RDF 기반 시맨틱 웹 시스템 설계

이종원 · 장기만 · 김경환 · 양새동 · 정회경

배재대학교 컴퓨터공학과

Design for RDF-based Semantic Web System

Jong-Won Lee · Ki-Man Jang · Kyng-Hwan Kim · Xitong Yang · Hoe-Kyung Jung

Department of Computer Engineering, PaiChai University

E-mail : starjwon@naver.com, {jangkiman, shwan10, withchyang1}@gmail.com, hkjung@pcu.ac.kr

요 약

현재의 웹은 점점 늘어가는 데이터로 인해 효율적인 검색과 관리가 어려워지고 있다. 이를 타개하기 위한 방법으로 차세대 웹인 시맨틱 웹 기술이 개발되고 있으나, 기존에 사용되고 있는 검색엔진들은 시맨틱 웹 기술을 도입하지 않음에도 압도적인 국내 사용률을 독점하고 있다. 이로 인해 시맨틱 웹에 대한 개발은 더뎌지고 있으며, 검색엔진을 사용하는 사용자들 역시 시맨틱 웹의 사용을 꺼려하고 있다.

본 논문에서는 현재 사용되고 있는 웹과 차세대 웹을 비교분석하며, 시맨틱 웹 기술을 사용하는 검색엔진이 기존 웹 기술을 사용하는 검색엔진에 비해 사용률이 왜 낮고, 무엇 때문에 비효율적인지 연구하였으며, RDF 기반으로 시맨틱 웹을 설계하여 효율성을 높일 해결방법을 제시한다.

ABSTRACT

It is difficult to effectively search and data management due to the increasing number of web is now. While Semantic Web technologies and the development of next-generation wepin this as a way to overcome them, and monopolize the domestic utilization is not overwhelming introduction to the Semantic Web technology is being used in existing search engines. This causes the development of the Semantic Web is becoming slower, and reluctant to use the Semantic Web users who use search engines as well.

In this paper, compared to the currently used web and the next generation of the web, and why utilization is low compared to the search engine you are using an existing Web technology that uses the Semantic Web technology is a search engine, what research was that the inefficient because, as a RDF-based Semantic suggest how to improve the efficiency solved by designing the web.

키워드

RDF, 시맨틱 웹, 크롤러, 검색엔진

Ⅰ. 서 론

현재 IT업계는 유용하고 효율적인 기술 개발에 힘쓰고 있고 그 원인은 데이터의 양이 급속도로 많아지고 사용할 수 있는 기술들이 늘어남에 있다. 그 중 데이터를 사용자에 맞게 가공하기 위해서 데이터들을 수집과 정제, 데이터베이스로 통합하는 ETL(Extraction, Transformation, Loading) 도구에 대한 연구는 차세대 웹 기술 개발과 확립에 필요하다. ETL 도구는 다양하며 Open Source 로도 제공되고 있고, 보통은 개발자가 개인의 프로그램을 만들어서 사용한다[1,2]. 이런 개인의 프로그램으로 가공한 데이터를 다수가 사용 가능한 데이터로 변환하기 위해서는 시맨틱 웹 기술 중 하나인 RDF(Resource Descripstion Framework) 기술을 사용해야 한다. RDF는 데이터를 정보로 가공하면서 해당 데이터가 어떤 데이터인지 서술하여 정보 제공자가 아닌 사용자가 그 정보를 사용할 때 일어날 수 있는 정보에 대한 혼란을 피하는 시맨틱 웹의 틀을 제공하게 되고 Ontology를 통해 정보 사이의 관계성을 부여, 컴퓨터가 이해할 수 있게 한다[3].

시맨틱 웹은 차세대 웹 기술로 불리며 현재의 웹 기술이 HTML 언어로 이루어져 있으며 인간의

편리성을 추구하게끔 구성되어 있다. 그러나 인터넷의 활용도가 높아지고 검색 정보량의 대령화로 인해 인간의 눈에 의한 정보관리 및 탐색의 어려움이 증가하게 되었고, 현재 웹의 표현으로 정보를 표현하는데 한계성을 띄게 되었다. 이러한 문제점들을 극복하기 위해서는 웹 문서에 대한 의미정보를 두어 검색 시 중복 검색이나 정확성이 떨어지는 데이터들을 제외시킨 검색 결과를 보여줌과 동시에 컴퓨터가 정보 자체를 이해하여 의미 있는 정보 추출이 이루어져야 한다.

본 논문에서는 RDF 기반의 Ontology 구축 시 컴퓨터 중심의 시맨틱 웹 기술의 문제점을 설명하고 그를 극복할 방법을 제시한다.

Ⅱ. 관련 연구

온톨로지를 기술하기 위한 언어로서 W3C에서 제정한 시맨틱 웹의 일종인 RDF와 RDF의 확장으로서 웹 온톨로지 구축을 위한 OWL, ISO에서 제정한 TopicMap 등이 있다.

2.1 RDF

RDF는 W3C에서 제정한 것으로서 기술하고자 하는 대상에 대한 부가정보, 데이터간의 상하 및 연관 관계 등을 기술하는 능력을 가진다. 데이터를 정의하고 데이터에 대한 설명이나 관계를 기술함으로써 온톨로지를 구축할 수 있는 방법을 제공한다.

RDF는 기본적으로 트리플 모델로 기술되는데 주어(Subject), 서술(Predicate), 목적(Object)으로 이루어져 있다. 주어란, 표현하고자 하는 데이터를 의미하며, 서술은 주어에 대해 기술하거나 주어와 목적의 관계를 의미한다. 목적이란 서술에 대한 내용이나 값을 의미하며, 각 내용들에 대해서 URI를 통해 기술할 수 있다[4,5].

2.2 시맨틱 웹

시맨틱 웹은 차세대 웹으로 표현되고 있으며, 인간의 언어를 이해하고 인간과 쉽게 의사사통이 가능해진 네트워크, 또한 컴퓨터 스스로 웹에 연결된 정보의 의미를 인식하고 사용자가 필요로 하는 정보를 검색하며 검색된 정보에서 지식을 유추할 수 있는 기능을 제공하는 지능형 웹 환경을 일컫는다. 등장배경으로는 인터넷의 활용도가 높아지고 검색 정보량의 대량화로 인해 인간의 눈에 의한 정보관리 및 탐색의 어려움이 증가하였고, HTML의 한계점에 따른 인터넷 상의 데이터 관리의 비효율성, 현재 웹의 표현으로 자원을 표현하는데 한계성이 존재하기 때문이다. 웹 검색 Agent가 문서로부터 의미를 자동 추출을 하지 못하고, 입력한 키워드나 주제 분야에 알맞은 URL 주소를 찾아주는 단순성에 그 문제점이 있다.

시맨틱 웹은 기존 웹과 같이 단어를 식별해서 관련된 사이트나 문서를 찾아줌과 동시에 새롭게 구성된 문서에 사물간의 관계를 명확히 기술하여 정확하고 의미있는 정보 제공에 목표가 있다[6].

Ⅲ. 제안하는 방법

3.1 시맨틱 웹의 계층 구조

시맨틱 웹의 핵심기술로써 첫째, 자원 서술 기술인 RDF를 통해 자원들을 구조화 시킨 뒤, 구조화된 자원들의 의미를 지정한다. 둘째, 지식 서술 기술은 개념의 체계적인 규정을 의미하며, 자원과 자원간의 관계, 용어와 용어간의 관계를 표현하는 컴퓨터 판독이 가능한 공식 규정을 말한다. 특정 주제에 관한 지식용어들의 집합으로서 이들 용어뿐만 아니라 용어간의 의미적 연결 관계와 간단한 추론 규칙을 포함하게 된다. 셋째, 인간을 대신하여 정보 자원을 수집, 검색하고 추론하여 메타데이터와 온톨로지를 이용해서 다른 Agent와 상호 정보 교환 등의 일을 수행하는 지능형 컴퓨터 프로그램이다. 자원 서술을 위한 RDF나 지식 서술을 위한 온톨로지가 비교적 정적인 구조를 가진 반면, Agent는 이러한 자원정보와 지식을 바탕으로 사용자의 요구에 맞게 정보를 추출하고 가공하여 제공하는 동적인 역할을 담당하게 된다.

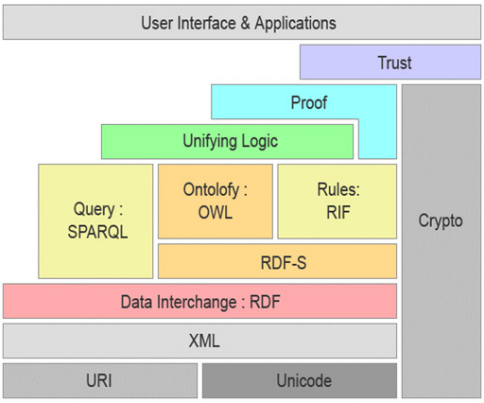

그림 1. 시맨틱 웹의 계층 구조

그림 1의 시맨틱 웹의 계층 구조는 URI로 자원을 표현하고, 각 나라의 언어가 다른 점을 Unicode의 사용으로 정보 전달에 있어서 생길 수 있는 문제점을 대비하고, XML을 이용해 데이터를 표현하게 된다. 그리고 자원들의 관계를 서술하는데 RDF가 이용되는데 RDF가 갖고 있는 문법보다 RDF-S가 갖는 문법이 보다 다양하게 자원들의 관계를 서술할 수 있다. 이 말은 RDF가 가지는 문제점을 말해준다. 자원간의 관계를 표현하는 데 있어서 정의할 수 있는 관계가 한정되어 있다는 것이다.

3.2 포털사이트와 시맨틱 웹의 관계

이러한 문제점은 실제로 국내 대형 포털사이트들의 사용률로 나타난다. 시맨틱 웹을 적극적으로 이용하여 검색엔진을 업데이트하는 곳보다 사용

자 중심의 검색결과를 나타내는 포털사이트의 사용률이 훨씬 높은 것이다. 포털사이트들의 사용률은 검색엔진의 활용도와 부가서비스에 달려있다. 차세대 웹으로 일컫는 시맨틱 웹을 사용함에도 불구하고 다른 포털사이트에 밀리는 이유로는 RDF를 통한 자원 서술에 한계점이 존재하고, 그 한계점을 극복하지 못하기 때문에 검색엔진으로써 경쟁력을 갖추지 못했다는 평이다. 아래의 표는 시맨틱 웹 사용여부와 국내에서 사용되는 포털사이트 순위에 관한 표이다.

표 1. 포털사이트 순위와 시맨틱 웹 사용여부

순위	시맨틱 웹 사용여부
1	부분적 사용
2	부분적 사용
3	전면적 사용

3.3 RDF 표현 규칙

그림 2. RDF 트리플 모델과 공백노드

본 논문에서 제안하는 방법은 그림 2와 같이 RDF의 표현 규칙에 사용자가 주로 사용하는 주어들의 관계를 공백노드를 통해 설정하고 결과로 볼 수 있게끔 하는 방법이다. 이는 현재 웹을 사용하고 있는 사용자들이 가장 원하는 검색엔진으로써의 기능이며 단순히 차세대 웹 기술을 구현한다 해서 사용률이 높아지는 것이 아니기 때문이다. 기존에 사용되던 RDF 트리플 모델에 사용자가 어떤 주어의 사용이 많았고 적었는지를 알 수 있게끔 로그 기록을 첨가하여 주어간의 관계를 정의한다면 현재 사용되고 있는 RDF 표현 규칙보다 다양하며 효율적인 표현 규칙이 된다.

Ⅳ. 결 론

크롤러가 데이터를 수집하면 수집된 데이터들을 바탕으로 관계를 정의하고 표현하여 의미 있는 정보 검색 및 결과를 보여주는 것이 시맨틱 웹이고, 현재의 웹이 가지는 단점들을 극복할 방법으로 평가되고 있다. HTML이 가지는 한계점으로는 많은 정보를 비효율적으로 사용자에게 보여주고, 인간의 눈에 의한 수동적인 정보관리 및 탐색이 어려워진다는 점이 있다. 정보가 많아지면 많아질수록 수동적인 방법으로는 관리나 탐색이 더욱 어려워질 것이다. 이러한 문제점을 극복하기 위해 시맨틱 웹은 많은 데이터들을 정리하기 위해 사용자가 원하는 데이터와 관련된 데이터만을 정리하여 보여주기 때문에 효율적이다. 그러나, 시맨틱 웹에서 데이터간의 관계를 정리하기 위해 사용되는 RDF가 가지는 유한성에 대한 문제점이 있다.

본 논문에서는 기존의 시맨틱 웹을 구현할 때 RDF 표현 규칙이 가지는 단점을 제시하였고, 해결 방안으로 사용자 중심의 기록을 추가하여 표현 규칙을 다양하고 효율적으로 사용하는 것이다.

향후 연구로는 RDF가 가지는 단점들에 대해 연구와 실험을 계속 진행하여 제시하고 있는 시스템의 유용성과 효율성 개선 방향에 대한 연구가 필요하다.

참고 문헌

[1] Hanhoon Kang, Seong Joon Yoo, Dongil Han, “Design and Implementation of Web Crawler Wrappers to Collect User Reviews on Shopping Mall with Various Hierarchical Tree Structure”, Korean Institute of Intelligent Systems, Vol.20, No.3, pp.318-325, 2010.6

[2] Chenghao Quan, Youngtak Lee, Youngjun Kim, Yongdoo Lee, “Design and Implementation of a High Performance Web Crawler“, Journal of the Korea Industrial Information Systems Research, Vol.8, No.4, pp.64-72, 2003.12

[3] Hoansuk Choi, Junyoung Lee, Nari Yang, Wooseop Rhee, “Ontology Based User-centric Service Environment for Context Aware IoT Services“, The Korea Contents Association, Vol.14, No.7, pp.29-44, 2014.7

[4] Junwon Jung, Hoyoung Jung, Jongnam Kim, Donghyuk Lim, HyoungJoo Kim, “A RDF based Ontology Management System“, Korean Institute of Information Scientists and Engineers, Vol.11, No.4, pp.381-392, 2005.8

[5] Jihyoung Park, Myungjin Lee, Juneseok Hong, “A study on Development of RDF Triple Storage System for Retrieval of Metadata in the Semantic Web“, Journal of Society for e-Business Studies, Vol.12, No.2, pp.291-304, 2007.5

[6] Byoungjun Kim, Deokmin Haam, Inchul Song, Kiyong Lee, Myoungho Kim, “A Method of Ranking Structured Queries for Keyword Search on Semantic Web Data“, Vol.11, No.4, pp.381-392, 2005.8“, Korean Institute of Information Scientists and Engineers, Vol.39, No.2, pp.138-146, 2012.4