제20회 한국정보처리학회 추계학술발표대회 논문집 제10권 제2호 (2003. 11)

시맨틱 웹을 위한 온톨로지 파서의 설계

이미경, 박천수, 손주찬*
* 한국 전자 통신 연구원 인터넷 컴퓨팅 연구부
e-mail : lmk63398@etri.re.kr

A Design of Ontology Parser for Semantic Web

Mi-Kyoung Lee Shu-Cheon Park Ju-Chan Sohn*
*Dept. of Internet Computing, ETRI

요 약

시맨틱 웹은 웹 상의 정보에 의미를 부여하여 컴퓨터가 문서의 의미를 해석할 수 있도록 하기 위한 목적으로 제안된 것이다. 시맨틱 웹의 잘 정의된 의미를 다루기 위해서 RDF/RDFS, DAML+OIL, OWL 등의 웹 온톨로지 언어가 필요하다. 본 논문에서는 시맨틱 웹에서 사용되는 온톨로지 문서들을 이용하는 온톨로지 기반 지식 엔진 시스템에서 코어 엔진의 Ontology Access Layer 에 해당되는 부분으로 웹 온톨로지 문서를 읽어서 Ontology Object Model 로 생성해주는 기능을 하는 온톨로지 파서를 설계하였다. 논문에서 설계한 온톨로지 파서는 RDF, DAML+OIL, OWL 웹 온톨로지 문서들을 파싱하여 Ontology Object Model 을 생성한다. 그리고 파싱에 필요한 API 를 제공해주며 문서를 읽고 저장해준다. 온톨로지 문서들의 Triple 값을 필요로 하는 시스템을 위해서 문서들의 Triple 형태의 결과 값도 제공해준다.

1. 서론

시맨틱 웹(Semantic web)은 웹 상의 정보에 잘 정의된 의미(semantic)를 부여함으로써 사람뿐만 아니라 컴퓨터도 쉽게 문서의 의미를 해석할 수 있도록 하여 컴퓨터를 이용한 정보의 검색 및 해석, 통합 등의 업무를 자동화하기 위한 목적으로 제안되었다. 이러한 "잘 정의된 의미"를 다루고자 하는 것이 바로 시맨틱 웹 온톨로지 언어의 역할이다[1]. 시맨틱 웹의 등장은 정보를 검색할 때 더욱 정확한 결과를 가져오고, 서로 다른 이형질 소스의 정보를 통합하고 비교한다. 그리고 어떤 리소스에 대해서도 의미적이고 기술적인 정보를 연관시키며, 웹 서비스의 자동화를 위해 웹에 세부 정보를 첨가시킨다.

Tim Berners-Lee 는 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 정의하였다[2]. 그림 1 은 시맨틱 웹의 계층 구조를 나타낸다.

웹 온톨로지 언어들을 처리하는 애플리케이션을 개발하기 위해서는 온톨로지 문서의 의미를 파악하기 위한 파서가 필요하다. 현재 RDF 파서들은 몇 개의 개발된 제품들이 있으나 DAML+OIL 이나 OWL 의 파싱을 제공하는 파서로는 JENA 파서를 들 수 있다. 하지만 JENA 파서는 속도가 느리고, OWL, DAML+OIL 문서에 대한 완벽한 API 를 지원하지 못하고 있다.

본 논문에서는 시맨틱 웹을 위한 온톨로지 언어들에 대해 살펴보고, 모든 형태의 웹 온톨로지 문서(OWL, RDF/RDFS, DAML+OIL)들을 파싱할 수 있는 온톨로지 파서를 설계한다.

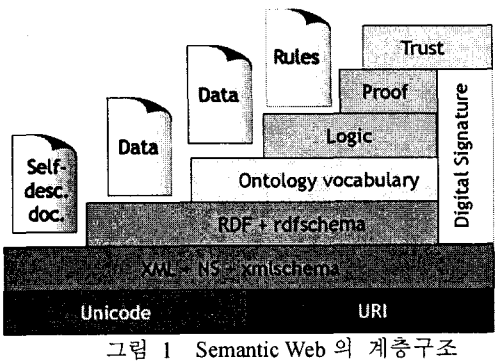

그림 1 Semantic Web 의 계층구조

2. 관련 연구

2.1. 온톨로지(Ontology)

Gruber 는 온톨로지를 "공유된 개념화(shared conceptualization)에 대한 정형화되고 명시적인 명세(formal and explicit specification)"라고 정의하였다[3]. 이 정의에서 개념화(Conceptualization)는 사람들이 사물에 대해 생각하는 바를 추상화한 모델로 특정한 분야에 국한시켜 논의된다. 명시적 명세(Explicit specification)란 개념의 타입이나 사용상의 제약 조건들이 명시적으로 정의한다는 말이다. 정형화(Formal)는 프로그램이 이해할 수 있는 온톨로지를 뜻하며, 여러 단계의 정형화가 존재할 수 있다. 공유된(shared) 개념은 어느 개인에게만 국한되는 것이 아니라 그룹 구성원이 모두 동의하는 개념을 말한다.

온톨로지는 웹 기반의 지식처리나 응용 프로그램 사이의 지식 공유, 재사용들을 가능하게 하는 아주 중요한 요소로 자리잡고 있다. 온톨로지에는 계층분류(taxonomy)와 추론규칙(inference rule)에 대한 정의가 포함된다. 계층 분류는 객체의 클래스(class)와 서브클래스(subclass), 그들간의 관계(relationship)를 정의한다. 온톨로지는 지식 공유와 재사용을 위해서 어떤 도메인에서 공통적으로 사용되는 어휘들의 집합을 개념적으로 표현하는 방법이다.

2.2. 웹 온톨로지 언어(Web Ontology Languages)

온톨로지를 표현하기 위해 스키마와 구문 구조 등을 정의한 언어가 온톨로지 언어(ontology language)이며, 현재 RDF/RDFS, DAML+OIL, OWL 등의 웹 온톨로지 언어가 정의되었다.

(1) RDF 와 RDFS

RDF(Resource Description Framework)는 W3C 의 가장 기본적 시맨틱 웹 언어로서 웹에 있는 자원에 관한 메타 데이터를 표현하기 위한 언어이다[3]. 메타 데이터는 데이터에 대한 데이터, 즉 어떤 객체나 리소스에 대한 서술적인 정보를 말한다. RDF 모델은 기본적으로 리소스(Resource), 특성(Property), 서술문(Statement)의 개념으로 구성된다. 서술문은 일반 문자의 주어(subject), 동사(predicate), 목적어(object)에 해당되는 것으로서 사람이나 웹 문서 등 특정 대상(object)이 특정 속성(attribute)에 대하여 특정 값(value)를 가지고 있는 상태를 표현하며 이것이 RDF 문의 기본 단위가 된다.

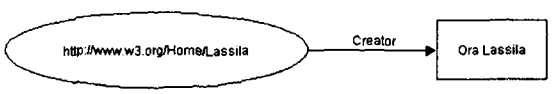

그림 2 RDF Statement

RDF 의 새로운 용어를 정의하기 위해서 RDF Schema 를 사용한다[5]. RDF 스키마는 특성에 대한 정의나 사용상의 제약 사항을 기술한 것으로 RDF 문을 구성하는 단어(term)를 정의하고 단어들의 세부적인 의미를 기술하고 있다.

(2) DAML+OIL

DAML+OIL[6]은 DAML(DARPA Agent Markup Language)과 OIL(Ontology Inference Layer)의 결합을 통하여 만들어졌다. DAML 은 시맨틱 언어로써, 서로 관련 있는 웹 페이지라도 서로 다른 의미를 사용하기 때문에 발생하는 의미론적 장벽을 해결하기 위해 만들어진 언어이다. DAML 이 나오기 전의 OIL 은 온톨로지를 정의하기 위해 개발된 언어이며, 이 언어만으로는 온톨로지를 표현하기에 부족하여 이를 보완하기 위해 두 언어가 결합되었다. DAML+OIL 웹 온톨로지 언어는 현재 W3C 에 의해서 OWL 웹 온톨로지 언어로 계승 발전되고 있다. 이 언어는 객체지향적인 구조를 가지며 클래스(class)와 속성(property)을 써서 표현된다. 클래스와 속성의 성격을 서술한 공리(axiom)의 집합으로 구성된다.

(3) OWL

W3C 의 Semantic Web activity 인 OWL 은 DAML+OIL 를 기반으로 발전된 형태로 시맨틱 웹을 위한 웹 온톨로지 표준 언어이다. OWL 은 DAML+OIL 과 유사한 형식을 가지며, DAML+OIL 의 네임 스페이스와 속성 클래스 이름 등을 변경하고 RDF/RDFS 의 변화를 수용하였다 [7]. OWL 은 DL(Description Logic)을 기반으로 만들어진 RDF 확장 언어이며, OWL 의 axioms 은 DAML+OIL 의 axioms 보다 더욱 풍부한 표현력을 가지고 있으며, 클래스나 속성 간의 subsumption 이나 equivalence 등의 다양한 성격을 선언하는데 사용된다. OWL 은 XML Schema 의 모든 데이터형을 지원하여 문자열, 십진수, 실수, 정수 범위 등을 사용할 수 있으며 RDFS 와 밀접하게 연관되어 있다. OWL 은 기본적으로 Description Logic 의 추론 능력과 표현력을 가지고 있기 때문에 DL 의 장점을 가지고 있으며, 현재 OWL 의 종류로는 OWL Lite, OWL DL, OWL Full 로 나눌 수 있다.

2.3. JENA 온톨로지 파서

JENA 파서는 RDF 문서를 읽어서 RDF 모델과 구조를 구현할 수 있는 파서로 잘 알려져 있다. JENA 파서는 ARP(Another RDF Parser)를 이용하여 RDF 그래프 형식으로 데이터를 읽어들이고 JENA 데이터 모델로 변환하여 처리된다[8]. JENA 는 기존의 RDF 파싱 뿐만 아니라 DAML+OIL, OWL, N3, DB 등의 데이터를 파싱해주고 RDF/XML, RDF/XML abbreviation, N-Triple 형태의 출력형태를 지원해준다. 그림 3 과 같이 온톨로지 문서의 파싱은 RDF 파서를 확장한 구조로 되어 있으며, RDF 파서(ARP)를 기반으로 하여 웹 온톨로지 문서를 파싱한 후, 그 결과 값을 다시 OWL, DAML+OIL 의 문법에 맞게 다시 파싱하는 구조를 가

진다.

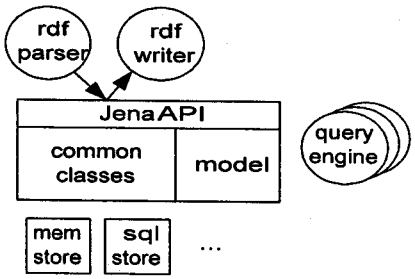

그림 3 Jena API 의 구조

현재 JENA2 는 시맨틱웹을 위한 자바 프레임웍을 지원하는 시스템으로 RDF/XML 파서인 ARP, 리즈닝 서브시스템, 온톨로지 서브시스템(OWL, DAML+OIL, RDFS), RDQL 쿼리 언어를 제공한다[9].

3. 온톨로지 파서의 설계

온톨로지기반 지식 엔진의 코어에 해당되는 부분은 Storage (DB, KB, web Ontologies 등)와 온톨로지 객체 모델과 연동하는 Ontology Access Layer, 그리고 온톨로지 객체 모델을 이용하는 Ontology Evaluation, Ontology Inference, Visualization Components 들로 나눌 수 있다. 본 논문에서 설계하는 온톨로지 파서는 Ontology Access Layer 에 해당되는 부분으로 Web Ontologies 나 DB, Repository 등에서 읽어온 데이터 값을 OOM (Ontology Object Model)형태로 메모리에 올려주는 역할을 한다. Ontology Evaluation, Ontology inference, Visualization components 들은 OOM 형태로 메모리에서 온톨로지 데이터들을 이용하게 된다. 온톨로지 파서는 온톨로지 저장소에 접근하여 데이터를 읽고, OOM 을 핸들링하고 온톨로지를 Import, Export 하는 기능을 하게 된다.

온톨로지 파서는 웹 온톨로지 문서인 RDF/RDFS, OWL, DAML+OIL 문서들을 읽어들여서 온톨로지 객체 모델로 변환한다. 기존에 나와있는 파서 중에서 온톨로지 문서의 파싱을 지원해주는 툴로는 HP 에서 개발한 JENA2 파서가 있다. 그러나 JENA 파서는 OWL, DAML+OIL 문서의 완벽한 핸들링을 지원해주지 못하고 처리 속도가 매우 느리다.

파서의 구성을 간단하게 살펴보면 그림 4 와 같다. 온톨로지 문서를 읽어들인 후, 온톨로지 언어별 문법에 의해 온톨로지 문서를 파싱하여 OOM 을 생성하고 파일에 읽고 쓰는 역할을 한다.

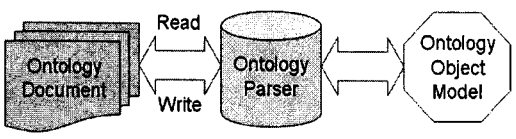

그림.4 온톨로지 파서의 구성

3.1. 시스템의 구조

온톨로지 파서의 구체적인 구조를 살펴보도록 하겠다. 파서는 기본적으로 어휘 분석 기능(Lexer), 구문 분석 기능(Parser), 코드 생성 기능(Code generator)을 가지게 된다. 본 시스템에서는 토큰 생성을 쉽게 하기 위하여 Lexer 모듈에서는 XML Parser 인 Xerces 를 사용하였고 파서를 생성하기 위해서 ANTLR(Another Tool for Language Recognition)[10]를 이용하였다. RDF 기반의 웹 온톨로지 문서(RDF, OWL, DAML+OIL)들의 어휘를 분석하기 위해서 XML 파서(Xerces)를 이용하여 SAX 이벤트를 발생시킨다. Ontology Lexer 는 어휘 분석의 기능을 하는 모듈로써, XML 파서에서 넘겨받은 SAX event 들을 이용하여 어휘 별로 토큰을 생성하고, 생성된 토큰을 Ontology Parser 모듈로 넘겨준다. Ontology Parser 모듈에서는 토큰들을 파싱 규칙에 의해서 구문을 분석하고 Triple 값을 생성한다. 파싱 규칙은 RDF 문법을 이용하여 EBNF 형식으로 구현하였다. 파싱 모듈에 의해 처리된 토큰은 semantic parsing 을 통해 각종 데이터를 저장한다. 저장된 데이터는 API 함수를 통하여 응용 프로그램들에 제공되고, Subject, Predicate, Object 의 Triple 값을 생성한다. Triple 값은 OOM 을 생성하거나 다른 시스템에서 이용되며, 우리가 개발한 OOM 은 RDF/RDFS, DAML+OIL, OWL 등의 웹 온톨로지 문서들을 모두 표현할 수 있는 통합된 형태의 온톨로지 객체 모델이다.

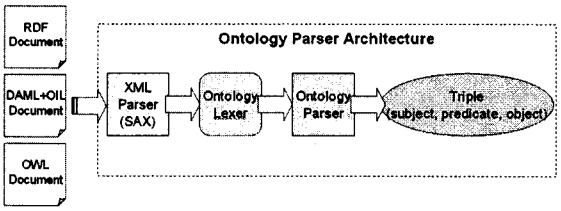

그림 5 온톨로지 파서의 구조

3.2. 시스템 모듈 구조

온톨로지 파서는 각 모듈별 특징에 따라 아래와 같이 크게 4 가지 모듈로 나눌 수 있으며, 각 모듈별 기능을 간단히 살펴보겠다.

(1) XML parsing 모듈

이 모듈은 웹 온톨로지 문서를 파싱하기 위한 전처리기의 개념으로 XML 파서를 이용하여 웹 온톨로지 문서를 파싱하는 역할을 한다. 파싱작업을 위해서는 아파치 그룹에서 개발한 Xerces 파서를 사용한다. 이 모듈에서는 이벤트 기반의 SAX 파싱 기법을 이용하며 startElement, endElement, Characters 의 이벤트가 발생하면 해당 문서에서 Qnames, Attribute Names, Attribute Values 를 Ontology Lexer 모듈로 넘겨준다.

(2) Ontology Lexer 모듈

온톨로지 Lexer 모듈은 어휘분석 모듈이다. SAX 이벤트로 넘겨진 데이터를 이용하여 어휘를 분석한 후 적합한 토큰을 형성하여 온톨로지 파싱 모듈로 넘겨주는 역할을 한다. 이 모듈에서는 온톨로지 언어의 어휘에 따라 구분되는 토큰의 종류를 미리 선언해둔다.

(3) Ontology parsing 모듈

이 모듈에서는 온톨로지 어휘 분석 모듈로부터 토큰을 하나씩 넘겨받아서 RDF/RDFS, OWL, DAML+OIL의 문법에 따라 미리 만들어 놓은 파싱 규칙을 이용하여 스택 연산을 수행하며, 파싱 테이블에서 검출된 룰에 의해 semantic parsing 모듈을 호출한다. semantic parsing 모듈은 현재의 정보를 이용하여 파싱 정보를 재구성하거나 API 를 위한 자료를 형성하고 예외처리 모듈을 이용하여 오류 메시지를 발생하는 역할을 한다.

(4) Triple 생성모듈

Triple 생성 모듈에서는 온톨로지 언어들을 Subject, Predicate, Object 의 Triple 값을 생성하는 역할을 한다. Triple 값은 OOM 과 Inference engine 에서 사용되며, Statement 에서 Subject, Predicate, Object 의 값을 생성할 수 있다.

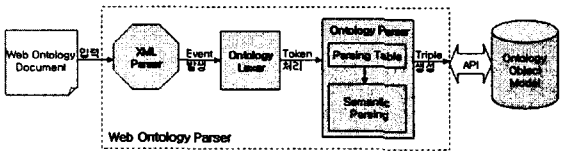

그림 6. 온톨로지 파서 모듈

4. 결론

지금의 웹보다 지능적인 시맨틱 웹을 만들기 위해서는 온톨로지가 필요하고, 온톨로지를 구축하기 위해서 웹 온톨로지 언어가 사용되고 있다. 온톨로지 언어로 구축된 온톨로지를 위한 프레임 웍을 위해서 본 논문에서는 웹 온톨로지 문서들을 파싱하는 온톨로지 파서를 설계하였다. 기존의 JENA 파서는 ARP 파서를 기반으로 이를 확장하여 온톨로지 언어별 파싱 기법을 구현한 반면, 본 논문에서 설계한 온톨로지 파서는 웹 온톨로지 문서가 온톨로지 구축 언어에 상관없이 OWL, RDF/RDFS, DAML+OIL 의 모든 온톨로지 언어의 파싱을 지원하고 공통의 온톨로지 객체 모델을 생성해준다. 우리가 개발한 온톨로지 객체 모델은 웹 온톨로지 문서들의 공통된 모델을 생성해주기 때문에 확장성이 뛰어나다.

온톨로지 파서는 웹 온톨로지 문서를 읽어들여서 어휘를 분석한 후, 웹 온톨로지 언어들의 문법을 이용해 만들어진 파싱 테이블을 거쳐서, 구문을 분석하고 그 결과로 Triple 형태의 Statement 를 생성해내며 온톨로지 객체 모델을 생성하기 위한 데이터 값을 넘겨주고 온톨로지를 위한 API 도 제공해준다.

참고문헌

[1] Asuncion Gomez-Perez and Oscar Corcho, "Ontology Language for the Semantic Web," IEEE Intelligent Systems, vol.17, no.1, January/February, 2002, pp.54-60

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

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

[4] Resource Description Framework(RDF), http://www.w3.org/RDF/

[5] RDF Vocabulary Description Language 1.0:RDF Schema, W3C Working Draft 23 January, 2003. http://www.w3.org/TR/rdf-schema/

[6] Debora L. McGuinness, Richard Fikes, James Hendler and Lynn Andrea Stein, "DAML+OIL:An Ontology Language for the Semantic Web," http://www.daml.org/2000/10/daml-ont.html

[7] Mike Dean et al.(Eds), "OWL Web Ontology Language Reference," W3C Candidate Recommendation 18 August, 2003.

[8] Brian McBride, "Jena:Implementing the RDF Model and Syntax Specification," Semantic Web Workshop, WWW2001, 21 December, 2001

[9] Jena2 – A Semantic Web Framework, http://www. hpl.hp.com/semweb/jena2.htm

[10] ANTLR(Another Tool for Language Recognition) Homepage, http://www.antlr.org