국내 기록관리학 연구동향 분석을 위한 토픽모델링 기법 비교*

- LDA와 HDP를 중심으로 -

Comparison of Topic Modeling Methods for Analyzing Research Trends of Archives Management in Korea: focused on LDA and HDP

박 준 형(JunHyeong Park)**
오 효 정(Hyo-Jung Oh)***

〈목 차〉

초 록

본 연구에서는 최근 각광을 받고 있는 텍스트마이닝 기법인 LDA 토픽모델링과 이를 변형한 HDP 토픽모델링을 적용하여 국내 기록관리학의 연구동향을 분석하고자 한다. 이를 위해 국내 기록관리학 관련 학술지 2종과 문헌정보학 관련 학술지 4종에서 1997년부터 2016년까지 발표된 기록관리학 관련 논문 1,027건을 수집하고 적절한 전처리과정을 거친 후 LDA 토픽모델링과 HDP 토픽모델링을 각각 수행하였다. 또한 토픽모델링 시각화 도구인 LDAvis를 활용하여 토픽별 거리를 가시적으로 표현하고 세부 대표 키워드를 분석하였다. 두 토픽모델링을 비교한 결과, LDA 토픽모델링은 전반적으로 해당 도메인을 대표하는 주요 키워드로 빈도수에 영향을 많이 받았으며, HDP 토픽모델링은 각 토픽별 특징을 파악할 수 있는 특수한 키워드가 많이 도출되었다. 이를 통해 LDA는 국내 기록관리학 내에 거시적으로 대표되는 주제들을, HDP는 세부 주제별 미시적인 핵심 키워드를 도출하는데 효과적임을 알 수 있었다.

키워드: 기록관리학, 연구동향, 토픽모델링, LDA, HDP

ABSTRACT

The purpose of this study is to analyze research trends of archives management in Korea by comparing LDA (Latent Semantic Allocation) topic modeling, which is the most famous method in text mining, and HDP (Hierarchical Dirichlet Process) topic modeling, which is developed LDA topic modeling. Firstly we collected 1,027 articles related to archives management from 1997 to 2016 in two journals related with archives management and four journals related with library and information science in Korea and performed several preprocessing steps. And then we conducted LDA and HDP topic modelings. For a more in-depth comparison analysis, we utilized LDAvis as a topic modeling visualization tool. At the results, LDA topic modeling was influenced by frequently keywords in all topics, whereas, HDP topic modeling showed specific keywords to easily identify the characteristics of each topic.

Keywords: Archives Management, Research Trends, Topic Modeling, LDA, HDP

* 이 논문은 2017년 한국도서관·정보학회 동계학술대회에 발표된 논문을 확장, 보완하였음
** 이 논문은 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2016S1A5B8913575)
*** 전북대학교 일반대학원 기록관리학과 석사(gulseori@naver.com) (제1저자)
전북대학교 기록관리학과 조교수, 문화융복합 아카이빙연구소 연구원(ohj@jbnu.ac.kr) (교신저자)
•논문접수: 2017년 11월 16일 •최초심사: 2017년 11월 25일 •게재확정: 2017년 12월 16일
•한국도서관·정보학회지 48(4), 235-258, 2017. [http://dx.doi.org/10.16981/kliss.48.201712.235]

Ⅰ. 서론

1. 연구배경 및 필요성

국내 기록관리학은 1999년 이후 공공기록관리법이 제정되고 기록학연구, 기록관리학회와 같은 기록관리 분야 학회가 탄생하면서 점차 발전하기 시작했다. 이후 컴퓨터와 정보통신망의 발달로 디지털 기반의 전자기록이 증가하면서 기존의 기록관리체계에 변화가 필요하게 되었다(손혜인 2016). 또한 국민들의 알권리를 존중하는 사회분위기가 강조되면서 기존의 보존과 전승 중심의 기록관리체계에서 기록정보 활용을 유도하는 서비스 중심의 기록관리체계로 변화하기 시작하였다(한국기록관리학회 2013). 이러한 시대적 변화에 따라 2006년 공공기록관리법이 개정되고 전자기록관리체계의 구축과 기록정보의 공개 및 활용의 확대에 대한 연구들이 활발하게 진행되었다. 특히 최근 데이터마이닝, 인공지능, 기계학습 등 IT기술이 발달하면서 정보의 규모와 발전속도가 빠르게 변화하는 시대에서 기존의 기록관리학 연구동향에 대한 보다 심층적인 분석과 향후 발전하게 될 주요 연구에 대한 파악이 필요하다.

한편, 텍스트마이닝 기법 중 하나인 토픽모델링은 최근 많은 연구에서 활용되면서 관심이 높아지고 있다. 토픽모델링은 방대한 양의 문서로부터 주요 주제를 추출하고 각 주제에 대응되는 문서를 식별하여 제공하는 방법이다(Blei, Ng and Jordan 2003). 토픽모델링 기법은 초기에 LSA(Latent Semantic Allocation), pLSA(Probablistic Latent Semantic Analysis) 등의 방법이 주로 활용되다가 이후 Blei, Ng and Jordan(2003)가 고안한 LDA(Latent Dirichlet Allocation)을 가장 널리 사용되고 있다. 최근에는 Teh et al.(2007)가 고안한 HDP(Hierarchical Dirichlet Process)가 새로운 토픽모델링 기법으로 떠오르고 있는데, LDA 토픽모델링은 사전에 이용자가 설정하는 토픽 개수가 결과에 많은 영향을 주지만 HDP 토픽모델링은 토픽 개수에 따른 결과의 변화가 상대적으로 적으며 사전에 토픽 개수를 설정하지 않아도 내부 알고리즘을 통해 적절한 결과를 출력하기 때문에 LDA 이후 새로운 주요 토픽모델링 기법으로 관심이 집중되고 있다(Wang, Paisley and Blei 2011). 하지만 아직까지 국내에서는 LDA 토픽모델링을 활용한 연구들이 주로 진행되었으며 HDP 토픽모델링을 활용한 연구는 매우 드물다.

이러한 맥락에서 본 연구에서는 국내 기록관리학의 연구동향을 분석하기 위해 LDA 토픽모델링과 이를 변형한 HDP 토픽모델링을 적용하고자 한다. 이를 위해서는 먼저 LDA와 HDP 토픽모델링 기법을 비교하여 그 특성을 파악하는 것이 선행되어야 한다. 특정 분야의 연구동향을 분석하는데 어떤 기법이 더 효과적인지 파악하기 위해 본 연구에서는 기록관리학

분야 학술지 2종과 문헌정보학 분야 학술지 4종을 중심으로 국내 기록관리학 관련 논문을 수집하고, LDA 토픽모델링과 HDP 토픽모델링을 각각 수행한 후, 두 토픽모델링의 결과를 비교함으로써 시사점을 도출하고자 한다.

2. 선행연구

최근 토픽모델링 기법을 활용한 연구가 다양한 분야에서 활발하게 진행되고 있다. 여론 분석이나 사회적 이슈 동향 파악, 특정 이슈를 도출하여 추적(tracking)하기 위한 연구 등 다양하게 활용되고 있는데, 특히 특정 분야의 연구동향을 파악하는데 주로 활용되고 있다(김남규 외 2017). 토픽모델링을 활용한 동향 분석 연구는 특정 학문 분야에 대한 연구동향 분석과 특정 기술에 대한 연구동향 분석으로 나눠서 살펴볼 수 있다. 먼저 특정 학문 분야에 대한 연구동향 분석을 살펴보면, 박자현, 송민(2013)은 1970년도부터 2012년도까지 국내 문헌정보학 분야의 발표 논문 초록 3,834건을 수집하여 LDA 토픽모델링을 수행한 후 도출된 연구주제를 문헌정보학 주제분류표와 비교·분석하여 국내 문헌정보학 연구동향을 분석하였다. 신규식, 최회련, 이홍철(2015)은 2006년 1월부터 2015년 6월까지의 신재생에너지학 관련 언론기사 51,558건을 대상으로 LDA 토픽모델링을 수행하여 국내 신재생에너지학 분야의 이슈 동향을 분석하였다. 진설아, 송민(2016)은 2009년도부터 2013년까지 정보학 분야 학술지 6,545개의 논문을 대상으로 LDA 토픽모델링을 활용하여 학술지의 학제성을 측정하였다.

세부 기술에 대한 연구동향 분석을 살펴보면, 박주섭, 홍순구, 김종원(2017)은 2000년부터 2016년까지 미국 특허 문서에서 “Artificial Intelligence” 키워드가 포함된 초록 14,187건을 대상으로 LDA 토픽모델링을 수행하여 AI(Artificial Intelligence) 기술의 동향과 예측을 분석하였다. 김태경, 최회련, 이홍철(2016)은 1990년 1월부터 2016년 7월까지 출원공개 된 미국, 한국, 중국특허 4,681건을 대상으로 LDA 토픽모델링을 수행하여 핀테크(Fintech) 기술의 동향을 분석하였다. 그 외에도 나상태 외(2017)는 1997년도부터 2016년도까지 Web of Science에서 “Smart Grid”키워드가 포함된 3,723건의 논문을 대상으로 LDA 토픽모델링을 시계열 토픽분석에 적합하게 확장한 기법인 DTM을 사용하여 스마트그리드 기술과 관련된 세부 연구동향을 분석하였다.

한편, 국내 기록관리학 연구동향 분석에 관한 선행연구는 기록학 전체 연구동향을 파악한 연구와 세부 주제의 연구동향에 대한 분석연구로 구분할 수 있는데(손혜인 2016), 본 연구가 기록학 전반에 걸친 전체 연구동향을 파악하기 위한 토픽모델링 기법 비교 연구이기 때문에 선행연구 분석은 기록학 전체를 대상으로 한 연구들을 중심으로 살펴보았다.

기록학 전체 연구동향 분석에 관한 연구는 대부분 기록관리학 분야 학술지와 문헌정보학 분야 학술지 내 기록관리학 관련 논문을 대상으로 진행하였는데 수집 및 분석대상 논문이 적은 순서로 살펴보면, 이재윤, 문주영, 김희정(2007)은 2001년부터 2006년까지 발행된 145편의 논문을 대상으로 문헌 클러스터링과 문헌 유사도 네트워크 분석을 활용하여 연구동향을 분석하였다. 김규환, 장보성, 이현정(2009)은 1999년부터 2008년까지 344편의 논문을 수집하고 논문제목의 구문과 의미구조를 분석하여 국내 기록관리학 분야 연구영역의 분포와 경향을 분석하였다. 김규환, 남영준(2009)은 1999년부터 2009년 9월까지 374편의 논문을 대상으로 주제영역과 연구배경 정보 등을 활용해 연구자 특성에 따른 기록관리학 분야 주제영역의 분포와 상관성을 분석하였다. 남태우, 이진영(2009)은 1997년부터 2007년까지 국내 기록관리학 분야 학술지 4종에서 수집한 399편의 학술논문을 대상으로 주제별 분포, 간행시기별 분포, 학회지별 분포, 연구자별 분포를 통계분석하였다. 최이랑(2015)는 2004년부터 2013년까지 479편의 논문을 대상으로 내용분석 및 네트워크 분석을 실시하여 국내 기록관리학 연구에서 가장 많이 등장한 키워드, 학술논문에 가장 많이 참여한 기관 등을 파악하였다. 손혜인, 남영준(2016)은 한국기록관리학회지와 기록학연구에서 2000년부터 2015년까지 게재된 681편의 논문을 중심으로 빈도분석과 네트워크 분석을 실시하였다. 그 결과, 두 학회지의 연구자 배경의 차이, 주제 변화 추이 등을 파악하였다.

이상에서 정리한 선행연구를 살펴보면 대부분의 연구들이 LDA 토픽모델링을 활용하여 해당 분야의 연구동향을 파악하였으며, 국내 기록관리학 연구동향 분석에 대한 선행연구에서는 문헌 클러스터링, 빈도분석, 네트워크 분석 등 다양한 기법을 활용하였지만 토픽모델링 기법을 활용한 연구동향 분석은 진행되지 않았다. 또한 단순히 특정 방법을 적용한 연구동향 결과의 분석에 그쳐 본 연구에서와 같이 다양한 방법을 적용한 결과들를 교차 비교함으로써 그 차이를 분석한 연구는 거의 전무하다.

Ⅱ. 이론적 배경

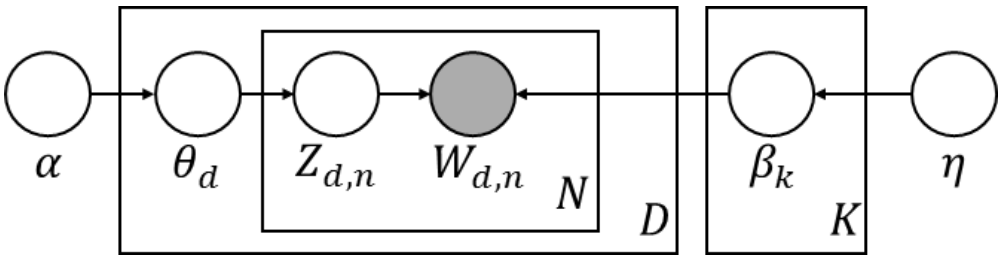
1. 토픽모델링

토픽모델링은 텍스트마이닝 기법 중 하나로 비구조화된 문서집합에서 잠재된 토픽들을 추출하주는 확률적 모델 알고리즘이다(Blei, Ng and Jordan 2003). 토픽모델링은 방대한 양의 문서집합에서 주요 토픽을 추출하고 각 토픽에 대응되는 문서를 식별하여 제공한다. 토픽모델링은 하나의 문서가 하나의 토픽으로만 할당되는 일반적인 군집화(clustering) 기법과

달리 하나의 문서가 여러 토픽에 동시에 대응될 수 있기 때문에 현실 세계의 모델링에 보다 적합한 기법으로 평가받고 있다(김남규, 이동훈, 최호창 2017). 기존의 정성적 분석 방법의 한계점을 극복하고 대량의 문서집합에서 잠재된 토픽을 찾아내는 기법으로 각광받고 있다(서성훈 2016). 본 절에서는 최근 각광받고 있는 LDA 토픽모델과 이를 변형한 방법인 HDP 모델링 방법에 대해 살펴보도록 한다.

가. LDA

LDA(Latent Dirichlet Allocation)는 토픽모델링의 가장 대표적인 방법론으로 이산 자료들에 대한 확률적 생성 모델로, 단어들의 확률을 이용하여 문서집합 내의 잠재된 토픽들을 찾아내는 기법이다(김태경, 최회련, 이홍철 2016). 초기에는 잠재의미분석(LSA: Latent Semantic Analysis)으로부터 시작하여 이를 변형한 확률 기반 잠재의미분석(pLSA: Probabilistic LSA) 기법으로 발전되어 사용되다가, 2003년 Blei가 고안한 LDA(Latent Dirichlet Allocation) 알고리즘을 발표된 이후 LDA가 토픽모델링의 주요 기법으로 사용되고 있다(남춘호 2016). LDA는 문서, 단어 등 관찰된 변수를 통해 문맥, 문서의 구조 등 보이지 않는 변수를 추론하는 방법으로 전체 문서집합의 주제, 각 문서별 주제 비율, 각 단어들이 각 주제에 포함될 확률 등을 파악할 수 있다(박자현, 송민 2013). <그림 1>은 LDA 그래프 모델이며 구체적인 내용은 다음과 같다(Blei 2012).



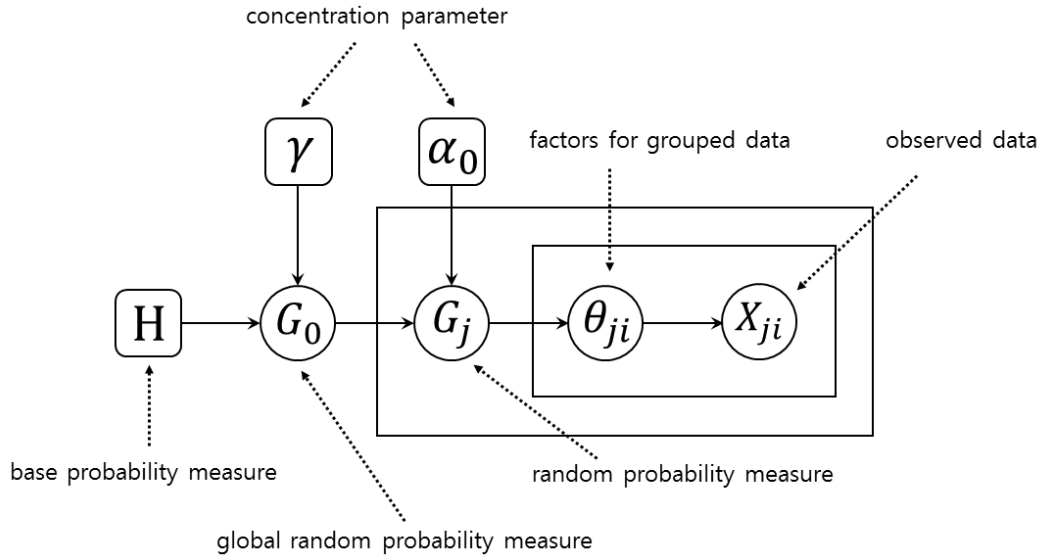
<그림 1> LDA 그래프 모델

<그림 1>을 살펴보면 K는 토픽의 개수, α는 θ값을 결정하는 파라미터이며 η는 β값을 결정하는 파라미터이다. θ는 문서별 토픽의 비율(topic proportions), β는 토픽별 단어 w의 생성비율(per-corpus topic distributions)이며 $Z_{d,n}$은 문서 d의 n번째 단어의 토픽, $W_{d,n}$은 문서 d의 n번째 단어로 문서에서 관측되는 변수(observed variable)를 의미한다(김상겸 2016). θ는 각 문서집합에 대한 주제 비율 값으로 Dirichlet 분포를 따르며 θ값에 따라 문서집합 내에 존재하는 단어들의 주제인, z가 결정된다. 또한 각 단어의 주제를 나타내는 값 Z와 토픽별 단어 생성비율인 β값에 따라 단어 W가 결정된다(박자현, 송민 2013). 이처럼

LDA는 사전에 설정한 파라미터 값에 따라 결과가 달라진다. 따라서 적절한 파라미터 값을 설정하지 않으면 적합한 결과를 얻을 수 없다는 한계점 있다. 또한 LDA는 1,000건 미만 소량의 문서집합에 대해 낮은 성능을 보이기 때문에 다른 분석 방법을 토픽모델링에 적용하여 이와 같은 한계점을 개선하고자 노력하고 있다(Loet Leydesdorff and Adina Nerghes 2017; 유소영 2015).

나. HDP

HDP(Hierarchical Dirichlet Process)는 Teh et al.(2007)이 고안한 토픽모델링 기법으로 Random Process에 기반한 DP(Dirichlet Process)를 계층적으로 적용하여 주요 토픽을 찾는 방법이다. DP는 앞서 설명한 LDA에서 사전에 정해진 주제 수 K값으로만 분포를 형성하는 디리클레 분포(Dirichlet Distribution)와 달리, K값을 사전에 설정하지 않아도 모분포에 따른 임의의 주제 개수를 가진 표본의 분포를 생성하는 방법이다.



<그림 2> HDP 그래프 모델

HDP 그래프 모델은 <그림 2>과 같이 표현할 수 있으며 구체적인 내용은 다음과 같다. X_{ji}는 문서집합 내 특정 문서에 대해 관찰된 단어이며 θ_{ji}는 특정 단어 X_{ji}에 대한 파라미터로 HDP 토픽모델링에서는 *factors*로 정의한다. G_j는 $\theta_j=(\theta_{j1},\theta_{j2},\ldots)$ 에 대한 사전 분포이며 α_0은 G_j의 집중 파라미터이다. G_0은 $G_j=(G_1,G_2,\ldots)$에 대한 랜덤 확률 측도이며 γ은 G_0의 집중 파라미터이다. H는 G_0에 대한 기본 확률 측도이다. 여기서 G_j은 G_0에 대해, G_0은

H에 대해 각각 DP를 생성한다. HDP 토픽모델링은 G_0과 H에 대한 DP를 하나 생성하여 이를 X_{ji}, θ_{ji}를 통해 생성된 사전 분포 G_j와 G_0에 대한 여러 개의 DP와 연결함으로써 적절한 K값과 주요 토픽을 찾는다.

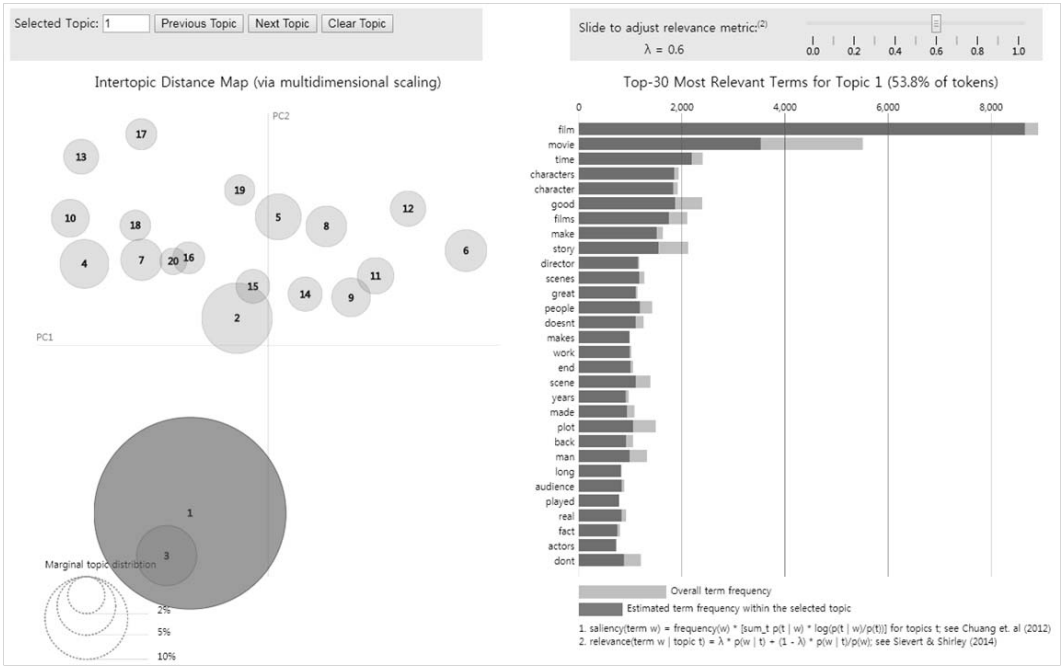
종합적으로 LDA 토픽모델링과 HDP 토픽모델링 기법을 비교해보면, LDA 토픽모델링은 사전에 이용자가 K, α, β값을 설정해줘야 하며 K값에 따라 토픽모델링 결과가 크게 달라지기 때문에 적절한 K값을 찾는 것이 중요하다. 반면 HDP 토픽모델링은 사전에 K값을 설정하지 않아도 DP를 통해 분석과정에서 적절한 K값을 찾아내 토픽모델링을 수행한다.

2. 시각화 도구

2003년 Blei가 LDA 토픽모델링 알고리즘을 제안하면서부터 다양한 학문분야에서 LDA 토픽모델링을 활용한 분석과 연구가 활발하게 진행되었다. 이러한 흐름에 따라 토픽모델링 결과를 시각화하는 도구에 관한 연구도 점차 증가하였다. 대부분의 토픽시각화에 대한 연구는 이용자에게 학습된 토픽모델 내 문서, 토픽, 용어에 대한 검색도구을 제공하는데 집중하였다(Sievert and Shirley 2014). 이러한 검색도구는 토픽에서 가장 가중치가 높은 용어 리스트, 전체 토픽에 대한 막대차트, 각 토픽 내 용어의 단어군집(Word Clouds), 각 문서 내 연관 토픽에 대한 파이차트를 제공하고 있다. 이러한 시각화 도구는 이용자가 직접 학습된 토픽모델링에 접근하여 원하는 개별 토픽과 용어를 직접 찾아볼 수 있어서 유용하지만 시각화를 통해 학습된 토픽모델링 결과 전체를 동시에 다양한 측면에서 살펴볼 수 없다는 한계점이 있다. Chuang, Manning and Heer(2012)는 학습된 토픽모델링의 토픽과 용어를 Matrix 형식으로 시각화하여 제공하는 Termite를 개발하였다. Termite는 Distinctiveness와 Saliency라는 측정 방법 사용하는데 전체 토픽에서 주요한 용어, 각 토픽과 주요한 용어 사이의 관계를 파악할 수 있다. 그러나 Termite는 전체 토픽에서 전반적으로 빈도수와 가중치가 높은 용어만을 보여주기 때문에 각 토픽마다 주요한 용어를 파악하기 어렵다.

한편, Sievert and Shirley(2014)는 기존의 토픽모델링 시각화 도구의 한계점을 보완하여 토픽과 단어의 관계를 전반적으로 살펴볼 수 있으며 각 토픽과 토픽 내 단어를 중요도에 따라 순위화하여 해당 문서집합에서 주요한 토픽, 단어를 쉽게 파악할 수 있는웹 기반 토픽모델링 시각화 도구인 LDAvis를 개발하였다.

LDAvis는 <그림 3>과 같이 두 부분으로 나눌 수 있다. 먼저, <그림 3>의 왼쪽 부분은 “Intertopic Distance Map”기능으로 학습된 토픽모델링의 전체 토픽을 2차원 척도로 나타낸 것이다. 여기서는 각 토픽의 연관성과 prevalence를 파악할 수 있는데, prevalence란 해당 토픽 내 용어가 전체 토픽에서 전반적으로 사용되는 용어이다. 각 토픽은 원으로 표현되며



<그림 3> LDAvis 예시

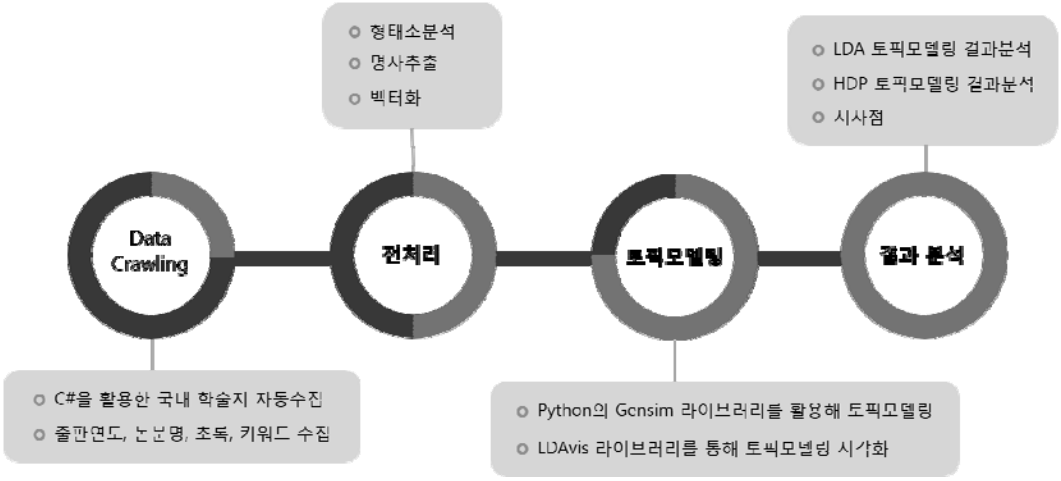
prevalence가 높을수록 원의 크기가 커진다. LDAvis는 prevalence가 높은 순서로 토픽을 정렬하여 제공하고 있다. 또한 각 토픽 사이의 거리는 토픽 사이의 연관성을 나타내며 거리가 가까울수록 토픽들의 연관성이 높으며 거리가 멀수록 토픽들의 연관성이 낮다. <그림 3>의 오른쪽 부분은 “Top-30 Most Salient Terms”과 “Top-30 Most Relevant Terms”기능을 제공하고 있다. “Top-30 Most Salient Terms”는 전체 토픽에서 가장 주요한 용어들을 보여주며 “Top-30 Most Relevant Terms Topic N”기능은 “Intertopic Distance Map”에서 특정 토픽을 클릭하면 해당 토픽에서 가장 주요한 용어들을 보여준다. 이는 토픽 내의 가장 주요한 용어들이 무엇인지 확인할 수 있으며 토픽 내의 주요한 용어들을 살펴봄으로 각 토픽이 가진 의미를 파악할 수 있다. 또한 각 용어를 선택하면 해당 용어를 포함한 다른 토픽들을 확인할 수 있어 선택한 용어의 prevalence를 파악할 수 있다.

이처럼 LDAvis는 기존의 토픽모델링 시각화 도구와 달리 학습된 토픽모델링을 다각적인 측면에서 동시에 시각화하여 토픽과 토픽, 토픽과 용어에 대한 관계를 쉽게 확인할 수 있다. 따라서 LDAvis는 학습된 토픽모델링을 단순히 시각화한 것에 머무는 것이 아니라 전체 토픽에 대한 전반적인 특징 파악 및 각 토픽의 개별적인 특징 파악이 동시에 가능하여 토픽을 쉽고 빠르게 이해할 수 있으며 토픽과 용어 사이의 관계를 통해 새로운 통찰(Insight)를 얻을 수 있다.

Ⅲ. 연구방법

1. 연구모델

본 연구에서 제안하는 국내 기록관리학 관련 연구동향 분석을 위한 토픽모델링 기법 적용 및 비교는 <그림 4>와 같이 진행되었다. 첫째, 기록관리학 관련 연구가 활발하게 진행되고 있는 기록관리학 학술지와 문헌정보학 학술지를 수집대상으로 선정하여, 국내학술지 원문DB 사이트인 DBPIA에서 학술지 최초 발간일부터 2016년까지 등재된 논문을 자동 수집하였다. 둘째, 수집된 데이터는 토픽모델링을 수행하는데 적합한 포맷으로 변경시키기 위해 형태소분석, 명사추출, 벡터화(Vectorization)과 같은 적절한 전처리 과정을 진행하였다. 셋째, 전처리 과정을 거친 데이터를 기반으로 LDA 토픽모델링과 HDP 토픽모델링을 각각 수행하고 LDAvis를 통해 각각 토픽모델링의 결과를 시각화하였다. 마지막으로 시각화를 통해 가시적으로 표현된 토픽들간의 거리와 세부 대표 키워드를 비교 · 분석하여 국내 기록관리학 연구동향 분석에 두 토픽모델링 기법의 특징과 차이를 분석하였다.



<그림 4> 연구과정 및 방법

2. 데이터수집

본 연구에서 기록관리학 관련 학술지인 ‘한국기록관리학회지’, ‘기록학연구’ 2종과 문헌정보학 관련 학술지인 ‘한국문헌정보학회지’, ‘한국도서관 · 정보학회지’, ‘한국비블리아학회지’,

‘정보관리학회지’ 4종으로, 총 6종의 국내학술지를 수집대상으로 선정하였다. “기록관리” 관련 연구는 주로 기록관리학 학술지에 많이 게재되지만 문헌정보학 학회지에서도 빈번히 게재되기 때문에 문헌정보학 학술지를 수집대상에 포함하였다. 수집기간은 각 학술지 발간일부터 2016년까지로 선정하였으며, 국내 학술DB 사이트인 DBPIA에서 논문명, 초록, 키워드, 논문저자 등 학술지에 게재된 모든 논문의 정보들을 자동수집하였다. 자동수집은 프로그래밍 언어인 C#을 기반으로 DBPIA 웹사이트 수집 크롤러를 직접 구현하여 사용하였다.

본 연구에서 수집한 데이터의 종합 통계는 <표 1>과 같다. 수집된 데이터 중에서 발간사, 목차, 서평 등 기록관리학 연구가 아닌 데이터는 사전에 연구대상에서 제외하였다. 세부 기록관리학 분야의 전문 학술지를 살펴보면 ‘한국기록관리학회지’는 발간연도인 2001년부터 2016년까지 339건의 논문을 수집하였고, ‘기록학연구’는 발간연도인 2000년부터 2016년까지 450건의 논문을 수집하였다. 문헌정보학 분야 학술지에서는 먼저 발간연도부터 2016년까지 전체 논문을 수집한 후 기록관리학 관련 논문만을 추출하기 위해서 논문명과 키워드에 “기록”, “아카이브”, “아카이빙”, “아키비스트”와 같은 단어가 포함된 논문을 추출하였다. ‘한국비블리아학회지’는 전체 수집 논문 791건 중 67건, ‘한국도서관 · 정보학회지’는 전체 수집 논문 1,562건 중 45건, ‘정보관리학회지’는 전체 수집 논문 1,242건 중 71건, ‘한국문헌정보학회지’는 전체 수집 논문 1,500건 중 55건이 기록관리학 관련 논문인 것을 확인할 수 있었다.

<표 1>은 본 연구의 실험집합의 통계를 나타낸 것으로, 전체 5,884건의 수집된 논문 중에서 기록관리학 관련 논문 1,027건을 대상으로 LDA 토픽모델링과 HDP 토픽모델링을 수행하여 각 모델링의 차이와 장 · 단점을 비교 분석하였다.

<표 1> 수집 데이터 종합 통계

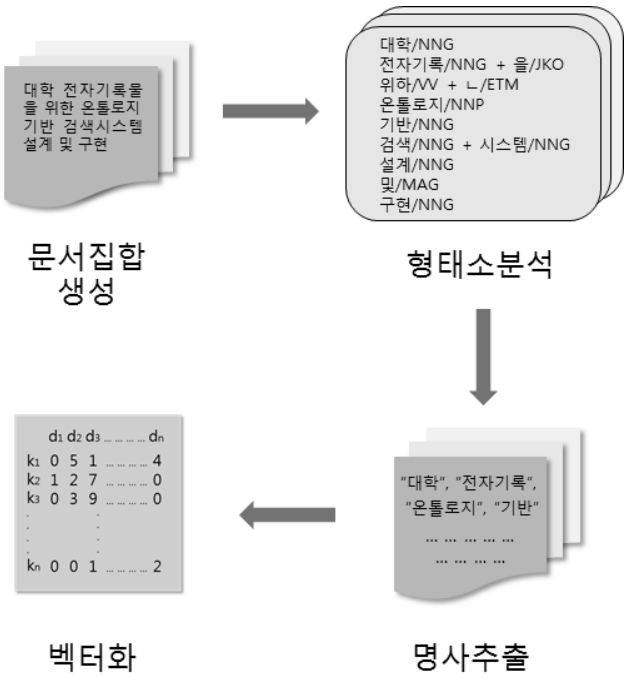
분야	학술지명	수집범위	전체 수집 논문 수	기록학 관련 논문 수
기록관리학	한국기록관리학회지	2001-2016	339	339
	기록학연구	2000-2016	450	450
문헌정보학	한국비블리아학회지	1972-2016	791	67
	한국도서관 · 정보학회지	1974-2016	1,562	45
	정보관리학회지	1984-2016	1,242	71
	한국문헌정보학회지	1970-2016	1,500	55
합계			5,884	1,027

3. 전처리작업

수집된 데이터는 토픽모델링 분석을 수행하기 전에 적절한 전처리과정이 수반되어야 한다.

연구내용분석을 위한 정확한 용어 추출 과정은 토픽모델링 수행 시 분석의 정확도 및 성능에 많은 영향을 주기 때문이다. 또한 수집된 데이터는 토픽모델링을 수행하기에 적합한 형식으로 변경해야 한다. <그림 5>는 본 연구에서 진행하는 전처리과정이며 구체적인 내용은 다음과 같다.

먼저, 수집한 기록관리학 관련 논문 1,027건의 논문명을 중심으로 문서집합을 생성하고 각 문서에 대해 형태소분석을 실시하였다. 형태소분석이 완료되면 분석된 문서에서 명사를 추출한다. 각 문서별로 명사추출을 완료하면 전체 문서집합은 벡터화(Vectorization)를 진행한다. 벡터화은 TF-IDF에 기반한 벡터 형식의 코퍼스(Corpus)를 생성하여 명사추출까지 완료된 전체 문서집합을 토픽모델링을 수행하는데 적합한 형식으로 변환하는 과정이다.



<그림 5> 전처리 과정

본 연구에서는 Python의 한국어 자연어처리 라이브러리인 Konlpy를 활용하여 문서집합에서 각 문서의 형태소분석 후 명사만 추출하였다. Konlpy 라이브러리는 Hannanum, Kkma, Komoran, Twitter, Mecab 5가지 형태소분석기를 제공하고 있다. 형태소분석기의 성능은 이후 진행되는 토픽모델링의 정확도와 성능에 있어 중요하다. 특히 형태소분석기가 복합명사 및 고유명사에 대한 처리를 어떻게 하는가에 따라 토픽모델링은 다른 결과를 가져온다.

<표 2> 각 형태소분석기 분석결과 예시

형태소분석기	예시) 대학특별사업단 기록물 관리 현황분석 및 개선방안 연구: J대학을 중심으로
Hannanum	**['대학특별사업단', '기록물', '관리', '현황분석', '개선방안', '연구', '대학', '중심']**
Kkma	['대학', '대학특별사업단', '특별', '사업단', '기록물', '관리', '현황', '현황분석', '분석', '개선', '개선방안', '방안', '연구', '대학', '중심']
Komoran	['대학', '특별', '사업단', '기록물', '관리', '현황', '분석', '개선', '방안', '연구']
Twitter	['대학', '특별', '사업', '단', '기록물', '관리', '현황', '분석', '및', '개선', '방안', '연구']
Mecab	['대학', '특별', '사업단', '기록물', '관리', '현황', '분석', '개선', '방안', '연구', '대학', '중심']

<표 2>는 Konlpy 라이브러리에서 제공하는 5가지 형태소분석기의 복합명사 처리결과 예시이다. <표 2>의 예시를 살펴보면 '대학특별사업단'이라는 복합명사가 문장에서 중요한 역할을 하고 있다. 이에 각 형태소분석기의 분석결과를 살펴보면 Hannanum과 Kkma는 '대학특별사업단'을 하나의 복합명사로 처리하였지만 Komoran, Twitter, Mecab은 이를 적절하게 처리하지 못한 것을 확인할 수 있다. 또한 Kkma는 '대학특별사업단'을 복합명사로 적절히 처리하였으나 '대학', '특별', '사업단'등으로도 명사를 추출한 것을 확인할 수 있다. 이러한 처리는 문서 내 반복되는 단어가 많아져서 이후 토픽모델링을 실행할 때 각 토픽의 특수한 키워드를 추출하기 어렵다. 따라서 본 연구의 형태소분석은 5가지 형태소분석기 중 복합명사를 적절하게 처리하는 Hannanum을 사용하여 진행하였다. 추출된 키워드 중 '연구', '중심', '분석' 등 모든 논문에 공통으로 출연하는 어휘는 불용어로 간주, 토픽모델링 입력에서는 제거하여 처리하였다.

4. 토픽모델링

전처리 과정을 통해 수집한 문서집합을 토픽모델링에 적합한 포맷으로 변환하면 동일한 문서집합을 대상으로 LDA 토픽모델링과 HDP 토픽모델링을 수행한다. 본 연구에서는 토픽모델링을 수행하기 위해 Python에서 지원하는 Gensim 라이브러리를 활용하였다. LDA 토픽모델링은 사전에 적절한 토픽 수를 설정해야 한다. 토픽 수를 너무 높게 설정하면 특별한 키워드가 없어 의미없는 토픽이 도출될 수 있으며 토픽 수를 적게 설정하면 한 토픽에 많은 키워드가 뭉쳐 토픽을 구분하기 어렵다. 이에 본 연구에서는 토픽 수를 5~20까지 설정한 후 각각 토픽모델링을 수행한 결과, 토픽 수 10개가 각 토픽을 적절하게 표현하는 것을 확인할 수 있었다. HDP 토픽모델링은 알고리즘 특성상 토픽 수를 설정할 필요가 없지만 본 연구에서는 두 토픽모델링의 적절한 비교를 위해 HDP 토픽모델링도 LDA 토픽모델링과 동일한 토

픽 수로 설정하여 분석을 진행하였다. 이후 Gensim의 토픽모델링 함수에 전처리된 문서집합과 토픽수, 반복횟수(1,000회)를 입력하고 다른 인수들은 기본값으로 설정하여 LDA 토픽모델링과 HDP 토픽모델링을 수행하였다. 수행된 결과는 Python의 토픽모델링 시각화 라이브러리인 LDAvis를 통해 각 토픽모델링 결과를 가시화하였다. LDAvis는 토픽모델링 결과를 html 형식의 파일로 제공하여 연구자가 웹브라우저를 통해 토픽모델링 결과를 쉽게 확인할 수 있다.

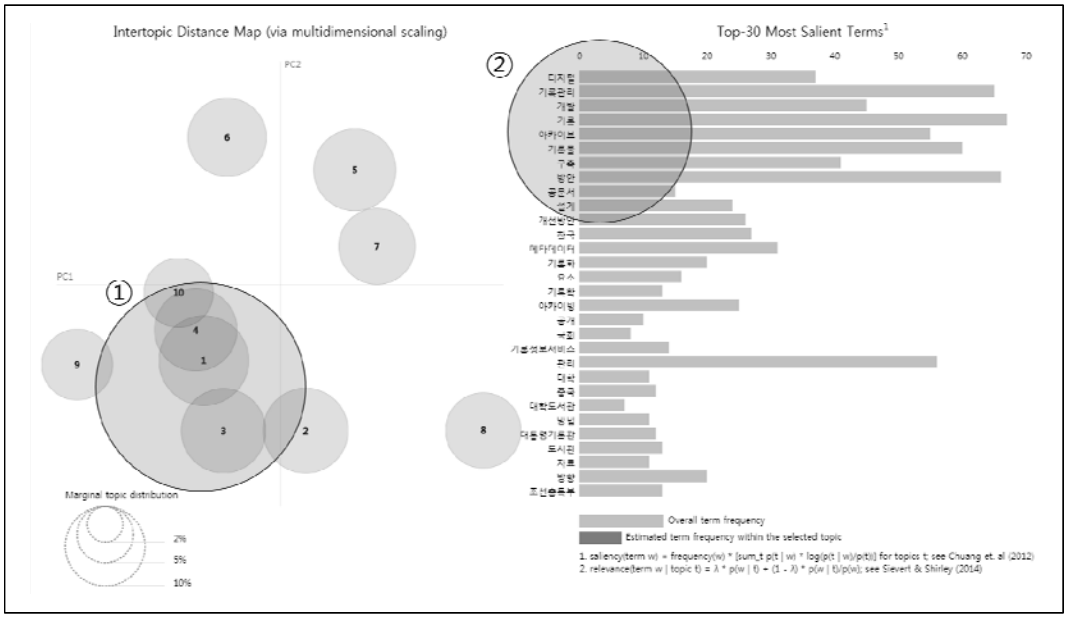
Ⅳ. 분석결과

1. LDA 토픽모델링 결과

LDA토픽모델링 결과를 LDAvis로 시각화하면 <그림 6>과 같다. LDAvis는 <그림 6>의 왼쪽 그림과 같이 전체 토픽에 대한 2차원 척도인 “Intertopic Distance Map”기능을 제공한다. 이 기능은 전체 토픽에 대해 전반적인 관계와 각 토픽 사이의 연관성을 파악할 수 있도록 도와준다. 또한 <그림 6>의 오른쪽 그림은 “Top-30 Most Salient Terms”기능으로 전체 토픽에서 가장 핵심적인 키워드를 확인할 수 있다.

이러한 LDAvis 기능을 통해 LDA토픽모델링 결과를 살펴보면, 먼저 LDA 토픽모델링의 Intertopic Distance Map은 ①과 같이 대부분의 토픽들이 특정 부분에 모여 있는 것을 확인할 수 있다. Topic1~4와 Topic9~10이 특정 부분에 집중되어 있으며 Topic5~8은 토픽이 집중된 부분에서 상대적으로 떨어져 있다. Topic5와 Topic7은 비교적 서로 가깝지만 Topic6, Topic8은 멀리 떨어져 있는 것을 확인할 수 있다. 이는 Topic1~4와 Topic9~10이 토픽 사이의 연관성이 높고 각 토픽 내 특정 키워드들이 많이 중복되고 있음을 나타낸다. Top-30 Most Salient Terms를 살펴보면 ②와 같이 “기록관리”, “기록”, “개발”, “아카이브”와 같은 키워드가 다른 키워드에 비해 전체 토픽에서 상당히 높은 빈도수로 출현하고 있는 것을 확인할 수 있다. 또한 그밖에 상위에 위치한 키워드를 살펴보면 “아카이빙”, “방안”, “기록물”, “관리”, “구축”등이 있다. 이처럼 LDA 토픽모델링은 전반적으로 논문제목에서 빈번히 사용하는 일반적인 키워드가 상위 키워드로 등장하였다.

한편, LDAvis 각 토픽을 클릭하면 해당 토픽에 대한 Top-30 Most Relevant Terms를 확인할 수 있다. 상대적으로 거리가 많이 떨어진 Topic6과 Topic8의 Top-30 Most Relevant Terms를 살펴보면, Topic6은 “기록정보서비스”, “이용자”, “콘텐츠”와 같은 키워드의 출현 빈도가 높고 Topic8는 “공문서”, “공공기관”, “한국국가기록연구원”과 같은 키워드



<그림 6> LDA 토픽모델링 시각화

가 토픽 내 상위 키워드로 등장한다. 반면에 가까운 거리에 위치한 Topic1~4와 Topic9~10의 키워드는 각 토픽에 유일하게 등장하는 키워드보다 대부분 Top-30 Most Salient Terms에 등장한 주요 키워드를 포함하고 있었다. <표 3>은 LDA토픽모델링에서 각 토픽별 상위 1~5순위의 주요키워드 및 가중치를 나타낸 것이다. <그림 6>의 Top-30 Most Salient Terms과 비교하여 살펴볼 때, 빈도수가 가장 많은 키워드인 “기록관리”, “기록”, “개발”, “아카이브” Topic1~10까지 거의 모든 토픽에서 가장 높은 순위에 위치하고 있으며 가중치도 다른 키워드들에 비해 상당히 높은 것을 확인할 수 있다. 또한 각 토픽의 1~5순위까지 주요키워드를 살펴보면 “기록물”, “기록화”등 일반적인 키워드가 높은 순위에 위치하고 있다.

<표 3> 상위 10개 각 토픽의 비율을 살펴보면, Topic1은 12.7%로 전체 토픽에서 가장 높은 비율을 차지하고 있으며 이후 토픽 순서대로 11.4%, 11.2%, 10.7%, 10.6%, 9.7%, 9.2%, 9%, 7.9%, 7.6%의 비율을 보이고 있다. 각 토픽 내 “기록관리”, “기록”, “개발”, “아카이브”와 같이 빈도가 많은 키워드와 Top-30 Most Salient Terms에 포함된 키워드의 가중치가 높을수록 토픽 비율이 높은 것을 확인할 수 있다.

<표 3>에 나타난 각 토픽의 상위 키워드를 통해 토픽별 주요 주제에 대해 분석해보면, Topic1은 “디지털”, “전자기록”, “메타데이터”, “개발”등 주로 ‘전자기록’이 주요 주제임을 확인할 수 있다. Topic2는 “대통령기록관”, “국가기록원”, “한국국가기록연구원”등 ‘국가기록원’에 관한 주제임을 알 수 있다. Topic3는 “구축”, “사례”, “활용”등의 키워드가 높은 순위에 나

<표 3> LDA 토픽별 주요키워드

순위	Topic1(12.7%)		Topic2(11.4%)		Topic3(11.2%)		Topic4(10.7%)		Topic5(10.6%)	
	키워드	가중치	키워드	가중치	키워드	가중치	키워드	가중치	키워드	가중치
1	**개발**	**0.027**	**기록**	**0.052**	**아카이브**	**0.039**	기록물	0.041	**아카이브**	**0.026**
2	**기록**	**0.023**	**기록관리**	**0.031**	**기록관리**	**0.037**	**개발**	**0.040**	기록화	0.024
3	디지털	0.023	대통령 기록관	0.012	구축	0.019	보존	0.018	**기록관리**	**0.015**
4	전자기록	0.015	국가기록원	0.008	사례	0.016	기록	0.008	운영	0.11
5	메타데이터	0.012	한국국가 기록연구원	0.008	활용	0.016	전자기록	0.008	아카이빙	0.011
순위	Topic6(9.7%)		Topic7(9.2%)		Topic8(9%)		Topic9(7.9%)		Topic10(7.6%)	
	키워드	가중치	키워드	가중치	키워드	가중치	키워드	가중치	키워드	가중치
1	**기록**	**0.018**	사례	0.020	**기록관리**	**0.047**	디지털	0.045	기록물	0.030
2	**아카이브**	**0.018**	**기록**	**0.013**	**기록**	**0.030**	**아카이브**	**0.026**	**기록관리**	**0.012**
3	이용자	0.016	중국	0.017	공문서	0.023	메타데이터	0.016	컴퓨팅	0.006
4	기록정보 서비스	0.013	한국	0.010	공공기관	0.10	표준	0.010	클라우드	0.006
5	콘텐츠	0.007	미국	0.008	한국국가기 록연구원	0.005	기술	0.010	전자기록 관리시스템	0.006

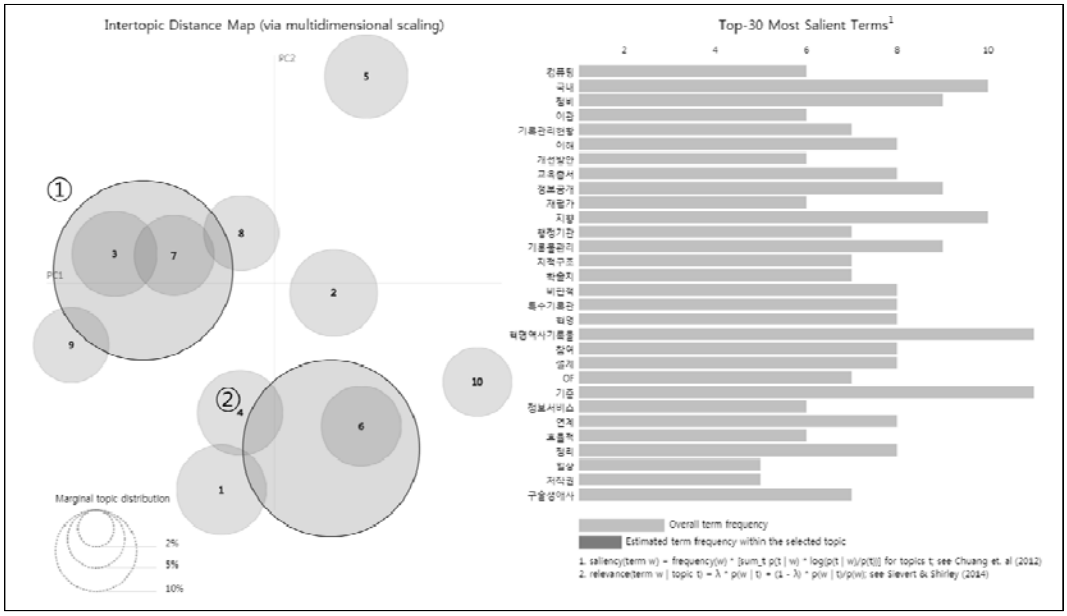
타나며 ‘아카이브 구축, 기록에 대한 활용’이 주요 주제이다. Topic4는 “보존”이라는 키워드가 다른 토픽보다 높은 순위에 나타나며 기록에 대한 보존이 주요 주제이다. Topic5는 “기록화”, “운영”, “아카이빙”등 ‘기록물의 아카이빙’이 주요 주제이다. Topic6은 “이용자”, “기록정보서비스”, “콘텐츠”등 ‘기록정보서비스’가 주요 주제이며 Topic7은 “사례”, “중국”, “한국”, “미국”등 ‘국내외 기록관리 현황, 사례’가 주요 주제이다. Topic8은 “공문서”, “공공기관”등 ‘공공기관의 기록’이 주요 주제이며 Topic9는 “메타데이터”, “표준”, “기술”등 ‘메타데이터와 기술’이 주요 주제이다. Topic10은 “컴퓨팅”, “클라우드”, “전자기록관리시스템”등 ‘전자기록관리시스템’이 주요 주제임을 파악할 수 있다.

토픽별 주요 주제를 종합적으로 살펴보면, 각 토픽별 키워드가 “전자기록”, “메타데이터”, “기록정보서비스”, “보존”등 대부분 기록관리학에서 주요 대주제로 사용하고 있는 키워드가 많기 때문에 각 토픽의 주제가 대부분 거시적인 것을 확인할 수 있다.

2. HDP 토픽모델링 결과

HDP 토픽모델링 결과를 LDAvis로 시각화하면 <그림 7>와 같다. <그림 7>의 왼쪽 그림을 살펴보면 HDP 토픽모델링은 LDA 토픽모델링과 같이 전체 토픽이 특정 부분에 과도하게 몰

려있지 않고 각 토픽들이 소영역을 이루며 전체적으로 퍼져 있는 것을 확인할 수 있다. 오른쪽 그림의 Top-30 Most Salient Terms를 살펴보면 HDP 토픽모델링은 “기록관리”, “기록”, “개발”, “아카이브”와 같이 전체 문서집합에서 단순히 빈도수만 높고 일반적인 의미를 가진 키워드보다 토픽의 특징과 주제를 파악할 수 있는 특수한 키워드가 높은 순위에 있는 것을 확인할 수 있다. 왼쪽 그림의 각 토픽별 Top-30 Most Relevant Terms를 살펴보면, LDA 토픽모델링에서는 대부분의 토픽들이 각 토픽의 특징을 나타내는 특수한 키워드, 유일한 키워드보다 일반적인 키워드가 중복되어 나타나지만 HDP 토픽모델링은 각 토픽이 중복된 키워드가 적고 유일하게 등장하는 키워드가 많이 나타난 것을 볼 수 있다.



<그림 7> HDP 토픽모델링 결과

마찬가지로 <표 4>의 HDP 토픽모델링에서 각 토픽별 상위 1~5순위의 주요키워드 및 가중치를 <표 3>과 비교해 살펴보면, HDP 토픽모델링은 LDA 토픽모델링처럼 빈도수가 높은 키워드나 일반적인 의미의 키워드가 전체 토픽의 상위 순위에 나타나지 않는다. 각 토픽의 상위 키워드들은 각 토픽의 고유한 특징을 파악할 수 있는 특수한 키워드가 주요 키워드로 위치하고 있다. 또한 각 토픽의 비율을 살펴보면, Topic1이 12.3%로 전체 토픽에서 가장 높은 비율을 차지하고 있으며 이후 Topic2은 11.5%, Topic3은 11.1%, Topic4은 10.9%, Topic5는 10.6%, Topic6과 Topic7은 9.7% Topic8은 8.6%, Topic9는 8.5%, Topic10은 7.2% 순의 비율을 이루고 있다.

<표 4> HDP 토픽별 주요키워드

순위	Topic1(12.3%)		Topic2(11.5%)		Topic3(11.1%)		Topic4(10.9%)		Topic5(10.6%)	
	키워드	가중치	키워드	가중치	키워드	가중치	키워드	가중치	키워드	가중치
1	적용방안	0.011	공문서	0.009	국제표준	0.007	이관	0.010	컴퓨팅	0.011
2	혁명역사 기록물	0.009	전자기록물	0.008	평가모델	0.006	관리제도	0.008	정보서비스	0.006
3	과학기술 분야	0.008	네트워크	0.006	포맷	0.005	분류방안	0.007	위키리크스	0.006
4	공연예술 기록	0.007	디지털	0.005	기술규칙	0.005	국가기록 관리체제	0.005	온톨로지	0.005
5	송전탑	0.006	시소러스	0.005	ICA	0.004	국가기록원	0.005	소셜미디어	0.005
순위	Topic6(9.7%)		Topic7(9.7%)		Topic8(8.6%)		Topic9(8.5%)		Topic10(7.2%)	
	키워드	가중치	키워드	가중치	키워드	가중치	키워드	가중치	키워드	가중치
1	기록관리 현황	0.011	서울특별시	0.008	기록물 관리	0.009	수집	0.008	우리나라	0.007
2	개선방안	0.009	전문성	0.006	정부산하 공공기관	0.007	세월호	0.008	대한제국 기록	0.007
3	관리제도	0.006	전문요원	0.006	레포지터리	0.007	공연예술 기록	0.006	공공기록물	0.006
4	운영현황	0.005	미국	0.006	스토리텔링	0.006	무형문화 유산	0.005	역사기록물	0.006
5	관리현황	0.004	기록관리 체제	0.006	매뉴스크립트	0.005	대학기록물	0.005	국회기록 정보	0.005

<표 4>에 나타난 각 토픽의 상위 키워드를 통해 토픽별 주요 주제에 대해 분석해보면, Topic1은 “적용방안”, “혁명역사기록물”, “과학기술분야”, “공연예술기록”, “송전탑”등의 키워드가 등장하는데 ‘다양한 민간기록의 아카이빙’이 주요 주제임을 파악할 수 있다. Topic2는 “공문서”, “전자기록물”, “네트워크”, “디지털”, “시소러스”등 ‘전자기록물과 디지털 기술’이 주요 주제이다. Topic3는 “국제표준”, “평가모델”, “포맷”, “기술규칙”, “ICA” 등 ‘기록관리에 관한 표준 및 규칙’이 주요 주제이다. Topic4는 “이관”, “관리제도”, “분류방안”, “국가기록관리체제”등 ‘공공기록물의 기록관리’가 주요 주제이다. Topic5는 “컴퓨팅”, “정보서비스”, “온톨로지”, “소셜미디어” 등 ‘기록정보서비스의 디지털 활용방안’이 주요 주제이다. Topic6는 “기록관리현황”, “개선방안”, “운영현황”, “관리현황”등 ‘기록관리현황 및 개선방안’이 주요 주제이다. Topic7은 “서울특별시”, “전문성”, “전문요원”, “미국”, “기록관리체제”등 ‘국내외 기록관리체제와 아키비스트의 전문성’이 주요 주제이다. Topic8은 “기록물관리”, “정부산하공공기관”, “레포지터리”, “스토리텔링”, “매뉴스크립트”등 ‘공공기관의 기록물 활용방안’이 주요 주제이다. Topic9는 “수집”, “세월호”, “공연예술기록”, “무형문화유산”, “대학기록물” 등 ‘다

양한 기록물의 수집 방안'이 주요 주제이다. Topic10은 "우리나라", "대한제국기록", "공공기록물", "역사기록물"등 '국내 역사기록물'이 주요 주제이다.

HDP는 각 토픽 내 키워드가 LDA와 비교할 때 상대적으로 세부적인 주제를 파악할 수 있는 키워드가 많다. 예를 들어, "공연예술기록", "혁명역사기록물", "무형문화유산"등과 같은 키워드는 최근 다양한 민간기록에 대한 기록화 연구가 진행되어지고 있는 것을 파악할 수 있다. 토픽별 주제를 종합적으로 살펴보면, HDP의 각 토픽의 주제가 LDA 보다 세부적인 의미로 분석되어지는 것을 확인할 수 있다.

3. 시사점

동일한 전처리과정을 거친 문서집합을 기반으로 토픽모델링을 수행했지만 LDA 토픽모델링과 HDP 토픽모델링은 그 결과가 상이하게 나타났다. 각 결과를 종합적으로 살펴보면 먼저, LDA 토픽모델링은 "기록관리", "기록", "개발", "아카이브"등과 같이 전체 문헌에서 빈도수가 높은 키워드의 영향을 많이 받았다. Top-30 Most Salient Terms에서는 대부분 "아카이빙", "방안", "기록물", "관리", "구축"과 같이 일반적인 단어들이 상위 키워드로 위치하고 있었으며 각 토픽별 주요키워드에서도 일반적인 키워드가 높은 가중치를 가지며 각 토픽에 중복되어 나타나고 있었다. 이러한 영향 때문에 Intertopic Distance Map을 살펴보면 전체 토픽들이 특정 위치에 몰려 있는 것을 확인할 수 있다.

반면 HDP 토픽모델링은 Top-30 Most Salient Terms에서 일반적인 키워드가 상위 순위에서 제외되었으며 각 토픽별 주요키워드는 다른 토픽과 중복되지 않으며 해당 토픽의 특징을 파악할 수 있는 특수한 키워드가 높은 순위에 있었다. 이로 인해 Intertopic Distance Map에서 확인할 수 있는 것처럼 HDP 토픽모델링은 LDA 토픽모델링보다 상대적으로 전체 토픽이 분산되어있고 일부 1~2개의 토픽들끼리 소영역을 이루고 있다.

LDA와 HDP 토픽별 키워드를 살펴보면, LDA는 각 토픽을 대표하는 키워드가 "전자기록", "메타데이터", "기록정보서비스", "보존"등 대부분 기록관리학 분야의 주요 대주제로 사용하는 거시적 의미의 키워드가 많이 존재하며 세부적인 의미를 파악할 수 있는 키워드가 적다. 반면에 HDP의 각 토픽은 세부적인 의미를 파악할 수 있는 키워드를 많이 포함하고 있다. 이와 같은 특징은 각 토픽의 주제 분석에 있어서 많은 영향을 미친다. <표 5>는 LDA와 HDP 토픽별 주요 주제를 정리한 것으로, 앞서 예시한 바와 같이 LDA 토픽모델링 결과에는 대부분 거시적 주제가, HDP는 LDA보다 세부적인 주제가 도출된 것을 확인할 수 있다.

토픽모델링을 활용한 연구동향 분석은 각 토픽마다 고유한 키워드를 통해 연구 주제 영역을 판단하고 이를 분석해 연구동향을 파악하는 것이다. 따라서 각 토픽들은 여러 토픽에 중

〈표 5〉 LDA와 HDP 토픽별 주제 비교

	LDA	HDP
Topic1	전자기록	민간기록의 아카이빙
Topic2	국가기록원	전자기록물과 디지털 기술
Topic3	아카이브 구축과 기록물 활용 방안	기록관리에 관한 표준 및 규칙
Topic4	보존	공공기록물의 기록관리
Topic5	기록물의 아카이빙	기록정보서비스의 디지털 활용방안
Topic6	기록정보서비스	기록관리현황 및 개선방안
Topic7	공공기관의 기록	국내외 기록관리체제와 아키비스트의 전문성
Topic8	국내외 기록관리 현황 및 사례	공공기관의 기록물 활용 방안
Topic9	메타데이터와 기술	다양한 기록물의 수집 방안
Topic10	전자기록관리시스템	국내 역사기록물

복되어 나타나는 키워드보다 해당 토픽의 특징을 나타내는 특수한 키워드를 많이 가지고 있어야 명확한 연구 주제 영역을 구분할 수 있다. 이에 HDP 토픽모델링은 일반적인 키워드, 중복된 키워드가 적으며 각 토픽별로 특징을 분명하게 파악할 수 있는 키워드가 뚜렷하게 나타나기 때문에 세부적인 연구동향 분석을 하는데 있어 LDA 토픽모델링보다 더욱 효율적일 것으로 판단된다. 반면 LDA는 해당 분야에서 공통적으로 다루고 있는 주제 키워드 분석에 유리해 거시적인 연구동향 파악에 적합한 방법으로 볼 수 있다.

Ⅴ. 결론

본 연구는 토픽모델링을 활용한 국내 기록관리학 연구동향 분석을 위해 데이터마이닝 기법 중 하나로 연구동향분석에 주로 활용되는 LDA 토픽모델링과 LDA 토픽모델링을 응용한 토픽모델링 방법인 HDP 토픽모델링을 적용, 그 결과를 비교 분석하였다. 먼저, 국내 기록관리학 분야 학술지 2종과 문헌정보학 분야 학술지 4종의 발간일부터 2016년까지 게재된 모든 논문의 논문명, 초록, 키워드 등을 자동수집하고 기록관리와 관련된 모든 논문을 추출하였다. 총 1,027건의 데이터를 수집하였고 형태소분석, 명사추출, 벡터화 등 적절한 전처리과정을 거쳤다. 전처리과정을 거친 문서집합은 토픽모델링에 적합한 포맷으로 변경한 후 LDA 토픽모델링, HDP 토픽모델링을 각각 적용하였다. 또한 토픽모델링 시각화 도구인 LDAvis를 활용하여 각 토픽모델링 결과를 시각화하였다.

LDA 토픽모델링과 HDP 토픽모델링을 수행한 결과, LDA 토픽모델링은 빈도수가 높은 키

워드에 많은 영향을 받았으며 각 토픽의 특징을 파악하기 어려운 일반적인 키워드가 많았다. 시각화를 살펴보면 대부분의 토픽이 특정 부분에 과도하게 집중되어 있다. 각 토픽마다 중복된 키워드와 일반적인 키워드가 많이 존재하기 때문에 각 토픽의 특징을 파악하기 어렵다. 반면에 HDP 토픽모델링은 빈도수가 높은 키워드에 상대적으로 영향을 적게 받으며 각 토픽마다 유일하게 등장하는 키워드가 많다. 또한 시각화를 살펴보면 LDA 토픽모델링과 같이 전체 토픽이 과도하게 특정 부분에 집중되어 있지 않고 일부 소영역을 이루며 고르게 분포되어 있다. 각 토픽의 특징을 파악할 수 있는 고유한 키워드가 많으면 연구 주제 영역을 뚜렷하게 구분할 수 있으며 이를 통해 연구동향을 파악할 수 있다. 이에 HDP 토픽모델링은 LDA 토픽모델링보다 세부 주제별 연구동향 분석을 하는데 더욱 효율적일 것으로 판단된다.

본 연구는 토픽모델링을 활용한 분석에서 주로 사용되었던 LDA 토픽모델링을 보완한 기법으로 국내 연구에서는 자주 사용되지 않았던 HDP 토픽모델링을 적용하여 차이점을 도출한 것에 의의가 있다. 본 연구의 향후 연구 방향으로는 두 토픽모델링 기반으로 도출된 토픽들에 대해 기록관리학에 관한 분류표, 기록관리학 분야 전문가의 조언 등을 통해 각 토픽에 대한 명칭과 주제를 정하여 연구동향을 심층적인 분석이 필요하며, 이후 시계열에 따른 토픽모델링 분석으로 확장하여 시간에 따른 국내 기록관리학 연구 변화와 각 시기마다 관심이 집중된 주제, 관심이 멀어진 주제 등을 파악하는 연구가 진행될 수 있다.

참고문헌

김규환, 장보성, 이현정. 2009. 우리나라 기록관리학 분야의 연구영역 분석 - 논문제목의 구문 및 의미 구조를 중심으로. 『한국문헌정보학회지』, 43(3): 417-439.

김규환, 남영준. 2009. 국내 기록관리학 분야 학회지 논문 분석을 통한 연구동향 연구. 『한국문헌정보학회지』, 43(4): 217-239.

김남규, 이동훈, 최호창, William Xiu Shun Wong. 2017. 텍스트 분석 기술 및 활용 동향. 『한국통신학회논문지』, 42(2): 471-492.

김상겸. 2016. 『토픽모델링을 이용한 국내 산업공학 연구동향 분석』. 석사학위논문, 서울과학기술대학교 일반대학원 데이터사이언스학과.

김태경, 최회련, 이홍철. 2016. 토픽 모델링을 이용한 핀테크 기술 동향 분석. 『한국산학기술학회논문지』, 17(11): 670-681.

나상태, 안주언, 정민호, 김자희. 2017. 동적 토픽분석을 활용한 스마트그리드 연구동향 분석. 『전기학회논문지』, 66(4): 613-620.

남태우, 이진영. 2009. 우리나라 기록관리학 연구 동향 분석.『한국도서관 · 정보학회지』, 40(2): 451-472.

남춘호. 2016. 일기자료 연구에서 토픽모델링 기법의 활용가능성 검토.『비교문화연구』, 22(1): 89-135.

박주섭, 홍순구, 김종원. 2017. 토픽모델링을 활용한 과학기술동향 및 예측에 관한 연구.『한국산업정보학회논문지』, 22(4): 19-28.

박자현, 송민. 2013. 토픽모델링을 활용한 국내 문헌정보학 연구동향 분석.『정보관리학회지』, 30(1): 7-32.

서성훈. 2016.『BM 특허 토픽 모델링을 이용한 핀테크 기술동향 분석』. 석사학위논문, 서울과학기술대학교 일반대학원 데이터사이언스학과.

손혜인, 남영준. 2016. 기록관리학 분야 국내 학술지의 연구동향에 관한 연구 -『한국기록관리학회지』와『기록학연구』를 중심으로.『정보관리학회지』, 33(1): 85-110.

신규식, 최회련, 이홍철. 2015. 신재생에너지 동향 파악을 위한 토픽 모형 분석.『한국산학기술학회논문지』, 16(9): 6411-6418.

유소영. 2015. 자아 중심 네트워크 분석과 동적 인용 네트워크를 활용한 토픽모델링 기반 연구동향 분석에 관한 연구.『정보관리학회지』, 32(1): 153-169.

이재윤, 문주영, 김희정. 2007. 텍스트 마이닝을 이용한 국내 기록관리학 분야 지적구조 분석.『한국문헌정보학회지』, 41(1): 345-372.

진설아, 송민. 2016. 토픽 모델링 기반 정보학 분야 학술지의 학제성 측정 연구.『정보관리학회지』, 33(1): 7-32.

최이랑. 2015. 국내 기록관리학 연구동향에 관한 연구 - 최근 10년간(2004-2013) 학술논문을 중심으로.『기록학연구』, 43: 147-177.

한국기록관리학회. 2013.『기록관리론 : 증거와 기억의 과학』. 성남: 아세아문화사.

Carson Sievert and Kenneth E. Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. *proceedings of workshop on interactive language learning, visualization, and interfaces,* Baltimore, Maryland.

Chong Wang, John Paisley and David M. Blei. 2011. Online Variational Inference for the Hierarchical Dirichlet Process. *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, Fort Lauderdale, FL.

David M. Blei. 2012. Provavilistic Topic Models. *Communications of the ACM*, 55(4): 77-84.

David M. Blei, Andrew Y. Ng and Michael I. Jordan. 2003. Latent Dirichlet Allocation.

Journal of Machine Learning Research, 3: 993-1022.

Gensim Home Page. <https://radimrehurek.com/gensim/> [cited 2017. 9. 15].

Jason Chuang, Christopher D. Manning and Jeffrey Heer. 2012. Termite: Visualization Techniques for Assessing Textual Topic Models. *Advanced Visual Interfaces,* 12: 21-25.

Konlpy Home Page. <http://konlpy-ko.readthedocs.io/ko/v0.4.3/> [cited 2017. 9. 15].

Loet Leydesdorff and and Adina Nerghes. 2017. Co-word Maps and Topic Modeling: A Comparison Using Small and Medium-Sized Corpora (N<1,000). *Journal of the Association for Information Science and Technology,* 68(4): 1024-1035.

Teh, Yee Whye, Michael I. Jordan, Matthew J. Beal and David M. Blei. 2007. Hierarchical Dirichlet processes. *Journal of the American Statistical Association,* 101(476): 1566-1581.

국한문 참고문헌의 영문 표기
(English translation / Romanization of reference originally written in Korean)

Choi, Yilang. 2015. A Study on the Research Trends of Archival Studies in Korea : Focused on Research Papers between 2004 and 2013. *The Korean Journal of Archival Studies,* 43, 147-177.

Gyu-Hwan Kim, Young-Joon Nam. 2009. A Study on the Research Trends of Records and Archives Management in Korea through an Analysis of Journal Articles. *Journal of The Korean Society for Library and Information Science*, 43(4): 217-239.

Gyu-Hwan Kim, Bo-Seong Jang, Hyun-Jung Yi. 2009. A Study on Intellectual Structure of Records Management and Archives in Korea : Based on Syntactic and Semantic Structure of Article Titles. *Journal of The Korean Society for Library and Information Science*, 43(3): 417-439.

Namgyu Kim, Donghoon Lee, Hochang Choi, Willam Xiu Shun Wong. 2017. Investigations on Techniques and Applications of Text Analytics. *The Journal of Communications and Information Sciences*, 42(2): 471-492.

Kim, Sang Kyoum. 2016. *A Study on the Research Trends in Domestic Industrial Engineering using Topic Modeling*. master's thesis, Seoul National University of Science and Technology, Seoul, Korea.

Seol A Jin, Min Song. 2016. Topic Modeling based Interdisoiplinarity Measurement in the Informatics Related Journals. *Journal of the Korean Society for Information Management,* 33(1): 7-32.

So-Young Yu. 2015. Combining Ego-centric Network Analysis and Dynamic Citation Network Analysis to Topic Modeling for Characterizing Research Trends. *Journal of the Korean Society for Information Management*, 32(1): 153-169.

TaeKyung Kim, HoeRyeon Choi, HongChul Lee. 2016. A Study on the Research Trends in Fintech using Topic Modeling. *The Journal of Korea Academy Industrial Cooperation Society,* 17(11): 670-681.

Jae-Yun Lee, Ju-Young Moon, Hee-Jung Kim. 2007. Examining the Intellectual Structure of Records Management & Archivial Science in Korea with Text Mining. *Journal of The Korean Society for Library and Information Science*, 41(1): 345-372.

Nahm, Choon-Ho. 2016. An Illustrative Application of Topic Modeling Method to a Farmer's Diary. *Cross-Cultural Studies,* 22(1): 89-135.

Sang-Tae Na, Joo-Eon Ahn, Min-Ho Jung, Ja-Hee Kim. 2017. Research Trend Analysis for Smart Grids Using Dynamic Topic Modeling. *The transactions of The Korean Institute of Electrical Engineers*, 66(4): 613-620.

Tea-Woo Nam, Jin-Young Lee. 2009. A Study on the Research Trends of Records and Archives Management in Korea. *Journal of Korean Library and Information Science Society,* 40(2): 451-472.

Ja-Hyun Park, Min Song. 2013. A Study on the Research Trends in Library & Infromation Science in Korea using Topic Modeling. *Journal of the Korean Society for Information Management*, 30(1): 7-32.

Park Ju Seop, Hong Soon-Goo, Kim Jong-Weon. 2017. A Study on Science Technology Trend and Prediction Using Topic Modeling. *Journal of the Korea Industrial Information Systems Research*, 22(4): 19-28.

Records Management & Archives Society Of Korea. 2013. *Records & Archives Management*. Seongnam: Asian cultural history.

Seo, Seong Hun. 2016. *Fintech trend analysis using topic modeling of BM patents*. master's thesis, Seoul National University of Science and Technology, Seoul, Korea.

Shin, Kyoo-Sik, Choi, Hoe-Ryeon, Lee, Hong-Chul. 2015. Topic Model Analysis of Research Trend on Renewable Energy. *The Journal of Korea Academy Industrial Cooperation Society*, 16(9): 6411-6418.

Hye In Sohn, Young Joon Nam. 2016. A Study on the Research Trends of Archives Management in Korea : Focused on the Journal of Records - Management & Archives Society of Korea and The Korean Journal of Archival Studies. *Journal of Korea Society for Information Management*, 33(1): 85-110.