국가과학기술 R&D 기반정보 온톨로지와 추론 모델링

강인수° 정한민 이승우 김평 성원경
한국과학기술정보연구원, NTIS 사업단
{dbaisk°, jhm, swlee, pyung, wksung}@kisti.re.kr

Semantic Web Ontology and Inference for Research Community

In-Su Kang°, Hanmin Jung, Seungwoo Lee, Pyung Kim, Wonkyung Sung
NTIS Division, Korea Institute of Science and Technology Information

요 약

과학기술 연구분야에서 인력, 기관 등의 연구 주체와 논문, 과제, 지적재산권 등의 성과에 대한 온톨로지는, 시맨틱웹 환경에서 이질적 과학기술 연구정보의 의미적 통합과 자동화된 유통, 그리고 암묵적 지식의 추론을 가능케 할 것이다. 이 논문에서는 현재 한국과학기술정보연구원에서 개발 중인 국가과학기술 R&D 기반정보 온톨로지를 소개하고, 그의 응용으로써 온톨로지에 내재된 암묵적 지식들을 규칙을 사용하여 추론하는 과정의 기술에 중점을 둔다. 상기 온톨로지는 인스턴스의 유일성 확보를 위해 URI(Uniform Resource Identifier)서버에 기반하여 온톨로지 인스턴스에 고유한 URI를 할당하는 데 중점을 두고 설계되었으며, 논문의 특정순위저자를 모델링한 저작자정보 클래스를 온톨로지 스키마 상에 명시적으로 표현한다는 특징이 있다.

1. 서 론

의미 표현이 결여된 현재의 웹은 지식 단위의 인식과 해석의 어려움으로 인해 기계에 의한 자동 처리가 쉽지 않다. 이러한 문제를 해결할 수 있는 차세대 웹의 비전으로 제시된 시맨틱웹은, 웹문서의 단위 정보에 해당 분야 온톨로지의 개념 클래스로 정의된 의미태그를 부착함으로써 이종 스키마를 갖는 정보의 의미적 통합 및 유통의 자동화와 명시적으로 표현되지 않은 암묵적 지식의 추론을 가능케 한다. 그러나, 시맨틱웹의 성공을 위한 중요한 전제 조건 중 하나는 개별 분야 온톨로지의 적절한 작성 및 왕성한 사용이다 [3]. 이러한 측면에서 본 논문은 과학기술 연구 분야의 온톨로지 설계 경험을 소개하고 그로부터 새롭게 얻어지는 지식의 추론 과정을 다룸으로써 향후 시맨택웹과 온톨로지 작성의 활성화에 기여하고자 한다.

본 논문에서 소개할 온톨로지는 현재 한국과학기술정보연구원에서 개발 중인 *국가과학기술 R&D 기반정보 온톨로지* (이하 *기반정보 온톨로지*)이다. 기존에 개발된 유사한 연구분야 온톨로지로는 AKT 온톨로지[1], SWRC 온톨로지[2][3], Bibster 온톨로지 [2]와 SWRC+COIN 온톨로지 [1] 등이 있으며, 이들 연구분야 온톨로지의 스키마를 구성하는 주요 클래스들은 인력, 기관, 과제, 논문 등이다. *기반정보 온톨로지*는 온톨로지 스키마 측면에서 이러한 기존 온톨로지들과 크게 다르지 않다.

*기반정보 온톨로지*가 기존 연구분야 온톨로지와 다른 주요한 차이점 중 하나는, 온톨로지에 표현될 인스턴스의 신원을 관리하기 위해 개별 클래스에 종속적인 URI (Uniform Resource Identifier)식별체계에 따라 고유한 인스턴스 ID를 할당한다는 데 있다. 인스턴스 ID의 생성, 저장, 검색 등의 관리를 위해 별도의 URI서버가 사용된다.

*기반정보 온톨로지*의 개별 인스턴스에 고유한 URI 값을 대응시키기 위해서는, 클래스 종속적인 URI식별체계를 적용하기 이전에, 연구 정보 데이터에 내재된 인스턴스에 대한 언어적 표현의 동의성과 다의성 문제를 먼저 해소할 필요가 있다. 위의 두 문제는 실세계의 인스턴스가 고유한 URI값 대신 동의성과 다의성을 내재하고 있는 자연언어로 표현된다는 데서 기인한다. 예를 들면, 두 논문 A와 B의 동일명 저자 "홍길동"은 동명이인일 수 있으며, 동일 기관을 지칭하기 위해 "한국과학기술원"이나 "KAIST"와 같이 서로 다른 표현을 사용할 수 있다. 이의 해결을 위해 본 연구팀에서는 동영다인의 중의성 해소 기법을 개발하여 사용하고 있으며, 이 논문에서는 더 이상 다루지 않는다.

기존의 연구 분야 온톨로지들에서 찾아보기 힘든 *기반정보 온톨로지*의 또 다른 특징은, 논문 작성 당시의 저자의 소속 기관을 표현하기 위해 "저작자정보"라는 다소 비직관적인 클래스를 도입한 것이다. 이렇게 함으로써 저자순위정보를 명시적으로 표현하고 기관의 논문실적을 정확히 산정할 수 있게 된다.

본 논문의 구성은 다음과 같다. 2장에서 *기반정보 온톨로지*의 개요를 기술하고, 3장에서 *기반정보 온톨로지*의 응용 예로써 온톨로지에 내재된 암묵적 지식을 추론하는 과정을 다룬다. 4장에서는 *기반정보 온톨로지*의 구축현황을 소개하고, 5장에서 결론을 맺는다. 논문에서 온톨로지 스키마 구성요소는 볼드체로 표시하였다.

2. 온톨로지 개요

2.1 온톨로지 스키마 - 클래스(class)와 속성(property)

*기반정보 온톨로지*는 현재 과학기술 연구 분야의 핵심 상위 객체들로 **인력, 기관, 과제, 저작물, 게재지, 토픽**을

1) http://www.aktors.org/publications/ontology/
2) http://ontoware.org/projects/swrc/

클래스로 포함하고 있으며, **저작물**은 **논문**, **특허**, **보고서**의 서브클래스를 갖고, **게재지**는 **학술지**와 **학술대회논문집**의 서브클래스로 세분된다.

다음은 *기반정보 온톨로지*의 핵심 상위 객체들간의 최소한의 객체관계속성을 나열한 것이다.

저작물 - hasCreatorsInformation (저작자정보를 갖는다) - 저작자정보
저작물 - hasOriginatedProject (유발과제를 갖는다) - 과제
저작물 - hasPublication (게재지를 갖는다) - 게재지
저작물 - hasTopic (토픽을 갖는다) - 토픽
저작자정보 - hasCreator (저작자를 갖는다) - 인력
저작자정보 - hasOrganizationOfCreator (저작당시기관을 가진다) - 기관
과제 - hasOrganizationOfFundingProject (발주기관을 가진다) - 기관
과제 - hasOrganizationOfPerformingProject (수탁기관을 가진다) - 기관
토픽 - hasSubTopic (서브토픽을 가진다) - 토픽
인력 - hasOrganizationOfPerson (현재소속기관을 가진다) - 기관

위에서 **저작자정보**라는 핵심 상위 객체가 하나 더 있는데, 이것은 **인력**과는 별도로 저작물 작성 당시의 저자를 표현하는 클래스이다. **저작자정보**는 데이터타입속성(datatype property)으로 **저자순위(orderOfCreator)**, **저자기여도(contributionWeightOfCreator)**를 가지며, **저작물**, **인력**, **기관** 등의 클래스와 객체관계속성(object property)을 가지는데 이는 각각 저작의 산물, 저작자, 저작물 작성 당시 저작자의 소속기관을 표현한다. 결국, **저작자정보**는 저작물 작성 시점의 저자에 종속적인 정보들을 표현하고 있다. [3]에서는 저작자정보로써 저자순위정보만을 표현하는 다른 해법을 제시하고 있으나, *기반정보 온톨로지*처럼 저자기여도나 저작당시 소속기관과 같은 부가적인 정보를 포용하는 응용에서는 [3]의 방법으로는 한계가 있다.

2.2 URI 지향적 온톨로지 인스턴스 표현

1장에서 기술한 바와 같이 연구분야 온톨로지에서는 인력이나 기관의 연구실적을 정확히 산정한다는 측면에서 특별히 인스턴스의 신원을 관리해 줄 필요가 있다. 이것은 비단 연구분야 온톨로지에 국한되는 것은 아닐 것이다. 이를 위해 *기반정보 온톨로지*에서는 인스턴스 ID를, 그것의 해당 클래스에 종속적인 URI할당지침에 따라 생성한다. 즉, **저작자정보**를 제외하고, 2.1절에 소개된 모든 핵심 상위 클래스들에는 별도의 URI할당지침이 관련되어 있다. 예를 들면, **저작물** 인스턴스에는, KOI[3] 식별체계를 간략화한 지침에 따라 URI를 부여하고, **인력** 인스턴스에는, 국가과학기술인력 종합정보시스템[4]에서 사용하는 인력식별체계를 근간으로 하여 만든 식별체계에 따라 URI를 부여한다. **기관** 인스턴스에는 학술진흥재단에서 부여한 기관별 코드를 기초로 하는 식별체계를 적용한다. 본 연구팀에서는 생성된 URI관리를 위해 별도의 URI서버를 개발하여 사용하고 있으며, URI서버와의 정보교환을 위해서는 Web Service를 사용한다.

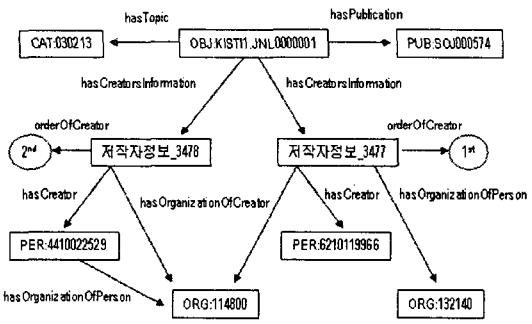
3) KOI(Knowledge Object Identifier): 한국과학기술정보연구원에서 개발한 과학기술지식정보를 연계/공유하기 위한 표준식별체계.

4) http://www.hrst.or.kr/

그림 1은 아래 논문게재건에 해당하는 *기반정보 온톨로지*의 인스턴스들과 그들간의 관계를 URI서버에 저장된 관련 데이터의 내용과 함께 보여준다.

논문제목: 분산 객체지향 데이타베이스에서의 트리 질의 최적화
저자: 김혁만(서울대학교), 이석호(서울대학교)
발행처/게재지: 한국정보과학회/정보과학회논문지
권(호)/페이지: 21(10) / 1849-1859
출판년도: 1994

그림에서 *저작자정보_3477*와 *저작자정보_3478*은 각각 상기 논문의 제1저자 *김혁만*과 제2저자 *이석호*에 해당하는 **저작자정보** 클래스의 인스턴스들이다. **저작자정보** 클래스에는 별도의 URI식별체계가 없으므로, 클래스명을 prefix로 하고 suffix에 자동 생성된 순차번호를 할당하여 인스턴스 ID를 생성하였다. 그림 1은 연구자 *김혁만(URI: 620119966)*이 논문작성 당시의 소속은 *서울대학교(URI:114800)*였으나 현재의 소속은 *국민대학교(URI:132140)*임을 나타내 주고 있다. 그림에서 알 수 있듯이 URI 지향적 기반정보 온톨로지에서는, *인력의 이름*, *저작물의 제목*, *저작물의 출판년도* 등과 같은 URI 종속적인 데이터를 URI서버에 저장해 두고 있으며, 온톨로지에는 URI값만을 인스턴스 ID로 표현해 두고 있다.



URI	유형	메타데이터
PER:4410022529	인력	이름: 이석호, 소속기관: ORG:114800
PER:6210119966	인력	이름: 김혁만, 소속기관: ORG:132140
OBJ:KISTI1.JNL.0000001	저작물	제목: 분산객체지향 데이타베이스에서의 트리 질의 최적화 게재지: PUB:SOJ000574, 연도: 1994, 권호:21(10), 페이지:1849-1859
ORG:114800	기관	기관명: 서울대학교
ORG:132140	기관	기관명: 국민대학교
PUB:SOJ000574	게재지	발행처: 한국정보과학회, 게재지: 정보과학회논문지
CAT:030213	토픽	용어명: 데이터베이스

그림 1. URI 지향적 기반정보 온톨로지 인스턴스 표현의 예 (표의 내용은 URI서버에 저장된 데이터이며, 개별 인스턴스의 생략된 namespace는 http://www.kisti.re.kr/isrl/NSTRDOnto.owl# 이다)

3. 기반정보 온톨로지에서의 추론의 예

시멘틱웹 온톨로지는 다양한 이질적 데이터 스키마를 가진 데이터소스들을 의미적으로 통합하고 유통시킬 수

있을 뿐만 아니라, 선언적 규칙을 사용하여 온톨로지 내에 명시적으로 표현되지 않은 암묵적 연관관계들을 추론해 내는 용도로 사용될 수 있다. 이 장에서는 연구분야 온톨로지의 한 예로 제시된 2장의 *기반정보 온톨로지*로부터, 규칙을 적용하여 온톨로지에 내재된 지식들을 추론해 내는 과정을 기술할 것이다. 이를 위해 먼저 *기반정보 온톨로지*에서 예상되는 규칙들을 정의하고 그 규칙에서 기술된 객체관계속성을 이용하여 *기반정보 온톨로지*에 내재된 객체 간의 관계들을 추출하는 RDQL (RDF Data Query Language) 질의의 예를 보일 것이다.

다음은 *기반정보 온톨로지*에 내재된 지식들을 추출하는 몇 가지 규칙들을 Jena[5] 추론엔진의 규칙기술형식에 따라 쓴 것이다. 각 규칙은 →를 중심으로 하여 *IF-THEN*형식으로 표현된 것이며, *Rule 2*에 사용된 *notEqual(;,)*은 통상의 추론엔진이 지원하는 built-in함수로써 그것의 두 매개변수가 같지 않다는 것을 의미한다.

```
Rule 1: (?x hasCreatorsInformation ?y) (?y hasCreator ?z) → (?x wasCreatedBy
        ?z)
Rule 2: (?x wasCreatedBy ?y) (?x wasCreatedBy ?z) notEqual(?y, ?z) → (?y
        isCoCreatorOf ?z)
Rule 3: (?x hasCreatorsInformation ?y) (?y hasCreator ?z) (?y orderOfCreator 1)
        → (?z isFirstCreatorOf ?x)
Rule 4: (?x hasCreatorsInformation ?y) (?y hasOrganizationOfCreator ?z) → (?x
        isOwnedByOrganization ?z)
Rule 5-1: (?x isCoCreatorOf ?y) → (?x isSameGroupWith ?y)
Rule 5-2: (?x isSameGroupWith ?y) (?y isSameGroupWith ?z) → (?x
        isSameGroupWith ?z)
```

각 규칙의 의미는 다음과 같다.

```
Rule 1: ?x라는 저작물이 ?z라는 인력을 저자로 갖는 어떤 저작자정보
        ?y를 가진다면, 그 저작물 ?x는 인력 ?z에 의해 작성된 것이다
Rule 2: 서로 다른 두 인력 ?y와 ?z가 동일한 저작물 ?x를 작성했다면
        ?y와 ?z는 각각 서로의 공저자이다
Rule 3: ?x라는 저작물이 ?z라는 인력을 제1저자로 갖는 어떤
        저작자정보 ?y를 가진다면, 저작물 ?x의 제1저자는 ?z이다.
Rule 4: ?x라는 저작물이 어떤 인력이 ?z라는 기관에 소속되었을 당시에
        작성한 저작물이라면, 저작물 ?x는 기관 ?z의 실적이다.
Rule 5-1: 공저자 관계에 있는 두 인력은 같은 연구자 그룹에 속한다.
Rule 5-2: ?y와 같은 연구자 그룹에 속하는 서로 다른 두 인력 ?x와 ?z는
        역시 같은 연구자 그룹에 속한다.
```

규칙 5-1과 5-2에서는, 저작물에 대해 단순히 직간접적 공저자관계에 있는 연구자들을 동일 연구자 그룹의 구성원으로 고려하고 있음을 주목하라.

위의 규칙들을 *기반정보 온톨로지*에 적용함으로써 아래와 같은 새로운 객체관계속성(object property)들을 얻을 수 있다.

```
저작물 - wasCreatedBy - 인력
인력 - isCoCreatorOf - 인력
인력 - isFirstCreatorOf - 저작물
저작물 - isOwnedByOrganization - 기관
인력 - isSameGroupWith - 인력
```

다음은, 위의 규칙들이 적용되어 확장된 *기반정보 온톨로지*에서 *데이터베이스*분야의 연구자들을 인력 URI값의 형식으로 추출해 내는 RDQL 질의의 예를 보여준다.

```
SELECT ?y
WHERE (?x wasCreatedBy ?y) (?x hasTopic "데이터베이스")
```

다음 RDQL질의는, 확장된 온톨로지에서 URI값 "*4410022529*"를 갖는 인력("*이석호*", 그림 1 참조)과 같은 연구자 그룹에 속하는 인력들을 인력 URI형식으로 출력해 준다.

```
SELECT ?y
WHERE ("4410022529 isSameGroupWith ?y)
```

4. 온톨로지 구축 현황

*기반정보 온톨로지*의 스키마 부분은 2장에 기술된 핵심 상위 클래스를 포함하여 현재 20개의 클래스와 39개 속성들(데이터타입속성 20개, 객체관계속성 19개)로 구성되어 있으며, 9개의 규칙이 기술되어 있다. 현재 *기반정보 온톨로지*의 인스턴스 부분은 IT 분야 5850건의 논문들로부터 구축되어 있다. 위 논문들은 2002년부터 2006년까지 한국정보과학회, 대한전자공학회, 한국HCI학회, 한국정보처리학회에서 주최한 학술대회논문집에서 추출한 것들이다. 구축된 인스턴스는 규칙(3장의 *Rule 1~4*)의 적용을 통해 얻어진 91,880개 triple을 포함하여 전체 181,979개의 RDF triple로 구성되어 있다. 현재, 이 triple 개수에는 **과제**나 **토픽** 클래스와 관련된 인스턴스는 포함하지 않은 것이다.

5. 결론

이 논문에서는 과학기술 연구분야의 온톨로지 설계 경험을 소개하고 그로부터 다양한 암묵적 지식들을 추론하는 과정을 기술하였다. 본 연구팀은 이러한 지식의 공유가 향후 국내 시맨틱웹 분야의 온톨로지 개발과 다양한 추론기반 응용시스템의 활성화에 기여할 수 있기를 기대한다.

참고문헌

[1] Bloehdorn, S., Haase, P., Hefke, M., Sure, Y., Tempich, C. (2005) "Intelligent Community Lifecycle Support", In *Proceedings of the 5th International Conference on Knowledge Management*, pp.278-285.

[2] Haase, P., Broekstra, J., Ehrig, M., Menken, M., Mika, P., Olko, M., Plechawski, M., Pyszlak, P., Schnizler, B., Siebes, R., Staab, S., Tempich, C. (2004) "Bibster - A Semantics-Based Bibliographic Peer-to-Peer System", In *Proceedings of the 3rd International Semantic Web Conference*, pp.122-136.

[3] Sure, Y., Bloehdorn, S., Haase, P., Hartmann, J., Oberle, D. (2005) "The SWRC Ontology - Semantic Web for Research Communities", In *Proceedings of the 12th Portuguese Conference on Artificial Intelligence*, pp.218-231.

5) http://jena.sourceforge.net/