구조화 된 데이터 기반의 웹 온톨로지 학습 및 확장 모델 설계

정혜진, 정동원

국립군산대학교 정보통계학과
e-mail:{xhyejin86x, djeong}@kunsan.ac.kr

A Design of Web Ontology Learning and Population Model based on Structured Data

Hyejin Jeong, Dongwon Jeong
Dept. of Informatics and Statistics, Kunsan National University

요 약

이 논문에서는 보다 풍부하고 정확한 정보를 제공하기 위한 구조화 된 데이터를 이용한 웹 온톨로지 확장(Population) 모델을 제안한다. 시맨틱 웹이 등장하면서 웹 온톨로지의 구축이 필수 요소가 되었으며, 더욱 정확하고 보다 풍부한 정보를 제공하기 위한 웹 온톨로지 생성 모델에 관한 연구의 필요성이 증가하였다. 이러한 요구 사항을 충족시키기 위해서는 첫 번째로, 일관성 있고 보편적인 개념을 이용한 웹 온톨로지 스키마 생성과 이를 기반으로 한 온톨로지 간 상호운용성 향상이 요구된다. 두 번째로, 보다 풍부한 정보 제공을 위해 정의된 온톨로지를 확장할 수 있는 방법 개발이 요구된다. 이 논문에서는 메타데이터 레지스트리(MDR, Metadata Registry)를 이용하여 생성된 구조화 된 데이터 기반의 온톨로지 학습 및 확장 모델을 제안한다. 특히 구조화 된 데이터에 대한 개념과 이를 기반으로 한 학습 및 확장의 특징 등에 대하여 기술하고 제안 모델을 위한 시스템 구조에 대하여 기술한다.

키워드 : 메타데이터 레지스트리, MDR, 시맨틱 웹, 웹 온톨로지, 학습, 확장

I. 서론

기존의 웹은 단순 키워드에 의한 정보 접근만을 허용 하였으며, 필요한 정보를 효과적으로 추출·해석·가공할 수 있는 기능은 제공하지 않았다. 따라서 기존의 웹을 확장하여 웹 정보에 잘 정의된 의미를 부여함으로써 사람뿐만 아니라 컴퓨터도 쉽게 문서의 의미를 해석할 수 있도록 하여 컴퓨터를 이용한 정보의 검색·해석·통합 등의 업무를 자동화하기 위한 시맨틱 웹이 1998년 팀 버너스-리에 의해 제안되었다[1-2].

시맨틱 웹을 구현하기 위해서는 다양한 기술이 요구되며 그 중에서 가장 기본적인 기술은 웹 리소스에 대한 정보와 자원 사이의 관계-의미 정보를 컴퓨터가 처리할 수 있도록 하기 위한 온톨로지의 구축이라 할 수 있다. 온톨로지의 관리 및 활용을 위한 다양한 편집기, 웹 온톨로지 기술 언어, 저장소 및 추론 엔진 등이 개발되었으며, 특히 W3C에서는 웹 온톨로지의 효율적인 생성을 위한 웹 온톨로지 기술 언어인 RDF(Resource Description Framework), RDFS(RDF Schema), OWL(Web Ontology Language) 등을 제안하였다[3-5].

앞서 기술하였듯이, 이미 시맨틱 웹을 구현하기 위한 다양한 기술들은 개발된 상태이다. 따라서 향후에는 보다 풍부한 정보 및 서비스 제공을 위한 웹 온톨로지의 보다 정확한 정의 및 정의된 온톨로지에 대한 확장 혹은 통합에 대한 연구가 요구된다. 특히 웹 온톨로지에 대한 보다 정확한 정의 및 확장을 위해서는 다음과 같은 요구 사항을 충족시켜야 한다.

- 일관성 있고 보편적인 개념을 이용한 웹 온톨로지 스키마 생성과 이를 기반으로 한 온톨로지 간 상호

※ 이 논문은 2008년 정부(교육과학기술부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구(KRF-2008-314-D00485)임.

운용성 향상

- 보다 풍부한 정보 제공을 위한 온톨로지 학습 및 확장 방법의 개발

이 논문에서는 앞서 언급한 요구 사항을 충족하기 위한 메타데이터 레지스트리(MDR, Metadata Registry) 기반의 웹 온톨로지 정의와 MDR 기반의 구조화 된 데이터를 이용한 온톨로지 확장 모델을 제안한다. MDR은 ISO/IEC에서 개발한 데이터베이스 간 상호운용성 향상을 위해 개발된 국제 표준(ISO/IEC 11179)으로서 이를 기반으로 한 많은 메타데이터 레지스트리들이 개발되었다[6-8]. MDR은 표준 개념, 즉 일관성 있는 의미 정보를 관리함으로써 이를 기반으로 생성된 데이터베이스 간 의미 공유 및 교환이 용이하다. 따라서 MDR 내의 표준 개념을 이용하여 보다 정확한 의미 교환이 가능한 웹 온톨로지 생성이 가능하다. 또한 MDR을 기반으로 생성된 데이터인 구조화 된 데이터를 이용한 기 생성된 온톨로지의 확장이 용이하다. 다시 말해, MDR의 표준 데이터 요소를 이용하여 정의한 데이터베이스의 데이터를 기존에 생성한 온톨로지에 추가함으로써 보다 풍부하고 정확한 정보 제공 및 서비스 개발이 가능함을 의미한다.

이러한 MDR의 특징과 장점을 이용하여, 이 논문에서는 MDR을 이용하여 생성된 구조화된 데이터 기반의 온톨로지 생성 및 확장을 위한 새로운 모델을 제안한다. 이 논문에서는 제안 모델의 구조, 프로세스 정의 등의 정의, 즉 설계에 초점을 둔다.

이 논문의 구성은 다음과 같다. 제2장에서는 MDR과 학습 및 확장에 대하여 기술한다. 제3장에서는 제안 모델의 전반적인 구조와 프로세스에 대하여 기술하고 제4장에서는 제안 모델의 평가에 대하여 기술한다. 마지막으로 제5장에서는 결론 및 향후 연구에 대하여 기술한다.

II. 관련 연구

이 장에서는 MDR의 기본 구조 및 특징에 대하여 소개하고 온톨로지 학습 및 확장을 위한 주요 기술에 대하여 기술한다.

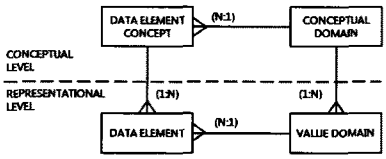
1. MDR

MDR(Metadata Registry)은 ISO/IEC JTC 1/SC 32에 의해 개발된 국제 표준으로서, 데이터 의미의 표현, 등록, 관리, 교환, 공유를 주목적으로 하며 레지스트리에 등록된 정보의 상호운용성 확보에 중요한 기반을 제공한다[6].

MDR에 의해 관리되는 데이터의 기본단위는 데이터 요소이며 3가지 구성요소로 이루어진다[7].

- 객체 클래스(Object Class)
- 특성(Property)
- 표현(Representation)

<그림 1>은 ISO/IEC 11179 MDR 모델로서 MDR에서 가장 중요한 개념적 모델의 요소인 개념 영역, 값의 영역, 데이터 요소 개념, 데이터 요소들 간의 관계를 보여준다[8].



<그림 1> ISO/IEC 11179 MDR 모델

데이터 요소는 데이터 요소 개념과 값의 영역으로 이루어지며, 여러 개의 데이터 요소는 같은 데이터 요소 개념 혹은 같은 표현을 공유할 수 있다. 값 영역은 데이터 요소와 관련 없이 독립적으로 관리될 수 있으며 동일한 값의 의미를 공유하면 이 값 영역들은 개념적으로 연관성을 가진다. 여러 개의 값 영역들은 동일한 개념적 영역을 공유할 수 있으며 데이터 요소 개념은 오직 하나의 개념 영역에 속한다.

MDR은 메타 모델의 형태로 기술되어 있다. 이는 구현 내용의 재사용과 공유를 용이하게 하기 때문이다. MDR의 메타 모델은 기능적으로 6분야로 나뉘며 다음과 같다.

- 관리와 식별
- 명명과 정의
- 분류
- 데이터 요소 개념
- 개념 영역/값 영역
- 데이터 요소

2. 학습 및 확장을 위한 주요 기술

MDR에 등록되어 있는 표준 데이터 요소는 온톨로지의 생성과 학습(Learning) 및 확장(Population)에 이용되어 온톨로지를 더욱 풍부하게 할 수 있다. 이때, 학습 및 확장을 하기 위하여 Full Text 분석을 통한 학습과 Logic 기반의 학습을 활용할 수 있다. 전자의 경우에는 Lucene을 이용하여 개발이 가능하며, 후자의 경우에는 Jena의 추론 엔진을 통한 개발이 가능하다.

Lucene은 Java를 이용하여 검색을 위한 인덱스를 작성하게 해주고 검색을 가능하게 하는 Full Text 검색 엔진이다.

일반적으로 Lucene은 데이터를 수집하는 부분은 담당하지 않으므로 색인·검색하고자 하는 자료는 텍스트 형태로 Lucene에 제공해야 한다. 즉, Lucene은 텍스트 형태로 변환할 수 있는 모든 파일을 색인하고 검색할 수 있다[9].

Jena는 HP 시맨틱 웹 연구소에 의해 개발된 시맨틱 웹 프레임워크로 RDF, RDFS 및 OWL을 위한 프로그래밍 환경과 기본적인 RDF 파서를 제공하며 내부적으로 Rule 기반의 추론 엔진을 포함하고 있다. Jena의 RDFS 추론 엔진은 대부분의 RDFS Entailment 계산이 가능하며 Full, Default, Simple의 세 가지 모드로 작동된다. OWL 추론 엔진은 OWL Lite를 지원하며 OWL Micro Reasoner, OWL Mini Reasoner, OWL Reasoner의 세 가지 추론 모드가 있다. 규칙 엔진은 전향 및 후향 추론 엔진(RETE 기반 전향 추론 엔

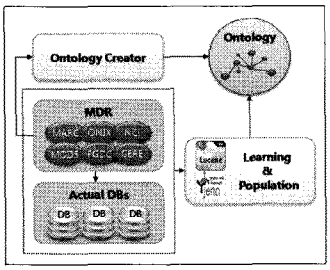
진, Datalog 기반 후향 추론 엔진)을 모두 포함하고 있으며, Triple 모델을 기반으로 한 술어 표현이 가능하다[10].

III. 제안 모델

이 장에서는 제안 모델의 개념, 시스템 구조 및 주요 프로세스에 대하여 기술한다.

1. 제안 모델의 개념

이 논문에서 제안하는 온톨로지 학습 및 확장 모델은 MDR을 기반으로 한다. 즉, MDR의 표준 요소를 이용하여 웹 온톨로지 스키마를 정의하며, 또한 MDR의 표준 요소를 이용하여 정의된 데이터베이스 내의 데이터 기반의 학습 및 확장 연산이 이루어진다. <그림 2>는 제안 모델의 전체적인 개념도를 보여준다.



<그림 2> 제안 모델의 개념도

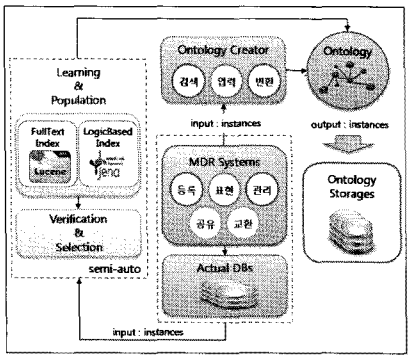
제안 모델은 MDR에 등록된 표준 데이터 요소를 이용하여 실제 데이터베이스 스키마의 필드를 정의한 경우와 그렇지 않은 경우를 고려한다.

먼저, 전자의 경우에는 Ontology Creator를 통하여 MDR에 등록되어 있는 표준 데이터 요소를 기반으로 온톨로지를 생성한다. 여기에서 표준 데이터 요소는 MDR의 개념에 대한 인스턴스를 가리킨다. 즉, 객체 클래스, 특성, 표현(데이터 요소 개념, 데이터 요소, 값 영역, 데이터 영역)에 대한 인스턴스를 말한다. 온톨로지를 생성한 후 학습 및 확장을 이용하여 MDR에 등록되어 있는 표준 데이터 요소를 Logic 기반의 학습을 통하여 기존 온톨로지와 함께 학습을 한다. 이 때 Logic 기반의 학습을 하기 위하여 Jena의 추론 엔진을 이용한다. 이를 기반으로 MDR에 등록되어 있는 표준 데이터 요소를 이용하여 실제 데이터베이스 스키마의 필드를 정의한 데이터베이스의 인스턴스를 기존에 정의해 놓은 온톨로지 스키마의 해당 개념이나 속성에 추가하는 확장 과정을 거쳐 보다 풍부하고 정확한 정보를 제공하기 위한 웹 온톨로지를 생성한다.

또한, 후자의 경우에는 실제 데이터베이스 스키마의 필드가 MDR의 표준 데이터 요소와 유사한지를 확인하여 전자의 경우와 같이 온톨로지를 생성하여 확장을 한다. MDR의 표준 데이터 요소와 실제 데이터베이스 스키마의 필드의 유사성을 확인하기 위해서 Full Text를 통한 학습을 하게 되며, 이 때 Lucene을 이용한다.

2. 시스템 구조

<그림 3>은 제안 모델을 위한 시스템 구조를 보여준다.



<그림 3> 제안 모델의 구조

MDR 시스템은 등록·표현·관리·공유·교환을 담당하는 부분으로 구성된다. MDR 시스템에 의해서 표준 데이터 요소가 등록 되며, MDR에 등록된 표준 데이터 요소를 이용하여 온톨로지를 생성하기 위한 Ontology Creator 모듈이 필요하다. 이 모듈은 검색·입력·변환을 담당하는 부분으로 구성된다.

또한, MDR에 등록된 표준 데이터 요소와 실제 데이터베이스의 인스턴스를 이용하여 웹 온톨로지를 풍부하게 하기 위해서 학습 및 확장 모듈이 필요하다. 이 모듈에서는 MDR에 등록되어 있는 표준 데이터 요소에 대하여 학습 및 확장을 하기 위한 Logic 기반의 학습을 하며, Jena의 추론 엔진을 이용한다.

이와 반대로, 데이터베이스 스키마 필드를 정의할 때 MDR에 등록되어 있는 표준 데이터 요소를 이용하지 않은 경우, MDR에 등록되어 있는 표준 데이터 요소와 유사한 것인지에 대하여 학습 및 확장을 하기 위해 Full Text 검색 엔진인 Lucene을 이용한다.

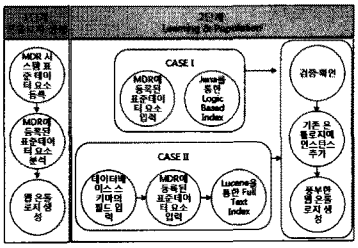
학습 및 확장 모듈은 사람이 직접 적합한 결과인지에 대한 검증·확인 작업을 해야 하므로 semi-auto 단계라고 볼 수 있다.

학습 및 확장을 하여 추출된 데이터 요소들은 보다 풍부하고 정확한 정보를 제공하기 위한 웹 온톨로지로 생성되며 온톨로지를 위한 별도의 저장소에 저장한다. 온톨로지를 위한 별도의 저장소는 대부분의 웹 온톨로지 저장소로 이용되고 있는 관계형 데이터베이스를 기반으로 한다.

이 논문에서 제안 모델은 MDR에 등록되어 있는 표준 데이터 요소를 기반으로 데이터베이스 스키마와 웹 온톨로지 스키마가 이미 정의되어 있다고 가정한다.

3. 프로세스

<그림 4>는 제안 모델의 프로세스를 보여주며 전체적 프로세스는 크게 2단계로 구성된다. 1단계는 온톨로지 생성 단계이며 2단계는 학습 및 확장을 위한 단계이다.



〈그림 4〉 프로세스

먼저, 1단계에서는 MDR System을 이용하여 MDR의 표준 데이터 요소를 등록하고 MDR에 등록된 표준 데이터 요소를 분석하여 웹 온톨로지를 생성한다.

2단계는 두 가지의 프로세스로 구분 할 수 있으며, 그 기준은 실제 데이터베이스가 MDR에 등록된 표준 데이터 요소를 이용하여 데이터베이스 스키마의 필드를 정의했는지의 여부이다. CASE I은 MDR에 등록된 표준 데이터 요소를 이용하여 정의하였을 경우로서, 먼저 MDR에 등록된 표준 데이터 요소를 Jena를 통한 Logic 기반의 학습 연산을 수행하고 추출된 데이터 요소들이 적합한지에 대한 검증·확인 작업을 거쳐 기존 온톨로지에 데이터베이스의 인스턴스를 추가하는 확장 연산을 수행하여 보다 풍부하고 정확한 정보를 제공하기 위한 웹 온톨로지를 생성한다.

CASE II는 MDR에 등록된 표준 데이터 요소를 이용하지 않고 데이터베이스 스키마를 정의하였을 경우로서, 데이터베이스 스키마의 필드와 MDR에 등록된 표준 데이터 요소가 유사한지를 확인하기 위하여 Lucene을 이용하여 Full Text 학습 연산을 수행한다. 유사 후보 데이터 요소들이 추출되며 확인 작업을 통한 확장 연산이 수행된다.

IV. 평가

이 장에서는 제안 모델의 평가에 대하여 기술한다. 현재까지는 MDR을 이용하여 생성된 구조화 된 데이터를 기반으로 보다 풍부하고 정확한 정보를 제공하기 위한 학습 및 확장 모델에 대한 연구들이 이루어지지 않았다. 따라서 이 논문의 비교 평가가 어렵기 때문에 제안 모델의 장점을 중심으로 기술한다.

제안 모델은 MDR을 이용하여 구조화된 데이터를 기반으로 웹 온톨로지 스키마를 정의하였기 때문에 웹 온톨로지 스키마 개념의 일관성을 향상시키며 보다 정확한 의미를 갖는 웹 온톨로지를 생성할 수 있다. 또한, 제안 모델은 학습 및 확장을 통하여 인스턴스를 수집하여 보다 풍부한 온톨로지를 생성할 수 있다. 따라서 직접 정보를 수집하여 인스턴스를 추가하는 것에 대비하여 보다 풍부한 정보를 수집할 수 있으며 보다 풍부한 정보를 수집하는 데에 드는 비용과 시간을 절감할 수 있다.

제안 모델의 장점을 요약하면 다음과 같다.

- 웹 온톨로지 스키마를 구성하는 개념의 일관성 향상
- 웹 온톨로지의 정보를 직접 추가해야 할 경우의 시간 및 비용 절감
- 보다 정확한 의미를 갖는 웹 온톨로지 생성
- 보다 풍부한 정보를 갖는 웹 온톨로지 생성

V. 결론 및 향후 연구

이 논문에서는 보다 풍부하고 정확한 정보를 제공하기 위하여 MDR을 이용하여 생성된 구조화된 데이터 기반의 온톨로지 학습 및 확장 모델을 제안하였다. 특히, 제안 모델의 개념과 시스템 구조, 주요 프로세스와 제안 모델이 제공하는 장점에 대하여 기술하였다. 그러나 이 논문에서는 제안 모델에 대한 설계에 초점을 두었으며 실제 구현에 대한 내용은 언급하지 않았다. 따라서 이 논문의 구현에 대한 연구가 향후 이루어져야 한다.

참고문헌

[1] T. Berners-Lee, "The World Wide Web: Past, Present and Future," IEEE, IEEE Computer Magazine, Vol. 29, No. 10, pp. 69-77, October 1996.

[2] Asuncion Gomez-Perez and Oscar Corcho, "Ontology Languages for the Semantic Web," IEEE Computer Society, IEEE Intelligent Systems, Vol. 17, No. 1, pp. 54-60, January/February 2002.

[3] World Wide Web Consortium. http://www.w3.org/, 2008.

[4] W3C, Resource Description Framework, http://www.w3.org/RDF/, 2008..

[5] W3C, Web Ontology Language, http://www.w3.org/2004/OWL/, 2008.

[6] 오삼균, "시맨틱 웹 기반 메타데이터 레지스트리 설계에 관한 연구," 한국도서관·정보학회, 한국도서관정보학회 하계학술발표회, pp. 7-36, 2005. 6.

[7] 백두권, "데이터 표준화와 메타데이터 레지스트리(MDR: MetaData Registry)," 한국정보통신기술협회, TTA 저널, 제2000권 제71호, 2000. 10.

[8] ISO/IEC JTC 1/SC 32/WG 2 Web site, http://metadata-standards.org/, 2008.

[9] 양단, 김양훈, 김국보, "루씬 라이브러리를 이용한 블로그 검색기 설계," 한국인터넷정보학회, 2008춘계학술발표대회, 제9권 제2호, pp. 415-420, 2008. 5.

[10] Jena Semantic Web Framework, http://jena.sourceforge.net/, 2008.