2001년 한국정보처리학회 추계 학술발표논문집 제8권 제2호

메타 데이타베이스를 이용한 퍼지 검색엔진의 설계 및 프로토타입 구현

유자영, 김남영, 박순철
전북대학교 정보통신공학과
e-mail:godpower@internet.chonbuk.ac.kr

Design and Prototype of Fuzzy Information Retrieval Engine with Meta Database

Ja-Young You, Nam-Young Kim, Soon C Park
Dept of Information and Communication, Chonbuk-Buk University

요약

현재 인터넷상에는 수많은 정보가 산재되어 있고, 사용자가 원하는 정보를 검색해주는 수많은 검색엔진들이 개발되어 사용되고 있다. 하지만 기존의 검색엔진은 사용자들이 입력한 질의어만을 가지고 단지 시소러스 사전만을 참조해서 검색결과를 나타내는 게 대부분이이서 사용자의 구미에 맞는 정보를 찾는데 어려운 점이 많았다. 이에 본 논문에서는 MetaDB안에 있는 보편적 Meta 데이터를 이용, 사용자의 간단한 정보 입력과 함께 퍼지연산을 적용시킨 매칭기법으로 사용자의 특성에 맞는 검색결과를 도출하는 퍼지 검색시스템을 제안한다.

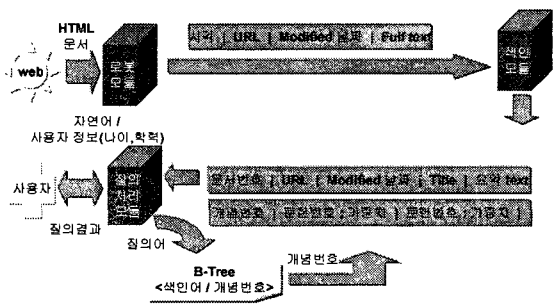
1. 서론

정보화 사회로 접어들며 정보량이 급증하면서, 정보의 생산에 대한 문제와 더불어 정보검색에 관한 문제가 사람들의 큰 관심사로 대두되고 있다. 이에 부응하여 수많은 검색엔진들이 저마다의 특징을 가지고 인터넷상의 정보를 보다 쉽게 검색해주고 있다. 그러나 기존의 검색엔진은 단지 사용자가 입력한 자연어 및 질의어를 파싱 및 형태소 분석하여 시소러스 사전을 참조하는 등 단지 그 질의어만을 가지고서 검색결과를 도출해내기 때문에 보다 사용자의 특성 즉, 나이 및 학력, 성별 등의 사용자 정보에 맞는 검색결과를 도출하기에는 무리가 있었다. 이에 본 논문에서는 질의어만이 아닌 사용자 정보와 그에 상응하는 소속척도가 입력된 MetaDB와, 문서들의 순위도 결정 시 사용하는 퍼지연산을 적용시켜 어느 한쪽만의 질의어에 치우치지 않게 함으로써, 보다 사용자의 특성에 맞는 보편적 검색결과를 도출할 수 있는 방법을 제안하고, 그에 따른 프로토타입을 구현하였다. [6,8,11,12]

2. 시스템 구조

2.1 전체 시스템 구조

본 시스템의 전체 구조도는 [그림1]과 같다.



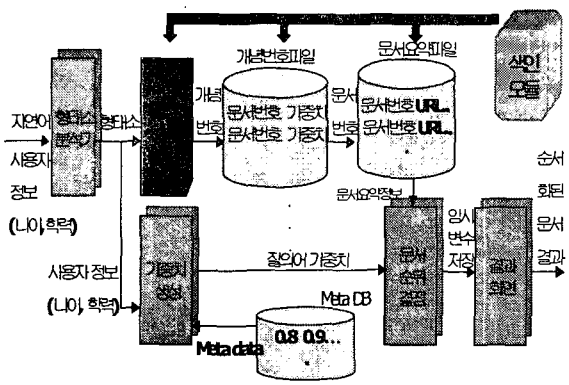
[그림 1] 전체 구조도

본 시스템은 기존의 대체적인 검색엔진과 같이 3가지 모듈로 이루어진다. 먼저 로봇모듈은 웹상에서 문서를 수집하여 , 태그를 제거하는 파싱역할까지 수행하고, 색인 모듈은 파싱된 문서들을 가지고서 Fast INV형식[1]을 적용시켜 색인트리와 각 색인어마다 연결된 개념번호를 가진 개념번호 파일 및 문

서요약정보 파일을 생성한다. 마지막으로 질의처리 모듈은 MetaDB와 퍼지연산을 적용시켜 사용자가 입력한 질의어와 사용자 정보를 이용해서 검색결과를 출력한다. [1,2,3]

2. 2 질의처리 시스템 구조

본 논문에서 제안한 질의 처리 시스템의 구조도는 [그림 2]와 같다.



[그림 2] 질의 처리 모듈 구조도

위에서 보는 바와 같이 입력된 자연어 형태의 질의 문장은 형태소 분석기를 거쳐 의미있는 형태소들로 분석되고 그 형태소를 이용하여 가중치 생성기에서 가중치를 생성한다. 그 가중치들과 키워드들을 이용하여 색인트리에서 개념번호 및 문서번호를 찾고, 이를 이용해서 검색된 문헌들로 퍼지 연산을 수행하여 순서화된 문서결과를 도출하게 된다.

```
keywordCount = inKeywordCount ;
if (userInformationCheck)
  repeat
      if(useRelevanceDic)
        read MetaDB개념번호 파일
        repeat
          keywordWeight = Max(AgeWeight,GradeWeight,,SexWeight)+
          (1-γ)(AgeWeight+GradeWeight+SexWeight);
          revanceKeywordCount = revanceKeywordCount-1 ;
        until revanceKeyword = 0 ;
      else
        keywordWeight = Max(0.7, 0.7, 0.7)+(1-γ)(0.7+0.7+0.7)/2;
        KeywordCount = KeywordCount-1
    NewKeywordCount = keywordCount + keywordCount*relevanceKeyword
  until KeywordCount = 0 ;
else
  keywordWeight = Max(0.7, 0.7, 0.7)+(1-γ)(0.7+0.7+0.7)/2
```

[그림 3] 질의어에 대한 가중치 설정 알고리즘

[그림 3]은 추출된 키워드들과 사용자 입력 정보를 가지고 질의어의 가중치와 MetaDB를 참조하는 연관어들의 가중치를 생성시키는 알고리즘을 나타낸 것이다. 위에서 표시한 바와 같이 질의어 가중치 합성부분에서 퍼지 연산을 적용시켜 어느 쪽에도 치우치지 않는 결과를 생성하도록 한다.

3. Meta Data 및 퍼지 연산기

3. 1 Meta Data

원래 Meta Data는 자연어 질의 처리 시에 참조하는 자체 데이터베이스로써 퍼지 비교 연산자가 퍼지 상수 그리고 Attribute간의 유사한 정도를 정의해 둔 데이터 베이스를 말한다. 이러한 일반적인 형태를 가진 것에는 실수값의 연속적인 구조를 표현하는 continuous한 형태, 실수값에서 이산적인 구조를 표현하는 index 형태, 행렬 연산을 통해 도출해낸 문서간의 유사도를 이용하여 원하는 소속척도값을 구하는 Matric 형태등 다양한 모델이 있다. 그리고 데이터베이스 분야와 인공지능 분야의 여러 이론들에 의해 도출된 변환 데이터가 정의된 데이터베이스 형태를 말하기도 한다. 이와 같은 데이터마이닝 기법에서처럼 확장된 메타데이타 형태로 말할 수 있다. [6,8,11,12] 본 논문에서는 일반적 형태인 continuous 형태와 index 형태의 이론적 바탕을 토대로 하여, 사람들을대상으로 직접 실시한 설문을 통해 일반적인 사람들의 특성을 메타데이타로 정리하여 그 데이터를 시스템에 적용시켰다.

3.2 index형태의 응용

3.2.1 소속척도 값의 정의

다음 [표 1]는 본 연구를 위해 행한 설문조사의 총인원 300명에 대한 상세적 현황과, 설문에서 정한 관심도와 그에 따른 가중치를 나타낸 것이다.

[표 1] 설문조사 참가 인원현황과 관심도에 따른 가중치

구분	초졸	중졸	고졸	대졸	대졸이상	계
10대	27	33				60
20대		2	19	42	7	70
30대	1	1	26	20	2	50
40대	13	2	21	13	1	50
50대	17	3	11	8	1	40
60대이상	16	1	10	3		30

관심도	가중치	인원수
매우관심(매우 좋아함)	5	α
관심(좋아함)	4	β
보통	3	γ
그다지 관심없음(별로 좋아하지 않음)	2	δ
관심없음	1	ε

각 질의어에 대한 소속척도값 U의 계산과정은 다음과 같다.

$$u=\frac{(\alpha\times5)+(\beta\times4)+(\gamma\times3)+(\delta\times2)+(\varepsilon\times1)}{N\times5} \cdots [식1]$$

$$U=u+(1-u)\times0.4\cdots \text{ [식2]}$$

식[1]에서 나온 값은 소속 구성수가 작을 경우 소속척도값이 너무 작은 값을 가지므로 실제 문서 검색

시 중요도가 너무 낮게 산정되어 거의 검색되지 않기 때문에, 식[2]에서 스케일링한 U의 값을 질의어의 소속척도 값으로 선정한다. 이와 같은 방법으로 나이, 학력, 성별등의 attribute에 해당하는 소속척도 값을 저장시킬 수 있다. [표 2]은 나이에 따른 정치의 관심도에 대한 해당 소속척도 값을 구한 것이다. attribute에 해당되는 사용자 입력 data는 문자이므로 대응되는 숫자로 변환시켜 내부적으로 처리한다.

[표 2] MetaData - 나이에 따른 정치 관심도

attribute	변환된 값	해당소속척도값
10대 이하	1	0.5
20대	2	0.7
30대	3	0.8
40대	4	0.9
50대	5	0.9
60대 이상	6	0.8

3.2.2 연관사전에의 적용

보다 나은 사용자 특성적인 검색결과를 도출하기 위해 하나의 키워드뿐만 아니라 관련어 사전을 이용하여 한 단계 더 정밀한 수준의 검색결과를 도출할 수 있다.

예를 들어 "컴퓨터'라는 키워드가 입력되었다면 그것과 관련된 단어들, 즉 '워드프로세서', '인터넷', '프로그래밍', '하드웨어' 등 많은 언어들이 관련된 언어들로써 관련어 사전에 등록될 수 있다. 그러면 이 단어들에게도 Meta DB가 연결되어 이들 단어들의 소속척도 값이 3.2.1과 같은 방법으로 할당되어 등록된다. 그렇게 등록된 Meta DB안의 질의어 처리할 때, 연관어로 간주되어 동시에 똑같이 메타DB를 참조하여 검색결과 출력에 관여하게 된다. 그림으로 살펴보면 다음과 같다. 참고로 아래에 나타난 소속척도값은 설문결과를 토대로 작성하였다.

[표 3] 연관어 사전과 나이에 관한 소속척도값

컴퓨터	워드프로세서
	인터넷
	프로그래밍
	하드웨어
	교육용 소프트웨어

구분	10대	20대	30대	40대	50대	60대
워드프로세서	0.7	0.4	0.7	0.65	0.25	0.19
인터넷	1	1	1	1	1	0.6
프로그래밍	0.6	0.95	0.5	0.3	0.5	0.03
하드웨어	0.25	0.4	0.35	0.25	0.25	0
교육용 S/W	0.2	0.15	0.7	0.55	0.65	0.55

3. 3 퍼지 결합 연산 이론 및 응용

1965년 zadeh교수에 의해 처음 퍼지 개념이 도입된 이래 많은 학자들이 퍼지 연산에 관한 식을 연구하였고, 현재에도 계속되고 있다. 특히 두 개의 피연산자를 결합하여 새로운 값을 산출하는 퍼지 and와 퍼지 or 방식에 많은 이론적 정립이 있었다. 그리고 and와, or 연산자에 대한 피연산자 의존 문제(Single Operand Dependancy problem)와 부정적 보상문제(Negative Composition Problem)에 대한 문제를 해결하기 위해 평균연산자가 도입되었고 본 연구에서는 [표 4]와 같은 연산자를 채택하였다.[5,8,14,15]

[표 4] Fuzzy and, Fuzzy or

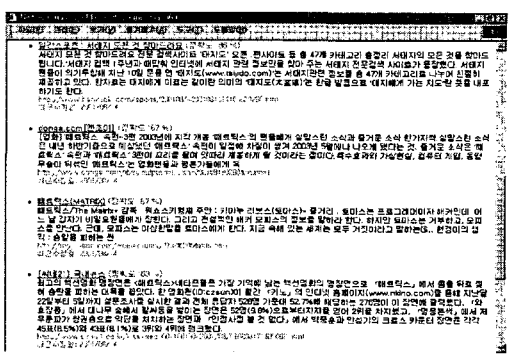
Fuzzy And	$Min(x,y)+(1-\lambda)(x+y)/2 \quad 0<\lambda<1$
Fuzzy Or	$Max(x,y)+(1-\lambda)(x+y)/2, \ 0<\lambda<1$

각 키워드들에 의해 병렬적으로 검색되어진 문서들은 서로간의 불리언 연산을 수행할 때, 이 연산자를 사용함으로써 전체 키워드 중에서 하나의 키워드만으로 검색된 문헌들이 전체를 좌지우지하지 않고 모든 키워드에 의해 상황에 맞는 최종적인 적당한 문헌의 중요도를 산정할 수 있도록 하였다.

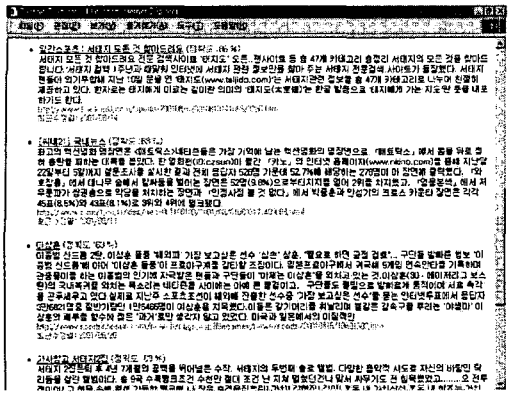
4. 프로토타입 구현

전체 시스템 중 로봇모듈과 질의처리 모듈은 자바로 구현하였으며 색인모듈은 C++로 구현되었다. 각 모듈별 인터페이스는 앞에서 [그림 1]에서 보는 바와 같이 파일단위로 정보를 주고받으며, 각 모듈은 상호 독립적으로 운용된다.

본 프로토타입은 전체적인 화면은 다음과 같다. 질의어로써 자연어를 입력받고 사용자에 따라 사용자 정보를 입력할 수 있도록 하였다. 사용자 정보 미입력시 해당 질의어들에 대한 소속척도 값은 0.7로 단일화하였다. 그리고 연관어 사전에 대한 수준을 조절함으로써 연관단어들까지의 Metadata도 참조할 수 있게 하였다. [그림 4]와 [그림 5]는 '컴퓨터에 대해서' 라는 같은 질의어에 다른 사용자 정보를 입력시켰을 경우 Meta 정보의 차이에 의해 다른 검색결과의 출력화면을 보여준다.



[그림 4] 20대, 고등학교졸, 여자로 검색한 경우



[그림 5] 60대, 초등학교졸, 여자로 검색한 경우

5. 결론 및 향후과제

나이와 학력 두가지 사용자 정보를 입력 MetaDB를 참조하여 검색한 결과 단순히 키워드만 입력했을때보다 사용자의 특성에 맞는 결과가 도출되었음을 알수 있었다. 본 연구에서는 나이와 학력 두가지 항목만으로 한정시켜 MetaData를 구성하였지만 보다 많은 질의어들에 대해 일반적으로 검색될 수 있도록 MetaDB의 확장이 필요한 동시에 데이터 마이닝 기법을 적용시켜 현재의 단순한 연관어 사전 수준을 벗어나 추론 기법을 이용한 MetaDB의 구축에도 보다 많은 연구가 필요할 것이다.

그리고 방대한 MetaDB안의 알맞는 소속척도 값을 계산하기 위한 수행 시간 또한 병렬적인 알고리즘을 이용하여 보다 짧은 시간안에 처리될수 있도록 하여 검색 시스템의 효율에 관심을 기울어야 할 것이며 보다 많은 일반적인 사용자가 수긍할 수 있는 MetaDB 설계와 구축에 많은 연구가 선행되어야 할 것이다.

참고문헌

[1] William B.Frakes and Richard Baeza-Yates., Information Retrieval.,Prentice Hall.,1992

[2] Richard Baeza-Yates.,Modern Information Retrieval., Addison Wesley.,1999

[3] Korth H. F. and Silberschatz, A., Database System Concepts, McGRAWHILL, N. Y., 1991.

[4] Date, C. J., An Introduction to Database Systems 1, Addison Wesley, 1985.

[5] Zadeh, L. A., "Fuzzy Sets", *Information Control 8*, pp.338-353, 1965.

[6] S.C. Park, C.S. Kim and D.S, Kim, Fuzzy Logic and its applications to engineering information sciences, and intelligent systems, KLUWER ACADEMIC PUBLICSHERS, 1995, pp. 407-415.

[7] Robert R. Korfhage., Information Storage and Retrieval.,1997

[8] 신세영, "인적사항 데이터베이스 조회를 위한 퍼지질의 시스템의 FSQL, 처리" 한국정보처리학회 학술발표논문지 제7권,제1호, 2000.4.

[9] 한국어정보처리연구소, C로 구현한 인터넷 정보검색 시스템, 1999

[10] 이광형, 오길록, 퍼지이론 및 응용 II권 홍릉과학출판사, 4장, 1991.

[11] 이도헌,이광형,김명호, "퍼지 질의를 위한 관계대수의 확장", 정보과학회논문지 제20권, 제2호, 1993.2.

[12] Buckles, B.P. and Petry, F. E., A Fuzzy Representation of Data for Relational Database, Fuzzy Sets and Systems 7, pp.213-226, 1982

[13] Lee K. H. and Oh,G.R., Fuzzy Sets and Applications 2, Hongreung Science Publishing Company,1991.

[14] Zadeh, L. A., "Outline of a New Approach to the Analysis of Complex Systems and Decision Process", *IEEE Trans. on Systems, Man, and Cybernetics SMC-3.*, 1973

[15] Zemankova, M. and Kandel, A., "Implementing Imprecision in Information systems", *Information*

[16] Michael T.Goodrich and Roberto Tamassia.,Data Structures and Algorithms in Java.,1999