

OntoThink-K®: DBMS 기반 추론 서비스

정한민^o 강인수 이미경 이승우 성원경
한국과학기술정보연구원 정보시스템연구팀
{jhm^o, dbaisk, jerryis, swlee, wksung}@kisti.re.kr

OntoThink-K®: An Inference Service Based on DBMS

Hanmin Jung^o, In-Su Kang, Mikyoung Lee, Seungwoo Lee, Won-Kyung Sung
ISRL, Korea Institute of Science & Technology Information

요 약

본 논문은 지식 기반 정보유통 플랫폼 OntoFrame-K® 상에서 추론을 이용하여 연구자 간 협업 서비스를 제공할 수 있도록 하는 DBMS 기반 추론 서비스 OntoThink-K®에 대해 기술한다. 본 추론 서비스는 URI 서버를 이용하여 RDF 트리플을 생성하고 추론 규칙에 의해 해당 트리플을 확장하며 SPARQL을 통해 질의 결과를 생성해낸다. 특히 이 모든 과정은 DBMS 기반으로 설계·구현되었는데 URI 서버와 성과·비성과 등록 인터페이스를 통해 별도의 추론 엔진을 사용하지 않고도 정확성이 보장되는 지식을 생성·관리할 수 있도록 하며, 불안정한 성능을 보이는 추론 엔진을 사용하지 않기 때문에 안정적인 성능을 보장할 수 있다는 데 그 특징이 있다. OntoThink-K®는 온톨로지 스키마 트리플, 인스턴스 트리플, 그리고 전방 추론을 통해 획득한 추가 트리플을 포함하는 확장 트리플을 기반 지식으로 하는데, 최종 사용되는 RDF 트리플의 크기는 지식 확장 이전 631,158개, 지식 확장 이후 1,112,100개이다.

1. 서론

시맨틱 웹에 대한 관심이 고조됨에 따라 시맨틱 웹 기술을 활용한 고도화된 서비스 구현 필요성이 점점 더 커지고 있다. 이에 따라 CS AKTive Space, Esperanto, OntoWeb 등 많은 온톨로지 기반 서비스가 출현하였다 [6] [10]. 우리는 시맨틱 웹 기술과 결합한 진보된 서비스를 제공하기 위해 연구자 간 협업을 지원할 수 있도록 지식 기반 정보유통 플랫폼 OntoFrame-K®를 구현하였다 [4]. 본 논문의 초점인 OntoThink-K®는 상기 플랫폼에 적용되어 추론 서비스를 제공하는 것을 목적으로 하며 DBMS 기반의 작업 흐름을 전반적으로 지원하는 동시에 연구자에게 연구자 네트워크, 연구자 정보, 연구 성과 맵, 통계 정보, 성과 정보, 기관 정보 등의 세부 서비스를 제공한다.

온톨로지를 이용한 서비스를 제공하는 방식은 다양한데 특히 새로운 지식을 내재된 지식을 통해 창출하는 추론은 고도화된 시맨틱 웹 기술들의 결합을 요구한다. 성능을 어느 정도 인정받고 있는 추론 엔진으로는 KAON2 (<http://kaon2.semanticweb.org/>), RacerPro (<http://www.franz.com/products/racer/>), OntoBroker (<http://ontobroker.semanticweb.org/>), Pellet (<http://www.mindswap.org/2003/pellet/>), Jena2 (<http://jena.sourceforge.net/>) 등이 있다. 그렇지만, 이들은 처리 영역 (DL 기반의 KAON2와

RacerPro, F-Logic 기반의 OntoBroker)에 있어서 부분적인 제약이 있거나, 인스턴스 적재 (Instance Loading)에 있어서 제약 (Jena2의 경우 OntoThink-K®에서 사용하는 일부 인스턴스를 적재하는 데, RacerPro의 경우 OWL Restriction을 포함한 온톨로지 스키마 로딩에 실패하였다.)이 있거나, 추론 성능에 있어서 제약 (Jena2의 경우 질의 ('?x <per:ancestor> ?y') 처리에 있어 대상 인스턴스 수가 2,000개를 넘기면 결과를 생성하지 못하며, Pellet이나 RacerPro의 경우 LUBM¹ Ontology를 대상으로 할 때 부분적으로 결과를 생성하지 못하였다.)이 있다 [7]. 외부에서 추론 엔진을 도입하는 경우 이러한 문제로 인해 안정적인 성능을 보장하기 힘들다.

이러한 환경을 극복하기 위해서는 OWL Restriction 처리에 대한 부담을 추론 엔진에서 배제하고 안정적인 추론을 위한 프레임워크를 구축할 필요가 있다. 본 연구에서는 OWL Restriction을 성과·비성과 등록 시 인터페이스를 통해 자연스럽게 검증하고 (예를 들어 논문의 경우 저자 수는 최소 1인 이상이어야 하며 (owl:minCardinality = 1), 해당 논문이 가질 수 있는 최대 주제 수는 3개 이내 (owl:maxCardinality = 3)이어야 하므로 성과 등록 인터페이스를 이에 맞게

¹ Lehigh University Benchmark

설계·구현한다는 의미이다.), DBMS 기반으로 추론 규칙을 해석하고 적용하는 추론 시스템을 확보하는 방향으로 문제를 해결하고자 한다. 결국 본 연구는 RDF(S) (Resource Description Framework Schema) 기반의 지식 생성, 전방 추론 (Forward-Chaining Inference)에 의한 지식 (트리플) 확장, 추론 서비스 제공에 이르는 전체 작업 흐름을 DBMS 상에서 구현함으로써 안정적이고 속도가 일정 범위 내에서 움직이는 추론 서비스를 추구한다.

2. 시스템 구성

OntoFrame-K[®] 상에서 추론 서비스를 담당하는 OntoThink-K[®]는 크게 지식 생성 파트, 지식 확장 파트, 추론 서비스 제공 파트로 구성된다.

지식 생성 파트는 스키마 트리플, 인스턴스 트리플 형식의 지식을 온톨로지와 URI 서버 [3]로부터 자동 생성하는 역할을 한다. DB-트리플 변환 규칙은 DBMS 기반 메타데이터를 온톨로지에 부합하는 RDF 트리플로 변환하는 데 이용되는 규칙으로 기존 유사한 연구로는 MT-Redland가 있다 [9]. 이것은 클래스와 속성을 RDF 트리플로 직접 매핑하는 인터페이스를 제공하는데, 본 연구와 같이 DBMS와 직접 연동하는 방식에 비해 스크립트 기반이라 속도가 느리다는 단점이 있다. 또한 매핑 정보들이 메타데이터와 동떨어져 기술됨으로 인해 매핑 규칙을 유지하고 관리하는 데 어려움이 따른다.

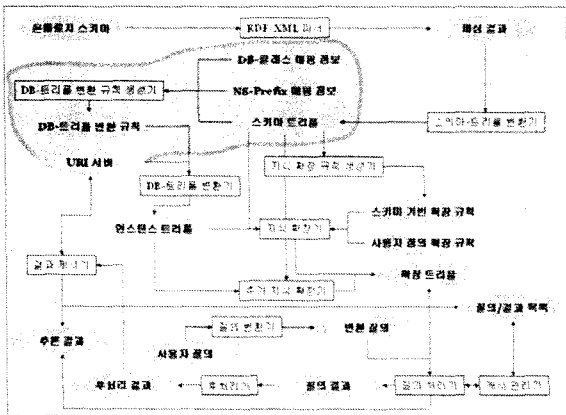


그림 1. OntoThink-K[®] 구성도

지식 확장 파트는 RDF(S)에 기반한 추론 규칙을 이용하여 전방 추론을 수행함으로써 이전에 생성된 지식을 확장하는 역할을 한다. 추론 규칙은 크게 RDF(S) 항의 규칙에 대응하는 스키마 기반 확장 규칙과 새로운 객체 관계 속성을 정의하는 사용자 정의 확장 규칙으로 나누어진다. 확장 트리플은 이러한 과정을 통해 획득된

지식으로 스키마와 무관한 (Schema-Oblivious) 형식으로 표현된다. DBMS 기반 RDF 트리플 표현 형식으로는 본 연구와 같이 하나의 테이블을 이용하여 트리플의 각 요소를 필드로 구성하는 스키마와 무관한 (Schema-Oblivious) 형식, 각 속성과 클래스를 개별 테이블로 구성하는 스키마 인지 (Schema-Aware) 형식, Range 유형 별로 클래스와 더불어 개별 테이블로 구성하는 혼용 (hybrid) 형식이 있다 [5] [8]. 성능 측면에서 볼 때 본 연구에서 사용한 스키마와 무관한 형식이 다른 형식에 비해 다소 떨어진다. 그렇지만, 본 연구의 초점이 아닌 부분으로써 온톨로지의 클래스 수가 작고 안정성을 우선적으로 고려하였기 때문에 스키마와 무관한 형식을 선택한 것이다. 향후 확장되는 지식 크기를 고려하여 스키마 인지 (Schema-Aware) 형식이나 혼용 (hybrid) 형식에서의 전환을 고려할 예정이다.

추론 서비스 제공 파트는 SPARQL (SPARQL Protocol And RDF Query Language) 형식의 질의를 RDF 트리플 상에서 처리하기 위해 SQL로 변환하고 추론 후 해당 결과를 가공하여 연구자에게 제공하는 역할을 한다.

2.1 지식 생성

지식 생성 파트는 DBMS 기반의 RDF 트리플을 생성·관리하는 저장소로서 URI 서버를 이용한다. URI 서버는 온톨로지에서 정의한 개체를 포함하여 성과물 메타데이터까지의 지식화를 위한 데이터를 가지고 있다. 본 시스템은 URI 서버 자체에서 스키마 트리플을 생성하지 않는 대신에 OWL로 기술된 온톨로지 스키마를 파싱하고 파싱 결과에 스키마-트리플 변환기를 적용하여 생성한다. 그림 2는 ARP (Another RDF Parser)²를 이용하여 온톨로지 스키마를 파싱한 결과 예를 보여준다.

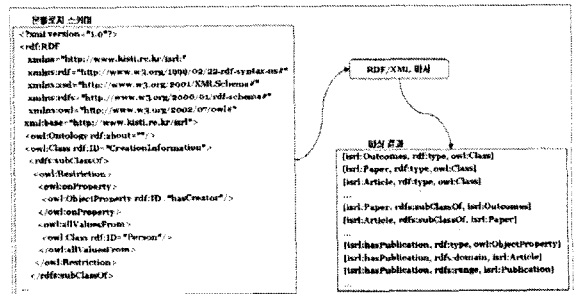


그림 2. 온톨로지 스키마의 파싱 결과 예

반면 인스턴스 트리플은 URI 서버 내 정보를 이용하여 자동 생성하는데 이때 이용하는 정보로는 URI 서버 내 저장소 테이블과 온톨로지 상의 클래스를 매핑시켜주는

² <http://www.hpl.hp.com/personal/jjc/arp/>

DB-클래스 매핑 정보, XML Namespace를 포함하는 Full-named URI와 Prefix를 매핑시켜주는 NS-Prefix 매핑 정보, 개체 별 필드로부터 인스턴스 트리플을 생성할 수 있도록 규칙을 정의한 DB-트리플 변환 규칙이 있다.

The screenshot shows a web interface for defining DB-class mappings. On the left, there's a dropdown menu for selecting a URI. The main area contains a table with columns for 'ClassName' and 'FieldName'. Below this, there's a section for '생성된' (Generated) triples, showing a list of URIs and their corresponding field names.

그림 3. 인력 URI에 대한 DB-클래스 매핑 정보 및 DB-트리플 변환 규칙 예

```
isri:OBJ_KIST11.PCD.123456 rdf:type isri:Outcomes
isri:OBJ_KIST11.PCD.123456 isri:hasCreationInformation isri:KIST11.PCD.123456_7010186243
isri:KIST11.PCD.123456_7010186243 rdf:type isri:CreationInformation
isri:KIST11.PCD.123456_7010186243 isri:orderOfCreator xsd:1
isri:KIST11.PCD.123456_7010186243 isri:hasCreator isri:PER_7010186243
isri:KIST11.PCD.123456_7010186243 isri:hasInstitutionOfCreator isri:INS_9R9048
isri:KIST11.PCD.123456_7010186243 isri:contributionWeightOfCreator xsd:0.93
isri:OBJ_KIST11.PCD.123456 isri:hasCreationInformation isri:KIST11.PCD.123456_6410136403
isri:KIST11.PCD.123456_6410136403 rdf:type isri:CreationInformation
isri:KIST11.PCD.123456_6410136403 isri:orderOfCreator xsd:2
isri:KIST11.PCD.123456_6410136403 isri:hasCreator isri:PER_6410136403
isri:KIST11.PCD.123456_6410136403 isri:hasInstitutionOfCreator isri:INS_9R9048
isri:KIST11.PCD.123456_6410136403 isri:contributionWeightOfCreator xsd:0.85
```

그림 4. 특정 성과물에 대한 인스턴스 트리플 예

그림 4는 두 명의 저자 ('PER_7010186243'과 'PER_6410136403')³를 가진 성과물 ('KIST11.PCD.123456')⁴ 정보를 표현하는 인스턴스 트리플을 보여준다. 해당 성과물 정보는 각 저자의 소속 기관 ('INS_9R9048')⁵, 성과물에 대한 가중치 ('0.93'과 '0.85')를 포함한다.

³ 인력 URI는 국가과학기술인력 종합정보시스템 ('http://www.hrst.or.kr')에서 사용하는 식별 체계를 근간으로 구성된다.

⁴ 객체 URI는 과학기술 지식 정보를 연계·공유하기 위한 표준 식별 체계인 KOI ('http://koix.kisti.re.kr:8081/user/root_Intro_1.jsp')를 근간으로 구성된다.

⁵ 기관 및 부서 URI는 한국학술진흥재단 ('http://www.krf.or.kr/')에서 사용하는 식별 체계를 근간으로 구성된다.

2.2 지식 확장

지식 확장 파트는 스키마 기반 확장 규칙과 사용자 정의 확장 규칙이라는 2개의 추론 규칙을 이용하여 전방 추론 방식으로 RDF 트리플을 추가로 생성한다. 스키마 트리플, 인스턴스 트리플, 상기 과정을 통해 추가 생성된 트리플 등 세 가지 트리플을 합쳐서 확장 트리플이라 명명한다.

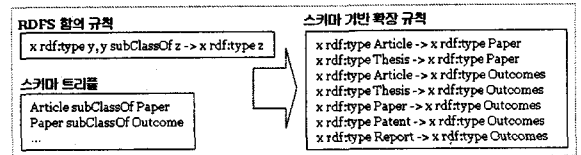


그림 5. 스키마 기반 확장 규칙 예

RDF(S) 합의 규칙과 스키마 트리플을 참조하여 인스턴스 트리플에 바로 적용될 수 있는 그림 5와 같은 스키마 기반 확장 규칙을 생성한다. 그림 상의 RDF(S) 합의 규칙은 RDF Semantics에서 제한한 합의 규칙 중 'rdfs9'에 해당한다⁶ [2] [5].

```
• Article
  - x hasPublication y, y yearOfPublication z -> x hasPublication:Year z
• Thesis
  - x yearOfPublishingThesis y -> x hasPublication:Year z
• Patent
  - x yearOfApplication y -> x hasPublication:Year y
• Report
  - x hasOriginatedProject y, y endingYearOfProject z -> x hasPublication:Year z
• Person
  - x hasCreationInformation y, y hasCreator z -> x createdBy:Person z
  - x hasInstitutionOfPerson y, y hasLocationOfInstitution z -> x locatedIn: z
  - x createdBy:Person y -> y creatorOf: x
• Institution
  - x hasCreationInformation y, y hasInstitutionOfCreator z -> x createdBy:Institution:z
• Location
  - x createdBy:Institution: y, y hasLocationOfInstitution z -> x createdBy:Location z
```

그림 6. 사용자 정의 확장 규칙 예

그림 6은 사용자 정의 확장 규칙의 예를 보여주는 데 'Person'과 'Location'에서의 'createdByPerson', 'createdByInstitution'은 온톨로지 스키마에서 정의된 객체 속성 관계 (Object Property)가 아닌 사용자 정의 확장 규칙에 의해 생성된 것으로서 'creatorOf'와 'createdByLocation'을 생성하는 데에도 이용된다. 추론 규칙을 사용하는 장점이 이러한 부분에 있는데, 새로운 관계를 추가로 정의하기 위해 DBMS 응용 시스템의 경우 스크립트를 갱신하고 필드를 추가로 정의하는 등의 복잡한 작업이 필요하지만 상기와 같이 추론 규칙을 이용하는 경우에는 간단한 규칙 추가로 해결할 수 있다. 현재 22개의 사용자 정의 확장 규칙을 적용하고 있다.

⁶ 본 연구는 스키마 기반 확장 규칙으로서 14개 RDF(S) 합의 규칙 중 현재 서비스에 필요한 'rdfs9'만을 사용한다.

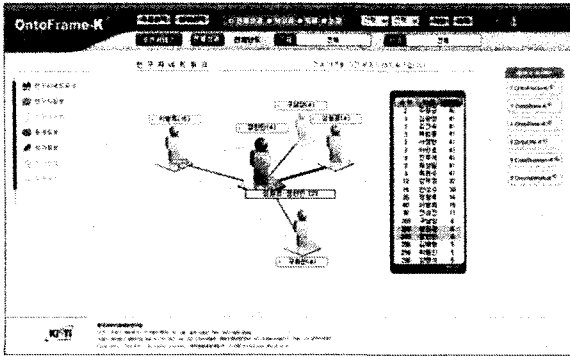


그림 10. 공저자 중심 연구자 네트워크를 보여주는 추론 서비스에

4. 구현 결과

추론 서비스 구현을 위해 KISTI 내부 성과 정보와 외부 성과 정보를 모두 수집·가공하였다. KISTI 내부의 경우 학술 대회 논문 1,172건 (2002년 ~ 2006년), 학술지 논문 360건 (2002년 ~ 2005년), 연구 보고서 734건 (1999년 ~ 2006년), 특허 64건 (2002년 ~ 2005년)을 포함하며, 외부의 경우 원문 포함 학술 대회 논문 8,704건을 포함한다⁹. 추론 서비스에서 이용하는 RDF 트리플 수는 1,112,100개로 (비성과 유형¹⁰: 285,982개, 성과 유형: 826,118개) 이는 지식 확장에 의해 획득된 확장 트리플을 포함하는 개수이다. [4]와 비교하여 확장 트리플 생성 비율을 살펴보면 그 비율이 상당히 작은 편¹¹인데 그 주된 이유는 첫째, 스키마 기반 확장 규칙에서 다루는 RDF(S) 함의 규칙이 현재 1개뿐이며, 둘째, Domain, Range, OWL Restriction 등을 다루지 않기 때문이다. 전자의 경우에는 나머지 13개 RDF(S) 함의 규칙들이 본 추론 서비스에서 별다른 필요성을 가지지 않아 생략한 것이고¹², 후자의 경우에는 성과·비성과 등록 인터페이스와 URI 서버를 통해 정합성을 검사한다 [1] [3]. 이러한 부분을 트리플 확장 대상으로 한다면 수백만 트리플이 확장 트리플로서 추가될 것이다.

⁹ 2차에 걸친 성과를 구축 결과이며, 현재 6차 성과를 구축이 진행 중이다.

¹⁰ 성과와 비성과로 유형을 구분하는 기준은 성과물 메타데이터를 직접 표현하는 트리플인지 (성과), 이와 별도로 기타 정보를 표현하는 트리플인지 (비성과)에 있다. 예를 들어, 특정 논문의 저자와 논문 내 소속 기관은 성과로서 다루어지며, 해당 저자의 현재 소속 기관 (Object Property에 해당)이나 이름 (Datatype Property에 해당)은 비성과로서 다루어진다.

¹¹ [4]의 경우 약 10배 수준이며 본 연구에서는 약 2배 수준이다.

¹² 일반적인 추론 엔진은 서비스에서의 필요성을 고려하지 못하기 때문에 나머지 규칙들에 대해서도 확장을 할 수밖에 없다.

5. 결론

기존에 시맨틱 웹 기술을 이용한 추론 서비스의 구현이 추론 엔진을 중심으로 하는 구조로 설계되어 안정성과 성능에서 제약을 가졌던 데 반해 본 연구에서는 RDF(S) 기반의 트리플로 표현된 지식을 생성하고, 추론 규칙을 이용하여 이를 확장하고, 성과·비성과 등록 인터페이스와 URI 서버를 통해 제약을 검사할 수 있도록 함으로써 기존 추론 서비스 구조를 변화시킬 수 있게 되었다. 특히 실제 데이터를 이용하여 다양한 추론 서비스를 구성함으로써 시맨틱 웹 기술을 필요로 하는 여러 응용 분야에 용이하게 적용될 수 있는 기반을 갖추었다는 데 그 중요성이 더해진다. 향후 지속적인 지식 추가에도 안정적인 성능을 보일 수 있도록 추론 시스템과 저장소 구조를 보완하고, 'YesKiSTI'와 같은 포털 서비스와 결합하여 시맨틱 웹 포털 구축을 시도할 예정이다.

참고문헌

[1] 강인수, 정한민, 이승우, 김평, 성원경, “국가과학기술 R&D 기반정보 온톨로지”, 한국콘텐츠학회 춘계종합학술대회, 2006.
 [2] 김기성, 유상원, 이태휘, 김형주, “RDF 스키마 함의 규칙 적용 순서를 이용한 RDFS 추론 엔진의 최적화”, 정보과학회 논문지: 데이터베이스 33(2), 2006.
 [3] 정한민, 강인수, 구희관, 이승우, 성원경, “URI 서버에 기반한 국가 R&D 기반정보 온톨로지 설계 및 구현”, 정보관리연구 37(2), 2006.
 [4] 정한민, 이미경, 성원경, 박동인, “OntoFrame-K: 연구자 간 협업 지원 서비스를 위한 시맨틱 웹 기반 정보 유통 플랫폼”, 한국컴퓨터종합학술대회, 2006.
 [5] S. Harris and N. Gibbins, “3store: Efficient Bulk RDF Storage”, In Proceedings of the 1st International Workshop on Practical and Scalable Semantic Systems, 2003.
 [6] H. Lausen, M. Stollberg, R. Hernández, Y. Ding, S. Han, and D. Fensel, “Semantic Web Portals – State of the Art Survey”, Technical Report (2004-04-03) of Digital Enterprise Research Institute (DERI), 2004.
 [7] Saltlux, “시맨틱 웹 기술 기반 추론 엔진 성능 평가”, KISTI Seminar (Closed), 2006.
 [8] Y. Theoharis, V. Christophides, and G. Karvounarakis, “Benchmarking Database Representations of RDF/S Stores”, In Proceedings of the 4th International Semantic Web Conference, 2005.
 [9] G. Williams, “MT-Redland: An RDF Storage Backend for Movable Type”, Workshop on Scripting for the Semantic Web (Collocated with ESWC 2005), 2005.
 [10] <http://triplestore.aktors.org/demo/AKTiveSpace/>