의사 결정 트리를 이용한 RDF 실체 뷰 관리 기법

Materialized View Management Scheme of RDF using Decision Tree

박 재 열*, 최 기 태*, 윤 상 원*, 임 종 태*, 복 경 수*,
이 병 엽**, 유 재 수*
충북대학교*, 배재대학교**

Park jae-yeol*, Choi ki-tae*, Yoon sang-won*,
Lim jong-tae*, Bok kyoung-soo*,
Lee byoung-yup**, Yoo jae-soo*
Chungbuk National Univ.*, Pai-Chai Univ.**

요약

본 논문에서는 의사 분산 트리를 이용하여 효율적으로 후보 실체 뷰를 선택하는 기법을 제안한다. 제안하는 기법은 후보 실체 뷰의 이득, 실체화 크기, 그리고 갱신율을 고려하여 의사 결정 트리로 구축한다. 의사 결정 트리를 이용하여 효율이 높은 후보 실체 뷰의 선택 및 빠른 교체 수행을 목적으로 한다.

Ⅰ. 서론

최근 시맨틱 웹의 발전과 함께 RDF 데이터의 양이 증가하고 있다. RDF 데이터들이 증가함에 따라 질의 처리 비용을 절감 하는 많은 연구들이 진행되고 있다. RDF는 데이터를 트리플(triple)로 표현한다. 트리플 구조는 (Subject, Predicate, Object)의 3개의 구성으로 이루어져 있다. 트리플은 주어(Subject)와 목적어(Object)를 노드(node)로 표현하고, 이 주어와 목적어의 술어(Predicate)를 나타내는 화살표(edge)로 연결하여 그래프화 할 수 있다.

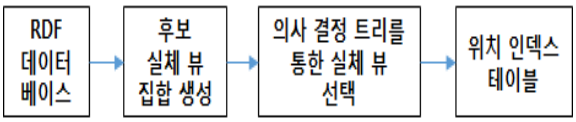
RDF 트리플 기반의 데이터베이스는 풍부한 데이터 표현을 할 수 있지만, 질의 처리시 많은 비용이 소모된다. 그중 가장 큰 비용을 차지하는 부분은 조인 연산이며, 최근 RDF 데이터베이스에서 조인 비용 감소를 위한 많은 연구들이 진행되었다. 첫 번째는 많은 트리플 데이터를 분산하여 저장하는 방법이다[1]. 두 번째는 트리플에 대한 색인 구조를 생성하여 질의에 대한 트리플 패턴에 대해 가장 적절한 색인을 선택하여 질의에 사용하는 방법이다[2]. 세 번째는 단축 경로 선택(Shortcut Selection)으로 조인의 횟수를 감소시키는 방법이다[3, 4]. [3]에서는 질의 빈도 및 저장 공간의 한도뿐만이 아니라 갱신 비용과 유지보수 비용도 같이 고려하고 있지만, 실체화 된 단축 경로의 교체 전략이 존재하지 않는다.

본 논문에서는 의사 결정 트리를 이용한 후보 실체 뷰 관리 기법을 제안한다. 이를 위해 제안하는 기법은 먼저 후보 실체 뷰(단축 경로) 이득(BenefitF), 실체 뷰 크기(Size), 그리고 갱신율(Update)을 고려하는 의사 결정 트리[4]를 구축한다. 또한 구축 된 의사 결정 트리를 통하여 효율성이 높은 순으로 후보 실체 뷰를 실체 뷰로 만든다. 위치 인덱스 테이블은 실체 뷰 교체시 필요 정보를 관리한다.

Ⅱ. 제안하는 실체 뷰 관리 기법

1. 전체 처리 과정

그림 1은 제안하는 기법의 전체 처리 과정을 보여준다. 구축 되어있는 RDF 그래프(RDF 트리플 데이터베이스)로부터 후보 실체 뷰(단축 경로) 집합을 생성한다. 생성된 후보 실체 뷰 집합은 의사 결정 트리의 3가지 속성(BenefitF, Size, Update)으로 분류되어 실체 뷰 선택이 이루어지고, 위치 인덱스 테이블에 실체 뷰 정보 저장 및 관리한다.



▶▶그림 1. 제안하는 기법의 처리과정

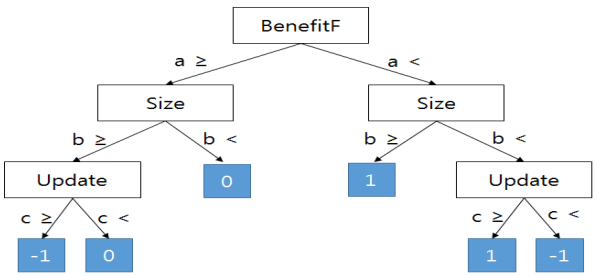
후보 실체 뷰는 RDF 그래프에서 길이(서로 다른 노드를 연결하는 엣지의 수)가 2 이상의 가능한 모든 부분 그래프의 가상 경로를 의미한다. 만들어진 후보 실체 뷰를 $subv_i$로 표현하고 단축 경로의 집합 SUBV는 $\{subv_1, subv_2, \ldots, subv_n\}$로 구성된다.

* 교신저자 : yjs@chungbuk.ac.kr

이 논문은 2013년도 정부(미래창조과학부)의 재원으로 한국연구재단(No.2013R1A2A2A01015710), 2014년도 정부(교육부)의 재원으로 한국연구재단(No2014R1A1A2055778) 및 교육부와 한국연구재단의 지역혁신인력양성사업으로 수행된 연구결과임 (No.2013H1B8A2032298)

2. 의사 결정 트리와 후보 실체 뷰 선택

본 논문에서는 그림 2에서 나타난 의사 결정 트리를 사용하여 실체 뷰를 선택한다. 실체화할 후보 실체 뷰 선택을 위하여 3가지 속성 실체 뷰 이득, 실체 뷰 크기, 갱신율를 의사 결정 트리에서 고려한다. 실체 뷰 이득은 4절에서 설명한다. 변수 a의 값은 전체 후보 실체 뷰의 평균 이득이다. 실체 뷰 크기는 후보 실체 뷰를 실체 뷰로 만들었을 때 크기이다. 변수 b의 값은 실체화 뷰가 저장될 저장 공간의 크기의 5%이다. 갱신율은 각 노드에 속하는 인스턴스들이 변경될 확률이다. 예를 들어 한 노드의 갱신율은 0.1이다. 0.1이라는 값은 질의를 10번 처리하였을 때 학번 노드에 갱신이 한번 일어난다는 것을 의미한다. 변수 c의 값은 전체 노드의 변화율의 합을 2로 나눈 값이다. 단, 전체 노드의 변화율은 1이하의 값이며, 하나의 노드의 갱신율은 0.3 이하의 값이다.



▶▶ 그림 2. 의사결정 트리

의사 결정 트리는 3개의 속성을 고려하여 후보 실체 뷰를 최종적으로 {1, 0, -1}의 값을 갖는 3개의 집합으로 분류하여 실체 뷰를 선택한다. 1에 해당되는 후보 실체 뷰는 메모리에 실체 뷰로 저장이 된다. -1에 해당되는 후보 실체 뷰는 현재 실체 뷰로 저장은 되어 있지만 추후 교체 상황 발생 시 우선적으로 선택된다. 0은 효율이 떨어지는 후보 실체 뷰의 집합으로 후보 실체 뷰의 가상 경로만 저장된다.

3. 위치 인덱스 테이블

그림 4은 후보 실체 뷰의 위치 인덱스 테이블이다. 테이블은 후보 실체 뷰의 이름과 의사 결정 트리에서 분류된 3개의 집합을 나타내는 Level, 후보 실체 뷰의 실체화 상태를 나타내는 State, 실체 뷰가 어디에 위치에 있는지 포인터로 관리하는 Location으로 이루어져있다. Level 에서는 실체 뷰(1), 후보 실체 뷰(-1), 가상 경로(0) 중 하나의 값이 들어가게 된다. 후보 실체 뷰를 이용하여 교체 상황 발생 시 교체 대상의 추가 연산 없이 빠른 접근을 통해 교체 비용을 감소시킨다.

SUBV	Level	State	Location
$subv_1$	-1	후보	위치 주소
$subv_2$	0	가상	위치 주소
$subv_3$	1	실체화	위치 주소

▶▶ 그림 4. 위치 인덱스 테이블

4. 이득 함수

후보 실체 뷰 선택 문제에서는 선택된 후보 실체 뷰가 최대의 이득을 가지도록 유지하는 것이 가장 중요하다. 이득을 구하는데 가장 중요한 것은 질의 처리 시간과 그 질의의 빈도이다. [3]에서는 후보 실체 뷰가 없이 질의 처리 비용에 후보 실체 뷰를 사용한 질의 처리 비용을 뺀 값에 빈도를 곱하여 이득을 구하였다. 본 논문에서는 정확성을 더욱 높이기 위하여 이득을 다음과 같이 계산한다.

$$BenefitF(subv_i) \sum_{q_k \in RQ_i} f_k * \frac{cost(q_k) - cost(q_k, shc_i)}{cost(q_k)} \quad (1)$$

cost(q_k)는 후보 실체 뷰 없이 질의 q_k를 처리하는 비용이고, cost(q_k, shc_i)는 후보 실체 뷰 $subv_i$를 사용하여 질의 q_k를 처리하는 비용이다. f_k는 질의 q_k에 해당하는 질의 빈도이다. RQ_i는 후보 실체 뷰 $subv_i$를 사용하여 질의 처리를 할 수 있는 질의의 집합이다.

Ⅲ. 결론

본 논문에서는 의사 분산 트리를 이용하여 후보 실체 뷰 선택 기법을 제안하였다. 제안하는 기법은 후보 실체 뷰 이득, 실체 뷰 크기, 그리고 갱신율을 고려하여 의사 분산 트리를 구축한다. 의사 분산 트리를 이용하기 때문에 저장 공간 및 갱신 비용의 측면에서 우수하다. 또한 위치 인덱스 테이블을 통하여 실체 뷰의 교체 비용을 감소시킨다. 향후 연구로는 본 논문에서 제안하는 기법의 우수성을 입증하기 위해 다양한 성능 평가를 수행할 예정이다.

▌참 고 문 헌▐

[1] 김천중 “대규모 RDF 데이터의 분산 저장을 위한 동적 분할 기법”, 한국정보과학회논문지, 제41권, 제12호, pp.1126-1135, 2014.

[2] Neumann, T. and Gerhard, W., “RDF-3X: a RISC-style Engine for RDF”, Proceedings of the VLDB Endowment Vol. 1, No. 1, pp.647-659, 2008

[3] 강승석 “트리플 데이터베이스 단축 경로 이득 함수와 구성 인자 실험 분석“, 한국전자거래학회지 19권, 제1호, pp.131-143, 2014.

[4] 장윤경 “데이터 웨어하우스에서 의사결정 트리를 이용한 실체화 뷰 선택 기법“, 한국정보처리학회 학술발표대회자료, 제13권, 제1호, pp.63-66 , 2006