자연언어 데이터베이스 인터페이스 시스템을 위한 프레임워크

임경업, 권혁철
부산대학교 컴퓨터공학과
{iku88, hckwon}@pusan.ac.kr

A Framework for Natural Language Database Interface System

Kyoungup Im, Hyuk-Chul Kwon
Dept of Computer Science, Pusan National University

요 약

자연언어 데이터베이스 인터페이스 시스템은 입력된 자연언어를 데이터베이스의 질의문(query)으로 바꿔주는 시스템으로, 데이터베이스에 잘 모르는 일반 사용자도 쉽게 데이터베이스를 이용할 수 있게 하는 장점이 있다. 본 논문에서는, 범용적인 분야의 자연언어 데이터베이스 인터페이스 시스템을 설계하기 위한 하나의 틀을 제안한다. 패턴 매칭과 구문 분석 기법을 동시에 사용하여 자연언어 처리 능력과 속도를 향상시켰으며, 패턴을 4개 분류로 나누어 의미 처리를 가능하게 하였다.

1. 서론

자연언어 데이터베이스 인터페이스 (Natural Language Database Interface, 이하 NLDBI) 시스템이란, 데이터베이스 기반 시스템에서, 사용자가 입력한 자연언어 질의문을 데이터베이스 질의문(query)으로 바꾸어주는 시스템이다[6]. 사용자에게 친숙한 자연언어 인터페이스를 가능케 하여, 데이터베이스를 좀 더 쉽고, 더 상세하게 이용할 수 있도록 도와준다.

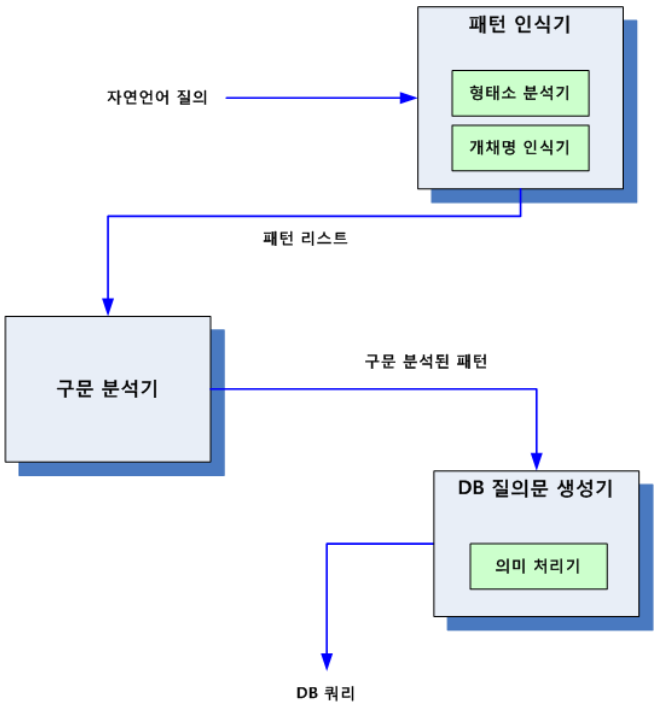
국외 NLDBI 시스템에 대한 연구는 1970년대 시작되었다[1]. 한정된 분야를 대상으로 시작된 국외 NLDBI 시스템은 질의응답 시스템(Question Answering)으로 확대되어 연구되고 있다[4]. 국내 NLDBI 시스템 연구는 1980년대부터 시작되어[2], AnyQuestion[7]과 같은 질의응답 시스템으로 확장되었지만, 기반 기술인 자연언어 구문 분석 기술의 미흡으로 패턴 매칭 기법에 의존하여[3][5] 아직 만족할만한 성과가 나오지 않고 있다. 패턴 매칭 기법은 시스템이 복잡해질수록 필요한 패턴의 수가 매우 늘어나고, 패턴 간의 충돌도 잦아지기 때문에 확장성이 떨어지는 단점을 가지고 있다.

본 연구에서는, 패턴 매칭과 간단한 구문 분석 기술을 이용한 NLDBI 시스템의 틀(framework)을 제시한다. 주어진 자연언어 질의 문장을 패턴 매칭하여 패턴화된 문장의 구성요소를 4개의 카테고리(category)로 분류한다. 이 카테고리를 이용하여 구문 분석을 쉽게 하고, 의미 처리를 가능하게 한다.

설명을 쉽게 하고자, [4]의 시스템을 예로 사용한다. 이 시스템은 학술자료 검색 데이터베이스를 위한 NLDBI 시스템이다.

2. 전체 구조

제안하는 프레임워크의 전체 구조는 (그림 1)과 같다.



(그림 1) 프레임워크의 전체 구조

입력된 하나의 자연언어 질의는 패턴 인식기에 의해 여러 개의 패턴으로 분석된다. 패턴 리스트는 구문 분석기에 의해 구문 분석 구조를 가지게 되며, 이를 바탕으로 데이터베이스(이하 DB) 질의문을 생성한다.

1) 패턴 인식기

본 연구에서는, 자연언어 질의 문장의 구성 요소들을 4개의 카테고리로 분류하였다. 4개의 분류는 <표 1>과 같다.

<표 1> 자연언어 질의 문장 구성 요소의 분류

분류	간략 설명
질의부	사용자가 알고자 하는 것. 최종적으로 생성될 DB 쿼리에서 SELECT 절에 들어갈 DB 필드(field)에 대한 정보를 담고 있음. DB schema에 의존함.
조건부	검색 조건. 최종적으로 생성될 DB 쿼리에서 WHERE 절에 들어갈 DB 필드(field)에 대한 정보를 담고 있음. DB schema에 의존함.
의미부	특수한 처리를 요구하는 자연언어 질의가 의미부에 해당함. 복잡한 DB 질의문이 요구됨.
연결부	질의부, 조건부, 의미부를 연결해 주는 것으로, 최종 DB 질의문에 영향을 주지 않음.

질의부란, 입력된 자연언어 질의 문장에서, 최종적으로 사용자가 묻고자 하는 정보를 나타낸 부분이다. [4]의 학술자료 검색 질의문을 예로 들어보자. 표본 DB가 <표 2>에 있다.

<표 2> 표본 DB

논문명	저자	연도	학회
논문A	홍길동	2005	한국정보처리학회
논문B	신사임당	2003	한국정보처리학회
논문C	이순신	1999	한국정보처리학회

(예 1) 홍길동이 2005년에 한국정보처리학회에서 발표한 논문은 무엇입니까?

(예 1)과 같은 자연언어 질의에서, 사용자가 최종적으로 궁금한 것은 논문(논문명)이다. 패턴 인식기는 '발표한 논문은 무엇입니까?'를 '질의부_논문명'이라는 패턴으로 묶는다. 이후, DB 쿼리 생성기에서 '질의부_논문명' 패턴을 만나면 SELECT 절의 내용을 선택하게 된다.

질의부에 속하는 패턴은 DB의 schema에 의존한다. 즉, DB 필드로 존재하는 정보만을 출력할 수 있게 된다. <표 2>의 DB에서, 질의부는 '질의부_논문명', '질의부_저자', '질의부_연도', '질의부_학회'의 4가지 패턴이 가능하다. "2005년 정보처리학회에 논문을 발표한 저자의 연락처는 무엇입니까?"와 같은 질의는 대답할 수 없다. DB에 해당 정보가 없기 때문이다.

조건부란, 입력된 자연언어 질의 문장에서, 결과를 한정해주는 조건이 나타난 부분이다. 앞의 (예 1)에서, '홍길동'과 '2005년', '한국정보처리학회'가 조건부에 해당한다. 이들은 DB 쿼리에서 WHERE 절에 조건으로 들어가게 된다.

조건부 역시, DB schema에 의존한다. 이는 WHERE 절에 이용해야 하기 때문이다. 다음과 같은 예를 보자.

(예 2) 2005년에 부산에서 발표한 논문은 무엇입니까?

(예 2)에서, '부산에서'는 조건부가 될 수 없다. DB에 적용할 수 없는 정보를 담고 있기 때문이다. <표 2>에서 가능한 조건부는 '조건부_논문명', '조건부_저자', '조건부_연도', '조건부_학회'가 된다.

질의부는 최종적으로 'DB의 어떤 필드를 묻는가'만 찾으면 되지만, 조건부는 '어떤 필드의 어떤 값인가'까지 찾아야 한다. 즉, 앞의 (예 1)에서, 질의부는 그저 '논문명'이라는 필드명만 찾으면 되지만, 조건부는 '저자=홍길동', '연도=2005', '학회=한국정보처리학회'임을 찾고, 그 정보를 구문 분석기에 전달해야 한다.

의미부는 특수 처리가 필요한 부분을 나타내는 부분이다. 가장 대표적인 예가 'NOT'인데, "홍길동이 발표하지 않은 논문은 무엇인가?"와 같은 질의문을 예로 들 수 있다. 이러한 의미부는 각 패턴에 맞춰 후처리를 해주어야 하며, DB 쿼리로 표현하기 불가능할 때도 있다.

연결부는 다른 3개의 분류에 속하지 않는 모든 패턴이 여기에 속한다. 생성 DB 쿼리에는 영향을 주지 않는다.

패턴 인식기는 먼저 주어진 자연언어 질의문을 형태소 분석하고, 개체명 인식기를 통해 고유 명사, 복합 명사 등을 인식한다. 각각의 패턴은 여러 개의 유한 오토마타(finite automata)로 구성되어 있는데, 이는 자연언어의 여러 형태가 하나의 패턴으로 묶이기 때문이다. 가령, "논문은 무엇입니까?", "논문은 무엇인가?", "논문은 뭐지?" 등은 모두 '질의부_논문명'으로 바뀌게 된다.

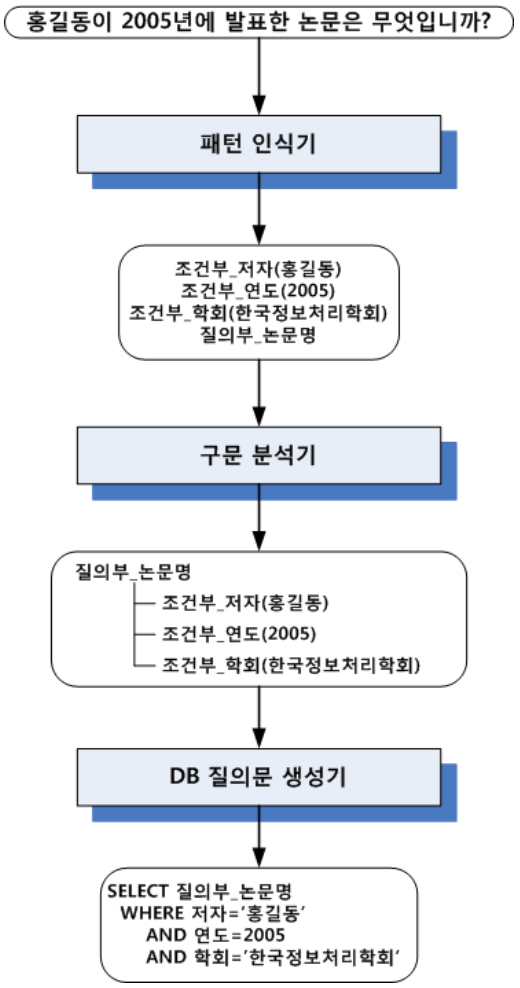
2) 구문 분석기

주어진 자연언어 질의가 4가지 분류의 패턴으로 바뀌므로 구문 분석 규칙이 매우 단순화된다. [4]에서 사용한 구문 분석 규칙은 의존 문법으로, 핵심 규칙은 '질의부'가 '조건부', '질의부', '연결부'를 지배하는 것이다. 이것을 기본으로 필요한 경우 패턴마다 예외 규칙을 적용하였다. 또한, 같은 조건부가 나열되는 경우를 위해, 같은 패턴의 조

건부가 나오면 지배할 수 있도록 하였는데, 접속 조사 등으로 연결되는 경우에 적용된다.

3) 질의문 생성기

구문 분석된 패턴을 분석하여 DB 질의문을 생성한다. 질의부 패턴이 있고, 질의부 패턴에 지배를 받는 의존소 패턴들을 찾아서 DB 질의문을 생성한다. 앞의 (예 1)에 대한 전체적 분석 과정이 (그림 2)에 나타나 있다. SQL query로 표현하였다. 편의상 FROM 절은 생략하도록 한다.



(그림 2) DB 질의문 생성 예

(그림 2)에 나타난 자연언어 질의문은 하나의 질의부 패턴을 가지고 있는 경우이다. 만약 질의부 패턴이 2개 이상 나오면, JOIN 연산을 사용하여 해결할 수 있다. 내포된 질의부에 대한 DB 질의문을 먼저 생성하고, 그 후에 바깥의 질의부에 대한 DB 질의문을 생성할 때, 가져올 테이블(FROM 절)을 내포된 질의부의 DB 질의로 설정하는 형식이다. 자세한 예는 [4]에 나타나있다.

의미 처리기는 의미부 패턴을 처리하는 부분이다. 적용되는 분야마다, 그리고 패턴마다 다르게 처리를 해 주어야 한다. 앞서 나왔던 'NOT'의 경우, 아래의 (예 3)을 보자.

(예 3) 홍길동이 발표하지 않은 논문은 무엇입니까?
(예 4) 홍길동이 2005년에 발표하지 않은 논문은 무엇입니까?

(예 3)의 경우, NOT은 조건절에 영향을 주게 되는데, SQL query로 표현하면 (Query 1)이 될 것이다. (예 4)와 같이, 조건부가 2개 이상인데 NOT이 나타난 경우는 어떻게 해야 할까? 대부분은 (Query 2)처럼, 가장 질의부에서 가까운 조건부_연도에만 NOT을 적용하는 것이 자연스러울 것이다. 하지만, 이는 절대적인 것은 아니며, 상황에 따라 다를 것이다.

(Query 1) SELECT 논문명 WHERE 저자 <> '홍길동'
(Query 2) SELECT 논문명 WHERE 저자 = '홍길동' AND 연도 <> 2005

또 다른 예로, DB에서 제공하는 COUNT() 함수를 이용하는 것이 있을 수 있다.

(예 5) 홍길동이 2005년에 발표한 논문의 개수는 얼마입니까?

(예 5)와 같은 질의문이 있다면, '개수'를 '의미부_개수' 패턴으로 인식한 후, (Query 3)과 같은 DB 질의문을 생성한 후에, 후처리로 (Query 4)와 같이 COUNT 함수를 적용하도록 한다.

(Query 3) SELECT 논문명 WHERE 저자='홍길동' AND 연도=2005
(Query 4) SELECT COUNT(논문명) WHERE 저자='홍길동' AND 연도=2005 GROUP BY 논문명

의미부에 대한 처리는 우선 어떤 '의미'를 처리를 할 건지 개발자가 설정해야 한다. 대상 '의미'를 정하고, 그것을 인식할 패턴을 만들고, 처리 함수를 만드는 것이다. 실제 사용하는 DB에서의 처리 가능 여부도 파악을 해야 한다. 좋은 의미 처리가 많을수록 사용자가 느끼는 편리함은 증가할 것이다.

3. 결론

본 논문에서는 자연언어 데이터베이스 인터페이스 시스템의 일반적인 틀(framework)을 소개하였다. DB 기반 시스템을 자연언어로 사용하는 것은 일반 사용자에게 매우 큰 편리함을 제공하며, 미래 사회에서 필수적인 기술이다. 자연언어 데이터베이스 인터페이스 시스템은 DB 기반 시스템의 전반부에 있는 전처리기로, 폭넓은 처리가 가능하면서도 빠른 속도로 처리할 수 있어야 한다.

단순히 패턴 매칭만 이용하면 확장성이 매우 떨어지는데, 이를 피하고자 구문 분석 기법을 도입했으며, 구문 분

석 규칙을 단순하게 하고자 패턴을 4가지로 분류하였으며, 의미부 패턴을 통해 다양한 의미 처리를 가능하게 하였다. 각 모듈의 단계를 명확히 하여, 시스템의 확장성과 유지보수성을 높였다. 패턴과 구문 분석 기법을 동시에 사용하여 처리 능력을 향상시키고 속도를 빠르게 하였다.

참고문헌

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

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

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

[4] 임경업, 이석형, 윤화묵, 권혁철, “학술자료 검색을 위한 자연언어 데이터베이스 인터페이스 시스템”, 제 30회 한국정보처리학회 추계학술발표대회, 2008년 11월

[5] 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

[6] 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

[7] AnyQuestion, http://anyq.etri.re.kr