관계형 데이터베이스의 자연어 인터페이스를 위한 확장된 데이터베이스 시멘틱 모델

정혜경*,배우정*,안동언**,이용석*

* 전북대학교 컴퓨터과학과
** 전북대학교 컴퓨터공학과

Extended Database Semantic Model for Natural Language Interface to Relational Database

H. K. Jeong*, W. J. Bae*, D. U. An**, and Y. S. Lee*

* Dept. of Computer Science, Chonbuk National University
** Dept. of Computer Engineering, Chonbuk National University

요 약

데이터베이스 사용자는 데이터베이스내에서 데이터를 검색하는 메카니즘과 원하는 데이터를 검색하기 위한 구체적인 질의 형태, 데이터베이스의 설계 과정에서 고려된 많은 묵시적인 의미 정보들을 인식하고 있어야 한다. 만일, 이들에 대한 정확한 인식이 이루어지지 않은채 요구된 질의는 잘못된 결과를 생성하게 된다.

데이터베이스에 대한 자연 언어 인터페이스는 이러한 세부 지식을 가지고 있지 않는 사용자에게 용이한 질의 환경을 제공해준다. 이를 위해 여러 자연 언어 인터페이스 시스템들이 개발되었다. 그러나 이 시스템들은 데이터베이스가 가지는 의미적 표현에 대한 구조적 제약성을 해소하지 못하였기 때문에 이 제약들이 사용자에게 그대로 남겨지고 있다는 문제점이 있다. 이러한 문제점은 근본적으로 자연언어와 데이터베이스의 시멘틱 모델간의 의미의 표현 레벨의 차이로 기인한다고 볼 수 있다.

본 논문은 이런 불일치 문제의 해결 방안으로 관계 데이터베이스내의 중요한 특성들을 구분하고, 이것을 표현할 수 있는 향상된 데이터베이스 시멘틱 모델에 대해 설명한다.

1. 서론

데이터베이스에 대한 자연 언어 인터페이스는 데이터베이스 시스템과 형식 질의(Formal Language)에 대한 지식을 가지고 있지 않는 사용자에게 사용하기 편리한 질의 환경을 제공해준다.

일반 사용자는 데이터베이스내에서 데이터를 검색하는 메카니즘과 원하는 데이터를 검색하기 위한 구체적인 질의 형태를 파악하고 있지 못하다. 또한, 데이터베이스내에는 데이터베이스의 설계 과정에서 고려된 많은 의미 정보들이 묵시적으로 내포되어 있어, 이들에 대한 정확한 인식이 이루어지지 않은채 요구된 질의는 잘못된 결과를 생성하게 된다[3].

이러한 문제를 해결하기 위해 Irus, Intellect, Ladder, NaturalLink 등의 자연 언어 인터페이스 시스템들이 개발되었다. 특히, 국내에서는 입력 질의의 유형에 따라 여러 패턴을 설정하여 이 패턴에 적절하게 대응하여 질의어로 변환하는 K-NLQ[11]와 키워드를 추출하는 방법으로 질의를 처리하는 방법[10] 등이 있다.

그러나 이 들은 데이터베이스가 가지는 의미적 표현에 대한 구조적 제약성을 해소하지 못하고, 이 제약들이 자연언어 인터페이스 사용자에게 그대로 남겨지고 있다는 문제점이 있다. 이러한 문제점은 근본적으로 자연언어와 데이터베이스의 시멘틱 모델간의 의미의 표현 레벨의 차이로 기인한다고 볼 수 있다.

본 논문은 이런 불일치 문제의 해결 방안으로 관계 데이터베이스내의 중요한 특성들을 표현할 수 있는 향상된 데이터베이스 시멘틱 모델에 대해 설명한다.

본 논문의 구성으로 2장에서는 데이터베이스의 주요 특성들을 기술하고, 이러한 특성을 반영하기 위한 향상된 데이터베이스 시멘틱 모델에 대해 기술한다. 3장에서는 이 모델을 기반으로 한 간단한 활용 예를 보이고, 마지막으로

결론 및 향후 이루어져야 할 연구과제에 대해 설명한다.

2. 향상된 데이터베이스 시멘틱 모델

관계형 데이터베이스는 엔티티와 이들간의 관계성을 릴레이션이라는 논리적인 구조를 통해 실세계의 복잡한 특성을 표현하고 있다. 이 장에서는 릴레이션과 이들에 대한 여러 시멘틱들을 보다 잘 표현하기 위한 구조들을 살펴보고, 자연 언어의 시멘틱을 위해 설계된 개념그래프가 데이터베이스의 시멘틱을 충분히 반영할 수 있도록 이에 필요한 지식에 대해서 고찰한다.

2.1 릴레이션

실세계는 엔티티 집합과 엔티티들간의 관계성으로 이루어진다. 엔티티는 여러 속성을 통해 세부적인 특성을 표현하며, 관계성은 특정한 상태에서 관련 엔티티들간의 사건을 표현한다. 관계형 데이터베이스에서는 엔티티 집합과 관계성을 모두 하나의 논리 구조인 릴레이션을 통해 표현하고 있다. 그러나, 이러한 릴레이션 구조에서 자연 언어 인터페이스에서 필요한 정보를 추출하기에는 적합하지 않다[1].

이러한 문제를 해결하기 위해 릴레이션의 표현으로 개념 그래프를 사용하고자 한다. 개념 그래프는 존재 그래프와 어의 네트워크에 기반하여 제안되었다. 개념 그래프는 논리적으로 간결하고 읽기 쉽고, 전산학적으로 다루기 쉬운 형태로 의미를 표현할 수 있다. 개념 그래프는 개념을 나타내는 개념 노드와 개념과 개념들사이의 개념적 관계를 나타내는 관계 노드를 아크로 연결하여 문자의 의미를 표현한다.

다음 데이터베이스를 고려하여 보자.

고객(고객, 고객도시)
지점(지점, 지점도시, 자산)
대출(지점, 대출계정, 대출받는자, 금액)
예금(지점, 예금계정, 예금자, 잔고)

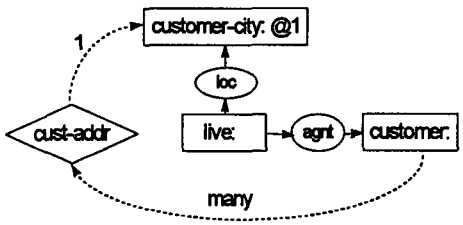
고객리스트 릴레이션의 시멘틱을 한국어 문장으로 나타내면 "모든 고객은 반드시 한 도시에서 살고있다."와 같다. 이 문장을 개념 그래프로 매핑시키면 다음과 같다.

```
[live] -
      (loc) → [customer-city:@1]
      (agnt) → [customer:∀]
```

여기서, 관계노드인 loc가 나타내는 것은 customer가 살고 있는 위치가 customer-city라는 것을 의미한다. 개념노드 [customer-city:@1]에서 기호 @1는 "반드시 한 개"에 해당하는 한정사이고, [customer:∀]에서 한정사 ∀는 이 범위내에 있는 모든 고객을 가리킨다.

그러나, 위의 개념 그래프에서 개념과 개념간의 관계성에 대해 아주 잘 기술하고 있지만 고객이 살고 있는 도시가 어디인가에 대해서는 나타나있지 않는다. 이를 보여주기 위해 릴레이션을 접근할 함수로 액티 노드를 사용하고, 이것과 개념간의 링크 관계를 표현한다.

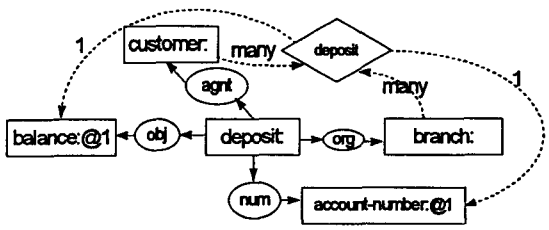
다음 그림은 고객의 도시와 고객이 어떻게 관련되어 있는가를 기술하며, 액티노드는 각 고객에 대한 도시를 저장하고 있는 릴레이션을 접근하는 함수이다.



(그림 1) customer에 대한 개념 스키마타

그림에서 @1과 ∀는 점선에 있는 단어 many, 1과 같은 정보를 의미한다. 점선의 화살표 방향을 통해 키 애트리뷰트는 customer가 된다.

다음은 엔티티 집합을 나타내는 릴레이션인 "예금" 릴레이션에 대해 고려하여 보자.



(그림 2) deposit에 대한 개념 스키마타

그림과 같이 "예금" 릴레이션을 나타내는 개념 그래프의 관계노드로 agnt, org, obj, num를 설정하였다.

이러한 변환 과정을 통해 데이터베이스와 자연언어간의 시멘틱 차이(sematic gap)을 줄일수 있다.

2.2 함수적 종속(Functional dependencies)

액터는 데이터베이스에서 개념의 참조값을 어떻게 찾을 것인가를 나타낸다. 이것은 어떤 속성이 키이고 독립변

수이며 또 어떤속성이 이 키에 대하여 함수적으로 종속적인지를 보여준다. 예를 들어, customer 엑터는 customer가 주키이고 나머지 city와 street는 이 키에 종속적이다. 따라서, customer와 city, stree간의 함수적 종속성은 "customer → city street"와 같이 표현된다. 또한, 한정사는 함수가 일대일, 다대일, 혹은 n대m 관계인지를 알려준다.

형식) (<엑터이름> (주요키) (종속키리스트))
예) (setq FD '(
(<cust-addrs> (customer)(city))
(<depositor> (customer branch)
(Account-Number Balance))))

2.3 형 계층구조 (Type hierarchy)

데이터베이스에서 엔티티의 형들은 일반적으로 생각할때 개념적인 단계를 이루로 있다. 예를들어, depositor (예금주), customer(고객), human(사람), animal(동물), alive-thing(생물체), entity(엔티티)와 같은 개념적인 계층구조를 가지고 있다.

```
(def-cnode entity)
(def-cnode alive-thing
 (INHERIT-FORM entity))
(def-cnode animal
 (INHERIT-FORM alive-thing))
```

2.4 도메인 역할(Domain roles)

두 개의 엔티티 형이 함수적으로 종속적인 관계를 나타내 주는 것 뿐만 아니라 개념 그래프는 이 종속성이 무슨 종속인지 종속성의 역할까지도 표현한다.

2.5 정의(Definition)

집성화와 개념 형, 관계 형, 엑터는 각각 또다른 개념 형, 관계 형, 엑터에 의해서 정의될 수 있다. 예를 들어, 개념 형 "customer"의 정의는 종차로 표현하는데 종은 보다 일반적인 개념 형인 "human"에서 유도되고, 특성에 해당하는 차는 "은행에서 거래하는 사람"이라고 정의할 수 있다.

2.6 스키마타(Schemata)

사람과 비슷하게 추론할 수 있도록 추론에 필요한 배경 지식을 표현하기 위한 기본구조를 스키마타라 부른다. 이 스키마타는 개념 그래프 종류중 세 번째 단계에 해당된다. 먼저, 첫째 단계인 임의 그래프는 어떤 제약조건도 부여되지 않고 만들어지는 개념그래프이다. 둘째 단계인 정규그래프는 선택적인 제약을 부여하여 개념그래프를 형성한다. 세 번째 단계인 스키마타는 실세계에 존재하는 사건과 엔티티와 속성에 대한 전형적인 모임들이 특정 도메인에 알맞는 지식으로 구체화되어 표현된다. 각 개념형에 대해서, 스키마타는 또 다른 개념형에 대하여 일반적으로 가질 수 있는 역할을 한다.

예를 들어, 스키마 Customer는 고객 이름 name, 고객의 주소 city, street 등을 갖는 배경 정보를 포함하고, 형 정의 Customer는 주로 정의하는 특징을 표현한다.

3. 활용 예

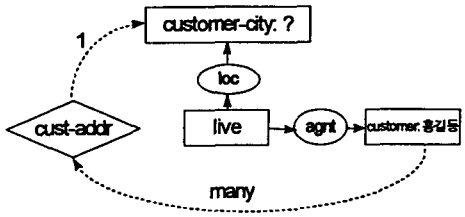
한국어 질의를 입력 받아 이 질의 결과를 얻기 위해 2장에서 설명한 데이터베이스 시멘틱에 관한 여러 배경 정보들이 요구된다. 이 장에서는 자연어에 대한 질의 인터페이싱 과정에서, 배경 정보를 자동적으로 결정하는 데이타베이스 추론과정에 관한 부분만을 예을 통해 보인다.

자료형과 암시적인 죠인을 찾기 위해 질의와 데이타베이스에 관한 배경 정보를 합친 추론 방법이 필요하다. 이 정보는 자연어 분석기, 추론기, 데이터베이스 시스템과 같은 질의 처리동안 각 단계마다 추가된다.

예를들어 "덕진지점에 예금한 홍길동은 어디에서 살고 있는가?" 질의를 생각해 보자. 먼저 질의 분석 결과 다음과 같은 개념 그래프가 생성된다.

```
[[live] -
    (loc) → [customer-city:?]
    (agnt) → [human:홍길동]] -
          (rec) → [[deposit] -
                (agnt) → [human:홍길동]    (1)
                (org) →[branch:덕진]]
```

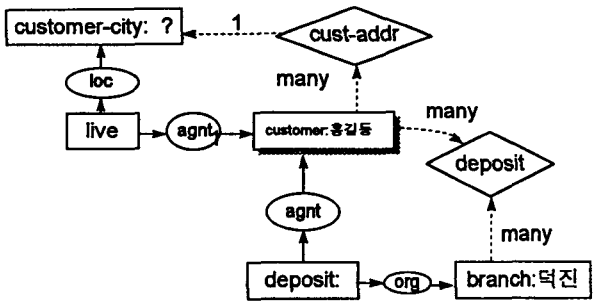
이 생성된 그래프에서 [city:?]의 참조값을 찾아야 한다. city에 대한 물음표가 어떤곳으로도 파급되지 않기 때문에 (1)과 죠인할 다른 스키마타를 찾아야 한다. [city:?]의 물음표는 cust-addr 스키마와 죠인된다. 죠인하기 전에 customer::홍길동에 대한 일치관계를 먼저 검사해야 한다. 만약 일치하면 이름 홍길동으로 대치할 수 있다.



(그림 3) 1 단계 작업 그래프

(1)의 결과에서 deposit 스키마를 필요로 하므로, 1단계

결과와 이 스키마가 죠인하면 2단계가 유도된다.



(그림 4) 2 단계 마지막으로 유도된 작업 그래프

deposit 스키마와 죠인된 후 customer-city의 참조값을 구할 수 있다. 이렇게 질의 결과값을 찾은 후 그림 4의 그래프는 더 이상 필요하지 않는다. 그러나, 기타 예금한 다른 사람들의 도시를 알고 싶으면 이 그래프를 새로운 엑터 <deposit-addr>로 정의하여 다시 이용할 수 있다.

4. 결론 및 향후 연구과제

데이터베이스에 대한 자연 언어 인터페이스는 데이터베이스 시스템과 형식 질의에 대한 지식을 가지고 있지 않는 사용자에게 사용하기 용이한 질의 환경을 제공해준다.

본 논문에서는 이러한 자연 언어 인터페이스를 위해 관계 데이터베이스내의 중요한 특성들을 분류하고, 이들을 표현할 수 있는 향상된 데이터베이스 시멘틱 모델을 제안하였다. 그리고, 데이터베이스 시멘틱 모델내의 여러 배경 정보들이 실제 질의 인터페이싱 과정에서 활용되는 데이타베이스 추론 과정에 관해 설명하였다.

이후 이루어져야 할 과제로써, 1) 제안된 모델에 대한 completeness와 soundness등의 검증과 이 모델을 기반으로 한 2) 데이터베이스 형식 질의로의 변환 메카니즘에 대한 연구가 추가로 필요하다.

참고문헌

[1] John F. Sowa, Conceptual Structure: Information Processing in Mind and Machine, Addison-wesley, 1992.

[2] Motoyuki ITOH, Oa HORII, Yukihiro ITOH, "Natural Language Interface to Relational Database," PRICAI, vol. 2, pp. 745-751, 1994.

[3] Henry F. Korth and Abraham Silberschatz, Database System Concepts, 2nd Edition, McGraw-Hill, 1993.

[4] Mohd Noor Mid Sap and D. R. McGregor, "Natural Language Interface to Database: State of the Art," Research Report of the Department of Computer Science, Univ. of Strathclye, 1993.

[5] Jinseok Chae, Sukho Lee, "Identifying Basic Patterns of Korean Natural Language Query," Proc. of NLPRS, pp. 606-611, Vol. 2, 1995.

[6] I. C. Park, W. J. Bae, Y. S. Lee, "A Study on Generation of Conceptual Graphs using Sentence Patterns in Korean," Proc. of NLPRS, pp. 685-690, Vol. 2, 1995.

[7] M. Bates, M. Moser, and D. Stallard, The IRUS Transportable Natural Language Database Interface, Expert Database Systems, Benjamin/Cummings Co. Ltd., pp. 617-630, 1987.

[8] G. Hendrix, E. Sacerdoti, E. Sagalowicz, and J. Slocum, "Developing a Natural Language Interface to Cmplex Data," ACM TODS, pp. 105-147, 1978.

[9] M. Templeton and J. Burger, "Problems in Natural Language Interface to DBMS with Examples from EUFID," Conf. on Applied Natural Language Processing, ACL and NRL, pp. 3-16, 1983.

[10] 김재문, 김재홍, 이상조, "키워드를 이용한 자연 언어 질의어 시스템 설계 및 구현", 한국정보과학회 논문지, 제 22권 1호, 1995.

[11] 채진석, 김성기, 이석호, "한국어 데이터베이스 질의 시스템의 설계 및 구현", 한국정보과학과 논문제, 제 20권 6호, 1993.