웹 문서를 자동으로 RDF로 변환하는 방법

구자훈 · 김우생
광운대학교 컴퓨터소프트웨어학과
E-mail : kjh880719@naver.com, kwsrain@kw.ac.kr

Method to transform Web Document into RDF

Jahoon Koo · Woosaeng Kim
Department of Computer Science and Engineering, Kwangwoon University

요 약

차세대 웹은 시멘틱 웹(Semantic Web)을 기반으로 발전할 것으로 예상된다. 시멘틱 웹을 구축하기 위해서는 특정 도메인의 정보를 온톨로지로 구축하는 작업이 필요하다. 그러나 이러한 작업은 많은 노력과 비용이 소요되는 전문가에 의한 수작업에 의존해야 하기 때문에 자동화된 방법으로 시멘틱 웹을 구축하는 방법이 요구된다. 이를 위해 본 논문에서는 웹의 정보를 자동으로 RDF 문서로 변환하는 방법을 제시한다. 웹상의 HTML 문서를 파싱하여 subject, predicate, object의 트리플을 추출한 후, Jena 자바 라이브러리를 통하여 RDF 문서로 변환 시키는 과정을 제시한다. 이러한 과정을 통해 자동화된 시멘틱 웹 구축의 가능성을 보인다.

1. 서 론

시맨틱 웹은 현재의 인터넷과 같은 분산 환경에서 리소스(웹 문서, 각종 파일, 서비스 등)에 대한 정보와 자원 사이의 관계-의미 정보를 기계(컴퓨터)가 처리할 수 있는 온톨로지 형태로 표현하고, 이를 자동화된 기계(컴퓨터)가 처리하도록 하는 프레임워크이자 기술이다. 시멘틱 웹은 RDF, RDF Schema,, OWL 등의 계층 구조를 이룬다[1]. 시맨틱 웹의 기반이 되는 RDF는 웹에 있는 임의의 자원을 subject, predicate, object의 트리플로 표현한다. 수작업으로 시멘틱 웹을 구축하는 것은 많은 비용과 노력이 필요하기 때문에, 본 논문에서는 기존 웹의 문서 정보를 자동으로 RDF로 변환하는 방법을 제시한다.

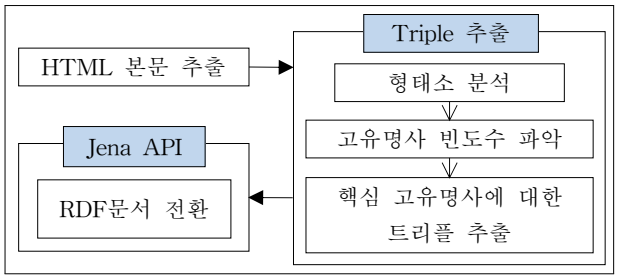
본 논문은 다음과 같이 구성된다. 2장에서 Jena 라이브러리를 살펴본다. 3장에서는 HTML문서가 RDF로 변환되는 과정을 설계하고, 4장에서는 이에 대한 구현 방법을 의사코드와 자바 소스코드의 일부를 통해 설명한다. 마지막으로 5장에서는 종합적인 결론과 향후과제를 제시한다.

2. 관련 연구

Jena는 HP 시맨틱 웹 연구소의 Brian McBride에 의해 개발된 시맨틱 웹 프레임워크이다. RDF, RDFS 및 OWL을 위한 프로그래밍 환경과 기본적인 RDF 파서를 제공하며 내부적으로 규칙(rule)기반의 추론엔진을 포함하고 있다[2]. Jena의 프레임워크는 온톨로지 생성, 편집, 저장 및 추론 등을 위한 다양한 API를 제공한다. 예를 들어, RDF API는 RDF의 Resource, Property, Literal 그리고 그들을 연결하는 Statement를 만드는 것을 가능하게 한다. 이렇게 구성된 Statement들의 집합을 RDF Model이라 하고, 이 RDF Model을 통하여 RDF가 생성된다.

3. 설 계

웹 상의 HTML 문서로 부터 RDF 생성의 전반적인 진행 과정은 그림 1과 같다. HTML 문서에서 본문이 되는 문장들을 추출한 후, 추출된 문장에 대한 형태소 분석으로 핵심이 되는 고유명사와 그와 관련된 트리플을 추출한 후, Jena를 사용해 RDF를 생성한다.



(그림 1) 자동 RDF 생성 과정

특정 도메인의 HTML 문서에서 추출된 문장들에 사용된 어휘를 분석하기위해, 'penn treebank'를 활용한 영문 형태소 분석기인 'Stanford POS Tagger' 자바 라이브러리를 사용하여 고유명사, 동사, 형용사 등을 찾아낸다[3,4]. 이를 위해 표 1에 의한 고유명사를 찾아내고 그 빈도수를 카운트한다. 빈도수가 가장 높은 고유명사를 핵심 고유명사라 칭하고, 이 핵심 고유명사가 포함된 문장을 추출한다. 핵심 고유명사를 Subject로 또 다른 고유명사는 Object로 선정하고, 이 둘과 연관된 동사를 Predicate로 하는 트리플을 구성한다. 마지막으로 Jena 자바 라이브러리를 사용해서 추출된 트리플들로 RDF를 생성한다.

(표 1) The Penn Treebank Tag Set

태그	품사 (Noun)	태그	품사 (Verb)
NN	Singular	VB	Base form
NNS	Plural	VBD	past tense
NP	singular Proper	VBG	present participle
NPS	plural Proper	VBN	past participle

4. 구 현

본 논문에서는 온라인 백과사전인 Wikipedia에서 "Korea"의 HTML 문서를 대상으로 자동 RDF 변환을 실험하였다. 해당 도메인을 가져와 본문을 구성하는 [p] tag의 문장을 추출한다. 그리고 이 문장에 사용된 어휘들에 대해 영문 형태소 분석기를 이용하여 각각의 품사를 추출한다. 표 2는 추출된 문장에 대한 영문 형태소 분석 결과를 보여준다.

<표 2> 형태소 분석기에 의한 처리 결과

Sentence 1 : Korea/NNP is/VBZ bordered/VBN by/IN China/NNP to/TO the/DT northwest/RB and/CC Russia/NNP to/TO the/DT northeast/NN

다음에 'NNP'해당하는 고유명사를 Hashtable을 사용하여 빈도수를 측정한다. 이를 통해 얻어진 데이터 값을 내림차순으로 정렬하고, 첫 번째 키 값을 핵심 고유 명사로 선정한다. 핵심 고유명사가 포함된 문장에 한하여, 핵심 고유명사는 subject에 할당되고, 핵심 고유명사 이후에 나오는 동사와 또 다른 고유명사에 대한 순서를 파악한 후에 동사를 predicate로, 고유명사를 Object로 구성해 트리플을 표 3의 방법으로 형성한다.

<표 3> 핵심 고유명사에 관한 트리플 구성

```
while sentence is not empty:
  if sentence contains Core_noun:
  // 문장 내 핵심 고유명사 유무 판단
    while split sentence with tags:
    // 문장을 태그 단위로 분할하고 순차적으로 접근
      if tag contains Core_noun:
      // 태그가 핵심 고유명사이면 subject에 할당
          subject <- remove tag noun
          next to tag
      else if tag contains 'VB' and subject isn't empty:
      // subject에 비어있지 않고 동사이면 predicate에 할당
        predicate <- remove tag proper noun
        next to tag
      else if tag contains 'NNP' and predicate is not empty:
      // predicate에 동사가 저장되어 있을 경우 Object에 할당
        object <- remove tag proper noun
        next to tag
      next to tag
    End while
  else next to sentence:
End while
```

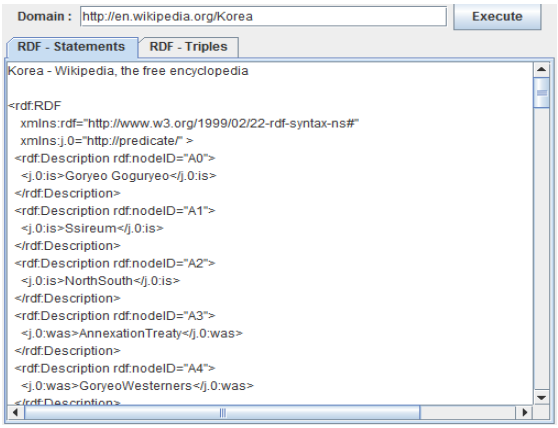
얻어진 트리플 집합으로 Jena 라이브러리를 사용하여 RDF를 생성한다. 표 4는 RDF를 생성하는 일부 코드이다. 먼저 빈 RDF Model을 생성한다. Resource와 Property는 URI 정보를 담고 있어야 하기 때문에, subject와 predicate에 각각 "http://subject/"와 "http://predicate"를 덧붙여 임의의 URI를 만들어 Resource와 Property에 할당하고, Literal에는 Object를 할당하는데, 만일 동일한 Property가 다수 존재 한다면 하나의 Property에 여러 개의 Object가 연결 될 수 있도록 해당 Property에 Object만 추가하여, Property의 중복을 줄인다. 이로써 트리플에 대응하는 RDF 문서가 생성된다.

<표 4> 추출된 트리플을 이용한 RDF 생성

```
Model model = ModelFactory.createDefaultModel();
while (statement) {
  Resource s = model.createResource(subject);
  Property p = model.createProperty(predicate);
  RDFNode o = model.createLiteral(object);

  if(s.hasProperty)
    s.addProperty(p, model.creteResource()
                          .addProperty(p. o));
  else s.addProperty(p, o);
}
```

그림 2는 앞서 선정한 도메인에 대하여 생성된 RDF의 일부 결과이다.



(그림 2) XML 형식의 RDF로 표현된 결과

5. 결론 및 향후과제

본 논문에서는 Wikipedia의 도메인을 이용하여 고유명사를 추출하고, 그 관계를 트리플로 구성하여 RDF를 생성하였다. 이번 연구를 통해 HTML 문서를 RDF로의 자동 전환함으로써 시맨틱 웹을 구성하기 위한 기반을 마련하였으나, 구성된 RDF들을 통합하는 연구가 필요하다. 또한 문서에서 핵심어와 이를 통한 중요한 트리플을 찾아내는 다양한 방법의 연구가 진행되어야 할 것이다.

참 고 문 헌

[1] http://en.wikipedia.org/wiki/Semantic_Web
[2] http://jena.apache.org
[3] http://nlp.stanford.edu/software/tagger.shtml
[4] 이효갑, 김관구, "특정 도메인 문서 내 관계 트리플 추출", 한국정보처리학회 춘계학술발표대회, 2010.04