JKIICE Journal of the Korea Institute of Information and Communication Engineering

한국정보통신학회논문지(J. Korea Inst. Inf. Commun. Eng.) Vol. 19, No. 4 : 871~876 Apr. 2015

시맨틱 웹 기반 사용자 중심 검색시스템에 관한 연구

김창수 · 이종원 · 정회경*

A Study on Semantic Web based User Oriented Retrieval System

Chang-Su Kim · Jong-Won Lee · Hoe-Kyung Jung*

Department of Computer Engineering, Paichai University, Daejeon 302-735, Korea

요 약

현재의 웹은 점점 늘어가는 데이터로 인해 효율적인 검색과 관리가 어려워지고 있다. 이를 개선하기 위한 방법으로 시맨틱 웹 기술이 개발되고 있다. 그러나 현재 사용되는 검색시스템들은 시맨틱 웹 기술을 도입하지 않음에도 압도적인 국내 사용률을 독점하고 있다. 이로 인해 시맨틱 웹에 대한 개발은 활성화 되지 않고 있으며, 검색시스템을 사용하는 사용자들 역시 시맨틱 웹의 사용률이 저조한 실정이다.

이에 본 논문에서는 현재 사용되고 있는 검색시스템을 분석하고, 제안하는 시스템의 온톨로지 구현 시 사용자가 사용한 데이터의 종류와 웹 서버의 게시판 사용 시 사용한 파일의 종류를 RDF 표현 규칙에 추가 설정하여 사용자 중심의 검색시스템을 설계 및 구현하였다.

ABSTRACT

Recently, the Web is becoming more difficult to manage efficiently retrieve with the increase of data. However, the retrieval systems that are currently used have not been applied to the Semantic Web technology. Thus, the development of the Semantic Web is not activated. User of the retrieval system also the Semantic Web usage is low is the situation.

In this paper, we are analyzed the retrieval system that is currently being used. we are proposed added the rule of the RDF representation during ontology implementation of the retrieval system. And we propose the user-centric of retrieval system design and implementation.

키워드 : 검색시스템, 시맨틱 웹, 온톨로지, RDF

Key word : Retrieval System, Semantic Web, Ontology, RDF

Received 02 January 2015, **Revised** 03 February 2015, **Accepted** 17 February 2015

* **Corresponding Author** Hoe-Kyung Jung(E-mail:hkjung@pcu.ac.kr, Tel:+82-42-520-5640)
Department of Computer Engineering, Paichai University, Daejeon 302-735, Korea

Open Access http://dx.doi.org/10.6109/jkiice.2015.19.4.871 print ISSN: 2234-4772 online ISSN: 2288-4165

Ⅰ. 서 론

현재 IT업계는 유용성과 효율성 측면의 기술 개발에 힘쓰고 있고 그 원인은 데이터의 양이 급속도로 많아지고 사용할 수 있는 기술들이 늘어남에 있다. 현재의 웹 기술은 HTML 언어로 이루어져 있으며 인간의 편리성을 추구하게끔 구성되어 있다. 그러나 인터넷의 활용도가 높아지고 검색 정보량의 대량화로 인해 인간의 시각에 의한 정보관리 및 탐색의 어려움이 증가하게 되었고, 현재 웹의 표현으로 정보를 표현하는데 한계성을 띄게 되었다. 이에 따라 기존 웹의 문제점을 해결하기 위한 방안으로 시맨틱 웹 기술이 대두되었다. 시맨틱 웹 기술의 핵심은 웹 문서에 대한 의미정보를 두어 검색 시 중복 검색이나 정확성이 떨어지는 데이터들을 제외시킨 검색 결과를 보여줌과 동시에 컴퓨터가 정보 자체를 이해하여 의미 있는 정보를 추출하는 데 있다[1,2].

본 논문에서는 현재 시맨틱 웹 기술을 사용하는 검색 시스템이 갖는 한계성에 대해 분석하고 해결방안을 제시한다. 현재의 시맨틱 웹 기술이 RDF를 기반으로 온톨로지를 구현하는데, RDF(Resource Definition Framework)는 자원과 자원 사이의 관계들을 스키마 정보로 설정하고, 이 값들을 중심으로 검색시스템이 동작하게 된다. 이 때, RDF로 기술하지 않아 스키마 정보에 없는 자원들의 관계 값은 검색결과로 나오지 않기 때문에 현재의 웹이 가지는 유한성을 해결하기 위한 방안인 시맨틱 웹 기술 또한 유한성을 갖게 된다. 이 문제를 해결하기 위해 컴퓨터 중심의 시맨틱 웹 기술의 문제점을 설명하고 그를 극복할 방법으로 RDF를 기반으로 사용자 중심의 검색시스템 설계 및 구현 방안을 제시한다[3,4].

Ⅱ. 관련 연구

온톨로지를 기술하기 위한 언어로서 W3C에서 제정한 RDF와 RDF의 확장으로서 웹 온톨로지 구현을 위한 OWL, ISO에서 제정한 TopicMap 등이 있다.

2.1. 시맨틱 웹

시맨틱 웹은 차세대 웹으로 표현되고 있으며, 인간의 언어를 이해하고 인간과 쉽게 의사소통이 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 URL주소를 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 의미 있는 정보 제공에 목표가 있다.

2.2. 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 구현할 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[5,6].

2.3. 검색시스템

검색시스템은 사용자가 찾고자하는 정보를 정확하게 찾는 것이 중요하다. 그러기 위해서는 크롤러가 데이터를 수집하고, 데이터베이스에 저장한 뒤 질의를 통해 저장된 데이터를 결과로 볼 수 있게끔 Indexing 작업이 필요하다. 현재의 검색시스템이 단어 그 자체와 관련된 데이터를 찾는 것이라면 차세대 웹은 단어가 가지는 의미를 검색시스템에서 이해하고 그와 관련된 데이터를 찾아주는 것이 목적이다.

Ⅲ. 시스템의 설계

제안하는 시스템은 크롤러가 수집한 데이터와 웹 서버에 등재된 게시물과 첨부파일들을 데이터베이스에 저장하여 온톨로지에서 Resource로 사용하되, 사용자가 어떤 Resource를 많이 사용하고 검색하였는지 Counting하여 우선순위를 매겨서 검색결과로 보여준다. 크롤러는 Eclipse 툴을 이용하여 Java로 구현하였고 HTML 크롤러와 URL 크롤러 두 가지이다. 크롤러가 수집한 데이터는 크롤러 전용으로 사용할 데이터베이스에 저장된다. 웹 서버는 PHP로 구현하였으며 부분적으로 HTML을 이용하였다. 게시판에 등재된 게시물과 첨부파일은 웹 서버용 데이터베이스에 저장되고 모든 데이터들은 온톨로지를 통해 Resource로 사용된다. 그림 1은 시스템 구성도이다.

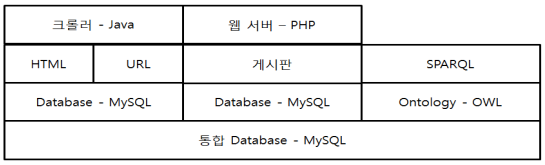

그림 1. 시스템의 구성도
Fig. 1 System Configuration

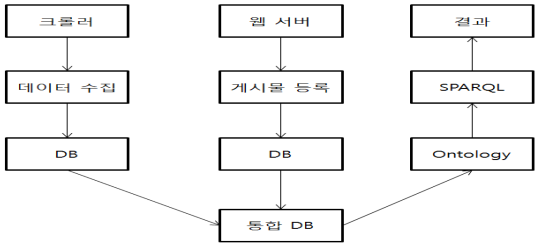

그림 2. 시스템의 흐름도
Fig. 2 System Processing

크롤러가 수집한 데이터를 사용할 수 있고, 웹 서버의 게시된 데이터를 사용할 수 있다. 크롤러와 웹 서버는 독립적으로 실행되며 온톨로지를 구현할 땐 통합 데이터베이스를 따로 사용함으로써 질의가 가능하게 하였다. 그림 2는 시스템 흐름도이다.

3.1. 크롤러 설계

크롤러는 데이터를 수집하는 작업을 하는 툴로써 수집 범위 설정과 수집 대상 설정으로 시작한다. 데이터 수집은 수집 범위가 좁을수록, 수집 대상이 적을수록 정확도가 높아진다. 그러나 정확도를 높이기 위해 앞에서 말한바와 같이 크롤러를 작동시키면 수집되는 데이터의 양이 적어 활용할 수 있는 범위 역시 좁아지게 된다. 이 문제점을 해결하기 위해 크롤러는 수집되는 데이터의 양과 정확도를 올리기 위해 이중 크롤러를 구현한다. 또한, 분산 처리로 작동하게 하여 효율성을 높인다.

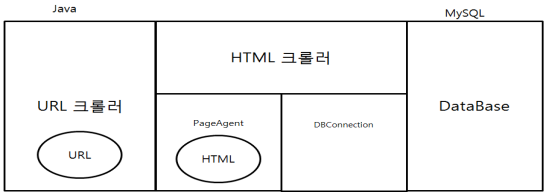

그림 3. 분산 처리 크롤러의 구성도
Fig. 3 Distribute Crawler Configuration

분산 처리 크롤러는 URL 크롤러가 사용자의 검색에 의해 가장 먼저 작동하며 시작된다. 해당 String 값이 있는 URL들을 URL List에서 검색한 뒤 그 결과 값을 출력한다. 다음으로 HTML 크롤러가 결과 값으로 나온 URL들의 HTML Source를 수집하여 데이터베이스에 저장한다. 저장된 HTML Source들은 구현된 온톨로지에서 Resource로 사용된다.

3.2. 온톨로지 설계

본 논문에서 제안하는 방법은 그림 4와 같이 RDF의 표현 규칙에 사용자가 주로 사용하는 주어들의 관계를 공백노드를 통해 설정하고 결과로 도출할 수 있는 방법이다. 이는 현재 웹을 사용하고 있는 사용자들이 가장 원하는 검색시스템의 기능이며 차세대 웹의 구현을 통해 검색시스템의 사용률을 높이는 것이 아니다. 기존에 사용되던 RDF 트리플 모델에 사용자의 주어 사용이 많고 적었는지를 알 수 있게 로그 기록을 첨가하여 주어 간의 관계를 정의한다면 현재 사용되고 있는 RDF 표현 규칙보다 다양하며 효율적인 표현 규칙이 된다.

RDF 표현 규칙을 사용자 중심으로 수정하면서 단순히 검색하려는 데이터와 관련된 데이터를 검색 결과로 보여주는 의미론적 검색 수준을 넘어 사용자가 자주 사용하던 데이터와 원하는 데이터, 이와 관련된 데이터를

보여줌으로써 시맨틱 웹의 궁극적인 목표를 기존의 사용되고 있는 웹보다 높은 수준으로 보여줄 수 있다.

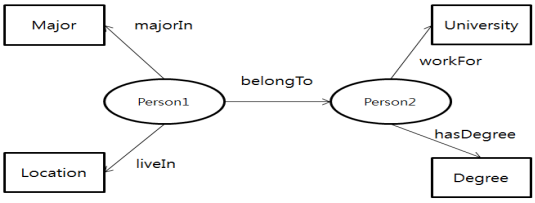

그림 4. RDF 트리플 모델과 공백노드
Fig. 4 RDF triple model and blank nodes

3.3. 데이터베이스 설계

데이터베이스는 Table 2개를 구현하여 크롤러와 웹 서버가 서로 다른 Table을 사용할 수 있게 한 뒤 데이터들을 저장 및 관리하게 된다. 크롤러 전용 Table1은 Field 값으로 URL, HTML Source, 방문여부를 저장하게 된다. URL 값은 테이블의 Key가 된다.

웹 서버 전용 Table2는 Field 값으로 게시한 글의 번호, 이름, 비밀번호, E-mail, 홈페이지, 제목, 내용, 조회수, Ip, 작성시간, 첨부파일을 저장하게 된다. 게시한 글의 번호는 테이블의 Key가 된다.

Ⅳ. 시스템의 구현

본 장에서는 제안 시스템의 구현을 다룬다. 구현 환경은 동일한 단일 컴퓨팅 환경이며, 표 1과 같다.

표 1. 구현 환경
Table. 1 Experimental Environment

CPU	I5-4670 @ 3.40GHZ
RAM	8.00GB
운영체제	Window7 64Bit

제안하는 시스템은 데이터를 검색 및 수집해올 크롤러와 사용자들이 사용할 웹 서버와 데이터베이스, 수집된 데이터나 사용자들이 사용한 데이터의 관계를 규정할 온톨로지로 구성된다.

4.1. 크롤러 구현

URL 크롤러는 원하는 문자열을 검색한 뒤 해당 문자열이 존재하는 URL값을 수집하여 결과로 나타내며, HTML 크롤러는 URL 크롤러의 결과 값들로 직접 이동해 HTML Source를 수집한다.

HTML 크롤러는 멀티 쓰레딩을 이용하여 데이터를 수집하는데, 필요한 기능들이 HTML 소스 중 어떤 태그들을 수집할 것인지를 결정하는 webCrawling Method, 검색되는 URL에서 링크가 있는지 확인하는 extractLinkFrom Method, 방문을 기피하는 링크인지 확인하는 badLink Method, 수집하지 않을 확장자를 설정하는 badUrl Method가 있다. 그림 6은 HTML 크롤러의 기능들이다.

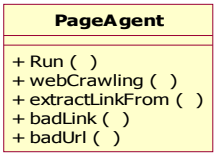

그림 5. HTML 크롤러의 기능
Fig. 5 HTML Crawler Features

4.2. 웹 서버 구현

웹 서버 전용 Table2의 구성은 그림 6과 같다.

	필드	종류	Collation	보기	Null	기본값	추가
☐	number	int(10)		UNSIGNED	아니오	None	auto_increment
☐	name	varchar(12)	utf8_bin		아니오	None	
☐	password	varchar(16)	utf8_bin		아니오	None	
☐	email	varchar(50)	utf8_bin		아니오	None	
☐	homepage	varchar(60)	utf8_bin		아니오	None	
☐	subject	varchar(60)	utf8_bin		아니오	None	
☐	memo	text	utf8_bin		아니오	None	
☐	count	smallint(5)		UNSIGNED	아니오	None	
☐	ip	varchar(15)	utf8_bin		아니오	None	
☐	writetime	int(10)		UNSIGNED	아니오	None	
☐	file_name1	varchar(255)	utf8_bin		예	NULL	
☐	s_file_name1	varchar(255)	utf8_bin		예	NULL	

그림 6. 웹 서버의 데이터베이스 Field들
Fig. 6 Web Server 데이터베이스 Fields

4.3. 온톨로지 구현

온톨로지는 Topbraid 툴을 이용하여 구현하였으며, 먼저 RDFS로 사용할 스키마를 설정하고, Subject와 Object로 사용할 Class는 OWL을 사용하여 생성하고, 스키마의 내용대로 RDFS를 설정하였다.

구현이 완성된 온톨로지는 크롤러의 수집된 데이터나 웹 서버에 등재된 게시물을 Resource로 동작하며 Topbraid로 사용자를 만들고, 그 사용자가 어떤 검색어와 Resource를 자주 사용했는지 Counting 함수를 통해 체크한 뒤 검색 결과로 가장 자주 사용했던 데이터를

우선순위로 보여주게 된다. 그림 7은 Topbraid 툴을 이용하여 온톨로지를 구현하기 위해 필요한 데이터들을 스키마로 설정한 것이다.

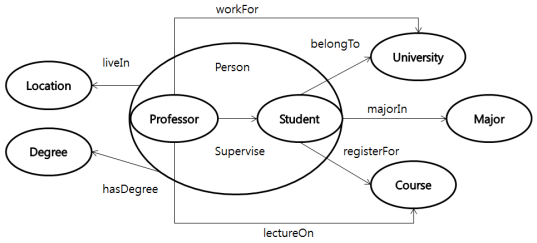

그림 7. 스키마 설정
Fig. 7 Schema Creation

RDF 표현 규칙은 트리플 모델로 사용하였고, 주어 값으로 학생을 설정한 뒤, 관계식 실험을 하였다. 표 2는 기존 시맨틱 웹에서 사용하는 트리플 모델을 정리한 것이고, 표 3은 제안하는 시스템의 트리플 모델을 정리한 것이다.

표 2. 기존 웹의 트리플 모델
Table. 2 Previous Triple Model

Subject	RDF	Object
Student1	supervise	Professor1
Student1	supervise	Professor2
Student2	majorIn	Computer
Student2	majorIn	literature

표 3. 제안 시스템의 실험
Table. 3 Propose System Experiment

Subject	Search1	Seaech2	Count
Student1	majorIn	supervise	3,1
Student2	majorIn	Course	3,2
Student3	majorIn	Course	3,2
Student4	liveIn	hasDegree	1,1

표 2는 주어, 관계, 목표로 이루어진 트리플 모델이고, 표 3은 주어, 관계 1, 관계 2, Count로 구성되어 있다. 관계 1과 관계 2는 검색어를 의미하며, 위와 같은 구성을 한 이유는 다음과 같다. 포털사이트에서 검색횟수를 Count하여 검색순위를 매긴 뒤 실시간으로 제공해주는 작업을 제안하는 시스템에서는 RDF 표현 규칙으로 지정한 것이다.

표 3을 보면 Count 결과로 majorIn 3회, Course 2회, supervise와 hasDegree 가 1회인 것을 알 수 있다. 학생들의 주된 검색이 전공분야와 교육과정인 것을 알 수 있다. 학생들의 데이터는 전공분야와 교육과정 관련의 데이터와 우선적으로 연결되어야 한다. 학생들의 트리플 모델은 표 4와 같이 변경되게 된다.

표 4. 제안 시스템의 실험 결과
Table. 4 Propose System Experiment Results

Subject	majorIn	Course
Student1	Electronic	Bachelor
Student2	Computer	Master
Student3	Computer	Doctor
Student4	Mechanical	Bachelor

표 4를 보면 실험을 통한 결과로 학생들의 기본 데이터로 전공분야와 교육과정이 연결되어 있는 것을 볼 수 있다. 이와 같은 결과는 기존 트리플 모델의 주어, 관계, 목표에 주어의 기본 정보를 더함으로써 주어가 어떤 데이터와 관련이 깊은지를 쉽게 알 수 있게 해준다. 주어와 관련이 깊은 데이터를 제시해줌으로써 RDF 표현 규칙이 가지는 유한성을 주어가 유동적으로 바뀜으로써 해결하였다.

Ⅴ. 결 론

시맨틱 웹은 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. 수동적인 정보관리의 어려움을 극복하기 위해 시맨틱 웹은 많은 데이터들을 효율적으로 정리하기 위해 사용자가 원하는 데이터와 관련된 데이터만을 정리하여 결과 값으로 도출한다. 컴퓨터 스스로 검색어에 대해 인식하여 검색어와 관련된 데이터들을 보여주는 시맨틱 웹은 데이터간의 관계를 정리하는 RDF 표현 규칙이 현재 검색시스템을 사용하는 사용자들의 중심이 되지 않는 문제점이 있다. 컴퓨터가 많은 양의 데이터 중에서 사용자가 원하는 데이터를 효율적으로 제공하기 위해서는 현재 사용되고 있는 검색시스템의 시스템 구성을 되짚어볼 필요가 있다. 가장 많이 사용되는 검색시스템들은 시맨틱 웹을 전면적으로 사용하

는 것보다 부분적으로 사용하여 사용자들의 요구에 대처하고 있다.

이에 본 논문에서는 기존의 시맨틱 웹을 구현할 때 생기는 문제점으로 RDF 표현 규칙이 가지는 비사용자 중심에 대해 분석하였고, 해결 방안으로 사용자의 데이터 사용 기록을 RDF 표현 규칙에 추가하여 사용자에 따른 유동적인 결과 값을 제공하는 검색시스템을 설계 및 구현하였다.

향후 연구로는 제안한 시스템의 효율성 검증을 위해 실험을 통한 데이터를 축적한 뒤 분석 및 다양한 분야로의 연구가 필요하다.

REFERENCES

[1] Gyujin Choi, Yunhee Son, Kyuchul Lee, "Schema-based Bitmap Join Indexing Method for Semantic Web Big Data Processing", *Korean Institute of Information Scientists and Engineers*, Vol.1, No.73, pp.181-185, 2013.6.

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

[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] Jiwoong Choi, Myungho Kim, "Navigator for OWL Ontologies Generated from Relational Databases", *The Korea Contents Association*, Vol.14, No.10,pp.438-453, 2014.10.

[5] Sungjae Jung, Taehong Kim, Seungwoo Lee, Hanmin Jung, "Ontology-based Efficient RDF Query Formuation and Processing", *The Korea Contents Association*, Vol.1, No.62, pp.173-175, 2013.11.

[6] Jihye Hong, Yongkoo Han, Youngkoo Lee, "A RDF data Management Method based on Associated Properties for Efficient Query Processing", *The Korea Contents Association*, Vol.39, No.2, pp.10-12, 2012.11.

김창수(Chang-Su Kim)

1996년 배재대학교 전자계산학과(이학사)
1998년 배재대학교 전자계산학과(이학석사)
2002년 배재대학교 컴퓨터공학과(공학박사)
2005년~ 2010년 청운대학교 인터넷학과
2013년~ 현재 배재대학교 컴퓨터공학과 조교수
※관심분야 : 멀티미디어문서정보처리, 차세대 인터넷, USN, 모바일 웹서비스

이종원(Jong-Won Lee)

2014년 배재대학교 컴퓨터공학과(공학사)
2014년 ~ 현재 배재대학교 컴퓨터공학과 석사과정
※관심분야 : Android, Java, Semantic Web

정회경(Hoe-Kyung Jung)

1985년 광운대학교 컴퓨터공학과(공학사)
1987년 광운대학교 컴퓨터공학과(공학석사)
1993년 광운대학교 컴퓨터공학과(공학박사)
1994년 ~ 현재 배재대학교 컴퓨터공학과 교수
※관심분야 : 멀티미디어 문서정보처리, XML, SVG, Web Services, Semantic Web, MPEG-21, Ubiquitous Computing, USN