시멘틱 검색을 위한 공리(axiom) 데이터베이스 구축의 개념적 모델

Developing Conceptual Model of Axiom Database for Semantic Search

조용훈, 한성대학교 대학원 문헌정보학과, jyhun1221@naver.com
서은경, 한성대학교 지식정보학부 교수, egseo@hansung.ac.kr

Yong-Hun JO, Graduate School of Library and Information Science, Hansung University
Eun-Kyung SEO, Prof., Division of Knowledge & Information, Hansung University

팀 버너스 리에 의해 '시멘틱 웹'은 1998년 제안되었으나 현재 새롭게 생성되고 있는 데이터 혹은 자연어 형식의 데이터를 시멘틱 검색을 위해 활용하기에는 아직까지 온톨로지 데이터베이스가 따라가지 못하고 있다. 이를 위해 온톨로지 구축의 구성요소인 공리(axiom)를 공공을 위한 데이터로 개발하여 시멘틱 검색에 활용하는 개념적 모델을 제안한다. 공리 데이터베이스는 단일 도메인에서 벗어난 시멘틱 검색을 위한 데이터베이스로서 도메인 온톨로지 구축에 기본적인 요소들을 제공하고, 이용자들이 시멘틱 검색을 통해 보다 만족한 정보검색을 할 수 있도록 한다. 또한 온톨로지 데이터를 확보하기 위해 정보생산자로부터 사전어휘에 대한 온톨로지 트리플을 생성하는 실험을 하였다. 온톨로지 자동구축에 대한 연구와 개발이 활발하지만 보편적 시멘틱 검색을 위해 정보생산자와 정보관리자가 많은 부분 데이터를 생성하고 검증해야할 필요가 있다.

1. 서 론

어느 산업분야에서든지 기반시설이 매우 중요하다. 예를 들어 자동차 산업에서는 자동차를 생산할 수 있는 시설이 매우 중요하고, 축산업에서는 가축을 키울 수 있는 시설이 필요하다. 하지만 시멘틱 정보검색분야에서는 기반 시설이라고 할 만한 부분이 마땅치 않다. 물론 시멘틱 검색을 위한 방법론은 이미 개발되어 이용자에게 서비스를 제공하고 있지만, 1998년 팀 버너스 리가 시멘틱 웹을 제시한 지 15년이 지난 지금도 시멘틱 검색에 필요한 기반 기술 구현이 쉽지는 않다. 특히 현재의 기술로는 자연어 처리부분과 데이터를 기반으로 한 추론기술이 아직 완벽하지 못하고 시스템이 자동적으로 온톨로지 데이터를 구축하지 못하여 고도의 노력이 필요한 수작업이 요구된다는 문제점을 가지고 있다.

이로 인하여 시멘틱 검색을 허용하는 시스템이 매우 적은 수이며 또한 특정 도메인 대상으로 한 실험적 시스템에서 허용하기 때문에 시멘틱 검색에 대해 많은 사람들이 알고 있더라도 정작 시멘틱 검색을 대부분 이용하지 못하고 있는 실정이다. 비록 특정기관시스템에서 시멘틱 검색을 구성했더라 하더라도 일반이용자들이 포털사이트처럼 쉽게 접근하고 검색하기란 쉽지 않다. 따라서 일반이용자들도 쉽게 시멘틱 검색을 이용할 수 있게 하기 위해서는 더 나아가서 즉 시멘틱 웹이 활성화되기 위해서는 온톨로지 구축이 용이해져

야 하고 온톨로지 대상이 보편화되어야 한다. 하지만 시멘틱 웹의 특성상 온톨로지는 특정 도메인을 대상으로 구축되기 때문에 보편화에 매우 취약하다는 문제점을 가진다.

이러한 문제점을 극복하기 위하여 본 연구는 온톨로지의 기본 구성요소인 공리(axiom)를 플랫폼으로 사용가능한 데이터베이스로 구축하여 도메인 온톨로지 구축시 공리 데이터베이스부터 기본 정보를 제공 받을 수 있고 시멘틱 검색에서 도메인 온톨로지간 정보공유가 가능하도록 개념적 모델을 제안하였다. 특히 정보생산자가 트리플태그를 이용하여 온톨로지 데이터를 생산하는 방법을 제안하여 보다 편리하고 보편적인 공리 데이터베이스를 구현할 수 있는 개념적 모델을 제안하였다.

2. 선행연구

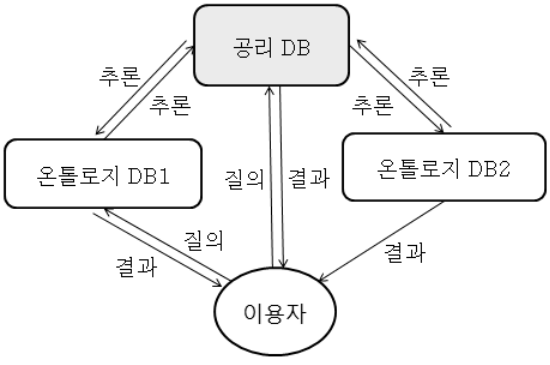
시멘틱 검색을 위한 온톨로지 데이터 구축에 대한 연구는 2000년대 초기에 시작되어 추론을 기반으로 한 시멘틱 검색기술을 개발하는 연구(하상범, 2004), 시멘틱 검색기술을 적용한 도메인 온톨로지 구축 사례연구(이기성, 2004), 그리고 온톨로지 기반의 검색시스템을 설계 및 구현하는 연구(김용, 2012; 양명석, 2012)으로 나눌 수 있다.

각각 시스템의 온톨로지는 특정 도메인이나 검색 목적에 맞추어져 설계되고 구현된 것이어서 보편적 검색을 위하여 온톨로지를 서로 연결하여 추론을 한다는 것은 쉽지 않다. 우선 시간과 비용을 들여 구축한 데이터를 공개하는 것이 쉽지 않고, 공개를 해도 온톨로지 공유를 위한 플랫폼이 없기 때문이다. 도메인 온톨로지에서 구축한 시소러스 변환 온톨로지 데이터는 LOD (Linked Open Data)에 공개·공유 하여 연계 및 공유하는 역할을 하지만(황미녕, 2012), 공리 데이터베이스도 LOD와 같은 역할을 해야 개방적으로 활용될 수 있다.

3. 공리 데이터베이스

온톨로지는 일반적으로 개념(concept), 속성(property), 관계(relationship), 제약조건(constraint), 공리(axiom)와 인스턴스(instance)로 구성되어 있다(최익규, 2013). 정보생산자는 온톨로지 구축시 개념, 속성, 관계, 인스턴스를 각 도메인과 이용자의 요구에 맞게 구성을 하게 된다. 특히 공리는 객관적 사실을 말하는 것으로 온톨로지 데이터베이스의 핵심이다. 하지만 제약조건과 공리는 추론을 하기위한 부분으로 일반적으로 개발자가 관계나 값에 대해 규정하는 정의로 정보생산자가 구성하지 못하고 있다.

이 연구의 목적은 특정 도메인만을 대상으로 하지 않는 공리 데이터베이스를 구현하여 각각의 특정 도메인에서의 추론을 수행하는데 필요한 데이터를 제공해주고 또 도메인과 도메인에 대한 추론이 서로 가능하도록 하여 온톨로지 데이터베이스 구축의 효율성을 최대화하는 것이다. 즉 <그림 1>과 같이 도메인 온톨로지 구축에 도움을 줄 수 있을 뿐 아니라 이용자가 직접 공리 데이터베이스를 검색하여 필요한 정보를 활용할 수 있게 하는 공리데이터베이스를 구축을 할 수 있는 방법론을 제시하는 것이다.



<그림 1> 공리 데이터베이스를 이용한 검색

이로써 기존에 시멘틱 검색을 실시했을 경우 이용자는 온톨로지 DB1에 구축되어 있는 정보만을 가지고 검색을 실행하며 대상 도메인에서만 추론한 결과 값을 이용자에게 보여주게 되지만, 공리DB를 구축하여 공리DB값을 추출해 보여줄 수도 있고, 공리DB를 통해 추론한 온톨로지 DB1을 이용하거나 새로운 온톨로지 DB2에 접속해 새로운 정보를 추론해 이용자에게 결과를 제공할 수 있는 것이다. 즉 각각의 도메인간 추론을 통해 시멘틱 검색이 가능하고, 공리 데이터베이스의 활용으로 다각적인 정보 수집도 가능하게 된다. 다음은 공리 데이터베이스를 이용하여 검색이 수행되는 예이다.

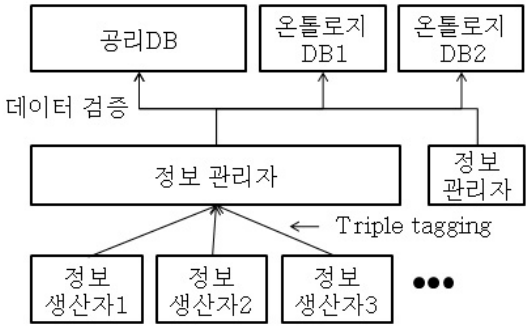
예시) 질의: 경복궁의 위치는?

- 공리온톨로지: 서울 종로구 사직로 161
- 문화재온톨로지: 서울 종로구 사직로 161, 한양(조선시대)
- 교통정보온톨로지: 서울 지하철 3호선 경복궁역 5번 출구

4. 트리플태그를 이용한 공리DB 구축

공리 데이터베이스를 구축하기 위해서는 많은 시간과 비용이 필요하다. 특정 도메인이 아닌 만큼 다양한 분야의 정보생산을 위한 정보생산자의 역할도 중요하고, 구축된 공리 데이터를 도메인 온톨로지와 연결하기 위한 정보관리 및 기준도 중요하다. 하지만 현실은 이렇게 많은 부분을 시간과 비용을 들여 하는 것이 쉽지 않다. 그래서 본 연구에서는 트리플태그를 사용한 방법을 제안하였다. 트리플태그는 아직 정보관리자에 의해 검증되지 않은 온톨로지 정보로 정보생산자가 태그와 같은 방법으로 스스로 트리플을 생성해내는 방법을 말한다. 웹 온톨로지 언어인 OWL(Web Ontology Language)에서는 기본적으로 RDF(Resource Description Framework)형식의 트리플(triple, 주어-술어-목적어)을 사용한다. 주어(Subject)는 사물의 이름에 해당하고, 술어(Predicate)는 기술하는 분야를 명시한다. 목적어(Object)는 술어에서 명시하는 값을 나타낸다. 이렇게 트리플 형식으로 표현했을 때, 특정 도메인의 온톨로지 구축이 가능하게 된다(김명관, 2009).

공리 데이터베이스는 기존의 온톨로지 프로그램을 통해 온톨로지를 구축하는 방법이 아니라 OWL의 트리플을 태그형식으로 바꾸어 정보생산자가 입력기를 통하여 입력됨으로서 생성될 수 있다. 즉 전문 데이터베이스의 각각의 한 문장에서 트리플태그를 추출하여 정보생산자가 태그를 만들게 되면 이후, 정보관리자가 트리플태그를 온톨로지 데이터베이스에 적용시켜 만드는 방식을 말한다. 이 과정을 그림으로 표현하면 다음 <그림 2>와 같다.



<그림 2> 트리플태그 구축 과정

트리플태그 구축에는 복수의 정보생산자와 복수의 정보관리자, 공리데이터베이스, 복수의 온톨로지DB가 필요하다. 정보생산자는 트리플태그를 찾아 정보관리자에게 전달하는 역할과 비정형 텍스트에 대한 분석을 직접적으로 하는 역할을 한다. 가장 말단에서 태그를 생성하는 모든 이용자도 정보생산자에 해당된다. 정보관리자는 정보생산자가 생산한 트리플태그를 수집하며, 기존의 데이터베이스와 비교

하여 트리플태그를 공리 데이터베이스와 온톨로지 데이터베이스로 나누어 각각 해당하는 데이터베이스에 입력한다. 공리 데이터베이스에 들어갈 트리플태그를 만들어 내거나 정보생산자에 의해 전달된 트리플태그를 검증이 정보관리자의 가장 중요한 역할이다.

본 연구는 트리플태그를 이용하여 직접 공리 온톨로지를 구축하는 실험을 실시해 보았다. 실험데이터로는 두산백과의 'doopedia'를 이용하였다. 백과사전은 정의가 명확하게 되어 있고, 백과사전 특성상 내용이 비정형 텍스트로 기술되어 있기 때문에 정보생산자가 트리플태그를 생성하기 용이하다는 장점을 가진다. 'doopedia'의 458,605개의 어휘 중 트리플태그 구축이 가능한 어휘를 22개 카테고리 분류에서 각각 3개씩 총 66개를 추출하였다. 트리플태그는 <그림 3>과 같은 방법으로 추출하였다.

삼성신화[三姓神話] : 제주도의 3성 씨족에 관한 시조 신화…제주도민들에게 구전으로 이어져온 신화로, 《고려사#》, 《영주지(瀛洲志)》 등에 그 기록이 전한다. 삼신인의 출현과 세 공주의 내도(內道), 거주 지역의 설정, 탐라국의 건국 등에 관한 이야기로 구성되어 있다….

↓ 문장단위 분할

1	제주도의 3성 씨족에 관한 시조 신화…
2	제주도민들에게 구전으로 이어져온 신화로…
3	삼신인의 출현과 세 공주의 내도(內道)…

↓ 트리플 태그 추출

주어(S)	술어(P)	목적어(O)	문장번호
삼성신화	시조신화	제주도	1
삼성신화	기록	고려사, 영주지	2
삼성신화	구성	삼신인의 출현과 …	3
…	…	…	…

<그림 3> 트리플태그를 이용한 속성추출

위의 방법으로 66개의 어휘에 대한 설명을 292문장으로 한 문장씩 분리해서 트리플태그를 추출해 보았다. 문장단위로 분할한 이유는 트리플 생성에 쓰인 관계의 근거를 문장단위로 다시 찾을 수 있도록 하기 위함이다. 문장단위로 분할된 데이터는 정보생산자에 의해서 트리플태그를 추출해 낸다. 추출한 어휘를 기본으로 해서 주어를 입력하고, 술어에는 속성명칭이 기술된다. 목적어는 술어에 대한 값으로 복수의 개체가 각각의 어휘에 대해 코드형식으로 입력되거나, 문장형식의 텍스트 데이터로 입력된다. 술어에 입력된 값이 코드일 경우 추론에 사용될 수 있지만, 데이터 형태인 경우에는 속성정보를 제공하는 술어를 이용해 추론을 할 수 없다.

각각의 문장에 대한 트리플태그를 추출한 결과를 분류별로 나열하면 <표 1>과 같다.

<표 1> 정보생산자에 의한 속성추출 결과

주제 분류	총 문장수	트리플태그가 있는 문장 수	트리플태그가 없는 문장 수	2개이상의 트리플태그를 갖는 문장 수
문화예술	6	6	0	3
문화유적	7	7	0	4
순수과학	7	7	0	3
기술과학	8	6	2	5
세계지명	8	8	0	1
자연지리	8	8	0	3
역사	10	7	3	0
인물	10	9	1	4
경제경영	11	8	3	1
IT	12	9	3	3
동물	12	8	4	4
스포츠	12	12	0	3
여행	12	7	5	2
철학	12	11	1	5
기관단체	14	12	2	4
지역	14	13	1	9
의학	16	9	7	2
사회과학	17	12	5	4
식물	17	14	3	8
교통통신	18	16	2	2
종교	23	10	13	2
생활	38	28	10	4
합계	292	227	65	76

22개 카테고리 분류에서 트리플태그는 총 327개 구축되었고, 트리플태그를 가진 문장은 227개가 있는 반면, 22%에 달하는 65문장은 트리플태그가 없다. 그 이유는 문장이 이야기를 구성하거나, 생성하려는 주어와 다른 설명을 하는 부분이기 때문이다. 또한 2개 이상의 문장을 갖는 문장수도 76문장 있었다. 한 문장에 트리플태그가 여러 개가 나오는 것은 문장 내 수식어구가 주어와 트리플태그를 구성하거나, 쉼표로 연결된 문항이 2개 이상의 트리플 태그를 생성하기 때문이다.

이렇게 추출한 태그는 정보관리자의 검수 및 검증과정을 거쳐 공리 데이터베이스로 구축되고, 속성에 따라서 도메인 온톨로지와 공유될 수 있을 것이다. ‘doopedia'를 통해 구축한 데이터는 불특정 다수의 이용자에게 시멘틱 검색 서비스를 제공할 수 있다. 또한 공리 데이터베이스로서 검증했을 경우, 도메인 온톨로지와 연동하여 이용자에게는 재현율을 높이고, 개발자에게는 초기 개발에 공리 데이터베이스로 온톨로지 구축에 활용하여 필요한 노동력을 줄일 수 있을 것으로 본다.

5. 결 론

본 연구에서는 공리데이터베이스를 구축하기 위한 실험 데이터를 만들어 보았다. 비전문가인 정보생산자가 66개의 어휘에 대해 평균 약 5개, 총 327개의 트리플태그를 추출하였다. 컴퓨터 프로그램을 통해서 추출하였을 경우, 1개의 어휘에서 2개 이상의 트리플 셋을 자동구축하기에는 많은 전제가 필요하다. 그렇기 때문에 아직까지는 정제되지 않은 비정형 텍스트를 온톨로지로 구축하기 위해서는 정보생산자와 정보관리자의 시간과 노력이 필요하다. 또한 트리플태그를 통해서 구축된 공리 데이터베이스가 스스로 추론을 통해 온톨로지 데이터를 자동 생산할 수 있는 프로그램 개발하는 연구도 함께 진행되어야 할 것이다.

참고문헌

김명관, 이영우. (2009). 웹 문서 정보추출과 자연어처리를 통한 온톨로지 자동구축에 관한연구.『한국인터넷방송통신·TV학회논문지』, 9(3), 61~67.

김용. (2012). 시멘틱 검색시스템 구축을 위한 요구사항 분석 및 설계에 관한 연구.『한국비블리아학회지』, 23(1), 91~111.

양명석 등저 (2012). 시멘틱 웹 기반의 이미지 검색을 이용한 비교 쇼핑 시스템.『정보관리학회지』, 29(4), 123~142.

이기성, 유영훈, 조근식. (2004). 시멘틱 웹 기반의 이미지 검색을 이용한 비교 쇼핑 시스템.『한국정보과학회』, 31(1B), 556~558.

최익규. (2013). 온톨로지 자동구축을 위한 지능형 계층 관계 추출 시스템 연구. 박사학위논문, 1~106.

하상범, 박영택. (2003). 개인화 에이전트를 이용한 시멘틱 웹서비스 검색.『한국정보과학회』, 30(2I), 124~126.

황미녕, 정도헌, 최성필, 조민희, 정한민. (2012). SKOS를 이용한 시소러스의 온톨로지 모델링과 LOD 공개.『한국정보과학회』, 39(1C), 92~94.

두산백과. www.doopedia.co.kr