

추상화 정보 데이터베이스 기반 협력적 질의 응답

허순영* · 이정환**

Cooperative Query Answering Based on Abstraction Database

Soon-Young Huh* · Jung-Whan Lee**

■ Abstract ■

Since query language is used as a handy tool to obtain information from a database, a more intelligent query answering system is needed to provide user-friendly and fault-tolerant human-machine interface. Frequently, database users prefer less rigid querying structure, one which allows for vagueness in composing queries, and want the system to understand the intent behind a query. When there is no matching data available, users would rather receive approximate answers than a null information response. This paper presents a knowledge abstraction database that facilitates the development of such a fault-tolerant and intelligent database system. The proposed knowledge abstraction database adopts a multilevel knowledge representation scheme called the knowledge abstraction hierarchy(KAH), extracts semantic data relationships from the underlying database, and provides query transformation mechanisms using query generalization and specialization steps. In cooperation with the underlying database, the knowledge abstraction database accepts vague queries and allows users to pose approximate queries as well as conceptually abstract queries. Specifically, four types of vague queries are discussed, including approximate selection, approximate join, conceptual selection, and conceptual join. A prototype system has been implemented at KAIST and is being tested with a personnel database system to demonstrate the usefulness and practicality of the knowledge abstraction database in ordinary database application systems.

1. 서론

데이터베이스로부터 다양한 정보를 획득하기 위

한 표준 도구로서 질의어가 널리 사용되어 왔다. 하지만 정확한 질의문을 작성하기 위해서는 질의 명세 문법은 물론이고 데이터베이스 스키마에 관

* 한국과학기술원 테크노경영대학원

** SK Telecom

한 지식을 필요로 하기 때문에 전문가를 비롯한 데이터베이스 사용자들은 질의 응답에 커다란 불편을 겪어 왔다. 또한 질의의 조건에 정확히 일치하는 답이 존재하지 않을 경우 아무런 정보도 제시해주지 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. 또한 데이터베이스 시스템이 모호하게 작성된 질의를 직접 처리할 수 있을 정도의 지능을 갖추게 된다면, 정보 검색시의 오류 처리와 의사결정 분석의 효과성이 크게 증대될 것이다. 협력적 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].

협력적 질의응답에서 지식 베이스는 기반 데이터베이스내의 원시 데이터 값들과 상위 개념 수준에서 추상화된 개념 사이에 존재하는 의미론적 관계를 관리하여 근사적 질의 처리를 수행할 수 있게 해주는 핵심적인 역할을 수행한다고 할 수 있다. 협력적 질의응답에 필요한 의미론적 지식을 관리하기 위하여, 의미론적 거리(Semantic Distance), 퍼지 집합(Fuzzy Set), 규칙(Rule), 개념적 분류(Conceptual Classification) 등을 기반으로 하는 다양한 지식 표현 기법들이 연구되어 왔다.

의미론적 거리 기법은 데이터 개체간의 유사성 정도를 수치화된 거리 개념으로 나타낸다[12,13,15]. 이 기법의 장점은 일단 데이터 개체간의 거리가 결정되면, 질의 완화 알고리즘을 쉽고 효율적으로 개발할 수 있다는 것이다. 각 애트리뷰트 도메인에 대해 다수의 거리 매트릭스가 이용 가능하며, 선택된 거리 매트릭스 집합을 사용자가 적절히 혼합하여 유사성 정도 값을 계산해냄으로써 질의 완화의 방향을 선택할 수 있도록 한다. 하지만 양적/질적 데이터 모두의 유사성을 수량화된 하나의 일률적인 척도로서 나타내기 힘들고, 다양한 데이

터 개체간의 의미론적 유사성을 평가하는 객관적인 기준이 존재하지 않는다는 한계점이 있다.

퍼지 데이터베이스 기법[1,19,25]은 각 도메인에 존재하는 데이터 개체들에 대하여 0과 1사이에 존재하는 유사성 정도값을 부여할 수 있다고 가정한다. 지식 베이스는 다양한 종류의 퍼지 정보, 예를 들어 혼합된 머리 색은 '갈색 0.6 검은색 0.4'로, 특정한 거주지는 '서울 혹은 부산'으로 표현하여 저장하게 된다. 릴레이션들은 퍼지값이나 도메인 개체들의 집합까지 저장할 수 있도록 확장되고, 사용자가 모호한 질의를 작성할 수 있도록 '~보다 크다'와 같은 퍼지 비교 연산자들이 제공된다.

형식상의 프레임웍으로 일차형 논리를 사용하는 규칙 기반 기법[3,9,11]은 기본형, 내장형, 혹은 유도형 술어들(predicates)의 집합에 대한 일차형 식을 이용하여 데이터 개체에 대한 의미론적 정보들과 무결성 제약을 표현한다. 이 기법에서는 기반 데이터베이스가 기본형 술어의 집합으로 간주되고, 데이터베이스에 대한 질의문은 찾고자 하는 정보들이 자유 변수로서 기술되는 술어 규칙으로 표현된다. 질의 처리는 해결(resolution)과 추론 메커니즘을 통해서 이루어지는데, 이때 필요한 질의 완화는 무결성 제약들을 조정함으로써 이루어진다. 규칙 기반 기법의 한계점은 질의완화 과정을 유도해나갈 체계적인 시스템이 부족하고, 질의 응답 과정의 직관성이 떨어진다는 것이다.

본 논문이 기반을 두고 있는 개념적 분류 기법에서는 원시 데이터 개체들로부터 여러 가지 추상화 정보를 추출해냄으로써 추상화된 클래스를 생성해 낸다. 각 클래스들은 서로 유사한 데이터 개체들을 포함하게 되고, 다른 클래스들과 부모/자식 관계를 맺음으로써 하나의 추상화 계층을 형성하게 된다[1,2,6,8,19,22]. 이 기법에서는 동일한 부모 클래스를 갖는 데이터 개체들은 서로 유사한 개체들에 해당하고 부모 클래스는 자식 클래스보다 더욱 일반화된 클래스로 여겨진다. 개념적 분류기법을 이용한 협력적 질의응답에서는 질의조건을 완화시켜 더 넓은 탐색 영역을 갖도록 질의를 변형시

김으로써 질의처리가 이루어진다. 특히 이 기법은 질적이고 분류 가능한 데이터 개체들에 적용할 경우에 많은 장점을 갖긴 하지만, 기존의 개념적 분류 기법들은 보다 다양한 정보를 관리하도록 개선되어야 한다. 다시 말해서, 데이터 값의 추상화 정보뿐 아니라 도메인의 추상화와 관련된 정보까지 포함할 수 있어야 한다. 그 결과 보다 효과적인 지식 관리는 물론 보다 다양한 유형의 질의를 처리할 수 있게 될 것이다. 기존 연구들은 대부분 기반 데이터베이스내의 데이터 값에 대한 추상값을 사용하여 값의 추상화 구조를 구축하는 데에 초점을 두었다. 만일 도메인 추상화 정보가 추가적으로 관리된다면, 도메인을 따라서 추상값 집합을 정렬시킴으로써 보다 체계적인 구조로 조직될 수 있고, 값의 추상화 정보와 도메인 추상화 정보 모두를 이용하여 보다 다양한 형태의 질의가 처리될 수 있다.

이 논문은 값의 추상화 계층뿐 아니라 도메인 추상화 계층을 포함하는 지식 추상화 계층(KAH: Knowledge Abstraction Hierarchy)을 제시함으로써 개념적 분류 프레임워크를 확장하는 것을 목표로 하고 있다. 값의 추상화 계층은 기반 데이터베이스내의 데이터 값들로부터 유도되는 추상값들로 구성되고, 일반화와 세분화 관계를 바탕으로 계층적 구조를 형성한다. 도메인 추상화 계층은 데이터 값들이 속해있는 도메인들로 구성되며, 수퍼 도메인/서브 도메인 관계를 바탕으로 구축되어 진다. 또한 이 논문에서는, KAH에 포함된 의미론적 지식을 관리하기 위하여 관계형 데이터 모델을 이용하여 지식 추상화 정보 데이터베이스를 구축한다. 마지막으로 지식 추상화 계층을 이용한 일반화와 세분화 과정을 바탕으로 다양한 질의처리를 수행해 나가는 협력적 질의응답 메커니즘이 제시된다. 이를 위하여 근사적 선택, 근사적 조인, 개념적 선택, 개념적 조인 등의 네 가지 유형의 메커니즘들이 논의된다. 한편, 일반적인 데이터베이스 응용 시스템에서의 KAH의 유용성을 검증하기 위한 프로토타입 시스템이 구현되어 테스트중에 있다

본 논문의 구성은 다음과 같다. 2장에서는 개념

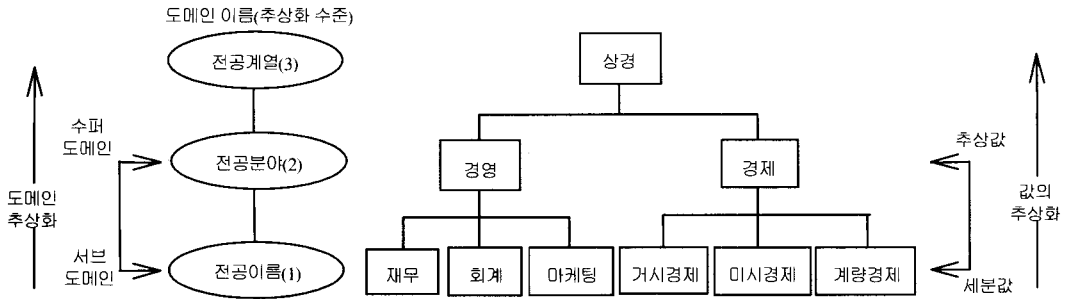
적 분류 프레임워크를 확장하기 위한 지식 표현 프레임워크인 지식 추상화 계층을 소개한다. 3장에서는 KAH에 포함된 의미론적 지식을 관리하기 위하여 관계형 데이터 모델을 바탕으로 지식 추상화 정보 데이터베이스를 구축한다. 4장은 추상화 정보 데이터베이스를 바탕으로 다양한 유형의 협력적 질의 응답을 수행하는 질의 처리 메커니즘이 논의된다. 마지막으로 5장에서 본 논문의 결론이 제시된다.

2. 지식 표현 프레임워크를 위한 지식 추상화 계층

다양한 데이터 모델링 기법들 중에서 데이터 추상화는 값들간의 의미론적 관계를 파악하기 위한 방법으로 널리 사용되어 왔다[21]. 데이터 추상화는 추상화 대상의 구체적인 세부 사항들 대신 추상화 된 개체 사이의 관계에 초점을 맞춘다. 따라서 그러한 데이터 추상화는 협력적 질의응답을 수행하기 위하여 추상화된 데이터 값들과 그리고 세분화된 데이터 값들을 다루는데 매우 유용한 것으로 알려져 있다. 본 논문에서 제시하는 지식 추상화 계층은 데이터 값들 사이의 의미론적 관계를 나타내는 지식 표현 프레임워크로서 데이터 추상화를 기본 개념으로 하고 있다.

지식 추상화 계층은 데이터 추상화를 이용하여 데이터 값과 데이터베이스에 대한 의미론적 지식을 다단계 구조로 표현한다. 데이터 추상화를 통하여 하나의 데이터 값을 추상값으로 일반화시킬 수 있고, 일반화된 추상값을 세분값으로 세분화시킬 수 있다. KAH는 추상값과 세분값 사이의 이러한 일반화/세분화 관계를 바탕으로 구축되는 값의 추상화 계층과 도메인 추상화 계층으로 구성된다.

값의 추상화 계층에서 한 값 u_2 가 다른 값 u_1 보다 좀 더 일반화된 값일 경우, u_2 는 계층 내에서 u_1 보다 상위의 수준에 존재하게 된다. [그림 1]에는 대학 전공에 관한 KAH가 나타나 있다. 그림의 예에서 '재무', '회계', '마케팅' 등은 대학에서 사용되는 전공 이름이다. '경영'은 세 가지 전공을 포괄하



[그림 1] 대학 전공에 관한 지식 추상화 계층

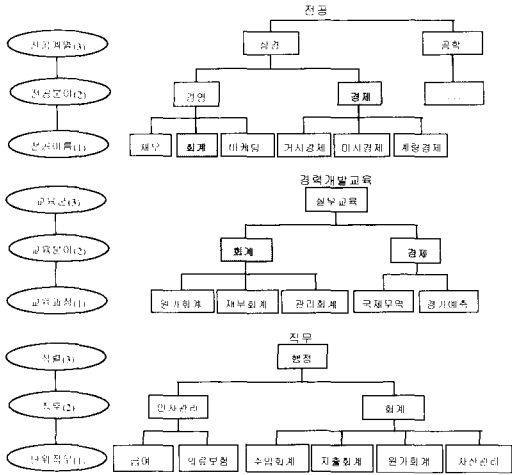
는 전공 분야의 한 값이며, 그들의 추상값으로서 더 높은 수준에 존재한다. 마찬가지로, '경제'는 '거시경제', '미시경제', '계량경제'를 포괄하는 또 다른 전공 영역의 값으로 그들의 추상값에 해당한다. 추상값 '경영'과 '경제'는 다시 더욱 추상화 되어 2단계 높은 수준에 존재하는 값 '상경'과도 추상화 관계를 갖는다. 이처럼 한 계층내에서의 상위 수준은 하위 수준보다 좀 더 추상화 된 데이터 표현을 가능하게 함으로써, 상위 수준에 존재하는 값은 하위 수준에 존재하는 값의 추상값으로 이해된다. 따라서 최상위 수준의 값은 계층 내에서 가장 추상화된 값인 반면, 최하위 수준의 값들은 가장 세분화된 값들이다.

추상값과 세분값 사이의 일반화/세분화는 IS-A 관계에 해당한다. 예를 들어, '재무'는 '경영'이고, '경영'은 '상경'이다. 또한 추상값과 그의 세분값 사이의 사상 원소수(mapping cardinality)는 일대다(one-to-many)로 제한한다. 따라서 '재무'는 '경영'이라는 하나의 추상값을 갖고, '경영'은 '재무', '회계', '마케팅'등의 세 개의 세분값을 갖는다. 한편, 도메인 추상화 계층은 값의 추상화 계층내의 모든 개별 값들이 속해있는 도메인들로 구성된다. 도메인 추상화 계층의 각 도메인은 계층내에서의 절대 위치를 갖게 되며, 따라서 값의 추상화 계층에 존재하는 값들의 추상화 수준은 자신이 속하는 도메인의 추상화 수준에 의해서 결정된다. 한 도메인 D_1 내의 모든 값들이 도메인 D_2 에 자신의 추상값을 가지고 있을 경우 도메인 D_2 는 D_1 의 수퍼 도메인

에, D_1 은 D_2 의 서브 도메인에 해당한다.

값의 추상화 계층에서처럼, 도메인 추상화 계층에서 수퍼 도메인은 서브 도메인보다 상위 수준에 존재하게 된다. 수퍼 도메인과 서브 도메인 사이의 이러한 관계는 값의 일반화/세분화에 대응하여 도메인의 일반화/세분화에 해당한다. 수퍼 도메인은 서브 도메인보다 더 일반화된 값들을 포함하고 있기 때문에 더 일반화 된 도메인이라고 할 수 있기 때문이다. [그림 1]의 세 가지 도메인중 '전공분야'는 '전공이름'의 수퍼 도메인이고 반면에 '전공계열'의 서브 도메인이다. 한편, 값과 도메인 사이에는 INSTANCE-OF 관계가 존재한다. 예를 들어, '재무', '회계', '마케팅', '거시경제' 등과 같이 1단계 수준에 존재하는 값들은 도메인 '전공이름'의 인스턴스들이다. 마찬가지로, '경영', '경제' 등과 같이 2단계 수준에 존재하는 값들은 도메인 '전공분야'의 인스턴스들에 해당한다.

수퍼 도메인과 서브 도메인 사이의 사상 원소수는 일대일(one-to-one)로 제한함으로써 한 서브 도메인은 인접한 하나의 수퍼 도메인을 가지며 반대로 하나의 수퍼 도메인은 인접한 하나의 서브 도메인을 갖는다. 인접한 수퍼 도메인과 서브 도메인의 추상화 수준 차이는 1이므로, 수퍼 도메인에 속하는 추상값을 서브 도메인에 속하는 세분값의 1단계 추상값이라 부른다. 마찬가지로, 추상화 수준의 차이가 n 인 더 높은 수준의 수퍼 도메인내의 추상값들은 서브 도메인내의 세분값의 n 단계 추상값에 해당한다.



[그림 2] 다수의 지식 추상화 계층의 예

KAH는 응용 문제에 따라 그 내용이 달라지게 되므로, [그림 2]와 같이 하나의 데이터베이스에 대하여 여러 개의 KAH 인스턴스들이 존재할 수 있게 된다. [그림 2]는 위에서부터 각각 '전공', '경력개발교육' 그리고 '직무' 계층을 나타낸다. KAH 설계자는 필요할 경우, 동일한 데이터베이스로부터 또 다른 KAH 인스턴스를 추가하게 된다. 다수의 KAH 인스턴스가 존재하는 상황에서, 한 도메인은 유일하게 하나의 계층에만 존재할 수 있지만, 한 값의 다수의 계층에 존재할 수 있으며, 따라서 다수의 도메인에 존재할 수 있다고 가정한다. 예를 들어, [그림 2]의 한 값 '경제'는 '전공'과 '경력개발교육' 2개의 계층에 존재하게 되고, 각각 '전공분야' 도메인과 '교육분야' 도메인에 존재한다. 이처럼 '경제'가 여러 계층에 존재할 수 있기 때문에 '경제'라는 값 이름만으로는 그의 추상값과 세분값을 결정하지 못하고 그 값의 도메인 정보가 추가적으로 필요하게 된다. 예를 들어, '경제'가 '전공분야' 도메인에 존재할 경우 1단계 추상값은 '상경'이 되고, '교육분야' 도메인에 존재할 경우 '실무교육'이 된다. 마찬가지로 주어진 값이 어느 도메인에 존재하느냐에 따라 1단계 세분값 집합이 결정된다. 이러한 특성을 추상값과 세분값의 도메인 종속성이라 부르기로 한다.

3. 지식 추상화 정보 데이터베이스

지식 추상화 정보 데이터베이스는 KAH에 포함된 의미론적 지식을 데이터베이스상에 실제로 관리하여 협력적 질의응답을 수행할 수 있게 해주는 협력적 지식 데이터베이스이다. 관계형 데이터 모델을 이용하여 지식 추상화 정보 데이터베이스의 스키마를 설계함으로써, 기반 데이터베이스와 그의 지식 추상화 정보 데이터베이스 모두를 관계형 질의어 등과 같은 하나의 형식론으로 다룰 수 있게 된다.

3.1 지식 추상화 정보 데이터베이스의 스키마

지식 추상화 정보 데이터베이스는 KAH에 포함된 의미론적 지식을 관리하기 위하여, 값의 추상화 정보, 도메인 추상화 정보, 그리고 추상화 계층과 기반 데이터베이스 사이의 대응 정보를 포함한 세 가지 종류의 정보를 추출해 낸다. 추상화 정보 데이터베이스는 이러한 정보들을 활용하여, 협력적 질의응답을 위한 일반화와 세분화 과정을 수행하게 된다. 지식 추상화 정보 데이터베이스를 구성하는 핵심 릴레이션은 DOMAIN_ABSTRACTION, VALUE_ABSTRACTION, 그리고 ATTRIBUTE_MAPPING이며, 그 스키마는 다음과 같다.

DOMAIN_ABSTRACTION(*Domain*,
Super_Domain, *Hierarchy*, *Abstraction_Level*)
 VALUE_ABSTRACTION(*Value*, *Domain*,
Abstract_Value)
 ATTRIBUTE_MAPPING(*Relation*, *Attribute*,
Domain)

[그림 3]은 [그림 2]의 세 가지 KAH 인스턴스에 포함된 정보를 관리하는 지식 추상화 정보 데이터베이스이다. 설명을 위하여 기반 데이터베이스내의 릴레이션과 그들의 애트리뷰트 이름들이 ATTRIBUTE_MAPPING 릴레이션에 인위적으로 추가되었다. DOMAIN_ABSTRACTION은 도메인 추상화 계층에 포함된 정보를 관리하기 위한 릴레이션

이다. 하나의 도메인은 자신의 이름과 슈퍼 도메인 이름, 계층 이름 그리고 계층 내에서의 위치를 나타내는 추상화 수준을 갖는다. 또한 슈퍼 도메인과 서브 도메인은 일대일(one-to-one) 대응관계를 갖기 때문에, 주어진 도메인의 이름만으로 슈퍼 도메인을 도메인을 결정할 수 있게 되어 도메인 이름이 키 애틀리뷰트가 된다.

● DOMAIN_ABSTRACTION

Domain	Super_Domain	Hierarchy	Abstraction Level
전공이름	전공분야	전공	1
전공분야	전공계열	전공	2
전공계열		전공	3
교육과정	교육분야	경력개발교육	1
교육분야	교육군	경력개발교육	2
교육군		경력개발교육	3
단위직무	직무	직무	1
직무	직렬	직무	2
직렬		직무	3
...

● VALUE_ABSTRACTION

Value	Domain	Abstract_Value
재무	전공이름	경영
회계	전공이름	경영
마케팅	전공이름	경영
거시경제	전공이름	경제
미시경제	전공이름	경제
계량경제	전공이름	경제
경영	전공분야	상경
경제	전공분야	상경
상경	전공계열	
원가회계	교육과정	회계
재무회계	교육과정	회계
관리회계	교육과정	회계
국제무역	교육과정	경제
경기예측	교육과정	경제
회계	교육분야	실무교육
경제	교육분야	실무교육
실무교육	교육군	
급여	단위직무	인사관리
의료보험	단위직무	인사관리
수입회계	단위직무	회계
지출회계	단위직무	회계
원가회계	단위직무	회계
자산관리	단위직무	회계
인사관리	직무	행정
회계	직무	행정
행정	직렬	
...

● ATTRIBUTE_MAPPING

Relation	Attribute	Domain
EMPLOYEE	EMP_ID	ID
EMPLOYEE	MAJOR	전공이름
COLLEGE_MAJOR	MAJOR	전공이름
TASK_HISTORY	EMP_ID	ID
TASK_HISTORY	TASK_PERFORMED	단위직무
TASK_MAJOR	TASK_NAME	단위직무
TASK_MAJOR	REQUIRED_MAJOR_AREA	전공분야
....

[그림 3] 지식 추상화 정보 데이터베이스의 구성 릴레이션

VALUE_ABSTRACTION은 값의 추상화 계층에 포함된 정보를 관리하기 위한 릴레이션이다. 하나의 값에 대하여 Value, Domain 애틀리뷰트는 자신의 이름과 자신이 속한 도메인 이름을 나타내고, Abstract_Value 애틀리뷰트는 1단계 추상값을 나타낸다. 하지만 다수의 계층이 존재하는 경우, 동일한 값의 이름이 서로 다르게 사용되어 서로 다른 계층, 서로 다른 도메인에 존재할 수 있으므로, 값의 이름과 자신의 도메인 이름이 VALUE_ABSTRACTION 릴레이션의 혼합키가 된다. 다시 말해서 추상값의 도메인 종속성으로 인하여, 값과 자신의 도메인 이름 모두가 추상값을 구하기 위해서 사용된다. 예를 들어, '전공' 계층에 존재하는 '회계'는 자신이 속한 도메인이 '전공이름'이기 때문에, VALUE_ABSTRACTION에서 1단계 추상값이 '경영'임을 알 수 있다. 동일한 방법으로, 한 추상값에 대해서 그의 세분값 집합을 구할 수 있게 해준다. 즉, '전공' 계층에 존재하는 '경영'은 '전공이름' 도메인에 존재하는 '재무', '회계', 그리고 '마케팅'을 세분값으로 갖는다. 추상화 관계의 관점에서 볼 때, VALUE_ABSTRACTION 릴레이션이나 DOMAIN_ABSTRACTION 릴레이션 모두 하나의 투플에 1단계 추상화 관계만을 표현한다. 임의의 n단계 추상값은 이러한 1단계 추상화 관계를 바탕으로 이행적으로 유도해 낼 수 있다.

ATTRIBUTE_MAPPING 릴레이션은 기반 데이터베이스에 존재하는 릴레이션 이름과 애틀리뷰트 이름, 그리고 그 애틀리뷰트가 갖는 도메인 이

름 사이의 관계를 나타낸다. 동일한 애트리뷰트 이름이 다수의 릴레이션에서 사용될 수 있기 때문에, 릴레이션과 애트리뷰트의 이름이 혼합키가 된다. ATTRIBUTE_MAPPING 릴레이션은 협력적 질의 응답의 질의분석 과정에서 사용된다.

이러한 릴레이션들은 관계형 스키마를 이용하여 서로 다른 추상화 수준에 존재하는 데이터 개체들 사이의 추상화 관계를 표현한 것이다. VALUE_ABSTRACTION 릴레이션은 추상값과 세분값 집합 사이에 존재하는 값의 추상화 관계를 나타낸다. 마찬가지로, DOMAIN_ABSTRACTION 릴레이션은 추상값의 도메인과 세분값의 도메인 사이에 존재하는 도메인 추상화 관계를 나타낸다. ATTRIBUTE_MAPPING 릴레이션은 이러한 추상화 정보를 기반 데이터베이스에 존재하는 릴레이션과 애트리뷰트에 관한 정보와 연결시킴으로써, 협력적 질의처리를 위해 질의를 분석할 수 있게 해준다. 구축된 지식 추상화 정보 데이터베이스는 기반 데이터베이스의 변화를 동적으로 처리할 수 있음은 물론, 협력적 질의응답을 위해 필요한 값의 일반화, 세분화 등과 같은 여러 가지 연산들을 수행할 수 있다.

3.2 값의 일반화 연산

추상화 정보 데이터베이스를 이용하여, 질의 조건에 나타난 값을 일반화시키는 것은, 그 값을 자신의 추상값으로 대체하는 것을 의미한다. KAH에서 상위의 추상화 수준에 존재하는 하나의 추상값은 하위 수준에 존재하는 다수의 세분값에 해당한다. 따라서, 협력적 질의응답에서 한 추상값을 사용해 질의를 수행하는 것은 다수의 세분값들을 사용해 질의하는 것과 동일한 의미를 갖게 된다. 예를 들어, [그림 2]에 나타난 '전공' 계층에 존재하는 '재무'를 고려해 보자. 1단계 일반화는 '재무' 외에 '회계', '마케팅'까지 포괄하는 1단계 추상값 '경영'을 제공한다. 2단계 일반화는 한 수준 더 추상화된 값인 2단계 추상값 '상경'을 반환한다. 2단계 추상

값 '상경'은 '경제'를 포함하여 더 넓은 범위의 단사값들을 포함하는 값이다.

한 값의 임의의 n 단계 일반화는 두 개의 추상화 정보 릴레이션, VALUE_ABSTRACTION과 DOMAIN_ABSTRACTION 모두를 필요로 한다. n 단계 추상값은 1-단계 일반화 과정을 n 번 반복함으로써 얻어질 수 있지만, 추상값의 도메인 종속성으로 인하여, 슈퍼 도메인 정보를 필요로 하는데, 이것은 DOMAIN_ABSTRACTION 릴레이션을 이용해서 얻어야만 하기 때문이다. 예를 들어, '경력개발교육' 계층의 '교육과정' 도메인에 속하는 '원가회계'의 2단계 추상값을 구해보자. 먼저 1단계 일반화 과정은 '원가회계'의 도메인이 '교육과정'임을 이용하여, VALUE_ABSTRACTION 릴레이션으로부터 1단계 추상값 '회계'를 찾아낸다. 하지만, 두 번째로 '회계'에 대한 1단계 일반화 과정을 수행하기 위해서는, '회계'의 도메인 이름을 알아야만 한다. 그것은 물론 '교육과정'의 슈퍼 도메인에 해당한다. 이러한 슈퍼 도메인 정보는 DOMAIN_ABSTRACTION 릴레이션을 이용하여 얻어질 수 있고, 그 결과 '교육분야'가 슈퍼 도메인이 된다. 3단계 추상값 역시, 2단계 추상값과 2단계 추상값이 존재하는 도메인 정보를 가지고 세 번째 1단계 일반화 과정을 수행함으로써 얻을 수 있다. 이러한 과정을 반복함으로써, 임의의 n 단계 추상값을 구할 수 있게 된다.

3.3 값의 세분화 연산

하나의 값은 자신보다 하위 수준에 존재하는 다수의 세분값들을 갖는다. 주어진 값에 대한 세분화는 KAH의 하위 수준으로 이동해 가면서 임의의 n 단계 세분값을 찾아내는 것을 말한다. 예를 들어, [그림 2]의 '전공' 계층에 존재하는 '상경'을 고려해 보자. '상경'에 대한 1단계 세분화 과정은 1단계 세분값으로 '경영'과 '경제'를 반환한다. 2단계 세분화는 '재무'에서 '계량경제'에 이르는 6개의 값들 즉, 2단계 세분값들을 반환하며, 그것들은 '경영'과 '경

제'의 1단계 세분값들에 해당한다.

값의 일반화와 마찬가지로, 세분화 역시 임의의 n 단계 세분값을 찾기 위하여 VALUE_ABSTRACTION 릴레이션과 DOMAIN_ABSTRACTION 릴레이션 모두를 사용한다. 한 값에 대한 1단계 세분값들을 찾기 위해서는 추상값과 세분값의 도메인 종속성으로 인하여, 그 세분값들이 존재하게 될 도메인을 먼저 알아내야 한다. 이것은 DOMAIN_ABSTRACTION 릴레이션을 이용하여 주어진 값이 존재하는 도메인의 서브 도메인을 찾아냄으로써 가능하게 된다. 일단 세분값들이 존재하게 될 도메인이 결정되면, VALUE_ABSTRACTION 릴레이션을 이용하여 주어진 값의 세분값들을 찾아낼 수 있다. 예를 들어, [그림 3]의 추상화 정보 데이터베이스를 이용하여 [그림 2]에서 '전공' KAH에 존재하는 '경제'의 세분값을 구해 보기로 하자. 설명한 것처럼, '경제'의 1단계 세분값을 구하기 위해서는, 세분값의 도메인 종속성으로 인하여 '경제'의 세분값들이 존재하게 될 도메인을 먼저 알아야 한다. 이는 '경제'가 속하는 도메인 '전공분야'의 서브 도메인에 해당하며, DOMAIN_ABSTRACTION 릴레이션을 이용하여 '전공이름'이 됨을 알 수 있다. 그 다음에, VALUE_ABSTRACTION 릴레이션에서 Domain과 Abstract_Value 애트리뷰트 값으로 각각 '전공이름'과 '경제'를 갖는 튜플들을 선택함으로써 '경제'의 세분값 '거시경제', '미시경제', '계량경제' 등을 찾아낸다. n 단계 세분화 과정은 1단계 세분화 과정을 동일하게 n 번 반복함으로써 이루어질 수 있다.

4. 협력적 질의응답 과정

협력적 질의응답에서, 데이터베이스내에 질의조건을 정확히 만족하는 답이 존재하지 않더라도 유사한 답이나 질의와 관련된 정보는 존재할 수 있을 것이다. 그러한 정보들은 것은 질의의 탐색조건을 완화하여 해 범위가 추가적인 정보까지 포함하도록 질의 조건을 확장함으로써 얻을 수 있게 된다. 협

력적 질의응답은 정보요구 형태와 지식 추상화 정도와 관련된 질의완화 정도에 따라 다양한 방법으로 이루어 질 수 있다. 우리는 지식 추상화 정보 데이터베이스를 이용한 일반화, 세분화 과정에 기초하여 네 가지 유형의 전형적인 협력적 질의응답 메커니즘을 제시하고자 한다. 그것은 근사적 선택, 개념적 선택, 근사적 조인, 그리고 개념적 조인 등이다. 또한 다른 연구들과 비교하여 본 연구에서 제시하는 추상화 정보 데이터베이스가 갖는 장점을 논의하고자 한다. 설명을 위해서 다음과 같이 단순화된 인사 정보 데이터베이스를 사용하기로 한다.

```
EMPLOYEE ((id, emp_name, dept, title))
TASK_HISTORY((id, beginning_date,
              ending_date, task_performed))
COLLEGE_MAJOR((id, major, entrance_data,
                graduation_date))
CAREER_PATH((task, prerequisite_task))
TASK_MAJOR((task, required_major_area))
```

EMPLOYEE 릴레이션은 모든 직원의 현재 직무상 지위에 관한 정보를 포함하고 TASK_HISTORY는 과거에 수행하였던 직무 기록을 포함한다. COLLEGE_MAJOR는 직원의 대학 전공 기록을 포함하고 있으며, CAREER_PATH는 여러 직무들간의 선행 관계에 대한 정의를 포함한다. 마지막으로 TASK_MAJOR는 개별 직무와 그 직무를 수행하는데 요구되는 대학 전공 영역 사이의 관계를 기술한다. 이러한 릴레이션들이 [그림 3]에 나타난 지식 추상화 정보 데이터베이스의 기반 데이터베이스를 구성하게 된다. 여기서 TASK_MAJOR의 required_major_area 애트리뷰트의 도메인이 '전공이름'이 아니라 '전공분야'임을 주목해야 한다. 왜냐하면 각 직무를 개별 전공 이름보다는 전공 이름의 범위를 포함하는 전공 영역과 연결시키는 것이 보다 간결하고, 또 어떤 면에서는 각 개별 직무를 세부적인 전공 이름과 연결시킨다면 실제적으로 너무 협소한 범위로 제한되기 때문이다.

4.1 근사적 선택

협력적 질의응답에서, 근사적인 해를 얻기 위하여 질의문에 나타난 하나의 애트리뷰트가 근사화될 수 있다. 근사적 선택이란, 질의문의 선택조건에 애트리뷰트와 찾고자 하는 목표값이 명시되었을 때, 그 애트리뷰트를 목표값과 유사한 값들의 범위로 추상화함으로써 질의조건을 완화시키고 탐색영역을 확장하는 것을 말한다. 예를 들어, 인사 관리자가 특정 직무를 수행하기 위한 '재무' 전공자를 찾고자 할 때, '재무'를 전공한 직원이 존재하지 않거나, 혹은 충분한 인원이 존재하지 않는 경우, 탐색영역을 확장하여, '재무'와 유사한 전공을 갖는 다른 직원들을 찾을 수 있을 것이다. 한 값의 근사값들을 찾는 것은 지식 추상화 정보 데이터베이스를 이용하여 그 값의 추상값을 찾아냄으로써 가능하게 된다. 따라서, 근사적 선택 조건 C가 주어졌을 때 다음과 같은 과정을 거쳐 질의완화가 이루어진다.

1. 질의조건 C을 정확히 만족하는 값을 찾는다. 만약 실패할 경우,
2. 근사화 연산자(eg., =?)에 의하여, KAH의 상위 수준으로 이동해 가면서 질의 조건에 명시된 목표값을 일반화 시켜 추상화 된 질의 조건을 만든다(질의 일반화 과정). 즉, $C \rightarrow C'$. 다음에,
3. KAH의 하위 수준으로 이동해 가면서 추상화 된 질의내의 추상값을 세분화 시켜, 세분화된 값으로 표현된 정규 질의 형태로 변형시킨다(질의 세분화 과정). 즉, $C' \rightarrow C''$.

이러한 단계를 통하여 다음과 같은 근사적 선택 질의가 처리될 수 있다.

- “전공학과가 재무이거나 그와 유사한 학과인 직원의 이름과 부서를 찾아라.”

```
select e.emp_name, e.dept
from employee e, college_major c
```

where c.major =? “재무” and e.id = c.id

근사적 선택 질의는 질의 완화 연산자 =?에 의해서 표현된다. 따라서 c.major=?“재무”의 의미는 college_major 릴레이션 내에서 major 애트리뷰트 값이 “재무” 이거나 그와 유사한 값을 갖는 튜플들을 찾는 것이다. 질의 완화 연산자가 의미를 갖기 위해서는, 질의에서 명시된 애트리뷰트와 목표값의 도메인이 동일해야만 한다. 이러한 의미에서 위의 질의문 예에서는 두 도메인이 ‘전공이름’으로 동일하기 때문에 근사적 선택 질의에 해당된다. college_major 릴레이션의 major 애트리뷰트의 도메인은 ATTRIBUTE_MAPPING 릴레이션으로부터, ‘재무’의 도메인은 VALUE_ABSTRACTION 릴레이션으로부터 ‘전공이름’임을 알 수 있다.

또한, 지식 추상화 정보 데이터베이스를 이용한 협력적 질의응답 과정에서 질의문에 명시된 애트리뷰트의 도메인 정보는 질의처리 과정에서 수행되는 일반화, 세분화 과정의 중심축 역할을 한다. 왜냐하면, 애트리뷰트 도메인 정보가 주어진 목표값과 애트리뷰트 값의 추상값과 세분값 집합을 구하기 위한 다양한 질의 완화 경로를 생성해 주기 때문이다. 앞에서 설명한 것처럼, 애트리뷰트의 도메인은 ATTRIBUTE_MAPPING 릴레이션을 통해 알 수 있고, 반면에 목표값의 도메인은 VALUE_ABSTRACTION 릴레이션을 사용하여 알아낼 수 있다. 그러한 도메인 정보는 다음의 연산자들을 사용하여 애트리뷰트, 목표값에 대한 일반화 세분화 과정을 수행하기 위해서 반드시 필요하게 된다.

- Get_Abstract_Value(value)는 주어진 값에 1단계 일반화 과정을 수행하여 1단계 추상값을 제공한다.
- Get_Abstract_Value(attribute)는 주어진 애트리뷰트가 가질 수 있는 모든 값들의 추상값들을 제공한다. 1단계 추상값과의 선택연산이나 조인 연산을 수행할 수 있게 해준다.
- Get_Specific_Value(value)는 주어진 값에 대한

1단계 세분화 과정을 수행하여 1단계 세분값 집합을 반환한다.

하나의 추상값은 자신보다 하위 수준에 존재하는 다수의 세분값들에 해당하므로, 1단계 추상값을 찾아 선택조건에 명시된 목표값의 근사값들을 생성할 수 있게 된다. 따라서, 다음과 같이 1단계 일반화된 질의를 생성해 낸다.

- “전공학과가 재무의 일반화된 학과영역에 포함되는 직원의 이름과 부서를 찾아라.”

```
select e.emp_name, e.dept
from employee e, college_major c
where c.major is-a Get_Abstract_Value("재무")
and e.id = c.id
```

is-a는 세분값과 추상값 사이의 일반화 관계를 나타낸다. 질의분석을 통하여 ‘재무’의 도메인이 ‘전공이름’임을 이용하여, 일반화된 질의문의 `Get_Abstract_Value("재무")`는 추상값 ‘경영’을 반환하며, 따라서 질의 조건은 다음과 같이 변환된다.

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

‘경영’을 세분화하였을 때, `c.major`가 세분화된 값의 집합 내에 존재하는 값을 취하는지 알아내기 위하여, **is-a** 연산자는 추상값 ‘경영’을 변환하여 다음과 같은 질의를 생성해 낸다.

- “전공학과가 경영의 세분학과 집합에 존재하는 직원의 이름과 부서를 찾아라.”

```
select e.emp_name, e.dept
from employee e, college_major c
where c.major in Get_Specific_Value("경영")
and e.id = c.id
```

추상값 ‘경영’의 1단계 세분화를 통하여 ‘재무’와

유사한 값들을 포함하는 세분값 집합 {재무, 회계, 마케팅}을 얻는다. 따라서, 다음과 같이 최종적으로 세분화된 질의문을 작성하여 정규 SQL 질의문을 통하여 처리한다.

```
select e.emp_name, e.dept
from employee e, college_major c
where c.major in ("재무", "회계", "마케팅")
and e.id = c.id
```

4.2 개념적 선택

질의 작성시 필요한 값들을 사용자가 명확하게 알고 있지 못할 경우에는 질의문을 정확하게 작성할 수 없게 된다. 이것은 데이터의 분류와 범주화가 강조되는 데이터베이스 시스템일 경우에 더욱 그러하다. 예를 들어, 인사 관리자가 일반적인 상경 분야를 전공한 사람들을 찾고 싶지만, 상경 분야의 대학 전공 이름에 익숙치 못하여 ‘회계’, ‘재무’, ‘마케팅’과 같은 구체적인 값들을 명확히 알지 못할 경우, 질의문을 정확하게 작성할 수 없게 된다. 이러한 경우에 사용자는 “일반적인 상경 분야에 속하는 전공을 갖는 직원을 찾아라”라는 보다 일반적이고 개념적인 질의를 작성하여 협력적 질의처리를 요구하게 된다. 개념적 질의는 선택조건 목표값이 애트리뷰트보다 추상화 수준이 높은 추상값인 경우에 해당하며, 추상값을 세분화하여 기존의 정규적인 질의로 변형시킨다. 따라서, 개념적 선택 조건 *C*가 주어졌을 때, 질의완화의 핵심 단계는 다음과 같다.

1. 질의문에 근사화 연산자가 존재하지 않기 때문에, 정규 질의로 가정한다. 따라서, 질의 조건 질의조건 *C*를 만족하는 정확한 값을 찾는다.
2. 만약 실패할 경우, 질의문에 나타난 개념적인 목표값의 도메인이 애트리뷰트의 도메인과 동일한 계층에 존재하는지를 파악한다.
3. KAH의 상위 수준으로 이동해 가면서 추상화

되어 있는 질의문에 명시된 추상값을 세분화 하여, 세분화된 값으로 표현된 정규 질의 형태로 변형시킨다(질의 세분화 과정). 즉, C' → C''.

이러한 단계를 통하여 다음과 같은 개념적 선택 질의가 처리될 수 있다.

- “상경 영역에 포함된 학과를 전공한 직원의 이름과 부서를 찾아라.”

```
select e.emp_name, e.dept
from employee e, college_major c
where c.major = “상경” and e.id = c.id
```

위에서 작성된 질의는 major 애트리뷰트 값으로 ‘상경’이 존재하지 않기 때문에 질의조건을 만족하는 답이 존재하지 않을 뿐만 아니라 major와 ‘상경’의 도메인이 서로 다르기 때문에, 정규 질의 응답에서는 이러한 질의를 잘못 작성된 질의로 해석한다. 하지만 협력적 질의응답 메커니즘은 지식 추상화 정보 데이터베이스를 이용하여 ‘상경’의 도메인이 major 애트리뷰트의 도메인의 수퍼 도메인임을 인식하고, 그러한 조건을 개념적 질의 조건으로 해석한다.

이러한 개념적 선택 조건은 질의문에 명시된 애트리뷰트와 목표값이 동일한 계층의 서로 다른 도메인에 존재하는 경우에 해당한다. 위의 질의에서 ATTRIBUTE_MAPPING 릴레이션을 이용하여 c.major 애트리뷰트의 도메인이 ‘전공이름’임을 알 수 있고, VALUE_ABSTRACTION 릴레이션을 이용하여 ‘상경’의 도메인이 ‘전공계열’임을 알 수 있다. 만약 개념적 조건으로 해석하는 것이 타당하려면, ‘상경’의 도메인이 c.major 애트리뷰트의 도메인의 수퍼 도메인이 되어야 한다. 수퍼 도메인과 서브 도메인의 그러한 일반화 관계는 DOMAIN_ABSTRACTION 릴레이션에서 파악이 가능하다. 일단 질의 조건이 개념적 조건으로 판단이 되면, 다음과 같이 세분화된 질의를 생성해 낸다.

- “상경의 세분학과를 전공한 직원의 이름과 부서를 찾아라.”

```
select e.emp_name, e.dept
from employee e, college_major c
where c.major in Get_Specialized_Value(“상경”) and e.id = c.id
```

추상값 ‘상경’의 1단계 세분화를 통하여 세분값 집합 (경영, 경제)를 찾아낸다.

```
select e.emp_name, e.dept
from employee e, college_major c
where c.major in (“경영”, “경제”) and e.id = c.id
```

하지만 계속해서 목표값 ‘경영’과 ‘경제’의 도메인이 c.major 애트리뷰트의 수퍼 도메인이므로, 두 도메인이 서로 동일한 도메인이 될 때까지 계층을 따라서 세분화가 계속 진행된다. 두 번째 질의 세분화의 결과로서 다음과 같은 정규 질의문이 생성되어 선택조건의 모든 값들이 동일한 도메인에 존재하게 된다.

```
select e.emp_name, e.dept
from employee e, college_major c
where c.major in (“재무”, “회계”, “마케팅”, “거시경제”, “미시경제”, “계량경제”) and e.id = c.id
```

4.3 근사적 조인

근사적 선택의 확장으로서, 근사적 조인 조건에서 두 조인 애트리뷰트의 애트리뷰트 값을 조인하는 대신, 조인 애트리뷰트의 값의 근사값을 통해 조인되어 진다. 예를 들어, 인사관리자가 ‘자산관리’ 직무의 선행 직무를 수행한 사람을 찾고자 한다고 가정하자. 만약 선행 직무 경험을 한 적당한 직원의 수가 충분치 않을 경우, 조건을 완화함으로써

써 대체 가능한 직원들도 찾아질 수 있다. 그러한 상황에서, 근사적 조인 질의가 ‘자산관리’ 직무의 선행 직무와 비슷한 직무를 경험한 사람들을 찾게 해 준다. 질의의 의미가 다소 복잡하지만, KAH의 관점에서 본다면 질의는 보다 단순하게 해석될 수 있다. 즉, 근사적 조인 질의는 두 개의 애트리뷰트를 그들의 추상값을 이용하여 조인하여 일상적인 조인의 결과보다 더 넓은 범위의 조인 결과를 생성하게 된다. 따라서 근사적 조인 조건 C가 주어졌을 때, 핵심적인 질의완화 단계는 다음과 같다.

1. 근사화 연산자(e.g., =?)에 의하여, KAH의 상위 수준으로 이동해 가면서 질의조건의 두 개의 조인 애트리뷰트 모두를 일반화하여 각각에 관련되는 추상값들을 취하도록 한다(질의 일반화 과정). 즉, $C \rightarrow C'$. 다음에,
2. 두 조인 애트리뷰트가 취하는 추상값들을 바탕으로 정규 조인 연산을 수행한다.

이러한 단계를 통하여 다음과 같은 근사적 조인 질의가 처리될 수 있다.

- “Asset Management 직무를 수행하기 위해 필요한 선행직무와 유사한 직무를 수행했던 경험이 있는 직원의 이름과 부서, 직위를 찾아라.”

```
select e.emp_name, e.dept, e.title
from employee e, task_history t, career_path c
where e.id = t.id and
      c.task = “자산관리” and
      t.task =? c.prerequisite_task
```

질의문에서 질의 완화 연산자 =?가 t.task와 c.prerequisite_task 사이에서 사용된다. 완화 연산자의 사용은 비교되는 두 개의 애트리뷰트 모두가 동일한 도메인을 갖는다는 면에서 근사적 선택 질의와 유사하다. 위의 예제에서 두 개의 도메인 모두는 ‘단위직무’이다. 다음에 두 개의 애트리뷰트들이 자신이 원래 취하는 값 대신 추상값을 취하도록 질의를 일

반화함으로써 다음과 같은 질의를 생성해 낸다.

- “이전에 수행한 직무가 포함된 영역이 ‘자산관리’의 선행직무가 포함된 영역과 동일한 직원의 이름과 부서, 직위를 찾아라.”

```
select e.emp_name, e.dept, e.title
from employee e, task_history t, career_path c
where e.id = t.id and
      c.task = “자산관리” and
      Get_Abstract_Value(t.task) =
      Get_Abstract_Value(c.prerequisite_task)
```

추상값을 바탕으로 조인 애트리뷰트를 조인하는 것은, 지식 추상화 정보 데이터베이스를 이용할 경우 여러 가지 방법으로 이루어질 수 있다. 한가지 직관적인 방법은 VALUE_ABSTRACTION 릴레이션으로부터 Value와 Abstract_Value 애트리뷰트만을 뽑아낸 ABSTRACTION 릴레이션을 이용하는 것이다. 위 예의 경우, ‘단위직무’ 도메인을 갖는 레코드들이 [그림 4]와 같은 ABSTRACTION 릴레이션으로 추출된다. ABSTRACTION 릴레이션은 각 세분값의 추상값을 포함하고 있기 때문에, 두개의 조인 애트리뷰트 사이의 매개체로서 사용될 수 있다. ABSTRACTION을 이용하여 다음과 같은 정규 질의를 생성해 낸다.

● ABSTRACTION

Value	Abstract_Value
급여	인사관리
의료보험	인사관리
수입회계	회계
지출회계	회계
원가회계	회계
자산관리	회계
.....	

[그림 4] Abstraction 릴레이션

```
select e.emp_name, e.dept, e.title
from employee e, task_history t, career_path c
      ABSTRACTION a1, ABSTRACTION a2
```

where e.id = t.id and
 c.task = "자산관리" and
 c.prerequisite_task = a1.value and t.task
 = a2.value and
 a1.abstract_value = a2.abstract_value

4.4 개념적 조인

마지막으로, 조인 조건의 두 개의 애트리뷰트가 서로 다른 도메인을 갖으며, 따라서 서로 다른 추상화 수준에 존재하는 경우를 생각할 수 있다. 개념적 조인은 한 조인 애트리뷰트의 도메인이 다른 조인 애트리뷰트의 수퍼 도메인이 되는 경우로서, 서로 다른 도메인을 갖는 애트리뷰트임에도 불구하고 조인연산이 이루어질 수 있다. 인사 관리 데이터베이스 예제에서, TASK_MAJOR 릴레이션은 각 직무에 요구되는 전공들을 관리한다. COLLEGE_MAJOR 릴레이션에서 required_major_area 애트리뷰트의 도메인은 major 애트리뷰트의 도메인보다 더욱 일반화된 애트리뷰트임에 주의해야 한다. 그러한 상황에서, 사용자가 특정 직무 '의료보험'을 수행하기 위해서 필요한 대학 전공을 갖는 직원을 찾고자 할 수 있다. 개념적 조인은 그러한 요구를 처리하기 위하여 수행된다. 두 개의 조인 애트리뷰트의 도메인이 서로 다를 경우 질의는 개념적 조인으로 해석되고, 일반화 과정을 통해서 두 개의 도메인을 동일한 도메인으로 변형시켜 정규 조인 질의를 생성한다. 따라서 개념적 조인 조건 C가 주어졌을 때, 질의완화의 핵심 단계는 다음과 같다.

1. 근사화 연산자가 존재하지 않기 때문에 정규 질의로 가정한다. 따라서 정규 조인 연산을 수행하여 질의 조건 C를 정확히 만족하는 값들을 탐색한다.
2. 만약 조인 연산이 실패하고 두개의 조인 애트리뷰트의 도메인이 수퍼 도메인/서브 도메인 관계에 있다면, 서브 도메인에 해당하는 애트리뷰트를 일반화하여 두 애트리뷰트가 동일

한 도메인을 갖도록 하여라.

3. 두 조인 애트리뷰트가 취하는 값에 의거하여 정규 조인 연산을 수행한다.

이러한 단계를 통하여 다음과 같은 개념적 조인 질의가 처리될 수 있다.

- "의료보험 직무를 수행하는데 필요한 학과를 전공한 직원의 이름과 부서를 찾아라."

```
select e.emp_name, e.dept
from employee e, task t, college_major c
where t.task_name = "의료보험" and
t.required_major_area = c.major and
e.id = c.id
```

위에서 작성된 질의는 두 조인 애트리뷰트의 도메인이 서로 다르고, 한 도메인 '전공분야'가 다른 도메인 '전공이름'의 수퍼 도메인이 되기 때문에 질의는 개념적 조인 질의로 해석된다. 따라서 더 낮은 수준의 도메인을 갖는 애트리뷰트 c.major에 대해서 일반화가 이루어져 다음과 같은 질의를 생성해 낸다.

- "전공한 학과가 포함된 영역이 의료보험 직무를 수행하는데 필요한 학과영역에 해당하는 직원의 이름과 부서를 찾아라."

```
select e.emp_name, e.dept
from employee e, task t, college_major c
where t.task_name = "의료보험" and
t.required_major_area =
Get_Abstract_Value(c.major) and
e.id = c.id
```

근사적 조인 메커니즘처럼, 두 개의 조인 애트리뷰트를 연결하기 위하여 ABSTRACTION 릴레이션이 사용되어 다음과 같은 질의를 생성해 낸다.

```
select e.emp_name, e.dept, e.title
from employee e, task t, college_major c
```

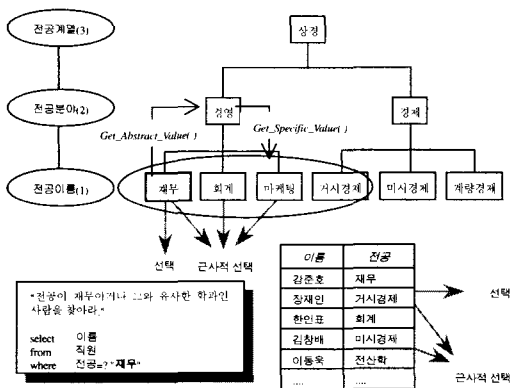
```

ABSTRACTION a1
where e.id = c.id and
      c.task_name = "의료보험" and
      c.major = a1.value and
      t.required_major_area =
      a1.abstract_value
    
```

본 연구에서 제시한 네 가지 질의 유형은 협력적 질의응답의 서로 다른 측면을 지원한다. 사용자의 정보 요구 형태와 KAH에 관한 지식의 정도에 따라, 하나의 질의에 네 가지 유형의 질의 조건이 혼합됨으로써 더 복잡하고 정교한 질의처리를 수행할 수 있다.

4.5 협력적 질의 응답 과정에 대한 예

앞에서 설명한 질의처리 과정이 기반 데이터베이스를 바탕으로 구축된 KAH와 관련해서 어떻게 수행되는지 근사적 선택, 근사적 조인 질의의 예를 통해 살펴보기로 한다. [그림 5]에 근사적 선택 질의에 대한 예가 나타나 있다.



[그림 5] 근사적 선택 질의처리 과정

근사적 선택 질의는 다음과 같은 질의 일반화, 질의 세분화 과정을 통해서 처리된다.

o 질의 일반화

```

select 이름
from 직원
where 전공 is-a Get_Abstract_Value("재무")

select 이름
from 직원
where 전공 is-a "경영"
    
```

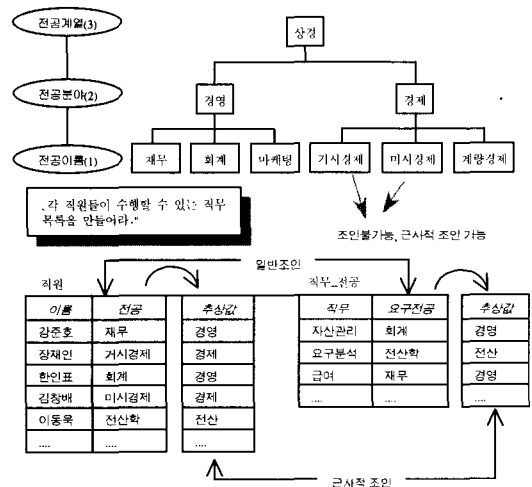
o 질의 세분화

```

select 이름
from 직원
where 전공 in Get_Specific_Value("경영")

select 이름
from 직원
where 전공 in ("재무", "회계", "마케팅")
    
```

다음의 [그림 6]은 각 직무를 수행하기 위해 필요한 전공을 갖는 직원들을 찾아내기 위하여 근사적 조인을 수행하는 예이다.



[그림 6] 근사적 조인 질의처리 과정

근사적 조인 질의는 다음과 같은 과정을 통해 처리되어 진다.

○ 근사적 조인 질의

```
select 이름, 직무
from 직원, 직무_전공
where 직원.전공=?직무_전공.요구전공
```

○ 질의 일반화

```
select 이름, 직무
from 직원, 직무_전공
where Get_Abstract_Value(직원.전공)=
       Get_Abstract_Value(직무_전공.요구전공)
```

앞에서 설명한 바와 같이 질의 세분화 과정을 수행하기 위해서 다음과 같은 ABSTRACTION 릴레이션을 이용하게 된다.

ABSTRACTION

Value	Abstract_Value
재무	경영
회계	경영
마케팅	경영
거시경제	경제
미시경제	경제
계량경제	경제

[그림 7] ABSTRACTION 릴레이션

○ 질의 세분화

```
select 이름, 직무
from 직원, 직무_전공, ABSTRACTION A1,
       ABSTRACTION A2
where 직원.전공= A1.value and 직무_전공.요구
       전공= A2.value and A1.Abstract_Value =
       A2.Abstract_Value
```

4.6 다른 연구와의 비교

이 절에서는 본 연구에서 제시하는 지식 추상화 정보 데이터베이스와 개념적 분류 기법, 객체 지향 데이터베이스 기법, 그리고 규칙 기반 기법을 포함

한 다른 기법들과 비교하고자 한다.

먼저, 지식 표현 스키마의 관점에서, 기존의 기법들은 KAH에서처럼 값의 추상화 계층과 도메인 추상화 계층 모두를 포함하지는 않는다. 타입 추상화 계층(type abstraction hierarchy)[7,9]과 같은 개념적 분류 기법은 값의 추상화 개념만으로 다단계 지식 표현 스키마를 제공한다. 슈퍼 타입/서브 타입 관계를 바탕으로 다수의 인스턴스 층(layer)을 가능케 하여 서브 타입의 인스턴스들이 슈퍼 타입 내에서 다른 인스턴스로 표현될 수 있도록 한다. 하지만, 도메인 추상화 개념은 명시적으로 포함되지 않으며, 타입 추상화 계층은 인스턴스 층들에 대해서만 일반화/세분화 관계를 적용할 뿐, 타입 층에는 적용되지 않는다. 반면에 KAH에서는 일반화/세분화 관계가 인스턴스 층과 도메인 층 모두에 적용된다. 따라서 서브 타입의 한 인스턴스는 슈퍼 타입에 속하는 한 추상 인스턴스를 갖게되며, 추상 인스턴스의 도메인은 서브 타입 인스턴스의 슈퍼 도메인으로 간주된다.

한편 객체 지향 데이터베이스 기법은 클래스 층과 인스턴스 층의 두 개의 층을 포함한다. 클래스 층에서는 동일한 특성을 갖는 개체들을 하나의 클래스로 그룹화하기 위하여 분류(classification) 개념을 사용한다. 클래스들은 이들 사이의 일반화/세분화 관계를 바탕으로 슈퍼 타입/서브 타입 관계를 갖게되며, 분류의 이러한 관점은 포함된 개체 인스턴스 값들 보다는 전체적인 타입(type)과 관련을 갖게 된다. 그러한 클래스 지향적 추상화는 KAH에서의 도메인 추상화 계층과 거의 동일하긴 하지만, 인스턴스 층에서는 한 클래스 내의 인스턴스들이 추상화 된 여러 개의 인스턴스 그룹으로 분류되어지는 값의 추상화와 같은 개념이 존재하지 않는다. 따라서 객체 지향 데이터베이스에서 서브 클래스에 속하는 세분화된 인스턴스 집합이 슈퍼 클래스에 속하는 추상화 된 개별 인스턴스들과 직접적으로 관련을 갖을 수 없다.

두 번째, 협력적 질의 처리 관점에서, 규칙 기반 기법[4,10,12]은 논리적 질의 완화 규칙 집합을 바

탕으로 한 추론 엔진을 사용한다. 하지만 질의완화 과정은 사용자들과의 상호 작용을 통한 제어를 수행할 수 있을 만큼 직관적이지는 못하다. 반면, KAH 기반 지식 추상화 데이터베이스는 좀더 인터랙티브하고 유연한 질의 변형 과정을 수행할 수 있다. 이때 값의 추상화 계층은 일반화와 세분화 과정을 통하여 질의 완화에 필요한 기초 정보를 제공하며, 도메인 추상화 계층은 값의 도메인, 도메인의 수퍼 도메인, 도메인의 추상화 수준과 같은 정보들을 사용하여 좀더 동적인 방법으로 질의처리를 수행할 수 있게 해준다. 추상화 수준에 관한 정보는 여러 도메인의 추상화 수준을 비교하여 개념적 선택, 개념적 조인과 같은 개념적 질의를 처리할 수 있게 해 주고, 질의 완화 시 원하는 정도까지 점진적으로 혹은 직접적으로 완화할 수 있게 하여 사용자에게 보다 유연한 질의완화를 제공한다. 이러한 특성으로 인하여, KAH 기반 지식 추상화 데이터베이스는 다양한 유형의 질의 처리를 수행할 수 있게 해준다. 다시 말해서 타입 추상화 계층 기법이 근사적 선택과 개념적 선택 질의만을 해결하는 반면, KAH 기반 기법은 근사적 조인과 개념적 조인 질의를 포함한 좀더 복잡한 질의를 처리할 수 있다.

5. 결 론

본 논문에서 제시하는 지식 추상화 정보 데이터베이스는 모호한 질의를 처리할 수 있으며, 질의조건을 정확히 만족하는 해가 존재하지 않을 경우 근사해를 제공함으로써, 관계형 데이터베이스 시스템에 지능적인 인터페이스를 제공하는 것을 목표로 하고 있다. 이를 위하여, 지식 추상화 정보 데이터베이스는 다단계 지식 표현 프레임워크인 지식 추상화 계층을 사용하여 기반 데이터베이스로부터 유도된 의미론적 지식을 관리한다. 지식 추상화 계층은 기반 데이터베이스로부터 값의 추상화 지식을 추출해냄으로써 구축되어진다. 값의 추상화 지식은 응용 문제에 따라 구체적인 내용이 결정되기 때

문에, 설계자는 질의의 용도와 조직의 사용 패턴을 연구하거나, 데이터베이스로부터의 추상화 지식의 의미론을 자신 나름대로 해석하거나, 데이터 값들의 분포와 빈도를 분석하여 애트리뷰트 값들을 계층적으로 그룹화해 나가는 상향식 분석을 수행하게 된다. 그 결과를 바탕으로, 일반화와 세분화 관계를 이용하여 추상값과 세분값 사이의 이원적 관계를 구성할 수 있고, 전체적인 계층을 점진적으로 구축할 수 있다. 구축된 계층의 최상위 수준에 존재하는 값은 가장 추상적인 값인 반면, 최하위 수준에 존재하는 값은 가장 세분화된 값이 된다. 하나의 층(layer)에 존재하는 값들은 하나의 도메인에 포함되어 동일한 추상화 수준을 갖게 된다. 값의 추상화 계층이 구축된 다음, 그 계층의 각 추상화 수준에 대응하는 도메인들을 결정함으로써 도메인 추상화 계층이 바로 구축될 수 있다. 만약 개별 추상화 수준 사이에 값의 추상화 관계가 적절하게 확립되었다면, 수퍼 도메인과 서브 도메인 관계 또한 값의 추상화 계층과 병렬적으로 적절하게 유도될 것이다.

지식 추상화 계층을 DOMAIN_ABSTRACTION, VALUE_ABSTRACTION, 그리고 ATTRIBUTE_MAPPING 등의 세 개의 릴레이션으로 변환함으로써 지식 추상화 정보 데이터베이스가 구축된다. 이러한 세 릴레이션은 구조적으로 단순하지만, 협력적 질의 응답의 핵심적 과정에 해당하는 데이터 값의 일반화와 세분화를 수행하기에 충분한 정보를 포함하고 있으며, 또한 지식관리 측면에서 보다 유연한 구조를 갖는다. 지식 추상화 정보 데이터베이스는 근사적 질의와 개념적 질의라는 두 가지 유형의 모호한 질의 조건을 처리하게 된다. 근사적 질의에서는 질의 조건에 명시된 값과 유사한 값들을 포함하도록 확장된다. 따라서, 명시된 값과 유사한 값들이 집합적으로 비교되고 선택될 수 있다. 개념적 질의에서는 사용자가 기반 데이터베이스에 존재하는 데이터 값들의 세부사항을 모르더라도 개념적인 수준에서 질의 조건을 작성할 수 있다. 이러한 질의 조건들은 기반 데이터베

이스 세부사항에 대한 사용자의 친숙성 정도에 따라 상호 보완적으로 이용된다. 한편, 이러한 두 가지 유형의 질의 조건은 다시 선택조건뿐 아니라 조인 조건에 적용되어 근사적 선택, 개념적 선택, 근사적 조인, 개념적 조인 등의 네 가지 유형의 혼합 조건을 만들어 낸다. 사용자의 필요와 지식 추상화 계층에 대한 이해의 정도 따라, 네 가지 유형의 질의 조건은 개별 질의로 작성되거나, 혹은 선택적으로 혼합되어 하나의 질의에 작성될 수 있다. 그 결과 지식 추상화 정보 데이터베이스는 스스로 오류 처리를 수행할 수 있으며, 좀 더 지능적인 특성을 갖게 된다. 지식 추상화 정보 데이터베이스의 유용성을 검증하기 위한 프로토타입 시스템이 구현되었으며, 현재 인사 정보 데이터베이스 시스템에 적용되어 시험 중에 있다.

본 논문을 통하여 근사적 질의를 보다 쉽게 수행할 수 있게 되었고, 협력적 질의응답의 투명성도 훨씬 향상되었다. 첫째, 사용자는 KAH 이름을 전부 기억할 필요없이 기반 데이터베이스의 릴레이션만을 알고 있어도 협력적 질의응답을 수행할 수가 있다. 지식 추상화 정보 데이터베이스는 ATTRIBUTE_MAPPING 릴레이션을 이용하여 질의문에 나타난 애트리뷰트를 스스로 분석하고 해당 도메인을 찾아냄으로써 질의처리에 필요한 KAH를 참조하게 된다. 그것은, 하나의 도메인은 반드시 하나의 KAH에만 유일하게 존재하기 때문에, 일단 도메인을 알게되면 그 도메인이 존재하는 KAH를 자동적으로 선택할 수 있기 때문이다. 둘째, 협력적 질의 응답의 과정에서, 사용자들은 질의 변환의 세부적인 단계를 기억하거나 이해할 필요가 없다. 질의 일반화와 세분화는 미리 정의된 자동적인 절차에 따라 진행될 수 있기 때문이다. 따라서, 지식 추상화 정보 데이터베이스는 질의 변환 과정에서 겉으로 드러나지 않게 되며, 사용자는 질의의 의도를 스스로 이해하여 데이터베이스로부터 풍부한 정보를 제공해주는 지능적 정보시스템을 이용할 수 있게 된다.

앞으로의 연구에서는, 관계 해석을 바탕으로

KAH와 질의 완화 과정에 대한 형식론을 확립할 것이고, 그를 바탕으로 본 논문에서 제시한 일반화/세분화와 같은 질의 완화 과정이 완전하고 정확한 것임을 증명하게 될 것이다. 또한 근사적 조건과 개념적 조건 모두가 퍼지 조건과 상당한 공통점을 가지고 있기 때문에 퍼지 집합 이론을 포함시켜 KAH의 의미론을 보다 풍부하게 하고 지식 추상화 정보 데이터베이스의 지능적인 성격을 향상시키고자 한다.

참 고 문 헌

- [1] Buckles, B. P. and F. E. Petry, A Fuzzy Representation of Data for Relational Databases, *Fuzzy Sets Systems*, Vol.7, No.3 (1982), pp.213-226.
- [2] Cai, Y., N. Cercone and J. Han, "Attribute-Oriented Induction in Relational Databases," *Knowledge Discovery in Databases*, AAAI Press/The MIT Press, California, 1993.
- [3] Chen, Q., W. Chu and R. Lee, "Providing Cooperative Answers via Knowledge-Based Type Abstraction and Refinement," *Methodologies for Intelligent Systems*, ICAIT, Knoxville, 1990.
- [4] Cholvy, L. and R. Demolombe, "Querying a Rule Base," *Proceedings of the First International Conference on Expert Database System*, (1986), pp.365-371.
- [5] Chu, W., H. Yang, K. Chiang, M. Minock, G. Chow and C. Larson, "CoBase: A Scalable and Extensible Cooperative Information System," *International Journal of Intelligence Information Systems*, Vol.6 (1996).
- [6] Chu, W., H. Yang and G. Chow, "A Cooperative Database System(CoBase) for Query

- Relaxation," *Proceedings of the Third International Conference on Artificial Intelligence Planning Systems*. Edinburgh, 1996.
- [7] Chu, W. and Q. Chen, "A Structured Approach for Cooperative Query Answering," *IEEE Transactions on Knowledge and Data Engineering*, Vol.6, No.5(1994), pp.738-749.
- [8] Chu, W., R. Lee and Q. Chen, "Using Type Inference and Induced Rules to Provide Intensional Answers," *Proceedings of the Seventh International Conference on Data Engineering*, Kobe, Japan, (1991), pp.396-403.
- [9] Chu, W., Q. Chen and R. Lee, "Cooperative Query Answering via Type Abstraction Hierarchy," *S.M. Deen, editor, Cooperative Knowledge Base System*, Elsevier Science Publishing Co., Inc., North-Holland, 1991.
- [10] Cuppens, F. and R. Demolombe, "Cooperative Answering : A Methodologies to Provide Intelligent Access to Databases," *Proceedings of the Second International Conference on Expert Database Systems*, (1989), pp.621-643.
- [11] Godfrey, P., J. Minker and L. Novik, "An Architecture for a Cooperative Database System," *Proceedings of the 1994 International Conference on Applications of Databases*, (1994), pp.3-24.
- [12] Hemerly, A., M. Casanova and A. Furtado, "Exploiting User Models to Avoid Misconstruals," *Nonstandard Queries and Nonstandard Answers*, Oxford Science Publications, London, 1994.
- [13] Ichikawa, T., "ARES : A Relational Database with The Capability of Performing Flexible Interpretation of Queries," *IEEE Transactions on Software Engineering*, Vol.SE-12, No.5(1986), pp.624-634.
- [14] Jain, A. K. and R. C. Dubes. *Algorithms for Cluster Analysis*, Prentice Hall, Englewood Cliffs, 1988.
- [15] Minock, M. J. and W. Chu, "Explanation for Cooperative Information Systems," *Proceedings of the Ninth International Symposium on Methodologies for Intelligent Systems*, (1996), pp.264-273.
- [16] Motro, A., "VAGUE : A User Interface to Relational Databases that Permits Vague Queries," *ACM Transactions on Office Information Systems*, Vol.6, No.3(1988), pp.187-214.
- [17] Motro, A., "FLEX : A Tolerent and Cooperative User Interface to Databases," *IEEE Transactions on Knowledge and Data Engineering*, Vol.2, No.2(1990), pp.231-246.
- [18] Motro, A., "Intensional Answers to Database Queries," *IEEE Transactions on Knowledge and Data Engineering*, Vol.6, No.3(1994), pp.444-454.
- [19] Prade, H. and C. Testemale, "Generalizing Database Relational Algebra for the Treatment of Incomplete or Uncertain Information and Vague Queries," *Information Science*, Vol.34, No.2(1984), pp.115-143.
- [20] Scheer, A., *Enterprise-Wide Data Modeling : Information Systems in Industry*, Springer-Verlag, 1989.
- [21] Shum, C. D. and R. R. Muntz, "An Information-Theoretic Study on Aggregate Responses," *Proceedings of the Fourteenth International Conference on Very Large Databases*, Morgan Kaughmann, California, (1988), pp.479-490.
- [22] Turban, E., *Decision Support and Expert Systems : Management Support Systems*,

- Macmillan, New York, 1988.
- [23] Ullman, J. D., *Database and Knowledge-Base Systems, Vol 1*, Computer Science Press, Maryland, 1987.
- [24] Vrbsky, S. V. and W. S. Liu, "APPROXIMATE-A Query Processor that Produces Monotonically Improving Approximate Answers," *IEEE Transactions on Knowledge and Data Engineering*, Vol.5, No.6(1993), pp.1056-1068.
- [25] Zemankova, M. and A. Kandel, "Implementing Imprecision in Information Systems," *Information Science*, Vol.37, No.1(1985), pp. 107-141.