SPARQL-to-SQL: 질의 성능 향상을 위한 캐시 관리자

김석현, 이상원
성균관대학교 전자전기컴퓨터공학과
e-mail : seokhyon@ece.skku.ac.kr

SPARQL-to-SQL: Cache Manager for Advanced Query Efficiency

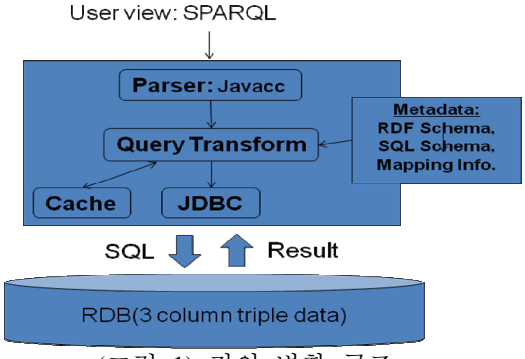
Seok-hyun Kim, Sang-Won Lee
Department of Electrical and Computer Engineering, Sungkyunkwan University

요 약

시맨틱(Semantic) 온톨로지(Ontology)에서 SPARQL 질의언어는 W3C 로부터 표준으로 제정된 이후부터 활발히 연구 되고 있다. 그리고 현재까지 온톨로지 기반 어플리케이션 개발이 다방면으로 진행되어 왔는데, 현재 개발된 온톨로지 어플리케이션들은 시맨틱 데이터 저장 및 질의 처리가 파일시스템 기반 및 데이터베이스 기반 방식으로 나누어 진다. 그 중 데이터베이스 기반 방식은 최근부터 연구가 진행되어 왔고 실제 개발된 어플리케이션도 있지만, 아직 질의 최적화 기술에 대해서는 개선할 수 있는 여지가 많다. 따라서 본 논문에서는 관계형 데이터 베이스를 기반한 온톨로지 데이터 저장 및 질의 처리 방법에서 캐시를 이용한 질의 속도 향상 방법을 제시 하도록 하겠다. SPQARQL 에서 변환된 SQL 질의 수행시 그 결과를 캐시하고, 후속 SQL 질의를 이전 질의와 비교하여 이전 SQL 질의와 일치하거나 그 결과가 포함 될 경우 캐시된 결과를 사용해 쿼리 속도를 향상 시킬 수 있다.

1. 서론

본 연구는 SPARQL-to-SQL 질의 변환 연구[1]에 데이터 베이스 캐시 관리자 연구[2]를 적용 하여 보다 효율적인 질의 처리 방법을 제시한다. [1]의 연구는 온톨로지 데이터를 관계형 데이터베이스에 저장한 후 SPARQL 질의를 SQL 질의로 변환하여 데이터 베이스에서 질의하는 방법을 제시 한다. 그리고 [2]의 연구는 특정한 SQL 질의 환경에서 캐시를 이용해 질의 속도를 향상 시킨다. 따라서 본 연구에서는 [1]의 SPARQL-to-SQL 질의 변환 연구와 [2]의 캐싱 방법을 연계한 질의 최적화 방법을 제시한다.



(그림 1) 질의 변환 구조

SPARQL 질의를 SQL 질의로 변환 수행하는 과정에서 캐시 사용이 적절한 이유는 다음과 같다.

1) SPARQL에서 변환된 SQL질의는 같은 형식의 패턴을 따르며 이 질의들은 일정한 정규화 형태로 표현이 가능하다.
2) SPARQL에서 변환된 SQL 질의는 대개 여러 횟수의 조인(Join) 연산을 필요로 하므로 질의 비용이 크다.

일반적으로 캐시를 사용할 경우, 캐시의 적중률(Hit-Ratio)이 높을수록 캐시 사용이 더욱 빈번해져 질의 수행에 효율적이다. 또한 SPARQL 을 SQL 로 변환하여 질의를 수행 할 경우, 변환된 SQL 질의는 항상 같은 형식을 가지게 되며, 따라서 먼저 수행된 질의와 이 후에 수행될 질의가 같거나 비슷해질 확률이 높아져 적중률이 증가하게 된다. 따라서 1)의 이유에서 설명한 것처럼 질의된 결과 캐시의 재사용률이 높아진다.

한편, 질의 시 1 회의 질의 수행 비용이 높다면, 질의를 수행하지 않고 미리 저장된 질의 결과를 사용할 수 있는 캐시를 사용하는 것이 효율적인데, 2)의 이유와 같이 SPARQL 에서 변형된 SQL 질의는 보통 수 차례 조인 연산을 필요로 하기 때문에 한번의 질의에 필요한 수행 비용이 크다. 따라서 이러한 질의 환경에서 캐시를 사용하면 질의의 속도 향상에 더욱 효율적이다.

본 논문에서는 SPARQL-to-SQL 질의 변환 연구와 관계형 데이터베이스에서의 캐시 사용에 대한 연구를 소개한다. 그리고 캐시를 사용하기 앞서 미리 저장된 질의 캐시와 이 후에 질의 결과의 포함 관계 여부를

* 본 연구는 지식경제부 및 정보통신산업진흥원의 대학 IT 연구센터 지원사업의 연구결과로 수행되었음 (NIPA-2009-(C1090-0902-0046))

구분하는 질의 포함 관계 판별에 대해 설명한다.

2. 관련 연구

2.1 SPARQL-to-SQL 질의 변환

현재 대부분의 온톨로지 시스템들의 경우, 파일 시스템 기반 데이터 저장 방식을 택하고 있다. 그러나 파일시스템을 이용해 새로운 시스템을 별도로 개발하는 것 보다는, 30 여 년간 발전되고 검증된 관계형 데이터 베이스를 기반으로 한 온톨로지 저장 및 질의/추론 시스템을 개발하는 것이 효과적인 접근 방법일 것이다. 이에 따라 연구 [1]은 SPARQL-to-SQL 질의 변환에서 온톨로지 데이터의 질의/추론과 같은 기능들을 기존의 관계형 데이터베이스를 이용해 질의/추론 하는 방법을 제시한다.

2.2 데이터 캐시 관리자

SPARQL-to-SQL 질의 변환 과정에서 캐싱을 사용한 연구는 [2]의 연구를 참고로 진행하였다. [2]의 연구는 SQL 질의 결과와 효율적인 캐시 교환 알고리즘을 제시 하는데, SPARQL-to-SQL 변환 연구에서의 SQL 질의 역시 [2]의 연구 환경에서 나오는 SQL 과 비슷한 형태를 따르기 때문에 [2]의 이론을 적용하여 질의결과를 캐싱 하는 것이 효율적이다.

3. 질의 포함 관계 판별

3.1 정규화 형태

SPARQL 에서 변환된 SQL 질의는 다음과 같은 정규화 형태를 따른다.

```
SELECT project-list
FROM   table t1, table t2, … , table tN
WHERE join-condition AND select-condition
```

(그림 2) 정규화 표현

(그림 2)의 정규화 표현에서 'project-list'는 원하는 열(Column)을 지정하여 그 결과를 출력할 수 있게 한다. 그리고 하나의 table 을 t1, t2, …, tN 으로 여러 번 셀프조인(Self-join) 한다. 'join-condition'은 두 테이블을 조인(Join) 하기 위한 조건이다. 그리고 'select-condition'은 여러 행(Tuple) 중에서 의미 있는 행을 가려내기 위한 조건이다. 'select-condition'에는 '=' 비교 연산자만 사용되며, 비교 연산자는 하나의 속성만을 가진다.

SPARQL 에서 변환된 SQL 이 위와 같은 일정한 형식의 정규화 형태를 따르기 때문에 사전에 캐시된 질의와 수행중인 질의의 관계 비교 작업이 매우 수월해 진다. 또한 중복되는 질의가 많아 캐시의 재사용률도 높아질 것이다.

3.2 질의 비교와 질의 종속

SPARQL 에서 변환되는 SQL 질의 수행시, 질의 캐시 재사용은 다음과 같은 질의 비교 절차를 따른다. 임의의 SPARQL 을 변환한 SQL 질의를 Q2 이라고 하자. 그런데 만약 미리 수행된 질의 Q1 의 결과가 질의 Q2 의 결과를 포함한다면, Q2 의 결과는 Q1 의 결과를 가공한 Q1' 에서 도출해 낼 수 있다

질의 포함 관계: Q2 ⊂ Q1

```
Q1: 모든 학생
SELECT DISTINCT t2.obj name
FROM table t1, table t2
WHERE t1.subj=t2.subj
AND t1.pred='rdf:type'
AND t1.obj='http://example.org#GraduateStudent'
AND t2.pred='ub:name'
```

```
Q2: 모든 학생 중 Mathmatics 과목 수강생
SELECT DISTINCT t2.obj name
FROM table t1, table t2
WHERE t1.subj=t2.subj
AND t1.pred='rdf:type'
AND t1.obj='http://example.org#GraduateStudent'
AND t1.pred='ub:takesCours'
AND t1.obj='http://www.University.edu/Mathmatics'
AND t2.pred='ub:name'
```

사전에 질의된 Q2 의 결과가 캐시되어 있고, 이후에 질의되는 Q1 을 수행시, Q1 와 Q2 를 비교하여 상관관계가 있을 경우, 이미 캐시된 Q2 의 질의 결과를 가공하여 Q1 의 질의 결과를 얻어낼 수 있으므로 질의 수행 시간을 향상시킬 수 있다. 위의 예와 같이 데이터 스키마(Schema)가 표준적 스키마를 따르고, 이에 수반하는 질의들이 대개 일정한 형태로 표현될 수 있을 경우 질의 포함관계 분석이 효율적임을 보여주는 연구는 다음과 같다[3].

5. 결 론

본 연구는 처음으로 SPARQL 을 SQL 로 변환 하여 질의하는 환경에서 캐시를 사용하는 연구로써, 캐시를 적용하는 과정에는 아직 더 많은 기술적인 연구가 필요하다. 향후 연구 과제로서 캐시된 데이터의 관리에 대한 연구가 필요한데, 캐시된 데이터의 업데이트시 이를 정적인 방법, 혹은 동적인 방법으로 업데이트 할 것인지에 대한 연구가 추가로 필요할 것이다.

이러한 질의 변환 과정에서 캐시 사용 시도는 아직 연구 단계에 놓여있는 SPARQL-to-SQL 질의 변환 방법에 있어 질의 수행 최적화를 이룰 수 있는 가능성을 열어줄 것이다.

참고문헌

[1] 김석현, 이상원, "관계형 데이터베이스를 이용한 온톨로지 데이터 저장 및 SQL 을 이용한 질의 처리 방법", 정보과학회 춘계 학술대회, 2009

[2] 심준호, "데이터웨어하우스 환경에서의 질의 처리 성능 향상을 위한 캐시 관리자", 정보과학회논문지 데이터베이스 제 30 권 4 호, 2003

[3] J.Shim, P. Scheuermann, and R. Vingralek, "Dynamic Caching of Query Results for Decision Support Systems", Proc. Of the 11th International Conference on Scientific and Statistical Database Management, IEEE Computer Society, 1999.