온톨로지 서버구축을 통한 시맨틱 웹 기반 정보검색 시스템 설계

양새동 · 김경환 · 김종문 · 김창수 · 정회경
배재대학교 컴퓨터공학과

A System Design for Search of Semantic Web-based Information through the Server Ontology

Xi-tong Yang · kyung-Hwan Kim · Jong-Moon Kim · Chang-Su Kim · Hoe-Kyung Jung
Department of Computer Engineering, PaiChai University
E-mail : withchyang1@gmail.com, shwan10@gmail.com, elcomtech@elcomtech.co.kr,
MIE-ddoja@pcu.ac.kr, hkjung@pcu.ac.kr

요 약

정보검색 시스템은 사용자가 검색하고자 하는 정보를 보다 정확하고 신속하게 전달하는 데 그 목적이 있다. 그러나 현재의 검색 시스템은 단순 구문 분석 방식으로 사용자가 원하는 정확한 정보를 제공하지 못하고 있다.

본 논문에서는 온톨로지 서버구축을 통한 정보검색 시스템을 제안한다. 제안하는 시스템은 시맨틱 웹 기반의 정보검색 기법을 이용하여 구조화된 문서뿐만 아니라 다양한 포맷의 데이터들의 처리를 극대화 시키고자 한다. 또한 상호 운용성 및 데이터 통합을 위해 RDF(Resource Description Framework) 방식의 문서저장을 지원하여 신속하고 정확한 정보검색이 가능하다. 이는 다양한 웹 브라우저를 지원하며 웹에서의 효율적인 데이터 검색 분야에 활용될 것이다.

ABSTRACT

The Information retrieval system is more accurate of the information for you want to search, and quickly delivered. But the current search system is a simple way to parse on users fail to provide accurate information.

This paper describes the ontology servers retrieve information through the system. Proposed system is Semantic Web-based information retrieval techniques in addition to structured documents using a variety of formats to maximize their data processing. In addition, interoperability and data integration RDF (Resource Description Framework) for saving documents by supporting rapid and accurate information retrieval. This supports a variety of Web browsers on the Web will be utilized in the field of efficient data retrieval.

키워드

Slug, Jena, Solr, RDF, Ontology

Ⅰ. 서론

IT 정보 기술의 발전과 함께 데이터양도 급격하게 증가하고 있다. 대용량의 데이터들 중에서 사용자가 필요한 정보들을 탐색하는데 어려움이 발생하고, 기존 검색 처리 기술의 한계에 도달하여 검색의 정확성이 저하되는 문제점이 발생하였다. 이러한 문제를 해결하기 위해서 시맨틱 웹 분야에 관심을 가지고 연구와 개발이 진행되고 있다[1]. 시맨틱 웹은 W3C의 워킹 그룹을 통하여 RDF와 GRDDL(Gleaning Resource Descriptions form Dialects of Languages), RDFa in XHTML, OWL 등을 개발했다[2].

본 논문에서는 시맨틱 웹 검색 시스템 구현을 위해 사용되는 크롤러와 RDF 프레임워크를 분석하고, 시맨틱 웹 기반의 검색 시스템 구현에 관한 방향을 제시한다.

Ⅱ. 관련 연구

시맨틱 기반의 검색 시스템에서 사용되는 도구들은 일반적인 웹 검색 시스템과 다르다. 본 장에서는 시맨틱 검색에서 사용할 수 있는 Slug[3]와 Jena[4], Lucene, Solr에 대해 기술한다.

2.1 Slug

Slug는 자바와 JenaAPI를 사용하여 구현한 웹 크롤러이며 검색, 처리 및 수집된 콘텐츠의 저장소 구성에 유연하게 구성할 수 있는 프레임워크이다. Slug는 메타데이터를 수집하고 RDF 어휘를 설명할 수 있게 제공하며 메타데이터 보고 및 HTTP 캐싱 데이터의 저장소를 통해 보다 효율적인 검색 크롤링 진행 상황을 분석 할 수 있다. 또한, RDF 문서의 탐색을 자동 처리하고 지속적인 모델을 사용하여 검색된 RDF 데이터들을 저장한다[5].

2.2 Jena

Jena는 시맨틱 웹에 대한 응용 프로그램의 작성을 위한 자바 프레임워크이며, RDF 저장소의 생성 및 조작을 위한 인터페이스와 클래스를 제공한다. 또한, OWL 기반의 온톨로지의 관리를 위한 클래스와 인터페이스를 제공하며 RDF API와 OWL API, SPARQL 쿼리 엔진을 지원한다.

2.3 Lucene과 Solr

Lucene은 자바로 작성된 텍스트 검색 엔진 라이브러리이며, 텍스트 색인 및 검색 기능을 추가할 수 있도록 지원한다. 또한, 크로스 플랫폼을 필요로 하는 응용프로그램에서 Lucene을 사용할 수 있으며, 다른 프로그래밍 언어의 인덱스를 호환한다.

Solr는 오픈 소스 엔터프라이즈 검색 플랫폼으로서 Lucene을 기반으로 개발되었다. Solr의 주요 기능은 전체텍스트 검색, 히트 강조, 데이터베이스 통합, 다양한 문서 처리 등이 있으며 메타데이터를 지원하고, Lucene을 사용하여 원하는 기능을 구현할 수 있다[6].

Ⅲ. 시스템 설계

제안하는 시스템의 구조는 그림 1과 같으며, 크롤러와 RDF 프레임워크, 검색 시스템으로 분류된다.

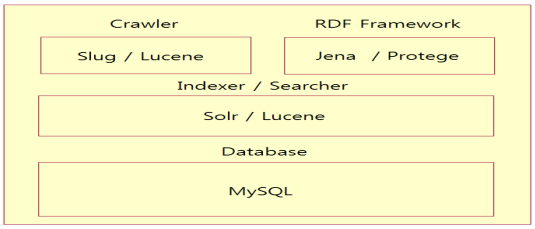

그림 1. 제안하는 시스템 구조도

3.1 정보 수집

Slug를 사용하여 인터넷에 존재하는 정보들을 추출한다. 추출한 정보들은 데이터베이스에 저장한다. 데이터베이스에 저장된 정보를 Jena에게 전송하고 Jena를 사용하여 RDF를 생성한다. 생성된 RDF는 포커스 크롤러를 사용하여 RDF를 다시 추출한다. 추출한 RDF의 추출 데이터는 Slug 웹 크롤러에 전송하여 로컬 캐시를 생성하고 데이터베이스에 저장한다.

3.2 온톨로지 관리 및 생성

온톨로지는 제안하는 검색 시스템의 기초이며, 온톨로지 요소들을 잘 설정하고 관리해야 검색의 성능을 향상시킬 수 있다. 온톨로지의 설계는 Protege를 사용하면 편리하게 작성할 수 있으며, Protege의 인터페이스는 그림 2와 같다.

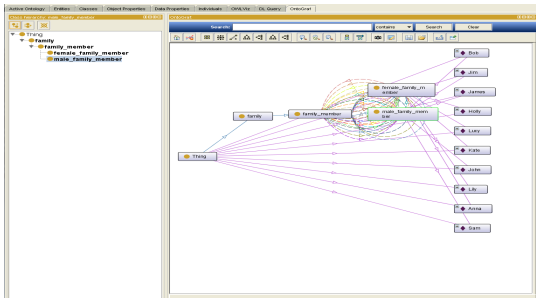

그림 2. Protege 인터페이스

시맨틱 정보 검색의 정확성을 향상시키긴 위해선 Jena의 RDF API와 OWL API를 사용하여 RDF와 OWL을 작성한다.

3.3 인덱싱 및 쿼리 처리

Slug에 의해 생성된 로컬 캐시와 RDF, OWL들은 Solr의 인덱싱 과정을 통해 쿼리 처리에 적합한 인덱스로 변환하고 Solr가 사용하는 데이터베이스에 저장한다. 그리고 사용자의 요구하는 쿼리에 맞게 결과 값을 웹 브라우저를 통해 전달하고, 검색에 사용된 키워드의 횟수와 사용자 의견을 반영하여 Solr에서 우선 순위를 재조정하여 정확성을 개선한다. 그림 3은 제안하는 시스템의 흐름도를 표현했다.

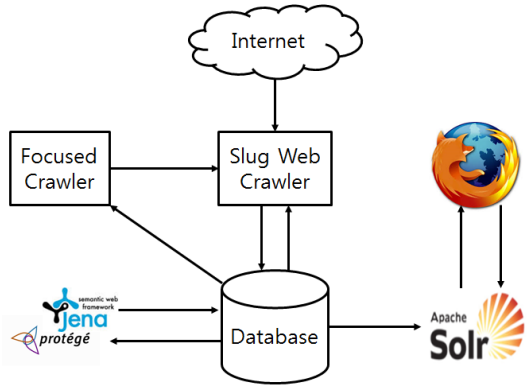

그림 3. 시맨틱 웹 기반 정보검색 시스템 흐름도

Ⅳ. 결 론

본 논문에서는 시맨틱 웹 기반의 정보 검색 시스템 모델을 제안했다. 웹 사이트의 메타 데이터 추출은 Slug를 사용하고, RDF 생성은 Jena를 사용했다. 생성된 RDF는 포커스 크롤러로 추출하여 Slug로 전송하고, Slug에서 로컬 캐시를 생성하여 데이터베이스에 저장한다. 로컬 캐시는 Solr에서 인덱싱하여 검색 시스템을 구축한다. 제안하는 시스템은 자바로 구현하여 기존 시스템에 비해 이식성과 확장성이 높다. 또한, 오픈 소스를 사용하여 기존 시스템에 비해 구현이 쉽다.

향후 연구로는 시맨틱 웹 검색의 정확성을 개선하는 방향에 대한 연구가 필요하다.

참고 문헌

[1] Berners-Lee, Tim, James Hendler, Ora Lassila, “The semantic web,“ Scientific american, pp.28-37, 2001

[2] http://www.w3.org/2001/sw/BestPractices/, 2014.4

[3] Dodds, Leigh, “Slug: A semantic web crawler,“ Proceedings of Jena User Conference, 2006.2

[4] Ganapathy, Gopinath, Sagayaraj, “To Generate the Ontology from Java Source Code‖,“ International Journal of Advanced Computer Science and Applications, Vol.2, No.2, pp.111-116, 2011

[5] Saxena, Priyanka, Masoud Nosrati, “Review of Slug Semantic Web,“ World Applied Programming, Vol.2, No.1, pp.34-37, 2012.1

[6] Devarakonda, Ranjeet, et al. “Mercury: reusable metadata management, data discovery and access system,“ Earth Science Informatics, Vol.3, pp.87-94, 2010.6