멀티데이터베이스 환경 하에서의 Description Logic을 이용한 의미상 질의 최적화*

이태웅[0] 권주홈 백두권
고려대학교 컴퓨터학과 소프트웨어 시스템 연구실
{minote[0], jkweon, haik}@software.korea.ac.kr

Semantic Query Optimization Using Description Logic in Mutidatabase Systems

Tae-Woong Lee[0] Ju-Hum Kweon Doo-Kwon Baik

Software System Lab. Dept of Computer Science & Engineering,
Korea University

요 약

물류 공급 관리 시스템과 같은 정보 통합 시스템은 분산되어 있는 데이터베이스들에 대해서 정보를 통합하여 사용자에게 보여준다. 이러한 정보 통합 시스템은 전역 질의를 생성하고 지역 질의로 변환하여 실행하기 전에 질의를 최적화할 필요성이 있다. 그런데, 단일데이터베이스 시스템에서의 질의 최적화 기법은 멀티데이터베이스 시스템에서 사용하기에는 부적절하다. 이는 분산된 데이터베이스 환경에서 오는 높은 연결 오버헤드, 높은 계산 시간, 데이터의 중복성 뿐만 아니라 의미 이질성 문제 때문에 기존의 최적화 방법은 사용하기가 어렵다. 이를 해결하기 위해서 의미상 질의 최적화 방법이 연구되어 왔다. 의미상 질의 최적화는 전역 질의보다 더 효과적으로 응답하고 의미상으로 동등한 질의로 변환하기 위해서 의미상 지식을 사용한다. 본 논문에서는 정보 통합 시스템에서 Description Logic(DL)을 이용하여 의미상 지식으로 사용할 지식 기반을 표현하고 이를 바탕으로 추론화된 지식을 이용하는 의미상 질의 최적화 방식을 제시한다.

1. 서 론

정보 통합 시스템은 분산되어 있는 이질적 데이터베이스들에 대해 데이터를 통합하여 사용자에게 전역뷰로 정보를 제공하는 멀티데이터베이스 시스템이다.[1] 이러한 시스템은 사용자에게 실제로 하나의 데이터베이스처럼 보여줄 수 있도록 통합된 인터페이스를 가지고 있고 전역 질의를 사용자에게 받아서 실행하게 된다.

이 전역 질의는 각각 데이터베이스의 로컬 스키마와 무관하게 작성된 질의이기 때문에 각각 데이터베이스의 스키마에 맞는 지역 질의로 변환시켜줄 필요가 있다. 이러한 전역 질의를 로컬 질의로 변환시키는 연구는 많이 이루어지고 있다.

사용자가 전역 질의가 생성되면 각각 데이터베이스의 스키마에 맞는 로컬 질의로 변환시키고, 각 로컬질의를 실행시킬 수 있도록 질의 플랜을 수립하게 된다. 이러한 질의 플랜에 있어서 또다른 필요한 과정은 질의 최적화 과정이다. 질의 최적화는 질의 수행에 있어서 모든 데이터베이스를 접근하여 처리하지 않고 보다 효율적으로 수행할 수 있도록 질의 최적화를 함으로써 질의 응답 시간을 줄일 수 있게 한다.

그러나, 로컬 데이터베이스의 질의 최적화와 달리 멀티데이터베이스에서의 질의 최적화는 분산된 데이터의 특성 때문에 기존 방법과 다른 방법이 요구하게 된다. 멀티데이터베이스 환경에서는 접근해야할 데이터가 서로 분산되어 있고, 스키마들 간의 이질성 문제를 가지고 있다. 따라서 높은 연결 오보헤드, 높은 계산 시간, 데이터의 중복성 뿐만 아니라 의미 이질성이 분산데이터베이스에서의 질의 최적화를 어렵게 하고 있다.

이러한 문제점을 해결하고 전역 질의를 각각 데이터베이스에 맞는 지역 질의로 최적화하기 위해서 의미상 질의 최적화 기법을 활용하고 있다.[2] 의미상 질의 최적화는 본래의 질의와 동등한 의미를 가지며 보다 효율적으로 처리할 수 있는 질의로 변환하는 과정을 말한다. 이러한 의미상 질의로 변환하기 위해서 의미상 정보를 필요하게 되는데 기존 연구에서는 룰 기반(Rule-Based)의 의미상 정보를 이용하였다.[4][5][6] 본 논문에서는 이와 달리 지식 기반(Knowledge-Based)으로 한 의미상 질의 최적화를 위해서 본 논문에서는 Description Logic을 이용하여 지식 기반을 구축하고 추론하는 방식으로 의미상 질의 최적화하는 기법을 제시한다.

DL을 이용한 의미상 질의 최적화는 질의 최적화에 필요한 지식 기반을 최소화시키며, 의미를 전역 영역과 지역 영역으로 분리함으로써 지역데이터베이스의 투명성을 보장한다.

2. 관련 연구

2.1 의미상 질의 최적화

두개의 질의가 무결성 제약 조건을 만족하는 데이터 모델 안에서 같은 결과를 반환된다면, 이 두 질의는 의미상으로 동등하다고 판단할 수 있다. 따라서, 의미상 질의 최적화이란 무결성 제약 조건을 표현한 의미상 정보를 이용하여 질

* 본 연구는 한국과학재단의 지원으로 수행되었음(2002~2003년)

의를 최적화하는 프로세스이다. 즉, 같은 결과를 가져오는 여러 의미상 질의 중에서 가장 낮은 비용을 가진 질의를 선택하는 과정이라고 할 수 있다.

의미상 질의 최적화 과정에서 본래 질의와 같은 결과를 가지지만 보다 효과적인 비용으로 실행할 수 있도록 다른 형태를 가진 질의로 변환하기 위해서 의미상 지식(Semantic Knowledge)을 이용하게 된다. 이 의미상 지식을 이용하는 방식에 대해서 룰 기반(Rule-Based) 표현이 많이 연구되고 있다.

2.2 Description Logic

Description Logic(DL)은 지식을 표현하기 위해 정형 기법을 이용한 로직의 한 종류이다. 이는 한 도메인 영역을 표현하기 위해서 개념(Concept)과 역할(Role)이라는 용어를 사용한다. 크게 개념은 각 클래스를 모델링하고 역할은 클래스간의 관계를 모델링한다. 이는 또다시 단순 개념(atomic concept)과 복합 개념(complex concept), 단순 역할(atomic role)과 복합 역할(atomic role)로 나눌 수 있는데 전자는 단순히 이름으로 표현한 것이고, 후자는 적절할 구성물을 통해서 이루어진 복합체이라고 할 수 있다.

DL로 표현한 지식 기반(Knowledge Base)은 두 개의 컴포넌트로 구성되며 각각 TBox와 ABox이라고 불러진다. TBox에서는 개념과 역할의 일반적인 프로퍼티로부터 전역적이고 포괄적인 선언(inclusion assertion)을 선언하는 반면에, ABox에는 개개 개념의 인스턴스(instance assertion)를 선언한다. 따라서, TBox는 일반적인 지식을 표현하고 ABox는 특수적인 지식을 표현하여 지식 기반을 구축할 수 있다.

이러한 지식 기반을 바탕으로 지식 간의 상호 관계를 명확히 할 수 있고 DL의 추론 기법을 이용하여 의미상 질의를 추론할 수 있는 기법으로 활용할 수 있다.[3]

3. Description Logic으로 표현한 지식 기반(Knowledge Base)

의미상 질의 최적화 과정에서 본래 질의와 의미상으로 동등하면서도 더 효과적으로 대답하는 질의로 변환하기 위해서 의미상 지식을 사용하게 된다. 이 의미상 지식에서 도메인 영역의 사실, 스키마 정보, 또는 제약 조건 등을 표현하고 있다. 의미상 지식를 표현하고 이용하는 시스템은 지식 기반 시스템(Knoewledge Base System)이라고 하며, 이러한 의미상 지식은 온톨리지와 같은 지식 표현 언어(Knowledge presentation Langauge)으로 표현한다. 이 의미상 지식을 표현하기 위해서 기존 의미상 질의 최적화는 룰 기반 언어을 많이 이용하고 있으나, 본 논문에서는 Descripton Logic 이라는 정형 언어으로 지식을 표현한다. 이 정형 언어는 또다른 지식을 추론할 수 있는 장점을 가지고 있다.

DL에서의 지식은 TBox와 ABox로 나누어서 표현한다. TBox는 도메인 영역 레벨(전역 데이터베이스) 측면에서 개념에 대한 정의를 선언하게 되며, ABox는 지역 데이터베이스 레벨 측면에서 스키마 정보, 또는 인스턴스에 대한 정의를 선헌한다. 따라서, TBox에서는 도메인 지식, 제약 조건 등을 가지고 있으며, ABox는 로컬 스키마에서의 테이블, 열, 메타데이터에 관한 지식을 가지고 있다. TBox와 ABox로의 경계는 의미상 질의 최적화하는데 있어서 명확한 기준점을 제공할 수 있게 한다.

4. Description Logic을 이용한 주요 의미상 질의 최적화 방법

지식 기반을 표현하기 위해 TBox는 개념 상의 선언을 정의하고 ABox는 인스턴스의 선언을 정의한다. 본 논문에서는 이러한 DL로 이루어진 지식 기반을 이용하여 다음과 같은 순서로 의미상 질의 최적화를 수행한다.

Step 1. 주어진 질의의 의미를 정의한 Tbox 개념을 추론한다.
Step 2. TBox와 상응하는 ABox의 인스턴스를 추론한다.
Step 3. Step 1,2를 기반으로 지역질의로 변환한다.

기본적으로 위의 알고리즘으로 진행되나, 간혹 Step1이나 Step 2 어느 하나가 생략된 채로 진행될 수 있다. 즉, 개념을 참조할 필요없이 인스턴스만 추론할 수도 있다. 의미상으로 표현된 기본적으로 질의를 최적화하기 위해서 지식 베이스와 DL을 이용한 추론 기법을 사용한다. 주요한 의미상 질의 최적화 방법을 설명하기 위해서 물류 체인 관리 도메인에 관한 예제를 가지고 설명한다.

4. 1 모순 감지

주어진 질의에서 지식 베이스 또는 무결성 조건에 의해서 모순이 발견된다면, 데이터베이스를 접근할 필요가 없이 즉시 널 값을 반환한다. 이 모순 감지는 지식 기반에 있는 의미 해석을 통해서 알 수 있다. 다음과 같은 예를 살펴보자.

*Select * From Retailer Where suppliedItem > 300;*

이라는 전역질의를 생성하고, TBox는 다음과 같이 구성되어 있다고 하자.

Wholesaler ⊏ ∃Supplied.Company ⊓ (≤ 100 Supplied.Item)
Supplied.Company ⊏ ∀Wholesaler ⊔ ∀Retailer
Retailer ⊏ ∃Supplied.Company ⊓ ¬Wholesaler

지식 기반에서 다음과 같은 지식을 표현하고 있다. 소매상은 도매상이 아닌 공급회사이며, 즉 도매상은 100개 이상되는 상품을 공급한다. 따라서, 이 지식 기반을 이용하여 소매상은 상품이 100 개 이하인 공급하는 회사임을 추론할 수 있다. TBox에서 추론 한 뒤, 본래 질의와 비교해보면, 질의가 모순된 질의임을 쉽게 알 수 있다. 따라서, ABox 과정으로 진행하거나 지역 데이터베이스의 접근할 필요없이 결과값을 널값으로 처리하게 된다.

4. 2 질의 재작성

질의 재작성은 의미상 지식을 이용한 질의 변환을 의미한다. 즉, 모순 감지를 시도하여 의미상 모순이 없다면, ABox에서 추론하여, 질의 변환이 일어나게 된다. 다음과 같은 질의를 살펴보자. 모든 회사 중에서 공급하는 제품이 200 개 이상인 회사를 검색하라는 전역 질의를 살펴보자. 다음과 같이 표현할 수 있다.

Select * From Company Where suppliedItem > 200

이 전역 질의가 실행된다면 모든 회사 데이터베이스를 다 검색하게 된다. 따라서, 모든 회사 데이터베이스를 검색하기에 비용과 시간이 너무 크다. 따라서, 다음과 같은 의미상 질의 최적화 과정을 진행한다. 위의 5.1의 TBox의 지식을 이용하면, Company이며 Supplied.Item이 100이상인 회사는 도매상임을 추론할 수있다. 따라서 공급하는 제품이 100개 이상인 공급회사는 도매상이라는 정의을 이용해서, 본 질의에서 회사 대신 도매상이라는 개념으로 대체하게 된다.

Select * From Wholesaler Where suppliedItem > 200

이는 모든 회사 데이터베이스를 접근할 필요 없이 단지 도매상 데이터베이스만 접근하기만 하면 된다. 따라서, 전역 질의보다 접근 액세스 비용이 대폭 줄일 수 있게 된다. 이 도매상 데이터베이스의 정보를 알기 위해서 ABox를 사용할 수 있는데 ABox는 각 데이터베이스의 인스턴스를 포함하고 있으므로, 이를 이용하여서 도매상인 데이터베이스를 쉽게 찾을 수 있다. 다음과 같은 ABox가 있다고 가정하자.

company(A) company(B)
supplied.company(A) supplied.company(B)
supplied.item(A, 150) supplied.item(B, 80)

A와 B이라는 인스턴스는 공급회사 데이터베이스임을 ABox에서 알수 있고 TBox에서의 Wholesaler ⊏ ∃ Supplied.Company ⊓ (≤ 100 Supplied.Item) 이라는 지식 표현에서 A 공급회사는 도매상이라는 사실을 추론할 수 있다. 반면에, B 공급회사는 도매상이라는 조건을 만족하지 못하므로 도매상이 아님을 알 수 있다. TBox에서 도매상 아닌 공급회사는 소매상이라는 정의에 의해서 B회사는 소매상이라는 것을 쉽게 추론할 수 있다. 따라서, 본 질의도 다음과 같은 지역질의로 변환하게 된다.

Select * From A Where suppliedItem > 200

따라서, 모든 데이터베이스를 접근할 필요성이 있는 본 질의가 TBox, ABox의 추론을 통해서 도매상 데이터베이스만 접근하는 질의로 변환하게 된다. 이 질의는 본 질의와 의미상 손실없이 같은 결과를 반환하게 된다. 따라서, 본래 질의보다 빠른 질의 최적화를 수행할 수 있게 된다.

5. 기존 룰 기반 질의 최적화 기법과 MBL 기반 질의 최적화 기법과의 비교

기존의 의미상 질의 최적화 기법은 지식을 표현하기 위해 룰 기반을 이용하였다. 이 룰 기반은 IF-THEN 문을 빈번한 사용에 의한 오버헤드가 큰 편이다. 또한 가능한 모든 규칙을 표현할 필요가 있기 때문에 지식 표현을 하기 위해서 보다 많은 메모리와 저장 공간이 필요하게 된다. 따라서 본 논문은 DL을 이용하여 추론 기법에 의한 지식 표현을 함으로써 IF-THEN 문의 사용을 최소화시키며, 룰 기반 지식 표현보다 적은 메모리 공간을 소모하게 된다. 따라서 룰 기반 의미상 질의 최적화보다 빠른 질의 최적화 수행이 가능하다. 또한, 룰 기반 의미상 질의 최적화와 달리 TBox와 ABox로 분류, 개념과 인스턴스의 계층화하여 룰 기반 추론에서 생겨날 수 있는 순환적 오류의 가능성를 줄일 수 있다. 순환적 오류이란 추론이 무한히 반복되는 경우를 의미한다. 룰 기반에서 같은 의미를 반복되어 추론할 가능성이 있다. 따라서 DL을 이용한 지식표현에서는 이런 무한 반복 가능성을 방지하고자, 추론은 TBox에서 시작하여 ABox에서 종결할 수 있게끔 한다. 즉 ABox에서 추론된 인스턴스를 찾을 수 있게 되면 추론을 끝나게 된다.

6. 결론 및 향후 과제

정보 통합 시스템 환경하에서 DL로 도메인 의미 정보를 포함하고 있는 지식 기반을 표현하고 이를 바탕으로 한 추론 기법을 이용하여 의미상 질의 최적화를 할 수 있다. 향후 멀티데이터베이스의 통합을 위한 물류 체인 관리 환경 도메인 바탕으로 전역 지식 모델을 DL으로 표현하는 방식을 연구한다. 그리고 이 환경에서 의미상 질의 최적화하기 위한 추론 기법을 보다 상세히 기술하고 질의 최적화 시뮬레이션의 프로토타입을 구축하여 평가할 필요가 있다.

참고문헌

[1] G.Wiederhold, "Mediators in the architecture of future information systems", IEEE Computer, 25(3):38-49, March, 1992.

[2] Chun-Nan Hsu and Craig A. Knblock "Semantic Query Optimization for Query Plans of Heterogeneous Multidatabase Systems", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL.12, No..6, November 2000.

[3] F.M.DONINI, M.LENZERINI, D.NARDI, A.SCHAERF "Reasoning in Description Logics", 1999.

[4] CLEMENT T.YU and WEI SUN "Automatic Knowledge Acquisition and Maintenance for Semantic Query Optimization", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, Vol.1, No.3, September 1989.

[5] MICHAEL SIEGEL, EDWARD SCIORE and SHARON SALVETER "A Method for Automatic Rule Derivation to Support Semantic Query Optimization", ACM Transactions on Database Systems, 17(4):563-600, Dec. 1992.

[6] Chun-Nan Hsu and Craig A. Knblock "Rule Induction for Semantic Query Optimization", Proc. 11th International Conference on Machine Learning,112-120,1994.