OWL/RDF(S) 도메인 온톨로지 검색 시스템 설계

황명권° 공현장 정관호 김판구
조선대학교 전자계산학과
{hmk2958°, kisofire, khjung, pkkim}@chosun.ac.kr

The Design of Retrieval System about OWL/RDF(S) Ontology

Myunggwon Hwang° Hyunjang Kong Kwanho Jung Pankoo Kim
Dept of Computer Science, Chosun University

요 약

본 논문은 웹에 산재되어 있는 OWL/RDF(S) 도메인 온톨로지에 대한 검색 시스템을 설계하여, 온톨로지의 재사용성을 극대화 하는데 그 목적이 있다. 컴퓨터와 인간이 정보를 공유하고, 의미적인 상호작용을 위한 시맨틱 웹에 대한 연구가 활발이 진행되고 있다. 시맨틱 웹을 실현하기 위해 개념들의 정의와 개념들간의 관계를 형성하는 온톨로지의 구축이 필수요소가 됨에 따라 온톨로지를 구축하기 위해 OWL, RDF(S) 그리고 DAML+OIL 등의 많은 온톨로지 언어가 개발되었고, 이들 언어를 기반으로 하는 Protege, OILEd와 KAON 등의 사용자들에게 온톨로지 구축의 편리성을 제공하는 온톨로지 구축 도구들도 50여가지 이상 개발되었다. 이러한 이유로 많은 온톨로지들이 개발되고 있다. 그렇지만 온톨로지의 가장 큰 특징은 동일 도메인의 온톨로지의 재사용인데, 산재되어 있는 온톨로지들을 검색하기 어렵고, 이들을 한데 모아놓은 저장소 또한 갖추어지지 않아 동일한 도메인 온톨로지가 존재할지라도 새롭게 온톨로지를 구축해야한다. 이에 본 논문에서는 웹상에 존재하는 온톨로지들의 검색을 용이하게 하여 지식 정보의 재사용을 최대화하기 위하여 본 연구를 진행하고 시스템을 설계하였다.

1. 서 론

1990년 10월 팀 버너스 리에 의해 웹이 개막된 이후 웹은 질과 양적으로 눈부신 성장을 거쳐 왔다. 1997년 W3C(World Wide Web Consortium)에서 최초로 RDF(Resource Description Framework) 모델과 구문에 관한 기술문서가 공표된 이후 컴퓨터와 사람이 지식을 공유함으로써 공동으로 업무를 처리할 수 있는 시맨틱 웹의 연구가 본격적으로 진행 되었다.[3] 시맨틱 웹 분야에서 핵심이 되는 것은 모든 사물의 개념정의인데, 이 부분을 온톨로지가 담당하고 있다. 이러한 온톨로지를 구축하기 위해 OWL(Web Ontology Language), RDF(S), DAML+OIL 등의 많은 온톨로지 언어가 개발되었다.[1] 그리고 이 언어들을 기반으로 Protege, OILEd, KAON등 사용자 중심의 인터페이스를 제공하는 온톨로지 구축도구들도 현재 50여 가지가 개발되었다.[4] 이처럼 시맨틱 웹의 핵심 요소인 온톨로지는 의료정보, 전자상거래, 인터넷 비즈니스 그리고 지식관리와 정보검색 분야에 활용되고, 앞으로 활용분야는 더욱 넓어질 것으로 예상되고 있다. 이와 같이 넓은 활용분야를 갖는 온톨로지는 여러 도메인에 대하여 온톨로지 연구자들에 의해 현재까지 많이 구축이 되었고, 많은 온톨로지 구축 도구들이 개발됨으로써 온톨로지 구축이 더욱 용이해졌으며, 앞으로도 도메인에 대한 온톨로지들이 지속적으로 개발될 것이다.

온톨로지의 특징중의 하나는 지식 정보의 재사용이라 할 수 있다.[2] [5]에서 설명하는 온톨로지 구축 과정에서도 같은 도메인에 대한 기존 온톨로지가 존재하는지의 여부가 온톨로지 구축의 큰 부분을 차지하고 있다. 하지만 온톨로지들이 여러 곳에 산재해 있고, 이를 검색하기 위한 시스템 연구가 많이 부족한 실정이어서 지식 정보의 재사용성이 현저하게 낮다.

이에 본 연구에서는 2004년 2월 W3C에서 국제 표준으로 제정하여 온톨로지를 구축하는데 가장 많이 사용하는 온톨로지 언어인 OWL/RDF(S)를 이용하여 구축된 도메인 온톨로지를 검색하기 위한 시스템을 설계하였다. 사용자는 본 시스템을 이용하여 검색하고자 하는 도메인을 정확하게 검색하여 온톨로지의 재사용성을 최대화할 수 있을 것으로 기대된다.

본 논문의 2장은 본 시스템의 설계에 필요한 관련연구에 대해 서술하고, 3장에서 본 논문의 핵심인 도메인 온톨로지 검색 시스템의 주요 모듈들을 상세하게 기술한 후, 4장에서 결론과 향후 연구 방향을 제시한다.

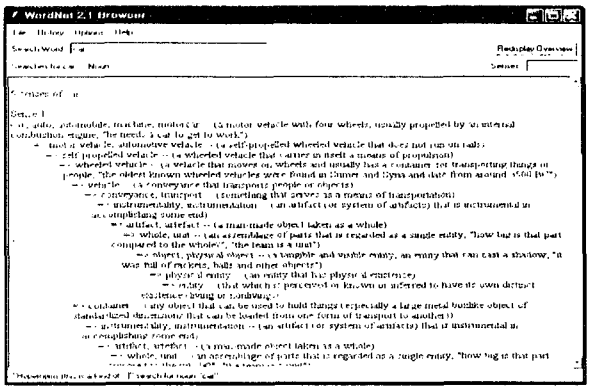
2. 관련 연구

기 구축된 온톨로지의 도메인 파악과 온톨로지의 완전성을 분석하기 위해서는 개념들의 매칭기준이 될만한 표본이 필요하고, 각 개념들을 비교하기 위해서는 개념들 사이의 유사도 측정이 필수이다. 본 시스템에서는 분석의 기준으로 워드넷을 사용하였고, 유사도를 측정하기 위해 Jaccard 유사도 측정 방법을 적용하였다.

2.1 워드넷(WordNet)

워드넷은 범용의 대형 온톨로지로서 미국의 프린스턴 대학(Princeton University)에서 개발되었고, 42,000개

이상의 어휘에 대해 정의하고 있다. 특히, 각 어휘에 대한 유의어, 반의어, 상/하위어에 대해 상세히 기술되어 있다.[7] 또한 자바 워드넷 라이브러리(Java Wordnet Library)가 제공되어 이를 응용한 많은 연구가 진행되고 있다. [그림 1]은 워드넷 실행 화면이다.



[그림 1] 워드넷 실행화면

2.2 Jaccard 유사도

개념들 사이의 유사도를 측정하기 위해서 많은 유사도 측정 수식이 정의되었지만, 본 시스템은 유사도 측정을 위해 Jaccard 수식을 이용하였다. 본 수식은 개념들의 유사도 측정이 용이하고, 가장 보편적으로 사용되고 있다.

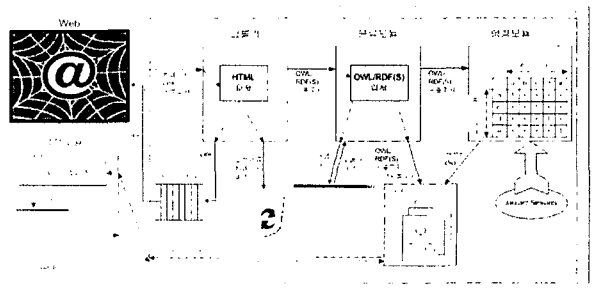
$$Jaccard - sim(c_1, c_2) = \frac{P(c_1 \cap c_2)}{P(c_1 \cup c_2)}$$

[수식 1] Jaccard 유사도 측정 수식

Jaccard 수식은 최소 0과 최대 1 사이의 값을 갖으며, 0은 두 개념이 서로 전혀 연관이 없음을 의미하고, 1은 두 개념이 서로 동의어임을 나타낸다. [6]

3. 시스템 구조

본 시스템은 웹에 있는 문서를 가져오는 크롤러(Crawler), 이들을 도메인별로 분류하는 분류모듈(Classifying Module), 분류된 문서들에 대한 순위를 부여하는 랭킹모듈(Ranking Module) 그리고 사용자에게 편리한 인터페이스를 제공하여 원하는 도메인 온톨로지를 검색할 수 있는 검색모듈(Retrieval Module)로 구성 되어있다.



[그림 2] 전체 시스템 구조

[그림 2]는 본 시스템의 전체 구조를 나타낸 것이다. 본 시스템의 전체 구조는 일반 웹 페이지 검색 시스템과 흡사하다. 하지만 온톨로지의 내부에 정의된 개념들을 분석하고, 도메인 내에 정의된 모든 개념들을 이용하여 랭킹정보를 갖기 때문에 각 모듈들의 기능들은 기존의 검색 시스템과는 확연히 구분된다.

3.1 OWL/RDF(S) 온톨로지 크롤러(OWL/RDF(S) Ontology Crawler)

크롤러는 웹에 있는 모든 OWL/RDF(S) 온톨로지를 저장소(Repository)로 가져오는 핵심적인 역할을 한다. OWL/RDF(S) 온톨로지를 가져오기 위해서는 먼저 HTML 문서를 분석한다. 그래서 크롤러 내부에는 HTML 파서가 있고, HTML 문서내의 링크 중에서 웹 페이지로 파악된 문서는 큐(Queue)에 저장하고, 분석을 완료한 웹 페이지나 온톨로지와 상관없는 문서는 폐기한다. 그리고 링크된 문서의 헤더분석을 통해 OWL/RDF(S) 온톨로지로 파악된 문서들은 도메인 분류모듈로 보내진다.

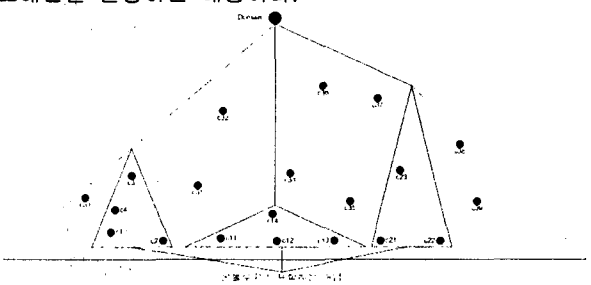
3.2 도메인 분류모듈(Domain Classifying Module)

분류모듈은 크롤러를 통해 웹에서 가져온 온톨로지를 분석하여 해당 도메인을 파악한다. 온톨로지가 정의하고 있는 개념들을 분석하기 위해 OWL/RDF(S) 파서가 있고, 파악된 개념들을 워드넷의 개념들과 매칭을 시킨다. 온톨로지의 도메인 개념을 파악하는 수식은 다음 [수식 2]와 같다.

$$DomainConcept = \max_{c \in S(c_1, c_2)} [-logP(c)]$$

[수식 2] 도메인 개념 정의 수식

위의 수식은 Resnik 방식으로, 온톨로지의 도메인 개념을 모두 포함하는 워드넷의 최소상위개념을 찾기 위한 수식이다. [그림 3]은 위 수식을 이용하여 온톨로지의 도메인을 결정하는 내용이다.



[그림 3] 도메인 개념 결정

[그림 3]과 같이 워드넷 내의 개념 중에서 온톨로지에 정의된 모든 개념들을 포함하는 최소의 상위개념을 [수식 2]를 적용하여 추출함으로써 도메인을 결정한다. 이런 방식으로 도메인이 파악되면 온톨로지 저장소로 온톨로지가 전달이 되고, 각 온톨로지들에 대한 색인(Index)을 위해 색인 온톨로지를 새롭게 구축한다. 색인 온톨로지는 워드넷과 개념 및 계층 구조가 같고, 속성은

"hasURI"와 "hasConsistency"를 갖고 있으며, 본 모듈에서 분석된 도메인 온톨로지는 "hasURI"의 값인 온톨로지의 위치와 함께 색인 온톨로지 개념의 인스턴스로 저장된다. 그런 다음 랭킹모듈로 분류된 도메인 온톨로지가 전달된다.

3.3 랭킹모듈(Ranking Modeule: 동일 도메인에 대한 우선순위 결정)

특정 도메인으로 분류된 온톨로지라 할지라도 그 내용의 완전성은 차이가 있다. 동일 도메인에 두개 이상의 온톨로지가 정의되어 있을 때, 검색 시에 순위를 부여함으로써 더 향상된 정보제공을 할 수 있다. 본 시스템에서는 워드넷 내에 일치하는 도메인의 개념들을 기준으로 도메인 온톨로지가 개념들을 얼마나 체계적으로 잘 정의하고 있는지 Jaccard 유사도 측정 수식을 사용하여 유사도를 측정하였다. [표 1]은 도메인이 'automobile'인 온톨로지를 워드넷의 'car'도메인과 매칭한 결과이다.

[표 1] 워드넷과 매칭

	car	sedan	bus	wagon	taxi	...
automobile	1	0.20	0.20	0.20	0	
sedan	0.20	1	0	0	0	
bus	0.20	0	1	0	0	
cab	0.20	0	0	0	1	

[표 1]의 결과를 이용하여 일치하는 개념들만을 이용하여 다시 Jaccard 유사도 수식을 통해 수치를 계산한다. 그리고 그 값을 색인 온톨로지의 각 개념에 대한 속성인 "hasConsistency"의 값으로 입력한다. [표 1]의 결과에 대하여 Jaccard 유사도 측정 수식을 이용하면, 워드넷의 'car'도메인에는 41개의 개념이 포함되어 있고, 'automobile' 온톨로지와 4개가 일치하므로 hasConsistency의 값은 약 0.100이 저장된다.

3.4 검색모듈(Retrieval Module)

검색모듈은 앞에서 설명한 과정을 통해 도메인으로 분류된 온톨로지들을 검색하는 부분이다. 사용자가 원하는 도메인 개념을 입력받아 워드넷과 매칭을 시킨다. 워드넷에는 모든 개념들의 Synset_ID를 갖는데, Synset_ID가 동일하다는 것은 동의어를 의미한다. 사용자가 입력한 개념에 해당하는 Synset_ID를 통해 동의어들을 파악하고, 이들 중 대표개념을 이용하여 색인 온톨로지를 검색한다. 색인 온톨로지는 워드넷의 대표개념들을 이용하여 상/하위 계층구조를 갖고 있고, 각 도메인 온톨로지의 주소는 도메인 개념의 인스턴스로 생성되어 있다. 생성된 인스턴스들은 자신의 주소와 완전성의 정도를 수치로 포함하고 있는데, 완전성을 이용하여 검색 결과인 인스턴스 이름과 주소를 우선순위에 따라 보여준다.

위의 전체적인 시스템의 과정을 살펴보면 [표 2]와 같다.

[표 2] 전체 시스템 과정

크롤러	1. HTML 파서에서 웹 페이지 분석 2. 링크된 주소를 큐에 저장 또는 폐기 3. OWL/RDF(S) 온톨로지를 분류모듈로 전송
분류 모듈	4. OWL/RDF(S) 파서에서 온톨로지 분석하여 워드넷과 개념 매칭 5. Resnik 방식을 이용하여 모든 개념을 포함하는 최소상위개념을 도메인으로 결정 6. 저장소로 온톨로지 저장, 색인 온톨로지 구축 7. 도메인 온톨로지를 랭킹모듈로 전송
랭킹 모듈	8. Jaccard 유사도 측정수식을 이용하여 온톨로지의 완전성 평가
검색 모듈	1. 입력된 도메인을 워드넷을 통해 대표개념 검색 2. 대표개념을 색인 온톨로지와 매칭한 결과들을 우선순위대로 보여줌

4. 결론 및 향후 연구 방향

본 논문은 온톨로지 내에 정의된 지식 정보들의 재사용을 최대화하기 위해 OWL/RDF(S)로 구축된 도메인 온톨로지의 전체 검색 시스템을 설계하였다. 시스템의 구성은 웹 문서를 가져오는 크롤러, 온톨로지의 도메인을 파악하는 분류모듈, 도메인으로 분류된 온톨로지의 완전성 정도를 분석하는 랭킹모듈 그리고 사용자가 검색을 하고 결과를 볼 수 있는 검색모듈로 구성이 되어 있다. 크롤러를 제외한 각 모듈에서 워드넷을 이용하여 개념들을 분석한다. 하지만 온톨로지는 개념들 뿐만 아니라 개념들 사이의 관계까지 고려해야 온톨로지의 완전함 정도를 정확하게 파악할 수 있다. 이는 향후에 연구해야할 과제이다.

참고 문헌

[1] 이재호, "시맨틱 웹의 온톨로지 언어", 정보과학회지, Vol.21, No.03, Pages: 18-27, 2003. 03

[2] Chintan Patel, Kaustubh Supekar, Yugyung Lee, E.K. Park, "OntoKhoj: a semantic web portal for ontology searching, ranking and classification", Proceedings of the 5th ACM international workshop on Web information and data management, Pages: 58-61, 2003

[3] http://www.w3.org

[4] Michael Denny, "Ontology editor survey results", http://xml.com/2002/11/06/Ontology_Editor_Survey.html, 2002. 06.

[5] Natalya Fridman Noy, Deborah L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology". Stanford Knowledge Systems Laboratory Technical Report KSL-01-05. 2001. 03.

[6] Hyunjang Kong, M.G. Hwang, P.K. Kim, "A New Methodology for Merging the Heterogeneous Domain Ontologies based on the WordNet", International Conference on Next Generation Web Services Practices, 2005. 08.

[7] http://wordnet.princeton.edu/