시맨틱 웹을 이용한 웹 변경 탐지 시스템의 설계

조부현°, 이복주
단국대학교 전자 컴퓨터공학과
{choboo°, blee}@dku.edu

Design of a Web Change Detection System Using the Semantic Web

Boohyun Cho° , Bogju Lee
Dept. of Electronic and Computer Engineering, Dankook University

요약

본 연구는 인공지능과 정보검색, 웹 기반 시스템 분야의 새로운 추세인 시맨틱 웹 기술을 이용하여 웹 문서의 변경을 자동으로 사용자에게 알려주는 시스템을 개발하고자 한다. 기존의 시스템의 문제점인 신택스 중심의 변화 탐지에서 벗어나 시맨틱 중심의 변화 탐지에 목표를 두어, 의미가 있는 변경 사항만을 찾아 알려주어 사용자에게 유용한 정보가 될 것 이다. 또한 특정 도메인에 중심 된 변경 사항을 사용자에게 알려준다면 더욱 유용한 시스템이 될 것 이다. 이를 위하여 특정 도메인과 시나리오를 가정한 온톨로지를 구축하고 이를 이용하여 시맨틱 중심의 변화 탐지를 가능하게 한다.

1. 서론

월드 와이드 웹의 발달로 인하여 개인은 정보 수집에 있어 웹에 많은 부분을 의존하고 있다. 따라서 개인의 입장에서 방문해서 확인해야 하는 웹 페이지의 수는 많아 질 수 밖에 없고 개인이 일일이 웹 페이지를 방문하여 변경된 새로운 정보를 확인하는 일은 큰 부담이 아닐 수 없다. 만약 사용자가 중요한 정보가 변경된 웹 페이지를 놓쳤을 경우 손실은 매우 클 것이다.

사용자가 관심이 있는 웹 페이지 변경의 자동 탐지에 관한 기존의 연구와 응용이 많이 있었으나 기존의 시스템이 서버 기반의 구조를 사용하여 탐지 시점을 제대로 잡지 못하고 사용자가 관심 있는 페이지를 일일이 입력해 주어야 하는 번거로움 등, 사용에 불편한 점이 많았다. 또한 기존의 시스템은 대상 웹 페이지의 과거와 현재 상태를 syntax 적으로 비교하여 사용자에게 알려주는 수준에 머물고 있다. 결과적으로 기존의 시스템에서 보고 되는 사항은 사용자 입장에서 그리 중요하지 않은 syntax적인 변화에 대한 부분이 많이 차지하고 있다. `본 연구는 인공지능과 정보검색, 웹 기반 시스템 분야의 새로운 추세인 시맨틱 웹 기술을 이용하여 웹 문서의 변경을 자동으로 사용자에게 알려주는 시스템을 개발하고자 한다. 기존의 시스템의 syntax적인 변화 탐지에서 벗어나 시맨틱한 변화 탐지를 목표를 맞춘다면 더욱 의미 있고 유용한 시스템이 될 것이다.

예를 들면 컴퓨터학 분야의 인물의 신상 정보 (직장의 이전, 승진, 학회 활동, 새로운 논문 발표)에 변화가 있을 때 자동으로 감지하여 관심 있는 사용자에게 알려주는 시스템을 생각해 볼 수 있다. 이를 위하여 인물 신상 정보를 설명하고, 시스템은 온톨로지 사이의 변형뿐만 아니라 도메인을 가정한 온톨로지를 구축하고 이를 이용하여 시맨틱 중심의 변화 탐지를 가능하게 한다.

2. 기존 시스템 구축 및 접근 방법

기존의 연구와 응용에 대해 특성과 역할에 따라 분류하여 각각의 버전간의 변화를 감지하여 찾아내는 Diff 부분, 변화 감지된 정보를 사용자에게 알려주는 방식에 관한 부분, 시맨틱 웹 관련 부분, 마지막으로 에이전트의 자동화된 서비스로 나누어 살펴본다.

첫째, 각각의 버전간의 변화를 감지하여 찾아내는 부분에서는 HTML 및 XML 문서의 변화 감지를 하는 알고리즘에 대한 기존연구로서 HTMLDiff[1]는 HTML 문서 사이의 정보 비교 뿐 아니라 태그까지 비교하여 결과를 얻어 낼 수 있다. 비교 시 사용자에게 의미가 없는 태그의 변화까지도 찾아내기 때문에 실질적으로 사용자에게 얻어지는 정보는 한계를 갖게 된다.

둘째, 변화 감지된 정보를 사용자에게 알려주는 방식에 관한 부분으로는 현재 웹 페이지에 대해서 HTML 문서의 변화를 알려주는 서비스를 운영하는 changedetection Service[2]와 그리고 InfoMinder[3] 등이 현재 서비스를 하고 있다. 위 서비스의 경우 사용자가 직접 받아볼 전자메일과 URL를 서버에 등록 시키면 탐지 간격은 최단기간은 일 단위이며 전자메일로 전송시켜주는 방식이다. 또한 서버 사이드 방식을 취하고 있어 사용자가 늘어났을 경우 서비스 비용이 상승하게 된다.

셋째, 시맨틱 웹 관련 부분으로서"Ontology versioning and change detection on the Web" [4] 사용자들의 디자인 다른 버전들이 가진 개념들 사이의 개념적인 관계를

시각화하기 위해, 차이점들을 찾아서 RDF[5] 기반의 온톨로지로 분류하는 융통성 있는 메커니즘을 이용한다.

마지막으로 에이전트를 통하여 사용자에게 자동으로 정보를 제공하는 기술이다. 한양대학교 웹 브라우징 에이전트 IWeBA[6]는 웹 문서에서 새로 추가되는 항목이나 특정 웹 문서의 변화를 사용자에게 자동적으로 알려 주는 기능을 탑재 하고 있다. 기존의 에이전트의 경우 미묘한 변화에 대해서도 반응하는 단점이 있는 반면 문서 내용의 변화에 중점을 주기 때문에 불필요한 변화에 대해서는 반응을 보이지 않는 특징이 있다.

3. 본 시스템 소개 및 접근 방법

다음은 본 연구에서 계획하는 시맨틱 웹을 이용한 의미 중심의 웹 변경 자동 탐지를 위한 시스템 아키텍쳐를 보이고 있다.

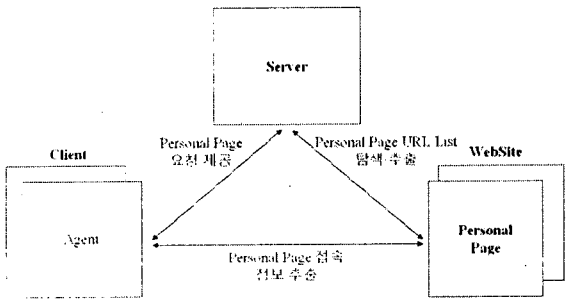

그림 1. System Architecture

Server는 컴퓨터학 분야 인물의 신상 정보 사이트의 URL를 찾아내고 이를 기반으로 Web page Storage를 구축한다. 이를 통하여 인물정보 사이트의 URL을 제공하여 Client는 보다 쉽고 정확하게 선택할 수 있다. Client 는 온톨로지 서버에 접속하여 인물에 대한 URL를 갱신하고 이를 통하여 인물정보 페이지에 접근하여 필요한 정보를 가져 오게 된다. 이 정보(HTML)를 XML로 변화시켜 의미 기반의 페이지로 변환하고, 문서간의 비교 시는 Ontology Prototype을 기반으로 하여 RDF기반에서 의미 기반 비교를 하게 된다.

본 연구는 다음의 관점에서 기존 방식과 다르다. 첫째 의미 기반(MetaData)에 의한 변화 감지한다. 기존의 방식은 HTML 문서 양식에 대한 변화 감지로 서비스를 이루고 있다. 현재 서비스를 제공하고 있는 업체들의 경우 두 버전의 HTML 문서에서 syntax를 분석하여 변화를 전자 메일로 통보한다. 본 연구는 Semantic 웹 기반으로 XML 문서 양식을 사용하여 MetaData만 분석함으로써 두 버전의 XML 문서 전체가 아닌 RDF 부분만 변화 감지 알고리즘이 동작하므로 성능을 향상 시킬 수 있다.

둘째, Server와 Client의 역할 분담을 통해 변화감지 간격을 단축시킬 수 있다. 기존의 방식은 Server-Side 에서 변화를 감지하여, Client에게 E-mail로 전송하는 방식이다. Client가 증가함에 따라 Server에서 각각의 Client별로 변화감지 알고리즘이 동작하기 때문에 Server의 역할이 증가한다. 따라서 변화를 감지하여 Client에게 전송하는 간격이 증가하게 된다. 본 연구는 기존의 Sever에서 동작한 변화감지 알고리즘을 Client의 Agent에 이식하여 변화 감지 간격을 최소화 시킬 수 있다.

4. 본 시스템 세부 구조

다음 그림 2은 컴퓨터학 분야 인물에 대한 정보 도메인을 가정한 온톨로지 prototype이다.

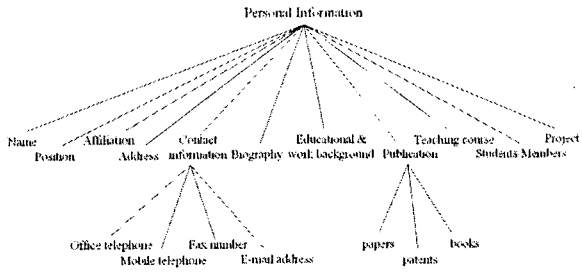

그림 2. 컴퓨터분야 인물에 대한 Prototype

이 prototype을 기반으로 한 XML Template을 두어 웹사이트로부터 받아온 정보를 XML문서로 변환 되어 저장소에 저장하게 된다. 이 때 버전 별로 구분을 두어 저장하게 되며 이 문서들은 변경 탐지를 위하여 아래 그림 3과 같은N Triple 형식의 RDF 문서로 변환되어 비교 하게 된다.

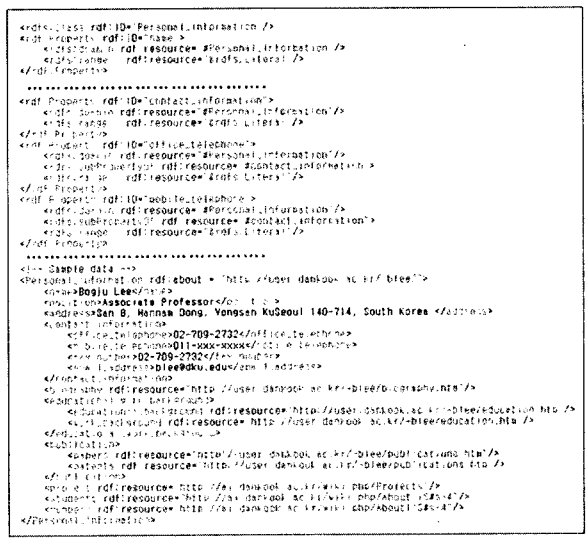

그림 3. Prototype을 기반으로 한 RDF 문서와 XML문서

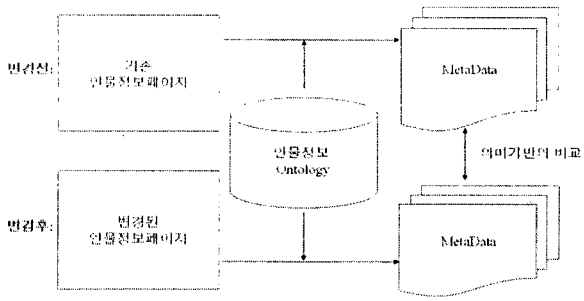

그림 4. 변화된 인물정보의 변화 탐지 구조

그림 4과 같이 두 버전간의 변화를 탐지하는 방식은 다음과 같다. HTML로 되어 있는 페이지에서 인물정보 Ontology를 기반으로 Metadata를 추출해 낸다. 이 정보는 XML형식의 RDF문서로 변환하게 되며, 기존에 저장되어 있는 XML형식의 RDF 문서와 의미기반의 변화를 비교하게 된다. 이를 통하여 기존 변화감지의 문제점인 정보와 관계없는 변화를 제거 할 수 있다.

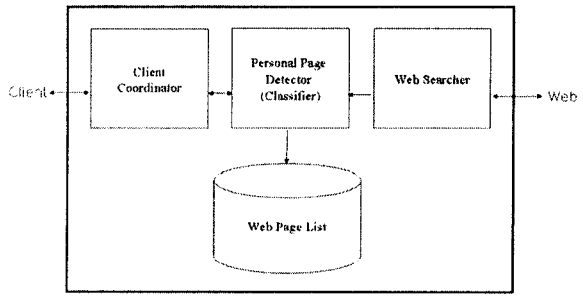

그림 5. Server Side Architecture

위의 그림과 같이 서버의 구조는 크게 3가지의 모듈로 이루어 지게 된다. Web Searcher는 Web에서 무작위로 탐색하여 페이지의 정보를 Personal Page Detector에게 제공하게 되고, 이 모듈은 Web Personal Page를 분류 및 추출하여 Storage에 저장하게 된다. Client의 요청에 의해 Client Coordinator는 Storage 에서 변화감지 대상 페이지의 URL를 제공하게 된다.

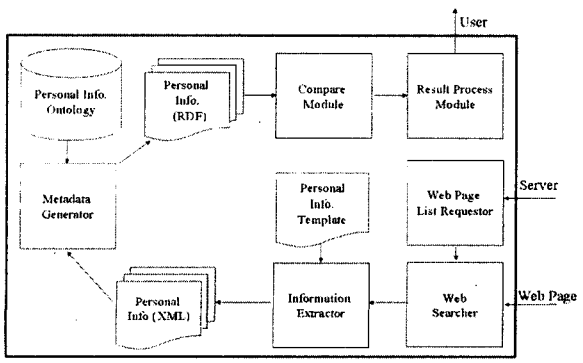

그림 6. Client Side Architecture

위 그림은 Client Side Architecture를 좀 더 자세히 보여주고 있다. 크게 5개의 프로세스로 이루어져 있으며 각각의 프로세스는 다음과 같다. Web Searcher는 Server에 접속하여 Client별로 인물정보 페이지 리스트를 가져오고 인물정보페이지 주소와 비교하여 갱신한다. 이 갱신된 주소를 이용하여 http 접속을 하고 페이지의 정보를 추출하여 Information Extractor에 제공해준다. 이 모듈은 추출한 정보를 XML Template를 참조하여 XML File를 작성하여 저장 후 MetaData Generator에게 프로세스가 이동한다. 이 모듈은 인물정보에 대한 Ontology를 기반으로 XML File 을 RDF Data로 생성을 하게 된다. RDF Data는 버전 별로 생성되어 Compare Module을 통하여 이 버전들간의 Meta Data를 뽑아 내여 비교하게 된다. Diff 알고리즘을 이용하여 찾아낸 변화를 Result Process Module을 통하여 사용자에게 알리게 된다.

5. 결론 및 향후 과제

본 논문은 시맨틱 웹이 제안하는 온톨로지를 기초로 하여 웹 페이지의 의미기반의 변화를 감지하고, 기존의 시스템에서 서버와 클라이언트의 역할을 분담 시스템을 설계하였다.

서버의 Personal Page Detector 모듈에서 무작위로 탐색 되어진 페이지의 정보를 분석하여 Machine Learning을 이용하여 인물정보 페이지를 자동으로 인식할 수 있도록 하여 보다 자동화 되고 신뢰성 있는 저장소를 구축해야 할 것이다.

XML에서 RDF로 변환의 경우 두 양식 모두 의미기반의 구조를 갖추고 있어 비교적 쉽게 변환이 가능하다. 하지만 클라이언트에서 추출해진 HTML에서 의미기반의 XML로 변환 시키는 문제는 정보 추출의 알고리즘을 도입해 적용해야 할 것이다.

6. 참고문헌

1.E. Berk "HtmlDiff: A Differencing Tool for HTML Documents", Student Project, Princeton University, http://www.htmldiff.com
2. http://www.changedetection.com/monitor.html
3. iMorph's InfoMinder, http://www.infominder.com/webminder/index.jsp
4.Michel Klein, Dieter Fensel Vrije Universiteit Amsterdam, Atanas Kiryakov, and Damyan Ognyanov OntoText Lab., Sirma AI Ltd. "Ontology versioning and change detection on the Web" http://gunther.smeal.psu.edu/klein02ontology.html
5. RDF (Resource Description Framework) http://www.w3.org/RDF/
6. 김태훈, 최중민, 한양대학교 전자계산학과, " An Intelligent Web Browsing Agent for User-oriented Internet Information Retrieval" July, 1997