# RDQL2SQL 기반의 효율적인 RDQL 질의 처리

김학수° 손진현
한양대학교 컴퓨터공학과
{hagsoo°, jhson}@cse.hanyang.ac.kr

# Efficient RDQL Query Processing based on RDQL2SQL

Hak Soo Kim° Jin Hyun Son
Department of Computer Science and Engineering, Hanyang University

요 약

최근 시맨틱 웹에 대한 관심이 증가하면서 W3C 표준으로 규정된 시맨틱 웹 온톨로지 언어(RDF, RDFS, OWL 등) 기반의 관련 기술에 대한 연구가 활발히 진행되고 있다. 그 중에서 시맨틱 웹 온톨로지 언어로 기술된 문서의 저장, 관리, 질의처리 기법에 대한 연구가 주목을 받고 있다. 이에 본 논문에서는 온톨로지 데이터에 대한 표준 질의 언어인 RDQL 을 기반으로 RDQL 질의를 효율적으로 처리하는 고성능 RDQL 질의 처리 엔진을 개발한다.

본 논문에서 제안하는 RDQL 질의 처리 엔진은 RDQL 질의를 대응하는 SQL 질의로 변환함으로써 기존의 관계형 데이터베이스 질의 처리 엔진(SQL 질의 처리 엔진)을 그대로 사용할 수 있다. 이 과정에서 메모리 사용량과 데이터베이스 접근을 최소화하는 고성능 RDQL 질의 처리 엔진을 개발한다. 궁극적으로 이러한 RDQL 질의 처리는 실시간 처리가 요구되는 로봇 환경뿐만 아니라 시맨틱 웹 애플리케이션에서 널리 활용될 수 있다.

## 1. 서론

온톨로지([1])는 시맨틱 웹의 등장과 함께 시맨틱 웹의 핵심 기술로서 연구되고 있는 주요 분야들 중의 하나이다. 지금까지 연구되어온 온톨로지 개념은 지식 정보를 표현하는데 가장 효율적이라고 고려되기 때문에 시맨틱 웹에서도 이러한 개념을 활용하여 웹 상에서의 시맨틱 정보 처리의 기반 기술로서 활용하고 있다. 이에 W3C 는 시맨틱 네트워크와 프레임 기반의 RDF (Resource Description Framework)([2])를 시맨틱 데이터 표준 표현 언어로 규정하고 있다. RDF 는 웹에 있는 임의의 자원을 (주어, 서술어, 목적어)의 세 원소로 구성된 트리플로 표현하며, 여기서 주어와 목적어는 그래프 노드로, 서술어는 에지(Edge)로 표현되는 RDF 그래프 모델을 구성할 수 있다. 결국, 웹 상에 존재하는 모든 자원들은 RDF 로 기술된 그래프 모델로 표현할 수 있다. 한편, RDF 는 단순히 존재하는 웹 자원을 표현하는 언어로 스키마 정보, 자원 사이의 관계 정보, 제약 정보 등에 대한 표현이 미약하다. 이에 W3C 에서는 RDF 기반의 온톨로지 언어를 완성하기 위해 RDFS([3])와 OWL([4])을 추가적으로 제정하였다. 그러므로, 본 논문에서는 시맨틱 데이터 표현 언어로 OWL 를 고려한다.

온톨로지 데이터를 대상으로 하는 시맨틱 질의 언어로 지금까지 약 20 여 개가 제안되었으며, 이 중에서 현재 W3C 에서 표준화 과정에 있는 시맨틱 질의 언어로는 RDQL([7])과 SPARQL([8])이 있는데 모두 SQL-유사 질의 언어이며 본 논문에서는 RDQL 을 대상으로 하는 시맨틱 질의 처리 엔진을 개발 하고자 한다.

논문의 구성은 다음과 같다. 시맨틱 질의 언어인 RDQL 에 대한 소개를 하고 Jena, Sesame 의 RDQL 질의 처리 엔진에 대한 관련연구를 2 장에서 소개한다. 3 장에서는 RDQL 질의 언어를 SQL 로 변환하는 기법을 소개한다. 마지막으로 4 장에서는 결론 및 향후 계획으로 마무리한다.

* 본 연구는 대학 IT 연구 센터 육성 및 지원 사업의 연구결과로 수행되었음.
* 본 연구는 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(R08-2003-000-10464-0)

## 2. 관련 연구

### 2.1 RDQL

RDQL 은 W3C 에서 표준으로 제정된 RDF 를 위한 질의 언어로서 SQL 과 유사한 구문을 제공한다. SELECT 절에서는 트리플 패턴에서 사용된 변수를 기술함으로써 변수의 결과값을 SQL 의 테이블과 같은 형태로 결과값을 얻도록 하고, WHERE 절에서는 트리플 패턴을 기술하여 RDF 그래프 모델상에 있는 특정한 트리플들을 검색할 수 있게 한다. 지문의 제약상 자세한 내용은 생략한다.

### 2.2 Jena2 와 Sesame 의 RDQL 질의 처리

Jena2([5])는 HP Laps 에서 시맨틱 웹 연구의 일환으로 연구 중에 있는 오픈 소스로서 시맨틱 웹 애플리케이션을 구축하기 위한 자바 프레임워크이다. 특징으로는 RDF, RDFS, DAML+OIL, OWL 에 대한 문서 저장 및 관리를 위한 API 를 제공하며 온톨로지 데이터에 대한 질의 언어로서 RDQL 를 지원한다. Jena2 에서 RDQL 질의 처리는 Jena 모델에 기반을 두고 있다. RDQL 파서를 통해서 WHERE 절의 트리플 패턴을 파싱한 다음에 Jena 모델의 그래프 검색을 통하여 질의 결과를 수집한다.

Aduna (NLnet Foundation 과 OntoText 에서 공동연구)에서 연구 중에 있는 Sesame([6])는 RDF 와 RDFS 를 저장, 질의, 추론하기 위한 오픈 소스 자바 프레임워크이다. 주요 특징으로는 RDF 와 RDFS 로 기술된 문서를 저장 관리하는 데이터베이스 시스템으로서 사용된다. 또한 자바 라이브러리를 제공하기 때문에 RDF 를 처리해야 할 필요가 있는 애플리케이션에서도 사용될 수 있다. 지원하는 질의 언어로는

RDQL, SeRQL, RQL 이 있다. 현재, Sesame 에서 RDQL 질의 처리는 RDQL 의 WHERE 절에 있는 각각의 트리플 패턴에 해당되는 결과를 SQL 질의를 이용하여 데이터베이스로부터 중간 결과를 수집한 다음에 이들을 조인하는 방식을 취한다. 이때 트리플 패턴의 결과 수집 순서는 질의 계획에 따라 중간 결과가 적게 생성되도록 하여 효율성을 증가시킨다. 또한, RDQL 에 대한 질의 모델을 생성하고 질의 옵티마이저 (Query Optimizer)를 통으로써 중간 결과가 적게 발생되도록 하여 조인의 효율성을 증가시키게 한다.

## 3. RDQL 질의 처리 엔진

본 논문에서 개발한 RDQL 질의 처리 엔진은 그림 1과 같이 3 단계로 구성된다.

단계 1: RDQL 파서(Parser)는 RDQL 질의를 입력으로 하여 RDF 그래프 모델을 생성한다. 이 단계에서는 기존의 RDQL 파서 모듈을 본 연구 환경에 적합하도록 재개발하였다.

단계 2: RDQL2SQL 은 RDQL 파서를 통해서 생성된 RDF 그래프 모델을 입력으로 하여 대응하는 SQL 질의 언어를 생성한다.

단계 3: RDQL2SQL 을 통해서 생성된 SQL 질의를 기존의 관계형 데이터베이스 질의 처리 엔진을 활용하여 처리 결과를 얻는다.


그림 1. 질의 처리 순서

### 3.1 단계 1: RDQL 파서

RDQL 파서는 입력으로 들어오는 RDQL 질의를 구문 검사를 통하여 RDF 그래프를 생성한다. 이 과정에서 RDQL 파서는 그림 2와 같이 RDQL 질의의 WHERE 절에 있는 각각의 트리플 패턴을 하나의 RDF 그래프로 변환한다. 트리플 패턴에 있는 주어(Subject)와 목적어(Object)는 그래프의 노드로 변환하고, 서술어(Predicate)는 에지(Edge)로 변환된다. 이와 같은 방법으로 모든 트리플 패턴들을 그래프 노드 및 에지로 변환한 후에 동일한 노드를 서로 연결하여 궁극적으로 하나의 RDF 그래프를 생성한다 (그림 2).

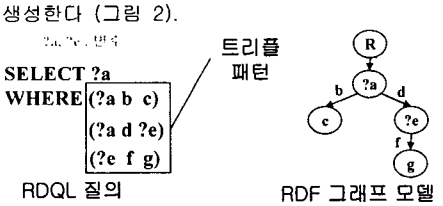



그림 2. RDF 그래프 모델

### 3.2 단계 2: RDQL2SQL 변환기

RDQL2SQL 변환기는 RDQL 질의 처리 엔진에서 핵심이 되는 모듈로서 RDF 그래프 모델에 대응하는 SQL 질의를 생성한다. 이 모듈은 아래에서 설명하는 변환 알고리즘에 의해 구현된다.

변환 알고리즘은 정의 1, 2 를 기반으로 RDF 그래프 모델의 깊이 우선 탐색을 통해서 이루어진다.

정의 1 : RDF 그래프를 이루는 트리플의 집합을 $T = \{t_1, t_2, \cdots, t_n\}$라고 할 때 T 의 각각의 요소는 하나의 SQL 질의로 변환된다. 즉, $t_1 \rightarrow s_1$, $t_2 \rightarrow s_2$, $\cdots$ $t_n \rightarrow s_n$ 으로 변환된다. 단, SQL 질의의 집합을 $S = \{s_1, s_2, \cdots, s_n\}$로 정의한다.

정의 2 : 2 개의 RDF 트리플이 주어(Subject) 또는 목적어(Object)를 공유할 때 아래의 패턴을 따른다.

- 패턴 1 : 2 개의 트리플이 동일한 주어를 가지는 경우 (그림 3의 Pattern 1)

  트리플 ①과 ②의 SQL 질의는 ①.subject = ②.subject 의 조건을 가지는 내부조인(Inner Join)으로 병합된다. 정의 1 에 의해 트리플 ①, ②로부터 변환된 SQL 질의를 각각 SQL(1), SQL(2)라고 할 때, 다음과 같이 정의한다.

  SQL(1) JOIN SQL(2)
  ON SQL(1).subject=SQL(2).subject

- 패턴 2 : 2 개의 트리플이 ①.object=②.subject 일 경우 (그림 3의 Pattern 2)

  트리플 ①과 ②의 SQL 질의는 ①.object = ②.subject 의 조건을 가지는 내부조인(Inner Join)으로 병합된다. 이를 다음과 같이 정의한다.

  SQL(1) JOIN SQL(2)
  ON SQL(1).object=SQL(2).subject

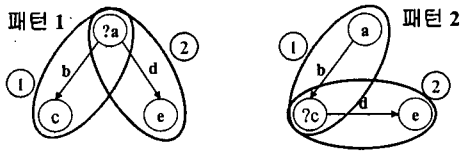



그림 3. 정의 2 의 패턴

그림 3에서 보는 것처럼, 정의 1 은 트리플 ①, ②에 대응하는 SQL 질의로 변환됨을 정의하며, 정의 2 는 RDF 그래프 모델이 SQL 질의로 변환될 때, 변환되는 패턴을 정의한다. 그림 3에서 패턴 1 은 SQL(1)과 SQL(2)가 주어인 " ?a" 를 서로 공유하기 때문에 " SQL(1).subject = SQL(2).subject" 의 조건을 가지는 내부조인으로 병합된다. 패턴 2 도 마찬가지로 내부조인의 조건으로 " SQL(1).object = SQL(2).subject" 를 가지며 SQL(1)과 SQL(2)과 내부조인으로 병합된다.

위와 같이 정의 1, 2 를 정의함으로써 RDF 그래프 모델에서 SQL 질의로 변환할 수 있는 기법을 제공해 줄 수 있다. RDF 그래프 모델과 이에 대응하는 SQL 질의가 동등한 의미를 가짐을 보장하기 위해서 이에 대한 증명이 필요하다.

정리 1 : RDF 그래프 모델은 정의 1, 2 에 의해 변환된 SQL 질의와 동등한 의미를 가진다.

증명 : RDF 그래프 모델에서 RDF 트리플은 2 개의 노드와 이를 연결하는 에지로 구성되어 있으며 이들 트리플들이 같은 노드를 공유함으로써 RDF 그래프 모델을 형성한다. 어떠한 노드나 에지도 독립적으로 존재할 수 없기 때문에 RDF 그래프 모델에서 트리플들은 정의 2 의 패턴 1 과 패턴 2 의 구조를 반드시 가지게 된다. 결과적으로 하나의 RDF 그래프 모델은 정의 1 과 정의 2 에 의해 SQL 질의로 변환될 수 있음을 의미한다.

정의 1 과 2 로부터 RDF 그래프 모델에 대응하는 SQL 질의로 변환하는 예제는 그림 4와 같다. 물론 변환된 SQL 질의는 정리 1 에 의해서 왼쪽의 RDF 그래프와 동등한 의미를 가지게 된다.

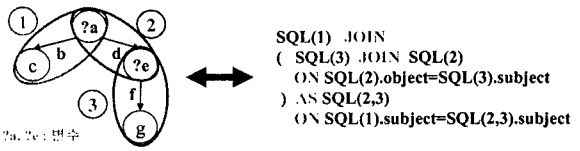
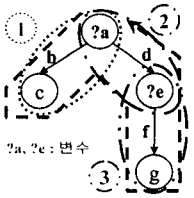



그림 4. RDF 그래프로부터 SQL 질의로 변환

지금까지 RDF 그래프에서 SQL 질의로 변환하기 위한 이론적인 면을 봤다. 여기서부터는 정의 1, 2 를 이용하여 SQL 질의로 변환하는 알고리즘을 소개한다.

알고리즘 : RDQL2SQL
입력 : RDQL 그래프 모델
출력 : SQL 질의

(1) RDF 그래프 모델에서 깊이 우선 탐색을 통해서 그래프의 단말노드 또는 SQL 질의로 변환된 트리플을 탐색한다.
(2) RDF 그래프의 단말 노드에 도달하면 정의 1 에 의해 (주어,서술어,목적어)에 해당하는 트리플을 SQL 질의로 변환한다.
(3) 변환해야 될 대상이 단말 노드가 아니고 정의 2 의 2 가지 패턴에 만족될 경우 두 개의 변환된 SQL 질의를 패턴 1, 패턴 2 에 맞게 변환한다.
(4) RDF 그래프의 탐색이 끝날 때 까지 (1), (2), (3)을 반복한다.




1. SQL(1) 생성
2. SQL(3) 생성
3. SQL(2) 생성
4. 정의 2의 패턴 2에 의해
   SQL(3) JOIN SQL(2) ON SQL(2).object=SQL(3).subject
5. 4에서 변환된 SQL문을 SQL(2,3)이라고 하면
   SQL(1) JOIN SQL(2,3) ON SQL(1).subject=SQL(2,3).subject

그림 5. RDQL2SQL 알고리즘의 예

그림 5는 알고리즘 RDQL2SQL 를 통해서 SQL 질의로 변환하는 과정을 보여준다. RDF 그래프 모델에서 깊이 우선 탐색이기 때문에, 먼저 ①이 정의 1 에 의해서 SQL(1)로 변환되고 그 다음 ③이 SQL(3)으로 변환된다. 다음에 ②가 SQL(2)로 변환되고 정의 2 의 패턴 2 에 의해 SQL(2)와 SQL(3)이 조인된다. 마지막으로 그림 5의 5 로 변환된다.

변환 알고리즘에서 SQL 질의로 변환되는 최소 단위는 트리플이다. RDF 그래프 모델에서 깊이 우선 탐색을 통해서 각각의 트리플이 SQL 질의로 변환된 후에는 트리플에 대응하는 SQL 질의들이 서로 내부조인으로 묶여져 병합된다(정의 2 에 의해).

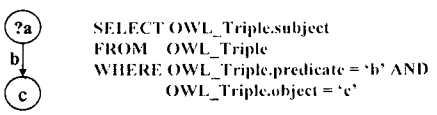



그림 6. 트리플로부터 SQL 질의로의 변환

RDQL 의 WHERE 절이 가질 수 있는 트리플 패턴은 (a,b,c), (?a,b,c), (a,?b,c), (a,b,?c), (?a,?b,c), (?a,b,?c), (a,?b,?c), (?a,?b,?c) 가 되며 트리플로부터 SQL 질의로의 변환은 간단한 규칙에 의해서 변환이 된다. 첫째, 변수(' ?' 로 시작하는 노드명)는 SQL 질의의 " SELECT" 부분에 위치한다. 둘째, 변수가 아닌 노드는 SQL 질의의 " WHERE" 절에 위치하고 2 개 이상의 노드일 경우 SQL 질의의 " AND" 에 묶여진다. 두 개의 규칙에 의해서 그림 6과 같이 SQL 질의로 변환된다. 노드 " ?a" 는 변수이므로 SELECT 에 위치한다. 즉 " ?a" 는 트리플의 주어에 해당되므로 " SELECT OWL_Triple.subject 로 변환된다. 그리고 간선 " b" , 노드 " c" 는 변수가 아니므로 WHERE 에 위치하여 그림 6과 같이 변환된다. 여기에서 OWL_Triple 는 시맨틱 웹 온톨로지 문서를 저장하는 시스템에 의존하게 된다. 이 논문에서 OWL_Triple 는 RDBMS 의 테이블로서 OWL 문서를 트리플(주어:subject, 서술어:predicate, 목적어:object)로 저장하기 위한 것이다. 정의 및 설명을 간략하고 명확하게 하기 위해 이름공간(namespace)는 고려하지 않았다. 이름공간을 고려하는 것은 간단하기 때문에 이 논문에서는 생략한다. 그래서 OWL_Triple 는 속성으로서 subject, predicate, object 를 가지는 테이블이다.

### 3.3 단계 3: 관계형 데이터베이스 질의 처리기

RDQL2SQL 변환기를 통해 생성된 SQL 질의는 기존의 관계형 데이터베이스 질의 처리기를 통하여 최종 질의 결과를 얻게 된다.

## 4. 결론 및 향후 계획

본 논문에서는 RDQL 기반의 시맨틱 질의를 효율적으로 처리하기 위해 RDQL 질의에 대응하는 SQL 질의로 변환하는 RDQL2SQL 를 기반으로 RDQL 질의 처리 엔진을 개발하였다. 이를 통하여 실시간 지능 로봇 플랫폼과 같은 환경에서 요구하는 효율적인 질의 처리를 지원할 수 있다. 향후에는 RDQL 질의가 변환된 SQL 질의를 더욱 효율적으로 처리하기 위해 기존의 관계형 데이터베이스 환경에서의 시맨틱 데이터에 대한 인덱싱(Indexing) 기법을 연구하고자 한다.

## 참 고 문 헌

[1] Dieter Fensel, "Ontologies : A Silver Bullet for Knowledge Management and Electronic Commerce", Springer-Verlag Berlin and Heidelberg GmbH & Co.K, 2004.
[2] Graham Klyne, Jeremy Carroll, "Resource Description Framework (RDF): Concepts and Abstract Syntax W3C Recommendation", Feb 2004. See http://www.w3.org/TR/rdf-concepts/.
[3] Dan Brickley, R.V. Guha, "RDF Vocabulary Description Language 1.0 : RDF Schema W3C Recommendation", Feb 2004. See http://www.w3.org/TR/rdf-schema/.
[4] Deborah L. McGuinness, Frank van Harmelen, "OWL Web Ontology Language Overview W3C Recommendation", Feb 2004. See http://www.w3.org/TR/owl-features/.
[5] Jeremy Carrol, Brian McBride, "The Jena Semantic Web Toolkit". HP-Labs, Bristol, 2001. See http://jena.sourceforge.net/.
[6] J. Broekstra, A. Kampman, F. van Harmelen, "Sesame: A Generic Architecture for Storing and Querying RDF and RDF Schema", In The Semantic Web - ISWC 2002, volume 2342 of Lecture Notes in Computer Science, pp. 54-68, 2002. See http://openrdf.org/.
[7] Andy Seaborne, HP Labs Bristol , "RDQL - A Query Language for RDF : W3C Member Submission", Jan 2004. See http://www.w3.org/Submission/2004/SUBM-RDQL-20040109/.
[8] Eric Prud' hommeaux, Andy Seaborne, " SPARQL : Query Language for RDF : W3c Working Draft" , 21 July, 2005