ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2015 December: 21(4): 79~92 http://dx.doi.org/10.13088/jiis.2015.21.4.079

온톨로지와 토픽모델링 기반 다차원 연계 지식맵 서비스 연구

정한조
한국과학기술정보연구원 (KISTI), 첨단정보융합본부, NTIS 센터
(hanjo.jeong@kisti.re.kr)

미래 핵심 가치 기술 발굴 및 탐색을 위해서는 범국가적인 국가R&D정보와 과학기술정보의 연계·융합이 필요하다. 본 논문에서는 국가R&D정보와 과학기술정보를 온톨로지와 토픽모델링을 사용하여 연계·융합하여 지식베이스를 구축한 방법론을 소개하고, 이를 기반으로 한 다차원 연계 지식맵 서비스를 소개한다. 국가R&D정보는 국가R&D과제와 참여인력, 해당 과제에 대한 성과 정보, 논문, 특허, 연구보고서 정보들을 포함한다. 과학기술정보는 논문, 특허, 동향 등의 과학기술 연구에 대한 기술 문서를 일컫는다. 본 논문에서는 지식베이스에서의 지식 처리 및 관리의 효율성을 높이기 위해 Lightweight 온톨로지를 사용한다. Lightweight 온톨로지는 국가R&D과제 참여자와 성과정보, 과학기술정보를 과제-성과 관계, 문서-저자 관계, 저자-소속기관 관계 등의 단순한 연관관계를 이용하여 국가R&D정보와 과학기술정보를 융합한다. 이러한 단순한 연관관계만을 이용함으로써 지식 처리의 효율성을 높이고 온톨로지 구축 과정을 자동화한다. 보다 구체적인 Concept 레벨에서의 온톨로지 구축을 위해 토픽모델링을 활용한다. 토픽모델링을 활용하여 국가R&D정보와 과학기술정보 문서들의 토픽 주제어를 추출하고 각 문서 간 연관관계를 추출한다. 일반적인 Concept 레벨에서의 Fully-Specified 온톨로지를 구축하기 위해서는 거의 100% 수동으로 해야 하기 때문에, 많은 시간과 비용이 소모된다. 본 연구에서는 이러한 수동적인 온톨로지 구축이 아닌 자동화된 온톨로지 구축을 위해 토픽모델링을 활용한다. 토픽모델링을 활용하여 온톨로지 구축에 필요한 문서와 토픽 키워드 간의 관계, 문서 간 의미 상 연관관계를 자동으로 추출한다. 마지막으로, 이와 같이 구축된 지식베이스의 트리플(Triple) 정보를 활용하여, 연구자들의 공동저자관계, 문서간의 공통주제어관계 등을 연구자, 주제어, 기관, 저널 등의 다차원 연관관계를 방사형 네트워크 형식을 이용하여 시각화한 지식맵 서비스들을 소개한다.

주제어 : 온톨로지, 토픽모델링, 지식베이스, 지식맵, 정보융합

논문접수일 : 2015년 11월 25일 논문수정일 : 2015년 12월 11일 게재확정일 : 2015년 12월 12일
교신저자 : 정한조

1. 개요

지식맵 (Knowledge Map)은 지식 분류 체계의 한 방법이라고 정의될 수 있다. 기존의 지식 기반 시스템은 주로 카테고리 (Category)나 택사노미 (Taxonomy) 같은 전통적인 지식 분류 체계를 이용한다. 이러한 전통적인 지식 분류 체계는 맥락 (Context)에 따라 변할 수 있는 지식을 고정적으로만 분류한다는 데에 문제점이 있다. 반면, 지식맵은 지식을 네트워크 형태의 유연한 구조로 표현하기 때문에 다양한 맥락 상에서의 지식을 표현할 수 있다. 또한, 지식맵은 네트워크와

* 본 연구는 서울과학기술대학교의 교내연구비로 수행되었음.

맵 형태로 지식을 표현할 수 있는 사용자 인터페이스 (User Interface) 또는 시각화 (Visualization) 체계로 정의될 수 있다(Howard, 1989; McCagg and Dansereau, 1991; Eppler, 2001; Kang et al., 2003). 지식은 일반적으로 지식관리시스템 (Knowledge Management System)에 처리되고 관리된다. 지식맵 기반의 지식관리시스템은 사용자에게 지식의 상호 유기적인 관계를 이용한 지식 네비게이션 (Navigation)을 가능하게 한다(Rao et al., 2012).

지식맵은 일반적으로 두 개의 형태로 분류된다. 첫번째는 한 기업이나 기관의 데이터를 지식 기반으로 저장/관리/처리 하는 일반적인 지식관리시스템 (Knowledge Management System) 에서 사용되는 형태이고, 두번째는 과학기술지식정보를 분석하고 표현하기 위해 사용된 형태이다. 지식관리시스템에서 사용되는 지식맵은 일반적으로 Business Process의 효율성을 높이기 위해 기업이나 기관의 내부 데이터나 프로세스를 표현하는 데 집중하는 반면(Businska et al., 2013), 과학기술지식정보를 분석 및 표현하는 용도의 지식맵은 주로 과학기술지식정보를 사용자가 효율적으로 직접 네비게이션 할 수 있는 형태의 구조를 설계하는 것에 집중된다(Klavans and Boyack, 2009; Leydesdorff and Rafols, 2009). 본 논문에서는 후자인 과학기술지식정보를 효율적이고 효과적으로 분석하고 표현할 수 있는 방사형 네트워크 기반의 지식맵 방법을 소개한다.

RDF(Resource Description Framework)(W3C RDF Working Group, 2014), RDFS (RDF Schema)(Brickley and Guha, 2014), OWL(Web Ontology Language)(W3CRDF Working Grou p,2012)같은 W3C (World Wide Web Consortium)에 의해 표준화된 시맨틱 웹 (Semantic Web) 기술 및 프레임워크로 인해, 보다 효율적이고 자동화 (Machine-Processible)된 방법으로 지식을 표현하고 처리할 수 있게 됐다. 그러나 이러한 Machine-Readable하고 표준화된 형태 (Format)의 문서는 일반적으로 기계가 읽고 처리하기 수월하고, 오히려 인간에게는 자연어 (Free-Text) 형태의 문서보다 가독성과 작성 용이성이 떨어진다. 게다가, 대부분의 사용자들은 이러한 시맨틱 데이터로부터 지식을 추출하기 위한 SPARQL (SPARQL Protocol and RDF Query Language) (Prud'hommeaux and Seaborne, 2008) 같은 질의를 사용하는 데 어려움이 있다. 특히, 인터넷 같은 분산 환경에서 각각의 도메인에 대한 온톨로지의 스키마와 Vocabulary를 숙지하고 있어야 가능한 일이기 때문이기도 하다. 지식맵은 사용자들에게 익숙하고 보다 자연스러운 네트워크 형태로 지식을 표현할 수 있기 때문에 사용자의 지식 활용성을 높일 수 있다.

본 연구에서는 온톨로지와 토픽모델링을 활용하여 국가 R&D 과제, 논문, 특허, 연구보고서 같은 국가 R&D 데이터를 연계·융합하고, 이를 트리플 데이터로 변환하여 지식베이스를 구축하는 지식 처리·관리 시스템을 소개한다. 추가로, 이러한 트리플 데이터를 활용하여 지식을 시각화하고 사용자의 지식 네비게이션의 편의성을 향상시킬 수 있는 네트워크 형태의 다차원 지식맵 서비스를 소개한다.

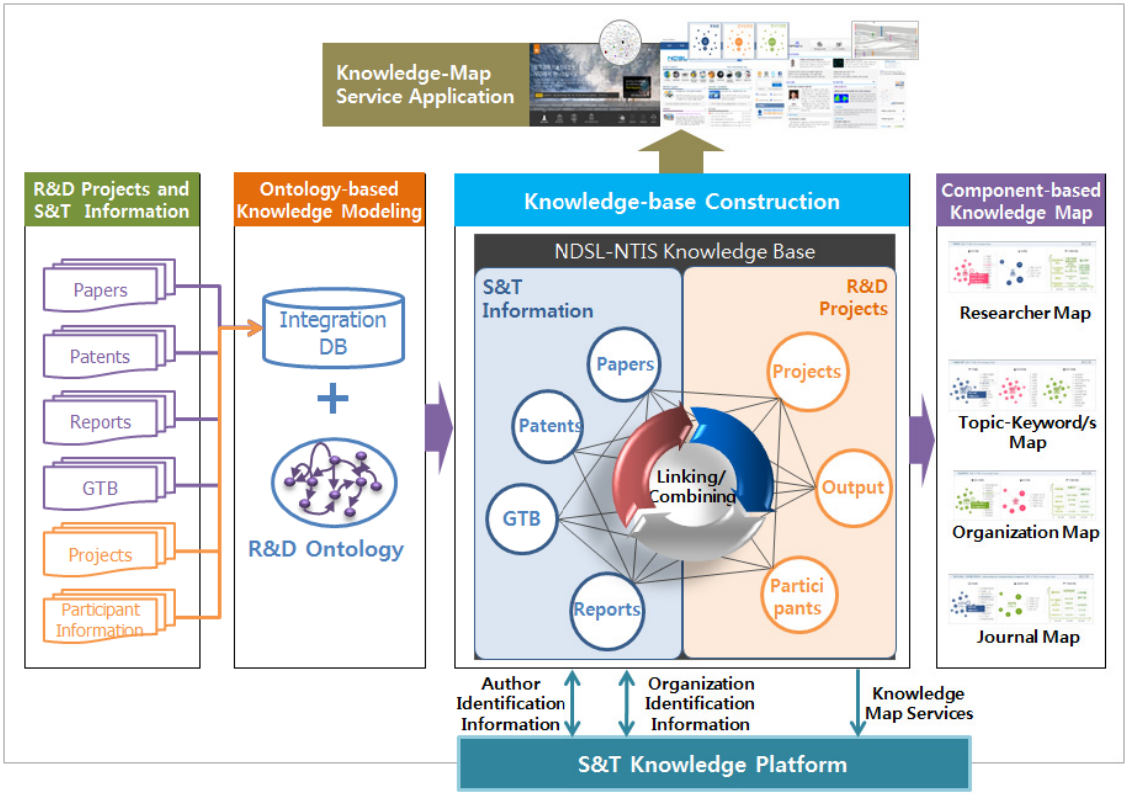
2. 온톨로지, 토픽모델링 기반 지식베이스

본 논문에서는 연구과제, 논문, 특허, 연구보고서 등의 국가 R&D 데이터에 대하여 온톨로지

와 토픽모델링을 사용하여 구축한 지식 기반의 지식맵 서비스 시스템을 소개한다. 이러한 시스템은 아래와 같은 목표를 가지고 있다.

1) 국가과학기술정보서비스 (NTIS: National Science & Technology Information Service)의 국가 R&D 데이터와 국가과학기술정보 제공 플랫폼 (NDSL: National Digital Science Library)의 과학기술정보를 융합하여 종합적인 정보를 제공한다.
2) 융합된 데이터에 대해, 의미 또는 토픽 기반의 검색을 제공한다.
3) 마지막으로, 시맨틱 분석과 지식 처리를 기반으로 지식맵 서비스를 제공한다.

<Figure 1>은 전체 시스템의 개요를 도식화한 것이다. R&D 과제 및 연구성과정보 등의 국가 R&D 정보는 NTIS로부터 주기적으로 업데이트되고, 연구 논문, 특허, 연구 보고서 등의 과학기술정보는 NDSL의 e-Gate를 통해 매일 업데이트된다. 업데이트 된 국가 R&D 정보와 과학기술정보는 통합 DB에 먼저 저장되고, S&T 지식 플랫폼의 저자 및 기관 등의 개체 식별 데이터를 통해 정제된다. 정제된 데이터는 토픽모델링과 국가 R&D 온톨로지 스키마를 통해 RDF 기반의 트리플 (N-Triple) 데이터로 변환된다. 마지막으로, 트리플 데이터로 형성된 지식베이스를 기반으로 컴포넌트 형태의 지식맵 서비스를 제

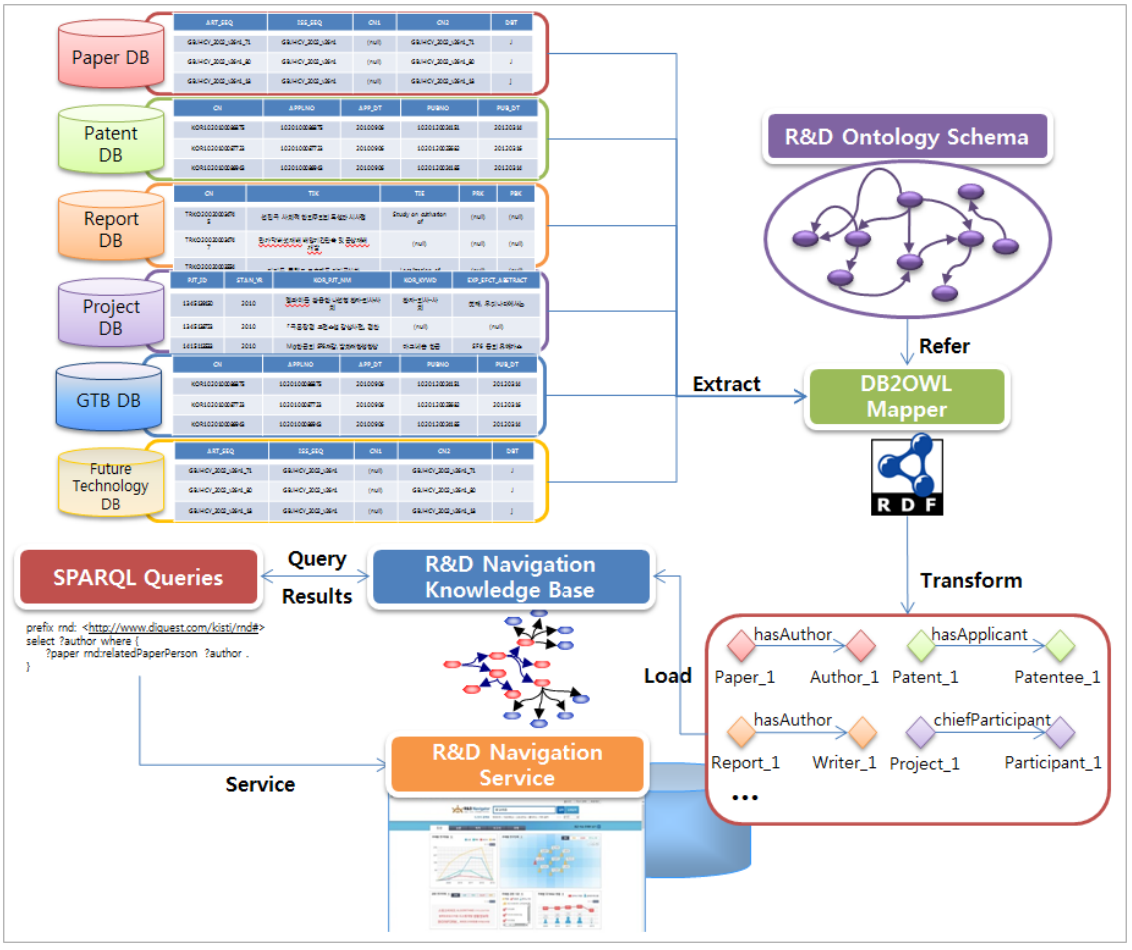


〈Figure 1〉 System Overview

공한다.

2.1. 지식베이스 생성

본 논문에서 온톨로지는 지식맵에서 국가 R&D 정보의 개체와 개체간의 관계를 표현하는데 이용된다. 토픽모델링을 사용하여 추출된 개체와 주제어 간의 연관관계 및 개체 간의 상호 유사도 기반의 연관관계도 표현한다. 이러한 단순한 연관관계를 표현하기 위해서 본 연구에서는 시스템의 효율성을 위해 Lightweight 온톨로지를 사용한다. Lightweight 온톨로지는 메타데이터와 같이 일반 온톨로지보다 상위적이고 일반적인 클래스와 관계로 데이터를 표현하여 온톨로지 데이터 구축과 처리의 효율성을 극대화한 것이다(Ahmad and Colomb, 2007; Morbach et al., 2009). <Figure 2>는 온톨로지 트리플 데이터가 어떻게 생성되고 처리되는 지의 과정을 표현한다. 먼저, 국가 R&D 정보와 과학기술정보가

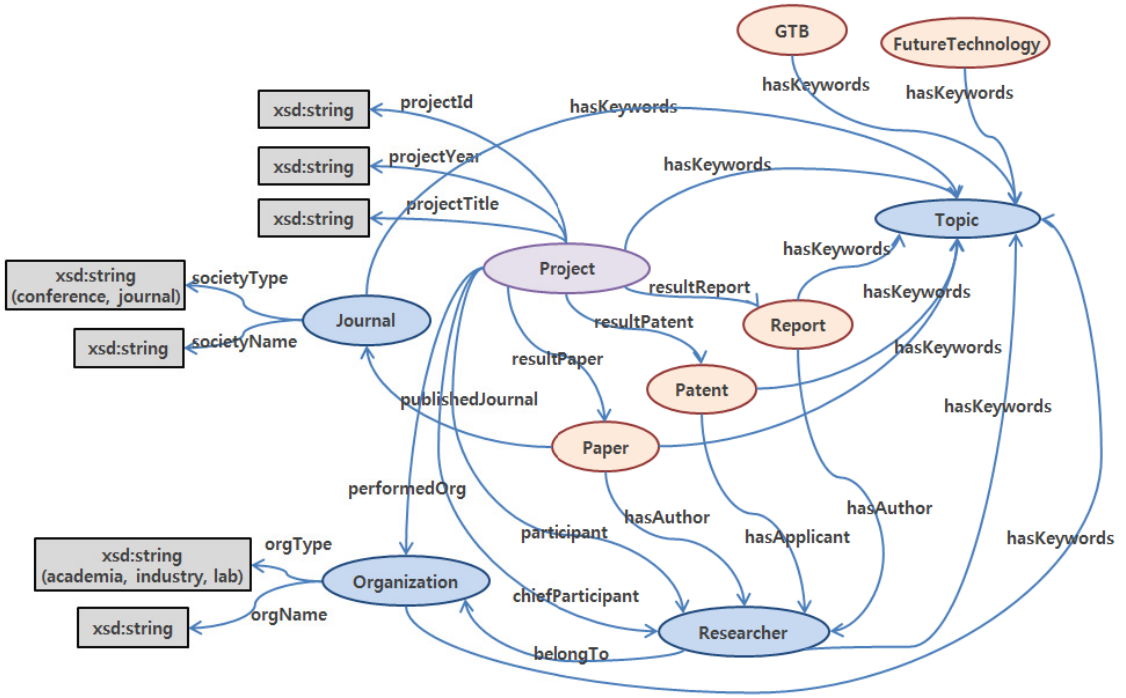


〈Figure 2〉 Construction Process of Knowledge Base

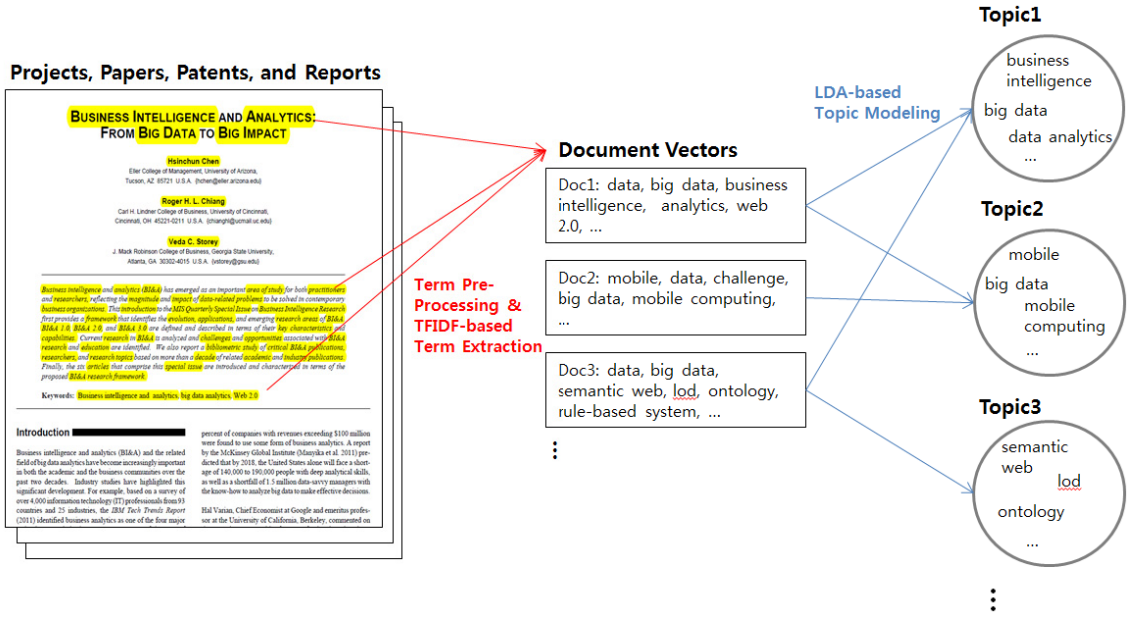
각각 NTIS와 NDSL에서 업데이트 되고, 이를 기반으로 통합 DB를 생성한다. 통합 DB로부터 국가 R&D 온톨로지 스키마와 DB2OWL Mapper를 사용하여 관계형 (Relational) 데이터를 트리플 데이터로 변환한다. 변환된 트리플 데이터는 온톨로지 저장소인 R&D 네비게이션 지식베이스로 적재된다. 마지막으로, 지식맵 서비스는 SPARQL Query Endpoint를 통해 필요한 트리플 데이터를 추출하여 사용한다.

2.2. 온톨로지 모델링

온톨로지 모델을 표현하기 위해서 W3C의 시맨틱 웹 표준 프레임워크 및 언어인 RDF/s 와 OWL을 사용한다. 지식을 효율적으로 처리하기 위해서 상위 개념 (High Level) 개체와 연관관계만을 표현한 Lightweight 온톨로지 형태로 온톨로지 모델을 설계한다. <Figure 3>은 국가 R&D 과제를 중심으로 과제의 연구성과로 연결된 과학기술정보, 연구자, 주제어 중심의 연관관계를 표현한 국가 R&D 온톨로지 모델을 나타낸다. Project 클래스는 국가 R&D 과제를 나타낸다. Paper 클래스, Patent 클래스, Report 클래스는 각각 논문, 특허, 연구보고서를 나타내고, Project 클래스와 국가 R&D 과제의 성과물로 연결된다. Researcher 클래스는 연구자를 나타내고 저자, 참여자 등의 역할 관계로 다른 국가 R&D 데이터와 연결된다. 또한, Researcher 클래스는 연구자의 소속기관을 나타내는 Organization 클래스와도 연결 된다. Organization 클래스는 학계, 산업계, 연구계 타입으로 분류된다.



〈Figure 3〉 National R&D Ontology Model



〈Figure 4〉 Topic Modeling-based Topic Extraction

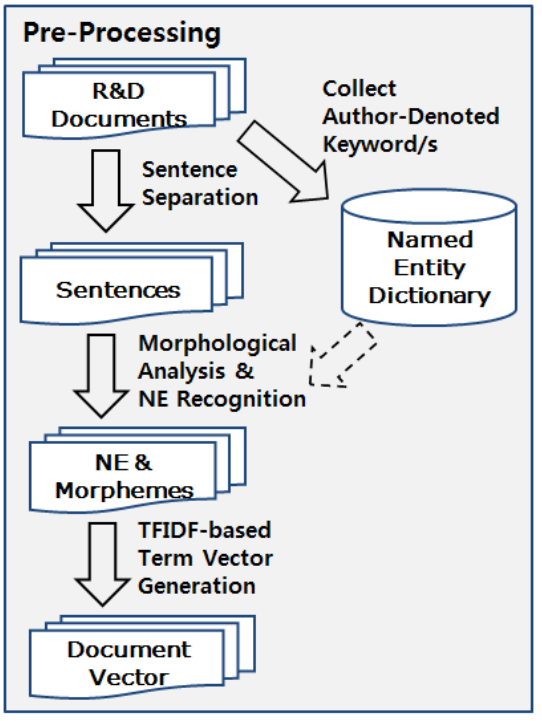
3. 토픽모델링 기반 토픽 추출

본 연구에서는 국가 R&D 문서의 주제어를 추출하기 위해 토픽모델링 방법을 이용한다. 토픽모델링을 수행하기 위해 <Figure 4>에 표현한 것처럼, 국가 R&D 원본 문서로부터 텍스트 전처리 과정 (Pre-Processing)을 먼저 수행한다. 전처리 과정에서 생성된 문서 벡터를 기반으로 토픽모델링의 대표적이고 효과적인 알고리즘인 Latent Dirichlet Analysis (LDA)(Blei et al., 2003)를 이용하여 각 문서와 연관된 토픽을 추출하고 이를 기반으로 코사인 유사도를 이용해서 주제어 및 문서, 단어 간 연관관계를 생성한다.

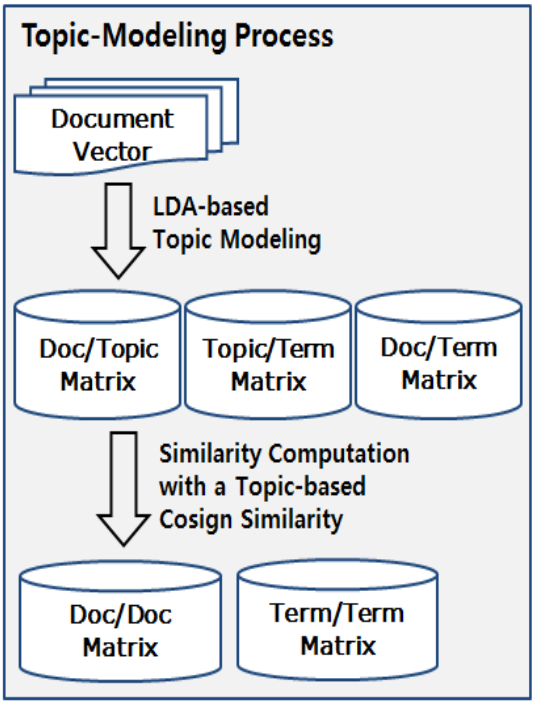
3.1. 텍스트 전처리 과정 (Pre-processing)

<Figure 5>에서 표현한 것처럼, 토픽모델링을 수행하기 위해서는 R&D 데이터가 텍스트 형태이므로 텍스트 전처리 과정이 필요하다. 먼저, 의미 있는 단위의 주제어를 추출하기 위해 연구자가 직접 입력한 키워드를 추출하여 개체명 사전을 구축한다. 예를 들어 “사물 인터넷”과 같은 복합어를 하나의 주제어로 추출하기 위해서는 개체명 사전에 해당 복합어가 추가되어있어야 한다. 그렇지 않으면, “사물”과 “인터넷”도 각각 의미 있는 형태소이므로 분리되어 추출될 수 있다. 이러한 형태소 분석과정에서의 복합어 처리를 통해, 과제, 논문의 본문 텍스트에서의 보다 의미 있는 단위의 주제어를 추출할 수 있다.

아래와 같이 추출 된 주제어를 기반으로 Term Frequency-Inverse Document Frequency (TFIDF) (Salton and Mcgill, 1986) 기반의 문서 벡터를 생성한다. Equation (1)은 문서 j에 대한 단어 i의 가중치($w_{i,j}$)를 TFIDF를 사용하여 계산



〈Figure 5〉 Pre-Processing



〈Figure 6〉 LDA-based Topic Modeling

한 식을 보여준다. $tf_{i,j}$ 는 문서 j에 나타나는 단어 i의 빈도를 나타내고, df_i 는 단어 i를 포함하는 문서의 수를 나타낸다. N은 전체 문서의 수를 나타낸다.

$$w_{i,j} = tf_{i,j} \times \log(\frac{N}{df_i}) \qquad (1)$$

3.2. LDA 기반 토픽모델링

<Figure 6>에 표현한 것처럼, 전처리 과정에서 생성한 문서 벡터를 이용하여, LDA 기반의 토픽 모델링을 활용하여 국가 R&D 데이터로부터 토픽 주제어를 추출한다. 추가로, LDA 결과로 생성된 행렬을 이용하여 문서간, 주제어간 유사도 행렬을 생성한다.

Equation (2)에 나타낸 것처럼, 토픽모델링은 문서 d에 단어 w가 나타날 조건부 확률 $p(w|d)$을 독립 다항 분포로 가정하고 은닉 변수 (hidden variable)인 토픽 z를 추가하여 단어-토픽 간 조건부 확률 $p(w/z)$와 토픽-문서 간 조건부 확율 $p(z/d)$로 분해하여 $p(w/d)$를 추정하는 방법론이다.

$$p(w \mid d) = \sum_{z} p(w \mid z)p(z \mid d) \qquad (2)$$

Equation (2)의 $p(w/d)$를 추정하기 위해서 샘플링을 이용한 대표적인 생성모델인 LDA를 사용하였다. 또한, LDA는 토픽모델링의 대표적인 알

고리즘 중의 하나인 Latent Semantic Indexing (LSI) (Hofmann, 1999)의 사전 확률 (prior probability)을 Dirichlet 분포로 가정한 확률적 모델과 동일하다고 볼 수 있다(Blei, 2012). 본 연구에서는 LDA를 수행하기 위해, Apache Hadoop 기반의 병렬처리를 지원하는 Apache Mahout 학습 라이브러리를 사용하였다. <Figure 6>에서 표현한 것처럼, LDA 수행 결과로 발생한 문서-토픽, 토픽-단어 행렬을 기반으로 문서-문서, 단어-단어의 유사도 행렬을 토픽 기반의 코사인 유사도를 이용하여 계산한다. 문서-문서 간의 유사도는 Equation (3)에 표현한 것처럼, 문서-토픽 행렬을 이용하여 토픽 기반의 문서 벡터를 생성한 후 코사인 유사도를 계산한다. D_i 와 D_j 는 각각 문서 i와 j의 토픽 가중치를 이용해서 생성한 문서벡터이다. 토픽 가중치는 LDA 결과로 생성된 문서-토픽 행렬의 원소 값을 이용한다.

$$\text{Doc}-\text{Doc Similarity} = \cos(\theta) = \frac{D_i \cdot D_j}{\|D_i\|\|D_j\|}, i \neq j \quad (3)$$

주제어를 나타내는 단어-단어 간의 유사도 행렬도 문서-문서 간의 유사도 행렬과 같이 Equation (4)에 표현한 것처럼, 토픽-단어 행렬을 이용하여 토픽 기반의 단어 벡터를 생성한 후 코사인 유사도를 계산한다. T_k 와 T_l 는 각각 단어 k와 l의 토픽 가중치를 이용해서 생성한 단어벡터를 나타낸다. 마찬가지로, 토픽 가중치는 LDA 결과로 생성된 토픽-단어 행렬의 원소 값을 이용한다.

$$\text{Term}-\text{Term Similarity} = \cos(\theta) = \frac{T_k \cdot T_l}{\|T_k\|\|T_l\|}, k \neq l \quad (4)$$

4. 지식맵 서비스

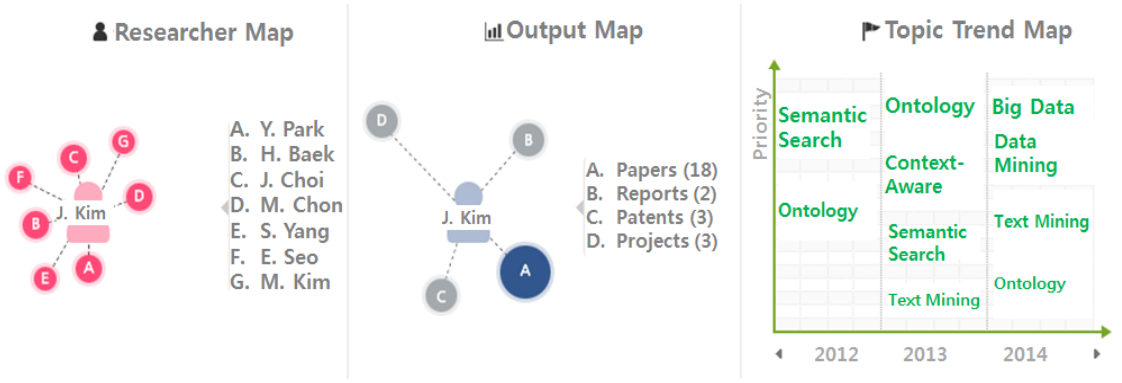
본 섹션에서는 온톨로지와 토픽모델링을 활용하여 생성한 지식베이스를 기반으로 효율적으로 지식을 시각화하고 사용자들의 효과적인 지식 네비게이션을 가능하게 하는 지식맵 서비스를 소개한다. <Table 1>은 지식베이스를 구축하는 사용된 국가 R&D 데이터와 생성된 트리플 데이터의 통계 정보를 나타낸다. 지식베이스의 트리플 데이터를 기반으로 연구자, 주제어, 연구기관, 학회 등의 다차원 기반의 지식맵 서비스를 생성한다.

4.1. 연구자 중심의 지식맵 서비스

국가 R&D 온톨로지 모델에서 연구자들은 논문, 특허 등의 공동 저자나 과제의 공동참여자

〈Table 1〉 Status of Knowledge Base Construction for National R&D Data

Data Type	Number of Source Data	Number of Triples
Research Projects	About 280,000	About 9M
Papers	About 1.5M	About 50M
Patent	About 2.5M	About 80M
Research Reports	About 90,000	About 5.7M
GTB (Global Trends Briefing)	About 160,000	About 5.4M
Researcher Profiles	About 390,000	About 2M



〈Figure 7〉 Researcher-Centric Knowledge Map Component

등에서 추출된 공동 저자 및 참여자 관계로 연결되어 있다. 이러한 연구자 네트워크를 표현하기 <Figure 7>과 같이 연구자 중심의 지식맵 컴포넌트를 생성하였다. 연구자 중심의 지식맵 서비스는 아래와 같이 3 종류의 지식맵으로 구성되었다.

1) 연구자의 공동저자, 공동참여자 관계를 기반으로 생성한 연구자 네트워크,
2) 선택된 연구자의 국가 R&D 연구성과정보,
3) 마지막으로, 선택된 연구자의 년 도별 국가 R&D 연구성과정보에서 추출된 주제어 기반의 트렌드 분석 정보.

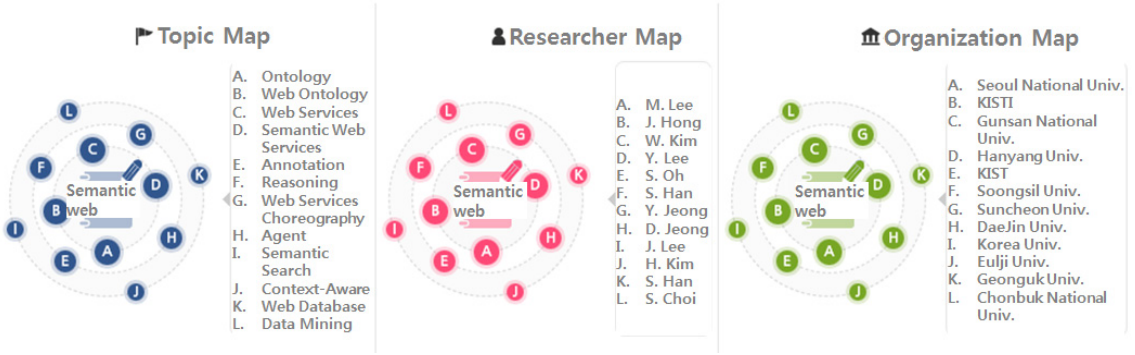
<Figure 7> 연구자 중심의 지식맵 컴포넌트에서 좌측의 연구자 맵에서 각각의 노드는 연구자를 나타내고 사용자는 각 노드를 클릭한 후 해당 연구자의 국가 R&D 연구성과정보를 보거나 선택 된 연구자를 중심으로 연구자 맵을 재생성할 수 있다. 추가로, 선택된 연구자와 연구자 맵의 중심 연구자와의 공동 연구성과정보도 볼 수 있다. 중앙의 연구성과정보 맵에서는 해당 연구자의 연구성과정보를 종류에 따라 검색할 수 있다. 마지막 트렌드 맵에서는 각 주제어를 중심으로 다음 섹션에 소개되는 주제어 중심의 지식맵으로 이동할 수 있다.

4.2. 주제어 중심의 지식맵 서비스

토픽모델링 결과로 추출된 주제어는 국가 R&D 온톨로지 모델에서 국가 R&D 데이터에 공동 주제로서 서로 연결되어 있다. 이러한 주제어 간의 연관관계를 <Figure 8>과 같이 주제어 중심의 지식맵 컴포넌트로 생성하였다. 주제어 중심의 지식맵 서비스는 아래와 같이 3 종류의 지식맵으로 구성되었다.

1) 선택된 주제어와 연관된 주제어 지식맵,
2) 선택된 주제어와 관련된 연구성과의 상위 연구자 지식맵,
3) 마지막으로, 선택된 주제어와 관련된 연구성과의 상위 연구 기관 지식맵.

연구자 맵과 같이 좌측의 주제어 맵에서 각각의 노드는 주제어를 나타내고 사용자는 각 노드를 클릭한 후 해당 주제어와 연관된 국가 R&D 연구성과정보를 보거나 선택 된 주제어를 중심으로 주제어 맵을 재생성할 수 있다. 선택된 주제어, 주제어 맵의 중심 주제어와 공통으로 연관된 연구성과정보도 볼 수 있다. 중앙의 연구자



〈Figure 8〉 Topic-Centric Knowledge Map Component

맵에서는 해당 주제어와 연관된 연구의 상위 연구자 지식맵을 나타낸다.

5. 결론

본 논문에서는 온톨로지와 토픽모델링을 활용한 지식맵 서비스를 소개하였다. Lightweight 온톨로지를 사용하여 국가 R&D 데이터의 공동 저자, 공동 참여자, 공통 토픽 등의 연관관계를 표현하였다. 온톨로지 데이터를 트리플 기반의 지식베이스로 구축하여 트리플 데이터의 유연성과 확장성을 이용해 다차원 연계 지식맵 서비스에서의 필요 데이터의 추출·활용 면에서 효율성을 높였다. 추가로, 국가 R&D 데이터의 의미 상 연관관계 및 주제어 추출을 위해 토픽모델링 방법을 활용하였다. 연구자들이 직접 입력한 주제어를 개체명사전에 포함하여 전처리 과정에서의 주제어 추출 정확도를 향상시켰고, 토픽모델링 결과 생성된 토픽 중심의 행렬을 이용하여 단어 중심이 아닌 토픽 중심의 코사인 유사도를 활용해서 의미 상 연관관계 추출의 정확도를 향상시켰다.

본 논문에서는 국가 R&D 정보를 토픽모델링으로 추출된 단순한 연관관계를 기반으로 생성한 온톨로지를 표현하였는데, 국가 R&D 원문 데이터에서 보다 세부적인 온톨로지 개체 및 관계를 추출하여 온톨로지를 구축하는 것이 필요해 보인다. 향후, 국가 R&D 원본 문서를 기반으로 이러한 세부적인 온톨로지를 구축하기 위해 보다 정교한 NLP (National Language Processing) 방법론을 연구할 계획이다. 이로 인해, 세부적인 온톨로지 구축이 가능하게 되면, 보다 세밀한 지식기반의 지식맵 서비스를 제공할 수 있어 보인다. 마지막으로, 본 논문에 소개한 지식맵 기반의 시각화 및 UI (User Interface) 방법론은 효율적이고 효과적인 지식 시각화와 지식 네비게이션 방법을 통해 일반적으로 사용하기 어려운 지식기반 시스템의 이용편의성 (easy of use) 을 높일 수 있다는 데 의의가 크다고 볼 수 있다.

참고문헌(References)

Ahmad, M. N. and R. M. Colomb, “Managing ontologies: a comparative study of ontology servers,” *Proceedings of the eighteenth*

conference on Australasian database, Vol.63 (2007), 13~22.

Blei, D. M., A. Y. Ng and M. I. Jordan, "Latent dirichlet allocation," *The Journal of machine Learning research,* Vol.3(2003), 993~1022.

Blei, D. M., "Probabilistic topic models," *Communications of the ACM*, Vol.55, No.4(2012), 77~84.

Brickley, D. and R. V. Guha, *RDF Schema 1.1*, W3C, 2014, Available at http://www.w3.org/TR/rdf-schema/ (Downloaded 14 December, 2015).

Businska, L., I. Supulniece and M. Kirikova, "On data, information, and knowledge representation in business process models," *Information Systems Development,* Springer New York, 2013, 613~627.

Eppler, M. J., "Making knowledge visible through intranet knowledge maps: concepts, elements, cases," *Proceedings of the 34th Annual Hawaii International Conference on System Sciences*, (2001), 9~18.

Hofmann, T., "Probabilistic latent semantic indexing," *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, (1999), 50~57.

Howard, R. A., "Knowledge maps," *Management science*, Vol.35, No.8(1989), 903~922.

Kang, I., Y. Park, and Y. Kim, "A framework for designing a workflow-based knowledge map," *Business process management journal*, Vol.9, No.3(2003), 281~294.

Klavans, R. and K. W. Boyack, "Toward a consensus map of science," *Journal of the American Society for information science and technology*, Vol.60, No.3(2009), 455~476.

Leydesdorff, L. and I. Rafols, "A global map of science based on the ISI subject categories," *Journal of the American Society for Information Science and Technology*, Vol.60, No.2(2009), 348~362.

McCagg, E. C. and D. F. Dansereau, "A convergent paradigm for examining knowledge mapping as a learning strategy," *The Journal of Educational Research*, Vol.84, No.6(1991), 317~324.

Morbach, J., A. Wiesner, and W. Marquardt, "OntoCAPE－A (re) usable ontology for computer-aided process engineering," *Computers & Chemical Engineering*, Vol.33, No.10 (2009), 1546~1556.

W3C RDF Working Group,*Resource Description Framework (RDF) 1.1*, W3C, 2014, Available at http://www.w3.org/RDF/(Downloaded14 December,2015).

Prud'hommeaux, E. and A. Seaborne, *SPARQL Query Language for RDF*, W3C, 2008, Available at http://www.w3.org/TR/rdf-sparql-query/ (Downloaded 14 December, 2015).

Rao, L., G. Mansingh and K. M. Osei-Bryson, "Building ontology based knowledge maps to assist business process re-engineering," *Decision Support Systems*, Vol.52, No.3(2012), 577~589.

Salton, G. and M. J. Mcgill, *Introduction to modern information retrieval*, McGraw-Hill, New York, 1986.

W3C OWL Working Group, *OWL2 Web Ontology Language (Second Edition)*, W3C, 2012, Available at http://www.w3.org/TR/2012/REC-owl2-overview-20121211/ (Downloaded 14 December, 2015).

Abstract

A Study on Ontology and Topic Modeling-based Multi-dimensional Knowledge Map Services

Hanjo Jeong*

Knowledge map is widely used to represent knowledge in many domains. This paper presents a method of integrating the national R&D data and assists of users to navigate the integrated data via using a knowledge map service. The knowledge map service is built by using a lightweight ontology and a topic modeling method. The national R&D data is integrated with the research project as its center, i.e., the other R&D data such as research papers, patents, and reports are connected with the research project as its outputs. The lightweight ontology is used to represent the simple relationships between the integrated data such as project-outputs relationships, document-author relationships, and document-topic relationships. Knowledge map enables us to infer further relationships such as co-author and co-topic relationships. To extract the relationships between the integrated data, a Relational Data-to-Triples transformer is implemented. Also, a topic modeling approach is introduced to extract the document-topic relationships. A triple store is used to manage and process the ontology data while preserving the network characteristics of knowledge map service.

Knowledge map can be divided into two types: one is a knowledge map used in the area of knowledge management to store, manage and process the organizations' data as knowledge, the other is a knowledge map for analyzing and representing knowledge extracted from the science & technology documents. This research focuses on the latter one. In this research, a knowledge map service is introduced for integrating the national R&D data obtained from National Digital Science Library (NDSL) and National Science & Technology Information Service (NTIS), which are two major repository and service of national R&D data servicing in Korea. A lightweight ontology is used to design and build a knowledge map. Using the lightweight ontology enables us to represent and process knowledge as a simple network and it fits in with the knowledge navigation and visualization characteristics of the knowledge map. The lightweight

* Corresponding author: Hanjo Jeong
NTIS Center, Division of Advanced Information Convergence, Korea Institute of Science and Technology Information (KISTI)
245, Daehak-ro, Yuseong-gu, Daejeon, 34141, Korea
Tel: +82-42-869-1861, Cell: +82-10-8453-2161, E-mail: hanjo.jeong@gmail.com

ontology is used to represent the entities and their relationships in the knowledge maps, and an ontology repository is created to store and process the ontology. In the ontologies, researchers are implicitly connected by the national R&D data as the author relationships and the performer relationships. A knowledge map for displaying researchers' network is created, and the researchers' network is created by the co-authoring relationships of the national R&D documents and the co-participation relationships of the national R&D projects.

To sum up, a knowledge map-service system based on topic modeling and ontology is introduced for processing knowledge about the national R&D data such as research projects, papers, patent, project reports, and Global Trends Briefing (GTB) data. The system has goals 1) to integrate the national R&D data obtained from NDSL and NTIS, 2) to provide a semantic & topic based information search on the integrated data, and 3) to provide a knowledge map services based on the semantic analysis and knowledge processing. The S&T information such as research papers, research reports, patents and GTB are daily updated from NDSL, and the R&D projects information including their participants and output information are updated from the NTIS. The S&T information and the national R&D information are obtained and integrated to the integrated database. Knowledge base is constructed by transforming the relational data into triples referencing R&D ontology. In addition, a topic modeling method is employed to extract the relationships between the S&T documents and topic keyword/s representing the documents. The topic modeling approach enables us to extract the relationships and topic keyword/s based on the semantics, not based on the simple keyword/s. Lastly, we show an experiment on the construction of the integrated knowledge base using the lightweight ontology and topic modeling, and the knowledge map services created based on the knowledge base are also introduced.

Key Words : Ontology; Topic Modeling; Knowledgebase; Knowledge Map; Information Integration

Received : November 25, 2015 Revised : December 11, 2015 Accepted : December 12, 2015
Corresponding Author : Hanjo Jeong

저 자 소 개

정 한 조

George Mason University에서 Information Systems 석사 및 Information Technology 박사를 취득하였다. LG 전자 CTO 부문 연구소에서 스마트 TV대상 BI 시스템 구축과 빅데이터 기반 앱 검색/추천 시스템, IoT 기반 스마트 기기 추천 시스템 등의 연구 및 개발에 참여하였고, 현재 한국과학기술정보연구원 (KISTI)에서 선임연구원으로 재직 중이다. 주요 연구 관심 분야는 빅데이터 기반의 데이터 마이닝, SNS 분석, 자연어 처리 및 검색/추천 시스템과 시맨틱 웹/온톨로지 Rule 기반 시스템, 기계학습 등의 Intelligence 기반 지식처리 시스템 등이다.