http://www.jsebs.org
ISSN: 2288-3908

The Journal of Society for e-Business Studies
Vol.21, No.1, February 2016, pp.119-129
http://dx.doi.org/10.7838/jsebs.2016.21.1.119

과학기술용어 간 관계 도출을 위한 토픽 분석 연구

Research of Topic Analysis for Extracting the Relationship between Science Data

김무철(Mucheol Kim)*

초 록

웹의 발달과 함께 많은 정보들이 쏟아지기 시작했다. 그에 따라서 사회 이슈들을 소셜 데이터로부터 추출하고, 이에 대한 해결 방법을 모색하는 연구에 대한 관심이 많아지고 있다. 이에 본 연구에서는 과학기술문헌들을 수집하고, 분석해서 이슈 토픽 별로 군집화 하는 연구를 수행한다. 이를 위해서 보건분야의 주요 용어들을 중심으로 수집하고, 효과적인 분석을 위한 데이터 처리 및 토픽들을 중심으로 군집화 연구를 수행한다. 그 결과, 연구 이슈들을 도출하고 사회 현상에 대한 해결 방안을 마련할 수 있는 토대를 구축하고자 한다.

ABSTRACT

With the development of web, amount of information are generated in social web. Then many researchers are focused on the extracting and analyzing social issues from various social data. The proposed approach performed gathering the science data and analyzing with LDA algorithm. It generated the clusters which represent the social topics related to 'health'. As a result, we could deduce the relationship between science data and social issues.

키워드 : 토픽 분석, 과학기술용어 분석, 소셜 네트워크 분석, 웹 기반 기술
Topic Analysis, Science Data Analysis, Web Technology, Social Network Analysis

* Department of Multimedia, Sungkyul University(mucheol.kim@gmail.com)
Received: 2016-02-04, Review completed: 2016-02-12, Accepted: 2016-02-22

1. 서 론

최근 웹 기술의 발전과 더불어 다양한 출처의 정보들이 웹을 통해서 공개되고 있으며, 그에 따라서 다양한 사회 이슈들이 사람들에게 신속하게 전달되고 있다. 정보의 양이 방대해짐에 따라서, 원하는 정보를 효과적으로 획득하기 위한 핵심문구 추출은 문서의 분류 및 군집, 그리고 정보 검색 분야에서 매우 중요하게 다루어지고 있다[1].

한편, 최근 정보기술 분야에서는 연구 및 서비스 결과가 사회의 필요를 충족하고 문제점 해결의 필요성이 대두되고 있다[2]. 다시 말하자면, 사회 문제와 동떨어진 정보기술은 최신 기술이라 하더라도 실험실 안에서만 적용되는 한계점을 극복하지 못하고 사장될 것이다. 따라서, 사회의 필요성을 충족시키고 더 나아가 사회문제 해결형 기술의 연구와 개발이 필요하게 되었다[3].

한편, 과학기술 분야에서는 다양한 R&D 사업을 통해서 논문, 특허, 연구보고서 등의 성과물을 도출하고 있다[4]. 이들은 실세계에서의 요구사항들을 도출하여 이론적, 기술적으로 정리하고 새로운 방법론을 제시하고 있다[5, 6]. 따라서, 신뢰성 있는 사회문제 해결의 열쇠가 될 수 있으며, 더 나아가 새로운 기술 및 이론의 시작점이 될 수 있다.

따라서, 본 연구에서는 사회문제 해결 방법을 도출하기 위해서 과학기술문헌들을 분석해서 주요 이슈들을 도출하고자 한다.

이를 위해서 과학기술문헌을 획득하기 위해서 한국과학기술정보연구원에서 제공하는 NDSL[7]의 논문 정보들을 수집하고, 이들이 가지고 있는 과학기술 이슈들의 토픽 분석을 수행한다. 그리고 분석한 결과를 바탕으로 과학기술용어들 간의 상관관계를 분석한다.

2. 관련 연구

LDA는 확률 모델을 기반으로 문서 집합의 이슈 정보들을 도출할 수 있는 토픽 모델링 알고리즘이다[8, 9]. 이는 문서 집합 내에서의 군집별 디리클레 분포를 통해서 각 문서별 용어들의 분포를 분석하여 군집에 포함 여부를 판단하는 알고리즘이다. 최근, 공학뿐만 아니라 인문사회학 등 다양한 학문 분야에서도 이슈 문서 도출 및 응용을 위해서 LDA 알고리즘이 적용되고 있다[2].

뿐만 아니라, 이는 단순히 문서 간 군집 분석에만 쓰이는데 그치지 않고 내용 기반 이미지 검색(Content-Based Image Retrieval), 음악 검색 에서도 적용되고 있다[10, 11].

정보검색 분야에서는 지속적으로 문서들을 분석해서 연관성을 찾고, 사용자 맞춤형 정보를 제공하기 위한 노력을 끊임없이 하고 있다[12]. 최근에는 소셜 네트워크 분석을 통해서 정보 간의 관계를 도출하고, 그 결과를 검색 결과에 반영하여 맞춤형 정보제공에 활용하는 연구들이 수행되고 있다[13, 14]. 또, 연관관계 분석을 사회 이슈에 적용하여 사용자 관점에서의 주제 군집화를 수행하기도 했다[15].

한편, 국내에서는 과학기술문헌의 수집 및 효과적인 제공을 위해서 많은 연구와 서비스들이 진행되고 있다. NDSL(National Digital Science Library)에서는 연구자들에게 국내외 학술저널 및 프로시딩 정보 및 원문을 제공해주는 서비스를 제공해주고 있다[4]. 또, 국가

R&D 정보들 간의 연관관계를 분석해서 현안 키워드를 도출하는 연구를 통해서 R&D 정보의 패키징 방안을 모색하기도 했다[3]. 그리고 분석의 신뢰성을 높이기 위해서 과학기술문헌으로부터(주제, 기술, 적용분야)의 핵심 이슈들을 추출하여 연구 결과의 동적 분석을 수행하는 연구[16]에 대한 검토가 필요하다.

3. 과학기술 데이터 수집 및 처리

3.1 과학기술 데이터 수집

본 연구에서는 과학기술 데이터 수집을 위해서 KISTI NDSL[7]에서 제공하는 Open API 서비스인 NOS(NDSL Open Service)를 이용해서 논문 정보를 수집했다. NOS는 키워드 기반 검색 결과를 제공하기 때문에 본 연구에서는 보건 분야의 주요 이슈키워드들을 이용해서 논문 정보들을 수집했다.

보건 분야 논문 수집을 위해서 노인, 농촌, 간호직, 구강 등의 주요 키워드들을 이용했다(<Table 1> 참조).

〈Table 1〉 Major Keywords for 'Health' Research Field

Major Keyword in 'Health'
Senior citizen, rural, public, mental, oral, safety, Medical Law, school, clinic, doctor, Safety, nurse, welfare

한편, 보건 분야의 주요 키워드들을 이용해서 논문을 수집한 현황은 <Table 2>와 같다. 주요 키워드인 보건과 관련된 논문은 약 6,600개이며, <Table 1>의 주요 키워드들을 이용해서 수집한 논문이 약 51,000개이다. 이들 중에서 중복을 제거했을 때, 32,173개의 논문이 수집되었다. 수집된 논문을 효과적으로 관리하기 위해서 본 연구에서는 논문의 관리번호, 게재년, 제목, 내용을 수집했다(관리 번호는 수집 대상이 된 NDSL의 관리를 위한 고유번호로 논문 간의 식별을 위해서 본 연구에서도 함께 수집했다).

〈Table 2〉 Status of Number of Papers in 'Health' Research Field

	Target	#Paper
1	Health	6,677
2	Major Keyword (Duplicated)	51,289
3	Major Keyword (Unique)	32,173

3.2 과학기술 데이터 전처리

〈Figure 1〉 Preprocessing for Science Data Analysis

한편 수집된 논문은 한글을 대상으로 하기 때문에 효과적인 정보의 관리 및 분석을 위해서 형태소 분석을 수행했으며, 이들 중에서 했다. 형태소 분석의 결과로 모든 용언들을 품사별로 분류했으며, 본 연구에서는 이들 중 명사들을 별도로 추출했다. 추출된 명사 중에는 본 연구의 초점을 맞추고 있는 용어 간 연결 관계들을 분석하기에 적합하지 않은 용어들을 다수 포함하고 있기 때문에 이에 대한 식별 및 불용어

사전(<Table 3> 참조)을 구축해야 한다.

이는 형태소 분석을 통해서 추출된 용어들 사이에서 본 연구에서 수행하고자 하는 용어 간 연계 관계 분석에 도움이 되지 않는 용어들에 대한 간섭을 줄이기 위한 단계이다. 예를 들면 추출된 용어들 중에 논문, 연구, 실험 등의 용어는 문서 내에 다수 존재할 뿐만 아니라 많은 문서에서 나타난다.

〈Table 3〉 Sample Results after Manual Stemming

	DF Factor	1st Manual Check	2nd Manual Check
Utilize (이용)	16808	Y	Y
Research (연구)	14310	Y	Y
Use (이용)	12325	Y	Y
Result (결과)	12170	Y	Y
Characteristic (특성)	11151	Y	Y
Paper (논문)	10446	Y	Y
Sensor (센서)	9229	N	N
Measure (측정)	8178	Y	Y
Analyze (분석)	7918	Y	Y
Methodology (방법)	7543	Y	Y

이와 같은 경우에는 용어 간 관계 정도를 평가하는 벡터 공간 모델(Vector Space Model)을 통해서 가중치를 낮출 수 있지만, 용어 간 연관 관계를 도출하는 과정에서 원치 않는 결과가 도출될 수 있다. 따라서, 본 연구에서는 불용어 사전을 구축하고 이에 해당하는 용어들이 연구 결과에 미치는 영향을 최소화시킨다.

그 결과 수집된 데이터에 대한 형태소 분석 결과 보건 분야의 논문에서 사용되고 있는 용어(명사)는 약 17,000건이 수집되었다. 논문 수에 대비해서 오히려 용어의 수가 적은 부분은 제목과 초록을 거쳐서 출현 빈도가 높은 용어들이 중복되는 경우가 다수 발견되기 때문이다. 즉, 비슷한 이슈를 담고 있는 논문이 많이 발견된다는 의미이기도 하다. 본 연구에서는 정제된 17,000건의 과학기술용어에 대해서 토픽 분석을 수행하고, 도출된 토픽을 통해서 용어 간 연관성 정도를 도출한다.

4. 과학기술용어 토픽 분석

본 연구에서는 제 3장에서 추출한 과학기술 용어들을 대상으로 토픽 분석을 수행했다. 토픽 분석을 위해서 용어 가중치의 연산을 위해서 벡터 공간 모델을 적용한다. 벡터 공간 모델의 TF(Term Frequency)와 IDF(Inverse Document Frequency)를 계산하여 이들을 각 용어의 가중치로 계산한다. 이를 기반으로 용어의 디리클레 분포 값을 이용하는 LDA 알고리즘을 통해서 과학기술문서 내에서 나타나는 이슈 토픽들을 도출하고, 그와 관련된 용어 및 문서들을 생성한다.

4.1 벡터공간 모델을 이용한 용어 가중치 연산

본 연구에서는 벡터 공간모델에서 TF-IDF 값을 통해서 용어 가중치를 결정한다. TF는 용어가 각 문서에 나타난 빈도수에 대한 지표로

각 문서에서 용어가 가지는 가중치를 의미한다. IDF는 각 용어가 나타나는 문서의 정도를 평가하는 지표로 용어가 데이터 집합 전체에 걸쳐서 평가되는 중요도에 대해 평가한다.

각각의 과학기술문서, 즉 연구논문은 그들이 가지고 있는 용어(명사)들의 집합으로 구성되어 있는데, 각각의 용어들은 해당 논문의 정체성을 표현한다(Eq. 참조).

$$\alpha_i = \{t_1, t_2, t_3, \cdots, t_n\} \quad (1)$$

이에 TF는 문서 내의 용어들의 빈도를 측정하는 TF 는 가장 많이 출현한 용어의 빈도에 대비해서 해당 문서 내의 빈도를 계산한다(Eq. 2).

$$TF(t_n, q_i) = \frac{f_{t_n, a_i}}{\max(f_{t_n, d} : t_n \in a_i)} \quad (2)$$

TF의 경우 IDF 값과의 조화, 그리고 일반화를 거치기 위해서 값의 범위를 조정하기도 하는데, 본 연구에서는 이와 같은 과정을 거치지 않는다.

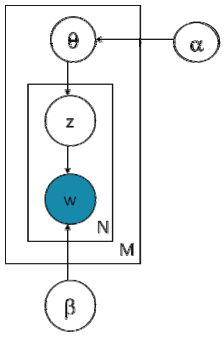
$$IDF(t_n, D) = \log \frac{N}{|\{a_i \in D : t_n \in a_i\}|} \quad (3)$$

한편, IDF는 용어가 출현하는 문서의 수 대비 전체 문서의 수를 연산하여 측정한다. 결과적으로 TF-IDF의 연산 값(TFIDF)은 TF와 IDF의 연산을 통해서 도출한다.

$$TFIDF(t_n, a_i, D) = TF(t_n, a_i) \times IDF(t_n, D) \quad (4)$$

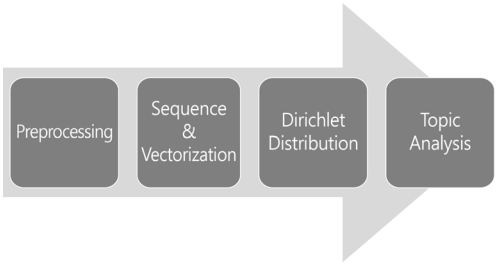
4.2 LDA 기반 토픽 분석

본 연구에서는 LDA 알고리즘[8]을 적용하여 과학기술문서에서 나타나는 용어들의 디리클레 분포를 연산하고 그 누적 분포를 이용해서 이슈 토픽을 도출한다. 이 과정에서 효과적인 처리 및 분석을 위해서 하둡 환경 하에서 사용 가능한 기계학습 라이브러리인 Mahout을 이용한다.



〈Figure 2〉 Essential of the LDA Algorithm

Mahout을 이용하기 위해서는 소스 데이터를 시퀀스 파일로 변환하고, 각각의 용어들에 대한 벡터화를 수행해야 한다(<Figure 1> 참조). 벡터화 과정에서 각각의 용어들의 가중치 입력을 위해서 본 연구에서는 TF-IDF 값을 사용했다. 이는 출현하는 용어의 빈도들을 기반으로 하는 디리클레 분포가 가지는 약점을 보완하기 위한 것으로 용어 별로 가중치가 다르게 적용되기 때문에, 잠재적 연결 관계를 도출함에 있어서도 가중치가 높은 용어들의 관계 출현 확률이 높아진다.



〈Figure 3〉 Process for Topic Analysis

한편, Mahout에서 제공하는 LDA 라이브러리는 CVB(Collapsed Variational Bayes)를 지원하는데 이는 LDA 알고리즘에 regular Variational Bayes와 Gibbs Sampling을 적용하여 군집화 과정의 복잡도를 개선한 알고리즘이다[17]. 이와 같은 CVB 알고리즘은 Gaussian 추정법(simple Gaussian approximation)을 이용해서 토픽 추출 결과의 계산 복잡도를 낮추고 정확도를 높인다.

〈Table 4〉 Various Parameters for LDA Algorithm

parameter	Description
-i	input path for document vectors
-dict	path to term-dictionary file(s), glob expression supported
-o	output path for topic-term distributions
-dt	output path for doc-topic distributions
-k	number of latent topics
-nt	number of unique features defined by input document vectors
-maxIter	max number of iterations
-mipd	max number of iterations per doc for learning

Mahout에서 LDA 알고리즘을 동작시키기 위해서는 <Table 4>의 파라미터들에 대한 설정이 필요하다. LDA 알고리즘 동작을 위해 필요한 입력 값, 출력 값들에 대한 설정은 물론이고, 추출할 토픽의 수와 입력되는 벡터에서의 고유한 용어들의 수, 그리고 연산의 반복 횟수를 설정한다. 본 연구에서는 토픽의 수를 반복하면서 적절한 결과 값을 얻고자 했다. 이를 위해서 과학기술문헌의 검색 결과 개선을 위해서 N-Gram을 적용하여(n = 2) 용어들 간의 유사도 평가를 수행했다.

5. 연구결과 및 분석

<Table 5>에서는 보건 분야의 수집 논문 전체를 대상으로 토픽 분석을 수행하였다. 실험 대상이 특정 분야를 대상으로 수집된 것이 아니고, 보건 분야의 주요 키워드를 기준으로 수집되었기 때문에 연구 결과물 역시도 다양한 분야의 논문이 수집되었다. 예를 들면, 의사라는 질의어는 병원의 의사로 쓰이는 한편, 의사소통을 의미하기도 하는데, 수집된 데이터들은 이와 같은 형태의 데이터들을 포함하고 있다.

〈Table 5〉 Issue Keywords and Associated Terms in Health Research Field

	Topic	Related Terms
1	signal (신호)	Channel, Frequency (채널, 주파수)
2	vessel(선박)	항공(Aircraft)
3	concrete (콘크리트)	구조물(Structure), 재료(material)
4	patient (환자)	Diagnostic, clinical, disease, radiation (진단, 임상, 질환, 방사선)
5	Cluster (군집)	Objects, water quality, water temperature, stream (개체, 수질, 수온, 하천)
6	environment-friendly (친환경)	Electricity, energy, Vehicle (전기, 에너지, 자동차)
7	senior citizen (노인)	Stress, suicide, patients (스트레스, 자살, 환자)
8	Consumer (소비자)	Design, content, game (디자인, 콘텐츠, 게임)
9	Processing (처리)	Growth, cultivation, content, soil (생육, 재배, 함량, 토양)
10	Farm village (농촌)	Multicultural families, Agricultural Management (다문화가정, 농업경영)

본 연구에서 시도하는 방법은 동음이의어뿐만 아니라 연계 용어들의 식별을 목적으로 하기 때문에, 이를 위한 별도의 데이터 정제 작업을 수행하지 않았다.

한편 실험 결과를 살펴보면 토픽 분석 군집을 10으로 했을 때, 해당 분야의 주요논문 이슈로는 '농촌', '노인', '환자', '친환경' 등이 나타났다. 이와 연계된 용어로는 '친환경'은 최근 기술적으로 이슈가 되고 있는 '전기', '에너지', '자동차' 등이 나타났다. '환자'와 관련된 연구 용어는 '진단', '임상', '질환', '방사선' 등을 포함하고 있었다.

〈Table 6〉 Focused Terms Related to '노인' and '농촌'

	A = Senior Citizen (노인)	B = Farm villa (농촌)
1	Stress (스트레스)	Multicultural families (다문화가정)
2	suicidal idea (자살생각)	Agricultural Management (농업경영)
3	tip of the tongue phenomenon (설단현상)	Viliage forecast (동네예보)
4	geriatric patients (노인환자)	Remodeling (리모델링)
5	elderly care facilities (노인요양시설)	Guideline (가이드라인)
6	male elderly (남성노인)	public construction industry (공공건설사업)
7	avocation (여가활동)	Separate order (분리발주)
8	cognitive function (인지기능)	Public Library (공공도서관)
9	Seniors Welfare Center(노인복지관)	Farmhouse (농가)
10	Smart Phone (스마트폰)	Farm product (농작물)

한편, 주요 토픽으로 도출된 '노인'과 '농촌'의 군집 구성을 구체적으로 살펴보면 <Table 6>과 같다. '노인'과 관련된 주요 논문 이슈들은 '스트레스', '자살', '환자' 등의 노인들의 사회적 이슈들을 다루고 있었다. 또, '요양시설', '여가활동', '복지관' 등의 노인들의 복지를 다루고 있는 이슈들도 함께 나타나고 있음을 확인할 수 있었다.

한편, '농촌'과 관련된 주요 이슈들을 살펴보자. 우선, 농업과 관련된 '농업경영', '농산물', '농가' 등의 이슈들이 도출되었다. 또, '농촌'의 '다문화가정'과 같은 사회학적 관점의 이슈들이 도출되기도 하였으며, '리모델링', '공공건설사업', '공공도서관'과 같은 농촌의 시설물 건설과 관련된 이슈들도 함께 등장하기도 했다.

이처럼 다양한 이슈들이 대표 키워드들과 함께 군집화되어 있기 때문에, 이를 추가 분석하기 위해서 본 연구에서는 최근 5년간의 연구 현황을 분석하기 위해서 특정 용어에 대해서 2010년~2014년의 연구 결과물을 추가 분석했다. <Table 7>과 <Table 8>은 해당 기간의 토픽 분석 결과를 나타내고 있다.

스트레스와 연계된 용어들의 시간 흐름별 분석 결과를 살펴보면, 2010년~2011년 사이에는 '스트레스'와 '건강상태'의 연계성이 강하게 나타나는 한편, '대학생'들의 '스트레스'에 대한 이슈도 나타남을 확인할 수 있었다. 한편, 2012년~2014년 사이에는 '스트레스'와 '의사소통' 간의 연관성이 강하게 나타났으며, '직무만족'과도 밀접하게 연계되어 출현하고 있음을 볼 수 있었다.

〈Table 7〉 Time Dimensional Clustering with '스트레스'

	Term1	Term2	Term3
2010	physical condition (건강상태)		
2011	college Student (대학생)	physical condition (건강상태)	
2012	job satisfaction (직무만족)		
2013	communication (의사소통)	job satisfaction (직무만족)	
2014	communication (의사소통)	job satisfaction (직무만족)	college Student (대학생)

〈Table 8〉 Time Dimensional Clustering with '저수지'

	Term1	Term 2	Term 3
2010	underwater (지하수)	pollutant (오염물질)	
2011	underwater (지하수)	Habitat (서식지)	
2012	sericultural farm housholds (양잠농가)		
2013	Ammonia (암모니아)	sericultural farm housholds (양잠농가)	Cocoon Crops (하추잠작)
2014	underwater (지하수)		

한편, <Table 8>에서는 '저수지'와 관련되어 지난 5년간의 이슈 용어들의 연계 현황들을 확인 할 수 있다. 전 기간에 걸쳐서 '저수지'의 이슈는 '지하수'와 연계되어 나타나고 있으며, 이와 관련하여 '오염물질'이 연계 용어로 도출됨을 확인할 수 있다. 또, 2012~2013년 사이에는 '양잠농가'가 '저수지'와의 연계성에 관심이 집중되었음을 확인할 수 있었는데, '하추참작', '암모니아' 등의 용어들이 이슈가 되었다.

위에서 살펴본 바와 같이 본 연구에서는 과학기술문헌을 통해서 연구 이슈들의 토픽 분석을 실시하였다. 본 연구의 결과는 기계학습의 결과물로써, 보건 분야의 주요 이슈들을 도출했다. 하지만, 이들을 통해서 용어들 간의 명시적 관계를 유추하는 것은 쉽지 않다. 따라서, 전문가 지식의 적용 및 기존에 존재하는 사전 등의 지식베이스를 함께 적용할 수 있는 방안이 모색되어야 한다.

6. 결 론

최근 소셜 데이터에 대한 관심이 높아짐에 따라서, 많은 연구자들이 이들로부터 사회 이슈의 효과적인 도출 및 분석에 관심을 가지기 시작했다. 한편, 사회 이슈들을 해결하기 위한 방법으로 과학기술문헌으로부터 신뢰성 있는 정보들을 도출하여 해결책을 마련하고자 하는 노력 역시도 지속하고 있다. 이에 본 연구에서는 과학기술문헌을 효과적으로 수집 및 관리하고, 이들이 다루고 있는 주제들을 검출하고 연계 관계를 분석하고자 했다. 이를 위해서 토픽 분석 기법을 이용하여 과학기술문헌 내의 용어들을 도출하고, 내재된 주제들 간의 연계 관계를 분석하기 위해서 LDA 알고리즘을 이용했으며, 그 결과 연구논문들이 다루고 있는 이슈들 간의 토픽 관계를 도출했다. 향후 연구

에서는 이들 간의 의미론적 관계는 물론이고 시간 흐름은 물론이고 이론이 기술로 전이되어가는 방향성 분석을 다루고자 한다.

References

[1] Wang, R., Liu, W., and McDonald, C., "Corpus-independent Generic Keyphrase Extraction Using Word Embedding Vectors," Software Engineering Research Conference, 2014.

[2] Jung, D., Kim, J., Kim, K., Hur, J., Ohn, B., and Kang, M., "A Proposal of a Keyword Extraction System for Detecting Social Issues," Journal of Intelligence and Information Systems, Vol. 19, No. 3, pp. 1-23, 2013.

[3] Hyun, Y., Han, H., Choi, H., Park, J., Lee, K., Kwak, K., and Kim, N., "Methodology Using Text Analysis for Packaging R&D Information Services on Pending National Issues," Journal of Information Technology Applications and Management, pp. 231-257, 2013.

[4] Kang, N., Cho, M., and Kwon, O., "A Relation Analysis between NDSL User Queries and Technical Terms," Journal of Information Management, Vol. 39, No. 3, pp. 163-177, 2008.

[5] Park, J. and Song, M., "A Study on the Research Trends in Library & Information Science in Korea using Topic Modeling," Journal of the Korean Society for Information Management, Vol. 30, No. 1, pp. 7-32, 2013.

[6] Kim, K. and Park, C., "Analysis of English abstracts in Journal of the Korean Data & Information Science Society using topicmodels and social network analysis," Journal of the Korean Data and Information Science Society, Vol. 26, No. 1, pp. 151-159, 2015.

[7] NDSL, http://www.ndsl.kr.

[8] Blei, D., A. Ng, M. Jordan, and J. Lafferty, "Latent Dirichlet Allocations," Journal of Machine Learning Research, Vol. 3, No. 4-5, pp. 993-1022, 2003.

[9] Blei, D. M., "Probabilistic topic models," Communications of the ACM, Vol. 55, No. 4, pp. 77-84, 2012.

[10] Misra, H., Anuj K. G., and Jose, J. M., "Topic Modeling for Content Based Image Retrieval," Multimedia Processing, Communication and Computing Applications. Springer India, pp. 63-76, 2013.

[11] Doulaty, M., Saz, O., and Hain, T., "Unsupervised Domain Discovery using Latent Dirichlet Allocation for Acoustic Modelling in Speech Recognition," in Proceedings of the 16th Annual Conference of the International Speech Communication Association (Interspeech), Dresden, Germany, 2015.

[12] Kim, S. and Kim., H., "Keyword Extraction from News Corpus using Modified TF-IDF," The Journal of Society for e-

Business Studies, Vol. 14, No. 4, pp. 59-73, 2009.

[13] Kim, M., Seo, J., Noh, S., and Han, S., "Identity management based social trust model for mediating information sharing and privacy enhancement," Security and Communication Networks, Vol. 5, No. 8, pp. 887-897, 2012.

[14] Oh, S., "A Model for Ranking Semantic Associations in a Social Network," The Journal of Society for e-Business Studies, Vol. 18, No. 3, pp. 93-105, 2013.

[15] Kim, J., Kim, N., Cho, Y., "User-Perspective Issue Clustering Using Multi-Layered Two-Mode Network Analysis," Journal of Intelligence and Information Systems, Vol. 20, No. 2, pp. 93-107, 2014.

[16] Gupta, S. and Manning, C. D., "Analyzing the Dynamics of Research by Extracting Key Aspects of Scientific Papers," In IJCNLP (pp. 1-9), 2011.

[17] Teh, Y. W., Newman, D., and Welling, M., "A collapsed variational Bayesian inference algorithm for latent Dirichlet allocation," In Advances in Neural Information Processing Systems, pp. 1353-1360, 2006.

저 자 소 개

김무철	(E-mail: mucheol.kim@gmail.com)
2012년	중앙대 컴퓨터공학과 (공학박사)
2011년~2014년	한국과학기술정보연구원 NTIS센터, 선임연구원
2014년~현재	성결대학교 멀티미디어공학과, 조교수
관심분야	정보검색, 소셜 네트워크, 웹서비스, 빅데이터