한국산학기술학회논문지
Vol. 10, No. 6, pp. 1220-1228, 2009

검색 키워드 확장을 이용한 온톨로지 자동 생성 시스템 개발

심준[1], 이홍철[1*]
[1]고려대학교 정보경영공학전문대학원 정보경영공학과

The Development of Automatic Ontology Generation System Using Extended Search Keywords

Joon Shim[1] and Hong-Chul Lee[1*]
[1]Department of Information Management Engineering, Korea University

요 약 시맨틱 웹의 핵심인 온톨로지는 주로 특정 도메인에 한정되거나 휴리스틱에 의존해 의미와 관계를 정의하여 생성하고 있다. 하지만 온톨로지의 생성은 매우 어려울 뿐만 아니라 많은 시간이 소요되는 작업이다. 특정 분야에서 사용되는 온톨로지와 달리 웹에서 사용되는 온톨로지는 지식 및 정보 표현의 범위가 한정적이지 않기 때문에 기존의 온톨로지 생성 방식으로는 정보를 표현하기가 어렵다. 따라서 온톨로지의 자동 생성은 시맨틱 웹의 구현에 있어서 매우 중요한 부분을 차지하게 된다. 본 논문에서는 웹 온톨로지를 자동으로 생성하기 위해서 형태소 분석을 이용하여 검색엔진에서 사용자들이 입력하는 검색 키워드로부터 색인어를 추출하고, 이와 관련된 키워드를 확장시켜 온톨로지를 생성하고 갱신하는 방법에 대하여 제안한다.

Abstract Ontologies, which are the core of the Semantic Web, are usually limited by specific domains or created by defining meanings and relationships that depend on the heuristic. However, the creation of an ontology is not only very difficult but also very time-consuming. In contrast with ontologies that are used in specific fields, an ontology for the Web entails an unlimited scope of knowledge and expression of information. Hence, it is hard to express information in the same way that is used to create ontologies in specific fields. Therefore, the automatic generation of an ontology takes very important role in the Semantic Web. In this paper, to make ontologies automatically, we suggest the methods to create and renew ontologies by expanding keywords related to the index-terms which are extracted from the search keywords which users input in the search engines by analyzing the morphemes.

Key Words : Semantic Web, Ontology creation, Ontology, Formal Concept Analysis (FCA)

1. 서론

월드 와이드 웹(WWW)은 수많은 사용자와 데이터를 만들어 냈고, 사람들이 정보를 얻고 처리하는 방식에 큰 변화를 일으켰다. 하지만 오늘날의 웹에 존재하는 데이터 구조는 인간위주로 구성되어 있다. 따라서 검색엔진은 사용자가 입력하는 검색 키워드의 의미를 이해하지 못하고, 단순히 키워드 일치 여부에 따른 결과를 반환하는 문제점을 갖고 있다. 이는 컴퓨터가 이해하고 처리할 수 있는 구조가 아닌 인간만 이해할 수 있는 표현위주의 언어(HTML)로 이루어져 있기 때문이다. 따라서 인간뿐만 아니라 컴퓨터가 이해하고 처리할 수 있도록 규격화 된 형식으로 데이터를 표현할 필요성이 있다.

WWW를 제안했던 Tim Berners-Lee는 표현 위주의 HTML의 문제점을 해결하기 위해서 데이터에 규격화된 의미를 부여하여 인간뿐만 아니라 컴퓨터도 데이터의 의미를 해석할 수 있도록 시맨틱 웹을 제안하였다. 시맨틱 웹은 컴퓨터가 실세계의 사물의 개념과 관계를 이해할

이 연구에 참여한 연구자(의 일부)는 '2단계BK21사업'의 지원비를 받았음.
*교신저자 : 이홍철(hclee@korea.ac.kr)
접수일 09년 05월 26일 수정일 09년 06월 09일 게재확정일 09년 06월 17일

수 있도록 데이터에 의미를 부여하여 사용자가 원하는 결과에 가까운 검색결과 제안과 자동화된 웹 환경을 만드는데 목표가 있다. 이를 실현하기 위하여 사물에 대한 개념을 정의하고 관계를 부여해주는 기술로 온톨로지가 있다.

온톨로지는 "개념화 된 것을 명시적으로 형식을 갖춰 구체화 한 것 (An ontology is an explicit specification of a conceptualization)"[1] 으로 정의되며, 시맨틱 웹의 실현에 있어서 핵심적 역할을 한다고 할 수 있다. Tim Berners-Lee는 AAAI 발표에서 "시맨틱 웹은 인공지능이 아니고, 인공지능은 시맨틱 웹이 아니며, 인공지능은 연구 분야고, 시맨틱 웹은 프로젝트이다. 인공지능은 시맨틱 웹에 많은 기여를 하였고, 마찬가지로 시맨틱 웹은 인공지능의 활동의 장이 될 수 있다."[2] 라는 발표를 하였다. 이것은 시맨틱과 웹의 개념을 분리하여 생긴 오해를 바로잡기 위한 것으로 시맨틱 웹은 기존의 웹을 발전시킨 환경일 뿐, 인공지능을 기반으로 한 웹 환경이 아니라는 것을 의미한다.

온톨로지 역시 인공지능, 소프트웨어 공학, 의료정보 분야 등에서 연구하는 개념과 생성방식은 웹을 위한 온톨로지(OWL, Web Ontology Language)와 많은 차이를 보인다. 현재의 온톨로지 생성 방법은 대부분 특정 도메인의 전문가가 휴리스틱에 의존하여 생성하게 되는데, 이는 해당분야의 다양하고 자세한 어휘 구현이 가능하며 정형적 의미에 충실한 온톨로지 생성이 가능하다. 하지만 휴리스틱에 의존하는 온톨로지 구축은 매우 어려운 작업일 뿐만 아니라 많은 시간을 소비하게 된다. 따라서 온톨로지를 자동으로 생성하는 것은 시맨틱 웹의 실현에 있어서 매우 중요한 역할을 하게 된다.

본 논문에서는 검색엔진에서 사용자들이 입력하는 검색 키워드를 확장시켜 웹에 존재하는 데이터로부터 웹 온톨로지로 자동 생성시키는 방법에 대하여 제안하고자 한다.

본 논문의 구성은 다음과 같다. 2장에서는 시맨틱 웹과 온톨로지 관련 연구 분야를 기술하고, 3장에서는 온톨로지를 자동으로 생성하기 위한 기반 기술들에 대하여 기술한다. 4장에서는 제안한 시스템의 설계와 구현에 대하여 기술하고, 마지막으로 5장에서는 결론 및 향후 연구 방향에 대하여 제시한다.

2. 관련 연구

2.1 시맨틱 웹과 온톨로지

시맨틱 웹은 웹에 존재하는 다양한 서비스와 데이터들이 상호운용 되기 위한, 그리고 인간과 컴퓨터가 서로 데이터를 명확히 정의하고 교환할 수 있는 공통 언어 및 아키텍처의 표준이라 할 수 있다. 시맨틱 웹은 모든 정보를 Triple 구조로 표현하게 되는데, 이는 <Subject, Predicate, Object>의 형태로 개념을 표현한다. Triple 구조는 RDF (Resource Description Framework) 언어를 기반으로 고유의 URI (Uniform Resource Identifier)를 갖으며, 각각의 namespace를 가질 수 있다. 시맨틱 웹의 모든 데이터는 이러한 Triple 구조를 갖고 있으며, 그래프 형태로 의미정보인 온톨로지를 표현한다.

온톨로지는 지식을 개념화 하고 이를 명세화 하는 것으로 정의되는데, 어휘 사전의 역할 이외에도 지식을 효과적으로 표현하기 위해서 정보에 의미를 부여하고, 정보간의 관계를 정의할 수 있다. 온톨로지는 웹 분야뿐만 아니라 자연어처리, 데이터베이스, 인공지능 등 다양한 분야에서 활발하게 연구가 이루어지고 있다. W3C (World Wide Web Consortium)에서는 OWL을 표준으로 권고하고 있으며, 이는 XML, RDF, RDF Schema등의 문법을 기반으로 객체와 객체간의 관계와 계층을 형식적인 방법으로 설명하고 의미를 표현하기 위한 문법을 가지고 있다. 온톨로지에서 지식의 표현은 Class, Relation, Function, Axiom, Instance 등의 요소를 이용해 형식화하여 표현하며, 이러한 관계를 기준으로 새로운 정보를 추론해낼 수 있다[3].

시맨틱 웹에서 온톨로지는 사물의 개념을 명확하게 정의하고 있기 때문에 개념의 모호성을 줄일 수 있으며, 기정의 된 온톨로지 정보를 중심으로 해당 자원과 유사한 자원 또는 관련도가 있는 자원의 제안이 가능하므로 질의에 대한 확장된 결과를 얻을 수 있다. 이는 시소러스(thesaurus)의 개념과 유사하지만 자원과 자원 사이의 계층관계 및 제약조건 등이 부여되었다는 점에서 차이를 보인다.

특정 분야에서만 사용되는 온톨로지와 달리 웹에서 사용되는 온톨로지는 지식 및 정보 표현의 범위가 한정적이지 않다는 문제가 발생한다. 따라서 웹을 위한 온톨로지를 구축하기 위해서는 현재의 웹에 존재하는 데이터를 기준으로 정보에 의미를 부여하고 관계를 정의하면서 확장시켜 나가야 한다.

2.2 온톨로지 관련 연구

온톨로지 생성 방법은 크게 두 가지로 나눌 수 있다. 첫 번째는 해당 도메인의 정형화된 온톨로지를 생성하여 적용시키는 방법이고, 두 번째는 간단한 정보만을 기술한 단순한 온톨로지로부터 메타데이터간의 의미관계를 조금씩 부여해 확장시키는 방법이다. 기존의 온톨로지 구축에

관한 연구의 상당 부분은 해당 도메인의 전문가들에 의하여 휴리스틱에 의존하는 수작업이 대부분을 차지하고 있다. 휴리스틱에 의한 온톨로지 구축은 정확하고 체계적인 온톨로지 구축이 가능하다는 장점이 있지만, 모든 사물에 대한 정보를 온톨로지로 구축하는 것은 현실적으로 불가능에 가까울 만큼 어렵고 많은 시간을 소요하게 된다.

휴리스틱에 의존하는 온톨로지 생성방법의 단점을 보완하기 위하여 온톨로지를 자동으로 생성하기 위한 연구가 많이 진행되고 있다. TextOntoEx는 Semantic Pattern을 기반으로 자연어 상태의 영문을 언어학적 분석에 의하여 자동으로 온톨로지를 생성하는 방법을 제안하였고[4], P. Clerkin 등은 개념 계층(Concept Hierarchies)을 이용한 온톨로지 생성방법에 대하여 제시하였다[5]. 도메인의 문서들로부터 기계적 학습과 통계적 방법에 의하여 ONTOSTRUCT를 통하여 온톨로지를 자동으로 생성하는 방법도 제안되었다[6].

이외에도 Decision Tree, Association Rules, Classification 등의 데이터 마이닝 기법을 통한 데이터베이스로부터의 도메인 온톨로지 생성 방법에 대한 연구가 많이 이루어지고 있으며, 단어의 개념과 계층적 관계를 표현해주는 시소러스(Thesaurus) 및 WordNet을 이용하여 언어학적 분석과 분류학적 분석에 의한 온톨로지를 구축하는 연구도 이루어지고 있다. 또한 지식이나 자료를 모델링하기 위한 자료 분석의 이론인 형식적 개념 분석(FCA, Formal Concept Analysis)을 이용하여 계층구조의 온톨로지 생성에 관한 연구도 진행되고 있다.

하지만 위에서 언급한 연구들은 대부분 도메인에 의존적이기 때문에 한정되지 않은 범위의 데이터를 다루는 웹에 적용시킬 온톨로지 생성 방법으로는 부족한 부분이 있다.

본 논문과 유사한 연구로 FCA와 개념간의 관련도 계산을 이용한 분류기법을 적용시킨 온톨로지 생성방법과[7] 온톨로지를 이용하여 웹 문서로부터 자연어처리(Natural Language Processing)를 통하여 지식을 추출하는 기법[8]등이 있다. 위의 연구들은 자연어처리를 통하여 주요 키워드를 추출해내고 WordNet, GATE[9]등의 사전 기반의 의미부여 기법을 사용하는 것에 있어서 유사점을 갖지만, 본 논문에서 제시하는 웹에 존재하는 자료를 기준으로 온톨로지를 생성하고 확장시켜나가는 방법에 있어서 차이를 보인다.

최근에는 미리 정의해둔 어휘를 사용하여 인간뿐만 아니라 컴퓨터도 처리가 가능할 수 있도록 하는 방향의 연구가 활발하게 진행되고 있다. 많이 사용되고 있는 메타데이터와 Semantic annotation 기술로는 Dublin Core[10], FOAF (Friend Of A Friend) [11], Microformats[12], RDFa[13], SIOC (Semantically-Interlinked Online Communities) [14], SKOS (Simple Knowledge Organization System)[15] 등이 있으며, 이러한 어휘들은 대부분 RDF를 기반으로 정의되어 있어서 RDF Application으로 표현되기도 한다. 이는 상당히 구체적이고 추상적인 개념을 미리 정의해 놓았기 때문에 온톨로지를 구축할 때 필요에 의해 사용하거나 기존의 HTML 문서에 추가하여 쉽게 데이터에 의미를 부여할 수 있다.

이러한 메타데이터를 사용하는 이유는 독자적으로 온톨로지를 정의하는 것은 매우 어렵지만, 해당 어휘들은 이미 표준으로 합의된 사항이므로 조금 더 쉽게 온톨로지를 구축하거나 데이터에 의미를 부여하고 공유할 수 있기 때문이다. 위에서 언급한 메타데이터의 특성은 표 1과 같다.

본 논문에서는 웹 온톨로지를 생성하기 위하여 메타검색엔진을 구축하고, 검색이 이루어질 때 마다 검색 키워드와 관련이 있는 키워드를 확장시킨다. 그리고 확장된 키워드에 의미를 부여하고 관계를 정의하여서 온톨로지를 생성하는 방법을 소개하고자 한다. 웹에서 사용될 온톨로지는 범위가 한정적이지 않고 데이터의 양이 매우 많기 때문에, 간단한 정보만을 기술한 단순한 온톨로지를 생성하고 메타데이터간의 의미관계를 조금씩 부여해 확장시키는 방법을 이용해 접근하였다.

[표 1] 메타데이터의 종류와 특성

Dublin Core	컨텐츠에 대한 저작 정보에 대한 기술
FOAF	자신과 주변사람에 대한 정보와 관계를 기술
Microformats	HTML 문서에 정의된 태그를 사용해 특정 정보를 메타 데이터 형태로 가공
RDFa	HTML 문서에 RDF를 삽입하기 위한 기술
SIOC	온라인 커뮤니티의 연결을 목표로 하는 프레임워크
SKOS	이미 구축된 시스템내의 컨셉트를 표현, 연결, 조합하기 위한 어휘

3. 온톨로지 자동 생성 기반 기술

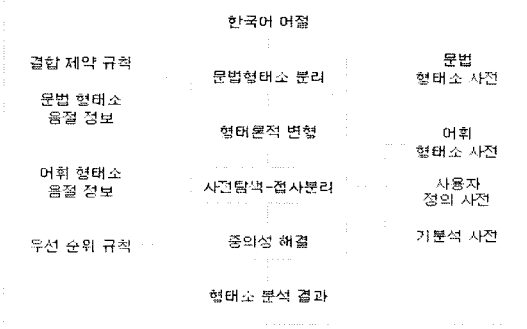
이번 장에서는 웹 온톨로지를 자동으로 생성하기 위하여 형태소 분석과 이를 통하여 추출된 색인어의 의미부여 및 FCA를 이용한 계층 구조 생성에 대하여 기술하도록 한다.

3.1 형태소 분석

영어의 경우 단어와 단어 사이의 경계를 공백으로 구

분할 수 있지만, 한국어의 경우 어절 단위로 띄어쓰기 때문에 단어 분할이 필요치는 않으나 붙여 쓴 복합명사를 단위명사들로 구분해야 하는 문제가 발생한다.

검색엔진에서 사용자들이 입력하는 키워드는 주로 검색하고자 하는 주요 키워드들로 이루어지거나 자연어 상태로 입력된다. 따라서 검색 키워드의 분석을 통하여 주요 키워드를 추출하고, 이를 기준으로 온톨로지에 Class를 생성 하게 된다. 형태소 분석을 위하여 KLT 2.1.0f [16] 가 사용되었으며 형태소 분석기의 구조는 그림 1과 같다.



[그림 1] 형태소 분석기 KLT 의 구조

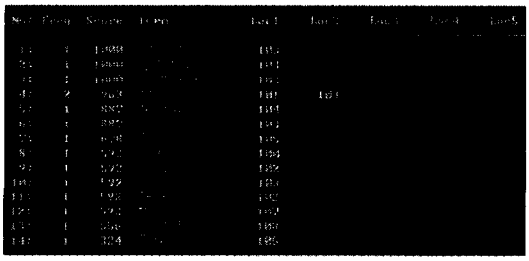
n 음절어에 공백을 i 개 삽입하는 경우의 수는 ${}_{n-1}C_i$ 이므로 최대로 생성될 수 있는 후보의 수는

$$\sum_{i=0}^{n-1} {}_{n-1}C_i = 2^{n-1}$$

가지이다. n 음절의 입력된 검색 키워드의 분리 가능한 형태소의 개수를 계산해보면, 길이가 i 음절인 형태소가 $(n-i+1)$ 가지이므로 최대

$$\sum_{i=1}^{n}(n-i+1) = \frac{n(n+1)}{2}$$

개의 형태소가 분리될 수 있다. 하지만 모든 분해 후보를 생성하는 것은 비효율적이므로 경험적으로 습득된 규칙을 적용하여 후보의 수를 줄여서 사용하게 된다[17].



[그림 2] 형태소 분석 결과

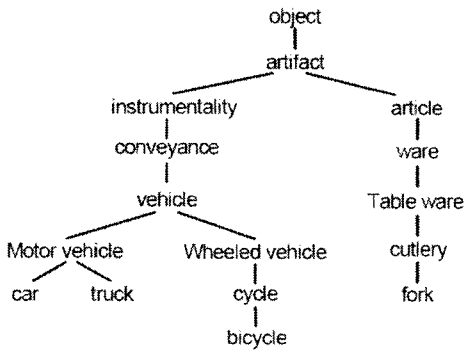
예를 들어, "고용보험및산업재해보상보험의보험료징수등에관한법률시행령" 이라는 띄어쓰기가 되어있지 않은 문장에 대하여 형태소 분석을 통하여 얻을 수 있는 색인어는 그림 2와 같다. 물론 실제 입력되는 검색 키워드의 경우 검색하고자 하는 키워드의 집합으로 이루어지는 경우가 대부분이므로 위의 예제보다 훨씬 단순한 구조를 갖게 된다. 이렇게 추출된 색인어를 기준으로 3.2절에서 다루게 되는 색인어에 의미를 부여하는 작업을 시행하게 된다.

3.2 추출된 색인어의 의미 분석

3.1 절의 형태소 분석을 통하여 추출된 색인어는 웹에 존재하는 데이터를 이용하여 정보를 재가공하고 이를 기준으로 온톨로지를 생성하게 된다. 추출된 색인어가 한국어인 경우 Naver, Daum, Google에서 제공하는 국어사전 및 백과사전 OpenAPI를 이용하여 단어의 품사 및 의미를 추출하고, 추출된 색인어가 영문인 경우 WordNet을 이용하여 의미와 동의어(synonymous term), 상위어(broad term), 하위어(narrow term), 관련어(related term) 등을 정의한다. 추출된 색인어에 의미를 부여하기 위하여 OpenAPI 또는 WordNet을 이용하는 것은 사전에 이미 정의되어 있는 정보를 이용하는 것이므로, 수작업으로 색인어에 단어의 의미를 부여하는 것보다 객관적이며 정확하다고 할 수 있다.

검색 키워드로부터 추출된 색인어 외에 동의어, 상위어, 하위어, 관련어 등을 함께 정의하는 이유는 사용자가 잘못된 단어를 입력하거나 그와 유사한 정보의 제안에 목적이 있다.

관련어는 추출된 색인어와 관련된 키워드를 OpenAPI를 이용하여 받아오게 되며, 이는 사용자들의 검색 패턴에 의한 유사한 키워드들을 제시하므로 보다 넓게 색인어의 확장을 가능하게 한다.



[그림 3] WordNet의 데이터 트리 예제

3.3 FCA를 이용한 계층구조 생성

FCA는 특정 도메인의 지식이나 데이터를 모델링하기 위한 방법론으로 자료 집합 사이의 개념적 구조를 조직화하기 위하여 수학적 사고로 접근한 자료 분석의 한 이론이다. FCA는 배경도(Formal Context), 개념(Formal Concept), 개념격자(Concept Lattice) 세 개의 기본 구성요소로 이루어진다[18].

FCA의 가장 기본적 자료구조인 배경도(Formal Context)는 문장 내에서 객체와 속성을 추출해낸 결과의 집합을 이야기한다. Formal Context K는 K=(G, M, I)로 정의되며, 객체(주어)들의 집합 G와 속성(서술어)들의 집합 M, 그리고 G와 M사이의 이항관계 I⊆G×M 로 구성된다. 이항관계 I는 "G의 원소 g는 M의 원소 m을 갖는다"는 것을 나타낸다. 위의 정의를 이용하여 두 집합 A', B'를 아래와 같이 정의하였다.

$$A' := \{m \in M \mid (g, m) \in I \text{ for all } g \in A\}$$
$$B' := \{g \in G \mid (g, m) \in I \text{ for all } m \in B\}$$

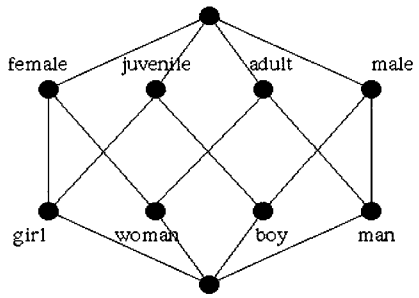
이 때 집합 A는 A⊆G이고, 집합 B는 B⊆M이다. 여기서 $A=B'$ and $B=A'$를 만족시킬 때, A는 extent라고 부르고 B는 intent라고 부른다.

위와 같은 정의를 바탕으로 문장으로부터 다양한 개념 (g, m)을 추출할 수 있다. 예를 들어, 그림 4에 나타난 Formal Context K=(G, M, I)는 객체집합 G={girl, woman, boy, man}와 속성집합 M={female, juvenile, adult, male} 그리고 관계집합 I={(girl, female), (girl, juvenile), (woman, female), (woman, adult), (boy, juvenile), (boy, male), (man, adult), (man, male)}로 표현할 수 있다. 이러한 개념들 사이에는 일종의 상-하위 관계에 따른 순서가 존재한다.

예를 들어, 임의의 개념 (X_1,Y_1)과 (X_2,Y_2)에 대하여 $(X_1,Y_1) \leq (X_2,Y_2) \Leftrightarrow X_1 \subseteq X_2$ ($\Leftrightarrow Y_2 \subseteq Y_1$) 일 때, 개념 (X_1,Y_1)은 개념 (X_2,Y_2)의 하위 개념이라고 하며, 반대로 개념 (X_2,Y_2)는 개념 (X_1,Y_1)의 상위 개념이라고 한다. 아래의 그림 4는 격자구조(Complete Lattice)를 이용하여 개념간의 상-하위 관계를 표현한 것이다.

이와 같이 추출된 색인어로부터 개념들을 구성하여 상-하위개념 관계를 구성함으로써, 격자구조를 구축할 수 있다. 추출된 개념들은 자연스럽게 객체집합이나 속성집합에 의한 계층적 관계가 형성이 되며, 이를 통하여 개념격자(Concept Lattice)를 구축할 수 있다[19,20].

	female	juvenile	adult	male
girl	x	x		
woman	x		x	
boy		x		x
man			x	x



[그림 4] Formal Concept Analysis의 예제

4. 제안한 시스템의 설계 및 구현

4.1 웹 온톨로지 구축 과정

포괄적인 온톨로지 구축과정은 목적 확인, 개념화, 기호화, 기존 온톨로지 통합, 평가, 문서화와 같은 과정으로 이루어져 있다.

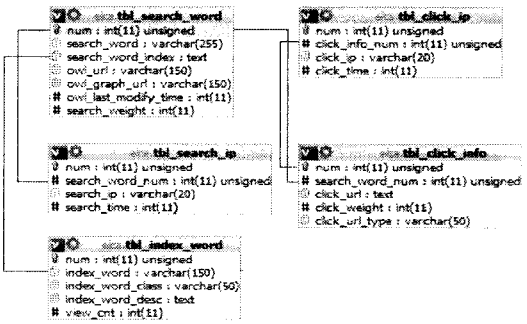
목적 확인은 온톨로지 구축의 목적을 분명히 하는 단계로 이용 대상과 특성을 파악하는 단계이다. 개념화 단계는 핵심적인 개념의 파악과 개념간의 관계를 어떻게 설정하게 되는지 확인하는 단계로, 개념과 용어를 정의하고 확인하는 작업을 하게 된다. 관련된 단어들을 최대한 나열하되, 중복개념은 배제하고 유사개념은 통합하거나 유사관계를 부여하여 나열한다.

본 논문에서는 사용자가 입력한 검색 키워드를 최상위의 계층으로 지정하고, 추출된 색인어와 관련 키워드들은 FCA분석을 통하여 하위 계층으로 분류하여 온톨로지를 생성한다. 추출된 색인어에 의미를 부여하기 위하여 색인어의 개수만큼 OpenAPI와 WordNet을 이용하여 질의를 하게 된다. 추출된 색인어와 관련된 키워드 정보를 받아서 확장을 위한 클래스로 정의하고, 색인어가 2.1절에서 언급한 메타데이터를 구성할 수 있는 경우 해당 어휘를 이용하여 정보를 추가적으로 구성하게 된다.

기호화 단계에서 생성할 온톨로지는 W3C에서 표준으로 권고하는 OWL을 기준으로 한다.

마지막으로 관리 단계에서는 통합 및 갱신 작업이 이루어지는데, 이는 다른 사용자에 의하여 입력된 유사한 키워드의 검색이나 동일한 키워드 검색 시 통합 또는 새

로운 개념을 추가하는 갱신작업을 수행하게 된다. 생성되는 온톨로지의 정보를 저장하기 위하여 MySQL을 사용하였으며, Database에는 온톨로지를 구성하는 색인어의 정보와 관련이 있는 온톨로지의 정보, 저장된 온톨로지 파일의 URI 정보 등을 담고 있다. 단순화 시킨 Database Schema는 그림 5와 같다.

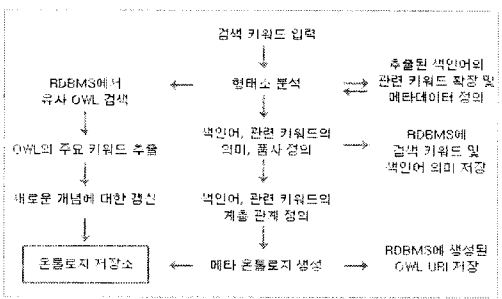


[그림 5] 온톨로지 정보 저장소 DB Schema

4.2 설계 및 구현

3 장에서 기술한 온톨로지 자동 생성 시스템의 기반 기술과 4.1절에서 언급한 웹 온톨로지 구축 과정을 기준으로 설계한 온톨로지 자동 생성 시스템의 아키텍처(architecture)는 그림 6과 같다.

개발한 시스템은 사용자가 입력한 검색 키워드의 형태소 분석을 통하여 색인어를 추출한다. 추출된 색인어는 OpenAPI와 WordNet을 이용하여 단어의 의미 및 품사 그리고 상위어, 하위어, 유사어 등의 개념을 함께 정의하게 되며, 이 정보는 추후 관리 및 갱신을 위하여 RDBMS에 함께 저장하도록 한다. 또한 OpenAPI의 관련 키워드 정보를 이용하여 해당 검색 키워드와 관련이 있는 정보 및 추출된 색인어와 관련이 있는 정보를 함께 표현하도록 한다.



[그림 6] 온톨로지 생성 시스템 아키텍처

계층관계를 정의하기 위하여 처음 입력된 검색 키워드를 최상위 계층으로 정의하고, 추출된 색인어는 FCA분석을 이용하여 최상위 계층의 하위 계층으로 정의한다. 추출된 색인어의 개수가 1개 이상인 경우, 각각을 형제(sibling)계층으로 정의하며 유사어, 상위어, 하위어 등은 각 형제 계층의 하위 계층으로 존재하게 된다. 색인어가 메타데이터 어휘로 표현이 가능한 경우 해당 어휘를 사용하여 추가적으로 온톨로지를 구성한다. 예를 들어, 가수 이름을 검색했을 경우 해당 가수의 프로필과 관련인물 정보를 FOAF로 구성하고, 발매한 음반의 정보를 Dublin Core로 표현할 수 있다.

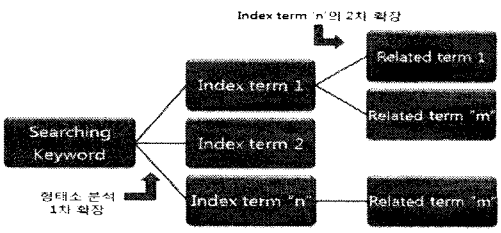
이런 단계를 기준으로 온톨로지를 생성하게 되며, 생성된 온톨로지는 온톨로지 저장소에 저장하여 관리하게 된다. 생성되었던 온톨로지를 구성하는 주요 정보와 URI를 RDBMS에 저장하고 있으므로, 다른 검색 키워드를 입력하더라도 형태소 분석을 통하여 색인어를 추출하고 기존에 있던 OWL파일에서 동일한 개념을 찾아서 연결하여 준다. 이는 검색 키워드 단위의 온톨로지를 생성하고 다른 검색 키워드를 구성하는 색인어와 색인어의 관계를 연결시키기 위함이다.

동일하거나 유사한 검색 키워드를 입력하였을 경우 RDBMS에서 기존의 온톨로지를 구성하는 정보를 검색하여 새롭게 변경되거나 추가될 부분, 삭제될 부분을 찾아서 갱신하도록 한다.

생성되는 온톨로지는 형태소 분석을 통하여 추출된 n개의 색인어 클래스와 OpenAPI를 통해 해당 색인어와 연관성을 갖는 키워드 클래스 m개를 갖게 된다. 온톨로지 내부에 생성되는 클래스의 개수 X는

$$X=(\sum_{i=0}^{n}\sum_{j=0}^{m}I_iRj)+1$$

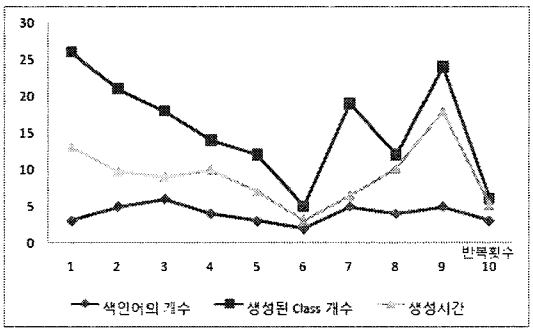
개가 생성된다. 여기서 I_i는 i 번째 추출된 색인어이며, R_j는 색인어 I_i에서 확장된 j 번째 관련 키워드이다. I_i는 관련된 데이터가 존재하는 경우에만 R_j로 확장되며, 입력된 검색 키워드 자체가 한 개의 클래스를 구성하게 되므로 생성되는 클래스의 수에 +1을 하였다. 본 논문에서는 R_j와 관련된 3차 확장 이후에 대해서는 다루지 않도록 하며, 2차 확장까지의 구조는 그림 7과 같다.



[그림 7] 키워드 확장을 통한 온톨로지 클래스 정의

그림 8과 표 2는 개발한 시스템을 이용해 임의의 키워드를 입력해서 온톨로지를 생성하고, 추출된 색인어와 생성된 클래스의 개수 및 생성에 소요된 시간을 표현하였다. 확장을 통하여 키워드가 많이 정의될수록 생성에 많은 시간이 소요되었으며, 평균적으로 약 9.11초가 걸렸다. 또한 OpenAPI를 이용하여 추출된 색인어의 확장을 통해서 형태소 분석으로 얻은 색인어의 개수보다 4배정도 많은 클래스를 얻을 수 있었다.

그림 9는 개발한 시스템을 이용하여 온톨로지를 생성한 것으로 검색 키워드를 입력하게 되면 해당 키워드를 확장하여 웹에 존재하는 데이터를 온톨로지로 만들고, 생성된 온톨로지를 RDF Triple로 변환하여 그림 10과 같은 그래프를 생성하도록 한다.



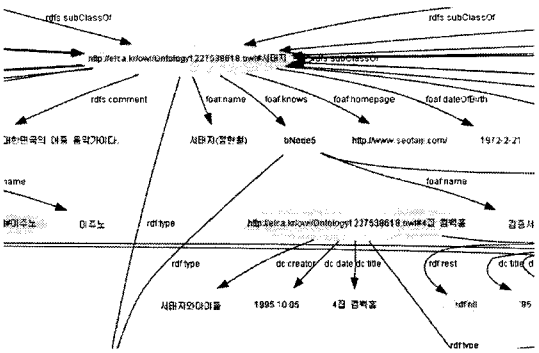
[그림 8] 온톨로지 생성시간 및 생성된 클래스의 개수

[표 2] 온톨로지 생성 실험 결과

시행횟수	색인어의 개수	생성된 Class 개수	생성시간
1	3	26	13.0333
2	5	21	9.6832
3	6	18	8.9326
4	4	14	9.9571
5	3	12	6.9571
6	2	5	3.0744
7	5	19	6.4016
8	4	12	10.1176
9	5	24	17.9018
10	3	6	5.1319

[그림 9] 온톨로지 자동 생성 시스템

생성된 온톨로지의 문법적 타당성을 검사하기 위하여 University of Maryland 에서 제공하는 Pellet OWL Reasoner를[21] 이용하여 유효성 검사를 시행하였다. 온톨로지의 유용성 또는 정확성을 정확하게 판단할 기준이 마련되어 있지 않고 온톨로지의 표현 부분은 상당히 주관적인 부분이므로, 본 논문에서는 문법적 타당성과 표현상의 오류여부만 검증하였으며 결과는 그림 11과 같다. 검사를 시행한 온톨로지의 문법적 결함이 없으면 그림과 같이 'Consistent: Yes' 라는 결과를 보여준다.



[그림 10] 생성된 OWL의 Graph

Results

Input file: http://elca.kr/owl/Ontology1227538618.owl
OWL Species: Full
DL Expressivity: ALC(D)
Consistent: Yes
Time: 2323 ms (Loading: 2067 Species Validation: 123 Consistency: 17 Classification: 112 Realization: 3)

[그림 11] 생성된 온톨로지의 유효성검사 결과

5. 결론 및 향후 연구 방향

기존의 온톨로지를 생성하여 적용시키는 방법은 대부분 해당 도메인 전문가의 수작업에 의존하기 때문에 풍부한 어휘와 의미적으로 충실한 온톨로지가 생성되는 장점을 갖지만, 많은 시간과 비용이 필요하고 구축이 어렵다는 단점이 있다. 본 논문에서는 시간적 소모를 줄이면서 온톨로지를 자동으로 생성하기 위하여 검색 키워드의 확장을 이용하여 온톨로지를 생성하는 방법에 대하여 제안하였다. 웹 기반 사전의 OpenAPI와 WordNet을 이용하여 의미와 계층적 관계를 정의하고, 색인어 간의 관계 및 관련 있는 OWL의 참조를 통하여 온톨로지 확장을 하고 있다. 또한 RDBMS에 클래스를 이루는 색인어와 확장된 키워드의 정보를 저장하여, 동일한 검색이 일어날 때 기존 온톨로지의 정보를 갱신할 수 있다.

기존의 정형화된 온톨로지를 생성하는 방법으로는 범위가 한정되지 않고 폭발적으로 자료가 늘어나는 웹 환경에서 모든 정보를 표현하는 온톨로지를 생성하는 것은 불가능하다. 하지만 본 논문에서 제안한 키워드 확장을 이용한 온톨로지 생성 방안은 간단한 정보를 기술한 단순한 온톨로지를 생성하고 메타데이터간의 의미관계를 조금씩 부여해 확장시키는 방법으로 해당 키워드의 최신 정보를 반영할 수 있고, 자동으로 생성되므로 많은 시간을 절약할 수 있다. 또한 웹상에 존재하는 데이터를 기준으로 새로운 정보를 계속 추가하고 변경되는 정보를 갱신하며 확장을 시켜나갈 수 있다는 장점을 갖고 있다.

본 논문에서는 색인어와 관련된 키워드의 정의에 있어서 범위 제한을 하였지만, 데이터가 존재하지 않을 때까지 관련 키워드의 범위를 넓히게 되면 더 많은 양의 클래스를 정의하고 기존의 온톨로지와의 연결이 가능하다.

웹에서 사용되는 온톨로지는 데이터의 규격화와 데이터를 서로 연결하는 Linked Data[22]의 역할을 한다고 할 수 있다. 이러한 데이터의 연결고리들로부터 원하는 정보를 찾기 위하여 RDF Query 언어인 SPARQL을 이용해 검색 시스템에 적용시킬 계획이다. 실제 검색엔진 모델에서 검색 키워드를 입력하게 되면 온톨로지를 자동으로 생성하고, 생성된 온톨로지를 바탕으로 검색 키워드와 일치하는 정보를 얻고자 한다. 이는 유사한 문서의 제안뿐만 아니라 보다 정확한 데이터의 검색이 가능할 것으로 기대된다.

참고문헌

[1] T. Gruber, "A Translation Approach to Portable Ontology Specifications", Knowledge Acquisition, Vol. 5, No. 2, pp. 199-220, 1993.

[2] T.B. Lee, "Artificial Intelligence and the Semantic Web", AAAI 2006 Keynote, July 2006, http://www.w3.org/2006/Talks/0718-aaai-tbl

[3] OWL Web Ontology Language Reference, February 2004, http://www.w3.org/TR/owl-ref

[4] M.Y. Dahab, H.A. Hassan, A.A. Rafea, "TextOntoEx: Automatic ontology construction from natural English text", Expert Systems with Applications, Vol. 34, No. 1, pp. 1474-1480, 2008.

[5] P. Clerkin, P. Cunningham, and C. Hayes, "Ontology Discovery for the Semantic Web Using Hierarchical Clustering", Semantic Web Mining Workshop, 2001.

[6] M. Degeratu, V. Hatzivassiloglou, "Building Automatically a Business Registration Ontology", ACM International Conference Proceeding Series, Vol. 129, pp. 1-7, 2002.

[7] S.S. Weng, H.J. Tsai, S.C. Liu, C.H. Hsu, "Ontology construction for information classification", Expert Systems with Applications, Vol. 31, No. 1, pp. 1-12, 2006.

[8] H. Alani, S.H. Kim, D.E. Millard, M.J. Weal, W. Hall, P.H. Lewis, N.R. Shadbolt, "Automatic Ontology-Based Knowledge Extraction from Web Documents", IEEE Intelligent Systems, Vol. 18, No. 1, pp. 14-21, 2003.

[9] GATE (General Architecture for Text Engineering), http://gate.ac.uk

[10] Dublin Core, http://dublincore.org

[11] FOAF (Friend Of A Friend), http://www.foaf-project.org

[12] Microformats, http://microformats.org

[13] RDFa, http://rdfa.info

[14] SIOC (Semantically-Interlinked Online Communities), http://sioc-project.org

[15] SKOS (Simple Knowledge Organization System), http://www.w3.org/2004/02/skos

[16] 국민대학교 한글공학-정보검색 연구소, http://nlp.kookmin.ac.kr

[17] 강승식, "한국어 복합명사 분해 알고리즘", 정보과학회논문지, Vol. 25, No. 1, pp. 172-182, 1998.

[18] 김미혜, "FCA 개념 망 기반 개인정보관리", 인터넷정보학회논문지, Vol. 6, No. 6, pp. 163-178, 2005.

[19] J. Eijck, J. Zwarts, "Formal Concept Analysis and

Prototypes", Workshop on the Potential of Cognitive Semantics for Ontologies, September 2004.
[20] B. Ganter, R. Wille, Formal Concept Analysis: Mathematical Foundations, Springer-Verlag, 1999.
[21] Pellet OWL Reasoner, http://www.mindswap.org/2003/pellet
[22] T.B. Lee, Linked Data, http://www.w3.org/DesignIssues/LinkedData.htm

심 준(Joon Shim) **[정회원]**

- 2007년 9월 ~ 현재 : 고려대학교 정보경영공학과 석사과정

<관심분야>
Semantic Web, Ontology

이 홍 철(Hong-Chul Lee) **[정회원]**

- 1983년 2월 : 고려대학교 산업공학 학사
- 1988년 2월 : Univ. of Texas 산업공학 석사
- 1993년 2월 : Texas A&M Univ. 산업공학박사
- 1996년 3월 ~ 현재 : 고려대학교 정보경영공학과 교수

<관심분야>
SCM, 생산 및 물류 정보시스템, PLM