20-2-R02

온톨로지 기반의 시맨틱 어노테이터 구현
(Implementation of Semantic Annotator Based on Ontology)

박 재 훈, 유 재 규, 전 양 승, 정 영 식, 한 성 국
원광대학교 컴퓨터공학과
(pjh98; jkyoo82; globaljeon; ysjeong; skhan)@wku.ac.k

요 약

이미 축적된 방대한 컨텐츠에 의미 기반 메타데이터 구축을 수작업으로 하는 것은 거의 불가능하다. 온톨로지나 시맨틱 웹 개념이 없는 일반 사용자가 자신의 컨텐츠에 효과적인 의미 정보를 부착을 하기란 매우 어렵다. 실질적인 KM, EDM, Semantic Portal, Semantic Search Engine 구현과 활성화를 위해서 필수적인 기술인 시맨틱 어노테이션은 자연언어 처리와 텍스트 마이닝 기술에 기반한다. 사람이름, 기업명, 주소 등의 개체명 인식과 사건, 원인, 결과, 상황 등의 정보 추출 기술에 기반하고 Semantic Annotation Tool은 다양한 온톨로지에 대한 적응력을 필요하며, 방대한 언어자원이 필수적이다. 이에 본 논문에서는 수많은 정보들에 태그를 붙여 컴퓨터가 의미처리를 할 수 있는 지식기반의 시맨틱 어노테이터를 구현한다.

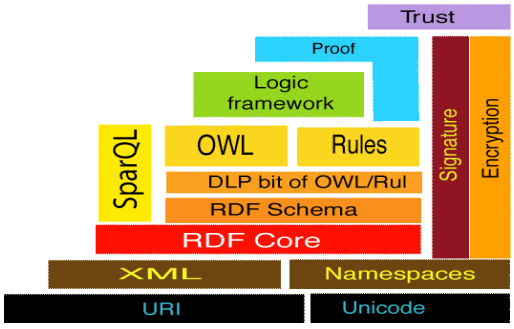
1. 서 론

1. 연구 배경

인터넷의 발전으로 기하급수적으로 늘어나고 있는 정보의 양은 사용자들에게 많은 지식과 다양한 서비스를 제공하고 있는 반면에 정보홍수라는 새로운 문제점을 초래하고 있다. 다양한 시도와 접근의 검색엔진이 개발되어 이러한 문제점을 해결하려고 시도하고 있지만 대부분의 검색엔진이 웹 문서의 내용보다는 단어나 구문 등의 키워드를 이용한 단편적인 방법으로 관련성을 검색하므로 사용자의 질의와는 관계없는 많은 문서를 결과로 가져올 수 있다. 이로 인해 사용자는 불필요한 정보를 걸러내느라 시간을 낭비하게 된다. 이런 문제점이 발생하는 가장 주된 원인은 현재의 웹이 사람을 위한 것이고 이를 위해 사람이 보고 잘 이해할 수 있도록 하기 위한 브라우저의 디스플레이 또는 레이아웃 기술에 초점을 맞추고 있다는 것이다. HTML 언어의 특징이 바로 이러한 디스플레이 용이라는 사실만 봐도 그러하다. HTML 을 이용하여 문서의 내용과 의미를 나타내는 시맨틱 정보를 표현하기가 어려우며, 따라서 사람이 아닌 프로그램 또는 소프트웨어 에이전트가 자동으로 문서로부터 의미를 추출하기가 어렵다. 시맨틱 웹은 메타데이터의 개념을 통하여 웹 문서에 시맨틱 정보를 덧붙이고 이를 이용하여 소프트웨어 에이전트가 이 의미 정보를 자동으로 추출할 수 있는 패러다임을 조성하는 것이다.

1.1.2. 문제 제기와 해결책

Tim Berners-Lee는 시맨틱 웹이 기존의 웹과 완전히 구별되는 새로운 웹의 개념이 아니라 현재 웹을 확장하여 웹에 올라오는 정보에 잘 정의된 의미를 부여하고 이를 통해 컴퓨터와 사람이 협동적으로 작업을 수행할 수 있도록 하는 패러다임이라고 그 역할을 정의하였다. 대표적인 월드와이드웹 표준화 단체인 W3C(World Wide Web Consortium)에서는 시맨틱 웹을 RDF나 다른 표준 기반으로 웹에 있는 데이터를 추상적으로 표현하는 것이라 정의하였다[1].



(그림 1) 시맨틱 웹의 계층구조

시맨틱 웹의 궁극적인 목적은 웹에 있는 정보를 컴퓨터가 좀 더 이해할 수 있도록 도와주는 표준과 기술을 개발하여 시맨틱 검색, 데이터 통합, 네비게이션, 태스크의 자동화 등을 지원하는 것이다. 시맨틱 웹을 실현하기 위한 다양한 접근방법이 제시되었다. 하지만 HTML을 기반으로 한 현재의 웹을 개선하는 기본 취지에서 보면 시맨틱 웹을 달성하기 위해 웹 프로토콜과 같은 하위 레벨의 개념을 정의하고 이 하위레벨을 이용하여 다음 레벨의 개념을 정의하는 계층구조를 설정하는 것이 일반적인 연구 방향이다.

온톨로지에 대한 정의는 여러 가지가 있지만 Gruber는 온톨로지를 “공유된 개념화(shared conceptualization)에 대한 정형화되고 명시적인 명세(formal and explicit specification)”라고 정의하였다. 온톨로지는 간단히 표현하면 단어와 관계들로 구성된 사전으로 어느 특정 도메인에 관련된 단어들을 계층적 구조로 표현하고 추가적으로 이를 확장할 수 있는 추론 규칙을 포함한다. 온톨로지의 역할 중 하나는 서로 다른 데이터베이스가 같은 개념에 대해서 서로 다른 단어나 식별자를 사용할 경우 이를 해결해주는데 있다. 온톨로지는 웹 기반의 지식처리나 응용 프로그램 사이의 지식 공유, 재사용을 가능하게 하는 아주 중요한 요소로 자리 잡고 있다[4].

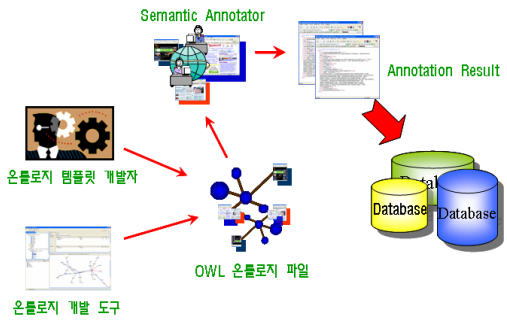
시맨틱 웹의 응용은 에이전트 기반의 웹 서비스 제공과 어노테이션 등과 같은 유용한 응용 프로그램의 개발로 요약된다. 어노테이션은 시맨틱 웹을 가장 쉽게 응용할 수 있는 매커니즘이다. 어노테이션은 이미 존재하는 웹 페이지에 대해 추가적인 설명을 덧붙여서 다시 웹에 공개하는 것으로 주로 정보 검색의 정확도를 높이는데 크게 기여할 수 있다.

2. 시맨틱 어노테이션

2.1. 어노테이션 방식

수많은 정보들에 태그를 붙여 컴퓨터가 의미처리를 할 수 있도록 지식베이스를 구축하는 것이 어노테이터의 목적이다. 이런 지식베이스는 사용자가 원하는 정확한 정보를 컴퓨터가 찾아줄 수 있다. 대표적인 어노테이션 방식은 다음과 같다.

첫째, 직접 온톨로지 문서에 어노테이션하여 저장하는 방식이다. 이 방식은 방대한 온톨로지 문서로 검색 시 불필요한 시간이 소모된다. 계속적인 파싱으로 인해 서버의 과부하를 발생하게 된다. 둘째, 다른 하나의



(그림 2) 어노테이션 시스템 구성도

XML 문서를 만들어 어노테이션 한 결과만을 저장하는 방식이다. 보다 가벼운 문서들로 인해 빠른 검색과 수정 및 관리가 용이하다. 또한 온톨로지 템플릿을 통해 많이 사용하는 주제의 틀을 만들어 배포한 뒤 사용자들에게 자신이 만들 온톨로지에 맞는 틀에 인스턴스를 추가하므로 편리하다. 셋째, 웹 페이지의 HTML 파일 자체에 어노테이션 한 결과를 입혀 쓰는 방식이다. 웹 페이지의 소스를 그대로 가져와서 따로 저장한다. HTML의 무질서한 태그, 의미를 내포하지 않은 보이기 위한 태그들로 인한 혼란을 야기한다.

2.2. 개선된 어노테이션 방식

본 논문에서 구현한 어노테이터의 어노테이션 방식은 기존의 방식 중 첫째와 셋째를 혼합한 방식이다. 어노테이션은 온톨로지 기반이고 웹 페이지의 텍스트를 블록 지정해 온톨로지에 드래그하는 방식이다. 온톨로지를 웹에서 표현하기 위한 OWL(Ontology Web Language) 파일을 파싱하여 클래스를 추출한 후 계층구조인 트리 형태로 출력한다.

온톨로지 개발도구를 이용하거나 온톨로지 템플릿 개발자는 유효한 온톨로지 파일을 제작하고 이를 시맨틱 어노테이터에서 사용한다. 시맨틱 어노테이터 프로그램에서 URL을 입력해 웹 사이트를 열고 원하는 단어나 문장을 드래그해서 온톨로지 클래스 트리에 드롭하면 자동으로 어노테이션 결과가 기록되는 방식이다. 웹 서핑을 하면서 드래그 앤 드롭으로 어노테이션 할 수 있으므로 사용자들에게는 보다 정확하고 명료한 컨텐츠를 제공할 수 있다.

3. 시맨틱 어노테이션 구현

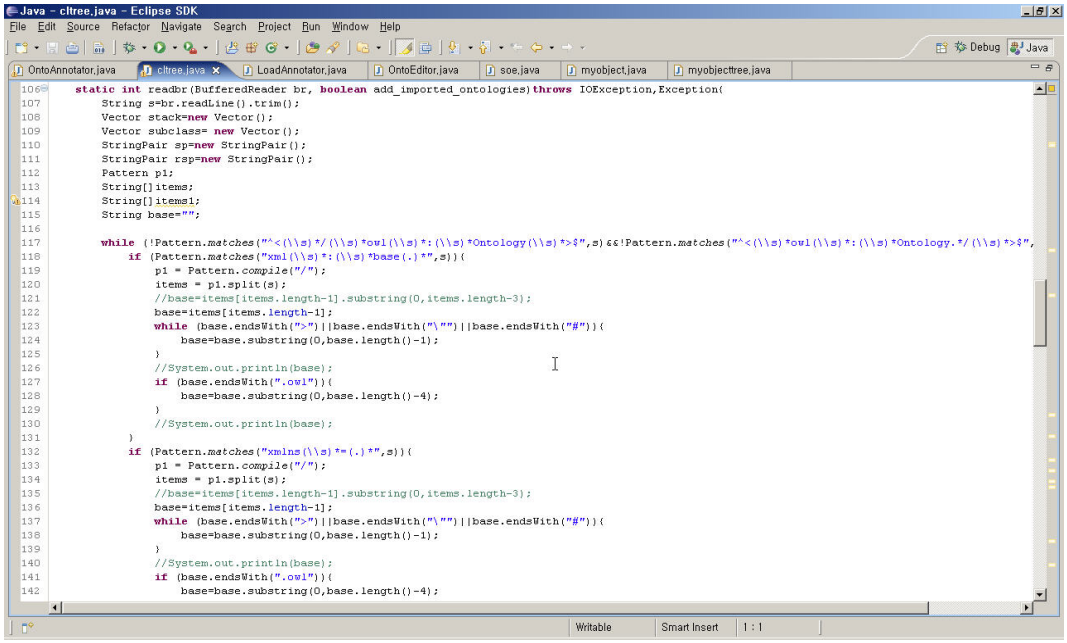
3.1. 온톨로지 파싱

온톨로지 기반의 어노테이션을 위해 온톨로지는 파싱되어 클래스를 추출하는 선행 작업이 필요하다.

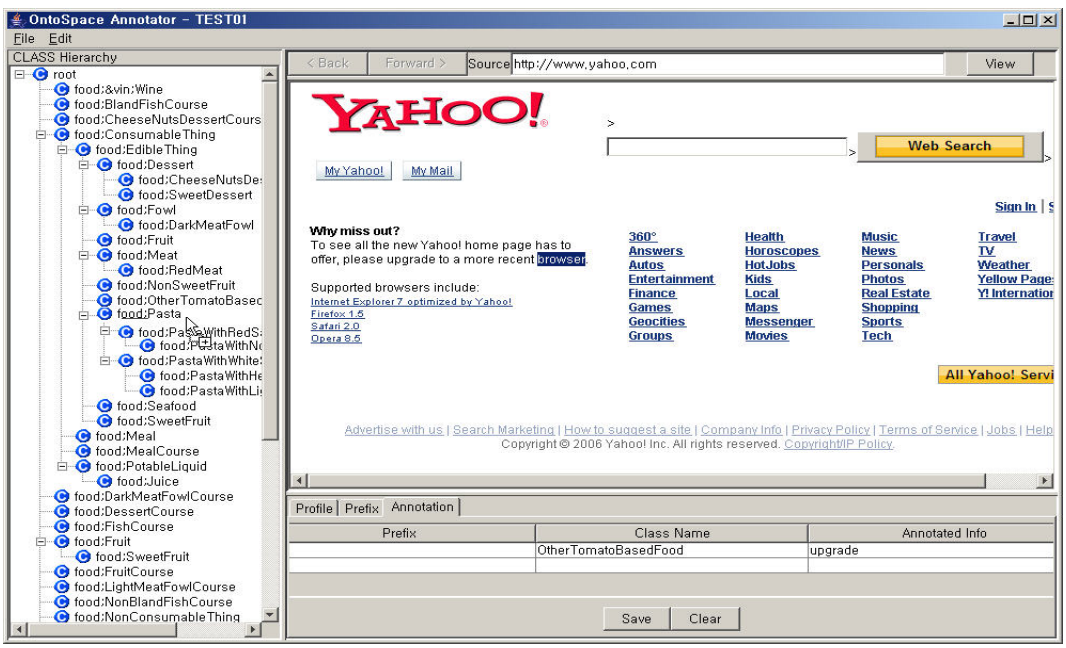
그림 3은 로컬 드라이브의 온톨로지 파일을 선택하면 온톨로지 파일 내용을 버퍼에 저장하고 읽어가며 클래스만 추출해 트리로 구성한다. 온톨로지는 계층구조로 이루어져 있기 때문에 온톨로지의 클래스 계층구조를 표현하는데 트리가 적합하기 때문이다. 온톨로지 파일을 파싱할 때 무결성을 보장하기 위해 온톨로지 파일은 유효성 체크가 되어 있는 온톨로지를 사용해야 한다. 본 논문에서는 food 온톨로지를 샘플로 사용해 클래스 트리 출력하고 이를 기반으로 어노테이션 결과를 저장한다. 어노테이션의 결과는 데이터베이스 형태로 저장된다.

그림 4는 시맨틱 어노테이션 프로그램을 이용해 드래그 앤 드롭 방식으로 어노테이션 하는 화면이다. 왼쪽의 트리는 food 온톨로지 파일을 클래스만 추출한 트리이다. 오른쪽 웹 브라우저에서 원하는 단어나 문장을 드래그한 후 왼쪽 트리에 드롭하면 온톨로지의 클래스명과 드래그 한 단어가 하단의 탭에 보여 진다. 하단 탭은 프로젝트의 정보와 어노테이션 결과를 보여준다.

어노테이션 결과는 사용자가 직접 정보를 입력하는 방식이 아닌 자동 입력이다. 온톨



(그림 3) 온톨로지 파일 파싱



(그림 4) 어노테이션 (드래그 & 드롭)

로는 의미를 가진 체계적인 구조이기 때문에 자동화 어노테이션 결과는 명확하다. 추후 검색 시스템에 활용할 경우 별도의 인덱싱 작업 필요 없이 효율적인 검색이 가능하다.

4. 결 론

본 논문에서 제시한 시맨틱 어노테이터는 온톨로지 파일을 기반으로 한다. 어노테이션을 하려면 프로그램을 시작한 후 매번 파일을 로딩하고 파싱해서 클래스 트리를 구성해야하는 시간적인 소모가 있다. 온톨로지 파일을 파싱해 클래스 트리를 구성할 때 데이터베이스에서 클래스 정보를 읽어와 트리를 구성한다면 시간을 절약하고 번거로움을 극복할 수 있다. 이를 위해 온톨로지 파일을 관계형 데이터베이스로 변환하는 연구가 필요하다. 온톨로지 파일의 문법적인 구성을 분석해 관계형 스키마를 추출하면 온톨로지 파일의 파싱을 통해 데이터베이스로 저장이 가능하다. 그리고 하나씩 어노테이션 하던 방식 외에 대량의 어노테이션 또는 자동 어노테이션 방식을 연구해 도입하면 전문지식이 없는 일반 사용자도 쉽게 어노테이션 할 수 있다. 어노테이션의 대상은 단어나 문장과 같은 텍스트 형식의 정보에만 국한되면 안 된다. 현재 웹에는 이미지 정보, 표 정보 등의 일정한 형식을 갖춘 정보들도 존재한다. 단순 텍스트 정보뿐만 아니라 이미지나 표와 같은 형식의 정보도 어노테이션 할 수 있는 연구가 필요하다. 본 연구는 아직 초기 단계이기에 많이 부족하지만 보완점을 연구하고 개선한다면 추후 다른 연구에 더 나은 결과를 이끌어 낼 수 있는 초석이 되는 연구였다.

Acknowledgement

이 논문은 2006년 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임(원광대학교, 헬스케어기술개발사업단).

참 고 문 헌

[1] Berners-Lee, T., Hendler, J. and Lassila, O., *The Semantic Web,* Scientific American, 2001.

[2] Decker, S., Melnik, S., van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., Erdmann, M. and Horrocks, I., "The Semantic Web: the roles of XML and RDF", *IEEE Internet Computing*, Vol. 4, No. 5, pp.63-73, 2000.

[3] Lassila, O., "Web metadata: a matter of semantics", *IEEE Internet Computing*, Vol. 2, No. 4, pp.30-37, 1998.

[4] Gruber, T., "A translation approach to portable ontologies", *Knowledge Acquisition*, Vol. 5, No. 2, pp.199-220, 1993.

[5] McGuiness, D., Fikes, R., Hendler, J. and Stein, L., "DAML+OIL: an ontology language for the Semantic Web", *IEEE Intelligent Systems*, Vol. 17, No. 5, pp. 72-80, 2002.

[6] McIlraith, S., Son, T. and Honglei, Z., "Semantic Web services", *IEEE Intelligent Systems*, Vol. 16, No. 2, pp.46-53, 2001.

[7] Euzenat, J., "Eight questions about Semantic Web annotations", *IEEE Intelligent Systems*, Vol. 17, No. 2, pp.55-62, 2002.

[8] Heflin, J., Hendler, J. and Luke, S., "SHOE: a knowledge representation language for Internet applications", tech. report CS-TR-4078. Dept. of Computer Science, Univ. of Maryland at College Park, 1999.

[9] Swartz, A.,"MusicBrainz: a semantic Web service", *IEEE Intelligent Systems*, Vol. 17, No. 1, pp.76-77, 2002.

박 재 훈

2004 원광대학교 컴퓨터공학과 (공학사)

2005~현재 원광대학교 컴퓨터공학과 (석사과정)

관심분야: 웹서비스, 온톨로지 공학

유 재 규

2002~현재 원광대학교 전기전자 및 정보공학부

관심분야: 시맨틱 웹서비스, 온톨로지 공학

전 양 승

2001 원광대학 컴퓨터공학과 (공학사)

2006 원광대학교 컴퓨터공학과 (석사과정)

2006~현재 원광대학교 컴퓨터공학과(박사과정)

관심분야: 시맨틱 웹서비스, 온톨로지 공학, 지능형 e-Business

정 영 식

1993 고려대학교 전산학(박사)

1993~현재 원광대학교 컴퓨터공학부 교수

1997 미시간 주립대학교 전산학과 객원교수

2004 웨인 주립대학교 컴퓨터공학과 객원교수

관심분야: 그리드컴퓨팅, LBS, 분산병렬처리

한 성 국

1979 인하대학교 전자공학과 (공학박사)

1984~현재 원광대학교 컴퓨터공학부 교수

1989 University of Pennsylvania 방문교수

2003~2004 University of Innsbruck와 DERI 연구교수

2004~현재 대한전자공학회 컴퓨터소사이어티 감사

2005~현재 한국정보과학회 호남·제주지부장

관심분야: 시맨틱 웹서비스, 온톨로지 공학, 웹서비스, 의료정보, e-Learning