데이터의 웹을 위한 상호연결된 대규모 온톨로지 네트워크 구축

(Constructing a Large Interlinked Ontology Network for the Web of Data)

강 신 재*
(Sin-Jae Kang)

요 약 본 논문에서는 국내외 대표적 온톨로지 지식베이스의 연결을 통하여 대규모 온톨로지망을 구축할 수 있는 방법론을 제시한다. 온톨로지는 일반에 공개되어 공유될 때 그 가치가 커지게 되므로, 국내의 대표적인 CoreOnto 온톨로지를 기존 온톨로지망에 연결하여 국내외적으로 공개하고 활용성을 높이고자 한다. YAGO 온톨로지는 Wikipedia의 카테고리 정보와 WordNet의 계층정보를 추출하여 구축되었으며, DBpedia 분류체계의 백본으로 활용되었다. 이에 기반하여 WordNet의 Synset을 매개로 하여 CoreOnto 온톨로지를 YAGO와 DBpedia 온톨로지에 연결할 수 있는 방법론을 제시하였다.

핵심주제어 : 온톨로지 네트워크, 온톨로지 매핑, CoreOnto, DBpedia, YAGO, WordNet

Abstract This paper presents a method of constructing a large interlinked ontology network for the Web of Data through the mapping among typical ontologies. When an ontology is open to the public, and more easily shared and used by people, its value is increased more and more. By linking CoreOnto, an IT core ontology constructed in Korea, to the worldwide ontology network, CoreOnto can be open to abroad and enhanced its usability. YAGO is an ontology constructed by combining category information of Wikipedia and taxonomy of WordNet, and used as the backbone of DBpedia, an ontology constructed by analyzing Wikipedia structure. So a mapping method is suggested by linking CoreOnto to YAGO and DBpedia through the synset of WordNet.

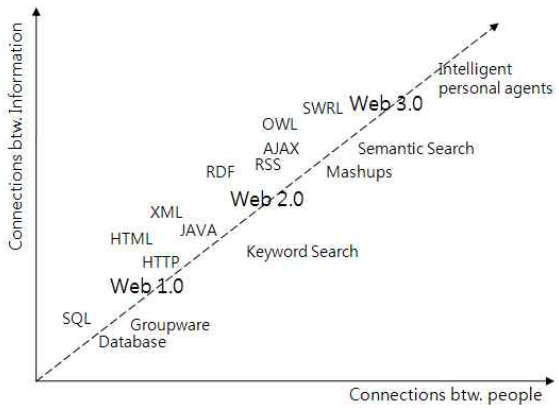
Key Words : Ontology Network, Ontology Mapping, CoreOnto, DBpedia, YAGO, WordNet

1. 서 론

인터넷 없이 살 수 없는 요즈음, 인터넷을 사용하는 주요한 이유 중 하나는 원하는 정보를 빠르게 찾는 것이다. 나아가 사용자가 정보를 요구하기 전에 사용자의 성향과 사용자가 처한 주변의 상황을 지능적으로 고려하여 정보를 의미적으로 검색하고 추천하는 것이야말로 현재의 웹이 궁극적으로 지향하는 한 방향이다. 키워드 검색 방식의 한계를 극복하기 위해서는 실세계 또는 특정 도메인에 속한 개념들의 정보를 체계적으로 기술한 온톨로지를 구축하고, 이를 기반

* 대구대학교 컴퓨터·IT공학부 교수

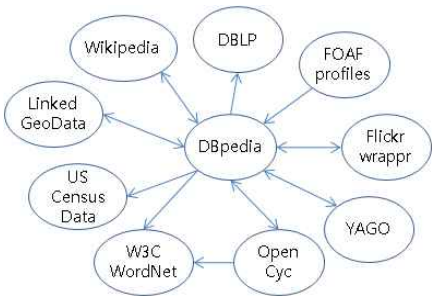
으로 웹에 표현되어 있는 정보들을 의미 태깅하는 기법이 필요하다. 이러한 기반이 마련되어야 지능적인 검색 및 추천 서비스의 개발이 가능하다. 현재 웹 기술의 수준은 그림 1과 같이 매쉬업 기술은 이미 활용되고 있으나, 의미기반 검색기술은 아직 개발 중인 Web 2.0(소셜 웹)과 Web 3.0(시맨틱 웹)의 사이에 위치한다고 볼 수 있다[1].



<그림 1> 웹 기술의 진화방향

웹 정보를 의미적으로 표현하기 위해서는 누구나 쉽게 접근해서 사용할 수 있고 신뢰할 수 있는 대규모의 온톨로지(ontology)가 필요하다. 온톨로지는 의미의 기본 단위인 개념과 그들 사이의 관계들로 표현되는데, 이렇게 표준화된 의미 단위와 표현방법을 사용함으로써 사람과 컴퓨터, 컴퓨터와 컴퓨터가 의미의 애매성 없이 정보를 해석할 수 있게 된다. 하지만 독자적으로 개발한 온톨로지는 구축 및 검증이 어려울뿐더러 모든 네티즌이 공감하고 공유할 수 있는 리소스가 아니기 때문에, 오히려 해당 응용서비스에서만 의미 있는 하나의 지식베이스라고 볼 수 있다. 따라서 본 연구에서는 독자적인 하나의 온톨로지를 구축하는 것이 아니라, 이미 개발된 국내외의 대표적인 온톨로지들을 연결하는 방법론을 제시하여 현실적이고 실용적인 방법으로 웹상에서 활용 가능한 범용 온톨로지 네크워크를 구축하는 방법론을 제시하고자 한다.

웹상에 존재하는 방대한 데이터가 서로 연결되지 않고 독립적으로 존재한다면 해당 사이트의 정보가 독립적으로 검색은 가능하겠지만, 관련된 모든 정보를 통합적으로 검색할 수는 없다. 이에 웹을 창안한 팀 버너스리 경이 '연결된 데이터'(linked data)의 구현이 필요함을 주장하였다[2]. '연결된 데이터'는 시맨틱 웹을 구현하기 위해서는 기본적으로 이루어져야 할 내용인데, URI(Uniform Resource Identifier)[3]/HTTP(Hypertext Transfer Protocol)/RDF(Resource Description Framework)[4] 등을 통하여 데이터를 노출/공유/연결할 수 있게 함으로써 이를 활용하여 다양한 지능적인 부가서비스를 개발할 수 있는 차세대 웹을 지향하고 있다. 위키피디아 사이트에 제시된 2009년 5월까지의 통계를 보면 웹상의 '연결된 데이터망'에는 총 42억 개의 RDF 트리플이 상호간 1억4천2백만 개의 링크로 연결되어 있다[5]. 다음은 '연결된 데이터망'에 포함된 대표적인 온톨로지인 WordNet[6], DBpedia[7], YAGO[8]를 중심으로 주요 웹 리소스 간 네트워크의 일부를 보여주고 있다[9]. 특히 DBpedia는 컴퓨터과학 분야의 도서목록 사이트인 독일의 DBLP, 미국의 인구조사 데이터, 지리정보 사이트 등 많은 다양한 오픈 데이터들과 연결되어 있어 그 활용가능성이 크다.



<그림 2> 연결된 데이터망

1) http://www.myplick.com/view/bYa3Nr7kudf/Nova-Spivvack-Understanding-the-Semantic-Web-and-Twine-Talk

2) http://www.w3.org/DesignIssues/LinkedData.html
3) http://www.w3.org/TR/uri-clarification/
4) http://www.w3.org/RDF/
5) http://en.wikipedia.org/wiki/Linked_data
6) http://wordnet.princeton.edu/
7) http://dbpedia.org/
8) http://www.mpi-inf.mpg.de/~suchanek/downloads/yago/
9) http://esw.w3.org/topic/SweoIG/TaskForces/CommunityProjects/LinkingOpenData#dbpedia-lod-cloud

이처럼 국외에서는 오픈 지식베이스간의 연결이나 병합이 활발하게 진행되고 있으나, 국내에서 개발된 온톨로지나 지식베이스를 '연결된 데이터망'에 연결하기 위한 시도는 아직 없는 상황이다. 국내에서 개발된 대표적인 온톨로지로는 CoreOnto[10]가 있는데, 이는 KAIST, ETRI를 중심으로 구축되고 있는 대규모 IT분야 온톨로지의 핵심 온톨로지이다.

이러한 국내외 상황에 근거하여 본 연구에서는 CoreOnto 온톨로지를 WordNet을 매개로 하여 DBpedia, YAGO 온톨로지에 연결함으로서 CoreOnto를 전 세계적인 '연결된 데이터망'에 포함시키고자 한다. 이를 통해 CoreOnto를 국외에 알리고 그 활용성을 높일 수 있겠다.

2장에서는 현존 대표 온톨로지와 관련 연구를 소개하고, 3장에서는 CoreOnto 온톨로지를 기존 온톨로지 네트워크에 연결하는 방법을 제시한다. 4장에서는 결론과 향후 연구계획을 제시한다.

2. 관련 연구

온톨로지는 일반에 공개되어 모든 사람에 의해 공유되어 사용될 때 그 가치가 커지게 된다. 국내의 대표적인 온톨로지인 CoreOnto를 국내외적으로 공개하고 활용성을 높이기 위해서 '연결된 데이터망'내의 WordNet과 DBpedia, YAGO 온톨로지에 연결하고자 한다. 이 장에서는 위에서 언급한 각 온톨로지의 개발배경 및 구축방법론을 소개하고 온톨로지간 조정(ontology mediation)을 위한 기존 접근법에 대해 설명한다.

2.1 WordNet

WordNet[1]은 프린스턴 대학에서 개발된 대규모의 영어 어휘 의미 목록 지식베이스이다. WordNet은 단어의 의미를 구분하기 위해 Synset이라는 유의어 집단을 정의하여 사용하고 있으며, Synset 사이에는 상위어, 하위어, 전체어, 부분어 등 다양한 의미 관계가 존재한다. WordNet 3.0에서는 117,798개의 개별 명사에 대해 총 82,115개의 Synset을 포함하고 있다.

영어를 대상으로 구현된 대부분의 자연어처리 응용 시스템에서는 단어 의미 구분 등의 처리를 하기 위해 WordNet을 활용하고 있으며, 이를 편리하게 하기 위해 WordNet 검색, 유사도 계산 API 등 많은 라이브러리와 소프트웨어 도구들이 개발되어 제공되고 있다. 따라서 온톨로지간 연결을 위해 WordNet의 Synset을 활용한다면 사람의 수작업을 줄이고 매핑 과정을 (반)자동화하는 데에 큰 도움을 얻을 수 있다.

2.2 DBpedia

DBpedia는 온라인 백과사전인 위키피디아(Wikipedia)에서 제공하는 인포박스(infoboxes) 정보를 분석하고 260만 개체 정보를 추출하여 구축한 지식베이스이다[2]. 인포박스는 테이블 형태로 정보를 제공하는 부분인데, 위키피디아 페이지별로 다른 템플릿을 사용하고 있기 때문에 일관된 방법으로 정보를 추출하기가 어렵다.

DBpedia 프로젝트에서는 인포박스의 분석을 위해 두 가지 방법을 사용하였다. 먼저 일반 인포박스(Generic Infobox) 추출방법은 해당 위키피디아 문서의 URI를 주어(subject)로 변환하고, 인포박스의 속성명을 술어(predicate)로, 그리고 인포박스의 속성값을 목적어(object)로 변환하여 하나의 RDF를 생성하는 방법이다. 이는 위키피디아 내의 모든 인포박스에 적용할 수 있는 장점이 있으나, 유사 속성명을 처리할 수 없다는 단점을 가지고 있다.

두 번째 방법인 매핑기반 인포박스 추출방법은 위키피디아 내 주요 인포박스 템플릿을 분석하고 이를 표현하기 위한 온톨로지를 정의하여 사용하는 방법이다. 350개의 템플릿을 분석하여 170개 클래스, 720개 속성, 55개의 데이터타입을 갖는 온톨로지를 정의하여 사용하였다. 이 방법은 고품질의 데이터 추출이 가능하나, 모든 위키피디아 웹 페이지에 적용할 수는 없는 문제점을 갖고 있다.

위키피디아로부터 추출된 개체들은 분류를 위

10) http://cscola.kaist.ac.kr/wiki/index.php/Overview

해 위키피디아 카테고리, YAGO, UMBEL, DBpedia 온톨로지의 개념분류에 할당되었다. 추출된 개체들은 Virtuoso라는 RDF 저장소에 저장되었으며, SPARQL로 검색하거나, RDF 술어의 종류별로 다운받아 사용할 수 있다. DBpedia 릴리스 3.2에 포함된 대표 클래스별 인스턴스의 수는 표 1과 같다.

<표 1> DBpedia 온톨로지 내 인스턴스의 수(릴리스 3.2 기준)

Class	Instances
Place	248,000
Person	214,000
Work	193,000
Species	90,000
Organization	76,000
Building	23,000
Resource(overall)	882,000

2.3 YAGO

YAGO는 Wikipedia 카테고리 계층의 말단 카테고리와 WordNet의 계층정보를 매핑하여 구축한 지식베이스이다[3]. 사람(person), 장소(location)와 같은 개체(entities) 정보를 170만 개 이상 포함하고 있으며, 두 개체가 하나의 관계(relation)를 맺고 있는 트리플 단위의 사실(facts) 정보는 1,500만 개 이상 포함하고 있다. YAGO에서는 사실이 발견된 출처 정보를 함께 저장하기 위해 사실 식별자(fact identifier)를 도입하고, 이를 통해 구체화 그래프(reification graph) 형태로 온톨로지를 표현할 수 있다. 예를 들어 “엘비스가 1935년에 태어났다는 사실을 위키피디아에서 발견했다”는 내용을 구체화 그래프로 표현하면 다음과 같이 표현할 수 있는데, 이는 RDFS[11]와 OWL[12]에서 허용하는 이진 관계(binary relations) 뿐만 아니라 다항 관계(n-ary relations)도 표현가능하게 한다.

#1 : Elvis BornInYear 1935

#2 : #1 FoundIn Wikipedia

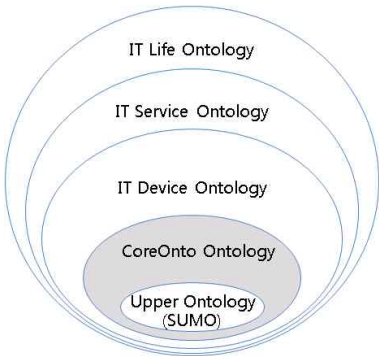
위에서 #1, #2는 사실 식별자에 해당하고, Elvis, Wikipedia 등은 개체 또는 인자(argument)에, BornInYear, FoundIn은 관계에 해당한다. YAGO 모델은 기본적으로 RDFS의 확장된 형태인데, RDFS의 Domain, Range, Type 관계와 subClassOf, subPropertyOf 관계를 포함하고 있으며, 여기에 비순환 전이 관계(acyclic transitivity relation)의 특징을 추가하여 확장되었다.

YAGO는 자동 구축된 여타의 지식베이스와 비교할 때 수작업으로 확인된 정확률이 95%에 이르며, DBpedia 온톨로지 분류체계의 백본으로 사용되기도 하였다.

2.4 CoreOnto

CoreOnto 온톨로지는 KAIST와 ETRI를 중심으로 국내에서 구축되고 있는 대규모 IT분야 온톨로지의 핵심부분이다[4]. SUMO 온톨로지[5]를 최상위 온톨로지로 삼고 있으며, 위키피디아에서 IT분야의 웹페이지 등을 분석하여 관련 어휘를 추출하고 의미 관계를 부가하여 온톨로지에 추가하는 방법으로 구축되었다. CoreOnto 프로젝트는 IT 분야에 범용적으로 활용이 가능한 IT 온톨로지를 구축하고 인터넷, 인트라넷, 유비쿼터스 환경에서 제공되는 각종 IT 서비스에 적용하여 단절없는(seamless) 서비스를 제공하는 것을 목표로 하고 있으며, 국가 차원에서 구축되어 국내 IT 분야뿐만 아니라 국제적으로도 표준적으로 활용할 수 있는 온톨로지를 구축하는 것을 목표로 하고 있다. CoreOnto는 차세대 이동통신, 홈네트워크, 디지털 TV/방송, 텔레매틱스, 지능형로봇, 차세대 PC, 임베디드 S/W 등에 공통적으로 적용될 수 있는 IT 온톨로지이다. 그림 3은 CoreOnto 온톨로지가 전체 IT 온톨로지에서 차지하고 있는 위치를 보여주고 있다. CoreOnto 온톨로지를 중심으로 IT 디바이스, 서비스, 라이프 온톨로지를 단계적으로 구축하여 확장하고 있다.

11) http://www.w3.org/TR/rdf-schema/
12) http://www.w3.org/TR/owl-features/



<그림 3> IT 온톨로지 구성

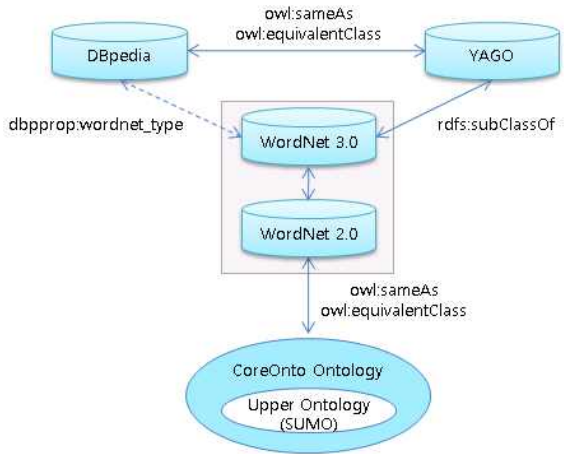
2.5 온톨로지 조정(Ontology Mediation)

온톨로지가 여럿 있을 때 취할 수 있는 방법으로는 온톨로지의 구조까지 완전히 병합하여 새로운 하나의 온톨로지를 만드는 방법(ontology merging)과, 각각의 온톨로지 구조는 그대로 두고 관련된 클래스간 매핑을 하여 온톨로지간 네트워크를 구성하는 방법(ontology mapping)이 있다[6]. 앞에서 언급한 WordNet, DBpedia, YAGO, CoreOnto 등의 온톨로지는 매우 방대한 크기를 가지고 있으므로 본 연구에서는 난이도와 효율성을 고려하여 후자의 방법을 취한다.

온톨로지간 관련된 클래스를 찾는 방법으로는 Edit Distance 등을 이용한 클래스명 어휘 유사도, WordNet 등을 이용한 클래스명 의미 유사도, 클래스 상하위관계 유사도, 상하위관계를 제외한 클래스 관계/속성 유사도 등이 있다[6,7]. 본 연구에서는 위 방법들을 혼합하여 적용한다.

3. 온톨로지 네트워크 구축

국내에서 대규모로 개발된 CoreOnto 온톨로지를 국제적으로 활성화되고 있는 데이터망에 연결하여 그 활용성을 높이고, 차세대 웹 서비스 개발을 위한 상호 시너지 효과를 일으키기 위해서 CoreOnto를 DBpedia와 YAGO에 연결하고자 한다. 이를 위한 기본적인 아이디어는 다음과 같다. YAGO 온톨로지는 WordNet의 분류체계를 기반으로 구축되었으며, DBpedia 분류체계의 백본으로 활용되었다. 따라서 WordNet의 Synset을 매개로 하여 CoreOnto 온톨로지와 YAGO/DBpedia 온톨로지 간 연결이 가능하게 된다. 다음 그림에서 전체적인 온톨로지 네트워크 구성방법과 각 온톨로지간 연결을 위해 사용되는 관계의 종류들을 제시하였다.



<그림 4> 온톨로지 네트워크 구성도

3.1 DBpedia-YAGO간 연결

DBpedia에 포함된 개체와 YAGO에 포함된 개체간 매핑은 URI를 제외한 개체명과 클래스명을 대상으로 Edit Distance 기법을 응용하여 구현한다. 하나의 문자열을 다른 문자열로 변환하기 위해 필요한 문자의 삽입, 삭제, 대체의 최소 연산수를 구하는 Levenshtein distance 기법[8]을 기본으로 하여, [9]에서 제시한 Edit distance 수식을 아래와 같이 변형, 적용하여 클래스간 유사도를 계산한다. 클래스명을 정규화하는 이유는 동일한 클래스가 단복수, 복합명사, 약어 사용여부에 따라 다르게 표현되는 경우를 모두 고려하여 유사도를 계산하기 위함이다. 이와 같은 과정을 거쳐 자동으로 매핑후보를 찾은 후, 최종적으로 수작업 검증하여 매핑을 하게 된다.

$$\mathrm{NameSim}(c1,c2) = \frac{|\{(c1,c2)|\delta(a,b) > h \ \ such\ that\ a \in tokens(c1)\ and\ b \in tokens(c2)\}|}{\max(|tokens(c1)|,|tokens(c2)|)}$$

c1, c2: 클래스명 or 개체명
tokens(c): 클래스 c를 정규화한 후보 토큰의 집합
$\delta(c1,c2)$: 클래스 c1과 c2 사이의 Levenshtein distance
h: 임계값

매핑 방법의 유용성을 확인하는 실험을 위해서 DBpedia와 YAGO 온톨로지에 존재하는 방대한 양의 개체(인스턴스)는 일단 제외하고, DBpedia 3.4 온톨로지[13] 내 204개의 클래스와 YAGO 계층체계[14] 내의 182,947개의 클래스를 대상으로 위에서 언급한 매핑 방법론을 적용해 보았다. 정확률과 적용률을 각각 높이기 위해 임계값 설정을 한 실험결과는 표 2와 표 3에 나타나 있다.

<표 2> 정확률을 높이기 위해 임계값 *h*를 높인 경우

매핑 대상 클래스수 (A)	204
임계값을 넘은 매핑 클래스수 (B)	26
정확한 매핑수 (C)	25
정확률 (C/B)	96.2 %
적용률 (B/A)	12.7 %

<표 3> 적용률을 높이기 위해 임계값을 적용하지 않은 경우

매핑 대상 클래스수 (A)	204
임계값을 넘은 매핑 클래스수 (B)	204
정확한 매핑수 (C)	50
정확률 (C/B)	24.5 %
적용률 (B/A)	100 %

실험결과와 같이 정확률을 높이려면 임계값을 높이면 되나, 적용률이 떨어지게 되고, 적용률을 높이려면 임계값을 낮추면 되지만 정확률이 떨어지게 된다. DBpedia와 YAGO간 연결은 수작업으로 최종 검증을 하기 때문에, 적용률을 높인 상태에서 매핑 후보를 찾고 이를 대상으로 수작업 검증하는 것이 가장 효율적인 접근방법이라 할 수 있겠다. 매핑 후보가 없는 상태에서 매핑을 하는 것은 어렵지만 매핑 후보가 있는 상태에서는 작업이 효율적으로 진행될 수 있기 때문이다.

정확률을 떨어뜨리는 또 다른 이유로 DBpedia 온톨로지에 존재하는 클래스에 대해 YAGO 온톨로지에 매핑에 적합한 클래스가 없을 수도 있다는 점이다. 이러한 경우에는 해당 클래스를 YAGO에 새로이 정의/추가하여 직접 매핑을 해주는 방법을 생각해 볼 수 있다. 실제로 YAGO 사이트에서 제공하는 DBpedia와의 매핑 도구[15]를 살펴보면, YAGO 클래스에 대응하는 DBpedia 클래스를 찾는 노력을 하지 않고, 모든 YAGO 클래스를 DBpedia 온톨로지에 정의하여 추가하고 매핑 관계를 맺어주는 접근법을 취하고 있다. 이는 구현이 쉬울 수는 있지만, 클래스의 중복 정의(동일 클래스가 단복수, 복합명사, 약어사용여부에 따라 다른 형태로 표현되는 경우 포함)라는 문제를 발생시키게 된다.

따라서 본 연구에서는 기존 정의된 클래스에서 먼저 매핑 대상을 찾고, 적합한 클래스가 없는 경우에 한하여 신규 클래스를 정의하고 매핑 관계를 맺어주는 방법을 취하여, 데이터의 중복 문제를 해결하고자 한다.

N-Triple 표기법[16]으로 나타낸 그림 5와 같이, 개체(instance)간 매핑은 owl:sameAs[17] 관계를 이용하여 연결되며, 클래스간 매핑은 owl:equivalentClass[18] 관계로 연결된다.

<그림 5> DBpedia와 YAGO의 매핑 예시

13) http://wiki.dbpedia.org/Downloads34
14) http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html
15) http://www.mpi-inf.mpg.de/yago-naga/yago/converters.zip
16) http://www.w3.org/TR/rdf-testcases/#ntriples
17) http://www.w3.org/TR/2004/REC-owl-features-20040210/#sameAs
18) http://www.w3.org/TR/2004/REC-owl-features-20040210/#equivalentClass

3.2 DBpedia-WordNet간 연결

DBpedia와 WordNet 간의 연결을 위해서 dbprop:wordnet_type[19] 관계가 이미 정의되어 있으나, 현재까지 구축된 DBpedia에는 이러한 관계 정보가 추가되지는 않았다. 본 연구에서는 YAGO를 통하여 DBpedia에 연결하는 방법을 취하고 있으므로 DBpedia-WordNet간 추가 연결을 위한 방법론은 제안하지 않는다.

3.3 YAGO-WordNet간 연결

YAGO에 포함된 위키피디아 카테고리 체계의 말단 부분은 구체적인 내용이 대부분이기 때문에 WordNet의 Synset과 하위어-상위어 관계(rdfs:subClassOf)의 형태로 이미 매핑되어 있다. 예를 들어 그림 6에서 제시된 위키피디아 말단 카테고리 중 "Capitals_in_Asia"는 WordNet의 Synset ID가 "08518505"인 "Capital" Synset의 하위클래스로 매핑되어 있으며, 이 관계의 신뢰도는 0.95임을 나타낸다.

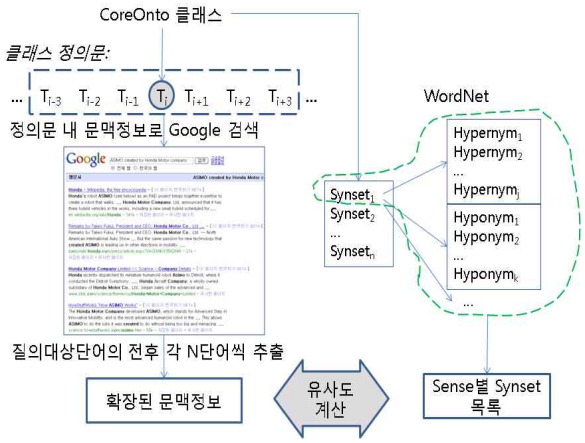
```
<?xml version="1.0"?>
<!-- This is a sample snippet from the subClassOf hierarchy of YAGO in RDFS -->
<!DOCTYPE rdf:RDF [<!ENTITY d "http://www.w3.org/2001/XMLSchema#">
                  <!ENTITY y "http://www.mpii.de/yago/resource/" >]>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
         xml:base="http://www.mpii.de/yago/resource"
         xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
         xmlns:y="http://www.mpii.de/yago/resource/">
<rdfs:Class rdf:about="&y;wikicategory_Capitals_in_Asia">
   <rdfs:subClassOf rdf:ID="f600000009" rdf:resource="&y;wordnet_capital_108518505"/>
</rdfs:Class>
<rdf:Description rdf:about="#f600000009">
   <y:confidence rdf:datatype="&d;double">0.9511911446218017</y:confidence>
</rdf:Description>
<rdfs:Class rdf:about="&y;wikicategory_Coastal_cities">
   <rdfs:subClassOf rdf:ID="f600000021" rdf:resource="&y;wordnet_city_108524735"/>
</rdfs:Class>
<rdf:Description rdf:about="#f600000021">
   <y:confidence rdf:datatype="&d;double">0.9511911446218017</y:confidence>
</rdf:Description>
```

<그림 6> YAGO 계층 체계 및 WordNet과의 매핑 정보

3.4 CoreOnto-WordNet간 연결

CoreOnto와 WordNet간 연결만 이루어지면 온톨로지 네트워크에서 CoreOnto와 DBpedia간 연결이 이루어지게 되므로 가장 중요한 부분이다. WordNet Synset을 매개로 CoreOnto 온톨로지를 연결하기 위해 고려할 사항은 다음과 같다.

CoreOnto는 WordNet 2.0 기반으로 되어 있는 반면, YAGO는 WordNet 3.0을 기반으로 구축되어 있어 WordNet 버전간 Synset ID의 매핑이 필요하다. 관련하여 [10]에서는 Synset간 관계 정보인 구조적 정보와 정의문(gloss), Synset에 포함되어 있는 단어정보를 활용하여 다양한 WordNet 버전간 Synset의 매핑작업을 하였다. 본 연구에서는 WordNet 2.0과 3.0의 매핑이 필요하므로, [10]의 결과물을 활용하여 변환하였다. 또한 CoreOnto 온톨로지의 상위 온톨로지로 사용된 SUMO[20]와 WordNet과의 매핑 정보는 오픈 소스의 형태로 공개되어 있다[21]. 따라서 CoreOnto 온톨로지에서 상위 온톨로지를 제외한 나머지 클래스들을 WordNet의 Synset으로 매핑하는 방법만 있으면 된다. 이는 단어의미중의성 해소 문제와 유사하므로, [11]에서 성능이 입증된 단어의미중의성 해소 기법을 변환하여 적용한다.



<그림 7> CoreOnto 클래스와 WordNet Synset 간 매핑 방법

19) DBpedia에서 정의한 속성(property)임

20) http://www.ontologyportal.org/

21) http://sigmakee.cvs.sourceforge.net/sigmakee/KBs/WordNetMappings/

CoreOnto 클래스 w의 직접 문맥정보(direct contextual information)로 w의 정의문에서 w의 전후 3개의 단어(CW, context window)를 추출한다. 그러나 이 직접 문맥정보는 양이 적어 추가의 문맥정보를 획득할 필요가 있다. 이를 위해 직접 문맥정보를 질의어로 Google 사이트를 검색하고 검색된 상위 100건의 결과 요약텍스트에서 간접 문맥정보를 추출한다. 간접 문맥정보의 추출을 위해 w와 w의 직접 문맥 정보로 추출된 각 단어 s에 대해 요약텍스트 내에서 s의 전후 3개의 단어를 추출하는 방식을 취한다(그림 7의 왼쪽 부분). 간접 문맥정보 추출 전에 Google에서 검색된 요약텍스트로부터 HTML 태그와 불용어(stopword)를 제거하는 전처리 과정이 필요하다. 추출된 간접 문맥정보 내의 각 단어에 대해 원형(root form)을 복원한 후 원형이 WordNet에 등록되어 있는지를 검사하여 WordNet에 등록되지 않은 단어는 문맥 단어로 고려하지 않는다. 이러한 과정을 통해, 각 클래스 w에 대해, w가 출현한 문맥과 유사한 문맥에서 출현한 공기(collocation)단어와 그 빈도수를 문맥정보로 획득할 수 있다.

각 클래스별 의미의 구분을 위해서는 의미에 해당하는 Synset 뿐만 아니라 상위어, 하위어, 전체어, 부분어에 해당하는 Synset들까지 모두 모아 Synset 목록을 구축한다(그림 7의 오른쪽 부분).

마지막으로 CoreOnto 클래스를 의미적으로 적합한 WordNet의 Synset과 매핑하기 위해서 정의문으로부터 유도된 문맥정보(단어목록)와 단어의미별 Synset 목록(단어목록)을 비교하여 유사도(단어 중복도)가 가장 높은 Synset을 선택하면 된다.

이러한 과정을 거치게 되면 최종적으로 CoreOnto 온톨로지 내 모든 클래스들이 WordNet에 매핑되게 되며, 결과적으로 YAGO와 DBpedia와도 연결이 되게 된다.

4. 결론 및 향후 연구과제

시맨틱 웹(semantic web)은 기계가 이해할 수 있는 형태로 웹상의 데이터를 표현함으로써 웹 서비스 간 검색(search)/조정(mediation)/실행(execution)이 자동으로 이루어지고 이를 기반으로 지능적인 웹 서비스가 개발될 수 있는 환경이다. '연결된 데이터망'은 시맨틱 웹으로 가기 위해 필수적인 과정인데, 온톨로지 네트워크의 구축이 이에 핵심적인 역할을 한다.

본 연구에서는 국내에서 대규모로 구축되고 있는 CoreOnto 온톨로지를 DBpedia와 YAGO 온톨로지에 WordNet의 Synset을 매개로 연결할 수 있는 방법론을 제시함으로써 국내 연구결과물을 국제적으로 널리 알리고 활용할 수 있는 계기를 마련하였다.

온톨로지 매핑결과를 검증하기 위해서는 매핑 정답이 포함되어 있는 평가 데이터 세트가 있어야 하나, 본 연구의 대상이 되는 CoreOnto, DBpedia, YAGO 온톨로지는 방대한 양의 최신 지식베이스이어서 이러한 데이터 세트가 아직 없고 또한 이를 활용한 웹 서비스(검색엔진 등)의 구현도 미미한 상황이다. 시간이 다소 걸릴 것으로 예상되나, 온톨로지 네트워크의 구축을 완료한 이후에 매핑 데이터를 샘플링하고 수작업으로 검증한 데이터 세트를 구축하고 나면 검증이 이루어질 수 있겠다. 검증을 위한 또 다른 방법으로 온톨로지 네트워크를 이용한 응용서비스의 개발을 들 수 있는데, 그림 1에서 언급한 의미 검색(semantic search) 또는 지능적인 개인 에이전트(intelligent personal agents) 서비스를 구현하여 독립적인 온톨로지를 사용하는 경우와 온톨로지 네트워크를 사용한 경우의 차이를 비교해 보는 것이다. 이는 향후 연구과제로 진행하고자 한다.

참 고 문 헌

[1] C. Fellbaum, *WordNet: An Electronic Lexical Database (Language, Speech, Communication)*, MIT Press, 1998.

[2] C. Bizer, J. Lehmann, G. Kobilarov, S. Auer, C. Becker, R. Cyganiak, and S. Hellmann, "DBpedia - A Crystallization Point for the Web of Data", Journal of

Web Semantics (JWS), vol. 7, no. 3, pp.154-165, 2009.
[3] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: A Core of Semantic Knowledge", 16th International World Wide Web Conference (WWW 2007), pp.697-706, Canada, 2007.
[4] J. Ahn, I. Moon, S. Nam, and K. Choi, "CoreOnto: a semi-automated ontology building toolkit", Proceedings of the 3rd Asian Semantic Web Conference (ASWC2008), Industry session, Bangkok, Thailand, 2008.
[5] I. Niles, and A. Pease, "Towards a Standard Upper Ontology", Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001), pp.2-9, Maine, 2001.
[6] J. Euzenat, and P. Shvaiko, *Ontology Matching*, Springer, 2007.
[7] P. Shvaiko, and J. Euzenat, "A survey of schema-based matching approaches," Journal on Data Semantics, vol. 4, pp.146-171, 2005.
[8] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," English Translation in Soviet Physics Doklady, vol. 10, no. 8, pp.707-710, 1966.
[9] 안진현, 최기선, "반복적 알고리즘을 이용한 온톨로지 매핑", 제21회 한글 및 한국어 정보처리 학술대회 논문집, pp.14-18, 2009.
[10] J. Daudé, L. Padró, and G. Rigau, "Mapping WordNets Using Structural Information", In Proceedings 38th Annual Meeting of the Association for Computational Linguistics (ACL00), pp.504-511, Hong Kong, 2000.
[11] 강신재, 강인수, "WordNet과 구글에 기반한 온톨로지 개체의 일반화", 한국지능시스템학회 논문지, 제19권, 3호, pp.363-370, 2009.

강 신 재 (Sin-Jae Kang)

- 종신회원
- 1995년 : 경북대학교 컴퓨터공학과 (공학사)
- 1997년 : 포항공과대학교 (POSTECH) 컴퓨터공학과 (공학석사)
- 2002년 : 포항공과대학교(POSTECH) 컴퓨터공학과 (공학박사)
- 1997년 ~ 1998년 : SK Telecom 정보기술연구원 연구원
- 2007년 : 오스트리아 University of Innsbruck, DERI 연구소 방문교수
- 2002년 ~ 현재 : 대구대학교 컴퓨터·IT공학부 부교수
- 관심분야 : Semantic Web, Social Web, Ontology Matching, Natural Language Processing

논 문 접 수 일 : 2009년 9월 22일
1차수정완료일 : 2009년 12월 30일
2차수정완료일 : 2010년 2월 1일
게 재 확 정 일 : 2010년 2월 10일