다중 온톨로지를 활용한 워드넷 확장

박경국*, 김진환*, 구태완**, 김병관*, 정연진*, 이광모*
*한림대학교 컴퓨터공학과
**대한상공회의소 강원인력개발원 정보기술과
e-mail : pkkstory@hallym.ac.kr

Wordnet Extension Using Multiple Ontology

Kyung-Kook Park*, Jin-Whan Kim*, Tae-Wan Gu**, Byung-Kwan Kim*,
Yeon-Jin Jung*, Kwang-Mo Lee*
*Dept. of Computer Engineering, Hallym University
**Dept. of , Information Technology, Gangwon Human Resources Development Institute

요 약

웹에 대한 사용자의 다양한 요구와 더불어 웹 서비스에 관한 연구가 활발히 진행되고 있다. 그 중 사용자가 원하는 정보를 정확하게 제공하기 위한 의미기반의 검색방법이 중요한 이슈로 등장하였다. 사용자 질의에 대한 의미 분석 및 검색과 온톨로지 구축의 기반으로는 어휘망이 사용된다. 그러나 어휘망은 작성 시기와 그 내용이 일반적인 내용으로 구성되어 전문적인 의미 검색으로 사용하기에는 부족함이 있다. 따라서 본 논문에서는 다중 온톨로지를 이용한 어휘망 확장을 제안한다.

1. 서론

정보검색은 급변하는 사회의 영향으로 수 많은 분야가 생성되고 각 분야별로 생성되는 다수의 정보를 빠르고 정확하게 제공받기 위해 필요하다. 그러나 다양한 정보의 내용이 방대해질수록 불필요한 정보들이 생겨나기 시작했으며 따라서 정확한 정보검색 결과 확보를 위하여 의미기반의 검색방법이 제안되었다[1].

이러한 의미기반의 검색방법은 사용자 질의를 분석하고 검색하는 것인데 이를 위하여 어휘망 또는 온톨로지를 활용하는 방법이 제안되었다. 그러나 온톨로지를 활용하여 모든 지식 영역을 구축하기에는 그 비용이 매우 높고 작성 시기와 그 내용이 일반적인 구성이라 사용에 제한이 있다[2]. 따라서 온톨로지의 재사용 및 확장, 그리고 제작에 관한 연구가 다양하게 진행되고 있으며[3], 온톨로지 구성과 의미 분석 및 검색의 기반으로는 어휘망이 사용된다. 대표적 어휘망으로는 워드넷이 있다. 워드넷은 단어와 단어 간의 관계를 나타내는 사전이나 보편적인 어휘들로 구성되어 있어 전문적인 용어는 드물게 나타난다. 따라서 워드넷을 보강 및 확장하기 위한 연구 또한 활발히 진행 중이다.

본 논문은 다중 온톨로지를 활용하여 워드넷을 확장 확장을 제안한 것이다. 사용되는 각각의 온톨로지에는 이미 사용자 질의 분석을 위한 의미가 잘 정의되어 있다. 따라서 어휘망 확장은 단어의 단일 개념이 아닌 의미 정보가 추가된 확장이 가능하게 된다.

2. 관련연구

2.1 온톨로지 언어

온톨로지 정보들은 컴퓨터가 이해할 수 있는 형태로 표현되는데 이러한 표현을 위해 사용되는 것이 온톨로지 언어다. 대표적인 온톨로지 언어에는 RDF (Resource Description Framework), RDFS(RDF Schema), DAML(DARPA Agent Markup Language), OIL (Ontology Inference Layer), DAML+OIL 및 OWL(Web Ontology Language)이 있다.

RDF 는 주어(Subject), 서술어(Property), 목적어(Object)로 온톨로지를 표현하며[4], RDFS 는 RDF 의 타입 및 관계를 나타내는 언어다[5]. DAML 은 미국 DARPA 에서 개발된 언어로 객체 및 객체들 간의 관계 묘사를 통해 온톨로지를 표현하고 웹사이트 들 간에 보다 높은 차원의 상호 운용성을 구축하도록 설계되었다[6].

OWL 은 RDF 와 DAML+OIL 에서 확장된 언어로서 보다 풍부한 어휘를 가지고 있어, 자원들의 개념 및 관계를 다양하고 세밀하게 기술하는 것이 가능하며 추론을 지원한다. 또한 W3C 가 OWL 을 표준 온톨로지 언어로 지정하여 계속해서 온톨로지 데이터들은 OWL 로 기술될 것이다[7, 8].

OWL 은 어휘를 구성하는 용어의 의미와 용어들간의 관계를 명시적으로 표현한 언어로써 도메인에 따른 다양한 형태의 요구사항을 수용하기 위해 3 종류의 SubLanguage 로 구성되는데 OWL Lite, OWL DL, OWL Full 로 분류된다. OWL Lite 는 기본적인 기능만을 제공하고, OWL DL 은 제약 사항에서만 사용할 수 있는 제한을 가지며, OWL Full 은 자유로운 표현력을 가진

다. OWL 은 서로 다른 개발자 및 사용자를 대상으로 하기 때문에 어떤 OWL 하위 언어가 주어진 요구사항에 최적인지 결정해야 한다

2.2 워드넷과 한국어 어휘망

워드넷(WordNet)은 1985 년 Princeton 대학 인지과학연구실의 G. Miller, Fellbaum 등 심리학자, 언어학자, 전산학자 등을 중심으로 구축해온 단어 간의 관계를 표현하는 형태로 어휘 의미가 계층적 구조로 이루어진 어휘망이다[2]. 일반적인 사전처럼 가나다순으로 제작된 것과는 달리 워드넷은 개념을 바탕으로 네트워크를 구축한 사전으로 개념 행렬을 기초로 정의 되었다. 워드넷의 관계는 동의관계, 반의관계, 상의관계, 하의관계, 분의관계, 양식관계, 함의관계로 이루어졌다.

영어 워드넷을 바탕으로 각국의 언어로 확장하는 연구가 진행되어 국내에서는 부산대학교 한국어 정보처리연구실에서 한국어 어휘망(KorLex)을 구축하는 연구를 진행하고 있다. 연구가 진행 중인 한국어 어휘망은 첫 번째로 영어 워드넷의 명사를 영-한 번역하여 구축한 다음 두 번째로 문제점을 분석하고 유형화하였다. 세 번째로는 구축된 영-한 번역 어휘망을 정제하여 한국어 어휘망을 구축하였다. 이렇게 구축된 어휘망은 정보검색, 자동번역, 문장분석 등과 온톨로지를 구축하기 위한 기반으로 사용될 수 있다[3].

본 논문은 부산대학교 한국어 정보 처리연구실에서 영-한 번역 어휘망을 정제하여 만든 신셋(Synset)을 사용한다. 이 신셋은 상하위관계를 갖는 워드넷의 명사를 번역하고 수 정한 데이터이다. 신셋의 구성은 신셋번호, 영어, 상위 번호, 하위번호, 번역어로 되었다. 예를 들어, <표 1>를 보면 상하위관계는 객체지향 프로그래밍언어를 기준으로 상위에는 '프로그래밍언어' 하위에는 '자바'로 구성된 것이다.

<표 1> 신셋의 상하위관계

상위용어	기준용어	하위용어
프로그래밍언어	객체지향프로그램	자바

신셋은 검색 키워드에 대한 상하위관련 용어들을 찾는 기준이 된다. 위드넷을 이용한 어휘망 구축은 워드넷의 명사를 한국어로 번역하고 상하위관계를 재정비한 것이다. 따라서 기존의 어휘들만 있으므로 확장을 위한 방법이 있어야 한다. 본 논문에서 제안하는 어휘망 확장은 온톨로지의 의미 정보를 이용하여 신셋에 새로운 용어들을 추가시켜 워드넷을 확장하였다. 확장의 범주는 컴퓨터 용어로 한정하여 추가하며 이용되는 온톨로지는 현재 웹 서비스 중인 컴퓨터 용어 사전의 구성을 기반으로 재구성한 온톨로지를 이용한다.

3. 다중 온톨로지 설계

온톨로지는 철학적 용어로 쓰였으나 전산학으로 되면서 시맨틱 웹을 표현하는 중요한 언어로 변화되었다. Thomas R. Gruber 는 온톨로지를 "An ontology is an explicit specification of a shared conceptualization (온톨로지는 개념화에 대한 명시적 명세이다)" 라고 정의하여 기술적 관점에서 정의했다[9]. 온톨로지가 명시적(explicit)이라는 것은 사용되는 개념의 유형과 개념의 사용에 대한 제한 조건(constraints)이 명시적으로 정의된다는 것이다. 개념화(conceptualization)는 온톨로지가 세상에 존재하는 현상의 추상적인 모델을 다룬다는 것을 의미한다. 다중 온톨로지는 동일지식에 의하여 구축된 온톨로지를 다른 온톨로지와 지식 공유를 하는 것이다. 이것은 데이터의 공유와 재사용을 위한 것이다. 본 논문은 용어사전 온톨로지를 활용하여 다중 온톨로지를 구성한다.

용어사전은 주요 검색 사이트에서 제공하고 있어서 사용자는 원하는 용어를 찾아 볼 수가 있다. 하지만 그 용어와 관련된 용어를 다시 찾아야 하는 불편함과 각 사이트 별로 용어 구성은 제 각각이고 다양해서 불편함을 준다. 그래서 용어사전을 온톨로지 언어로 변환하고 의미가 비슷한 용어도 함께 보여주면 지능화된 용어사전이 된다. 아직 표준화된 용어사전 OWL 은 없기 때문에 용어 검색을 위하여 용어사전 OWL 을 구성해야 한다.

분야별 용어는 방대하기 때문에 모든 용어를 온톨로지화 하여 쓰기에는 어려운 점이 있다[10]. 그래서 컴퓨터 용어로 데이터를 한정시키고 구성하여 결과를 보여줄 수 있도록 용어사전 온톨로지를 만들어야 한다. 본 논문에서는 네이버 용어사전과 팀즈 코리아를 기본 데이터로 하여 온톨로지를 구성한다. 기본 데이터는 검색 사이트의 용어와 내용을 가지고 있기 때문에 이를 토대로 용어사전 온톨로지를 설계하는 것이다.

3.1 용어사전 구성

각 사이트 별로 용어사전 구성에 대해서 명세화해야 하는데 본 논문은 DTD 로 정의를 한다. DTD(Document Type Definition)는 XML 문서의 형태를 일관된 구조로 정의하는 문서이다[11]. 용어 사이트 페이지는 주제별로 분류되어 있는데 용어사전 온톨로지를 만들기 전에 미리 골격을 만드는 작업이다. 각 사이트의 온톨로지는 구성이 달라지므로 다중 온톨로지에서 검색을 할 수 있는 시험 배경이 된다. 웹 온톨로지 언어인 OWL 은 기본적으로 XML 문법을 따르므로 DTD 로 명세화 후에 온톨로지를 구성해야 한다. <그림 1>는 팀즈 사이트의 용어사전 구성 형식을 DTD 로 정의한 결과다.

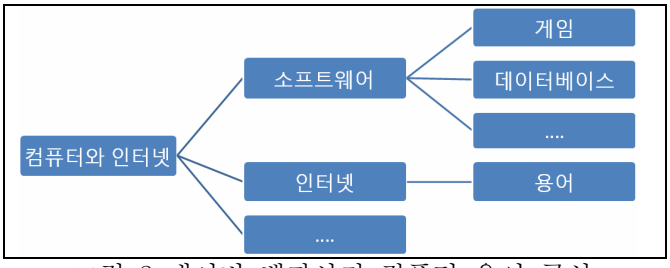
```
<!ELEMENT TermsDictionaryCategory
(Software*, PC*, Internet*, Networking*, ComputerBasic*,
Communication*, Hardware*)>
<!ELEMENT Software
```

```
(Application*, Database*, OS*, Programing*)>
<!ELEMENT PersnalComputer
(PC*, MutimediaGraphic*, Device* )>
<!ELEMENT Internet(InternetTech*,InternetSevice*)>
<!ELEMENT Networking
(NetworkHardware*, NetworkSoftware*, Security*)>
<!ELEMENT ComputerBasic(ComputerGenaral*, Standard*)>
<!ELEMENT Communication
(Circuit*, Wireless*, HighNetwork*, Transmission*,
NetworkConnecting*)>
<!ELEMENT Hardware
(HardwareGenaral* ,Electronic *, Microprocessor*)>
```

<그림 1> 컴퓨터 용어사전 텀즈 DTD

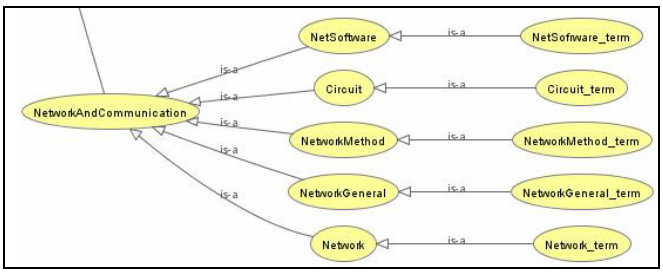
3.2 용어사전 온톨로지 및 클래스

명세화된 DTD 를 기준으로 용어사전 온톨로지를 생성시킨다. 검색 사이트의 사전 구성은 클래스와 용어, 속성 관계를 OWL 로 만든다. 예를 들어, 네이버 백과사전 컴퓨터 용어 구성은 <그림 2>과 같이 구조화된 디렉토리를 갖고 있다. 이 구조에 컴퓨터 용어를 넣으면 컴퓨터 용어사전 온톨로지가 된다. 각 용어 사이트마다 구성이 상이하므로 온톨로지는 별개로 구성된다.



<그림 2>네이버 백과사전 컴퓨터 용어 구성

온톨로지 저작 도구인 Protégé-2000[12]을 사용하여 네이버 백과사전 컴퓨터 용어와 텀즈 코리아 용어의 구성을 각각 컴퓨터 용어사전 OWL 로 두 개의 파일을 만들고 문법 검증을 거쳐 유효한 데이터인지 확인을 한다. <그림 3>는 네이버 백과사전 컴퓨터 용어를 OWL 로 변환시킨 관계도 이다. 각 클래스들은 속성값으로 클래스 간의 관계와 어느 분류에 속하는지 알 수가 있며 하단클래스는 인스턴스를 가진다.



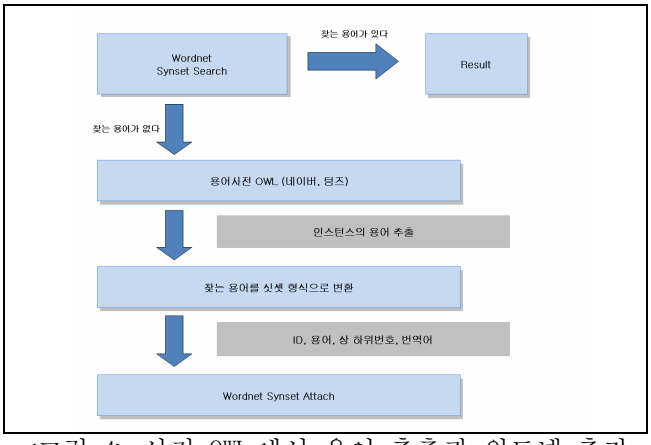
<그림 3> 컴퓨터 용어사전 OWL 관계도 일부

인스턴스는 용어, 요약, 본문으로 구성되며 실질적인 컴퓨터 용어가 들어간다. 인스턴스의 용어는 신셋과 비교대상이 되며 신셋에 찾는 용어가 없으면 네이버 컴퓨터 용어 OWL 의 인스턴스 용어를 찾아 추가하여 어휘망을 확장한다.

4. 어휘망 확장

신셋을 활용한 검색은 찾는 용어 간의 상하위 관계를 의미적으로 알 수가 있다. <표 1>에서 '객체지향프로그램'을 찾으면 상위용어는 '프로그래밍언어', 하위용어로 '자바'를 나타내고 있다. 이것을 컴퓨터 용어사전 OWL 에서 데이터를 가져와서 추출하면 된다. 용어사전 OWL 은 Jena2 와 SPARQL 를 이용하여 검색 및 추출을 한다. Jena2 는 온톨로지 관리 시스템으로 RDF, OWL 모델을 지원하는 개발 툴이고[13], SPARQL 은 W3C 가 제안하는 RDF 모델에서 데이터를 검색하는 표준 질의이다[14]. 신셋과 용어 사전 OWL 에서 나온 검색 결과를 비교하여 신셋에 필요한 용어를 추가시켜 어휘망 확장을 구현할 수 있다. <그림 4>은 신셋에 없는 용어를 추가시키는 것이다.

본 논문에서는 각 용어사전 OWL 의 구성이 다르므로 같은 용어가 두 개의 사전에서 나오면 우선순위를 네이버 용어사전 OWL 에 기준을 둔다. 왜냐하면 네이버 사전은 업데이트가 주기적으로 되는 반면 텀즈 사전은 불규칙적이므로 안정적인 사전을 우선으로 하는 것이다.

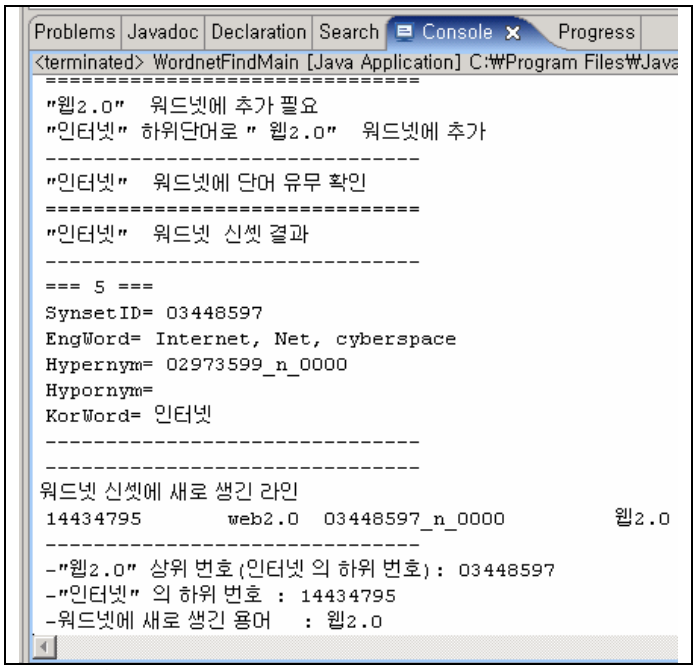


<그림 4> 사전 OWL 에서 용어 추출과 워드넷 추가

4.1 온톨로지를 활용한 어휘망 확장 구현

어휘망 확장 방법은 우선 찾는 용어가 신셋에 있는지 확인해야 한다. 만약 없으면 네이버, 텀즈 용어사전 OWL 에서 용어를 찾아 신셋에 추가시킨다. 추가 시 신셋 형식이 신셋번호, 영어, 상위번호, 하위번호, 번역어로 구성되어 있으므로 새로운 용어에 대하여 신셋 형식을 갖추어야 한다. 우선 신셋번호가 만들어져야 하는데 현재 신셋번호는 워드넷의 상하관계를 의미간의 거리가 측정되어 생성되어 있지만 본 논문에서는 컴퓨터 용어만을 선정하여 확장시키는 것이므로 단어간의 거리 측정보다는 상하위관계를 중요시한다. 새로 추가될 신셋번호는 마지막 신셋번호에서 10 씩 카운트 하여

일정한 간격을 두었다. 영어와 번역어 위치에는 검색 용어의 영어와 한글을 넣고 상위번호는 사전의 디렉토리 구조에서 용어를 포함하는 클래스의 ID 를 따른다. 만약 클래스의 ID 가 신셋에 있으면 번호를 추출하여 상위번호로 적용시키면 된다. 하지만 각 사전 구조와 클래스가 다르므로 클래스의 ID 가 없을 때가 있는데 현 클래스의 상위클래스의 ID 를 상위번호로 취하면 된다. 클래스 ID 의 신셋번호를 상위번호로 적용했으므로 클래스 ID 신셋라인 중 하위번호에는 현재 추가시킬 용어의 번호를 넣으면 새로운 신셋라인이 생기고 확장의 과정을 만드는 것이다. 예들 들어, 최근에 쓰이는 용어인 '웹 2.0(Web2.0)' 을 신셋에서 검색하면 없다. 각 용어사전 OWL 에서 인스턴스의 용어 '웹 2.0' 을 찾는다. 그러면 각 용어사전 OWL 에서 '웹 2.0' 을 검색할 수 있는데, 네이버 용어사전 OWL 은 Internet 의 인스턴스로 '웹 2.0' 이 포함되어 있고 텀즈 용어사전 OWL 은 InternetTech 의 인스턴스로 '웹 2.0' 이 포함되어 있다. 두 사전에서 네이버 사전에 우선 순위가 있으므로 신셋에 네이버 사전의 '웹 2.0' 을 추가하게 된다. '웹 2.0' 의 클래스 ID 는 Internet 이므로 Internet 신셋번호를 상위어로 하고, 영어와 번역어 위치에 'Web2.0' 과 '웹 2.0' 을 생성한다. 마지막으로 Internet 의 하위번호로 '웹 2.0' 의 신셋번호를 추가시키면 된다. <그림 5>은 신셋에 '웹 2.0' 을 추가시킨 결과이다.



<그림 5> 워드넷에 새로운 용어 추가 결과

5. 결론 및 향후 과제

사회의 다양성으로 새로운 용어들이 출현하고 수용해야 하는데 워드넷은 연구 성향이 있으므로 신생 용어 수용이 늦게 된다. 본 논문은 웹 사이트의 사전에서 컴퓨터 용어 사전 OWL 를 구축하고 워드넷에 없는 용어를 OWL 에서 추출하여 위드넷 신셋 형식을 갖추고 상하위관계를 가지는 컴퓨터 용어만을 추가시켜 어휘망 확장을 시켰다. 하지만 분야별 용어는 계속 생성되고 정립이 된 용어 사전 OWL 은 아직 미비하므로 분야별 전문 용어 사전 OWL 구축과 어휘망 확장을 위한 방법에 대한 연구가 필요하다.

참고문헌

[1] Tim Berners-Lee, James Hendler and Ora Lassila, "The Semantic Web," Scientific Amercan, Vol. 284, No. 5, pp. 34-43, 2001.

[2]Miller, G. et al., (1990), "Introduction to WordNet: an on-line lexical database." International Journal of Lexicography 3(4), pp.235-244.

[3]정홍석, "용어추천기능을 가진 온톨로지 편집기의 설계와구현",http://klpl.re.pusan.ac.kr/graduates/hsjung/

[4]W3C, "Resource Description Framework (RDF): Concepts and Abstract Syntax." URL: http://www.w3.org/TR/rdf-concepts/

[5]W3C, "RDF Vocabulary Description Language 1.0: RDF Schema."URL: http://www.w3.org/TR/rdf-schema/

[6]DARPA, "DARPA Agent Markup Language (DAML)." URL: http://www.daml.org

[7]W3C, "OWL Web Ontology Language Guide." URL: http://www.w3.org/TR/owl-guide/

[8]W3C, "OWL Web Ontology Language Overview." URL: http://www.w3.org/TR/owl-features/

[9] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. In Formal Ontology in Conceptual Analysis and knowledge representation. Kluwer Academic Press, 1993. also Technical Report KSL 93-04, Stanford University.

[10]권혁철, "시맨틱웹의 가능성과 한계", 한국과학기술정보연구원: 지식정보인프라지 통권 15 호, 2004.

[11]Extensible Markup Language (XML) 1.0 (Fourth Edition), http://www.w3.org/TR/2006/REC-xml-20060816

[12]Protege(2000). The Protégé Project. http://protege.stanford.edu

[13]Brain McBride, "An introduction to RDF and the Jena RDF API", http://jena.sourceforge.net/tutorial/RDF_API/index.html, 2006. 12

[14] Eric Prudhommeaux, Andy Seaborne, "SPARQL Query Language for RDF", W3C, October 2004