

시맨틱 웹에서 온토로지를 기반한 Annotation 시스템

강상구⁰ 양재영 최중민
한양대학교 컴퓨터공학과
{sgkang⁰, jyyang, jmchoi}@cse.hanyang.ac.kr

An Ontology-based Annotation System for Semantic Web

Sangu Kang⁰ Jaeyoung Yang Joongmin Choi
Dept. of Computer Science and Engineering, Hanyang University

요 약

시맨틱 웹은 인간이 이해하는 것처럼 웹 문서의 의미를 컴퓨터가 이해할 수 있도록 하는데 있다. 이를 위해 본 논문에서는 Annotation Editor를 사용하여 논문에 대한 RDF 메타데이터의 자동 생성 방법을 제안한다. 사용자가 논문을 주석 처리할 때, 문서에 대한 특징을 추출하고 온토로지 인터페이스를 사용하여 문서를 분류한다. 구현된 시스템을 통해 사용자는 추출된 메타데이터를 메타데이터 뷰를 통해 수정하고 RDF Store로 저장할 수 있으며, 주석 뷰를 통하여 수동으로 RDF 메타데이터를 입력할 수 있다. 본 논문은 검색 엔진을 통하여 논문 검색 시 전체 내용보다 RDF 메타데이터 정보만으로 효율적인 검색을 할 수 있는 방법에 초점을 둔다.

1. 서 론

최근 들어 인터넷의 발전으로 엄청나게 늘어나고 있는 정보의 양은 사용자들에게 많은 지식과 다양한 서비스를 제공하고 있는 반면에 정보 과다(information overload)라는 새로운 문제점을 가지게 되었다. 검색엔진이 개발되어 이러한 문제점을 해결하려고 시도하고 있지만 대부분의 검색엔진이 웹 문서의 내용보다는 단어나 구문 등 단편적인 방법으로 관련성을 검사하므로 사용자가 의미적으로 원하는 문서의 검색이 어려운 실정이다.

따라서 본 연구에서는 이러한 문제점을 해결하기 위해 Annotation Editor를 사용해 웹 문서에 RDF 메타데이터(metadata)를 생성함으로써, 의미 기반으로 정보를 검색할 수 있도록 한다.

본 논문의 구성은 다음과 같다. 2장에서는 관련연구를 설명하고 3장에서는 시맨틱 웹(semantic web)에 대한 내용을 설명한다. 4장에서는 주석(annotation) 시스템의 전체적인 구성과 각 모듈에 대한 기능을 설명하고 마지막으로 5장에서는 연구 내용을 요약하고 앞으로의 향후 과제와 함께 결론을 내린다.

2. 관련연구

Annotea[1]는 공유된 주석들과 함께 W3C 협력 환경을 증진시키는 LEAD(Live Early Adoption and Demonstration) 프로젝트이다. 이 프로젝트는 두 가지 중요한 목적이 있는데, 첫 번째는 공유된 웹 문서들을 유

지하는 환경의 불을 개발하기 위해 W3C에 협력을 유지하는 것이고, 두 번째는 주석된 문서 내에서 주석의 위치를 알기 위한 XPointer, 메타데이터와 같은 주석들을 표현하기 위해 RDF를 기반한 주석 스키마(schema), 그리고 XLink와 HTTP등 존재하는 W3C 기술들을 재 사용하는 것이다. Annotea는 주석 서버와 클라이언트에서 실행하는 W3C의 Amaya editor/browser로 구성된다.

CREAM(Creating Relational, Annotation-based Metadata)[2]은 관계 메타데이터(relational metadata) 구조를 제공하는 주석 환경의 프레임워크이다. 즉 메타데이터는 클래스 인스턴스와 속성 인스턴스 그리고 관계 인스턴스로 구성된다. 이 인스턴스들은 더블린 코어(dublin core) 같은 고정된 구조 보다는 도메인 온토로지(ontology)에 기반한다. 이 프레임워크를 실행한 것이 Ont-O-Mat인데 이것은 컴포넌트 기반이며, 온토로지를 통한 주석 불어이다.

SHOE(Simple HTML Ontology Extensions)[3]는 온토로지에 기반한 지식 표현 언어이며, 의미 정보를 웹 페이지에 삽입할 수 있도록 필요한 태그들을 추가한 HTML의 superset이다. SHOE 태그들은 두개의 카테고리로 나누어지는데, 첫 번째는 온토로지 생성을 위한 태그이다. SHOE에서의 온토로지는 객체를 정의하고 그것의 의미를 파악할 수 있는 규칙의 집합으로 정의할 수 있다. 두 번째는 웹 문서에 주석을 달기 위해 사용되는 태그이다. 태그의 주된 역할은 웹 문서에 적합한 온토로지의 설정 및 데이터 선언 그리고 온토로지를 기반으로한 명제의 표현에 있다. SHOE는 Knowledge Annotator, Expose, Knowledge

Base, SHOE Search로 구성된다.

3. 시맨틱 웹

시맨틱 웹[4]은 광범위한 범위에서 기계들에 의해 의미를 쉽게 처리할 수 있는 서로 연결된 정보의 그물망이다. 즉, 인간이 이해하는 것처럼 웹에 존재하는 데이터들의 의미를 컴퓨터도 이해할 수 있다는 것이다. 이러한 웹은 WWW, URL, HTTP, 그리고 HTML의 창안자인 팀 버너스리(Tim Berners-Lee)에 의해 고안되었다. 현재의 시맨틱 웹은 독립적인 서비스들보다도 더 큰 기능성과 상호 운용성을 제공하는 정보 매개자(information broker), 검색 에이전트(search agent), 그리고 정보 필터(information filter)와 같은 지능적인 서비스들을 개발하는데 초점을 두고 있다.

3.1 메타데이터

웹은 원래 인간이 사용하기 위해 작성된 것이며, 웹에 있는 모든 것을 읽을 수 있긴 하지만 이 데이터를 기계가 이해할 수는 없다. 웹에서 모든 것을 자동화하는 것은 매우 어렵고, 그리고 웹이 포함하고 있는 정보량이 많기 때문에 수작업으로 이것을 관리하는 것은 불가능하다. 시맨틱 웹에서 제시한 해결책은 웹에 수록된 정보를 기술하기 위해 메타데이터를 사용하는 것이다. 메타데이터는 일반적으로 “데이터를 위한 데이터”라고 정의한다.

3.2 RDF(Resource Description Framework)

인터넷 특히 웹의 등장으로 엄청난 양의 자원(resource)이 생산 이용되고 있기 때문에 이러한 정보 바다에서 필요한 정보만을 선택할 수 있기 위해 자료에 대해 기술한 메타데이터의 역할은 매우 중요하다. RDF[5]는 메타데이터의 기술과 교환을 위한 구조로, 웹상의 메타데이터를 지원하는데 필요한 구조를 정의하기 위해 W3C(World Wide Web Consortium)에서 제안한 표준이다. 그리고 RDF는 인터넷 상에 존재하는 상이한 성격의 메타데이터간의 상호 운용이 가능하도록 하는데 그 목적이 있다.

3.3 RDF 구문

RDF 데이터 모델[6]은 메타데이터를 사용하고 정의하기 위한 추상적이고 개념적인 구조를 제공하고 있는데, 이를 표현하는 구문 구조는 RDF를 통해 표현된 내용들을 기계가 읽을 수 있는 형태로 변환할 수 있어야 한다. RDF에서 사용하고 있는 표현 구문은 XML이며, RDF와 XML은 상호 보완적이다.

3.4 RDF 스키마

RDF 스키마[7]는 자원들 사이의 속성 및 관계성을 표현하는 계층에 대한 정보의 집합으로, RDF 자원의 클래스에 대한 속성을 표현한다. RDF 스키마를 식별하기 위해 XML 이름 공간을 사용하며, URI를 이용하여 RDF 스키마를 인간과 기계가 동시에 처리할 수 있도록 한다. 본 논문에서는 논문에 대해 RDF 스키마를 정의해서 사용한다.

4. 시스템 구현

본 논문이 제안하는 주석 시스템 구조는 그림 1과 같다. 시스템에서 사용자가 논문을 요청하면 파일 형식이 PDF인 경우는 HTML 변환하고 전처리 과정을 거쳐 문서에서 특징(feature)을 추출한다.

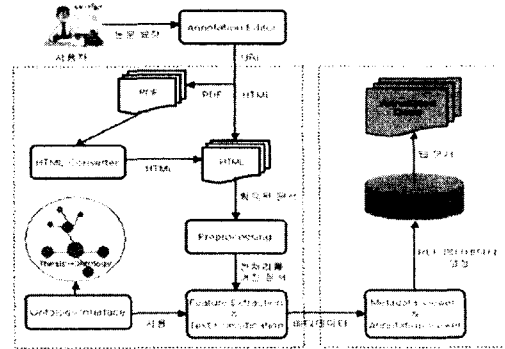


그림 1 Annotation Architecture

Annotation Editor는 그림 2와 같이 온토로지를 사용한다. 여기에서 온토로지는 “ 흥미 있는 도메인(domain)에 대한 공유된 개념화의 기술(description)”로 정의한다. 인공지능 관련 논문에 대해 XML 기반으로 온토로지를 수동 구축 하였으며, 각 노드는 개념이며 이것은 개념과 특징, 관계(relation) 그리고 제약조건(constraint)으로 이루어진다. 사용자의 요청에 의해 추출된 특징은 온토로지 인터페이스를 사용해 문서를 분류하고 생성된 메타데이터는 메타데이터 뷰를 통해 상태를 확인할 수 있다. 이렇게 생성된 메타데이터가 정확하게 추출되고 분류 되었으면 RDF Store로 저장한다. 그러면 RDF 메타데이터가 생성 된다.

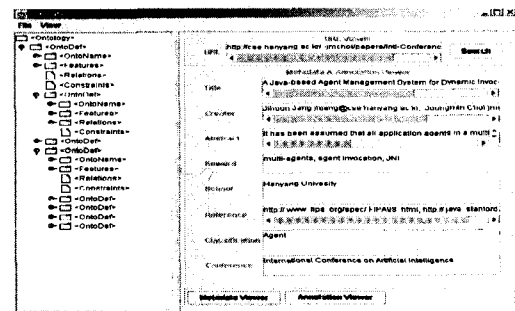


그림 2 Annotation Editor

그림 3은 논문에 대하여 Annotation Editor를 통하여 RDF 메타데이터를 생성한 인스턴스들이다. 이렇게 논문에 주석을 부여 함으로써 웹 로봇이 문서 전체 내용보다는 주석 처리된 부분만을 사용하여 더 정확한 의미를 얻을 수 있다.

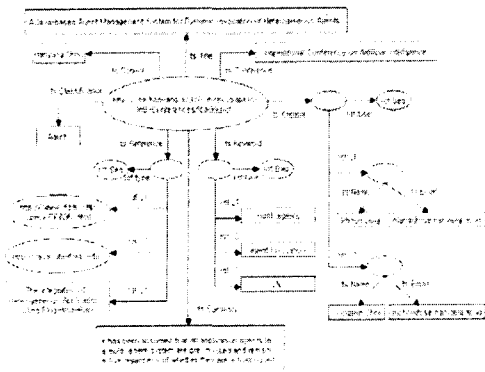


그림 3 An Instance of RDF Models

아래 그림 4는 위의 그림 3의 데이터 모델을 RDF로 표현한 것이다. 본 논문에서는 RDF 메타데이터와 논문의 URI 이렇게 두 파일을 RDF Store에 저장한다. 현재의 웹 문서는 저자(author)만이 수정할 수 있어서 다른 사용자가 문서에 대해 주석 처리하더라도 여러 사용자들이 같이 공유할 수가 없다. 이러한 문제점을 해결하고자 본 논문에서는 앞에서 온토로지에 의해 분류된 것을 기반으로 만든 웹 문서에 논문의 URI와 그 속에 RDF 메타데이터를 같이 삽입한다. 삽입된 RDF 메타데이터를 통해서 내용에 대한 의미 파악 및 카테고리 정보를 알 수 있다.

```
<?xml version="1.0"?>
<rdf RDF
  xmlns:rd="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:ts="http://13lab.hanyang.ac.kr/~spkano/thesis-rdf-schema#"
  <rdf:Description rdf:about="http://cse.hanyang.ac.kr/~imchol/papers/AMI-Confereces/ci99.pdf">
  <ts:Creator>
    <rdf:Seq>
      <rdf:li rdf:resource="#A1"/>
      <rdf:li rdf:resource="#A2"/>
    </rdf:Seq>
    </ts:Creator>
    <ts:Title>
    <ts:Title>
    </ts:Title>
    <ts:Keyword>
    <rdf:Bag>
      <rdf:li multi-agents </rdf:li>
      <rdf:li agent invocation </rdf:li>
      <rdf:li JINI </rdf:li>
    </rdf:Bag>
    </ts:Keyword>
    <ts:reference>
    </ts:reference>
    <ts:School>Hanyang University</ts:School>
    <ts:Conference> International Conference on Artificial Intelligence </ts:Conference>
    <ts:Classification>Agent</ts:Classification>
    </rdf:Description>
    <rdf:Description rdf:about="#A1">
      <ts:Name>Jihoon Jang</ts:Name>
      <ts:Email>jihoon@cse.hanyang.ac.kr</ts:Email>
    </rdf:Description>
    <rdf:Description rdf:about="#A2">
      <ts:Name>Joonamin Choi</ts:Name>
      <ts:Email>jmchoi@cse.hanyang.ac.kr</ts:Email>
    </rdf:Description>
  </rdf:Description>
</rdf:RDF>
```

그림 4 RDF Metadata

5. 결론 및 향후과제

본 논문에서는 Annotation Editor를 사용하여 논문을 요청하고, 전처리 과정을 통해 획득된 문서에서 특징을 추출한다. 그리고 온토로지 인터페이스를 사용해서 논문을 분류하며, 추출된 메타데이터는 메타데이터 뷰를 통해 수

정하고 RDF Store로 저장할 수 있도록 구현하였다. RDF 메타데이터로 주석 처리된 웹 문서는 웹 로봇이 검색 시 의미적으로 원하는 정보를 쉽게 검색할 수 있도록 성능을 높여 준다.

앞으로의 향후과제는 인공지능 관련 논문뿐만 아니라 다른 웹 문서에도 RDF 메타데이터를 자동 생성할 수 있도록 하는 것이다.

참고 문헌

[1] José Kahan, Marja-Riitta Koivunen, Eric Prud'Hommeaux, Ralph R. Swick, Annotea: An Open RDF Infrastructure for Shared Web Annotations, In *Proceedings of the WWW10 International Conference*, Hong Kong, May 2001.

[2] Siegfried Handschuh, Steffen Staab, Authoring and Annotation of Web Pages in CREAM, In *Proceedings of the 11th International World Wide Web Conference, WWW 2002*, Honolulu, Hawaii, May 7-11, 2002.

[3] Jeff Heflin, James Hendler, Semantic Interoperability on the Web, In *Proceedings of Extreme Markup Languages 2000*, pp. 111-120, 2000.

[4] Stefan Decker, Sergey Melnik, Frank Van Harmelen, Dieter Fensel, Michel Klein, Jeen Broekstra, Michael Erdmann, Ian Horrocks, The Semantic Web: The Roles of XML and RDF, *IEEE Internet Computing*, Vol. 4(5), pp. 63-74, Sept/Oct 2000.

[5] Johan Hjelm, *Creating the Semantic Web with RDF: Professional Developer's Guide*, WILEY Press 2001.

[6] Ora Lassila, Ralph R. Swick, Resource Description Framework (RDF) Model and Syntax Specification, Technical Report, W3C Recommendation. <http://www.w3.org/TR/REC-rdf-syntax>, 1999.

[7] Dan Brickley, R.V. Guha, Resource Description Framework (RDF) Schema Specification, Technical Report, W3C Candidate Recommendation. <http://www.w3.org/TR/2000/CR-rdf-schema-20000327>, 2000.