데이터 추상화와 퍼지 관계를 이용한 근사적 질의응답에 관한 연구

허순영, 문개현*

한국과학기술원 테크노경영대학원
130-012 서울 동대문구 청량리동 207-43

요약

본 논문은 데이터베이스에 존재하는 데이터 값들 사이의 유사성에 관한 지식을 이용하여 사용자가 요구한 정확한 답뿐 아니라 그와 유사한 답까지 제공해 줄 수 있는 근사적 질의처리 기법을 제시한다. 이를 위하여, 계량적인 방법에 해당하는 퍼지 관계와 비계량적인 방법에 해당하는 데이터 추상화를 하나로 통합한 유사성 표현 프레임웍을 제시하고 그를 이용한 지식 베이스를 설계한다.

1. 서론

데이터베이스로부터 다양한 정보를 획득하기 위한 표준 도구로서 질의어가 널리 사용되어 왔다. 하지만 질의의 조건에 정확히 일치하는 답이 존재하지 않을 경우 아무런 정보도 제시해 주지 못함에 따라 만족할 만한 결과를 얻을 수 없게 된다. 이러한 경우에 만일 데이터베이스 시스템이 근사해를 제공해 줄 수 있다면 질의의 효과성을 더욱 높일 수 있을 것이다.

근사적 질의응답은 이처럼 정확하게 작성되지 못한 질의문의 의도를 분석하여 질의를 처리하고, 사용자가 원하는 답이 존재하지 않을 경우 질의의 조건을 완화하여 원래 질의에 대한 근사해를 제공해 주는 지능적 시스템의 개발에 초점을 맞춘다[7, 8, 9, 10, 15, 17, 18, 24].

이를 위하여 다양한 질의처리 기법이 제시되어 왔지만 기존의 기법들은 처리 가능한 질의 유형이 제한되어 있으며, 질의완화 과정에서 사용자와의 interaction이 부족하다는 단점을 가지고 있다.

근사적 질의처리에 필요한 데이터 값들간의 유사성을 표현하기 위하여 크게 계량적(metric) 기법을 이용한 방법과 비계량적 기법을 이용한 방법이 사용되어 왔다. 본 논문은 두 가지 기법의 장점을 모두 흡수하여 하나의 통합된 질의처리를 수행하기 위한 기법을 제시함을 목적으로 한다. 비계량적 기법을 수용하기 위해서 데이터 추상화를 바탕으로 한 추상화 계층(abstraction hierarchy)을 이용하였고, 계량적 기법을 수용하기 위해서 퍼지 관계를 이용하였다. 표현된 데이터에 관한 의미론적 지식은 관계형 데이터베이스를 이용해 구축된 지식베이스에 저장되어 사용자와의 상호작용을 통해 질의처리에 이용된다. 구축된 지식베이스는 기존의 기법들보다 지식의 추가, 삭제, 갱신 등의 지식 관리 측면에서 많은 장점을 가질 뿐 아니라 기존의 근사적 질의처리 방법보다 다양한 유형의 질의처리를 수행할 수

있게 되고 그로 인하여 질의처리와 의사결정의 효율성을 높일 수 있게 된다.

2. 데이터 추상화를 이용한 지식 추상화 계층

지식 추상화 계층(KAH)은 하나의 지식 표현 프레임웍으로서, 데이터 추상화를 통하여 데이터 값과 데이터베이스에 대한 의미론적 지식을 다단계 구조로 표현할 수 있게 해 준다. 데이터 추상화에서 하나의 데이터 값은 추상값으로 일반화될 수 있고, 일반화된 추상값은 세분값으로 세분화될 수 있다. KAH는 추상값과 세분값 사이의 이러한 일반화/세분화 관계를 바탕으로 구축되어 진다. 그림 1에는 대학 전공에 관한 KAH가 나타나 있다.

KAH는 값의 추상화 계층과 도메인 추상화 계층의 두 가지 추상화 계층으로 구성된다. 그림 1은 대학 전공들을 전공 이름(MAJOR_NAME), 전공 분야(MAJOR_AREA), 전공 그룹(MAJOR_GROUP)등의 다단계 추상화 수준으로 분류한다. 값의 추상화 계층에서 Finance, Accounting, Marketing등은 전공 이름(MAJOR_NAME)이다. Management는 세가지 전공을 포괄하는 전공 영역(MAJOR_AREA)의 한 값이며, 그들의 추상값으로서 더 높은 수준에 존재한다. 마찬가지로, Economics는 Macro Economics, Micro Economics, Econometrics를 포괄하는 또 다른 전공 영역의 값으로 그들의 추상값에 해당한다. 추상값 Management와 Economics는 다시 더욱 추상화 되어 2단계 높은 수준에 존재하는 값 Business와도 추상화 관계를 갖는다. 이처럼 한 계층내에서의 상위 수준은 하위 수준보다 좀 더 추상화 된 데이터 표현을 가능하게 함으로써, 상위 수준에 존재하는 값은 하위 수준에 존재하는 값의 추상값으로 이해된다. 따라서 최상위 수준의 값은 계층 내에서 가장 추상화된 값인 반면, 최하위 수준의 값들은 가장 세분화된 값들이다. 전체 추상화 수준의 깊이는 사용자가 문제를 적절하게 묘사하기 위해서 어느 정도의 추상화가 필요한지를 판단하여 자유롭게 결정할 수 있다.

도메인 추상화 계층은 값의 추상화 계층내의 모든 개별 값들이 속해있는 도메인들로 구성된다. 도메인 추상화 계층에 존재하는 하나의 도메인은 유일한 이름과 계층 내에서의 절대 위치를 가지고 있다. 또한 도메인 이름은 값의 추상화 계층에 존재하는 값들의 추상화 수준을 결정할 수 있게 해준다.

3. 퍼지 관계를 반영한 지식 추상화 계층

데이터 추상화 기법을 기반으로 하여 개발된 지식 추상화 계층에 계량적 기법을 추가하기 위하여 퍼지 관계를 사용하여 KAH를 FKAH로 확장하

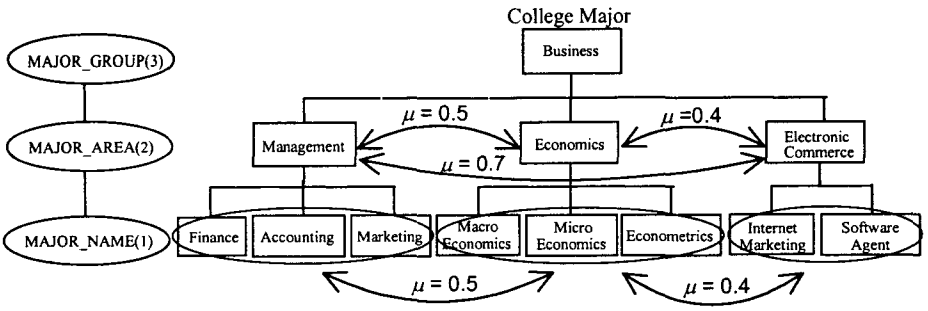

그림 1. 지식 추상화 계층

게 된다. 일반적으로 관계란 두 집합의 카티션 프로덕트의 부분집합으로 정의할 수 있다. 그림 1에는 다음과 같은 값들 사이의 유사성 관계가 존재한다.

R={(Finance, Accounting), (Finance, Marketing), (Accounting, Marketing), (Micro Economics, Macro Economics), (Micro Economics, Econometrics), (Macro Economics, Econometrics), (Internet Marketing, Software Agent), (Management, Economics), (Management, Electronic Commerce), (Economics, Electronic Commerce)}.

여기서 각각의 순서쌍들의 관계 R에 대한 소속 함수 μ_R: A×B → [0,1] 를 정의함으로써 퍼지 관계 R을 정의할 수 있다. 이 때, 소속 함수는 순서쌍을 이루는 두 원소 사이의 유사성의 정도로 해석할 수 있다.

퍼지 관계를 이용하여 KAH에 존재하는 값들 사이의 유사성의 정도를 정의할 수 있다.

- v_1, v_2, v_3이 동일한 추상값을 갖더라도 $\mu_R(v_1, v_2)$과 $\mu_R(v_1, v_3)$ 이 다를 수 있다.
- v_1, v_2가 동일한 1단계 추상값을 갖고 v_1, v_3 이 다른 1단계 추상값을 갖으면 $\mu_R(v_1, v_2) > \mu_R(v_1, v_3)$이고 그 역도 참이다.
- $\mu_R(v_1, v_2)$ 은 v_1과 v_2의 세분값 집합 사이의 유사성 정도를 의미하기도 한다.

첫번째 가정은 동일한 추상값을 갖는 값들이더라도 유사성의 정도가 다를 수 있음을 의미한다. 두 번째 가정은 동일한 추상값을 갖는 값들은 그렇지 않은 값들 보다 유사성의 정도가 큼을 의미한다.

이 같은 가정은 동일한 추상값을 갖는 값들 사이에 대해서만 유사성의 정도를 부여함으로써 만족 시킬 수 있다. 이때 가정 3에 의하여 유사성의 정도가 정의되지 않은 값들 사이의 유사성도 계산해 낼 수 있게 된다.

4. 퍼지 관계를 반영한 지식 추상화 계층

인사 관리자가 특정 직무를 담당하기 위한 Finance 전공자를 찾는다고 할 때, Finance를 전공한 직원이 존재하지 않거나 혹은 충분한 인원이 존재하지 않는 경우, 탐색 영역을 확장하여 관련된 전공을 갖는 다른 직원들을 찾을 수 있을 것이다. 한편, KAH에서 한 값의 근사값들을 찾는 것은 그 값의 추상값을 찾아냄으로써 가능하게 된다. 왜냐하면 그 추상값의 세분값들은 주어진 값과 서로 유사한 값들이기 때문이다. 따라서, 근사적 선택 조건이 주어졌을 때, 질의완화의 핵심적인 과정은 다음과 같게 된다.

Approximate Selection

Step 1. Original Query

```
select   e.emp_name, e.dept
from     employee e, college_major c
where    c.major =(1:6) "Finance" and e.id = c.id
```

Step 2. Query Generalization

```
select   e.emp_name, e.dept
from     employee e, college_major c
where    c.major is-a Generalize("Finance", 1:6) and e.id = c.id

select   e.emp_name, e.dept
from     employee e, college_major c
where    c.major is-a "Management" or c.major is-a "Electronic Commerce"
         and e.id = c.id
```

Step 3. Query Specialization

```
select   e.emp_name, e.dept
from     employee e, college_major c
where    c.major in Specialize("Management") or
         c.major in Specialize("Electronic Commerce") and e.id = c.id

select   e. emp_name, e.dept
from     employee e, college_major c
where    c.major in ("Finance", "Accounting", "Marketing", "Internet
         Marketing", "Software Agent") and e.id = c.id
```

그림 2. 질의 완화 과정

질의문에서, c.major=?(1:6)"Finance"의 의미는 major 애트리뷰트 값이 "Finance" 이거나 그와 유사한 근사값인 투플들을 찾는 것이다. 즉 1단계 추상화를 통해 얻을 수 있는 값들과, 1단계 추상값과 0.6 이상의 퍼지 관계를 갖는 값들까지도 찾아내는 것이다.

상위 수준에 존재하는 추상값은 하위 수준에 존재하는 다수의 세분값들에 대응하므로, 1단계 추상값을 찾는 것은 선택조건에 명시된 목표값의 근사값들을 생성할 수 있게 한다. 따라서, 1단계 일반

화된 질의를 생성해 낸다.

is-a는 세분값과 추상값 사이의 일반화 관계를 나타낸다. 질의분석을 통하여 Finance의 도메인이 MAJOR_NAME임을 알고 있으므로, 일반화된 질의문의 Generalize("Finance", 1:6)는 추상값 Management와 Electronic Commerce를 반환한다.

추상값 Management의 1단계 세분화는 Finance와 유사한 값들을 포함하는 세분값 집합 {Finance, Accounting, Marketing, internet Marketing, Software Agent}을 반환한다. 따라서, 최종적으로 세분화된 질의문이 작성되고 정규적인 SQL 질의문을 통하여 처리될 수 있다.

4. 결론

본 논문에서는 질의조건을 완화하고 정확한 답뿐 아니라 유사한 답까지 제시해 줄 수 있는 근사적 질의처리 기법을 제시하였다. 이를 위하여 데이터 추상화와 퍼지 관계를 이용하여 데이터 값들 사이의 유사성을 표현하기 위한 FKAH를 개발하였다. 제시된 FKAH는 기존의 근사적 질의처리 방법보다 좀더 다양한 질의를 처리할 수 있을 뿐 아니라 사용자와의 상호작용도 높임으로써 보다 효과적인 질의처리를 수행할 수 있게 된다.

참고문헌

[1] Chu, W., Yang, H., and Chow, G. (1996) A Cooperative Database System (CoBase) for Query Relaxation. Proc. of the Third International Conference on Artificial Intelligence Planning Systems

[2] Chu, W. and Chen, Q. (1994) A Structured Approach for Cooperative Query Answering. IEEE Transactions on Knowledge and Data Engineering, 6(5), 738-749

[3] Cuppens, F. and Demolombe, R. (1989) Cooperative Answering: A Methodologies to Provide Intelligent Access to Databases. Proc. of 2nd International Conference on Expert Database Systems, 621-643

[4] Godfrey, P., Minker, J., and Novik, L. (1994) An Architecture for a Cooperative Database System. Proc. of the 1994 International Conference on Applications of Databases

[5] Hemerly, A., Casanova, M., and Furtado, A. (1994) Exploiting User Models to Avoid Misconstruals. Nonstandard Queries and Nonstandard Answers, Oxford Science Publications

[6] Minock, M. J. and Chu, W. (1996) Explanation for Cooperative Information Systems. Proc. of Ninth International Symposium on Methodologies for Intelligent Systems

[7] Motro, A. (1990) FLEX: A Tolerent and Cooperative User Interface to Databases. IEEE Transactions on Knowledge and Data Engineering, 2(2), 231-246

[8] Motro, A. (1990) Accommodating Inprecision in Database Systems: Issues and Solutions. Data Engineering, 13(4) 29-34

[9] Shenoi, S. and Melton, A. (1992) Functional Dependencies and Normal Forms in the Fuzzy Relational Data Model. Information Sciences, 1-28

[10] Takahashi, Y. (1993) Fuzzy Database Query Languages and Their Relational Completeness Theorem. IEEE Trans. Knowledge and Data Engineering, 5(1), 122-125

[11] Umano, M. and Ezawa, Y. (1991) Implementation of SQL-type Data Manipulation Language for Fuzzy Relational Databases. Proc. of IFSA Brussels.

[12] Vrbsky, S. V. and Liu, W. S. (1993) APPROXIMATE-A Query Processor that Produces Monotonically Improving Approximate Answers. IEEE Transactions on Knowledge and Data Engineering, 5(6)

[13] Wong, M. H. and Leung, K. S. (1990) A Fuzzy Database Query Language. Information Systems, 15(5), 583-590