대용량 온톨로지 데이터의 가시화 연구

정성문*, 이정훈*, 한욱신*+
*경북대학교 IT 대학 컴퓨터학부
e-mail : {smchung, jhlee}@www-db.knu.ac.kr, wshan@knu.ac.kr

A Study on Visualization for Large Ontology Data

Sung-Moon Chung*, Jeong-Hoon Lee*, Wook-Shin Han*+
*School of Computer Science and Engineering, Kyungpook National University

요 약

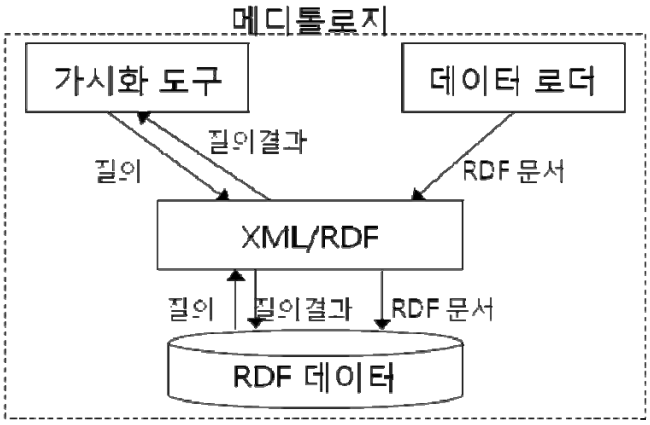
온톨로지는 정보들간의 체계 및 상호작용을 표현하고, 이를 통해 사용자들에게 유용한 지식을 제공하는 툴로 정보과학, 전자상거래, 및 의료 분야 등에서 널리 활용되고 있다. 본 논문에서는 온톨로지 데이터베이스 관리 시스템인 XML/RDF 를 이용하여 대규모의 온톨로지 데이터를 효율적으로 처리하고 가시화하는 방안에 대해 연구한다.

1. 서론

온톨로지는 정보들간의 체계 및 상호작용을 표현하고, 이를 통해 사용자들에게 유용한 지식을 제공하는 툴로 정보과학, 전자상거래, 및 의료 분야 등에서 널리 활용되고 있다. 그러나, 온톨로지를 다루는 기존 연구들은 메인 메모리 기반의 데이터 처리 및 처리 결과의 가시화를 수행하고 있어, 대용량의 온톨로지 데이터를 다루기 어려운 문제가 있다. 즉, 디스크에 저장된 온톨로지 데이터 전체를 메인 메모리에 로드(load)한 뒤, 로드된 데이터에 대해 질의 처리 및 가공을 수행하므로, 데이터를 처리하기 전까지 준비 시간(preparation time)이 매우 길어 비효율적일 뿐만 아니라, 처리할 수 있는 데이터의 양이 메인 메모리의 용량에 의해 제한된다. 본 논문에서는 온톨로지 데이터베이스 관리 시스템인 XML/RDF [1]를 이용하여 온톨로지 데이터를 메인 메모리에 모두 로드(load)하지 않고도, 대용량의 온톨로지 데이터를 효율적으로 처리하고 가시화하는 방안에 대해 연구한다. 그리고, 기존의 메모리 기반 시스템인 프로테제 [2]의 TripleStore [3]와의 성능 비교를 통해 제안한 방법의 우수성을 보인다.

2. XML/RDF 기반의 온톨로지 처리 및 가시화

본 논문에서는 RDF 데이터베이스 관리 시스템인 XML/RDF 를 이용하여 RDF 형식을 따르는 트리플(triple) 형태의 온톨로지 데이터와 스키마 정보를 저장하고, 온톨로지 데이터에 대한 검색이 필요할 때에만 XML/RDF 에 저장된 데이터에 접근하여 원하는 검색 결과를 얻어 이를 가시화하는 디스크 기반의 접근 방식을 취한다. 그림 1 은 XML/RDF 기반의 온톨로지 데이터 처리를 위해 제안한 방법을 따르는 시스템인 메디톨로지의 시스템 아키텍처를 나타낸다. 그림에서 보는 바와 같이 메디톨로지는 RDF 형태로 저장된 데이터로부터 온톨로지 데이터의 스키마를 읽어 들이고, 사용자가 요구할 때마다 XML/RDF 를 검색하여 결과를 스키마에 맞도록 변환한 뒤에 이를 가시화하는 방식으로 동작한다.



(그림 1) 메디톨로지의 시스템 아키텍처

3. 성능 평가

본 절에서는 디스크를 기반으로 온톨로지 데이터를 처리를 수행하는 메디톨로지와 메모리를 기반으로 온톨로지 데이터를 처리하는 기존 방법인 프로테제간의 성능을 비교한 실험 결과를 보여준다.

.

+ 교신저자.
*본 연구는 지식경제부 및 한국산업기술평가관리원의 지식서비스 산업원천기술개발사업의 일환으로 수행하였음.[KI10033545]

3.1. 실험 환경 및 데이터

실험에서는 합성(syntactic) 데이터로 벤치마크(benchmark)를 위해 많은 연구에서 사용되고 있는 LUMB [4] 데이터를 사용한다. 이 데이터는 메디톨로지의 온톨로지 저장시스템인 XML/RDF 와 프로테제의 온톨로지 저장시스템인 TripleStore [3]에 각각 미리 저장되어 있다고 가정한다. 실험은 데이터의 크기를 20, 40, 80, …, 640, 1280 Mbytes 로 2 배씩 증가시켜가며 수행하였으며, 척도로는 데이터를 처리하기 전까지의 준비 시간(preparation time)과 이때 비교 대상들의 메모리 사용량(memory usage)이다. 메디톨로지와 프로테제에서 준비 시간과 준비 시간을 구성하는 요소들에 대한 설명은 각각 표 1 및 표 2 와 같다.

표 1. 메디톨로지와 프로테제의 준비 시간.

실험 비교 대상	준비 시간
메디톨로지	스키마 정보의 시각화 시간
프로테제	메모리 로드 시간 + 온톨로지 요약정보의 시각화 시간

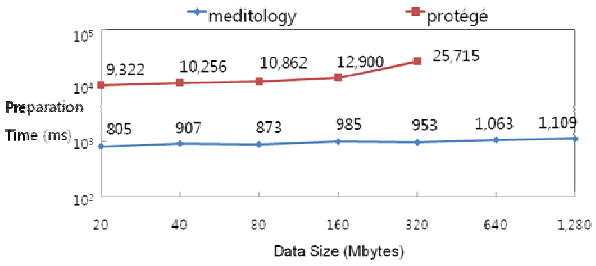
표 2. 준비 시간의 구성요소 설명.

구성요소	설명
스키마 정보의 시각화 시간	메디톨로지에서 온톨로지의 스키마 정보를 XML/RDF 로부터 읽어 들이고, 이를 가시화하는데 걸리는 시간
메모리 로드 시간	프로테제에서 모든 온톨로지 데이터를 TripleStore 로부터 메인 메모리로 읽어 들이면서 온톨로지 객체를 생성하는데 걸리는 시간
온톨로지 요약정보의 시각화 시간	프로테제에서 생성된 온톨로지 객체로부터 온톨로지 요약정보를 추출하고 이를 가시화하는데 걸리는 시간

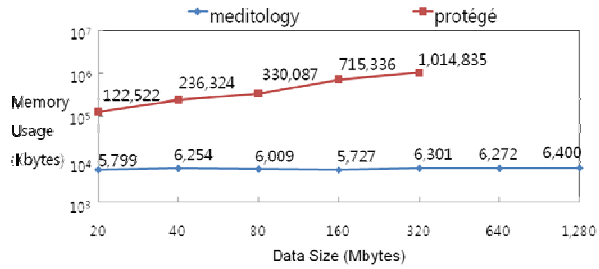
3.2. 실험 결과

실험 결과에서는 합성 데이터인 LUBM 데이터에 대해 메디톨로지와 프로테제의 준비 시간 및 이때의 메모리 사용량을 비교하였다. 대용량 데이터를 대상으로 한 실험에서 메모리 부족으로 인해, 프로테제가 동작하지 않는 경우는 결과를 그래프상에 표시하지 않았다.

그림 2(a)에서 보는 바와 같이 메디톨로지는 프로테제에 비해 11.3~27.0 배 빠르며, 1GBbytes 이상의 대용량 온톨로지 데이터도 처리가 가능하다. 또한, 데이터 크기의 증가에 따른 준비 시간의 증가는 304ms 로 무시할 수 있을 정도로 작다.



(a) 준비시간 비교.



(b) 메모리 사용량 비교.

(그림 2) LUBM 데이터에 대한 성능 비교.

그림 2(b)는 LUBM 데이터에 대한 메모리 사용량 비교 실험의 결과를 보여 준다. 메디톨로지가 프로테제에 비해 21.1 ~ 161.1 배 작은 메모리를 사용한다. 또한, 320Mbytes 이상의 온톨로지 데이터를 사용하는 경우, 프로테제는 동작하지 않는 반면, 메디톨로지는 1GBytes 이상의 데이터에 대해서도 동작한다. 결론적으로 제안하는 시스템인 메디톨로지는 기존의 메모리 기반 시스템인 프로테제에 비해 작은 메모리를 사용하여, 온톨로지 데이터를 빠르게 처리할 수 있다. 또한 프로테제가 처리할 수 없는 대용량의 온톨로지 데이터도 효율적으로 처리 가능하다.

4. 결론

본 논문에서는 XML/RDF 를 저장 기반 시스템으로 대용량 온톨로지 데이터를 처리하는 디스크 기반의 온톨로지 데이터 처리 방안을 제시하였다. 그리고, 기존의 메인 메모리 기반 시스템과 성능평가 수행을 통해 제안한 방법이 효율적이며, 대용량의 온톨로지 데이터를 처리할 수 있음을 보였다.

참고문헌

[1] J. Lee et al., “Processing SPARQL Queries with Regular Expressions in RDF Databases.” *BMC Bioinformatics*, 12(Suppl 2):S6, 2011.
[2] Protégé. version 3.4. http://Protégé.stanford.edu/.
[3] Triplestore http://en.wikipedia.org/wiki/Triplestore/.
[4] Y.Guo et al., “Lubm: A benchmark for owl knowledge base systems.” *J.Web Semantics*, 3(2-3):158-182, 2005.
[5] C.Bizer et al., “DBpedia-a crystallization point for the web data,” *J.Web Semantics*, 7(3):154-165, 2009.