# 정보검색을 위한 자연언어 질의어의 불리언 질의로의 변환

서광준*, 나동열**, 최기선*
한국과학기술원 전산학과*, 연세대학교**

# A System for converting natural language queries into boolean queries for Information Retrieval

Kwang-Jun Seo*, Dong-yeol Ra**, Kye-sun Choi*
Department of Computer Science-KAIST*, Yonsei Univ.**

## 요 약

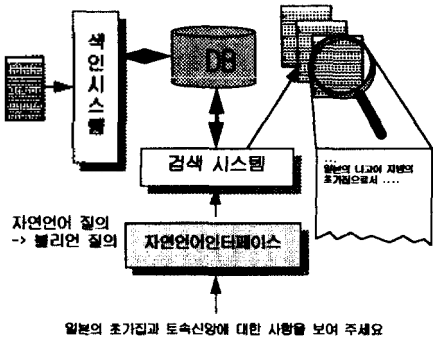
자연언어 인터페이스는 초보자나 비숙련가의 입장에서는 새로운 시스템의 적용에 있어서 어떤 학습도 필요하지 않다는 강력한 장점이 있다. 이 연구에서는 불리언 질의를 처리하는 정보검색 시스템의 자연언어 인터페이스를 구현하였다. 즉, 한국어 자연언어 질의를 불리언 질의로 변환해주는 시스템이다. 접근 방법은 먼저 자연언어 질의를 구문 해석한 후에, 그 결과인 문장의 의존 구조와 불용어 정보를 사용하여 기본적인 불리언 질의를 만든 다음, 시소러스를 이용하여 불리언 질의를 확장한다. 여기에서 사용한 구문 해석 방법은 의존 문법에 기반한 [서광준93] 방법이다. 변환 시스템은 SPARC-II 호환기종에서 구현되었으며, 약 5만 단어의 사전을 사용한다. 가공된 120 개의 질의를 대상으로 실험한 결과, 전체 소요 시간은 13.5(질의당 0.11)초가 걸렸다. 그리고, 변환된 불리언 연산식중에 110 개(91.6%)가 적절하게 변환된 것으로 조사되었다.

## 1. 서 론

정보서비스의 필요성이 커짐에 따라 정보를 얻고자하는 이용자들의 수와 정보 이용량 또한 급격히 증가하고 있는 추세이다. 특히, 일반인들을 대상으로 하는, PC 통신을 이용한 전문 데이타베이스 서비스에 있어서도 이와 같은 현상이 두드러진다. 이와 같은 상황에서, 사용자가 원하는 정보를 쉽고, 정확하게 찾아주는, 사용자 중심이며 지능을 갖춘 사용자 인터페이스 개발의 중요성이 강조되고 있다. 이러한 시각에서 보면, 자연언어 인터페이스는 초보자나 비숙련가의 입장에서는 새로운 시스템의 적용에 있어서 어떤 학습도 필요하지 않다는 것과, 전문가 입장에서도 자기가 필요한 자연언어 명령어만을 간결하게 사용하여 원하는 결과를 얻을 수 있게 하는 등, 자연언어의 특징인 유연성(flexibility), 명료성(succintness), 풍부한 표현력(expressiveness) 등을 그대로 살려줄 수 있다는 장점으로 정보 서비스의 사용자 인터페이스에 유용하게 사용될 수 있다.

그러나, 자연언어 인터페이스는 개발 비용이 비교적 많이 든다. 개발자의 입장에서 보면 자연언어처리에 내재되어 있는 형태소, 구문 및 의미분석 단계에서 각종 애매성 처리에 어려움이 있으며 자연언어의 방대함에서 오는 대량 사전의 필요성 등의 난제가 산적해 있는 등, 해결해야 될 많은 문제점들이 있다. 이와 같은 어려움에도 불구하고 자연언어 인터페이스는 강한 매력과 잇점들이 있으므로, 차세대 사용자 인터페이스로 많은 연구가 진행되고 있으며 외국의 경우 몇몇 상용화된 시스템들이 등장하고 있는 실정이다. 상업적으로 사용가능한 자연언어 인터페이스 시스템들은 대부분 데이타 베이스의 전위 시스템으로서 개발되었으며 사용자들이 잘 정제된 질의만을 한다는 것을 가정하고 있다.

이 연구에서는 정형, 비정형 데이타를 모두 수용하며, 불리언 질의(Boolean Query)를 처리하는 정보검색 시스템의 간단한 자연언어 인터페이스를 구현하였다. 즉, <그림 1>과 같이 한국어 자연언어 질의를 검색 시스템의 질의어인 불리언으로 변환하는 시스템을 개발하였다. 이 논문에서는 이에 관한 연구 사항을 소개한다. 논문의 구성은 II 장에서 시스템의 입력과 출력을 살펴보고 변환 예를 들어본다. III 장에서는 변환 방법을 기술한다. IV 장에서는 구현 환경과 실험 및 결과에 대해서 기술하며 V 장에서는 결론을 맺는다.




<그림 1> 불리언 질의 정보 검색 시스템

## II. 입력과 출력

자연언어 인터페이스의 입력은 자연언어 질의이며, 출력은 불리언 질의이다. <그림 2>에 입력과 출력의 몇가지 예를 기술하였다.

예1. 김영삼 대통령의 공약은 무엇인가.
=> 김영삼 AND 대통령 AND 공약
예2. 천리안, 하이텔
=> 천리안 OR 하이텔
예3. 고향이 대전이 아닌 국회의원은 누구인가?
=> 고향 NOT 대전 AND 국회의원
예4. 저자가 김대중씨인 정치 관련 책
=> (2=김대중) AND 정치
예5. 1994년에 출판된 소프트웨어에 관련된 책은?
=> (3=94/01/01:94/12/31) AND 소프트웨어

<그림 2> 입력과 출력의 예

### 2.1 자연언어 질의

자연언어 질의는 정형화된 것이 아니라 찾고자 하는 정보를 일상 생활에서 흔히 쓰는 문장으로 표현한 것을 말한다. 여기에서는 자연언어 검색식을 일반 문장 형태와 몇 개의 단어가 SPACE, COMMA(,) 등으로 연결된 단순나열형(예2)으로 구분한다. 또한, 일반 문장형태는 다시 완전한 문장(예1,3)과 단축된 문장(예4,5)으로 구분할 수 있다.

### 2.2 불리언 연산식

불리언 연산식은 찾고자하는 데이타를 표현하는 검색어와 이것들의 결합관계를 기술하는 불리언 연산자, 그리고 정형 데이타 필드에 관한 사항을 기술하는 필드식으로 구성된다. 또한, 필드식은 특정한 정형 데이타 필드를 나타내는 항목번호와 그 필드가 만족해야하는 사항을 기술하기 위한 비교자와 단위어로 구성된다. 구체적인 연산식의 구 구조는 다음과 같다.

```
연산식 -> (연산식 연산자 연산식) |
          필드식 |
          검색어
필드식 -> (항목번호 비교자 단위어) |
          (항목번호 = 단위어 : 단위어)
연산자 -> AND | OR | NOT
비교자 -> = | > | < | => | =<
단위어 -> 숫자 | 날자 | 문자열
```

<그림 3> 불리언 연산식 구조

불리언 연산자에 대한 자세한 사항은 [salton 89]에 잘 기술되어 있다. 필드식은 정형 데이타 필드의 값의 정의에 사용된다. 특히, 첫번째 규칙은 특정 값이나 크기의 지정을 기술하며, 두번째 규칙은 범위를 기술한다. 예를들면, (예3)에서는 저자 필드(항목번호인 2)가 특정 값인 "김대중"임을, (예5)에서는 출판일자 필드(항목번호 3)은 범위 값을 가짐을 알 수 있다.

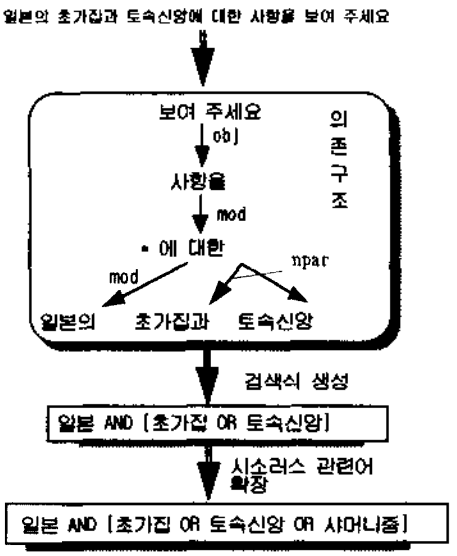
## III. 변환

사용자의 자연언어 질의를 불리언 질의로 적절하게 변환하는 것은 매우 어렵다. 상당한 의미 해석 기술이 필요할 뿐 아니라, 사람에 따라서 동일한 질의를 이해함에 차이가 있기 때문이다. 여기에서는 구문 해석 수준의 기술로서 구문 패턴에 의해 간단하게 변환하는 방법을 연구하였다.

불리언 연산식은 검색어(항목번호, 단위어 포함)와 연산자(비교자 포함), 그리고 결합구조 (nested structure)로 구성된다. 자연언어 질의를 불리언으로 변환하기 위해서는 이러한 정보들을 자연언어 질의로부터 획득해야한다.

검색어와 연산자를 결정하기 위해서 형태소 해석 정보가 필요하며, 결합 구조를 결정하기 위해서는 자연언어 질의의 구문 구조가 필요하다. 따라서, 구문 해석 수준의 자연언어 해석 기술이 필요함을 알 수 있다. 또한, 불필요한 검색어를 제거하기 위한 불용어 정보와 검색어들을 확장하기 위해 시소러스 정보를 이용한다.

구문 해석은 [서광준93]의 방법을 사용하였다. 이 방법은 의존문법을 기반으로 하며, 특히 문장을 구성하는 요소들 사이의 의존 관계를 퍼지(fuzzy)하게 표현하여 견고성과 최적해 선정의 용이성을 제공한다.




<그림 4> 변환 과정

기본적인 변환 방법은 <그림 4>와 같이 먼저 자연언어 질의를 구문 해석하고, 그 결과인 의존 구조를 탐색하면서 불용어 사전을 참조하여 초기 불리언 질의를 만든 다음, 시소러스를 이용하여 불리언 질의를 확장한다.

### 3.1 검색어 인식

자연언어 질의에서 불리언 질의에 사용되는 검색어를 인식하는 방법은 다음과 같다.

① 문장 검색어 : 문장의 의존 구조에 나타나는 모든 명사들 중에서 불용어가 아닌 단어(필드항목 제외)

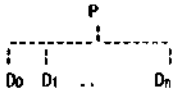
② 확장 검색어 : ①의 검색어의 시소러스 상에서 관계된 단어

이러한 인식을 위해서 두 가지의 다른 선처리가 필요하다. 그 하나는 정형 데이타 필드를 나타내는 단어(필드단어)를 인식하여 그 단어를 항목번호로 대치하는 것이다. 필드단어는 일반 검색어로 사용될 수 없고 사용될 경우는 항목번호로 대치되어 필드식에만 나타난다는 것을 가정한다. 다른 하나는 질의어에 나타나는 날짜, 단위 등을 단위어를 인식하여 정형화된 형태로 재구성한다. 날짜의 경우에는 다음과 같은 예를 들 수 있다.

1994년 12월 10일 => #y:94/12/10
1994년 12월 => #y:94/12/01:94/12/31
12월 10일 => #y:94/12/10

### 3.2 연산식 생성

연산식 생성은 연산자와 결합 구조의 결정에 의해 이루어 진다. 먼저, 입력이 <그림 2>의 (예3) 처럼 나열형 질의인 경우에는 검색어들을 각각 OR로 연결한 식의 형태로 변환한다. 그리고, 문장형태인 경우에는 질의의 구문분석 결과인 의존 구조의 각 노드에서 <그림 5>와 같은 규칙을 적용함으로써 수행된다. 규칙에 사용되는 용어들과 기호들에 관한 자세한 설명은 [서광준 93]을 참조하기 바란다. 규칙들은 각 노드(P)의 자식 노드(Child Node)들에 의해 이미 생성된 연산식($D_0$-$D_n$)에 의해 다음과 같이 정의된다.




<그림 5> 각 노드에서의 규칙 적용 틀

① P가 검색어일때는 "[$D_0$ AND $D_1$ AND ... $D_n$ AND P]" 가 생성된다.

예) 도시
프랑스의 => [프랑스 * 도시]

② P 의 실질 형태소가 "아니-","제외하-" 등의 NOT을 유발하는 단어이고, $D_i$가 주어나 목적어 일때,

"[$D_0$ AND $D_1$ AND ...NOT $D_i$ AND .. $D_n$]"을 생성한다.

예) 제외하고
서울에서 호텔을 => [서울 NOT 호텔]

③ P 가 "$"(지정사,"-이다" 형태)이고, $D_i$ 가 필드 검색어인 경우에, $D_i$ 가 주어이고 $D_j$ 가 보어인 경우와 $D_i$ 가 보어이고 $D_j$ 가 주어인 경우는 다음을 생성한다.

"[$D_i$ = $D_j$] AND $D_k$ ] k = 1.. n,k != i != j"

예) "$"이다
저자(2)가 홍길동 => [저자(2) = 홍길동]

④ P의 실질형태소가 "*",("와" 같은 병렬 명사구)일때, P 의 부모 노드의 실질 형태소가 "대결", "회의" 등과 같이 공동격을 하위격으로

가지면, "[$D_0$ AND $D_1$ AND ... AND $D_n$]",
아니면, "[$D_0$ OR $D_1$ OR ... OR $D_n$]" 를 생성한다.

예) 대결
"*"
서봉수와 이창호의 => [서봉수 AND 이창호]

⑤ 기타 모든 경우에는 "[$D_0$ AND $D_1$ AND ... AND $D_n$]"를 생성한다.

예) 잔다
광장에서 배낭족들이 => [광장 AND 배낭족]

위에서 기술한 생성 규칙은 많은 보완이 필요하며, 지속적인 연구가 필요하다. 예를들어, 필드식의 생성은 규칙 ③에서 기술한 것 이외에도 다양한 경우가 있을 수 있다. 즉, "홍길동이 쓴 책"인 경우도 마찬가지로 "[저자(2) = 홍길동]"의 결과가 나와야 한다. 그러나, 아직까지는 구문해석 수준에서 이런 다양한 경우에 대한 정형화가 어렵다. 그리고, 범위의 인식에 대한 능력이 없다. 즉, "20 평에서 30 평 사이의 아파트" 라는 질의에 "아파트 * 평수=20;30"과 같은

질의가 만들어져야 하지만, 마찬가지로 범위를 나타내는 구문적인 패턴의 정형화가 어려운 문제이다.

### 3.3 시소러스 확장

시소러스란 용어의 관련 단어들을 체계적으로 분류, 정리해 놓은 사전으로 단어의 동의어(Synonym), 광의어(Broad Term), 협의어(Narrow Term), 관련어(Related Term), 전거어(Authority) 등을 수록한 것이다. 여기에서는 관련어와 유사어 및 전거어에 해당하는 Term들만을 이용하여 OR 연산으로 검색식을 확장한다. 일반적으로 DB에 접근하는 사용자는 찾고자하는 정보에 대해 자세히 기술하지 않고 매우 피상적으로 기술하는 경향이 있으므로 두가지 관점에서 시소러스 확장이 고려되어야 한다. 첫째 질의어가 너무 광범위한 대상을 지칭할 경우 범위를 축소시킬 필요가 있고 둘째 범위가 너무 좁을 경우에는 범위를 넓힐 필요가 있다. 본 연구에서는 첫번째 경우는 고려하지 않고 아래에서 보이는 바와 같이 검색대상을 넓히는 경우만 고려하였다.

예) 철도로 여행을 하려면
=> 철도 * 여행
=> (철도+기차+열차) * (여행 + 관광)
철도의 관련어 여행의 동의어

## IV. 실험 및 분석

본 시스템은 SPARC-II 호환기종에서 구현되었다. 시스템에 사용한 사전 크기는 5 만여 단어이며 UNIX의 DBM으로 구축하였다. 실험용 대상 질의문은 데이콤 개발실에서 작성한 120 개의 예상 질의로서 평균 문장길이가 4.5 어절이며, 철자 오류, 띄어쓰기 등의 가능한 오류를 포함하고 있다.

Unix의 time 명령어로 측정한 전체 질의 처리 시간은 cpu를 98% 점유한 상태로, real(반응 시간)이 13.5(질의당 0.11)초, user 7.4(질의당 0.061)초, sys 5.6(질의당 0.046)초가 걸렸다. 수행 시간은 형태소 해석 단계에서 가장 많이 소요되는 것으로 나타났다. 구문 해석은 질의의 특성상 문장의 길이가 작은 것을 고려해서 근사해를 간단하고 빠른 알고리즘을 사용하여 찾기때문에 형태소 해석보다는 상대적으로 수행 시간이 빨랐다.

변환된 불리언 연산식중에 110 개(91.6%)가 적절하게 변환된 것으로 조사되었다. 적절하지 못한 결과를 보면 질의 해석에서 잘못된 경우가 대부분이었고, 불리언 생성에서 실패한 경우는 드물었다. 그리고, 질의 해석에서 잘못된 경우도 올바른 불리언 생성에 성공하는 경우도 있었다.

## V. 결 론

본 연구에서는 불리언 질의를 처리하는 정보검색 시스템의 자연언어 인터페이스를 연구하고 그 결과로서를 이용한 불리언 질의어 변환 시스템의 프로토타입을 구현하였다. 이 시스템은 사용자가 자연언어를 쓸 수 있도록 하기 위해 자연언어 질의어를 불리언 질의어로 변환하는 기능을 수행한다. 변환을 위한 구문해석은 [서광준93]을 이용하였었고, 구문해석의 결과인 자연언어 질의의 의존구조로부터 불리언 질의어를 생성하는 규칙들을 개발하였으며, 구문 해석으로는 완전히 해결할 수 없는 질의어 유형이 존재함을 발견하였다. 필드 연산, 범위 연산등은 상당한 수준의 이해가 필요하며 이는 전반적인 자연언어 이해 기술의 향상 없이는 달성하기 어려운 것이다.

이상적인 불리언 질의어 생성기는 시소러스나 의미사전 혹은 특정 분야에 대한 지식까지 활용하여 필드연산에 필요한 정보까지도 완벽하게 채워 주는 것이다. 현재로서는 구문 단계에서 분석을 끝내기 때문에, 일차적으로 구문 분석만으로 생성가능한 불리언 질의어와, 구문분석으로는 생성할 수 없는 질의어의 유형을 정확히 분류해내고, 그 다음에 구문분석의 수준을 넘어서는 더 정교한 불리언 질의어를 생성하기위해 필요한 정보들을 파악하고 구축하는 연구가 진행되어야 할 것이다.

## 참고 문헌

[서광준 93] 서광준, 최기선, "어절 사이의 퍼지 의존 관계를 이용한 구문 해석기", 제 20 회 한국정보과학회 가을 학술발표 논문집, pp 1151-1154, 1993.

[BART 85] M. Bartschi,"An Overview of Information Retrieval Subjects", IEEE Computer, May 1985.

[SALTON 89] G.Salton,"Automatic Text Processing", Addison-Wesley Publishing Company, 1989.

[한국과 92] 한국과학기술원, "자연언어 인터페이스를 위한 도구 환경의 연구개발", 과학재단, 1992.

[한국과 93] 한국과학기술원, "자능형 정보검색에 관한 연구", 한국통신, 1993.