 서울:홍릉과학출판사, 1991.

[2] Cooper, W. S. "Exploiting the Maximum Entropy Principle to Increase Retrieval Effectiveness," *Journal of the American Society for Information Science* Vol. 34 No. 1 pp. 34-39, 1983.

[3] Cooper, W. S. "Getting Beyond Boole," *Information Processing and Management* Vol. 24, No. 3, pp. 243-248, 1988.

[4] 정영미. 정보검색론. 서울:구미무역, 1988.

[5] 강일중. 용어간 관계를 이용한 검색문헌의 순위부여에 관한 연구. 서울:연세대학교 대학원 문헌정보학과. 석사학위논문. 1990.

[6] Salton, G. and McGill, M. J. *Introduction to Modern Information Retrieval* New York:McGraw Hill, 1983. pp. 52-117.

[7] Belkin, N. J. and Croft, W. B. "Retrieval Techniques," *Annual Review of Information Science and Technology* Vol. 22, pp. 109-146, 1987.

[8] 김영귀. "완전매치와 부분매치기법에 관한 연구," 정보관리학회지. Vol. 7, No. 1, pp. 79-95, 1990.

[9] 이순재. 1989. "정보검색 시스템에 Fuzzy Set 이론의 적용," 도서관 정보학연구(경북대학교 대학원 도서관 정보학과) Vol. 1, pp. 201-236, 1989.

[10] Fox, E. A. and Koll, M. B. "Practical Enhanced Boolean Retrieval:Experiments with the SMART and SIRE Systems," *Information Processing and Management* Vol. 24, No. 3, pp. 257-267, 1988.

[11] Ogawa, Y. et al. "A Fuzzy Document Retrieval System Using the Keyword Connection Matrix and a Learning Method," *Fuzzy Sets and Systems*

Vol. 39, pp. 163-179, 1991.

[12] Buell, D. A. and Kraft, D. H. "Threshold Values and Boolean Retrieval systems," *Information Processing and Management* Vol. 17, No. 3, pp. 127-136, 1981.

[13] Radecki, T. "Outline of a Fuzzy Logic Approach to Information Retrieval," *International Journal of Man-Machine Studies* Vol. 14, pp. 169- 178, 1981.

[14] Radecki, T. "A Theoretical Background for Applying Fuzzy Set Theory in Information Retrieval." *Fuzzy Sets and Systems* Vol. 10, pp. 169-183, 1983.

[15] Murai, T. et al. "A Modeling of Search Oriented Thesaurus Use Based on Multivalued Logical Inference," *Information Science* Vol. 43, pp. 185-212, 1988.

[16] Miyamoto, S. "Information Retrieval Based on Fuzzy Associations," *Fuzzy Sets and systems* Vol. 38, pp. 191-205, 1990.

[17] Nomoto, K. et al. "A Document Retrieval System Based on Citations Using Fuzzy Graphs," *Fuzzy Sets and Systems* Vol. 38, pp. 207-222, 1990.

[18] Enrado, Patty. "Fuzzy Retrieval," *AI Expert* Vol. 10, No. 1, p. 48. 1995.

[19] Cox, Earl. "Relational database queries using fuzzy logic," *AI Expert* Vol. 10, No. 1, pp. 22-30, 1995.

[20] Cox, Earl and Goetz, Martin M. "Querying massive databases using natural language concepts," *Computer World* Vol. 25, No. 10, pp. 71, 1991.

[21] 이승채. "퍼지개념을 적용한 질의식의 분석과 문헌정보 검색에 관한 연구," 도서관학(한국도서관학회) 제21집, pp. 249-290, 1991.

[22] Murai, T. et al. "A Fuzzy Document Retrieval Method Based on Two-Valued Indexing," *Fuzzy Sets and systems* Vol. 38, pp. 191-205, 1989.

[23] Buell, D. A. "A Problem in Information Retrieval with Fuzzy Sets," *Journal of the American Society for Information Science* Vol. 36, No. 6, pp. 398-401, 1985.

[24] Chiaramella, Y & Defunde, B. "A Prototype of an Intelligent System for Information Retrieval," *Information Processing and Management* Vol. 23, No. 4, pp. 285-303, 1987.

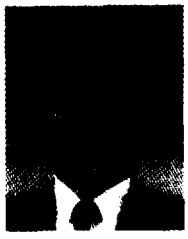
[25] Belkin, N. J., et al. "ASK for Information Retrieval: Part I," *Journal of Documentation* Vol. 38, No. 2, pp. 61-71, 1982.

김 철

1982년 전남대학교 계산통계학과(학사)
1985년 전남대학교 대학원 계산통계학과(이학석사)
1991년 전남대학교 대학원 전산통계학과 박사수료
1992년~현재 광주교육대학교 전산교육과 조교수
관심분야: 퍼지문헌정보검색, 컴퓨터교육, 멀티미디어시스템 등

이 승 채

1984년 전남대학교 문헌정보학과(학사)
1987년 연세대학교 대학원 문헌정보학과(석사)
1995년 연세대학교 대학원 문헌정보학과(박사)
관심분야: 퍼지문헌정보검색, 정보서비스, 디지탈 라이브러리 등



김 병 기

1978년 전남대학교 수학교육과(학사)
1980년 전남대학교 대학원수학과(이학석사)
1981년~현재 전남대학교 전산학과 교수
관심분야: 소프트웨어 공학, 신경망 컴퓨터, 초고속 정보통신 등