

XMDR 기반 온톨로지 구축 방법에 대한 실험 및 평가

이석훈*, 정동원**†, 김장원*, 백두권*†

*고려대학교 컴퓨터·전파통신공학과

**군산대학교 정보통계학과

e-mail : leha82@korea.ac.kr, djeong@kunsan.ac.kr, ikaros1223@korea.ac.kr, baikdk@korea.ac.kr

Experiment and Evaluation of the XMDR-based Ontology Building Method

Sukhoon Lee*, Dongwon Jeong†**, Jangwon Kim*, Doo-Kwon Baik*†

*Dept. of Computer and Radio Communications Engineering, Korea University

**Dept. of Informatics and Statistics, Kunsan National University

요 약

온톨로지 간 이질성 문제를 해결하고 상호운용성을 향상시키기 위한 연구가 진행되어 왔으며, 최근 XMDR에 기반한 온톨로지 구축 방법이 제안되었으나 기존 연구와의 비교 평가가 부족하여 장점을 정확하게 보이지 못하였다. 따라서 이 논문에서는 XMDR 기반 온톨로지 구축 방법의 장점을 보다 명확하게 보이기 위해 정량적인 평가를 수행한다. 이를 위해 실제 온톨로지를 구축하고, 구축된 온톨로지는 온톨로지 참조 기반 온톨로지 구축 방법, 사전 참조 기반 온톨로지 구축 방법, 기존 방법론을 이용한 온톨로지 구축 방법을 평가 대상으로 하여 5가지 평가 지표로 분석된다. 평가 지표로는 구축된 온톨로지의 어휘 및 구조의 일관성 비교를 위하여 어휘 및 구조의 빈도수 평균과 엔트로피를 사용하고 구축 비용의 평가를 위하여 각 온톨로지의 구축 시간을 사용한다. 이러한 실험 및 평가의 결과로써, 온톨로지 참조 기반의 온톨로지 구축 방법은 다른 온톨로지 구축 방법들에 비해 온톨로지 어휘 및 구조가 일관적이고 효율적임을 보인다.

1. 서론

지식을 표현하기 위한 한 방법으로 온톨로지가 사용되고 있다. 온톨로지에 대한 연구가 계속되면서 온톨로지는 시맨틱 웹, 클라우드 컴퓨팅, 웹 서비스 등의 영역에서도 중요한 역할을 하는 기술이 되었다[1]. 온톨로지는 그 표현이 다양하고, 같은 온톨로지를 구축하는 사람들 사이의 지식 수준 차이에 따라 그 어휘와 구조가 달라진다. 따라서 온톨로지 공학, 온톨로지 구축 방법론 등의 잘 정의된(well-defined) 온톨로지를 구축하기 위한 연구가 진행 되었다[2]. 그러나 기존 온톨로지 구축 방법론들은 온톨로지를 구축하기 위한 가이드라인 정도만 제공할 뿐, 웹 자원과 함께 온톨로지의 폭발적인 증가로 인해 야기되는 문제점들을 해결하지는 못한다. 특히, 온톨로지 간 이질성, 온톨로지 기반 서비스의 상호운용성 저하와 같은 문제점들은 온톨로지 학습 및 확장을 더욱 어렵게 한다.

이러한 문제들을 해결하기 위하여 XMDR (eXtended Metadata Registry)에 기반한 온톨로지 구축 방법과

같은 연구가 제안되었다[3]. 그러나 이 연구에서는 기존 접근 방법과의 정성적인 평가 결과만을 보였을 뿐, 실질적인 온톨로지 구축 실험을 통한 구체적인 정량 평가 결과를 제공하지 않는다.

이 논문은 XMDR에 기반한 온톨로지 구축 방법에 대한 정량 평가를 위해 온톨로지를 실제로 구축하고 구축된 온톨로지에 대한 평가를 한다. 먼저 온톨로지 전문가들이 XMDR에 기반한 온톨로지 구축 방법과 비교 대상이 될 만한 다른 온톨로지 구축 방법들을 이용하여 동일한 영역에서 동일한 목적으로 온톨로지를 구축한다. 그리고 구축된 온톨로지를 구축 방법 별로 온톨로지들이 어느 정도의 이질성을 지니는지, 즉 어휘와 구조가 얼마나 일치하고, 얼마나 다른지를 정량적으로 평가한다. 이러한 실험 및 평가를 통하여 XMDR 기반 온톨로지 구축 방법이 다른 방법과 비교하여 얼마나 효율적이고 상호운용성 측면에서 나은 성능을 보이는지를 명확하게 확인할 수 있다.

이 논문의 구성은 다음과 같다. 2장에서는 온톨로지 구축 방법과 온톨로지 평가에 대한 관련연구를 기술한다. 3장에서는 온톨로지를 구축 방법에 따라 온톨로지를 구축하는 실험과 실험의 결과를 분석하고 평가하는 내용을 기술한다. 마지막으로 4장에서는 결론을 기술한다.

이 연구에 참여한 연구자는 '2 단계 BK21 사업'의 지원을 받았으며, 지식경제부의 기술혁신사업-표준기술력향상사업 일환으로 수행되었음. 또한 이 논문은 2010년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임(No.2010-0010981).

† 공동 교신 저자(Co-Corresponding authors)

2. 관련 연구

이 장에서는 평가 대상이 되는 온톨로지 구축에 관련한 연구와 온톨로지 평가에 관련한 연구를 소개한다.

2.1 온톨로지 구축 관련 연구

XMDR (eXtended Metadata Registry)은 국제 표준 ISO/IEC 11179 MDR (Metadata Registry)의 확장된 버전이다[4]. [3]은 XMDR에 기반하여 XMDR에 등록된 공통 개념을 온톨로지 구축하는 방법을 제안 하였다. 이 방법은 XMDR에 개념에 대한 어휘가 정의되어 있고, 개념들 사이의 관계와 속성 또한 정의되어 있어 이를 사용하여 온톨로지를 구축하는 방법이다.

[5]는 온톨로지 구축 시 워드넷의 어휘를 이용하여 구축하는 방법으로 워드넷에서 구축된 어휘들은 그 뜻과 의미가 정확하고 어휘간의 관계(동의어, 동의어, 반의어 등)가 명확히 정의되어 있어 정확한 어휘의 온톨로지 구축이 가능하다. 그러나 워드넷은 용어의 어휘적 특성에 대한 정의만 되어 있을 뿐 특정 영역에서 공통 개념과 개념들 사이의 관계를 정의하지 않기 때문에 구축된 온톨로지의 구조적인 이질성 문제가 발생할 수 있다.

Ontology Development 101은 온톨로지 개발 방법론의 하나로써, 미국 스탠포드 대학에서 개발한 온톨로지 구축 방법이다[6]. 개발하고자 하는 분야의 전문가가 직접 온톨로지를 개발하는 방법이며 전문가는 7 단계를 거쳐 온톨로지를 개발한다. 하지만 이 방법은 전문가마다 지식 수준이 다르기 때문에, 각 시스템마다 다른 전문가가 온톨로지를 정의 할 경우 상호운용성의 문제가 발생할 수 있다.

2.2 온톨로지 평가 관련 연구

구축된 온톨로지의 품질을 평가하기 위한 다양한 연구가 이루어져 왔다. 온톨로지 평가는 크게 4가지 접근법과 6가지의 단계로 구분된다[7]. <표1>은 각 단계별 온톨로지 평가 방법들의 비교를 나타낸다.

<표 1> 단계별 온톨로지 평가 방법 비교

Approach Level	Golden Standard	Application- based	Data- driven	Assessment by humans
Lexical, vocabulary, concept, data	√	√	√	√
Hierarchy, taxonomy	√	√	√	√
Other semantic relations	√	√	√	√
Context, application		√		√
Syntactic	√			√
Structure, architecture, design				√

- Golden Standard - 잘 정의되었다고 판단되는 고품질의 표준(golden standard)을 선정해서 구축한 온톨로지를 비교하는 방법이다. 구축된

온톨로지의 어휘, 상속 관계, 다른 의미 관계, 문법에 관한 비교가 가능하다.

- Application-based - 실제 적용될 수 있는 응용 프로그램에 구축된 온톨로지를 적용함으로써, 제대로 적용이 되었는지를 보고 평가하는 방법이다. 구축된 온톨로지의 어휘, 상속 관계, 다른 의미 관계, 문맥에 대한 비교가 가능하다.
- Data-driven - 온톨로지가 참조하는 문제 영역에 대한 실제 데이터와 구축된 온톨로지를 비교하여 얼마나 적용할 수 있는지를 평가한다. 구축된 온톨로지의 어휘, 상속 관계, 다른 의미 관계에 대한 비교가 가능하다.
- Assessment by humans - 구축된 온톨로지 영역의 전문가 혹은 온톨로지 전문가가 직접 온톨로지를 평가하는 방법이다. 구축된 온톨로지의 어휘, 상속 관계, 다른 의미 관계, 문맥, 문법, 구조 등에 대한 평가가 가능하다.

이 절에서 소개한 관련 연구는 구축된 온톨로지의 품질에 대한 평가 방법이다. 그러나 이 논문은 XMDR에 기반한 온톨로지 구축 방법에 대한 평가에 초점을 두고 있다. 따라서 이 논문은 온톨로지 구축 방법 관련 연구들을 평가 대상으로 선정하고, 각 방법에 의해 구축된 온톨로지와 XMDR 기반 온톨로지 구축 방법에 의해 구축된 온톨로지의 품질을 비교 평가한다. 이를 통해 XMDR 기반 온톨로지 구축 방법의 장점 및 단점을 정량적이고 명확하게 확인할 수 있다.

방법론에 의해 구축된 온톨로지 간의 이질성과 상호운용성이 얼마나 높은지를 평가하기 위하여, 온톨로지 평가 관련 연구의 Data-driven 접근법을 이용한다. 따라서 각 방법론에 의해 구축된 온톨로지 간 어휘, 상속 관계, 다른 의미 관계의 이질성 정도를 평가 항목으로 사용하여 평가한다.

3. 실험 및 평가

이 절에서는 평가 대상과 평가 지표를 정의하고 실험 방법을 소개한다. 그리고 실험 후, 평가 대상을 평가 지표에 따라 평가한 분석 결과에 대하여 기술한다.

3.1 평가 대상

평가 대상은 관련 연구의 온톨로지 구축 방법들을 선정하여 3가지의 온톨로지 구축 방법들로 정의 하였다. 각 평가 대상들은 구축 방법에 따라 온톨로지를 구축하여 평가에 이용된다.

- 온톨로지 참조 기반 온톨로지 구축 방법 - 이미 구축된 잘 정의된 온톨로지를 재사용하고 참조하여 온톨로지를 구축한다. XMDR에 기반한 온톨로지 구축 방법은 잘 정의된 공통 개념을 이용하여 구축하므로 이 방법과 동일한 성능을 지닌다.

- 사전 참조 기반 온톨로지 구축 방법 - 사전에 있는 용어를 이용하여 온톨로지를 구축하는 방법이다. 사전은 용어에 대한 명확한 정의와 정확한 어휘의 사용이 가능하다는 특징을 가진다. 워드넷은 그 형태나 구조가 사전과 유사하다. 따라서 이 방법은 워드넷을 참조하여 온톨로지를 구축한다.
- 기존 방법론을 이용한 온톨로지 구축 방법 - 기존의 다양한 온톨로지 구축 방법론 중 실험의 일관성을 위하여 Ontology Development 101 방법을 이용하여 온톨로지를 구축한다[6].
- 구조의 빈도수 평균 - 각 평가 대상에서 개념 별 트리플 구조를 분석하여 그 트리플 별 빈도수의 평균을 계산한다. 예를 들면 “교수와 학생 간의 수학 관계” 라는 개념의 트리플 구조가 있다고 한다면 ‘teach(professor, student)’와 ‘learn_from(student, professor)’와 같이 서로 다른 구조가 존재할 수 있다. 각 구축된 온톨로지마다 트리플 구조들의 개념을 추출하여 개념 별 트리플 빈도수를 측정한다. 같은 개념이지만 다른 트리플의 구조를 띤 경우도 이질성 및 상호운용성에 영향을 끼치기 때문이다. 평가 대상의 구조의 빈도수 평균 역시 평균이 높을수록 빈도수 높은 트리플이 많아 나은 성능을 보인다.
- 구조의 엔트로피 - 각 평가 대상에서 구조의 개념 별로 트리플의 빈도수를 구하여 엔트로피를 계산한다. 트리플 구조가 개념에 대하여 얼마나 일관되어 있는지를 나타낸다. 트리플의 빈도수 엔트로피 공식은 어휘의 빈도수 엔트로피 공식과 동일하게 이용한다. 각 평가 대상을 A_i 라고 한다면 각 개념 트리플들에 대한 엔트로피의 효율 $efficiency(D_j)$, 각 개념 트리플 별 엔트로피는 그 개념의 빈도수 $|c|$ 와 각 트리플의 빈도수 c_k 에 의해 결정된다. 평가 대상의 엔트로피가 0에 가까울수록 트리플 구조가 일관적임을 의미한다.
- 평균 구축 시간 - 각 평가 대상 별로 온톨로지를 구축한 시간을 측정하여 평균을 구한다. 평가 대상의 평균 구축 시간이 적을수록 구축 비용이 적어 효율적임을 의미한다.

3.2 평가 지표

각 평가 대상에 대한 평가를 위하여 5가지 평가 지표를 선정하였다. 각 평가 지표는 각 평가 대상에 의해 구축된 온톨로지의 어휘와 구조의 일관성 그리고 효율성을 평가한다.

- 어휘의 빈도수 평균 - 각 평가 대상에서 사용된 어휘의 빈도수를 구하고 각 어휘 별 빈도수의 평균을 구한다. 평가 대상의 평균이 높을 수록 빈도수 높은 어휘가 많아 더 나은 성능을 보인다.
- 어휘의 엔트로피 - 각 평가 대상에서 개념 별 어휘들의 빈도수를 이용하여 엔트로피를 계산한다. 엔트로피는 각 평가 대상의 어휘들이 개념에 대해 어휘들의 일관성 여부를 의미한다. (그림 1)은 어휘의 빈도수 엔트로피 공식을 나타낸다[8,9]. 각 평가 대상을 A_i 라고 한다면 각 평가 대상 별 엔트로피는 각각의 개념들 D_j 에 대한 엔트로피의 효율 $efficiency(D_j)$ 와 그 빈도수에 따른 가중치의 곱을 더함으로써 구할 수 있다. 각 개념 별 엔트로피의 효율 $efficiency(D_j)$ 는 그 개념의 빈도수 $|c|$ 와 각 어휘들 c_k 의 빈도수에 의해 결정된다. 엔트로피는 0 ~ 1의 범위를 가진다. 평가 대상의 엔트로피가 0에 가까울 수록 어휘의 사용이 일관적이다.

$$entropy_{A_i}(D) = \sum_{j=1}^v \frac{|D_j|}{|D|} \times efficiency(D_j)$$

$$efficiency(D_j) = - \sum_{k=1}^{|c|} \frac{Pr(c_k) \log_2 Pr(c_k)}{\log_2 |c|}$$

$$\sum_{k=1}^{|c|} Pr(c_k) = 1$$

(그림 1) 어휘의 빈도수 엔트로피 공식

예를 들어, ‘수업’이라는 개념에 대해 ‘Class’와 ‘Lecture’라는 어휘의 빈도수가 각각 15번, 5번으로 측정 되었다면, ‘수업’이라는 개념에 대해 $Pr('Class') = \frac{15}{20} = 0.75$, $Pr('Lecture') = \frac{5}{20} = 0.25$ 이므로 $efficiency('수업')$ 은 0.81로 계산 된다. 엔트로피는 전체 어휘의 수에 대한 개념의 개수와 비례하게 $efficiency(D_j)$ 를 적용하여 계산된다.

3.3 실험 방법

실험은 전문가들이 3개의 평가 대상의 방법을 모두 이용하여 다음과 같이 주어진 문제에 대한 온톨로지를 구축하는 방법으로 이루어진다.

“학교에서 수강신청을 위한 시스템을 구축한다고 할 때, 학생과 수업과 교수님과의 관계가 잘 드러나도록 온톨로지를 구축하시오.”

전문가들은 3개 조로 나뉜다. 각 조는 구축 방법 별로 대학, 회사, 병원의 3 영역에서 2문제씩 총 6문제에 대한 온톨로지를 구축한다. 각 조가 모든 구축 방법을 이용하여 온톨로지를 구축하되 구축 방법 이용 시 온톨로지의 영역이 서로 겹치지 않도록 문제를 분배 한다.

실험 후 전문가들이 구축한 온톨로지를 평가 대상인 온톨로지 구축 방법 별로 분류하고, 평가 지표를 측정하는 것으로 각 평가 대상을 평가한다.

3.4 실험 결과

실험을 통하여 12명의 전문가로부터 총 72개의 온톨로지를 수집하였다. 72개의 구축된 온톨로지서

총 536개의 어휘와 626개의 트리플 구조를 추출하여 분석하였다.

<표 2>는 평가 지표에 따른 평가 대상들의 분석 결과를 보인다. 평가 대상들은 평가 지표의 결과 값의 상대적인 크기에 의해 성능이 평가된다.

<표 2> 실험 분석 결과

접근 방법 평가 지표	온톨로지 참조 기반	사전 참조 기반	기존 방법론 이용
어휘 빈도수 평균	2.48개	1.54개	1.65개
어휘 엔트로피	0.414	0.666	0.559
구조 빈도수 평균	3.17개	2.52개	2.89개
구조 엔트로피	0.108	0.153	0.174
평균 구축시간	6분50초	9분28초	10분29초

어휘의 빈도수 평균은 온톨로지 참조 기반 방법이 평균 2.48개로 1.54개와 1.65개로 측정된 사전 참조 기반 방법과 기존 방법론 이용 보다 많은 평균 빈도수를 가진다. 이는 평균적으로 온톨로지 참조 기반 방법에 의해 구축된 온톨로지가 빈도수가 높은 어휘를 많이 가지고 있다는 것을 의미한다. 즉 평균적으로 동일한 어휘를 많이 사용한다는 의미이다.

어휘 엔트로피는 온톨로지 참조 기반 방법이 0.414로 가장 낮게 측정되었고, 사전 참조 기반 방법과 기존 방법론 이용은 각각 0.666, 0.559로 측정되었다. 이것은 온톨로지 참조 기반 방법이 다른 두 방법보다 일관적이고 통일된 어휘를 사용함을 의미한다.

트리플 구조 역시 온톨로지 참조 기반 방법은 3.17개의 평균 빈도수 값이 측정되어 다른 두 방법 보다 높게 측정되었다. 트리플 구조의 엔트로피 또한 온톨로지 참조 기반 방법은 0.108로 측정 되어 다른 두 방법보다 낮은 엔트로피 값을 가진다. 이는 온톨로지 참조 기반 방법이 어휘뿐 아니라 개념간의 관계성 역시 일관적이고 동일한 구조를 더 많이 지니는 것을 의미한다.

마지막 평가 지표인 평균 구축 시간은 온톨로지 참조 기반 방법이 6분50초, 사전 참조 기반 방법이 9분28초, 기존 방법론 이용이 10분29초로 측정되어 온톨로지 참조 기반 방법이 가장 적은 비용으로 구축함을 알 수 있다.

분석 결과는 온톨로지 참조 기반 방법이 모든 평가 지표에서 다른 온톨로지 구축 기반 방법보다 우수한 성능을 지니는 것을 보인다. 특히 온톨로지 참조 기반 방법은 짧은 시간에 온톨로지의 어휘 및 구조를 일관성 있게 구축하기 때문에 온톨로지 구축에 따른 비용을 줄이고 다른 온톨로지들 간의 낮은 이질성과 높은 상호운용성을 지닌다.

사전 참조 기반 방법은 기존의 방법론 이용과 비교하여 총 5개의 평가 지표 중 구조 엔트로피와 평균 구축 시간을 제외한 나머지 3개의 평가 지표에서 기존 방법론 이용 방법보다 성능이 떨어지는 것으로 측정되었다. 이것은 워드넷이 어휘가 특정 영역에 한정되지 않고 동의어, 유사어

등 광범위한 데이터 양으로 인해 어휘 선택의 폭이 넓어지기 때문으로 보인다.

4. 결론

이 논문에서는 XMDR에 기반한 온톨로지 구축 방법을 평가하기 위하여 온톨로지를 구축하는 실험을 하고 구축된 온톨로지를 평가 하였다. XMDR에 기반한 온톨로지 구축 방법은 XMDR에서 관리되는 공통 개념을 이용하여 온톨로지를 구축하므로 온톨로지 참조 기반 온톨로지 구축 방법을 평가 대상으로 선정하였다. 또 비교를 위하여 사전 참조 기반 온톨로지 구축 방법과 기존 방법을 이용한 온톨로지 구축 방법을 평가 대상으로 선정하였다. 어휘와 구조의 빈도수 평균 및 엔트로피 그리고 평균 구축 시간을 평가 지표로 선정하였다.

총 12명의 전문가들이 실험을 진행하여 72개의 온톨로지를 구축하였다. 구축한 온톨로지를 분석하여 평가 대상들에 대한 정량 평가를 수행하였다. 온톨로지 참조 기반 온톨로지 구축 방법은 모든 평가 지표에서 다른 온톨로지 구축 방법보다 우수한 성능을 보였다.

향후 연구로는 XMDR에 기반한 온톨로지 구축 방법에 의해 구축된 온톨로지를 이용하여 인스턴스의 학습 및 확장 방법에 대한 연구가 요구된다.

참고문헌

- [1] M. C. Daconta, L. J. Obrst, K. T. Smith, The Semantic Web : A Guide to the Future of XML, Web Services, and Knowledge Management, Wiley Publishing, Indiana, 2003.
- [2] R. Mizoguchi, M. Ikeda, "Towards Ontology Engineering," The Institute of Scientific and Industrial Research, Technical Report AI-TR-96-1, I.S.I.R., Osaka University, 1996.
- [3] 이석훈, 정동원, 백두권, "XMDR에 기반한 온톨로지 구축 방법", 한국정보과학회, 2009년 가을 학술발표 논문집, Vol. 36, No. 2(C), 2009.
- [4] ISO/IEC JTC1/SC32, ISO/IEC 11179 Metadata Registries(MDR) Part 3, 3rd Edition. 2009.
- [5] H. Hu, X.Y. Du, D.Y. Liu and J.H. Ouyang, "Ontology Learning using Wordnet Lexicon," Computational Methods, 2006.
- [6] N. F. Noy, D. L. McGuinness, "Ontology Development 101: A Guide to Creating Your First Ontology," Stanford Knowledge Systems Laboratory, Technical Report KSL-01-05, Stanford Medical Informatics, Technical Report SMI-2001-0880, Mar. 2001.
- [7] Janez Brank, Marko Grobelnik, Dunja Mladenić, "A survey of ontology evaluation techniques," Proceedings of the Conference on Data Mining and Data Warehouses, 2005.
- [8] Bing Liu, "Web Data Mining – Exploring Hyperlinks, Contents, and Usage Data," Chapter 3, Springer, 2007.
- [9] Wikipedia, Entropy (information theory), [http://en.wikipedia.org/wiki/Entropy_\(information_theory\)](http://en.wikipedia.org/wiki/Entropy_(information_theory))