특집 3

데이터베이스 검색을 위한 자연어 질의 인터페이스

채진석
서울대학교 컴퓨터공학과

1. 서론

인공 언어인 프로그래밍 언어에 대하여 인간이 일상 생활에서 사용하는 언어를 자연어라고 하는데, 인간은 자연어를 사용하여 인간 상호간의 의사 소통을 하고 정보를 보존하며 넓은 분야의 다양한 생각들을 표현하고 복잡하고 심오한 사상을 전달할 수 있다. 인간이 습득하고 사용하는 이러한 언어적 지식을 컴퓨터에 이식함으로써 인간이 언어를 사용하여 얻는 많은 이점을 컴퓨터에서도 얻으려는 연구 분야가 자연어 처리(Natural Language Processing) 분야이다. 이러한 자연어 처리 분야의 대표적인 응용 분야로는 기계 번역(Machine Translation), 정보 검색(Information Retrieval), 인간-컴퓨터 인터페이스(Human-Computer Interface) 등이 있다.

또한, 컴퓨터를 일상 생활에 이용하는 여러 분야 중에서 실생활에 유용한 자료를 관리해 주는 데이터베이스 관리 시스템(DBMS)은 날로 이용자가 많아지고 중요성도 높아지고 있다. 그러나 사용자가 데이터베이스에서 필요한 자료를 검색하려면 SQL이나 OQL 등과 같은 데이터베이스 질의어에 정통해야만 한다. 그러나 데이터베이스 질의어에 익숙하지 않은 사용자에게는 아무리 간단한 질의어라도 그것을 배우고 이해하여 적절하게 사용하는 데에는 많은 시간과 노력이 요구된다.

또한 데이터베이스 질의어는 일반적인 프로그래밍 언어인 C나 Pascal과는 달리 컴퓨터 자체에 대한 이해뿐만이 아니라 데이터베이스에 대한 이해가 있어야 배울 수 있기 때문에 초보들의 어려움을 가중시키고 있다.

따라서 기존의 데이터베이스 질의어가 가지고 있는 제약 조건을 대폭 완화시켜 초보자라도 쉽게 배워 사용할 수 있는 자연어 질의 인터페이스는 미래의 DBMS가 가져야 하는 필수적인 요소가 될 것으로 기대된다. 이러한 자연어 질의 인터페이스는 데이터베이스 질의어에 익숙하지 않은 사용자뿐만 아니라 데이터베이스 질의어를 공부하고자 하는 학습자를 위한 학습 도구로도 사용될 수 있을 것이다.

2. 데이터베이스 검색 도구와 자연어 질의 인터페이스

현재까지 일반 사용자들이 DBMS나 정보 검

색 시스템을 쉽게 접근할 수 있게 도와주도록 개발된 도구로는 다음과 같은 것들이 있다.

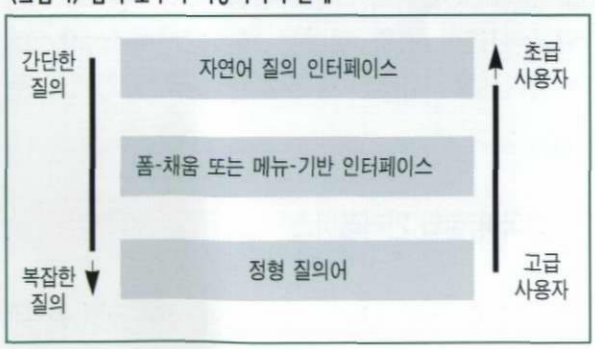
- 정형 질의어: SQL, QUEL, OQL 등
- 폼-채움 또는 메뉴-기반 인터페이스: CUPID
- 예를 통한 질의 작성: QBE
- 대화식 질의 명세: RENDEZVOUS, CO-OP 등
- 자연어 질의: TEAM, IRUS, INTELLECT, NHI, KID 등
- 자연어 질의 안내 기능을 통한 질의 작성: PRED, KDA 등

데이터베이스에 저장되어 있는 정보를 검색하는 가장 일반적인 방법은 정형 질의어를 사용하는 것인다. 이러한 정형 질의어를 사용하면 아무리 복잡한 질의도 표현할 수 있기 때문에 사용자가 원하는 정보를 가장 효과적으로 검색할 수 있다.

그러나 일반 사용자가 정형 질의어를 사용하기 위해서는 기본이 되는 데이터 모델과 특정 데이터베이스의 구성 및 개념에 대한 이해가 선행되어야 한다.

따라서 이러한 기본 지식이 없는 사용자들이 정형 질의어를 사용하는 데에는 많은 시간과 노력이 요구된다.

〈그림 1〉 검색 도구와 사용자와의 관계



이에 비해 폼-채움(Form-Filling)이나 메뉴-기반(Menu-Based) 기법들은 정형 질의어를 사용하는 것보다는 쉽게 데이터베이스를 접근할 수 있는 방법을 마련하여 주고 있지만 다음과 같은 경우에는 자연어를 사용하는 것이 보다 효율적이다.

- 정보의 구성이나 검색 절차가 복잡한 경우
- 검색 효율을 위하여 일상적인 개념과는 상이한 개념을 사용하여 데이터베이스를 구축한 경우
- 사용자의 요구가 복잡하여 정형 질의어를 사용하더라도 표현이 어려운 경우

이러한 검색 도구와 사용자와 관계는 〈그림 1〉과 같다. 자연어 질의 인터페이스는 자연어의 모호성과 자연어 분석의 어려움 때문에 간단한 질의를 주로 하는 초급 사용자에게 적합한 인터페이스라고 할 수 있고, 그보다는 복잡한 질의를 해야 하는 사용자에게는 폼-채움이나 메뉴-기반 인터페이스가 적합하며, 정교한 질의를 통한 복잡한 검색 연산이 필요한 고급 사용자는 정형 질의어를 사용해야 할 것이다.

자연어를 사용하여 데이터베이스에 저장되어 있는 정보를 검색함으로써 얻을 수 있는 이점은 다음과 같다.

- 데이터베이스의 실제 저장된 내용을 직접 언급할 수 있다.
- 데이터베이스의 저장 구조와는 상관없이 정보를 접근할 수 있는 방법을 제공한다.
- 적은 양의 학습만으로 데이터베이스를 접근할 수 있다.

자연어 질의 인터페이스를 사용하기 위해서는 질의 작성에 관한 기본적인 교육이 필요하게 되

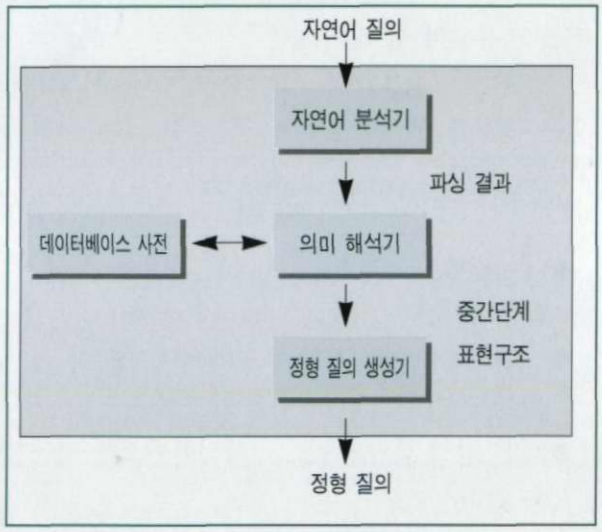
는데, 약 1시간 정도의 사전 교육만으로도 다양한 자연어 질의를 사용하여 원하는 정보를 검색할 수 있었다는 실험 결과도 있다.

그러나 자연어라는 것은 반드시 모호성(Ambiguity)을 내포하고 있으므로, 자연어를 사용하는 모든 시스템은 반드시 이 모호성 문제를 해결해야만 한다. 모호성이란 자연어의 일반적인 성질 중의 하나로 어느 정도는 문맥 정보에 의해서 해결될 수 있다.

또한 문장의 구문 구조도 모호성을 가질 수 있는데 이것을 자연어를 파싱하는데 어려움을 주며, 문장 속에 있는 지시 대명사나 부사구 등은 정량화된 논리를 사용하는 인공 언어에 비해 여러 가지 해석을 가능하게 한다. 자연어 질의 인터페이스에서는 이러한 일반적인 자연어 처리에서 나타나는 모호성뿐만 아니라 도메인 지정이 불가능함에 따른 모호성도 발생한다. 예를 들어 대학에서 사용되는 학사 관리 데이터베이스에 Q1과 같은 자연어 질의를 했다고 하자.

Q1: "홍길동이 소속된 학과를 보여라."

〈그림 2〉 자연어 질의 인터페이스의 시스템 구성



이러한 질의는 자연어 질의 인터페이스가 구축되어 있는 지식 베이스를 사용하여 "홍길동"의 도메인이 사람 이름이라는 것을 알았다고 해도 그것이 학생인지 교수인지 직원인지를 알 수 없으므로 올바른 질의를 만들지 못한다.

이러한 질의는 문맥 정보로도 해결할 수 없고 사용자에게 정확한 도메인의 명세를 요구하여 해결해야만 한다. 즉, 사용자는 Q1 대신 Q1′과 같이 질의해야만 올바른 결과를 얻을 수 있는 것이다.

Q1′: "홍길동 학생이 소속된 학과를 보여라."

따라서 이러한 모호성에 대한 처리 여부는 자연어 질의 인터페이스의 성능을 결정하는 매우 중요한 요소 중의 하나이다.

3. 자연어 질의 인터페이스의 시스템 구성

자연어 질의 인터페이스의 전형적인 시스템 구성은 〈그림 2〉와 같다.

사용자에 의해 입력된 자연어 질의는 자연어 분석기에 의해 형태소 분석과 구문 분석이 이루어지며 구문 분석 결과로 파싱 트리가 만들어진다. 파싱 결과는 의미 해석기에 의해 데이터베이스 사전에 있는 스키마에 관한 지식을 참조하여 데이터베이스 환경에서의 의미로 해석되며 이러한 결과는 중간 단계의 지식을 표현하는 구조로 나타내어 진다.

중간 단계 지식 표현 구조는 정형 질의 생성기를 통해 SQL이나 OQL 등의 정형 질의로 변환되고 이러한 정형 질의는 데이터베이스 시스템에 전달되어 사용자가 원하는 자료를 검색해 준다.

4. 한국어 질의 인터페이스

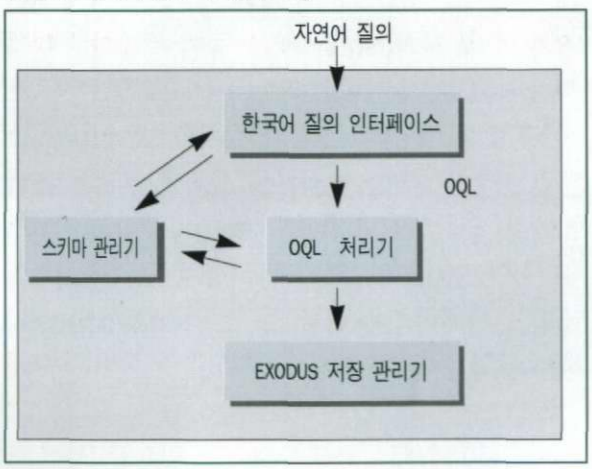
한국어 질의 인터페이스는 한국어 질의를 데

이터베이스 정형 질의어인 SQL이나 OQL 등으로 변환해 주는 시스템으로 아직 국내에서는 이에 대한 연구가 활발하지 못하며 대학의 연구실에서 개발한 프로토타입이 몇 개 있을 뿐이다. 여기서는 서울대학교에서 개발하고 있는 객체 지향 데이터베이스를 위한 한국어 질의 시스템에 대해 간략히 설명하고자 한다.

한국어 질의 시스템은 한국어 질의 인터페이스를 통해 입력된 한국어 질의를 ODMG에서 제안한 객체 지향 데이터베이스를 위한 질의어인 OQL로 변환해 주고, 이것을 OQL 처리기를 통해 최적화하며, 최적화된 질의를 저장 시스템인 EXODUS 저장 관리기에 전달하여 원하는 결과를 검색할 수 있게 해 주는 시스템이다.

이러한 한국어 질의 시스템의 시스템 구성을 〈그림 3〉에 보였다. 여기서 한국어 질의 인터페이스와 OQL 처리기는 스키마 관리기가 관리하는 스키마 정보를 이용하여 질의 처리를 수행한다.

〈그림 3〉 한국어 질의 시스템의 구성



입력된 한국어 질의는 형태소 분석과 구문 분석을 통해 파싱이 이루어져 파싱 결과인 논항 구조를 생성한다. 이러한 논항 구조는 질의 분해기에 의해 최소한의 의미를 가지는 단위인 질의구로 분해되고, 규칙 베이스를 참조하여 각 질의구별로 질의 프레임을 생성해 준다. 질의 프레임은 OQL의 표현력에 근접하기 위해 select 프레임, from 프레임, where 프레임, group by 프레임, order by 프레임으로 나누어 중간 단계의 지식을 표현하고 있다. 이렇게 표현된 질의 프레임은 OQL 생성기를 통해 OQL로 변환된다.

예를 들어 예제 질의 Q2는 다음과 같은 단계를 거쳐 OQL로 변환된다.

Q2: 키가 165보다 크고 "데이터베이스"를 수강하는 학생의 이름과 나이를 검색하라.

Q2를 자연어 분석한 결과로 만들어지는 파싱 결과인 논항 구조와 논항 구조를 질의구로 분해한 결과, 질의구로부터 생성된 질의 프레임, 질의 프레임을 통합하여 만들어진 OQL은 다음과 같다.

● 논항 구조

```
HEAD → VERB 검색하
  └ OBJECT → NOUN 나이
  └ NCONJUNCT → NOUN 이름
    └ MODIFIERS → NOUN 학생
      └ MODIFIERS → VERB 수강하
        └ MODIFIERS → NOUN "데이터베이스"
        └ VCONJUNCT → VERB 크
          └ SUBJECT → NOUN 키
          └ MODIFIERS → NOUN 165
```

● 질의구

```
QP1: HEAD → VERB 검색하
     └ OBJECT → NOUN 나이
       └ MODIFIERS → NOUN 학생
QP2: HEAD → VERB 검색하
```

```
    └ NCONJUNCT → NOUN 이름
      └ MODIFIERS → NOUN 학생
QP3: QP-HEAD → NOUN 학생
    └ MODIFIERS → VERB 수강하
      └ MODIFIERS → NOUN "데이터베이스"
QP4: QP-HEAD → NOUN 학생
    └ VCONJUNCT → VERB 크
      └ SUBJECT → NOUN 키
      └ MODIFIERS → NOUN 165
```

● 질의 프레임

```
Q-Frame1: (select_frame_list: S-Frame-list1, from_frame_list:
          F-Frame-list1, where_frame_list: W-Frame-list1)
S-Frame-list1: ( S-Frame11: (class: Student, attr: name),
                 S-Frame21: (class: Student, attr: age) )
F-Frame-list1: ( F-Frame11: (class: Student, variable: s),
                 F-Frame21: (class: Student, attr: enrolls,
                            variable: e ) )
W-Frame-list1: ( W-Frame11: (class: Student, attr: enrolls),
                 W-Frame12: (class: Enrollment, attr: course),
                 W-Frame13: (class: Course, attr: name, operator
                            : =, value: "데이터베이스"),
                 W-Frame21: (class: Student, attr: height,
                            operator: >, value: 165) )
```

● OQL

```
select s.name, s.age
from Student s, s.enrolls e
where e.course.name = "데이터베이스"
 and s.height > 165
```

5. 결론 및 향후 전망

이 글에서는 데이터베이스 검색을 도와주는 자연어 질의 인터페이스의 현황과 객체 지향 데이터베이스를 위한 한국어 질의 인터페이스의 처리 과정을 예를 통해 설명하였다. 현재까지 영어 등의 외국어를 위한 자연어 질의 인터페이스 분야는 비교적 활발하게 연구되었지만 한국어를 위한 자연어 질의 인터페이스에 대한 연구는 부진하였다.

그것의 가장 큰 이유는 한국어가 근본적으로 많은 모호성을 가지고 있어 성능이 우수한 자연어 분석기를 만들기 어렵기 때문이고, 또 다른 이유는 자연어 질의 인터페이스를 개발한다 해도 그것이 실용적인지 여부를 평가하기가 쉽지 않기 때문이다.

따라서 자연어 질의 인터페이스의 성능을 향상시키기 위해서는 자연어 질의 인터페이스 전용의 자연어 분석기가 개발되어야 하며, 그것의 성능을 평가할 수 있는 기법에 대한 연구가 필요하다.

현재까지는 비교적 간단한 자연어 질의에 대한 연구가 이루어졌지만 추후 존재 정량자나 전체 정량자를 포함하는 질의와 메소드를 포함하는 질의, 중첩 질의(Subquery) 등의 복잡하고 다양한 질의를 처리하는 연구가 수행되어야 한다. 또한 관계 데이터베이스나 객체 지향 데이터베이스뿐만 아니라, 정형 질의어를 사용해 정보를 검색하기가 용이하지 않은 멀티미디어 데이터베이스나 공간(Spatial) 데이터베이스를 위한 자연어 질의 인터페이스에 대한 연구도 필요하리라 생각된다. DC