전통적인 정보검색기법과 퍼지정보검색의 비교 연구

울산대학교 의학도서관

최 홍 식

A Comparative Study on the Traditional IR Methods and Fuzzy IR

Hung-Sik Choi

Medical Library, Ulsan University

I. 서 론

정보검색시스템은 대상문헌의 주제분석과 이용자의 요구분석을 통하여 양자간의 커뮤니케이션을 증진시키는 시스템이다. 현행 정보검색 시스템에서 가장 널리 사용되고 있는 불리안검색은 이용자의 정보요구를 정확하고 간단하게 표현할 수 있다는 장점이 있다. 그러나 탐색어로 표현되는 각 개념의 상대적인 중요도를 나타내지 못한다는 점, 문헌과 질문간에 유사도의 크기에 따라 검색문헌을 등급화 할 수 없다는 점, 탐색문헌과 완전히 일치되는 문헌만이 검색되기 때문에 부분적으로 일치하는 문헌은 검색할 수 없다는 점 등의 단점이 있다.

이러한 불리안검색의 문제점을 해결하기 위하여 용어의 출현빈도수를 통계적으로 처리하여 가중치를 부여하거나 확률적으로 연산처리하는 검색기법이 모색되고 있다. 그러나 정보검색은 본질적으로 불확실성과 불완전성이 내포되어 있기 때문에 이와같은 통계적인 기법이나 확률을 적용한 검색기법은 전통적인 검색기법의 문제점을 완전히 해결하지 못하고 있다. 따라서 정보검색에 수반되는 애매모호함을 처리하기 위해서는 보다 새로운 대안이 필요한데 이러한 정보처리의 이론적인 바탕이 되고 있는 것이 퍼지(Fuzzy) 이론이다.

따라서 본고에서는 전통적인 정보검색 기법의 문제점을 고찰하고 이러한 문제점을 해결하기 위한 대안으로써 퍼지집합을 적용한 정보검색의 이론적 배경과 국내외 연구동향을 살펴보고 이를 상호 비교하여 봄으로써 앞으로의 정보검색시스템이 나아가야할 방안을 모색하고자 한다.

II. 전통적 검색기법의 문제점

Blair는 정보검색의 기법을 12가지 즉, 단일디스크립터(Single Descriptors), 다중디스크립터(Multiple Descriptor), 임계치 부여(Cut-off), 문헌등급화(Ranking), 질의어가중치(Weighted Requests), 색인어 가중치(Weighted Indexing), 질의어 및 색인어 가중치(Weighted Request and indexing), 코사인계수(Cosin Rule), 불리안검색(Boolean Request), 전문탐색(Full Text), 이진 시소러스(Binary Thesaurus), 가중치 시소러스(Weighted Thesaurus)로 구분하여 장단점을 비교한 바 있다(Blair 1990). 이를 크게 대별하여 이용자의 탐색어와 문헌을 표현하는 색인어를 비교하는 기법으로 정의하면 완전매치와 부분매치 기법으로 대별할 수 있다(Belkin, 1987; Shoval, 1985; 김영귀, 1990). 이는 검색된 문헌의 집합이 질문과 정확히 일치하는가 아니면 부분적으로 일치하는가에 따라 완전과 부분으로 구분한 것이다. 완전매치기법으로는 불리안검색이 대표적이며 부분매치기법으로는 불리안논리의 문제점을 향상시키기 위한 방안으로써 확률검색, 벡터공간검색등이 있다. 이

를 간략하게 살펴보면 다음과 같다.

1. 불리안 검색

불리안검색은 현행 정보검색 시스템에서 가장 널리 사용되고 있는 전통적인 검색기법으로서 이용자의 정보요구를 논리적(AND), 논리합(OR), 논리부정(NOT)과 같은 집합연산자를 조합하여 검색하는 기법이다. 불리안논리에 의한 탐색문은 이용자가 작성하기가 편리하고 컴퓨터 처리가 용이하며 이용자가 정보요구를 표현할 때 탐색패턴을 공식화 할 수 있어서 배우기 쉽다는 장점이 있다(Bookstein, A., 1978). 즉 탐색문이 정보요구를 나타내는 탐색어와 그들간의 논리적 관계로 구성되므로 정보요구를 비교적 정확하고 간단하게 표현할 수 있어 이용자가 작성하기가 편리하고, 대부분의 이용자들의 정보요구에 대한 개념형성시 조작의 융통성이 높으며 입력요건이 단순하므로 배우기가 쉽다. 그러나 이러한 장점에도 불구하고 많은 사람들이 불리안검색의 문제점을 지적하고 있다. 여러 선행연구에서 지적되고 있는 주요 문제점을 요약하면 다음과 같다.

① 색인어와 탐색어가 완전히 일치하는 문헌만 검색되고 부분적으로 일치하는 문헌은 검색이 불가능하다. (Cooper, 1983)

② 집합연산자 AND, OR에 의한 연산결과를 이해하지 못함으로서 이용자 등에게 혼동을 야기 시킨다 (Salton, 1985). 즉, OR연산자인 경우에 오직 하나의 용어만 포함된 문헌일지라도 모든 용어를 포함한 문헌과 동이하게 검색된다. AND연산자인 경우에는 하나의 용어만 부족 한 문헌일지라도 탐색어가 전혀 포함되지 않은 문헌과 마찬가지로 검색되지 않는다.

③ 이용자나 색인자가 질의어나 색인레코드에 관련된 주제어의 상대적인 중요성을 표현할 수가 없다 (Bookstein, 1980; 1981)

④ 문헌과 문헌간의 관계나 유사도, 그리고 색인어와 탐색어간의 관계를 고려하지 않기 때문에 검색된 문헌의 등급화가 불가능하다(Salton, 1975)

이상과 같은 단점에도 불구하고 이 기법이 현행시스템에서 널리 사용되고 있는 것은 다른 시스템으로 달리 전환하려 하여도 경제적으로 실행할 수 없을 정도로 투자가 많으며, 또 대안적으로 제기되고 있는 기법들이 소규모 환경에서만 시험되고 대규모 환경에서는 시험되지 않았고, 그 대안적 기법의 연구 결과들이 기대치를 정당화할 정도로 충분하지 않았기 때문이라고 할 수 있다.

2. 확률 검색

위와같은 불리안검색의 문제점을 해결하기 위해서 색인어와 탐색어에 개념의 중요도를 나타내는 가중치를 부여하고 색인어와 탐색어의 유사성 정도에 따라 검색문헌을 등급화하는 확률검색 기법이 등장하였다. 이러한 가중치를 부여하는 목적은 첫째, 특정한 색인어의 가중치나 혹은 그 가중치의 합이 일정치 이상인 문헌만 검색하므로서 탐색문과 어느정도 관련성이 높은 문헌을 검색하여 결과적으로는 정도율(Precision)을 높이기 위한 것이고 둘째, 검색문헌을 가중치의 합이나 유사성계수의 크기 순서로 출력함으로써 부적합한 문헌을 모두 살펴보아야 할 이용자의 시간과 노력을 감소시키기 위한 것이다. 확률검색은 질문에 대한 각 문헌이 적합할 확률(Probability of Relevance)과 부적합할 확률(Probability of non-Relevance)을 계산하여 적합할 확률이 부적합할 확률보다 큰 문헌순으로 검색하는 기법이다. 1960년에 Maron과 Kuhns는 최초로 색인자가 문헌의 색인어에 가중치를 부여하는 확률개념을 정보검색에 적용하였다(Maron & Kuhns, 1960). 그후에 Losee와 Robertson은 탐색요구에 대한 시스템의 응답이 이용자에게 적합할 확률에 따라 검색된 문헌을 등급화하는 확률등급원칙(Probability Ranking Principle)을 제시하였다. 확률검색은 불리안연산자를 사용하지 않기 때문에 불리안검색에 비판적인 견해를 가진 사람들에게 환영을 받았고, 문헌의 등급화가 가능하며 이용자의 피드백정보의 활용이 용이 하다는 장점이 있으나 다음과 같은 문제점이 지적되고 있다.

① 적합문헌의 결정은 확률에 근거를 두기 때문에 근본적으로 오차를 내포하고 있다.

② AND와 OR연산자를 동일하게 처리하고 있다 (Bookstein,1985)

③ 탐색문이 이용자에 의해 직접 작성되는 것이 아니라 의사결정함수가 시스템에 의해 적합성 피드백을 거쳐 작성된다는 불편한 점이 있다(Wong, 1989)

이러한 확률검색은 이론적으로는 매력이 있는 검색모델이지만 실제상에 있어서 적합한 정보의 획득이 어

려운 문제로 인하여 사용상 문제가 뒤따른다 할 수 있다.

3. 벡터공간을 이용한 검색

벡터공간을 이용한 정보검색은 전통적인 불리안검색의 또다른 대안으로서 불리안연산 없이 색인어와 탐색어 간의 유사성을 계산하여 크기에 따라 검색하는 기법이다. Raghavan은 벡터공간을 이용한 정보검색의 개념을 모델화 하였으며 Jones는 문헌벡터와 탐색어 벡터간의 유사성을 측정하기 위하여 코사인계를 사용하는 방안을 제시하기도 하였다.(Raghavan, 1986; Johnes, 1987).

색인어와 탐색어간의 유사성에 따라 검색된 문헌을 등급화할 수 있으며 일차적으로 검색된 문헌중에서 이용자의 적합성 판정에 따른 피드백정보를 이용하여 탐색문의 수정이 가능하다. 또한 탐색어가 불리안연산자 없이 작성되기 때문에 이용자가 간단하게 탐색문을 작성할 수 있으며 검색된 문헌의 수를 통제 할 수 있다. 그러나 벡터를 이용한 정보검색은 탐색어의 모든 용어는 각각 독립된 것으로 처리하기 때문에 유사성 계수를 계산하는데 AND와 OR를 동일한 것으로 처리하며 NOT는 처리가 불가능하다는 문제점이 있다. 따라서 불리안논리의 문제점은 해결할 수 있으나 또다른 문제점을 야기 시키고 있다.

III. 퍼지 정보검색

Fox와 Koll은 이용자가 표현한 정보요구와 관련된 용어가 문헌에 포함되어 있지 않더라도 관련문헌을 찾아낼 수 있어야 하며, 비록 문헌에 포함되어 있더라도 부적합한 문헌이면 검색하지 않거나 그 순위를 낮출 수 있어야 만이 효율적인 정보검색 시스템이라 강조하고 있다. 이러한 목적을 달성하기 위해 지금까지 채택된 방법으로는 확률과 통계이론을 들 수 있다. 그러나 이러한 기법은 전술한 바와같이 불확실성과 불완전성이 내포되어 있는 정보검색의 근본적인 문제를 해결하지 못하고 있다. 따라서 불확실성이 내포된 색인이나 이용자의 정보요구를 효과적으로 처리하기 위해서 퍼지집합을 응용한 정보검색시스템이 점점 확대되고 있다.

1. 퍼지집합의 이론적 배경

Zadeh교수에 의해 처음으로 소개된 퍼지집합이론은 보통집합이론을 일반화 한 것이다(김성혁, 1990). 퍼지집합이론은 부분집합의 소속도(Membership)를 인정함으써 전통적인 일반 집합이론을 퍼지집합의 특수한 경우로 생각하고 있다. 전통적인 집합이론에서 특정 대상(Objects)은 집합의 요소든지 원소가 아니든지 둘중의 하나다. 다시말해서 이 집합이론은 집합의 요소인지 아닌지를 구분하는 경계선이 분명하다. 그러나 퍼지집합에서는 소속도를 인정함으로써 경계선을 분명하게 말할 수 없게 되었다. 이와같이 집합의 경계선이 불분명한 것을 퍼지라 한다. 각 퍼지집합에는 대상 전체의 각 구성요소에 0과 1사이의 값을 부여하고, 이는 집합과 각 구성요소간에 관련정도를 나타낸다. 집합의 소속도와 구성요소간에 관련정도를 나타내는 표(Table) 또는 규칙을 소속함수(Membership Funtion)라 한다(Bookstein, 1985).퍼지집합에서는 각 구성요소마다 소속함수 값을 함께 표시해 주어야 한다.

퍼지집합의 연산은 소속함수를 사용하여 다음과 같이 정의할 수 있다(Bookstein, 1980).

① 포함관계: 집합A는 집합B의 부분집합으로 문헌집단D에서 모든 x에 대해 $A \subseteq B$는 $m_A(x) \leq m_B(x)$를 나타낸다.

② 합집합: A와 B의 합집합(A OR B)는 AUB로 표시하고 $m_{A \cup B}(x)$의 소속함수 값은 $\max[m_A(x), m_B(x)]$로 정의한다.

③ 교집합: A와 B의 교집합(A AND B)는 $A \cap B$로 나타내고 소속함수 값은 $\min[m_A(x), m_B]$로 정의한다.

④ 여집합: 퍼지집합 A의 여집합은 $1-m_A(x)$로 정의한다.

위의 퍼지집합연산 정의에 의하면 불리안 대수에서 정의 되었던 연산정의는 그대로 만족되는 것을 알수 있다.(그러나 퍼지 여집합은 성립하지 않는다)

2. 퍼지집합과 정보검색

정보검색에 있어서, 문헌 퍼지집합은 각 용어와 연관되어 있다. 소속함수는 각 용어와 관련된 각 문헌의 범위 즉, 관련정도에 따라 0과 1사이의 값을 부여한

다. 전통적인 불리안 검색시스템과는 대조적으로 용어와 전혀 관련이없는 문헌과 용어와 핵심적으로 관련된 문헌간에 연속적인 특성이 있다. 이는 적합문헌과 부적합 문헌간에 경계선을 명확하게 구분할 수가 없다. 이런 의미에서 색인어에 적합한 문헌과 부적합한 문헌을 구분하는 경계선을 퍼지라 할 수 있다. 불리안검색시스템과 마찬가지로 색인어레코드를 우선 작성한다. 퍼지검색시스템에서 색인작성자는 단순히 문헌에 색인어를 할당하는데 그치지 않고 할당한 용어와 문헌간에 관련된 정도를 표시한다(Bookstein, 1985). 예를들어 문헌에 색인어를 부여할 때 색인작성자는 그 문헌이 해당 용어와 가장 많이 관련된 경우에는 1이라는 값을 부여하고, 해당 용어와 최소한의 관련이 있을 경우에는 0.1의 값을 주며, 그 이외의 경우에는 관련정도를 판단하여 적당한 중간값을 부여한다. 문헌에 부여되지 않은 색인어는 관련도가 제로(0)의 가중치를 갖고 있음을 의미하기 때문에 모든 색인어가 문헌에 할당된 가중치가 존재함을 알 수 있다. 이와같은 방식으로 일단 모든 문헌에 대해 색인작성이 이루어 지고나면 문헌관점에서 색인어 관점으로 소속함수가 정의된다. 이 소속함수를 통해 퍼지집합이 결정된다. 정보검색에 퍼지니스의 중간값을 인정함으로서 색인작성자는 불합리하고 절대적인 YES/NO로 결정하는 대신에 문헌에 적용되는 용어를 부분적으로 표현할 수 있다. 만일 소속함수값을 0과1 2개의 값으로만 제한하면 전통적인 일반집합과 동일한 결과를 초래한다. Zadeh의 논문이 발표된 직후 Bellman & Giertz는 비판받고 있는 불리안 공리를 충족시키기 위해서는 Zadeh가 정의한 합집합, 교집합연산이 유효함을 입증하였다(Bellman & Giertz, 1973). 또한 정보검색(IR) 차원에서 Bookstein은 간단한 예를들어 연산정의를 증명하였으며 퍼지 여집합에 대한 Zadeh의 정의를 뒷받침하는 논증을 제시하기도 하였다.

전술한 퍼지집합과 퍼지연산에 대한 정의가 이루어지면서 정보검색 과정에 큰 발전을 가져왔다. 탐색질의는 전통적인 방법과 마찬가지로 집합과 집합연산에 의해 처리되며 검색결과는 문헌집합으로써 이용자에게 제공된다. 그러나 퍼지검색시스템에서는 퍼지연산에 기초하고 검색결과 또한 퍼지집합이다. 퍼지연산은 문헌과 색인어간의 소속함수값을 통해 결정되며 검색결과는 문헌과 탐색질의어간의 소속함수로 정의할 수 있다. 검색된 집합의 소속함수값은 이용자에게 제시되는 문헌을 등급화하는데 사용되며, 가장 적합한 문헌(소속함수 값이 가장 큰 문헌)이 가장 먼저 나타난다.

예를들어, 문헌d1과 d2가 색인어 t1과 t2로 다음과 같이 색인되어 있다고 가정하자:

d1={(t1,.5),(t2,.8)}

d2={(t1,.9),(t2,.1)}

이를 색인어에 대한 집합형태로 전환하면을 다음과 같다:

t1={(d1,.5),(d2,.9)}

t2={(d1,.8),(d2,.1)}

여기서 이용자의 탐색질의가 "t1 AND t2"라면 이에 대한 응답으로 다음과 같은 집합이 검색된다.

(d1,.5),(d2,.1)

여기서 d1의 경우 0.5와 0.8중 최소값이 0.5이기 때문에 0.5의 가중치를 부여한다. 또한 검색된 문헌의 강도(Strength) 즉, 관련정도를 이용자에게 표시할 수 있다. 따라서 위의 예에서 검색된 문헌의 등급화는 d2보다 d1이먼저 배열된다.

3. 퍼지 정보검색의 연구동향

지금까지 정보검색 분야에 퍼지집합 이론을 적용한 연구는 ① 퍼지 색인시스템 ② 퍼지 시소러스 ③ 퍼지탐색어를 이용하거나 자연어 탐색어를 퍼지추론으로 처리하는 기법으로 대별할 수 있다. 정보검색에 퍼지이론을 적용한 국내외 연구동향을 살펴보고자 한다.

(1) 해외 연구동향: 먼저 해외 연구들은 퍼지색인시스템, 퍼지시소러스, 키워드관계행렬, 인용을 이용한 퍼지정보검색시스템등이 있다.

퍼지집합을 정보검색에 적용하려는 초창기의 시도로는 Negoita(1973)가 있다. Negoita의 연구목적은 퍼지논리 차원에서 적합성의 개념을 규명하는데 있었다. 특히, Negoita는 명제논리(Logic of Proposition)차원에서 정보검색과정을 규명하여 2價論理로부터 多値論理로 일반화하였다. Negoita & Flonder는 색인어에 대한 퍼지할당의 개념을 더욱 확대하였으며 이들의 연구와 더불어 Tahani, Radecki의 연구는 퍼지집합을 적용하여 정보검색 과정을 분석한 최초의 연구였다.

Tahani는 4차원(X,D,Q,r)으로 정의하여 정보검색 시스템의 수학적 모델을 제시하였다. 여기서 X는 문

현표현물 또는 색인레코드, D는 디스크립터 집합, Q는 탐색문 집합, r은 q와 x간의 관련정도에 따른 매칭함수를 나타낸다. 또한 매칭함수값에 따라 검색문헌을 등급화하는 퍼지리스트[L(d>]의 개념을 소개하였다(Tahani, 1976).

1976년에 Radecki는 퍼지시소러스를 이용한 정보검색시스템의 모델을 제시하였다. 그는 색인어 및 탐색어에 중요도를 표시할 수 있으며 색인어와 탐색어간에 관계를 퍼지시소러스를 통해 형식화하는 방안을 제시하였다(Radecki, 1976). 그러나 이들의 연구논문은 공식을 중심으로 기술되어 있기 때문에 수학적 배경이 없는 비전문가는 이해하기 어렵다는 문제점이 있다.

또한 Radecki는 1981년에 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 달라질 수 있으므로 객관성 유지에 문제가 있다(김현희, 1993).

그 이후의 연구로 Choros & Danilowitz는 퍼지검색시스템내에서 이용자의 피이드백에 따라 색인레코드의 자동 수정이 가능함을 증명하였다. 검색된 문헌에 대해서 그 문헌이 적합한지의 여부를 판단한 결과에 따라 색인레코드중에서 탐색질의와 일치하는 용어의 가중치를 약간씩 높이거나 낮추는 것이다. 이 기법은 논리부정(NOT)에 해당하는 용어의 처리가 곤란하다. 또한 이 기법은 또한 형식이 다른 동등한 탐색질의에 대해 그 검색결과가 상이하다는 문제점이 있다.

검색된 문헌의 이에 대한 대안으로 적합성 판정에 따라 탐색질의어를 수정하는 대신에 검색된 문헌의 색인레코드 자체를 수정하는 대안적인 방법이 있다. 예를들어, 각 색인어를 대상으로 적합한 문헌의 색인레코드와 부적합한 문헌의 색인레코드에 대한 평균가중치를 계산할 수 있다. 적합한 문헌으로 부터 도출되는 평균값을 위한 함수가 적합한 문헌의 가중치에 추가된다. 부적합한 문헌에 대해서도 동일한 과정이 수행된다. 이 기법은 각 문헌을 벡터로 처리하여 탐색질의에 적합성 여부의 평가기준이 되는 평균벡터 또는 센트로이드를 계산한다. 여기서의 수정은 적합한 문헌에 적합한 문헌에 대한 센트로이드함수가 추가되는 과정이 수반된다. 물론 부적합한 문헌에 대한 수정도 동일한 과정이 수행된다. 이러한 방법으로 탐색질의에 포함된 색인어 뿐만아니라 모든 색인어에 대해 새로운 색인레코드가 형성된다(Bookstein, 1985).

지금까지 색인과정에서 수반되는 퍼지니스를 표현하고 불리안 탐색문의 상대적 중요도 표시를 위한 시도였다. 그러나 이러한 시도를 충족시키기란 대단한 어려움이 따른다는 것이 입증되었다.

Radecki와 Buell & Kraft는 기준치(Threshold Value; λ)를 설정하는 방안을 제시하였다. 이 기준치는 탐색질의어에서 각 구성요소가 갖는 상대적 중요도를 표시한다. 처리단계에서 구성요소에 대한 소속함수 값이 기준치(λ)이상인 문헌만이 퍼지집합으로 검색된다. 공식으로 나타내면, 퍼지집합A에 대해 문헌d가 $A(d)>\lambda$면 A(d)의 관련도는 $A\lambda$로 나타내며 기준치 이하면 관련도는 0이다. 여기서 A(d)는 원래의 퍼지집합 A에 대한 문헌d의 소속도를 나타낸다. 용어의 중요도가 떨어질수록 기준치 λ의 값을 높여 줌으로써 그 용어의 핵심적인 문헌만이 제시된다. 이러한 공식은 퍼지검색시스템으로써 대수학적 특성을 충족 시켜주지만 λ가 중요도의 척도로 간주된다면 정보검색에서 모든 조건을 만족시키지는 못한다. 예를들어 A_i은 탐색질의(A_i OR B)에서는 탐색결과에 영향을 끼치지 않는다. 그러나 탐색질의(A_i AND B)에서 A_i은 탐색결과에 매우 강한 영향을 끼친다. 앞서 언급한 퍼지교집합의 정의에 의하면 A_i은 교집합내에 모든 문헌을 제로(0)의 수준으로 만들어 버린다. 이러한 문제를 해결하기 위해 Bookstein(1980)은 중요도 변수(Parameter)인 λ를 다음과 같이 사용하였다. 합집합 연산의 멤버인 경우에는 집합의 소속함수값에 λ를 곱하고, 교집합 연산의 멤버인 경우에는 소속도값을 λ로 나눈다(나눈값의 몫이 1보다 클 경우에는 1로 한다). 이렇게 함으로써 용어의 중요도 개념과 일치하고 불리안 대수와 관련된 구조적 특성도 만족 시킨다. 그러나 이 정의는 Waller & Kraft가 말한 분리성에 위배된다. 따라서 Kantor(1981)는 대수학적 차원에서 퍼지집합 검색을 구체적으로 분석하고 불리안 대수의 공리는 일관성있는 시스템을 위해 희생되어야 한다고 결론짓고 있다.

일본 Tsukuba 대학 교수인 Miyamoto는 단어의 동시출현빈도와 퍼지집합연산에 기초한 정보검색시스템을 제안하여 과학정보처리센터에 적용하였다(Miya-

moto, 1990). 정보검색시 탐색어를 포함하는 문헌은 물론 탐색어의 동의어, 관련어 등을 포함한 문헌을 검색하기 위하여 시소러스를 이용한다. 여기서의 시소러스는 전통적인 시소러스와는 달리 용어와의 관계를 나타내는 가중치가 단어에 부여된다.

1991년에 Ogawa는 통계정보를 이용하여 퍼지시소러스와 유사한 키워드 관계행렬을 통한 정보검색을 제안하였다. 두개의 색인어가 동시에 포함된 문헌빈도수 즉, 키워드의 초기관계값을 측정할 수 있는 공식을 제시하였다. 키워드 관계행렬을 이용하여 문헌집단의 이진색인을 퍼지색인으로 변환하고 문헌은 질문에 대한 관련정도에 따라 등급화 된다(Ogawa, 1991).

Nomoto는 1990년에 퍼지그래프를 이용하여 인용에 기초한 정보검색시스템을 구축하였다. 이들은 문헌간의 관련도를 그래프의 정점과 변이 각각 문헌과 인용을 나타내는 유향퍼지그래프를 통해서 계산하였다.

(2) 국내 연구동향: 국내의 연구동향을 살펴보면 1989년에 이순재는 퍼지이론을 적용한 정보검색시스템의 기본개념을 고찰하였고(이순재, 1989), 1990년에 조혜민은 전통적인 불리안검색시스템의 단점을 보완하기 위하여 불리안검색시스템에 가중치를 부여하는 시스템을 설계하였다(조혜민, 1990). 색인어와 탐색질의어에 모두 가중치를 부여하고 질의에 대한 각 문헌의 관련정도를 계산하는데 퍼지집합을 이용하였다.

이승채는 1991년에 퍼지집합 이론을 적용한 정보검색시스템을 구현함으로서 용어와 이를 통해 표현된 질의식과 문헌간의 관계를 일반화 하고자 하였으며, 색인어와 문헌에 관계값으로서 이치관계가 아닌 퍼지관계를 채택함으로서 수행하였다. 본 연구에서 색인어들과 불리안 연산자인 AND,OR, NOT로 이들 색인어들을 조합한 질의식을 통해 실험한 결과 PC환경에서의 실험적 환경에서 기존의 일반집합이론에 의한 검색실험보다 우수한 성능을 보였고, 특히 재현율과 정확률을 측정한 결과는 퍼지 정보검색시스템이 효율적이라고 결론짓고 있다.

1993년 김현희, 배금표는 이진색인체제를 유지하면서 퍼지시소러스를 통해 퍼지정보검색시스템을 구현할 수 있는 시스템을 구축하고 그 검색결과를 불리안검색결과와 비교/분석하였다(김현희, 1993). 형태소해석방법에 의한 자동색인기능, 통계정보와 퍼지집합연산을 응용한 퍼지시소러스 화일 구축기능, 퍼지집합 검색기능을 갖는 퍼지정보검색시스템을 구현하고 재현율과 정확률을 통해 불리안시스템과 비교한 결과 재현율은 퍼지검색이 75%로 불리안보다 15% 높았고, 정확률의 경우 불리안검색이 73%로 퍼지검색보다 4% 높다는 결론을 얻었다.

1993년 이준호 등은 전통적인 Max/Nin 연산자의 부정적 특성을 지적하고 퍼지연산자의 특성을 분석함으로서 높은 검색효율을 제공할 수 있는 긍정적 보상연산자를 정의하고 이들을 퍼지집합 모델의 불리안연산자 계산식으로 사용할 것을 제안하였다(이준호등, 1993). 또한 현재까지 개발된 긍정적 보상연산자는 탐색어의 중요도를 왜곡하여 문서값을 계산하기 때문에 긍정적 보상연산자를 다항연산식으로 확장하였다. 긍정적 보상연산자를 사용하는 퍼지집합 모델이 다른 연산자를 사용하는 모델보다 높은 검색효율을 제공하고 긍정적 보상 연산자의 다항연산식이 이항 연산식보다 검색효율이 우수함을 실험을 통해 입증하고 있다.

1993년에 복명희는 퍼지이론을 적용한 기존의 검색기법은 색인어와 탐색어의 유사성 계수를 계산하는데 있어서 Max/Min집합연산을 따르는 단점을 지적하고 이를 개선하기 위해서 새로운 형태의 매칭함수에 따라 유사성 계수를 계산하여 최종적으로 검색문헌의 순위가 부여되는 알고리즘을 제안하였다(복명희, 1993). 또한 소규모 문헌집단을 대상으로 실험을 하고 내적계수와 코사인 계수 그리고 전통적인 퍼지집합 검색방법으로 유사성을 계산하고 그 결과를 이용자의 적합성 판단과 비교/평가 하였다. 실험결과 퍼지색인과 퍼지탐색문, 그리고 퍼지시소러스를 이용한 확장된 탐색문까지 모두 처리가 가능하며,퍼지색인어와 퍼지탐색어의 유사성 계산에 있어서 실제로 이용자의 적합도 판단 절차보다 근접하는 함수식에 의해 검색의 효율을 향상 시킬수 있으며 기존의 퍼지검색에서 해결하기 못했던 NOT의 처리가 가능하다는 결론을 얻고 있다.

IV. 결 론

기존의 검색기법인 불리안검색은 이용자들에게 익숙하고 컴퓨터 처리가 용이하다는 장점이 있는 반면 단어와 단어 사이의 중요도를 표시하지 못하고 완전히 일치되는 문헌만이 검색되는 등의 몇가지 단점들이 있다. 이러한 문제점을 해결할 수 있는 하나의 방안으로

퍼지집합검색이 많은 사람들에 의하여 연구되고 있다.

전통적인 집합이론의 일반화로서 고안된 퍼지집합이론이 우리에게 관심을 끄는 이유는 인간의 애매한 표현을 처리할 수 있기 때문이다. 즉 엄밀하지 않은 용어들이 잘 정제된 수학적 방법으로 표현되고 처리되기 때문일 것이다. 그러나 퍼지집합이론에 근거한 검색시스템은 근본적으로 부울구조안에서 검색을 향상시킨다는 점에서 한계를 가질 수 밖에 없다고 하겠다. 이를 포함하여 전술한 여러가지 문제점을 해결하기 위하여 많은 연구를 통하여 규명되어야 할 점을 종합하여 요약하면

첫째, 정확한 색인어를 추출할 수 있어야 하고 그 과정이 자동적으로 수행되어야 한다.

분명하고 정확한 색인어를 추출하여야 함은 물론이고 대량으로 입수되는 자료들을 자동으로 입력하고 색인하는 일까지 수행하여야 만이 현실적으로 실용성과 경쟁력을 가진다 하겠다.

둘째, 가중치 부여에 대한 분명한 기준이 제시되어야 한다.

문헌에 할당된 디스크립터에 정확한 가중치를 부여하는 것은 매우 어려운 작업이며 색인자의 주관에 따라 많이 달라질 수 있을 뿐만 아니라 객관성유지에 문제가 있으므로 어떤 기준이 마련되어 적합한 가중치가 부여될 수 있게 하여야 한다.

셋째, 정확한 색인어와 적합한 가중치가 주워지면 실제의 검색환경에서 적용하여 봄으로써 보다 완벽한 이론으로 정립하고 현장기술로 적용하기 위하여 지속적인 연구가 요망되어야 한다.

넷째, 대부분의 상업용 데이타베이스가 이진색인체계를 기초로 하고 있어서 퍼지색인체계가 필요한 퍼지집합검색의 실용화를 위하여 좀 더 효율적인 방법이 요망되어야 한다.

참 고 문 헌

1) 김성혁: "전문가 대체시스템에서의 퍼지 추론에 관한 연구", 정보관리학회지 *7(1): 68-78, 1990*

2) 김영귀: "완전매치와 부분매치 검색기법에 관한 연구" 정보관리학회지 *7(1): 79-95, 1990*

3) 김현희: 계량정보학, 서울: 구미무역, *1993, p247*

4) 배금표: "퍼지정보검색시스템의 검색효율에 관한 연구", 정보관리학회지, *10(1): 31-54, 1993*

5) 복명희: "퍼지집합에 의한 검색문헌의 순위부여 연구", 중앙대학교 대학원 문헌정보학과 석사학위논문, *1993*

6) 이순재: "정보검색시스템에 *Fuzzy Set*이론의 적용", 도서관. 정보학연구, *1: 201-234, 1989*

7) 이승채: "퍼지개념을 적용한 질의식의 분석과 문헌정보검색에 관한 연구", 도서관학 *21: 249-289, 1991*

8) 이준호 등: "퍼지집합 모델의 검색효율 개선을 위한 퍼지 연산자의 분석", 정보관리학회지 *10(1): 53-63, 1993*

9) 조혜민: "퍼지논리를 이용한 가중치 부울정보검색시스템", 서강대학교 공공정책대학원, 석사학위논문 *1990*

10) Belkin NJ, Croft WB: "*Retrival Techniques*" *ARIST 22: 109-131, 1987*

11) Bellman R, Giertz M: *on the Analytic Formalism of the Theoty of Fuzzy Set, Information Science 5: 149-156, 1973*

12) Blair DC: "*Language and Representarion in Information Retrival*" *New York: Elsevier Science Pub., 1990, pp27-67*

13) Bookstein A: "*On the Perils of Merging Boolena and Weighted Retrival System,*" *JASIS 29(3): 156, 1978*

14) "*Fuzzy Request*", *JASIS 31(3): 240-24, 1980*

15) "*Fuzzy Request: an Approach to Weighted Boolean Search*" *JASIS July, 1980, pp241-242*

16) "*A Comparison of two Systems of Weighted Boolean Retrival*", *JASIS 32(4): 275, 1981*

17) "*Probability and Fuzzy Set Application to Information Retrival*" *ARIST 20: 120, 1985*

18) Cooper WS: "*Exploiting the Maximum Entropy Principle to Increase Retrival Effecieness*" *JASIS, 34(1): 275, 1983*

19) Fox EA, Koll MB: "*Practical enhanced Boolean retrieval: experiences with the SMART and SIRE systems*", *IPM 24(3): 260, 1988*

20) Jones W: "*Picture of Relevance: A Geometric Analysis of Similiarity Measure*" *JASIS 38(6): 423, 1987*

21) Maron ME, Kuhns J: "*On Relevance, Probabilistic Indexing and Information Retrival*" *JASIS, 7(3): 216-244, 1960*

22) Miyamoto S: "*Information Retrival Based on Fuzzy Association, Fuzzy Sets and Systems*", *38: 191-205, 1990*

23) Negoita C: "*on the Application of the Fuzzy Set Separation Theorem for Automatic Classification in IR, Information Science 5: 279-286, 1973*

24) Ogawa Y: *"a Fuzzy Document Retrival System Using the Keyword Connection Matrix and a Learning Method, Fuzzy Sets and System, 39: 163-179, 1991*

25) Radecki T: *"Mathematical Model of information Retrival System Based on a Conceptof Fuzzy Thesaurus", IPM 12(5): 313-318, 1976*

26) Salton G: *"Advanced Feedback Method in Information Retrival", JASIS 36(3): 204, 1985*

27) Shoval P: *"Principles Procedures and Rules in an Expert System for Information Retrival" IPM 26(6): 476, 1985*

28) Tahani V: *"a Fuzzy Model of Document Retrival System", IPM 12(3): 177-188, 1976*

29) Wong S: *"A Probability Distribution Model for Information Retrival", IPM 25(1): 39, 1989*