http://dx.doi.org/10.5392/JKCA.2012.12.12.035

연구 개발 트렌드 분석을 위한 기술 지식 온톨로지 구축

Ontology Construction of Technological Knowledge for R&D Trend Analysis

황미녕*, 이승우*, 조민희*, 김순영**, 최성필*, 정한민*
한국과학기술정보연구원 소프트웨어연구실*, 해외정보실**

Mi-Nyeong Hwang(mnhwang@kisti.re.kr)*, Seungwoo Lee(swlee@kisti.re.kr)*,
Minhee Cho(mini@kisti.re.kr)*, Soon Young Kim(maya@kisti.re.kr)**,
Sung-Pil Choi(spchoi@kisti.re.kr)*, Hanmin Jung(jhm@kisti.re.kr)*

요약

과학기술 분야 연구자들은 이전 연구와 개발 결과에 대한 조사 연구에 많은 시간을 소비한다. 또한, 연구자들은 유리한 입지를 성공적으로 차지하기 위해 일반적으로 학술 논문, 특허, 최근 연구 동향에 대한 웹 문서 등의 다양한 학술 자원을 분석하여 새롭게 등장하는 연구 주제를 선점하려고 한다. 하지만 키워드 기반의 정보 검색이나 참고문헌 정보에 근거한 연관 문서 추출 방법을 사용해서는 방대한 문헌에서 투자 가능한 연구 주제를 효율적으로 찾는 일이 쉽지 않다. 본 논문에서는 대규모 기술 문헌 자료에서 추출되는 기술, 제품, 연구 주체 간의 의미론적으로 연결된 정보를 효율적으로 생성, 저장하고 활용할 수 있는 방법을 제안한다. 세부적으로 텍스트 마이닝 기술을 활용하여 문헌에서 나타나는 주요 개체들과 연관 관계를 추출하여 시맨틱 웹 환경에 적용 가능한 기술 지식으로 생성하는데 적합한 온톨로지를 구축한다. 이렇게 구축된 온톨로지는 연관 관계를 가진 기술 지식 탐색을 지원하기에 연구 개발 트렌드 예측 및 분석 서비스인 InSciTe Adaptive에 사용되었다.

■ 중심어 : | 기술 지식 | 온톨로지 구축 | 시맨틱 웹 | 정보 시스템 |

Abstract

Researchers and scientists spend huge amount of time in analyzing the previous studies and their results. In order to timely take the advantageous position, they usually analyze various resources such as paper, patents, and Web documents on recent research issues to preoccupy newly emerging technologies. However, it is difficult to select invest-worthy research fields out of huge corpus by using the traditional information search based on keywords and bibliographic information. In this paper, we propose a method for efficient creation, storage, and utilization of semantically relevant information among technologies, products and research agents extracted from 'big data' by using text mining. In order to implement the proposed method, we designed an ontology that creates technological knowledge for semantic web environment based on the relationships extracted by text mining techniques. The ontology was utilized for InSciTe Adaptive, a R&D trends analysis and forecast service which supports the search for the relevant technological knowledge.

■ keyword : | Technological Knowledge | Ontology Construction | Semantic Web | Information System |

접수번호 : #121009-004
접수일자 : 2012년 10월 09일
심사완료일 : 2012년 12월 13일
교신저자 : 최성필, e-mail : spchoi@kisti.re.kr

1. 서 론

대용량의 정보를 체계적으로 구조화하여 사용자에게 의미 있는 지식을 제공하는 방법론에 대한 연구가 활발히 진행되고 있다. 지식정보처리 관점에서는 다양한 분야의 지식이 담겨 있는 문서 데이터에서 개체를 자동으로 추출하여 개체명을 인식하고, 그 중에서 핵심 개체를 발견하여 추출되는 개체들의 연관 관계를 분석한다. 그 이후 이들 정보들 사이에서 새로운 지식을 자동으로 추론하는 일련의 지식정보화 과정을 통해 사용자에게 의미 있고 정확한 지식을 분석하여 제공할 수 있다[1].

여기에서 데이터, 정보, 지식의 차이는 다음과 같다. 데이터는 가공되지 않는 사실이고 정보는 데이터를 1차 가공한 값이며, 지식은 정보를 집적하며 체계화하여 미래의 사용에 대비하여 보편성을 갖도록 한 형태로 볼 수 있다[2]. 즉, 문서를 구성하고 있는 단어들 그 자체는 데이터이고, 이 단어들의 개체를 인식하고 개체 간의 관계를 파악하는 것은 정보가 된다. 이 정보들을 연관시켜 새로운 정보를 획득하면 지식이 된다. 이러한 지식 획득의 근간이 되는 문서 데이터의 기하급수적인 증가 추세는 과학기술 분야의 연구자들에게도 연구 주제의 선정 및 관심 기술의 동향 파악을 위해 데이터에서 지식을 추론해내는 지능형 정보 서비스의 필요성을 유발시키는 계기가 되었다[3].

본 논문에서는 다양한 과학기술문헌으로부터 텍스트 마이닝 기법을 이용하여 추출한 개체 정보 및 이들 개체 간의 관계 정보 등의 관계 정보를 '기술 지식(Technological Knowledge)'으로 정의한다. 이렇게 생성되는 기술 지식을 효율적으로 관리하고, 재사용하며, 공유하기 위해서는 2001년 팀 버너스리(Tim Berners-Lee)에 의해 제시된 시맨틱 웹 기술이 적합하다[4]. 시맨틱 웹은 기계가 이해할 수 있는 의미를 기반으로 하여 의미적 상호 운용성(Interoperability)을 실현하여 다양한 정보 자원의 처리 자동화, 데이터 통합 및 재사용성을 기계가 스스로 수행하여 인간과 컴퓨터 간의 효과적인 협력체계를 구축하기 위한 기술이다.

이와 같은 시맨틱 웹 서비스에는 시맨틱 웹에 알맞은 데이터를 생성하는데 필요한 기술, 방대해진 데이터를 다룰 수 있는 빅데이터 처리 기술, 도출된 결론을 사용자에게 전달하는 인터페이스 기술이 필요하다. 본 논문에서는 기술 지식의 활용성 극대화를 위해 기반이 되는 온톨로지 설계 과정을 소개하고, 이를 기초로 하여 문헌에서 추출된 기술 지식을 온톨로지 정보 시스템으로 구축하는 방법에 대해 설명하고자 한다.

2. 관련 연구

2.1 RDF 데이터 모델

본 논문에서는 RDF를 사용하여 온톨로지를 구축한다. RDF(Resource Description Framework)은 W3C에서 제정한 것으로 특정 리소스를 정의하고 그 리소스에 대한 설명이나 관계를 기술함으로써 온톨로지를 구축할 수 있는 방법을 제공한다[5]. RDF는 기본적으로 주어부(Subject), 서술부(Predicate), 목적부(Object)의 트리플(Triple) 모델로 기술된다. 주어부는 표현하고자 하는 리소스를 의미하며 서술부는 주어부에 대해 기술하거나 주어부와 목적부의 관계를 의미한다. 목적부란 서술부에 대한 값이나 내용을 의미한다.

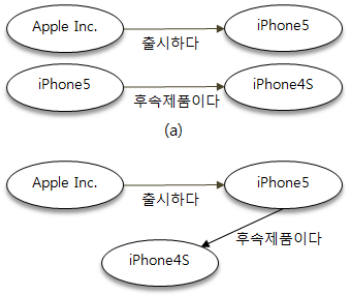


그림 1. 트리플 표현 예(a)와 그래프로 병합된 두 트리플 예(b)

예를 들어, 트리플 형식으로 "Apple Inc.는 iPhone5를 출시하다."라는 사실을 [그림 1](a)와 같이 표현할 수 있다. 이 트리플은 "iPhone5는 iPhone4S의 후속제품이다." 트리플과 'iPhone5'라는 공통 노드를 연결함으로써 방향성이 있는 [그림 1](b)와 함께 그래프로 표현된다. 여기서 각 리소스는 URI(Uniform Resource Identifier)

라는 식별자를 사용하여 구별한다. 시맨틱 웹은 이러한 트리플 구조에 기반을 두어 방향성이 있는 그래프 형태의 의미 정보로 표현되는데 이를 기술 지식의 의미 관계 표현에 적용하기로 한다.

2.2 온톨로지

온톨로지란 도메인 내에서 공유되는 데이터들의 개념화한 형식적이고 명백한 규정이며 이는 특정 분야에서 사용되는 표준 어휘들의 집합이라고 할 수 있다. 즉, 온톨로지는 도메인 내의 지식을 개념화하고 이를 명세화하는 것으로 정의된다. 도메인 내의 지식의 연결뿐만 아니라 지식이 문법과 어휘 같은 언어적인 구조의 연관 관계를 이해할 수 있도록 설계되어있다. 따라서 온톨로지는 다양한 도메인에 적용이 가능하도록 표준을 제시함으로써 웹 문서에 나타난 지식을 표현, 공유, 재사용하는 것을 목적으로 두고 있다. 의미에 따른 추론을 하기 위해서 온톨로지는 시맨틱 웹의 중요한 기반이 된다.

온톨로지는 개념의 정의인 클래스(Class), 개념의 구체화된 개체인 인스턴스(Instance), 클래스나 인스턴스 사이에 존재하는 관계나 클래스나 인스턴스의 특정한 성질, 성향을 나타내기 위해 특정한 값과 연결시킨 속성(Property), 추론 규칙인 공리(Axiom) 등으로 구성된다. 또한, 구축 범위에 따라 일반적인 개념을 대상으로 구축하는 일반 온톨로지(Generic or common-sense ontology)와 특정 영역에서 유효한 지식들을 대상으로 구축하는 도메인 온톨로지(Domain ontology)로 구분된다. 도메인 온톨로지는 인공지능, 정보검색, 유비쿼터스, 전자상거래, 의학 분야에서 활발하게 구축되고 있다[6-8].

2.3 트리플 데이터 생성 및 변환

시맨틱 웹 데이터를 생성하기 위해서는 처음부터 시맨틱 웹 데이터 모델에 맞는 RDF 데이터를 입력하거나 기존의 RDB 형태로 저장되어 있는 데이터를 RDF 형태로 변환하는 방법이 있다[9].

문서에서 텍스트 마이닝 기법을 이용하여 개체 추출, 관계 추출, 이벤트 추출의 과정을 거쳐 기술 지식을 추출하게 된다. 이 과정에서 지명, 인명, 기관명 등의 개체가 사전을 참조하거나 규칙화된 패턴을 통해 추출된다. 이 때 추출된 용어가 동의어인 경우에 명확성을 규명하여 URI를 부여한 후, 온톨로지를 이용하여 개체와 관계를 매핑하여 RDF 트리플을 생성해낸다. 기술지식을 추출하는 대상인 문헌 데이터는 RDB에 저장되어 있으므로 RDB-to-트리플 매핑 규칙에 의해 변환되어 트리플 저장소에 저장된다. 이 두 과정에서 각 개체들의 URI는 통제되고 문헌에서 추출된 개체들과 문헌과의 연관 정보까지 포함하여 저장소에 구축된다.

다음 장에서는 문헌에서 텍스트 마이닝을 이용하여 추출된 개체 정보와 이들 개체 간의 의미적 연관 관계를 표현하기 위한 기술 지식 도메인 온톨로지를 설계하고 이를 바탕으로 기술 지식 정보 체계를 구축하는 과정에 대해서 서술하기로 한다.

3. 기술 지식 온톨로지 스키마 구축

온톨로지 구축을 위한 방법론으로는 목적, 분야에 맞게 다양하게 제시되어 왔다. 대표적인 방법론으로 목적 설정, 온톨로지 구축, 평가, 문서화의 4단계로 진행하는 Enterprise Methodology 방법론, Competency question을 정의한 다음 온톨로지를 통하여 질문에 대한 답을 할 수 있는지의 평가 방법을 통해 온톨로지를 구축하는 TOVE(TOronto Virtual Enterprise) 방법론, 소프트웨어 개발 생명 주기를 기반으로 온톨로지 생성 단계를 정의한 METHONTOLOGY 방법론, 지식 관리 프로세스와 지식 메타 프로세스 기반으로 세부 단계로 나뉘어서 진행하는 On-To-Knowledge 방법론, 클래스 정의 및 속성 정의 등으로 구체적인 세부 프로세스를 수행함으로써 온톨로지를 설계하는 Ontology Development 101 등이 있다[10-14].

본 연구의 기술 지식 온톨로지는 문헌의 내용에서 주요한 개체들을 추출하여 각 개체들의 의미를 인식하고, 이들 개체들 사이의 연관 관계를 정의하는 의미적 지식 네트워크 구축을 목적으로 하기에 보편적인 온톨로지 구축 방법인 Ontology Development 101을 토대로 기술 지식 온톨로지를 설계하기로 한다[14].

이 방법론은 온톨로지의 클래스와 그 관계들에 초점을 두고 의사 결정 시스템을 설계하는데 그 7단계를 다음과 같이 설명하고 있다.

1. 온톨로지의 도메인과 범위를 결정한다.
2. 기존 온톨로지의 재사용을 고려한다.
3. 온톨로지에 있어서 중요한 용어들을 열거한다.
4. 클래스 간의 계층을 정의한다.
5. 클래스의 속성들을 정의한다.
6. 속성을 정제한다.
7. 인스턴스를 생성한다.

3.1 도메인과 범위 결정

첫 번째 단계로 구축하고자 하는 온톨로지의 도메인과 범위를 결정해야 한다. 기술 지식을 추출하는 대상인 과학기술 문헌으로 논문, 특허, 웹 자원을 선정하고 [표 1], 이들 콘텐츠 유형의 자료가 가지는 서지 메타데이터들과 텍스트에서 추출해낸 기술 지식들을 온톨로지 구축 범위로 한다.

표 1. 기술 지식 추출 대상이 되는 과학기술 문헌 종류

도메인	수집처
논문[1]	해외 프로시딩, 저널
특허	국제공개, 미국공개, 미국등록, 유럽공개 특허
웹 자원	IDC Press Release, Wikipedia, InformationWeek, Gizmag, TechnologyReview, Ieee Spectrum, TechnewsWorld, DiscoverMagazine, NewYork Times, BBC, Fox News, CNN, Thomson Reuters, USA Today, EtnTws

웹에서 수집하는 자원은 뉴스(NewYork Times, BBC, Fox News, CNN, USA TODAY, EtnEws), 매거진(InformationWeek, Gizmag, TechnologyReview, Ieee Spectrum, TechnewsWorld, DiscoverMagazine), 보고서(IDC Press Release, Thomson Reuters), 사전(Wikipedia) 분야에서 과학, 비즈니스 섹션으로 한정하여 자원을 수집하도록 한다. 또한 정보 추출 과정에서 개체 추출의 정확성을 높이기 위해 외부에서 수집한 전거 데이터도 포함한다[15].

3.2 기존 온톨로지의 재사용

기존에 존재하는 온톨로지를 재사용하는 것을 고려한다는 원칙은 구축된 온톨로지의 메타데이터의 상호운용성을 높여주고 다른 온톨로지와 호환성을 고려할 때 필요하다. 기구축된 학술문헌 온톨로지는 논문과 특허를 대상으로 하여 기술용어를 추출하고, 논문, 특허의 서지 메타 데이터인 인명, 기관, 국가와 연관 관계를 매핑하였다[16]. 본 연구는 이 학술문헌 온톨로지를 기반으로 사용하되 웹 자원 도메인을 추가하고, 기존의 기술용어를 기술명 및 제품명으로 추출 대상 개체를 구분한 후, 이들 개체 사이의 의미 연관 관계를 세분화하여 추가 정의한다. 지명과 관련하여 도메인 온톨로지인 GeoName[2]을 재사용한다.

3.3 용어 열거

과학기술 문헌의 서지 메타데이터 요소들을 모두 열거하여 개념과 용어를 조직화하는 과정을 거친다. 이들 문헌에서 추출하는 개체들은 인명(Person), 위치(Location), 기관(Organization), 용어(Term: 기술명, 제품명), 시간(Time) 정보로 한정한다. 그리하여 기초 자료를 바탕으로 공통적으로 쓰이는 개념이나 특성, 개별적으로 쓰이는 개념이나 특성을 파악하고, 각각의 세부 영역 정보가 다른 세부 영역 정보와 어떠한 관계를 맺고 있는지를 열거해본다.

3.4 클래스 및 계층구조 정의

다음 단계로 온톨로지의 클래스와 클래스 간의 계층관계를 정의한다. 이전 단계에서 열거된 용어들 중에서 클래스를 선정하고 클래스 사이의 관계를 정의하는 단계이다. 기술 지식을 추출하는 대상이 되는 문헌에는 논문, 특허, 웹 자원으로 이들은 각각 Article, Patent, WebArticle 클래스로 정의하고, 이들을 Document 클래스의 하위 클래스로 개념 간의 계층적 구조를 설정한다. Article 클래스와 연관되어있는 출간 정보에 관한

1 논문과 특허 데이터는 국가과학기술정보센터(NDSL)로부터 수집하였음

2 http://www.geonames.org/ontology/

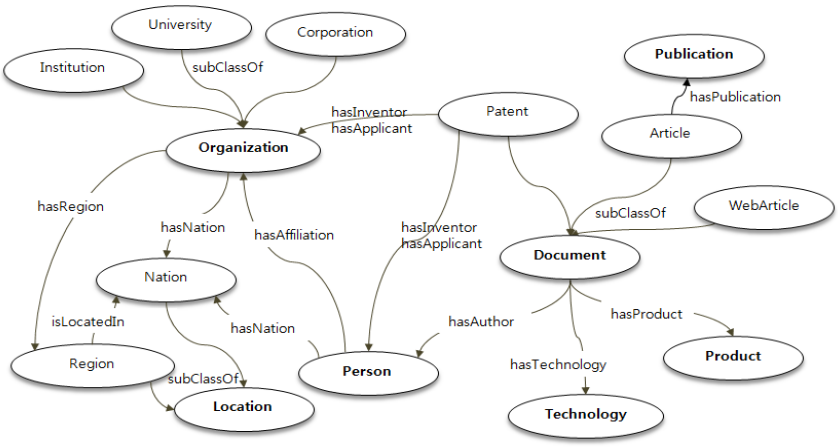


그림 2. 클래스와 클래스 간의 관계 정의

Publication 클래스를 정의한다. 그리고 문헌에서 정보 추출을 통해 얻어지는 개체인 기술명은 Technology 클래스로, 제품명은 Product 클래스로, 인명은 Person 클래스로 정의하고, 연구소, 회사, 대학은 Organization 클래스의 하위 클래스인 Institution, Corporation, University 클래스로 계층 구조를 설정한다. 이를 도식화하면 [그림 2]와 같다. 이 그림은 클래스 간의 계층구조를 확인할 수 있으며 클래스와 클래스 간의 기본적인 관계가 정의되어 있다. 예를 들어, 학술논문은 Publication과 hasPublication 관계를 가지고, 저자 정보에서 Person과 Organization과의 hasAffiliation 관계 및 Location을 서지 데이터로부터 추출이 가능하며, 초록에서 정보 추출 과정을 거쳐 Technology, Product, Person, Organization, Location 등의 기술 지식을 얻게 된다. 이 중에서 Technology와 Product는 추출된 문헌과의 관계(hasTechnology, hasProduct)를 매핑하여 해당 개체의 원문 출처 정보를 추적할 수 있게 하여, 정보 추출 시스템의 성능을 향상시킬 수 있는 피드백 역할을 수행한다.

3.5 클래스의 속성 정의

클래스와 클래스 간의 계층 구조의 정의와 함께 클래스의 속성을 생성, 수정, 삭제하는 단계이다. 속성의 종류에는 데이터타입 속성(DatatypeProperty), 객체속성(ObjectProperty), 주석 속성(Annotation Property)으로 구분한다. 객체 속성은 인스턴스와 인스턴스를 연결하기 위함이며, 데이터타입 속성은 인스턴스와 값, 주석 속성은 인스턴스의 값이 주석의 형태를 지니는 경우를 포현한다. 하나의 속성은 domain과 range의 트리플 관계를 구성하게 된다. domain은 속성이 사용될 대상이 되는 클래스 범주를 가리키며, range는 속성의 값을 가질 수 있는 범위를 한정한다. Article 클래스는 제목, 초록, 주제어, 발행 타입, 유형, 식별번호, 발행연도, 발행일자 등의 속성을 가지며, Patent 클래스는 제목, 요약, 자료구분, 출원연도, 출원일자, 출원번호, 등록일자, 등록번호, IPC 코드 등의 속성을 가진다. WebArticle 클래스는 제목, 본문, 분류, URL, 수집 사이트명, 카테고리, 섹션, 키워드, 발행 연도, 발행날짜, 수정 날짜의 속성을 데이터타입 속성으로 정의한다.

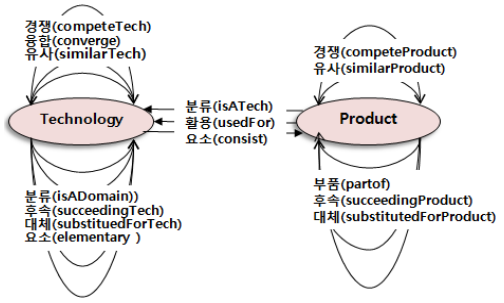


그림 3. Technology, Product 클래스 간에 정의된 객체 속성 정의

객체속성은 클래스에 포함되는 인스턴스 간(동일한 인스턴스나 서로 다른 인스턴스)의 관계를 표현하는 것으로 [그림 3]에서는 Technology 클래스와 Product 클래스 사이에 정의된 속성들을 정의한 것이다. 활용 관계는 [P-usedFor-T], 요소 관계는 [T-elementary-T]의 트리플로 나타낼 수 있다(P는 Product, T는 Technology의 약자). 경쟁 관계는 기술과 기술, 제품과 제품 사이에 모두 나타나지만, 객체 속성 이름을 구분이 가능하도록 명명하여 [T-competeTech-T]와 [P-competeProduct-P]로 표현한다. 이와 같이 동일한 관계이지만 domain과 range에 따라 변이가 있는 이름을 부여하게 되면 인스턴스 검색 시 domain과 range를 지정 하지 않고 질의어를 작성하여도 원하는 인스턴스 결과를 획득할 수 있다. 기술 간의 경쟁 관계나 제품 간의 경쟁 관계는 [A-competeTech-B]이면, [B-competeTech-A]도 만족하는 대칭성(Symmetry)을 갖기 때문에 Symmetric property를 추가로 부여한다. 기술 동향을 파악하고 통찰력을 얻기 위해 주요 요소인 기술, 제품, 주체(기관, 인명 등) 간 다양한 관계들을 38가지의 속성으로 정의하였다[17].

인스턴스 간의 속성 중 특허취득(patentTech), 출시(launch), 발표(announce) 등의 관계 정보는 해당 행위가 이루어진 시점의 정보를 획득하는 것이 기술 동향의 전/후를 추론하기 위해 필요하다. 이 경우에는 해당 트리플(S-P-O)의 Statement를 S로 하고, O는 date, P는 eventTime의 데이터타입 속성으로 표현하는 것이 가능하다. 실례로, "On 31 August 2010, HP announced collaboration with Hynix to bring memristor to high volume manufacturing step."는 HP와 Hynix가 협력관계가 되었다는 것을 알리는 문장인데, 그 시점이 2010년 8월 31일이라는 시간 정보가 중요하다. [HP-collaborate-Hynix]가 다시 S가 되고 [S-eventTime-2010/08/31]로 트리플로 표현할 수 있는 것을 [그림 4]를 통해 확인할 수 있다. 더 나아가 이 트리플 정보가 'TechNews'로부터 추출되었다는 컨텍스트 정보까지 객체 속성으로 정의함으로써 트리플을 이용한 SPOCT 구체화가 가능해진다(S:Subject, P:Predicate, O:Object, C:Context, T:Time)[18].

```
@prefix onto: <http://isrl.kisti.re.kr/ontology#> .

onto:Corporation_85 onto:collaborate onto:Corporation_308 .
onto:Triple_13646834 rdf:subject onto:Corporation_85 ;
    rdf:predicate onto:collaborate ;
    rdf:object onto:Corporation_308 ;
    onto:eventTime "20100831"^^xsd:date ;
    onto:context onto:TechNews .
```

그림 4. SPOCT를 표현한 트리플 데이터

3.6 속성 정제

온톨로지를 구성하고 모델링을 완성하기까지 지속적으로 정제 과정을 거치게 된다. 일차적으로 모델링 도구 내에서 트리플 관계 생성에서나 논리적 오류에 의한 속성 값들의 정제를 거친 후 인스턴스를 생성하고 질의, 추론을 해 나가는 과정에서도 정제 작업은 계속된다. 실제 기술 지식과 문헌 메타 데이터의 인스턴스 및 트리플의 생성 과정과 서비스를 위한 SPARQL 질의를 시행하는 과정에서도 지속적인 정제 과정을 진행하고, 추론 결과의 피드백을 통해서도 정제 과정은 계속된다.

3.7 인스턴스 생성

마지막 단계로는 문헌의 메타 데이터와 문헌으로부터 추출된 실제 기술 지식과 외부에서 수집한 전거 데이터(도시명과 국가명)를 바탕으로 매핑에 의한 변환으로 인스턴스를 일괄 생성한다. 지금까지 설계한 온톨로지를 이용하여 'Clouding Computing' 인스턴스와 연관된 트리플들을 도식화하면 [그림 5]와 같다. 'Clouding Computing'의 요소 기술로는 'Hadoop'이 있고, 'Hadoop'은 'Map Reduce'와 요소관계를 맺고 있다. 'Clouding Computing'에 속하는 제품으로는 'Amazon EC3', 이 제품을 판매하는 기업은 'Amazon'이다. 'Amazon EC3' 인스턴스는 WebArticle1에서 추출되었음을 hasProduct 객체 속성을 통해 알 수 있다. 이러한 인스턴스들은 고유한 URI(Uniform Resource Identifier)를 할당 받은 후 시맨틱 저장소에 저장된다.

이와 같은 단계를 거쳐서 생성된 온톨로지 스키마는 [표 2]와 같다. 도메인의 범위, 클래스의 정의, 관계 속성의 세부 정의 등의 단계를 거치면서 온톨로지 구성이 순조롭게 진행되었고, 온톨로지 편집 도구인 Protégé

4.0[3]을 활용하여 온톨로지 스키마를 제작하였다. 다음 장에서는 이 온톨로지를 기반으로 하여 실제 인스턴스를 생성하여 기술 지식 정보 체계를 구축하는 과정에 대해 기술하고자 한다.

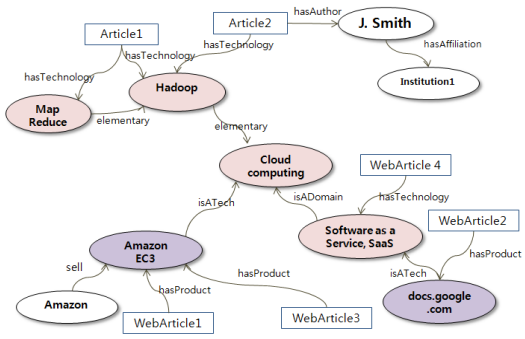


그림 5. 'Cloud Computing' 인스턴스와 연관된 트리플의 그래프 표현

표 2. 기술 지식 생성을 위한 온톨로지

```
....(중략)
<owl:ObjectProperty rdf:about="#competeTech">
    <rdf:type rdf:resource="&owl;SymmetricProperty"/>
        <rdfs:range rdf:resource="#Technology"/>
        <rdfs:domain rdf:resource="#Technology"/>
    </owl:ObjectProperty>

<owl:ObjectProperty rdf:about="#consist">
    <rdfs:domain rdf:resource="#Technology"/>
    <rdfs:range rdf:resource="#Product"/>
    <owl:inverseOf rdf:resource="#usedFor"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:about="#elementary">
    <rdf:type rdf:resource="&owl;TransitiveProperty"/>
        <rdfs:range rdf:resource="#Technology"/>
        <rdfs:domain rdf:resource="#Technology"/>
    </owl:ObjectProperty>

<owl:ObjectProperty rdf:about="#develop">
    <rdfs:domain rdf:resource="#Organization"/>
        <rdfs:range rdf:resource="#Technology"/>
        <rdfs:subPropertyOf rdf:resource="#own"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:about="#isADomain">
    <rdf:type rdf:resource="&owl;TransitiveProperty"/>
        <rdfs:range rdf:resource="#Technology"/>
        <rdfs:domain rdf:resource="#Technology"/>
    </owl:ObjectProperty>

<owl:ObjectProperty rdf:about="#isATech">
      <rdfs:domain rdf:resource="#Product"/>
      <rdfs:range rdf:resource="#Technology"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:about="#launch">
      <rdfs:domain rdf:resource="#Organization"/>
      <rdfs:range rdf:resource="#Technology"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:about="#succeedingTech">
      <rdfs:domain rdf:resource="#Technology"/>
      <rdfs:range rdf:resource="#Technology"/>
      <rdfs:subPropertyOf rdf:resource="#competeTech"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:about="#own">
      <rdfs:domain rdf:resource="#Organization"/>
      <rdfs:range rdf:resource="#Technology"/>
</owl:ObjectProperty>

<owl:ObjectProperty rdf:about="#hasTechnology">
      <rdfs:domain rdf:resource="#Document"/>
      <rdfs:range rdf:resource="#Technology"/>
</owl:ObjectProperty>

<owl:Class rdf:about="#Technology"/>
<owl:Class rdf:about="#Product"/>
<owl:Class rdf:about="#Organization"/>
<owl:Class rdf:about="#Corporation">
  <rdfs:subClassOf rdf:resource="#Organization"/>
</owl:Class>
<owl:Class rdf:about="#Institution">
        <rdfs:subClassOf rdf:resource="#Organization"/>
</owl:Class>
.....(중략)
```

4. 기술 지식 온톨로지 인스턴스 생성

기술 지식을 추출하는 대상으로 선정된 논문, 특허, 웹 자원을 수집한다. 이 수집된 문헌 자원의 현황은 [표 3]에서 확인할 수 있다.

표 3. 기술 지식 추출 대상으로 수집한 과학기술 문헌 자원 현황(2001년~2012년8월)

수집 분야	건수
웹 자원	5,261,949
논문	8,135,432
특허	7,444,046
총 합계	20,841,427

웹 자원은 제목과 본문, 논문과 특허 데이터는 제목과 초록에서 기술 개체를 인식하고 관계를 추출하는 과정을 거쳐 생성된다. 개체 인식 과정은 문장 분리, 언어 분석, 기술 개체 후보 인식, 사전기반 기술 개체 인식, 통계기반 기술 개체 인식 등의 총 5단계를 거친다. 관계 추출의 단계는 전처리 단계, 구문 분석 단계, 술어-논항

3 http://protege.stanford.edu

구조 변환 단계, 관계 추출의 4단계로 나뉘어서 진행된다[19].

표 4. 인스턴스 생성 현황

Class	Instance 건수
Technology	316,143
Product	40,976
Person	6,746,620
Institution	11,884
University	15,076
Corporation	398,472
Patent	7,444,046
Article	8,135,432
Web Article	5,261,949
Publication	16,906
Nation	245
Region	1,194,925

이러한 추출 과정을 통해서 생성된 기술 지식의 인스턴스는 [표 4]에서 보듯이 Technology 316,143건, Product 40,976건으로 정보 추출 과정을 통해 생성된 것이고, Nation, Region 인스턴스는 GeoName에서 수집한 전거 데이터로 정보 추출 단계에서 위치명 인식을 위한 사전으로 사용되었다. 수집된 문헌 자원은 중복되지 않았기에 수집 건수와 생성된 인스턴스의 생성 건수가 동일하다. 이러한 인스턴스들을 RDF 트리플로 표현한 총 수는 425,252,384건이다.

[표 5]에서는 정보 추출 과정을 통해 생성되는 기관, 기술, 제품 인스턴스 간의 주요 객체 속성과 실제로 추출된 트리플의 건수를 표기하고 있다. 기술의 동향 파악을 위한 관계 추출 어휘 중에서 경쟁, 동종 관계에 비해 기술 발달의 전후를 분석할 수 있는 근거가 되는 후속, 대체 관계는 비교적 적은 수가 추출되었음을 알 수 있다.

이렇게 생성된 트리플 데이터는 시맨틱 지식 저장소에 적재되는데 저장소로는 Native 기반의 시맨틱 서비스 프레임워크인 SEMON을 사용하였다[20]. 트리플 저장소와 추론 엔진이 통합된 시스템인 SEMON에 적재한 트리플 데이터를 검색하기 위해서는 SPARQL 질의어를 사용한다[21]. [표 6]에서 연관 정보를 가져오는 SPARQL 질의어 타입 3가지를 실례로 제시하고 있다. (a)에서는 특정 기술과 관련된 요소 기술 목록을 얻어오는 질의어이며 (b)는 두 기업이 공통적으로 보유하고 있는 기술 목록을 가져오는 것이고 (c)는 두 기업이 보유하고 있는 기술들 중에서 서로 경쟁 관계를 가지는 기술 목록을 찾는 질의어이다.

표 5. 주요 인스턴스들의 연관 관계를 나타내는 객체 속성의 종류와 실제 추출된 트리플 건수 (T:Technology, P:Product, O:Organization의 약자)

ObjectProperty		Domain	Range	특성	건수	ObjectProperty		Domain	Range	특성	건수
분류	isADomain	T	T		67,033	출시	launch	O	T		7,204
경쟁	competeTech	T	T	Symm	19,802	특허취득	patentTech	O	T		10
후속	succeedingTech	T	T	Symm	129	소유	ownProduct	O	P		319,798
대체	substitutedForTech	T	T		876	사용	useProduct	O	P		8,049
요소	elementary	T	T		95,554	판매	sell	O	P		4,770
동종	similarTech	T	T	Symm	30,496	발표	announce	O	P		10,318
분류	isATech	P	T		374,048	생산	produce	O	P		320,575
요소	consistTech	P	T		78,710	경쟁	competeOrg	O	O	Symm	40,259
요소	consistProduct	T	T		8,259	동종	isSimilarOrg	O	O	Symm	40,259
부품	partOf	P	P		37,387	협력	collaborate	O	O	Symm	11,082
경쟁	competeProduct	P	P	Symm	72,294	법적다툼	competeByLaw	O	O	Symm	1,776
동종	similarProduct	P	P	Symm	72,294	법적소송	sue	O	O		930
후속	succeedingProduct	P	P		121	투자	invest	O	O		451
대체	substitutedForProduct	P	P		31	인수	takeover	O	O		10,016
보유	own	O	T		241,932	고객	hasCustomer	O	O		7,113
사용	useTech	O	T		19,481	설립	found	O	O		295
개발	develop	O	T		3,553	보완	supplement	O	O	Symm	13,832

표 6. SPARQL 질의어 예제

(a)

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX onto: <http://isrl.kisti.re.kr/ontology#>

SELECT ?xname
WHERE {
   onto:Technology_46 onto:elementary ?x .
   ?x onto:prefLabel ?xname .
}
```

(b)

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX onto: <http://isrl.kisti.re.kr/ontology#>

SELECT  ?z ?zname
   onto:Corporation_1  onto:own ?z .
   onto:Corporation_2  onto:own ?z .
   ?z  onto:prefLabel ?zname .
}
```

(c)

```
PREFIX rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX onto: <http://isrl.kisti.re.kr/ontology#>

SELECT ?x ?xname ?y ?yname
WHERE {
          onto:Corporation_164 onto:own ?x .
          onto:Corporation_181 onto:own ?y .
          ?x onto:competeTech ?y .
          ?x onto:prefLabel ?xname .
     ?y onto:prefLabel ?yname
}
```

5. 결 론

본 논문을 통해 과학기술 문헌에서 텍스트 마이닝 기술을 이용하여 개체 및 관계를 파악하여 자동으로 추출된 기술 지식을 시맨틱 웹 기술에 적용하기 위해 적합한 온톨로지를 설계하는 과정에 대해 살펴보고, 실제 인스턴스, 트리플 데이터를 시맨틱 저장소에 적재하여 정보 시스템을 구축하는 과정에 대해 설명하였다. 기술, 제품, 연구주체 간 다양한 관계들을 객체 속성으로 정의하여 온톨로지로 설계하는 것은 인과 관계 등의 의미 연관 관계 분석을 가능하게 지원해준다.

이와 같이 온톨로지를 기반으로 하여 기술 지식 정보 체계를 구축하게 되면 이들 지식 간의 상호참조적 네비게이션이 가능하게 될 뿐 아니라, 이러한 기술 지식 간의 네트워크 관계에 의해 추가적인 지식 간의 관계를 추론할 수 있다. 향후에는 온톨로지로 생성된 기술 지식들을 링크드 데이터(Linked Data)로 공개하고 기존의 링크드 데이터와의 상호 연계도 지원하여 기술 지식의 재사용과 공유를 지원하기 위한 연구가 필요할 것이다.

또한, 본 기술 지식 정보 체계를 기반으로 하여 R&D 전략을 수립하고자 하는 사용자에게 기술, 연구주체와 연구 성과의 다양한 조합을 통해 통찰력을 제공하기 위한 서비스 간 연계, 융합 및 예측기반 분석 서비스인 InSciTe Adaptive를 제공할 예정이다.

참 고 문 헌

[1] M. Blume, “Automatic entity disambiguation: Benefits to NER, relation extraction, link analysis, and inference,” International Conference on Intelligence Analysis, 2005.

[2] http://en.wikipedia.org/wiki/Data#Meaning_of_data.2C_information_and_knowledge

[3] 정한민, 김진형, 정도헌, 조민희, 송사광, 이승우, 이상환, “사용자 적응적 가이드 방식의 R&D 기획 시스템에 대하여”, 컴퓨터종합학술대회, pp.411-413, 2012.

[4] T. Berners-Lee, J. Hendler, and O. Lassila, “The Semantic Web: A new form of Web content that is meaningful to computers will unleash a revolution of new possibilities,” Scientific American, May 2001.

[5] http://www.w3.org/RDF/

[6] 최호섭, 임지희, 배영준, 최수일, 옥철영, “온톨로지 구축 방법과 사례”, 정보과학회지, 제24권, 제4호, pp.31-44, 2006.

[7] 한국정보문화진흥원, *국가지식정보 온톨로지 표준 개발*, 2006.

[8] 민병원, 오용선, “U-Health 개인 맞춤형 질병예측 기법의 개선”, 한국콘텐츠학회 논문지, 제10권, 제10호, pp.54-67, 2010.

[9] C. Blakeley, “RDF Views of SQL Data(Declarative SQL Schema to RDF Mapping),” OpenLink Software, 2007.

[10] M. Uschold and M. King, “Towards A Methodology for Building Ontologies,” IJCAI-95 Workshop on Basic Ontological Issues in Knowledge Sharing, 1995.

[11] M. Gruninger and M. S. Fox, “Methodology for the Design and Evaluation of ontologies,” IJCAI-95 Workshop on Basic ontological Issues in Knowledge Sharing, 1995.

[12] M. F. Lopez, A. Gomez-Perez, and J. P. Sierra, “Building a Chemical Ontology Using Methontology and the Ontology Design Environment,” IEEE Intelligent Systems, Vol.14, No.1, 1999.

[13] S. Staab, H. Schnurr, R. Studer, and Y. Sure, “Knowledge processes and ontologies,” IEEE Intelligent Systems, Special Issue on Knowledges —anagement, Vol.16, No.1, 2001.

[14] N. F. Noy and D. L. McGuinness, “Ontology Development 101: A Guide to Creating Your First Ontology,” Stanford Knowledge Systems Laboratory Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880, 2001.

[15] 황미녕, 김태홍, 최성필, 조민희, 홍순찬, 정한민, “DBpedia를 이용한 공개 정보 수집 방법”, 2012년도 한국인터넷정보학회 하계학술발표대회 논문집, 제13권, 제1호, pp.75-76. 2012.

[16] http://www.ontoframe.kr/sw/UseCases/InSciTe.html

[17] 조민희, 이승우, 송사광, 이진희, 구희관, 홍순찬, 정한민, “R&D 기획 지원을 위한 개체-관계 모델링”, 2012년도 한국인터넷정보학회 하계학술발표대회 논문집, 제13권, 제1호, pp.137-138, 2012.

[18] M. N. Hwang, D. M. Seo, S. W. Lee, M. H. Cho, S. K. Song, J. H Lee, S. C. Hong, S. P. Choi, and H. M Jung, “Ontology Model of Technical Knowledge for Analytics,” International Conference on Smart Media and Applications, pp.66-67, 2012.

[19] 최성필, 최윤수, 전홍우, 정창후, 송사광, 정한민, “SINDI-WALKS: 과학기술지식발견 워크벤치”, 한국정보과학회 2012 한국컴퓨터종합학술대회 논문집, 제39권, 제1호, pp.279-281, 2012.

[20] http://www.diquest.com

[21] http://www.w3.org/TR/rdf-sparql-query/

저 자 소 개

황 미 녕(Mi-Nyeong Hwang) 정회원

- 2000년 : 부산대학교 전자계산학과(이학사)
- 2002년 : 부산대학교 전자계산학과(석사)
- 2002년 ~ 현재 : 한국과학기술정보연구원 소프트웨어연구실 선임연구원

<관심분야> : 시맨틱 웹, 온톨로지, 데이터마이닝

이 승 우(Seungwoo Lee) 정회원

- 1999년 : 포스텍 컴퓨터공학과(석사)
- 2005년 : 포스텍 컴퓨퍼공학과(박사)
- 2006년 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 자연어처리, 정보추출, 시맨틱 웹, 정보분석, 빅데이터

조 민 희(Minhee Cho) 정회원

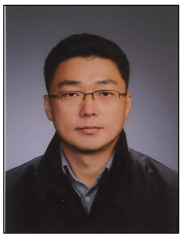
- 2003년 : 연세대학교 전산학과(이학사)
- 2005년 : 연세대학교 전산학과(석사)
- 2005년 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 자연어처리, 텍스트마이닝

김 순 영(Soon Young Kim) 정회원

- 1992년 2월 : 충남대 문헌정보과
- 1996년 6월 ~ 2005년 12월 : KAIST 도서관
- 2006년 1월 ~ 현재 : 한국과학기술정보연구원 해외정보실 선임연구원

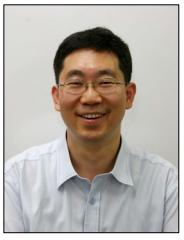
최 성 필(Sung-Pil Choi) 정회원



- 1998년 : 부산대학교 전자계산학과(석사)
- 2012년 : 한국과학기술원 정보통신공학과(박사)
- 1998년 ~ 현재 : 한국과학기술정보연구원 선임연구원

<관심분야> : 기계학습, 정보검색, 자연어처리, 정보추출, 텍스트마이닝

정 한 민(Hanmin Jung) 정회원



- 2003년 : 포항공과대학교 컴퓨터공학과(박사)
- 2004년 ~ 현재 : 한국과학기술정보연구원 책임연구원/소프트웨어연구실 실장
- 2004년 ~ 현재 : 과학기술연합대학원대학교 겸임교수

<관심분야> : 시맨틱 웹, 정보검색, 자연어처리, HCI