학술자료 검색을 위한 자연언어 데이터베이스 인터페이스 시스템[1)]

임경업*, 이석형**, 윤화묵**, 권혁철*
*부산대학교 컴퓨터공학과
**한국과학기술정보연구원
*{iku88, hckwon}@pusan.ac.kr
**{skyi, hmyoon}@kisti.re.kr

Natural Language Database Interface System for Scholar Search

Kyoungup Im*, Seok-Hyoung Lee**, Hwa-Mook Yoon**, Hyuk-Chul Kwon*
*Dept of Computer Science, Pusan National University
**Korea Institute of Science and Technology Information

요 약

자연언어 데이터베이스 인터페이스는 자연언어를 데이터베이스의 쿼리(query)로 바꿔주는 시스템이다. 이를 통해, 데이터베이스에 잘 모르는 일반 사용자도 쉽게 데이터베이스를 이용할 수 있다. 본 논문에서는, 학술자료 검색에 사용되는 자연언어 데이터베이스 인터페이스 시스템을 소개한다. 패턴과 구문분석 기법을 동시에 사용하여 속도와 확장성을 모두 만족하게 한다.

1. 서론

자연언어 인터페이스(Natural Language Interface)란 어떤 시스템에 접근할 때, 자연언어를 이용할 수 있게 하는 인터페이스이다. 자연언어 데이터베이스 인터페이스(NLDBI, Natural Language Database Interface, 이하 NLDBI)는 사용자가 자연언어를 통해 데이터베이스에 접근할 수 있게 해준다[12]. 데이터베이스(Database)는 편리하지만 정형화된 별도의 인터페이스를 가지고 있어, 일반 사용자들이 사용하기에 부담스럽다. 자연언어 인터페이스를 통해 일반 사용자들이 데이터베이스를 부담없이 효율적으로 사용하도록 유도할 수 있고, 이것은 정보화 사회를 위한 근본적이 바탕이 된다고 하겠다[1].

선진 외국에서 NLDBI 시스템들은 이미 1970년대부터 개발되었다[1]. 초기의 NLDBI는 주로 한정된 도메인(specific domain)을 대상으로 개발되었으며, 점차 그 도메인을 확장적용하도록 연구가 진행되었다[12]. 현재는 주로 질의응답 시스템(Question Answering System)과 같은 범용 도메인에서 기본적으로 사용되고 있다.

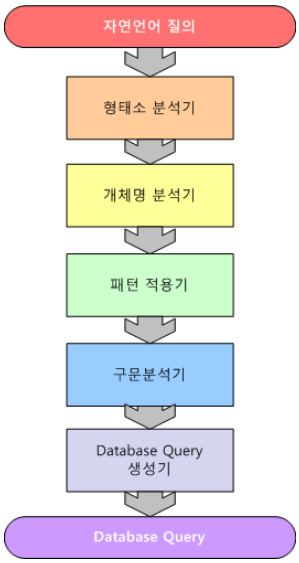
국내의 NLDBI 시스템들은 1980년대부터 개발되었다[5]. 초창기의 국내 NLDBI 시스템들은 자연언어 질의를 데이터베이스 쿼리(query)로 생성하는 전체적인 과정을 중점적으로 연구하였다[5][6][9]. 하지만, 국내의 한국어를 대상으로 한 NLDBI 시스템들의 개발은 국외에 미치지 못하는데, 자연언어 인터페이스의 핵심 부분인 자연언어 구문분석 기술이 아직 실용적인 수준에 이르지 못했기 때문으로 보인다. 2000년대에 들어서 질의응답 시스템들의 연구와 함께 NLDBI가 적용되고 있지만, 구문분석 정보를 활용하기보다는 패턴을 이용하는 것에 그치고 있다[7][11]. 패턴을 이용하면 구문분석 과정이 필요 없어 더욱 빠른 속도의 처리가 가능하지만, 매우 많은 패턴이 필요하며, 그 때문에 확장이 어렵다. 구문분석을 할 경우 이러한 문제가 해결되지만, 속도 문제와 아직 만족할 만한 일반적인 한국어 구문분석기가 개발되지 못하였다는 문제가 있다.

본 논문에서 제안하는 시스템은 NLDBI 시스템의 일종으로, 패턴 기법과 구문분석 기법을 동시에 적용한다. 우선 입력된 자연언어 질의문을 패턴화하여 몇 개의 묶음으로 만든 후, 패턴의 리스트에 대해 구문분석을 시도한다. 또한, 학술자료 검색이라는 제한된 도메인(specific domain)을 위한 시스템이므로, 구문분석 과정의 모호성을 크게 줄일 수 있었다. 학술자료 검색은 논문의 저자나 제목, 학회 등에 대한 정보를 찾는 것을 말한다.

2. 시스템 구조

본 논문에서 제안하는 시스템은 속도와 확장성 두 가지를 모두 만족시키기 위해 패턴 기법과 구문분석 기법을 혼합시켰다. 자연언어 질의를 형태소 분석한 결과에 바로 구문분석을 하는 것이 아니라, 주어진 도메인(domain)에 맞게 패턴으로 묶은 후 패턴들을 구문 분석한다. 시스템의 전체적 구조가 (그림 1)에 나타나있다.

1) 본 연구는 한국과학기술정보연구원의 지원으로 이루어졌음.



(그림 1) 시스템 구조

1) 형태소 분석기

입력된 자연언어 질의를 형태소 분석한다. 패턴을 적용할 때 단순히 스트링 매칭(string matching)을 하는 것보다 형태소 분석 정보를 가지고 적용하는 것이 더욱 정확하기 때문이다.

2) 개체명 분석기

학술자료 검색 대상 데이터베이스에서 필요한 개체명을 미리 추출한다. 주요 추출 대상으로 저자, 학회명 등이 있다. 개체명을 구분하고자, 기존의 품사 분류를 하위범주화한다. 개체명은 모두 고유 명사이므로, 고유 명사의 하위범주로 사람이름, 학회이름 등을 만들고, 개체명 분석기의 결과를 이용해 해당 고유명사의 품사 분류를 하위범주화한다.

3) 패턴 적용기

본 논문에서 제안하는 시스템은 3종류의 패턴을 사용하는데, 질의부, 조건부, 연결부가 그것이다. 각 패턴은 FSA(Finite State Automata)로 구성되어 있다.

질의부는 사용자가 궁금한 정보에 대한 패턴이다. 예를 들어, '홍길동이 2005년에 발표한 논문은 무엇입니까?'라는 자연언어 질의에서, 사용자가 최종적으로 궁금한 정보는 '논문'이다. 다른 예로, '2005년에 C학회에 논문을 제출한 사람은 누구입니까?'라는 질의문에서는, '사람'(저자)이 질의부가 된다.

조건부는 질의에서 나타난 단서이다. 처음 예에서, '홍길동'과 '2005년'이 단서가 된다. 즉, 질의부를 수식하여 그 결과를 제한하는 부분이다. 두 번째 예에서, '2005년'과 'C학회'가 조건부에 해당한다.

연결부는 질의부와 조건부를 제외하고, 자연스러운 문장을 위해 적용된 부분을 말한다. 실질적으로 생성할 데이터베이스 쿼리와는 관계가 없으며, 질의부나 조건부 사이에서 연결해주는 역할을 한다.

질의부와 조건부는 학술자료 검색 대상 데이터베이스의 필드에 따라 만들어진다. 예를 들어, 데이터베이스에 '저자명'에 대한 필드가 있고, '학회명'에 대한 필드명이 있다면, 이 두 필드는 모두 질의부로도 사용될 수 있고, 조건부로도 사용될 수 있다. 사용자가 질의하는 것은 데이터베이스의 내용일 것이며, 그것을 찾기 위한 정보 역시 데이터베이스의 내용일 것이기 때문이다.

다만, 패턴의 내용은 다르다. 저자명이 질의부로 쓰일 때는 저자/사람/글쓴이 등의 단어가 중심이 되겠지만, 조건부로 쓰일 때는 개체명 분석기로 분석한 '사람이름'이라는 분류정보가 중심이 될 것이다. 이처럼 패턴을 적용할 때에는 실제 텍스트의 스트링과 구문분석기에서 사용하는 분류 정보가 사용된다.

패턴을 구축하고자, 먼저 가능한 질의문 자체를 만들었다. 질의문은 "(저자명)이 (학회명)에 발표한 논문은 무엇입니까?"와 같이, 데이터베이스의 필드명을 이용하여 만들었다. 만든 질의문을 바탕으로 패턴을 구축했다.

학술질의 검색을 대상으로 하는 데이터베이스 필드의 종류는 약 80여 개이며, 현재 구축된 패턴은 165개이다. 데이터베이스의 필드 중, 내부적으로만 사용하는 필드들은 패턴구축에서 제외하였다. 대상 데이터베이스가 정해져 있고, 패턴을 바탕으로 구문분석을 하기 때문에 패턴의 수가 매우 적은 것을 알 수 있다. [11]에서는 3,254개의 패턴이 적용되었다.

4) 구문분석기

패턴으로 표현된 자연언어 질의에 대해 구문분석을 시도한다. 의존문법을 사용하며, [8]의 구문분석기를 사용한다. 다만, 구문분석의 모호성을 줄이고자 다음의 사항을 제약한다.

제약 사항: 구문분석 결과가 여러 개일 경우, 의존거리의 합이 가장 작은 구문분석 결과를 선택한다.

예를 들어, "A B C"의 구문분석 결과로 [[A B] C]와 [A [B C]] 두 개가 있을 때, 전자는 B가 A를 지배하고, C가 B를 지배하므로 각각 의존거리가 1이고 합은 2이다. 후자는 C가 B를 지배하고 C가 A를 지배하는데, C가 B를 지배하는 경우의 의존거리는 1이고 C가 A를 지배하는 경우의 의존거리는 2이다. 따라서 후자의 의존거리의 합은 3이며, 최종적으로 전자가 결과로 선택된다.

의존문법에서의 '지배'는 수식의 역관계라고도 설명할 수 있다. 즉, 수식 거리가 가까울수록 정답이라고 가정한다. 실제 자연언어에서는 아닌 경우도 빈번하게 발생하지만, 학술자료 검색 질의에서는 거의 일어나지 않는다. 그

이유는, 학술자료 검색의 질의는 어떤 질의부를 조건부들이 수식하는 형태로 구성되기 때문이다. 이 경우 의존거리가 먼 경우는 사람이 보기에도 모호성이 있기 때문에, 사용자들 역시 될 수 있으면 그런 질의를 사용하지 않는다.

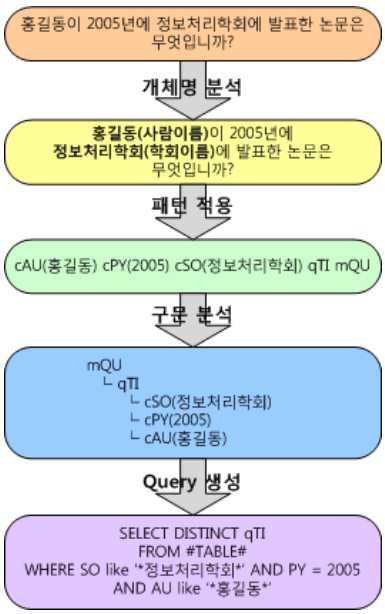
구문분석에 사용되는 의존규칙은 주로 질의부가 조건부를 지배하는 종류와, 연결부가 질의부와 조건부를 지배/의존하는 규칙들로 이루어졌다. 질의부에 따라서 조건부를 지배하지 못하는 일도 있는데, 예를 들면 '학회명'의 질의부는 '키워드'의 조건부를 지배할 수 없다. 대상으로 하는 데이터베이스에는 '논문'의 키워드 정보는 있으나, '학회명'의 키워드 정보는 없기 때문이다.

5) Database Query 생성기

패턴들의 구문 정보를 바탕으로 데이터베이스 쿼리를 생성한다. 현재 시스템은 SQL 쿼리를 출력하고 있다. 패턴 중 질의부는 '사용자가 찾는 정보'에 해당하므로 'SELECT' 절에 들어가며, 조건부는 'WHERE' 절에 사용된다.

'FROM' 절에 들어가는 테이블은 후처리로 따로 처리한다. 사용자는 데이터베이스 구조를 모르기 때문에, 어떤 테이블이 있는지, 어떤 테이블에서 찾아야 할지 직접적으로 알지 못한다. 따라서, 별도의 인터페이스를 통해 테이블을 선택해야 할 것이다. 다만, 경우에 따라 질의문 자체에 테이블에 대한 정보가 있을 수도 있는데, 이럴 때는 미리 선택을 한다.

쿼리를 생성하는 모듈을 분리함으로써, 다양한 종류의 데이터베이스 쿼리를 지원할 수 있다.



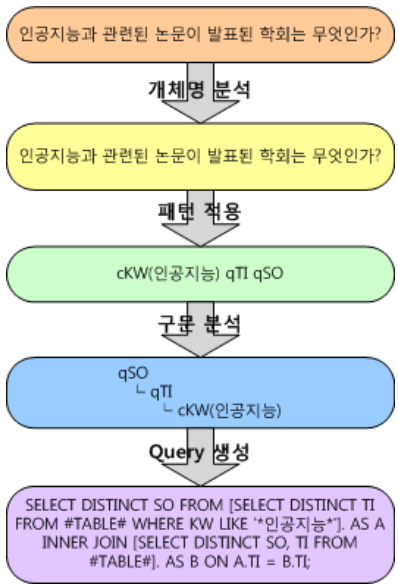
(그림 2) Query 생성 과정 예 - 단문

이상의 과정을 통해 자연언어 질의를 SQL 쿼리(query)로 변환하는 예가 (그림 2)에 나타나있다. 형태소 분석 과정은 생략했다.

cAU, cPY, cSO 등과 같이 c로 시작하는 패턴은 조건부를 의미하며, qTI와 같이 q로 시작하는 패턴은 질의부를 의미한다. mQU와 같이 m으로 시작하는 패턴은 연결부에 해당한다. AU는 저자명, PY는 발행연도, SO는 학회, TI는 논문명에 해당하는 데이터베이스 필드명이다.

(그림 2)의 예는 단문이다. 본 논문에서 제안하는 시스템은 복문에 대해서도 지원한다. 학술검색 질의에서, 복문의 형태는 질의부가 2개 이상인 경우이다. 이 경우, 앞쪽 질의부를 이용해 Query를 만들고, 그 결과를 이용해 다시 Query를 만들어야 한다. 방법은 JOIN 연산을 이용하는 방법과 저장 프로시저(procedure)를 이용하는 방법이 있는데, 본 시스템에서는 JOIN 연산을 이용하고 있다. 호환성이 좋기 때문이다. 복문의 예는 (그림 3)에 나타나있다. 질의부의 중첩에 따른 예를 보이고자 조건부를 단순화했다.

KW는 논문의 키워드가 저장된 데이터베이스 필드이다.



(그림 3) Query 생성 과정 예 - 복문

3. 결론 및 향후 과제

NLDBI는 Database에 잘 모르는 일반 사용자도 편리하게 사용할 수 있도록 한다. 자연언어 인터페이스는 미래 정보화 사회의 핵심 기술이다. 본 논문에서는 학술자료 검색이라는 제한된 도메인의 데이터베이스에 자연언어로 질의하는 시스템을 소개하였다. 패턴과 구문분석을 동시에

적용하여 속도와 확장성을 모두 만족하게 하고자 하였으며, 데이터베이스에서 제공되는 정보에 맞게 구문분석 문법을 제한하였다. 구문분석 기법을 적용하여 복문의 자연언어 질의 역시 쉽게 SQL 쿼리(query)로 만들 수 있었다.

본 논문에서 최종적으로 생성하는 SQL 쿼리(query)는 아직 최적화 기법이 적용되지 않았다. 각 테이블과 스키마에 맞춘 쿼리(query) 최적화 모듈이 개발되어야 할 것이다. 인터넷의 사용자들은 단 수초의 기다림도 허용하지 않기 때문이다[7].

각 패턴의 FSA 확장은 지속적으로 계속되어야 하는 작업이다. 실제 입력되는 다양한 질의문을 바탕으로 대부분의 자연언어 질의를 패턴으로 만들 수 있어야 한다.

참고문헌

[1] 김한우, “데이터베이스 검색을 위한 자연언어 인터페이스 시스템” 데이터베이스월드, 한국데이터베이스진흥센터, 1995년 7월

[2] 박기선, 정혜경, 이근용, 이용석, "자연어 질의 문맥 구조를 이용한 효과적인 정보검색" 한국인터넷정보학회 2005 정기총회 및 추계학술발표대회 제6권 제2호, pp.427-431, 한국인터넷정보학회, 2005년 11월

[3] 박미화, 원형석, 이원일, 이근배, "구문 분석에 기반한 자연어 질의로부터의 불리언 질의 생성" 제 10회 한글 및 한국어 정보처리 학술대회, pp.73-79, 한국정보과학회/한국인지과학회, 1998년 10월

[4] 박세영, "[기술해설]멀티미디어 정보검색에서의 한국어 정보처리" 정보과학회지 제12권 제8호, pp.60-66, 한국정보과학회, 1994년 9월

[5] 이석호, 임해철, 김성기, "자연 한글 질의어 처리를 위한 인터페이스의 설계및 구현" 한국정보과학회 1984년도 가을 학술발표논문집 제11권 제2호, pp.190-195, 한국정보과학회, 1984년 10월

[6] 이석호, 김성기, "자연 한글 질의어 처리를 위한 인터페이스의 설계 및 구현" 정보과학회논문지 제12권 제1호, pp.31-44, 한국정보과학회, 1985년 2월

[7] 이승우, 이근배, "유한패턴매칭을 이용한 자연어 질의응답 시스템" 정보과학회지 제22권 제4호, pp.21-27, 한국정보과학회, 2004년 4월

[8] 임경업, 정영임, 권혁철, "한국어 어휘의미망에 기반한 논항 정보를 이용한 의존문법 구문분석기의 구현," 제19회 한글 및 한국어 정보처리학회, 2007, pp. 158-164.

[9] 채진석, 김성기, 이석호, "한국어 데이타베이스 질의 시스템의 설계 및 구현" 정보과학회논문지 '제20권 제6호, pp.810-820, 한국정보과학회, 1993년 6월

[10] 채진석, 이석호, "객체 지향 데이타베이스를 위한 한국어 질의 인터페이스에서의 경로식 처리" 정보과학회논문지 제22권 제8호, pp.1193-1202, 한국정보과학회, 1995년 8월

[11] Hyo-Jung Oh, Chung-Hee Lee, Changki Lee, Ji-Hyun Wang, Yi-Gyu Hwang, Hyeon-Jin Kim and Myung-Gil Jang, "Heterogeneous Answer Acquisition Methods in Encyclopedia QA", LNCS Volume 4224/2006, pp.346-354, 2006

[12] In-Su Kang, Seung-Hoon Na, Jong-Hyeok Lee, "Conceptual Schema Approach to Natural Language Database Access" Proceedings of the Australasian Language Technology Workshop Volume 1, December 2003