

학술연구정보 분석을 위한 OntoFrame[®] 추론 서비스 질의 설계

Designing Reasoning Service Queries of OntoFrame[®] for Analyzing Academic Research Information

이승우, 김평, 강인수, 정한민, 이미경, 성원경
한국과학기술정보연구원 정보서비스연구팀

Seungwoo Lee, Pyung Kim, In-Su Kang, Hanmin Jung,
Mikyong Lee, Won-Kyung Sung
ISRLab, KISTI

요약

논문이나 보고서, 특허 등의 연구 성과물에 대한 분석 정보는 미래를 위한 연구 정책의 방향을 설정하는 지표로서 중요할 뿐만 아니라, 연구자들간의 연구 성과물에 대한 활발한 정보 공유 및 유통과 협업을 위해서도 중요한 자료이다. 본 논문에서는 연구 성과물에 대한 분석 정보를 쉽고 정확하게, 그리고 체계적으로 제공하기 위한 한 방법으로 온톨로지에 기반한 지식화와 추론을 통한 지식의 확장을 통해 학술연구정보에 대한 분석 서비스를 제공하는 방법을 기술한다. 특히, 연구 성과물 정보로부터 얻을 수 있는 분석 정보를 정의하고, 정의된 온톨로지에 기반하여 가공된 지식으로부터 분석 정보를 얻기 위한 질의와 그 후처리 과정을 중심으로 소개하고자 한다.

Abstract

Analyzed information from research outcomes such as papers, reports and patents is important data as an indicator for setting up the direction of future research policy, as well as is essential for active sharing and dissemination of research information and cooperative work between researchers. In this paper, as a method for presenting the analyzed information easily, accurately and systematically, we describe a method that processes the information about research outcomes into ontology-based knowledge and expands the knowledge by reasoning. Especially, we focus on the definition of analyzed information from research outcomes and queries and postprocess for obtaining the analyzed information from the ontology-based knowledge.

I. 서론

논문이나 보고서, 특허 등의 연구 성과물에 대한 분석 정보는 미래를 위한 연구 정책의 방향을 설정하는 지표로서 중요하다. 특히, 연구 성과물에 대한 메타 데이터를 분석하게 되면, 최근에 어떤 분야의 연구가 활발한지, 그런 분야의 연구를 수행하는 연구자로는 누가 있는지, 어떤 연구자들이 서로 활발히 공동 연구를 수행하는지 등의 분석 정보를 얻을 수 있다. 이런 분석 정보는 연구 정책 수립을 위해서 뿐만 아니라, 연구자들 사이에 연구 정보를 서로 공유하거나 유통시키고, 나아가 연구자들 사이의 협업을 유도하는 데에 활용될 수 있다.

그러면, 이러한 분석 정보를 얻기 위해 연구 성과물에 대한 메타 데이터를 어떻게 처리할 것인가를 결정해야 할 것이다. 기본적으로 생각할 수 있는 방법은 정의된 분석 정보 서비스에 특화된 DB 스키마를 설계하고 이 스키마에 맞게 메타 데이터를 구축하는 방법이다. 그러나, 이 방법은 해당 분석 정보 서비스만을 위해서는 가장 효율적일 수 있으나, 서비스에 추가 또는 변경 요구가 있을 경우 DB 스키마를 변경해야 하는 등의 융통성이 부족하다는 단점이 생길 수 있다. 이에 대한 대안으로 제시할 수 있는 방법이 본 논문에서 이용하는 온톨로지와

시맨틱 웹 기술을 활용하는 방법이다.

온톨로지는 개념을 표현하는 클래스와 클래스 사이의 관계를 표현하는 속성으로 기술되는데, 연구 성과물에 대한 메타 데이터를 온톨로지로 지식화할 수 있다[1]. 또한, 시맨틱 웹 기술에 기반한 추론을 통해 온톨로지로 표현된 명시적 지식으로부터 암묵적인 지식을 유도할 수 있다. 뿐만 아니라, 규칙을 기술하여 추가적인 속성(관계)을 정의하고 확장할 수도 있다. 이렇게 연구 성과물의 메타 데이터를 온톨로지로 지식화하고 추론을 통해 확장한 다음에는 SPARQL과 같은 온톨로지 질의 언어를 사용하여 다양한 분석 정보를 얻을 수 있다.

본 논문에서는 연구 성과물에 대한 분석 정보를 쉽고 정확하게, 그리고 체계적으로 제공하기 위한 한 방법으로 온톨로지에 기반한 지식화와 추론을 통한 지식의 확장을 통해 학술연구정보에 대한 분석 서비스를 제공하는 방법을 기술한다. 특히, 연구 성과물 정보로부터 얻을 수 있는 분석 정보를 정의하고, 정의된 온톨로지에 기반하여 가공된 지식으로부터 분석 정보를 얻기 위한 SPARQL 질의와 그 후처리 과정을 중심으로 소개하고자 한다.

II. 학술 연구 정보의 지식화

연구 성과물에 대한 메타 데이터는 국가과학기술 R&D 기반정보 온톨로지[1]를 기반으로 지식화될 수 있다. 이 기반정보 온톨로지는 과학기술 연구 분야의 핵심 상위 객체들인 인력, 기관, 지역, 과제, 연구성과물, 연구영역, 게재지 등의 클래스를 갖고 있으며, 연구성과물은 논문, 저서, 특허, 보고서 등의 하위 클래스를 갖고, 게재지는 학술지와 학술대회논문집을 하위 클래스로 가지며, 연구영역은 주제영역과 분야영역을 하위 클래스로 갖는다. 또 하나의 클래스로 저작자 정보는 연구 성과물의 저작 당시의 저자에 대한 정보를 표현하는 클래스이다. 이상과 같은 기반정보 온톨로지의 일부를 그림 1에 나타내었다. 그림 1에서 실선 화살표는 객체속성 (object property)을, 점선 화살표는 subClassOf 관계를 가리키며, 데이터속성 (datatype property; 예를 들면, Person 클래스의 *engNameOfPerson*)은 지면 관계상 생략하였다.

그림 1은 기반정보 온톨로지의 스키마를 나타내며, 이 온톨로지의 인스턴스는 연구 성과물들의 메타 데이터로부터 생성된다. 즉, 논문의 경우, 논문의 제목, 저자 이름 및 저자의 소속 기관, 논문의 게재지 및 게재 연도, 논문의 주제 및 분야 등의 메타 데이터가 온톨로지 스키마 정의에 따라 URI 서버를 통해 주어-술어-목적어 형식의 트리플 형태로 지식화된다[3].

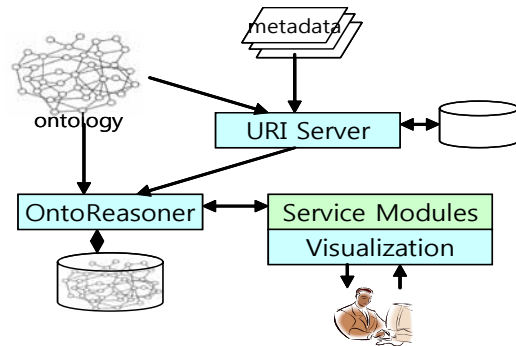


▶▶ 그림 1. 국가과학기술 R&D 기반정보 온톨로지

III. 정보유통 플랫폼 OntoFrame®

온톨로지를 기반으로 지식화된 학술연구정보는 추론 모듈인 OntoReasoner를 통해 암묵적인 지식과 규칙에 의한 지식을 확장한다. OntoReasoner는 또한 사용자의 SPARQL 질의에 대해 구축된 지식을 바탕으로 응답을 한다. 서비스 모듈 (service modules)은 학술연구 정보에 대한 다양한 분석 정보를 사용자에게 제공하기 위해 적합한 SPARQL 질의를 생성하고 OntoReasoner의 응답을 받아 후처리 과정을 통해 한번

더 가공한 후 시각화(visualization) 모듈에 전달하며, 시각화 모듈은 분석 정보를 사용자에게 보기 좋은 형태로 제시한다. 이러한 과정은 그림 2에 설명되어 있으며, 정보유통 플랫폼인 OntoFrame®[2]의 기반을 구성한다. 서비스 모듈에 대한 보다 자세한 설명은 4장에서 기술한다.



▶▶ 그림 2. 정보유통 플랫폼 OntoFrame®

IV. 학술연구정보 분석 서비스 질의 설계

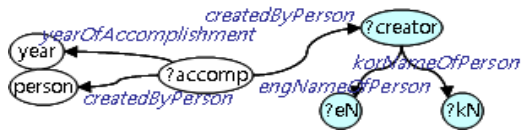
이 장에서는 기반정보 온톨로지에 따라 구축된 연구성과물에 대한 지식으로부터 얻을 수 있는 분석 정보와 이를 얻기 위한 SPARQL 질의 및 후처리 과정에 대해 기술한다. 여러 가지 가능한 분석 정보 중에서, 본 논문에서는 인력정보와 기관 정보, 지역정보, 주제정보, 성과물정보 등에 대한 분석 정보를 다루고자 한다.

1. 인력정보 분석

• 연구자별 연구자 네트워크

특정 연구자와 공저자 관계 혹은 인용 관계를 맺고 있는 연구자들을 찾는 서비스이다. 먼저, 공저자 관계의 경우, 특정 연구자(person)와 성과물의 저작연도(year)를 조건으로 하여 공저자 관계에 있는 연구자의 URI(?creator)와 영문(?eN) 혹은 한글이름(?kN)을 구할 수 있다. 이를 그래프 질의로 표현한 것이 그림 3)이며, 이에 대한 SPARQL 질의의 예를 그림 4에 보였다. SPARQL 질의 예에서 person과 year와 같은 상수는 FILTER 절로 처리되며, 관계가 없을 수도 있는 영문 및 한글 이름은 OPTIONAL 절로 처리된다.

1) '?'표시가 있는 노드는 변수를, 그렇지 않은 노드는 상수를 가리키며, 색깔 표시가 있는 변수 노드는 질의에서 얻고자 하는 값을 나타낸다. 속성의 namespace는 생략하였다.
2) 지면 관계상 이후의 서비스에 대한 SPARQL 예시는 생략한다.

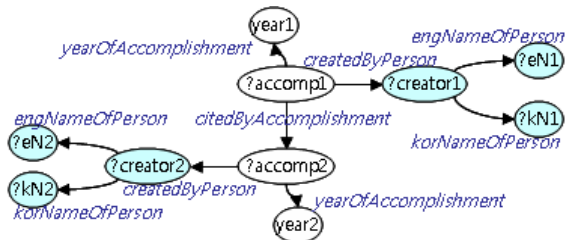


▶▶ 그림 3. 공저자 관계 네트워크-그래프 질의

```

SELECT ?creator ?eN ?kN
WHERE {
  ?outcome isrl:yearOfAccomplishment ?year .
  FILTER (?year>=2000 && ?year<=2007) .
  ?outcome isrl:createdByPerson ?person .
  FILTER(?person=isrl:PER_00000000000000000034) .
  ?outcome isrl:createdByPerson ?creator .
  FILTER (?creator!=isrl:PER_00000000000000000034) .
  OPTIONAL{?creator isrl:engNameOfPerson ?eN . }
  OPTIONAL{?creator isrl:korNameOfPerson ?kN . }
}
    
```

▶▶ 그림 4. 공저자 관계 네트워크-SPARQL 질의



▶▶ 그림 5. 인용 관계 네트워크-그래프 질의

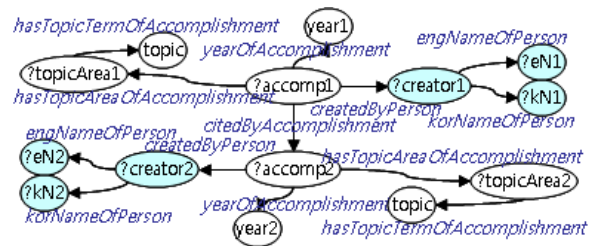
인용 관계의 경우, 연구 성과를 통해 특정 연구자(person)와 인용 혹은 피인용 관계에 있는 연구자의 URI(?creator1 혹은 ?creator2)와 영문 혹은 한글이름을 구할 수 있다. 이를 그래프 질의로 표현한 것이 그림 5이며, 여기서 ?creator1 혹은 ?creator2가 person에 해당할 때이다. 인용과 피인용의 구분은 후처리 과정에서 처리된다.

● 주제별 연구자 네트워크 및 그룹

특정 주제(topic)와 연도(year)를 조건으로 성과물 기반의 공저자 관계 혹은 인용 관계를 맺고 있는 연구자 쌍(?creator1, ?creator2)을 구한다. 공저자 관계의 경우(그림 6), 이 쌍은 저자순위(?order1<?order2)에 의한 순서를 구분하며, 인용 관계의 경우(그림 7), 인용과 피인용에 의한 순서를 구분한다. 연구자 그룹은 네트워크를 탐색하는 후처리 과정을 통해 얻어진다.



▶▶ 그림 6. 주제별 연구자 네트워크 - 공저자 관계



▶▶ 그림 7. 주제별 연구자 네트워크 - 인용 관계

● 주제별 연구 전문가

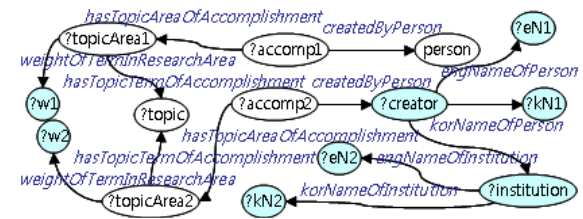
특정 주제(topic)를 연구하는 연구자(?creator)와 소속기관(?institution)을 얻고(그림 8), 후처리 과정에서 해당 주제에 대한 연구 성과물 건수를 기준으로 전문도를 계산하여 누가 해당 주제에 대한 연구 전문가인지 분석할 수 있다.



▶▶ 그림 8. 주제별 연구 전문가

● 주제기반 유사 연구자

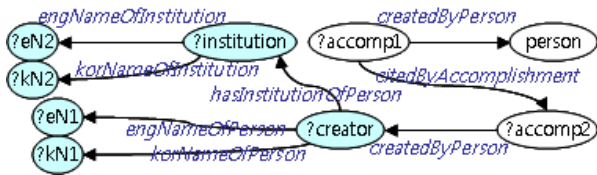
특정 연구자(person)와 동일한 주제(?topic)를 연구하는 연구자(?creator)와 현재 소속기관(?institution)을 분석할 수 있다. 또한, 후처리 과정에서는 주제에 대한 가중치(?w1, ?w2)를 기준으로 유사도를 계산하여 정렬을 제공한다(그림 9).



▶▶ 그림 9. 주제기반 유사 연구자

● 인용 및 피인용 연구자

특정 연구자(person)의 성과를 인용하는 연구자(?creator)와 현재 소속기관(?institution)을 그림 10에서와 같은 질의를 통해 얻을 수 있다. 역으로, 'citedByAccomplishment' 속성의 화살표 방향을 바꾸면 특정 연구자가 인용하고 있는 연구자(피인용 연구자)도 구할 수 있다.



▶▶ 그림 10. 인용 연구자

• 연구자의 연도/주제별 성과물 통계

특정 연구자(person)가 몇 연도(?year)에 어떤 주제(?topic)에 대해 몇 건의 성과물(?accomp)을 저작하였는지를 분석할 수 있다(그림 11). 성과물의 건수를 세는 부분은 후처리 과정에서 처리된다.

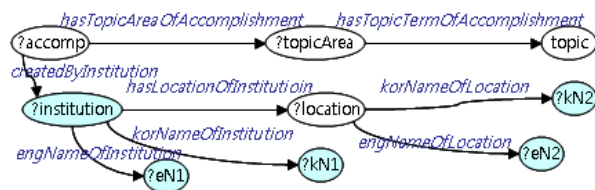


▶▶ 그림 11. 연구자의 연도/주제별 성과물

2. 기관정보 분석

• 주제별 연구 전문 기관

특정 주제(topic)에 대한 연구 전문 기관이란 그 주제에 관한 연구 성과물을 많이 도출한 기관이라 할 수 있다. 따라서, 특정 주제에 관한 연구 성과물의 저자가 속한 기관(?institution)을 찾으면 된다(그림 12). 후처리 과정에서는 그러한 성과물의 건수를 기관별로 센다.

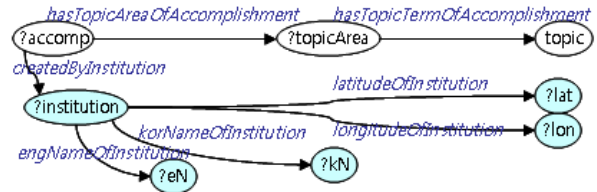


▶▶ 그림 12. 주제별 연구 전문 기관

3. 지역정보 분석

• 주제별 연구 지역

주제별 연구 전문 기관을 분석하는 것과 같은 방법으로 기관이 위치한 지역의 GPS좌표(?lat, ?lon)를 이용하여 연구 전문 지역을 분석할 수 있다(그림 13). 연구 지역을 분석하는 단위는 그림 13에서처럼 기관(?institution)이 될 수도 있고, 연구자나 성과물이 될 수도 있다. 이렇게 분석된 정보는 구글맵과 같은 mash-up 서비스와 연동하여 지도 위에 표시하는 것도 가능하며, 시각화 모듈에서 처리된다.

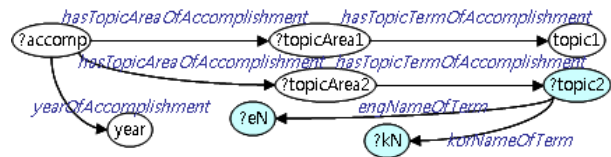


▶▶ 그림 13. 주제별 지역의 연구 기관

4. 주제정보 분석

• 연관 주제

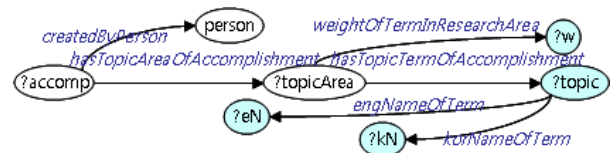
연관 주제는 특정 주제(topic1)와 함께 거론되어 연관성을 갖는 주제들(?topic2)을 찾는 분석 서비스를 말한다. 연구 성과물은 대개 하나 이상의 주제를 담고 있으며, 한 성과물에서 다루고 있는 주제들 사이에는 연관성이 있다고 가정할 수 있다. 이를 바탕으로 연관 주제를 찾는 질의는 그림 14와 같다. 연도(year) 범위를 제약하여 연도별로 연관 주제의 추이를 살펴볼 수도 있다. 주제 사이의 연관성 정도를 계산하기 위해 각 주제 쌍이 나타나는 회수를 세는 것은 후처리 과정에서 처리된다.



▶▶ 그림 14. 연관 주제

• 연구자의 연구 주제

특정 연구자(person)가 주로 연구하는 주제가 무엇인지를 분석하는 서비스로, 그림 15와 같이, 그 연구자가 저작한 성과물(?accomp)의 주제(?topic)를 분석함으로써 얻을 수 있다. 후처리 과정에서는 각 주제의 빈도수를 계산하여 주제를 순위화할 수 있다.

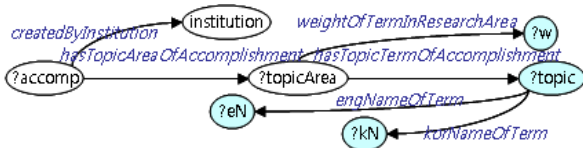


▶▶ 그림 15. 연구자의 연구 주제

• 기관의 연구 주제

연구자의 연구 주제는 그 연구자의 저작 당시의 소속기관의 연구 주제로 전이될 수 있다. 따라서, 특정 기관(institution)의 연구 주제는 그 기관에서 저작된 성과물(?accomp)의 주제(?topic)를 분석함으로써 얻을 수 있다(그림 16). 후처리 과정

에서는 각 주제의 빈도수를 계산하여 주제를 순위화할 수 있다.



▶▶ 그림 16. 기관의 연구 주제

• 주제별 연구 통계

각 주제가 연도별로 얼마나 연구되고 있는지를 분석한다. 특정 주제(topic)를 다루고 있는 연구 성과물(?accomp)이 연도(?year)별로 나타나는 건수를 그림 17과 같은 질의를 통해 얻을 수 있다. 여기서, 건수를 세는 부분은 후처리 과정이 맡는다.



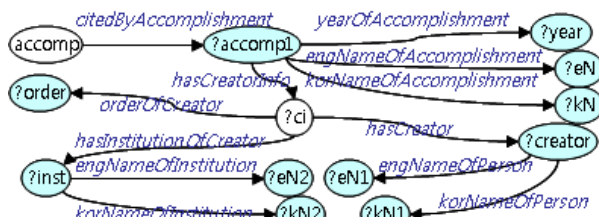
▶▶ 그림 17. 주제의 연도별 성과물 통계

5. 성과물정보 분석

• 인용 및 피인용 성과물

각 연구 성과물을 인용하고 있는 또는 역으로 각 연구 성과물이 인용하고 있는 연구 성과물 정보(연도, 저자 및 소속 기관 포함)를 분석할 수 있다. 그림 18은 전자의 경우(인용 관계)를 설명하고 있으며, 후자의 경우(피인용 관계)는 'citedByAccomplishment' 속성의 화살표 방향을 바꾸면 얻어질 수 있다.

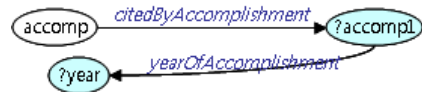
그림 18에서 보는 바와 같이, 특정 연구 성과물(accomp)을 인용하고 있는 성과물(?accompl)과 저작 연도(?year)를 얻고, 다시 그 성과물의 저자(?creator)와 저자 순위(?order), 저작 당시의 소속 기관(?inst)을 얻을 수 있다. 여기서, 저자 순위는 성과물의 저자가 둘 이상일 때, 저자 순위에 맞게 제시하기 위해서 필요한 정보이다.



▶▶ 그림 18. 인용 성과물 및 저자, 소속기관

• 성과물의 피인용 통계

연구 성과물의 가치를 평가하는 잣대로 흔히 사용되는 피인용 지수를 계산할 수 있다. 그림 19와 같이, 특정 성과물(accomp)에 대해, 연도(?year)별로 이를 인용하는 성과물(?accompl)의 건수를 얻음으로써 가능하다. 여기서, 건수를 세는 부분은 후처리 과정이 맡는다.



▶▶ 그림 19. 연도별 인용 성과물

V. 결 론

지금까지, 학술연구정보에 대한 다양한 분석 정보를 제공하기 위한 OntoFrame® 기반의 질의 설계와 후처리에 대해 설명하였다. 연구 성과물에 대한 메타 데이터는 온톨로지 기반으로 지식화되고 추론을 통해 그 지식을 확장된다. 이렇게 구축된 지식으로부터 SPARQL 질의와 그 후처리를 통해 다양한 분석 정보를 제공할 수 있었다. 앞으로는 여기서 설계된 SPARQL 질의를 보다 효율적으로 처리하여 응답 속도를 개선하는 방안을 모색할 필요가 있다.

■ 참 고 문 헌 ■

- [1] 강인수, 정한민, 이승우, 김평, 성원경, “국가과학기술 R&D 기반 정보 온톨로지와 추론 모델링”, 2006 한국컴퓨터종합학술대회 발표논문집, 2006.
- [2] 정한민, 이미경, 성원경, 박동인, “OntoFrame-K: 연구자 간 협업 지원 서비스를 위한 시멘틱 웹 기반 정보 유통 플랫폼”, 2006 한국컴퓨터종합학술대회 발표논문집, 2006
- [3] 정한민, 강인수, 구희관, 이승우, 성원경, “URI 서버에 기반한 국가 R&D 기반정보 온톨로지 설계 및 구현”, 정보관리연구 37권 2호, pp109-136, 2006.