링크드 데이터를 위한 대용량 RDF 저장 및 검색 시스템

이용주
경북대학교 IT대학 컴퓨터학부
e-mail:yongju@knu.ac.kr

A Large-scale RDF Storage and Retrieval System for Linked Data

Yong-Ju Lee
School of Computer Science and Engineering, Kyungpook National University

요 약

본 논문에서는 링크드 데이터를 위한 대용량 RDF 저장 및 검색 시스템을 제안한다. 현재 링크드 데이터에 대한 핵심 이슈는 링크드 데이터의 효율적인 저장과 검색, 그리고 활용 애플리케이션 개발이다. 제안 시스템은 저장 관리자, 인덱스 구조, 그리고 질의 처리기로 구성되어 있다. 저장 관리자는 대용량 RDF 데이터를 처리하기 위해 그래프 데이터베이스에 데이터를 분산 저장하며, 인덱스 구조는 다차원 히스토그램, 보조 인덱싱, 그리고 그래프 인덱싱 기법이 구현된다. 질의 처리기는 SPARQL 또는 NoSQL 질의를 사용하여 질의 최적화 및 랭킹기법이 적용된 RDF 트리플 검색을 수행한다.

1. 서론

최근 빅데이터(big data)의 가치와 중요성이 널리 인식되면서 빅데이터가 정치, 경제, 사회, 문화, 과학 등 전 영역에 걸쳐 활용되고 있다. 이러한 환경하에서 정부 각 부처에서는 그들이 보유한 데이터의 효율적인 공개 및 재사용 증진을 위하고, 정부 3.0 공공정보 개방운동의 일환으로 Open API 구축을 활발히 추진하고 있다[1]. 상업적으로도 지난 몇년전부터 구글, 야후, 이베이, 아마존과 같은 웹 주요 벤더들은 웹 2.0의 대표적 산물인 Open API를 통해 그들의 자원을 외부로 공개하기 시작하였으며, 대표적인 Open API 웹 사이트인 ProgrammableWeb에서는 2016년 8월 17일 기준으로 15,610개의 API를 제공하고 있다. 이러한 수많은 Open API들은 두 가지 이상의 서로 다른 자원을 섞어서 완전히 새로운 가치의 컨텐츠를 만드는 매쉬업(mashup) 개발을 촉진시켰으나 Open API는 중복된 데이터들을 분리된 사일로(silos) 속으로 방치할 수 밖에 없었으며, 매쉬업 개발자들은 그들의 애플리케이션 개발을 위해 특정 Open API에 접속해서 그들의 요구에 맞는 메소드를 선택하여 사용해야만 했다[2].

반면에, 링크드 데이터는 관련된 데이터를 서로 링크함으로써 다음과 같은 장점을 얻을 수 있다. 1) 내가 만든 데이터가 아니더라도 외부에 개방된 데이터를 연결하면 웹을 하나의 거대한 지식베이스처럼 사용할 수 있다[3]. 2) 링크드 데이터를 통해 공개된 데이터를 이용하면 자신이 원하는 데이터가 이미 존재하는지, 있다면 어디에 존재하는지를 알 수 있으므로 시스템의 사일로 문제에 의해 발생될 수 있는 불필요한 데이터 중복 문제를 해결할 수 있다[4]. 3) 시맨틱 웹 표준인 RDF 형태로 발행되므로 마치 웹을 하나의 거대한 글로벌 데이터베이스처럼 질의하고 이용할 수 있다. 4) URI, RDF, SPARQL 등 표준에 의존한다. 5) 데이터 링크를 따라가는 것에 의해 런타임으로 새로운 데이터의 발견이 가능하고, 웹상에 새로운 데이터 소스가 계속 생산됨에 따라 데이터들이 새로운 가치를 더할 수 있다.

현재 링크드 데이터에 대한 핵심 이슈는 링크드 데이터의 효율적인 저장과 검색, 그리고 활용 애플리케이션 개발이다. 따라서 본 연구에서는 웹상에 데이터 의미적 연결을 통해 구조화되어 있는 가장 최선의 방법인 링크드 데이터에 대한 초기 연구로써 저장 및 검색 시스템을 제안한다.

2. 링크드 데이터 국내외 구축현황

링크드 데이터는 2007년에 20억개의 RDF 트리플과 2백만개의 링크에서 2011년엔 316억개의 RDF 트리플과 5억개의 링크로 급속히 증가하였다. 2014년도에는 2011년도 295개 데이터셋이 1,014개 데이터셋으로 엄청난 규모의 초대용량 빅데이터로 발전하였다.

해외 구축현황을 살펴보면, 2009년 미국정부는 공공데이터 개방사이트(http://data.gov)를 개설하고 공공정보에 대한 접근을 용이하게 하여 국민의 창의적인 혁신을 지원하고 있다. 2010년 영국도 이와 유사한 플랫폼인 CKAN (http://ckan.org)를 오픈소스로 공개했다. 디비피디아(http:

이 논문은 2016년도 정부(교육부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No. 2016R1D1B02008553).

//dbpedia.org)는 위키피디아로부터 RDF 데이터를 추출하여 웹에서 이들 정보들이 활용 가능하도록 하였고, 현재 약 400만개의 개념(things)이 기술되어 있으며 영어 외에 119개 언어로 구성된 정보를 제공하고 있다. BBC 방송은 미디어 분야 중 가장 적극적으로 링크드 데이터를 직접 활용하고 업무영역까지 확장하여 실제 사용자들에게 BBC가 생산하는 막대한 정보들을 상호 연계·공유된 형태로 제공하고 있다. 연이어, 뉴욕타임즈가 링크드 데이터 프로젝트를 수행하였고, 미의회도서관이 LCSH 링크드 데이터를 구축하였으며, 미국의 전자제품 소매회사인 베스트바이가 GoodRelations 온톨로지를 이용해 제품 정보를 링크드 데이터 형식으로 발행하였다. 최근 들어 도서관 영역에서 링크드 데이터를 적용하는 사례가 늘고 있는데, 특히 각국의 국립중앙도서관들이 앞장서고 있다.

우리나라도 최근 분야별로 링크드 데이터 구축사업을 추진하고 있으나 아직까지는 대부분 사업의 초기단계이고 그 효과를 검증하기 위한 시범사업 수준에 머무르고 있다. 한국과학기술정보연구원(KISTI)은 논문, 특허, 보고서, 동향정보 등을 서비스하고 있는 NDSL 데이터를 일부 RDF 트리플로 변환하여 링크드 데이터 서비스를 제공하고 있다. 최근에는 서울시의 열린데이터광장(http://data.seoul.go.kr)이나 한국정보화진흥원(NIA)의 국가 공유자원 포털(http://data. go.kr)을 통해 각 부처의 다양한 데이터를 통합 구축하여 제공하고 있다. 또한 NIA는 링크드 오픈데이터 시범사업인 공공DB피디아(http://lod.data.go.kr)를 통해 공공데이터를 대상으로 링크드 데이터를 시범적으로 구축하고 있다. 2014년 12월 NIA, KISTI, 국립수목원, 국립중앙도서관, 국사편찬위원회, 서울특별시, 특허청, 그리고 한국교육학술정보원과 링크드 데이터 협력에 관한 업무협약(MOU)를 체결하였으나, 국내의 경우 링크트 데이터는 아직 시작단계로써 구축분야 및 연결이 잘 맞지 않아 기술 저변확대가 절실히 필요한 상황이다.

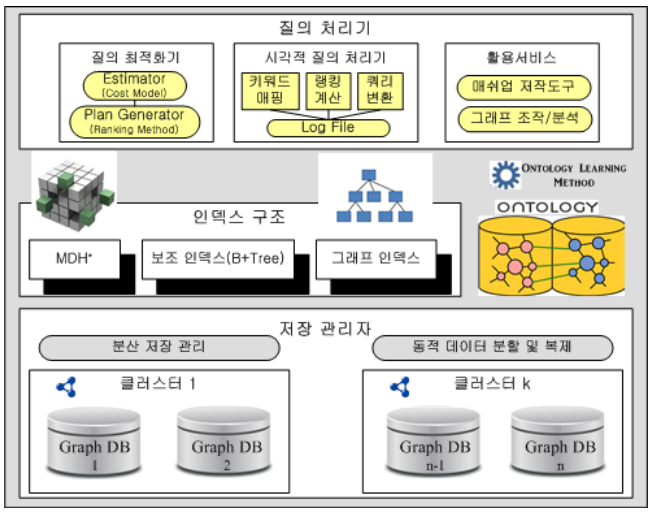
3. 대용량 RDF 저장 및 검색 시스템

본 연구에서 제안하는 시스템은 누구나 자유롭게 링크드 데이터를 발행하고 연계·활용할 수 있는 링크드 데이터 기반 저장·검색 시스템이다. 이를 통해 쉽게 창의적인 아이디어를 개발·공유할 수 있는 차세대 웹 환경을 제공한다. 특히 다음과 같은 링크드 데이터 생명주기(life cycle)에 맞추어 관련기술들을 개발함으로써 실질적이고 유용한 시맨틱 웹 3.0 개발 도구를 지원한다.

- 데이터 수집: 정형/비정형 데이터를 RDF 데이터로 변환하여 저장
- 온톨로지 구축: 자동 온톨로지 학습방법에 의한 상위레벨 OWL 온톨로지 구축
- 저장/색인: 초대용량 링크드 데이터를 위한 압축된 데이터 저장 및 인덱싱 기법 개발, 그리고 데이터 연관성과 이질적 분산 특성을 고려한 그래프 데이터베이스 기술 개발
- 검색/분석: 사용자를 위한 시각적인 링크드 데이터 탐색 및 강력한 검색 질의기 개발, 그리고 사용자를 위한 직관적인 분석 도구 개발

이러한 생명주기 각 단계는 개별적으로 분리되어 존재하지 않고 상호보완적으로 서로 연계되어 전체적인 문제를 해결해 나간다.

(그림 1)은 대용량 링크드 데이터 저장·검색 시스템의 구성요소와 그들 간의 관계를 나타낸 것이다. 저장 관리자는 대용량 RDF 데이터에 대한 분산 저장을 위해 그래프 데이터베이스에 데이터를 분할하고 부하 분산을 처리하기 위해 동적 데이터 분할 및 복제를 수행한다. 인덱스 구조는 확장된 다차원 히스토그램, 보조 인덱싱, 그래프 인덱싱 방법을 적용하고, 링크드 데이터 기반 온톨로지 학습방법을 적용한다. 질의 처리기는 SPARQL 또는 NoSQL 질의를 사용하여 질의 최적화, 랭킹기법이 적용된 RDF 트리플 검색을 수행한다. 활용서비스는 매쉬업 저작도구 및 그래프 데이터베이스 조작/분석을 수행한다.



(그림 1) 대용량 링크드 데이터 저장·검색 시스템

4. 결론

본 논문에서는 링크드 데이터의 국내외 구축현황을 분석하고, 링크드 데이터를 위한 대용량 RDF 저장 및 검색 시스템을 제안한다. 제안된 시스템은 누구나 자유롭게 링크드 데이터를 발행하고 연계·활용할 수 있는 링크드 데이터 기반 저장·검색 시스템이다.

참고문헌

[1] 행정안전부, 정부3.0 추진 기본계획, 2013.
[2] 이영환, 웹 3.0 세상을 바꾸고 있다, 보문각, 2010.
[3] S. Auer, et. al, “Introduction to Linked Data and its lifecycle on the Web,” Reasoning Web 2013, LNCS 8067 Tutorial, 2013, pp. 1-90.
[4] C. Bizer, “The emerging Web of Linked Data,” Intelligent Systems, vol. 24, no. 5, 2009, pp. 87-92.