경영정보학연구
제1권 제1호
1991년 7월

퍼지 논리를 이용한 사용자 중심적인 Full-Text 검색방법에 관한 연구

이 원 부*

Consideration of a Robust Search Methodology that could be used in Full-Text Information Retrieval Systems

The primary purpose of this study was to investigate a robust search methodology that could be used in full-text information retrieval systems. A robust search methodology is one that can be easily used by a variety of users (particularly naive users) and it will give them comparable search performance regardless of their different expertise or interests. In order to develop a possibly robust search methodology, a fully functional prototype of a fuzzy knowledge based information retrieval system was developed. Also, an experiment that used this prototype information retreival system was designed to investigate the performance of that search methodology over a small exploratory sample of user queries. To probe the relatonships between the possibly robust search performance and the query organization using fuzzy inference logic, the search performance of a shallow query structure was analyzed. Consequently the following several noteworthy findings were obtained: 1) the hierachical (tree type) query structure might be a better query organization than the linear type query structure 2) comparing with the complex tree query structure, the simple tree query structure that has at most three levels of query might provide better search performance 3) the fuzzy search methodology that employs a proper levels of cut-off value might provide more efficient search performance than the boolean search methodology. Even though indings could not be statistically verified because the experiments were done using a single replication, it is worth noting, however, that the research findings provided valuable information for developing a possibly robust search methodology in full-text information retrieval.

I. 서 론

우리는 현재 정보폭발 시대에 살고 있다. 매일 매일 수많은 종류의 새로운 정보가 생성이 되고 있으며 또한 이에 못지않게 기존 정보의 수정 정보가 쏟아지고 있다. 이러한 정보를 전달해 주는 매체들은 무척 다양하지만 대부분의 경우 자연 언어로 쓰여진 문서들(text)이다. 따

* 동국대학교 경상대학 정보관리학과

라서 정보 폭발 시대에 있어서 효과적인 정보 관리는 결국 자연 언어로 쓰여진 문서 정보 처리(저장 및 검색)에 그 궁극적인 성공의 여부가 달려 있다고 해도 과언이 아니게 되었다.

자연 언어란 본래 자연 발생적으로 생성되어진 지역적 집단 사회에서 의사 소통용으로 쓰여지던 생활 언어를 말한다. 이러한 자연 언어는 인간 생활의 발달 과정과 더불어 점차 복잡, 고도화되어서 이제는 생활 용어뿐만아니라 각 전문 분야의 지식 전달 및 교환 과정에까지 사용되게 되었다. 자연 언어의 예를 들어 보면 한국어, 영어 및 독일어등 현재 인류의 각 종족 및 사회에서 쓰여지는 언어들이 있다.

자연 언어는 그 내용의 표현 및 이해의 관점이 아주 다양하다.

특수 목적으로 개발이 되어진 인위적인 언어(예 : 컴퓨터 언어)에 비한다면 자연 언어는 의미의 표현이나 수용이 불확실하다. 예를들면 인위적인 언어에서는 동일한 개념을 표현해주는 방법이 대부분의 경우 매우 제한적인 반면 자연 언어에서는 그 표시 방법의 제약이 없다. 따라서 자연 언어에서는 표현의 다양함에 비례한 내용의 불확실성이 아주 크다.

전기에서 언급한 바와 같이 문서정보는 자연 언어로 쓰여있다. 또한 자연 언어는 그 내용이나 표현에 있어서 본질적인 불확실성을 갖고 있다. 따라서 효과적인 문서 정보의 처리는 결국 이 자연 언어의 본질적인 불확실성을 어떻게 극복하느냐에 그 성공여부가 달려 있다.

일반적으로 문서정보처리란 자연 언어로 쓰여진 모든 문서 형태의 정보를 일정한 주제별로 분류 및 정리하여 일정 장소에 저장시킨후 특정 주제에 따라 관련 정보들을 신속하게 찾아보는 일련의 연관된 작업을 의미한다.

현재 도서관학이나 전산학 또는 기타 관련 분야에서 제반 문서 정보 처리 기법들이 개발되어 일반 관리 활동이나 산업 분야에서 널리 활용되고 있다.

Ⅱ. 연구 범위 및 목적

본 연구에서는 연속적인 문서 정보 처리 과정(색인, 저장 및 검색)중에서 검색 과정을 주된 연구대상으로 한다. 물론 문서정보의 색인 과정이나 저장 방법등이 궁극적으로 검색 방법으로 연결되어 검색 결과에 영향을 미치게 되지만 본 연구에서는 일단 문서 정보의 검색 과정을 주된 연구 분야로하고 이에 관련된 문서의 색인이나 저장 절차를 부수적인 관련 연구 분야로 하였다.

본 연구의 1)일반적인 목적은 자연 언어로 쓰여진 문서 정보의 내재적인 불확실성을 극복하여 사용자들에게 그들의 검색에 대한 전문성을 불문하고 주어진 특정 주제에 맞는 문서 정보들을 전체 문서 정보군으로 부터 효과적이고 효율적으로 인출하여 줄 수 있는 문서 정보 검색 방법을 고찰해 보는 것이다.

이를 위하여 본연구에서는 2)기술적 목적으로서 퍼지논리(fuzzy logic)를 이용한 트리(tree) 유형의 비교적 사용자들이 사용하기가 간단한 검색(query)구조를 채택한 문서 정보 검색 시스템을 개발 하여 상이한 검색 전문성을 갖는 여러부류의 사용자들이 그들의 다양한 검색 목적에 맞추어 용이하게 사용할수 있도록 한다.

Ⅲ. 문서정보 검색방법에 대한 배경연구

1. 색인화일검색(Index file search) 對 전문검색(Full-text search)

일반적으로 문서정보 검색은 색인화일검색방법을 주된 근간으로 하고 있다. 이 방법에서는 발생되는(또는 수정되는) 모든 문서정보들은 인간 문서분류자(indexer)에 의해 관련 키워드(keyword)별로 우선 색인이 된 후 실제적인 정보검색은 색인에 사용된 키워드들을 모아 놓은 색인화일을 대상으로 이루어진다 〔Cogner, 1977; Faloutsos, 1985; Hawkins, 1982; Heaps, 1978; Kent, 1963; Moulton, 1979; Takle, 1980〕. 따라서 이 방법을 사용하면 문서정보 검색의 신속성을 도모할수있게 된다. 하지만 이 방법에서는, 모든 문서 정보들이 강제적으로 사전 색인이 되어 검색용 색인화일(index file)이 생성 및 유지되어야 하므로 값비싼 인간 문서분류자를 계속적으로 고용해야 한다거나 또는 색인화일 유지를 위한 경비 지출이 큰 문제점이 되고있다.

또한 인간 문서분류자들의 상이한 자질및 경험등에 기인한 색인의 변화는 경우에 따라서는 전반적인 검색 시스템의 운용에 큰 영향을 미치게 된다. 더구나 색인화일의 사전적인 강제적 구축은 종종 사후의 실제적 검색과정과 연결이 단절되어 검색의 탄력성을 저해할수 있다. 이와같은 제반 문제점을 보완하기 위해, 검색 대상 문서정보들의 사전적인 색인을 요하지 않는 전문(full-text)검색 방법이 개발 되었다 〔Stafill and Kahle, 1986; Tong and Shapiro, 1985a; Tong and Shapiro, 1985b; Blair and Maron, 1985; Salton, 1983〕. Full-text 검색 방법이란 실제적인 검색 과정의 대상이 문서정보의 원문들로써 검색문서 정보들의 사전적인 색인과정이 전혀 필요없는 검색방법이다.

이 방법에서는 일단 발생되는 모든 문서 정보들은 사전적인 색인를 거치지 않고 직접데이터베이스로 입력이 되며, 실제적인 검색은 데이터베이스에 수록된 문서정보의 원문들을 대상으로 이루어진다. 이 방법을 사용하면 사용자들은 그들의 검색경험이나 전문성에 비교적 구애됨이 없이 그들의 고유한 검색용 단어들을 사용하여 원하는 문서정보를 마음대로 찾아볼 수 있게된다. 이 방법에서는 문서정보들이 사전적으로 색인 되어있는 요약된 색인화일을 사용하지 않으므로 사용자들은 색인에 사용된 색인 용어들이나 사전적으로 구축된 특정 검색 시스템에 대한 사용절차를 미리 숙지할 필요가 없이 그들의 임의의 키워드를 사용하여 검색을 할 수 있게 된다.

그러나 위와같은 장점이 있는 반면, 이 방법을 쓰게되면 검색을 위해서 데이터베이스내에 있는 모든 문서정보의 원문들을 매번 일일이 읽어야 하므로 검색시간의 지연이라는 단점이 있다. 이를 극복하기 위해서 현재 여러 가지 방법들이 활발하게 연구되고 있다.

그 중에서 가장 많이 연구되는 방법으로써 문서정보의 원문의 축약이 있다 〔Salton, 1968 and 1983〕. 이 방법은 문서정보의 원문중에서 직접적으로 검색의 결과에 영향을 주지않는 不用語나 조사 또는 동사의 변형들을 제외하여 문서정보의 원문의 양을 줄여 전체적인 검색시간의 단축을 도모하는 것이다.

본 연구에서는 사용자 위주의 보다 탄력적인 검색 방법 고찰을 위해 검색 시간의 지연이 있더라도 full-text 검색 방법을 중점적으로 고찰한다.

2. Boolean 검색 對 퍼지논리 (Fuzzy logic) 검색

현재까지의 일반적인 문서정보 검색방법은 개별 검색 대상 문서들의 검색 주제에 대한 관련성 여부를 오직 可 또는 否로만 판정하는 Boolean검색 방법이었다〔Kent, 1963; Moulton, 1979; Tackle, 1980〕. 이 방법에서는 개별문서정보들의 특정주제에 대한 차별적 관련성이 파악이 되지 못하고 오직 특정 검색 주제에 대해 관련이 된다거나 또는 관련이 없다는식의 양극적인 관련여부만 파악이 된다. 그러므로 이러한 검색방법은 사용자들에게 때로는 너무 제한적인 검색 결과를 보여주거나 또는 아주 방만한 결과를 보여 주게 된다.

이러한 문제점을 보완하기 위하여 검색대상 문서들의 일정 주제에 대한 절대적인 관련성 평가가 아닌 상대적 관련성 파악이 가능한 퍼지검색 방법이 개발되었다〔Yager, 1978; Tong and Shapiro, 1985b; Miyamoto and Nakayama, 1986; Kantor, 1981; Buell, 1982; Zenner, 1985〕.

이 방법은 모든 검색 대상 문서들을 그 내용에 따라 일정 검색 주제에 대해 상대적으로 그 관련성을 구분하고 또한 사용자들로 하여금 임의적인 인출수위(cut-off) 값을 사용하게 하여 탄력적인 검색결과를 얻을수 있도록 하게 한다. 즉 문서 정보들을 "아주", "꽤", "상당히", "거의", "전혀", "약간" 등과 같은 퍼지수식개념을 사용하여 그들의 주제에 대한 개별적 관련성을 구분평가하여 궁극적으로 사용자들로 하여금 선별적인 문서정보 인출을 할 수 있게 하는것을 의미한다.

따라서 이 퍼지 검색방법에서는 boolean 검색방법과는 달리 사용자들이 검색 주제어를 선택할시(검색을 구성할 시) 그들의 주제에 대한 상대적인 중요도를 퍼지개념을 이용하여 미리 구별적으로 표시해 주어야 한다〔Baldwin, 1979a and 1979b; Zadeh, 1975〕.

Boolean 검색과 퍼지검색의 검색주제어 구성의 상이함을 예를들어보면 다음과 같다〈그림 1〉.

상기와 같은 상이한 주제어 선정 방법을 택함으로써, Boolean 검색에서는 검색의 결과로써 문서정보들의 특정 검색 주제에 대한 부정적인 관련 여부만 파악할 수 있는 검색 인출 여부 리스트를 얻을수 있는 반면 퍼지 검색에서는 모든 문서 정보들의 검색 주제에 대한 상대적 관련도를 비교하여 볼 수 있는 검색 관련도 리스트를 얻을수 있게 된다〈그림 2〉.

Boolean 검색	검색 주제	"주식 시장"
	검색 주제어 구성	「 주식 + 사채 + 이자 + 배당금 + 투자 」
Fuzzy 검색	검색 주제	"주식 시장"
	검색 주제어 구성	「 주식(1.0) + 사채(0.7) + 이자(0.8)+배당금(0.9) + 투자(0.5)」
· Fuzzy 검색 주제어에 첨가된 숫자들은 해당 주제어의 검색 주제에 대한 상대적인 중요도를 표시해 주고 있다.		

그림 1. Boolean 및 퍼지 검색의 주제어 구성의 예

Boolean 검색		Fuzzy 검색	
문서	관련도	문서	관련도
1	1.0	1	0.8
2	0.0	2	0.3
3	1.0	3	0.7

그림 2. Boolean 및 퍼지 검색의 검색 결과 list의 예

3. 단순트리검색(Simple tree search) 對 복합트리검색(Complex tree search)

일련된 검색과정에 있어 가장 중요한 작업중의 하나는 적절한 주제어의 초기 선정및 그들의 논리적 연결이다. 이를 위하여 일반적인 검색 시스템에서는 주제어(index word)들의 선형적인 연결(linear connection) 구조나 또는 계층적인 주제어 연결(hierachical connection) 구조가 많이 이용되고 있다.

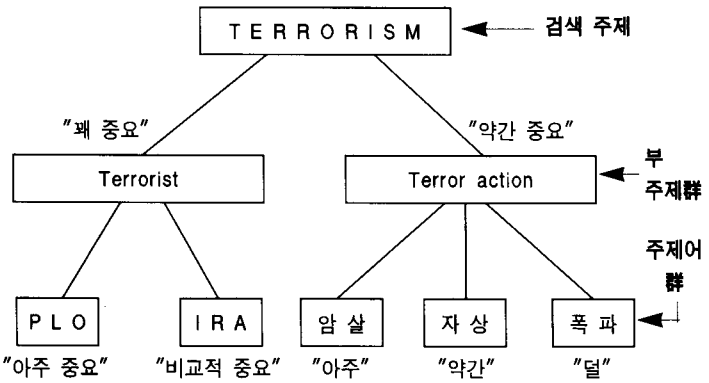


그림 3. Query tree를 이용한 계층적 주제어 생성 및 연결의 예

특히 계층적 트리형태(tree type)의 연결구조는 여러 검색 주제어들을 논리적인 의사결정수목(decision tree)형태의 조직을 통하여 서로 수평적으로 또는 수직적으로 연결시켜 궁극적으로 하나의 통일된 검색 주제를 구성케 함으로써 검색 사용자의 입장에서 보면 이 트리 구조를 이용함으로써 특정검색주제 (root node)로부터 파생적 연상 작용을 통하여 비교적 용이하게 논리적인 주제어생성 (bottom nodes)을 해낼 수 있다〈그림 3〉.

본 연구에서는 트리형태의 계층적인 주제어 연결 구조 방법(query tree organization)을 주된 연구대상으로 한다. 계층적 검색트리는 사용자에 따라 그 구성의 복잡성이 상이하며 또한 이 트리의 복잡성은 궁극적으로 검색 결과에 상당한 영향을 미치게 된다.

일반적으로 복잡한 트리는 대상 주제어 수가 비교적 많으며 그들의 주제에 대한 논리적 연결이 수직적으로나 수평적으로 비교적 길게 된다. 따라서 복잡한 검색 트리에서는 사용자들이 다양한 주제어들을 선정함으로써 그들의 미세한 검색요구를 검색결과에 정밀하게 반영할 수있게 되는 장점이 있는 반면 트리의 복잡성에 기인한 논리적 결론 도출상의 오류의 축적은 크게된다.

특히 퍼지검색 방법에서 주제어들의 검색 주제에 대한 비교적 주관적이고 애매모호한 중요도 구별은 검색 트리가 복잡해 질수록 최종 검색결론에 대한 논리적 오류의 축적이 크게 된다.

검색 트리의 폭과 깊이에 따른 검색 구조의 복잡성과 결론 도출 과정에서 축적되는 논리적 오류(error)의 량(量)과의 일반적인 상관 관계는 Suen and Wang〔1985〕과 Tong and Shapiro〔1985〕의 연구들에서 실증적으로 규명되었다. 그들의 연구 결과에 따르면, 불확신한 상태를 표시하

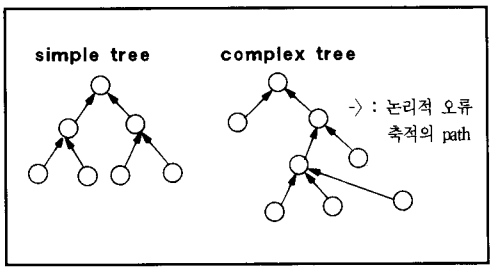


그림 4. 검색 트리에서의 논리적 오류의 축적

는 많은 의사 결정 변수(node)들로 구성된 의사결정수목에 있어서, 트리가 종적으로나 횡적으로 점점 커질수록 최종 결론 도출에 있어 각 변수들에서 생성된 논리적 오류의 축적은 점점 커지게 된다〈그림 4〉.

본 연구에서는 검색 트리상의 논리적 오류의 축적을 가급적 피하기 위해 복잡한 검색트리를 쓰지 않는 비교적 간단한 구조의 검색트리를 사용한 초보적 사용자 위주의 검색 방법을 고찰하고자 한다.

이를 위하여 본 연구에서는 특정 검색 주제를 포함하여 4 단계이상의 트리구조를 갖는 검색을 복합트리검색라고 간주하여 그 미만(3단계 이하)의 계층을 갖는 단순트리검색와 구별하였

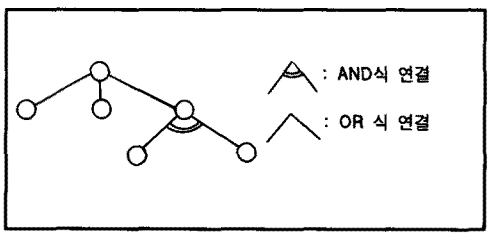


그림 5. 3 단계로 구성된 단순 트리 검색

다. 검색 트리의 복잡성의 판정의 기준으로 3 단계을 사용한 이유로써는 일반적인 트리식의 검색구축에 있어 3 단계는 검색대상 주제어들을 상이한 논리적 연결자(예;AND, OR NOT등)로 동시에 연결시킬 수 있는 최소한의 단계이기 때문이다〈그림 5〉.

4. 앞으로의 연구방향

본 연구의 주된 연구목표는 사용자들에게 그들의 전문적 검색능력에 비교적 무관하게 두루두루 고른(less varied) 검색결과를 보여주는 문서정보 검색 방법을 고찰해 보는 것이다.

이를 위하여 본 연구에서는 3장에서 살펴본 여러가지 검색관련 기본 지식을 토대로 다음과 같은 실험적 연구 주제를 정립하였다.

1) 트리를 이용한 계층적 검색은 단선적 검색에 비해 사용자들로 하여금 그들의 검색에 대한 전문성을 불문하고 보다 용이하게 논리적인 주제어 생성및 연결을 가능하게 해 줄 수 있는가 ?

2) 적절한(비교적 높은) 인출수위값을 사용하는 단순 트리 퍼지 검색 방법은 Boolean 검색 방법에 비해 보다 효율적인 검색 결과를 제공해 줄 수 있는가 ?

3) 단순트리를 이용한 퍼지검색 방법은 사용자들의 검색에 대한 전문성에 관계없이 그들에게 두루 고른(less varied) 검색 결과를 제공해 줄 수 있는가 ?

상기와 같은 연구주제들은 본 연구를 위해 실제로 개발된 단순트리검색(simple tree query)를 사용하는 전문(full-text) 문서정보 검색 시스템을 실험 대상으로 그 신빙성이 검증되었다.

상기 연구주제들에서 거론된 검색의 효과성, 및 효율성 그리고 사용자간의 검색결과의 상이성(variation) 등은 본 연구에서 각기 recall, precision, 그리고 HRM(Heuristic Robustness Measure)라는 평가비율로 측정되었다.

5. 단순 트리 구조를 이용한 퍼지 전문 검색 시스템

4장에서 언급된 연구 주제들을 검증하기 위하여 본 연구에서는 퍼지 논리을 이용한 실제적인 전문 검색 시스템이 개발 되었다.

이 시스템에 사용되어진 주제어 연결 구조로서는 단순트리 검색이 사용 되었다〈그림 6〉.

검색 시스템은 PASCAL 로 프로그램 되었으며 IBM PC AT 급의 하드웨어에서 구현 되었다.

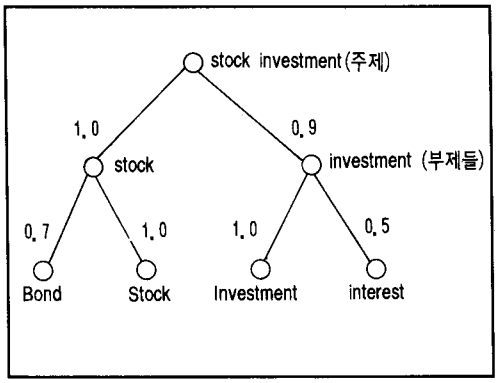


그림 6. Simple tree query의 예

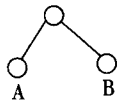
검색 트리상의 각 노드(부제 및 주제어들)들의 퍼지 값들을 논리적으로 처리하기 위하여 사용되어진 퍼지논리은 H. Reichenbach와 Zadeh 교수등이 개발한 20개의 복합 퍼지 추론 이론들이다〈그림 7〉.

상기 도표에서 보면, 퍼지 검색 트리상의 각 노드(부제 및 주제어들)에서 생성되는 퍼지 값(주제 관련도)들은 2 가지 서로 다른 계산 방법들(수평적 및 수직적 방법)의 연결적조합을 통하여 최종 결론에 연결되어 진다.

먼저 수평적 퍼지값 계산 방법을 살펴 보면, 이 계산 방법은 검색 트리의 각 노드들을 동일한 parent node단위로 구분하여 그들의 퍼지 값들을 횡적으로 서로 연결시켜가며 개개의 단위 parent node의 퍼지 값들을 계산하는 방법으로서 connective_disconnective 방법이라고 부른다.

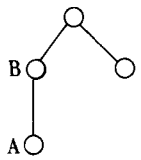
복합 Fuzzy이론

A. 수평적 추론 방법(Connective_Disconnective추론)



	V(A and B)	V(A or B)
방법1	T〔V(A), V(B)〕	S〔V(A), V(B)〕
방법2	Max〔0, V(A)+V(B)-1〕	Min〔1, V(A)+V(B)〕
방법3	V(A)*V(B)	V(A)+V(B)-V(A)*V(B)
방법4	Min〔V(A), V(B)〕	Max〔V(A), V(B)〕

B. 수직적 추론 방법(Detachment-Implication추론)



	Detachment	Implication(==〉)
방법1	V(B)=Min〔V(A), V(A==〉B)〕	Min〔V(A), V(B)〕
방법2	V(B)=Min〔1, V(A)+V(B)〕 if V(A)+V(A==〉B) 〉 1 = 0 otherwise	Max〔1-V(A), V(B)〕
방법3	V(B)=V(A)*V(A==〉B)	Min〔1, V(B)/V(A0〕
방법4	V(B)=Max〔0, V(A)+V(A==〉B)-1〕	Min〔1, 1-V(A)+V(B)〕
방법5	V(B)=Max〔0, (V(A)+V(A==〉B)-1/V(B)	1-V(A)+V(A)*V(B)

그림 7. Reichenbach와 Zadeh 교수등이 개발한 퍼지 이론의 예

다음으로 수직적 퍼지 값 계산 방법은 개개 노드의 퍼지 값들을 하부 노드에서 직접 입력되었거나 또는 중간적으로 계산되어 올라오는 (또는 입력되는) 종속 퍼지 값들을 이용하여 수직적으로 수정 계산하는 방법으로서 detachment_implication 계산방법이라고 부른다.

이 두가지 서로 다른 퍼지 계산 방법은 서로 한 조(pair)로 결합되어 전체적인 검색 트리의 퍼지 값을 계산할 수 있게 된다.

본 연구에서는 4 개의 수평적 계산 방법 및 5 개의 수직적 계산 방법을 이용하여 이들의 조합으로 가능한 총 20 개의 퍼지 계산 방법(pair)을 대상으로 연구를 진행하였다[Tong & Shapiro, 1985].

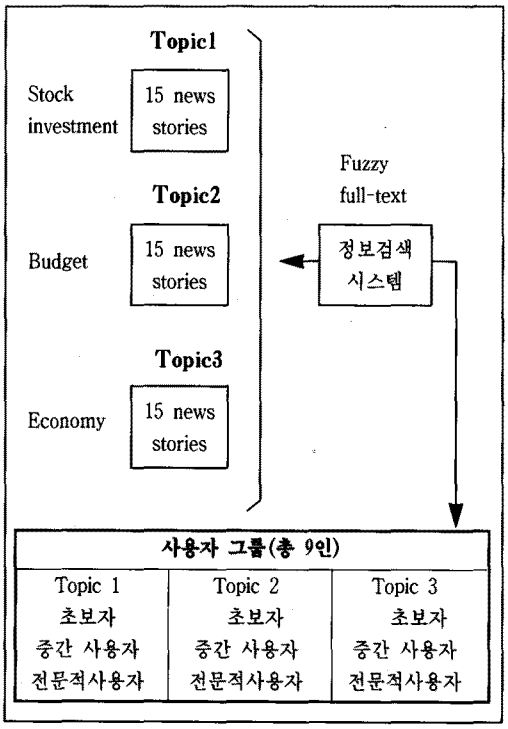


그림 8. Fuzzy full-text 검색 실험의 개요

6. 단순 트리를 이용한 퍼지 검색 평가 실험

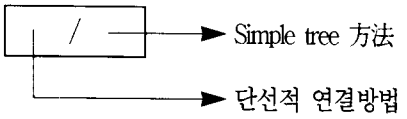
본 연구에서 정립된 제반 연구 주제들의 신빙성을 실증적으로 검증하기 위해 다음과 같은 개요로 프로토타입적인 문서정보 검색실험이 진행 되었다. 먼저 실험적 검색주제로써 "주식투자", "예산제도", 및 "일반경제"의 서로 관련성이 있는 주제들이 선정되어, 본 연구대상 검색 방법의 검색성과가 보다 차별적이고 효과적으로 판별되게 하였다. 다음으로 실험대상 사용자들로써는 각 검색주제 별로 초보자, 중급 사용자 그리고 전문가적 사용자로 구성된 3인의 사용자 그룹이 선정 되었다. 또한 주제어 연결 방법으로써는 단순트리를 이용한 계층적 검색 구조가 사용 되었다 〈그림 5〉. 실험적 검색대상 문서정보 데이터베이스는, AP 통신에서 CompuServe를 통하여 수신한 전문으로 된 뉴스들 중에서 수작업을 통하여 15개씩 선정된 "투자관계", "예산제도" 및 "일반 경제"관련 1000 단어 내의 news story들로 구성 되었다. 전체적인 실험의 개요를 도표로 나타내면 다음과 같다 〈그림 8〉.

7. 실험결과 요약

1) 트리를 이용한 계층적 주제어 연결방법은 사용자들로 하여금 비교적 용이하게 논리적인 주제어 생성 및 연결을 가능하게 해줄 수 있다.

일련의 연결된 전체적인 정보 검색과정에 있어 가장 중요하고 어려운 과정은 검색초기의 적절한 주제어 선택과 그들의 논리적 연결과정이다. 일반적으로 선형적인 검색 구조를 사용

		Topic		
		주식투자	일반경제	예산제도
사용자 그룹	초보자	中 / 上	下 / 上	下 / 上
	비초보자	上 / 上	上 / 上	上 / 上



* 上 : good 中 : average 下 : poor

그림 9. 초보자 그룹과 비초보자 그룹의 주제 생성의 논리성 분석

하는 대부분의 사용자의 경우 그들의 검색 주제어 선정 및 연결은 아무런 논리적 지침이 없는 무작위적인 주제어 선택과 그들의 단선적(linear) 연결이 대부분이다.

반면 일정검색 주제에 대한 검색를 이용한 계층적 주제어 생성 및 연결방법은 사용자들로 하여금 거의 반 강제적으로 검색생성에 대한 검색 생성에 대한 논리적 사고력을 동원케하여 궁극적으로는 초보적 사용자라 하더라도 비교적 논리적이고 효과적인 주제어 생성 및 연결을 가능하게 할 수 있다.

단선적 연결방법과 계층적 트리를 이용한 주제어 연결방법의 논리적 성과의 차이를 파악하기 위하여 본 연구에서는 다음과 같은 실험을 행하였다.

먼저 사용자들을 각 주제별로 초보자 그룹과(3인) 과 비초보자 그룹(6인) 으로 구분하여 초보자 그룹과 비초보자들간의 주제어 생성과정의 논리성을 비교분석하여 보았다〈그림 9〉.

상기 도표에서 사용자들의 검색 구성의 논리성은 1) 주어진 검색 주제에 대한 일관성 있고 체계적인 부제(sub_topic) 및 주제어 생성 정도 및 2) 검색 주제어들의 합리적 연결성 여부에 따라 평가되었다.

사용자들의 검색의 논리성은 미국의 대학 도서관에서 오랜 기간의 검색 보조 경험과 전문적 검색 능력을 갖춘 숙련된 검색 전문가 2人이 공동으로 평가하였다.

위 도표에서 보면 단선적인 주제어 연결방법에 비해 트리를 이용한 계층적인 주제어 생성 및 연결방법이 사용자들의 검색에 대한 전문성을 불문하고 보다 만족할 만한 논리적인 주제어 생성 맡 연결성을 보여주고 있다.

2) 적절한 인출 수위를 사용하는 단순 트리퍼지 검색방법은 Boolean 검색에 비해 비교적 효율적인 검색결과를 보여준다.

일반적으로 모든 검색결과는 검색의 인출성(recall) 과 검색의 정밀성(precision) 이라는 2가지 서로 다른 정보 인출 비율로 측정되어진다 〔Blair and Maron, 1985; R. Tong and shapiro, 1985〕.

Recall이란 검색의 효과성을 표시해 주는 비율이고 precision이란 검색의 효율성을 나타내어 준다. 검색의 효과성(Recall) 은 일정 검색의 결과 얼마나 원하는 문서정보를 빠뜨리지 않고

		Search methods		increase or
		fuzzy	boolean	decresae
평균 인출 문서의 수	corect	12.60	14.0	+1.40
	total	20.60	33.4	+12.8
검색 결과 비율	recall	0.84	0.93	+0.09
	precision	0.61	0.42	0.19

그림 10. Fuzzy 검색방법과 boolean 검색방법을 이용한 평균적 검색결과 요약

잘 인출 했는지의 여부를 판단 해주는 비율로써 그 비율은 총 인출 정보중에서 정확하게 인출되어진 정보의 수를 꼭 인출 되어져야 할 정보의 총수로 나눈 것이다. 반면 검색의 정밀성(precision)은 일정 검색의 결과 얼마나 정확히 꼭 필요한 정보만을 인출 했는지 여부를 나타내주는 비율로써 정확히 인출되어진 정보의 수를 검색의 결과로써 인출된 문서 정보의 총수로 나눈 것이다.

예를들면 어떤 검색의 경우 총 10개의 인출되어야 할 정보가 있는 경우, 만약 15개의 정보가 실제로 인출되어 이중 7개의 정보만이 정확히 인출되어진 정보로 판명되었다면 검색의 효과성은 7/10이 되며 검색의 정밀성은 7/15이 된다.

본 연구에서는 이 두가지 비율을 고려하여 단순 검색 트리 를 이용한 검색 방법이 boolean 검색방법에 대해 얼마나 효율적인 검색결과를 제공하는지의 여부에 대한 실험을 행하였다〈그림 10〉.

일반적으로 퍼지 검색에 있어서 높은 인출 수위값는 낮은 인출 수위값에 반면 보다 상대적으로 정밀한 검색 결과를 보인다. 반면 정보의 전반적 인출성에서는 열세한 결과를 나타낸다.

이에 본 연구에서는 퍼지 검색 방법의 장점인 검색의 정밀성에 촛점을 맞추어 "적절한" 인출 수위값은 검색의 비교적 높은 정밀성을 유지해 주는 값으로서 간주하였으며 이에 본 연구에서는 연구의 편의상 0.7 이상의 인출 수위값들을 적절한 인출 수위값로 가정하였다. 실제적으로 본 연구에서는 연구결과의 객관성을 높이기위해 0.7과 0.9의 2 가지 상이한 인출 수위값이 사용되었다. 위의 도표에서 퍼지 검색에 나타난 숫자들은 이 두가지 인출 수위값을 사용한 평균적인 검색 결과들이다. 또한 이 실험에 사용된 퍼지 논리은 총 20개의 복합 논리들로써 상기 도표에는 이들의 평균적 검색결과가 요약되어있다.

위 도표에서 보면 적절한 인출수위를 사용하는 퍼지 검색 방법은 boolean 검색방법에 비해 검색의 높은 효율성을 보이는 반면 약간 낮은 검색의 효과성을 보이고 있다.

Boolean 검색방법은 평균적으로 약 1.4개의 관련 문서정보를 더 얻기위해 무려 12.8개라는 총체적 문서 정보들을 추가로 인출하고 있다.

이러한 비효율적인 boolean 검색은 퍼지 검색에 비해 비록 궁극적으로는 약간의 검색의 효과성을 향상시켰다 하더라도 그 실질적 효용성은 없게 된다. 왜냐하면 이질적이고 복합적인 대량 정보의 폭발적 생성에 대처하는 현재의 문서 정보 검색에 있어 불필요한 정보의 인출을 효과적으로 억제해주는 검색의 효율성(정밀성)은 검색의 효과성(인출성)에 비해 극히 그 지니는 의미가 지대하기 때문이다.

3) 단순 트리 구조를 이용한 퍼지 검색방법은 사용자들에게 그들의 검색에 대한 전문성 여부에 관계없이 비교적 고른 (less varied)검색 결과를 제공한다.

좋은 검색 방법이란 사용자의 검색에 대한 전문성에 상관없이 사용자 모두에게 골고루 만족스러운 검색결과를 제공할 수 있어야 한다.

즉 상이한 검색능력을 보유한 사용자들간에 비교적 차이가 적은 검색 결과를 보장해 줄 수 있는 방법이 좋은 검색 방법이 된다.

본 연구에서는 단순 트리를 이용한 퍼지 검색 방법을 개발하여 이 방법이 사용자들에게 비교적 덜 가변적인 검색결과를 제공해 주는 방법으로 일단 간주 되었다. 이의 실증적 검증을

HRM = MAD + 1/3 MDAD
·MAD = (Σabs(Xi-X) / N)
N -> 변수의 수 Xi -> 변수 i의 값 X -> 변수의 평균 값
·MDAD = Median of | Xi- M |
M -> 변수들의 Median Xi -> 변수 i 의 값

그림 11. HRM 생성 공식

위하여 본 실험에서는 각 검색 주제별로 사용자 그룹을 초보자, 중급 사용자 및 전문적 사용자 그룹으로 나누어, 이 나누어진 그룹들의 검색결과를 상호비교하여 그들 검색결과의 상호간의 상이점을 실증적으로 비교하여 보았다.

본 연구에서, 3인의 상이한 전문성을 가지는 검색 사용자간의 검색결과의 차이는 "Hueristic Robustness Measure(HRM)"라는 휴리스틱척도로 측정 되었다. HRM은 통계학적 개념인 MAD(Mean absolutue deviation)와 MDAD (Median of absolute deviation)가 합성된 척도로써 본 연구에서 다음과 같은 공식으로 고안되었다(S. Kachigan, 1986)<그림 11>.

MAD는 본래 통계학적으로 측정변수들의 평균적 산포도를 나타내며 일반적으로 변수들중의 양 극단적(minimum or maxmum)인 값에 상당히 많은 영향을 받는다. 반면 MDAD는 양극의 극단적인 값에 비교적 영향을 받지않는 변수들의 중앙값을 표시해 주는 반면 변수들의 전체적인 평균적인 분포도를 나타내지 못한다. HRM은 이 두가지 서로 다른 변수들의 산포도 측정에 관한 척도를 서로 상호 보완적으로 고려해 준다.

HRM은 3人의 서로다른 검색결과의 차이점(variation)을 측정해 주는 척도로서, 만약 각각의 검색 결과들을 두자리의 소숫점 숫자들(ex, 0.12, 0.34,. ...)로 표시한다면 전체적으로 101*101*101개의 상이한 검색 결과의 조합이 가능하고 각각의 상이한 3개의 검색 조합은 각기 고유한 한 개씩의 HRM 값을 갖게된다.

이 상이한 HRM 값들은 전체적인 하나의 분배 곡선을 이루게 되며 본 실험에서는 이것을 HRM의 frequency 분배곡선 이라고 명명하여 임의적으로 계산되는 특정 HRM의 값을 평가하는 기준 곡선으로 한다<그림 12>.

상기 곡선에서 보면, 왼쪽에 있는 HRM 값일수

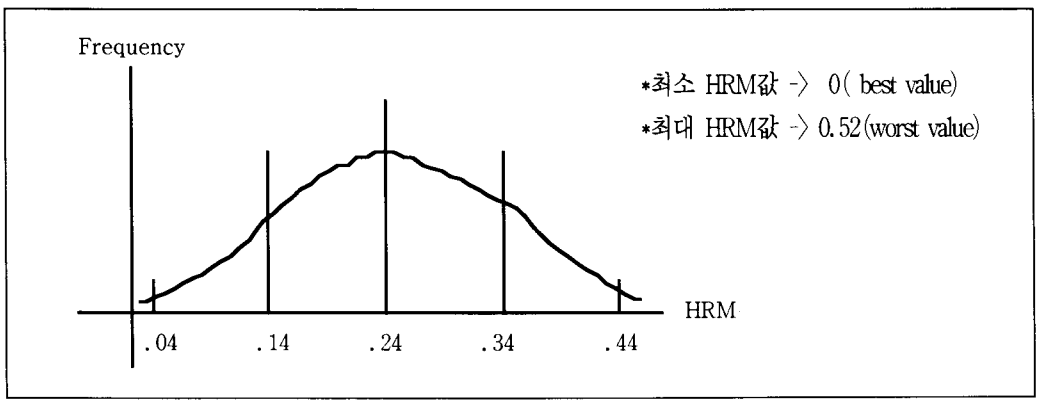


그림 12. HRM frequency distribution curve

Recall

		Topic			Average
		Stock	Budget	Economy	
검 색 사용자	초보자	0.65	0.82	0.63	0.70
	중 급	0.96	0.77	0.73	0.82
	전문가	0.97	0.77	0.74	0.83
HRM		0.16	0.12	0.20	0.16

Precision

		Topic			Average
		Stock	Budget	Economy	
검 색 사용자	초보자	0.50	0.51	0.80	0.60
	중 급	0.70	0.79	0.43	0.64
	전문가	0.70	0.52	0.49	0.57
HRM		0.10	0.15	0.20	0.15

그림 13. Recall과 Precision의 평균적 HRM 요약

록 사용자들의 검색 결과간의 차이성(variation)이 점점 적어진다는 것을 나타내고 있다.

단순 트리를 채택한 퍼지 검색 방법을 사용하여 사용자들이 얻은 상이한 검색 결과의 HRM 값들을 요약해 보면 다음과 같다〈그림 13〉.

위의 도표를 보면 단순 트리를 이용한 퍼지 검색의 결과, 검색의 효과성의 경우 사용자의 검색 전문성이 높을 수록 검색의 결과가 좋았지만 전체 사용자들의 검색결과들의 절대적 차이는 별로 크지 않다.

즉 검색의 효과성의 평균 HRM 값은 0.16으로써 전체적인 HRM 분배곡선의 top 10 %안에 들어갈 수 있는 좋은 고른 분포를 보여주고 있다. 반면 검색의 정밀성의 경우 특이하게도 사용자의 검색 전문성과는 무관하게 전문적 사용자의 검색 결과가 가장 저조했다. 그 이유로써는 전문적 사용가일수록 사용 주제어가 다양해지고 또한 그들의 논리적 연결과정이 복잡해지며 따라서 축적되는 논리적 오차들이 검색 결과에 반영되어 궁극적으로 필요없는 문서 정보들의 대량인출이 있었기 때문인 것으로 보인다.

특히 본 실험에서는 실험 대상 검색 주제들이 서로 내용적으로 중복이 많이되어, 전문적 사용자에의해 선정되어진 특정 주제 인출용 주제어들이 검색되어야할 주제뿐만 아니라 타 주제에 관계된 문장들까지 중복적으로 인출하였기 때문이다. 따라서 전문적 사용자의 검색의 정밀성에 비해 비교적 단순한 주제어를 사용하는 비 전문적 사용자의 검색의 정밀성이 우월하게 되었다. 하지만 세 사용자 그룹간의 절대적 검색 결과는 서로 큰 상이성이 없는 것으로 나타났다.

즉 검색의 정밀성의 평균 HRM 값은 0.15로써

검색의 효과성의 평균 HRM 값과 함께 상위 10 % 안에 들어가는 좋은 산포도를 나타내주고 있다. 즉 결론적으로 본 연구의 실험을 통하여 보면 단순트리를 이용한 퍼지 검색 방법은 검색종류(recall 및 precision) 사용자들의 검색에 대한 전문성 여부를 불문하고 비교적 상이성이 적은 검색 결과를 사용자에게 제공해 주고 있다.

8. 프로토타입적 실험에 대한 검증(verification)

본 연구에서 실행된 단순 트리를 이용한 퍼지 검색 결과에 대한 프로토타입적 실험은 그 검색 대상 news story들의 수(45개)의 제한성으로 인해 실험 결과의 신빙성이 취약하였다. 이의 보강을 위해 본 연구에서는 AP 통신으로 부터 무작위적으로 2000 여개의 news story들을 추가로 수집하여 프로토타입 적 실험에서 사용되었던 동일 실험사항으로 일련의 검증적(verified) 실험을 다시하였다.

그 결과에 따르면 검증적 실험은 단순 트리를 사용한 퍼지 검색 방법의 성과 분석에 대한 프로토타입 실험에서 발견되었던 결과와 거의 흡사한 내용을 보여 주었다 [W. Lee, 1989]. 검증적 실험에 대한 자세한 결과 분석은 그 대부분이 프로토타입 실험의 분석과 중복이 되므로 본 논문에서는 생략키로 한다.

9. 결 론

본 연구의 주된 목적은 1)특별한 검색 전문가의 도움없이도 누구든지 용이하게 사용할 수 있고 동시에 2)전문적 검색능력에 관계없이 사용자 모두에게 대체적으로 만족할만한 검색 결과를 두루 제공해 줄 수 있는 효과적이고 효율적인 문서 검색 방법을 고찰해 보는 것이다. 이를 위하여 본 연구에서는 제반문서 검색 방법들을 고려하여 선험적인 전략적 연구 주제들을 설정하였다. 이 선험적 연구 주제들은 본 연구를 위하여 실제적으로 개발된 트리 형태의 단순한 검색 구조를 사용하는 퍼지 전문 검색 시스템과 AP 통신에서 제공된 45개의 news story들을 대상으로한 프로토타입적인 검색 실험을 통하여 그 신빙성이 다음과 같이 실증되었다.

첫째, 트리를 이용한 계층적 주제어 연결방법은 단선적 연결 방법에 비해 사용자들로 하여금 보다 용이하게 논리적인 주제어 생성 및 연결을 가능케 해 준다.

둘째, 비교적 높은 인출 수위를 사용하는 단순 트리 퍼지 검색은 Boolean 검색에 비해 효율적인 검색결과를 보여준다.

세째, 단순 트리 구조를 이용한 퍼지 검색방법은 사용자들로 하여금 그들의 검색에 대한 전문성 여부에 관계없이 비교적 고른 (less varied) 검색 결과를 제공한다.

본 연구의 프로토타입 검색 실험은 실험 대상 문서 정보(news story)의 수적제한으로 말미암아 그 검증의 신빙성이 취약하다.

이의 보강을 위해 본 연구에서는 추가로 대규모의 문서 정보군을 검색 대상으로한 확인(verification) 실험이 행해졌으며 그 결과 프로토타입적인 실험 결과의 대부분은 긍정적으로 평가되어 그 검증의 보강을 가지게 되었다. 하지만 본 연구의 실증적 가설들이 프로토타입 및 확인 실험을 통하여 신빙성있게 확인되었다 하더라도 실험 자체에 내재된 다른 통계적 취약성(소수의 검색 대상 사용자 선정 및 검색 주제의 제한성등) 때문에 여전히 객관적인 통계

적 증명은 아직 되지 못하고 있다. 이를 위하여 본 연구에 이은 추후 보강적인 통계적 실험 절차가 요구되어 진다.

참 고 문 헌

Baldwin, J. "Fuzzy Logic And Its Application To Fuzzy Reasoning," Advances in *Fuzzy Set Theory And Applications*, Amsterdam, North-Holland, 1979a, pp. 93-116.

Baldwin, J. "A New Approach To Approximate Reasoning Using A Fuzzy Logic,"*Fuzzy Sets And Systems*, Volume 2, No. 4, 1979b, pp. 309-325.

Blair, D. and Maron, M. "An Evaluation of Retrieval Effectiveness For A Full Text Document Retrieval System,"*ACM Communication*, March, Volume28, No. 3, 1985, pp. 289-299.

Buell, D. "An Analysis Of Some Fuzzy Subset Applications To Information Retrieval Systems," *Fuzzy Sets and Systems*, Volume 7, No. 1, 1982, pp. 35-42.

Conger, D. "Restricting Searches In DIALOG," *Online 1*, October, 1977, pp. 68-77.

Faloutsos, C. "Access Methods for Text," *Computing Surveys*, Volume17, No. 1 March, 1985, pp. 49-74.

Hawkins, D. "ON-Line Bibliographic Search Strategy Development," *Online 6*, May, 1982, pp. 12-19.

Heaps, H. *Information Retrieval and Theoretical Aspects*, Academic Press, New York. 1978,.

Kachigan, S. *Statical Analysis:An Interdisciplinary Introduction To Univariate And Multivariate Methods*, Radius Press, New York, 1986, pp. 56.

Kantor, P. "The logic of Weighted Queries," *IEEE Transactions on Systems, Man, and Cybernetics*, volume SMC-11, No. 12, December, 1981, pp. 816-821.

Kent, A. *Text Book On Mechanized information* Retrieval, John Wiley and Sons, New York. 1963,.

Lee, W. "Consideration of A Query Methodology to Identify Natural Language Texts That Correspond to Specitied Topics," Ph. D thesis, University of Cincinnati, 1989, pp. 199-228.

Miyamoto, S. and Nakayama, K. "Fuzzy information Retrieval Based on A Fuzzy Pseudothesaurus,"*IEEE Transactions on Systems, Man , and Cybernetics*, Volume SMC-16, No. 2, March/April, 1986, pp. 278-282.

Moulton, J. "Dow Jones News/Retrieval, "*Database* 2, March, 1979, pp. 54-65.

Salton, G. *Automatic INformation Organization And Retrieval*, McGraw-Hill Book Company, N. Y. 1968,.

Salton, G. *Introduction To Modern Information Retrieval*, McGraw-Hill, New York 1983,.

Stafill, C. and Kahle, B. " Parallel Free-Text Search On The Connection Machine System," *ACM Communications*, Volume 29, Number 12, December, 1986, pp. 1229-1239.

Suen, C. and Wang, Q. "Analysis And Design Of Decision Tree Based On Entropy Reduction And Its Application To Large Character Set Recognition," *IEEE Transactions on Pattern Analysis and machine intelligence*, Volume PAMI-6, July, 1984, pp. 406-417.

Takle, Q. "STAIRS Search Strategy: Ideas and Opinions," *Online Review* 4, June, 1980, pp. 163-168.

Tong, R. and Shapiro, D. "Experimental Investigation Of Uncertainty In a Rule_Based System For Information Retrieval," *International Journal of Man-Machine Studies*, Volume 22, 1985a, pp. 265-282.

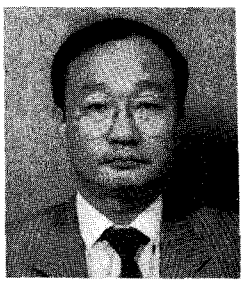
Tong, R. and Shapiro, D. "RUBRIC:An Environment For Full Text Information Retrieval, "Proceedings, 8TH International ACM Conference on R & D Information Retrieval, Montreal, June. 1985b,.

Yager, R. "A Logical On-Line Bibliographic Searching: An Application Of Fuzzy Sets, "*IEEE Transactions on System, Man , and Cybernetics*, Volume SMC-10, No. 1, January, 1980, pp. 51-53.

Zenner, R. "A New Approach To Information Retrieval Systems Using Fuzzy Expressions, "*Fuzzy Sets and System*, Volume 17, North-Holland, Amsterdam, 1985, pp. 9-22.

Zadeh, L. "The Concept Of A Lingustic Variable And Its Application To Approximate Reasoning, "*Information Science*, Volume 8, 1975, pp. 199-249.

◈ 저자소개 ◈



저자 이원부는 1977년 연세대 경영학과를 졸업후 한국외환은행 및 삼성물산에 근무하다 1982년 도미하여 미국 보스톤대학 및 신시내티 대학에서 경영정보학 석·박사 학위를 취득하였으며 주요 연구분야는 퍼지이론을 이용한 DB관리 및 전문가시스템 개발이다.