토픽모델링을 활용한 국내 문헌정보학 연구동향 분석*

A Study on the Research Trends in Library & Information Science in Korea using Topic Modeling

박자현 (Ja-Hyun Park)**
송 민 (Min Song)***

초 록

본 연구는 국내 문헌정보학 분야의 연구동향을 규명하기 위하여 문헌정보학 주요 학술지인, 정보관리학회지, 한국문헌정보학회지, 한국도서관·정보학회지, 한국비블리아학회지의 1970년도부터 2012년도까지 발표 논문 초록을 수집하여 LDA(Latent Dirichlet Allocation)기반의 토픽 모델링 실험을 수행하였다. 그 결과를 종합하면 다음과 같다. 첫째, 토픽모델링 실험에서 도출된 연구주제를 문헌정보학 주제분류표와 비교·분석한 결과, '정보학'영역의 디지털도서관, 이용연구, 인터넷, 전문가시스템, 계량정보학, 자동화, 정보검색, 정보시스템, '도서관 서비스'영역의 정보서비스, 도서관 유형별 서비스, 이용자 교육/정보리터러시, 서비스 평가, '문헌정보학 기초'영역의 도서관과 사회, 전문성, '자료조직'영역의 분류, 편목, 메타데이터, '도서관 경영'영역의 도서관 평가, 장서개발/관리, '서지학'영역의 고서지, '도서관 체제'영역의 도서관 및 정보정책, '출판'영역의 도서/출판, '기록관리학'영역의 하위주제 등과 연결할 수 있었다. 또한 가장 많은 연구주제가 발견된 학문영역은 정보학과 도서관서비스로 나타났다. 둘째, 문헌정보학의 주요 연구주제에서 도서관 유형별 서비스 및 평가, 인터넷, 메타데이터의 연구주제는 상승세를 보였으나, 도서, 분류, 편목, 고서지에 관한 연구주제는 하강세를 보였다. 셋째, 학술지를 구분하여 비교·분석한 결과, 정보관리학회지는 도서관에 관한 연구주제보다 정보학에 관한 연구주제가 많이 출현하였고, 한국문헌정보학회지와 한국도서관·정보학회지, 한국비블리아학회지는 도서관에 관한 연구주제가 정보학에 관한 주제보다 많이 나타났다.

ABSTRACT

The goal of the present study is to identify the topic trend in the field of library and information science in Korea. To this end, we collected titles and abstracts of the papers published in four major journals such as Journal of the Korean Society for information Management, Journal of the Korean Society for Library and Information Science, Journal of Korean Library and Information Science Society, and Journal of the Korean BIBLIA Society for library and Information Science during 1970 and 2012. After that, we applied the well-received topic modeling technique, Latent Dirichlet Allocation(LDA), to the collected data sets. The research findings of the study are as follows: 1) Comparison of the extracted topics by LDA with the subject headings of library and information science shows that there are several distinct sub-research domains strongly tied with the field. Those include library and society in the domain of "introduction to library and information science," professionalism, library and information policy in the domain of "library system," library evaluation in the domain of "library management," collection development and management, information service in the domain of "library service," services by library type, user training/information literacy, service evaluation, classification/cataloging/meta-data in the domain of "document organization," bibliometrics/digital libraries/user study/internet/expert system/information retrieval/information system in the domain of "information science," antique documents in the domain of "bibliography," books/publications in the domain of "publication," and archival study. The results indicate that among these sub-domains, information science and library services are two most focused domains. Second, we observe that there is the growing trend in the research topics such as service and evaluation by library type, internet, and meta-data, but the research topics such as book, classification, and cataloging reveal the declining trend. Third, analysis by journal show that in Journal of the Korean Society for information Management, information science related topics appear more frequently than library science related topics whereas library science related topics are more popular in the other three journals studied in this paper.

키워드: 문헌정보학, 연구동향, 토픽모델링, 텍스트 마이닝, LDA
library and information science, research trends, topic modeling, text mining, latent Dirichlet allocation

---

* 본 연구는 2012년 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2012-2012S13A2033291).
** 연세대학교 문헌정보학과 일반대학원(parkja04@ksdc.re.kr) (제1저자)
*** 연세대학교 문헌정보학과 교수(min.song@yonsei.ac.kr) (교신저자)
■ 논문접수일자 : 2013년 1월 30일 ■ 최초심사일자 : 2013년 2월 15일 ■ 게재확정일자 : 2013년 3월 17일
■ 정보관리학회지, 30(1), 7-32, 2013. 〔http://dx.doi.org/10.3743/KOSIM.2013.30.1.007〕

1. 서 론

1.1 연구배경

우리나라에서 1957년 연세대학교에 도서관학과(현 문헌정보학과)가 설립되면서 현대적 의미의 문헌정보학 교육이 도입이 된 후 반세기에 가까운 시간이 흘렀다. 이 과정에서 국내의 문헌정보학계는 수많은 학문적 발전을 이루었으며, 정보환경의 급격한 변화와 이로 인한 혼란과 갈등 속에서도 성장을 거듭하여 왔다.

이와 같은 발전에 견인차의 역할을 한 것은 1970년에 창립한 한국도서관학회(현 한국문헌정보학회)를 시발점으로 하여 이후 설립된 한국정보관리학회, 한국도서관·정보학회, 한국비블리아학회 등 문헌정보학 분야 주요 학회들의 활발한 연구 활동이라고 할 수 있다. 이들은 국내 문헌정보학의 대표적인 학술지를 발간할 뿐만 아니라 연구자들이 상호간의 연구결과를 공유할 수 있도록 정보교환의 장을 마련하는 등 다양한 연구 사업을 수행함으로써 학문 발전에 기여하고 있다.

한편, 국내 문헌정보학의 연구 활동이 활발해짐에 따라 연구자들은 문헌정보학이 어떤 주제를 중심으로 연구를 수행하고 있고, 문헌정보학자의 관심분야가 어떻게 변화하고 있으며 학문적 유행은 어떤 양상을 보이는지 등 그 현상을 밝히기 위해 다각적인 측면에서의 연구동향 분석을 시도하였다. 그러나 이러한 연구 중 다수는 문헌정보학의 특정 분야에 집중하여 분석하거나, 제한된 기간에 발표한 논문을 대상으로 분석하였기에 문헌정보학 영역 전반의 동향을 파악하는 것과는 거리가 있다. 또한 문헌정보학의 연구 동향 분석은 양적분석과 주제별 내용분석이 주를 이루었으나, 연구 주제를 파악하는 과정에서 연구자마다 다른 기준에 근거하여 논문을 분류하는 경향을 보이며, 분석도구를 사용하는 경우는 그 방법이 상이하고 일관되지 않아 논문마다 도출된 연구결과를 통합적으로 비교하여 분석하기가 어려운 실정이다. 뿐만 아니라 국내 문헌정보학의 연구동향 분석은 연구자가 직접 내용을 분석하고 판단해내는 작업을 수행한 연구가 많았다. 그렇기에 보다 정확한 결과가 나올 수도 있지만, 한편으로는 연구자의 주관적 가치와 개인적인 의견이 반영될 위험성이 있으며 많은 분량의 데이터를 소화하기 어렵다는 단점이 있다.

본 연구는 양적분석과 기존의 내용분석이 갖는 단점을 텍스트 마이닝 기법을 적용하여 보완하고자 한다. Kao와 Poteet(2007)는 텍스트 마이닝을 자유롭거나 비 구조화된 텍스트로부터 흥미롭고 일상적이지 않은 지식을 발견, 추출하는 것이라고 정의하였다. 텍스트 마이닝 기법을 활용한다면, 문헌정보학의 주요 학술지가 창간한 1970년대부터 현재까지 40년간에 걸친 우리나라 문헌정보학 연구의 방대한 텍스트 데이터를 대상으로 객관적인 분석이 가능해지고 연구의 주제와 동향을 다각도로 조명해볼 수 있다. 국내에서 문헌정보학의 연구 동향을 텍스트 마이닝 기법을 적용하여 분석하는 연구는 문헌 클러스터링, 프로파일링 기법, 네트워크 텍스트 분석 등의 방법으로 연구되었다. 그러나 아직까지 텍스트 마이닝 기법 중 문헌에서 숨겨져 있는 주제들을 찾아내기 위해 개발된 통계 추론 모델인, 토픽모델링을 활용하여 우리나라 문헌정보학 연구주제의 변화와 그 양상을 종합적으로 분

석한 연구는 수행된 바 없다.

따라서 본 연구는 문헌정보학 분야 주요 학술지인, 정보관리학회지, 한국문헌정보학회지, 한국도서관·정보학회지, 한국비블리아학회지의 1970년대부터 2012년도까지 발표 논문 초록을 수집하여 LDA(Latent Dirichlet Allocation) 기반의 토픽 모델링 실험을 수행함으로써 문헌정보학자들이 관심을 갖는 주요 연구주제가 무엇인지 규명하고자 한다. 이를 통해 연구동향 분석뿐만 아니라, 계량정보학의 내용분석에 대하여 새로운 방법론을 제시하고자 한다.

1.2 선행연구 개관

1.2.1 문헌정보학 연구동향 관련 연구

문헌정보학 연구동향 분석에 대한 선행연구에서 문헌정보학의 학위논문과 학술지논문 가운데 이 모두를 분석한 연구는 소수이며, 대부분 한 가지를 분석하고 있다. 그 중 학술지논문을 포함하여 분석한 연구는 한상완, 조인숙(1996), 서은경(1997), 정진식(2001), 손정표(2003), 오세훈(2005), 서은경(2010), 정재영, 박진희(2011) 등의 연구가 있다.

한상완과 조인숙(1996)은 우리나라 문헌정보학분야의 학술지논문의 내용분석을 위해 한국문헌정보학회지, 정보관리학회지 등 4개 학술지에 게재된 논문의 내용을 연구자, 논문주제 및 참고문헌의 세 측면으로 나누어 조사하였다. 그 결과 우리나라 주제별 분포는 문헌정보학분야의 관종별 도서관 중 대학도서관(17.0%)이, 정보학분야는 정보검색(22.7%), 서지학분야는 고활자, 판본학 등을 포함한 고서지(72.4%)분야라는 연구 결과를 제시하였다. 즉 한국문헌정보학회지와 도서관 논집에는 문헌정보학, 정보학 및 서지학의 논문이 종합적으로 게재하고 있으며, 정보관리학회지에는 정보학의 논문이, 서지학연구에는 서지학에 관한 논문만이 게재되고 있음을 발견하였다.

서은경(1997)은 정보학분야의 연구동향을 규명하기 위하여 정보학분야의 연구영역, 정보학 연구의 타학문 주제의존도, 그리고 정보학의 학제적 구조 및 변화를 분석하였다. 내용분석을 통한 개념적 방법과 인용패턴을 이용한 계량적 방법을 이용하여 한국과 미국의 연구동향과 연구의 변화를 조사하였다. 그 결과, 정보학자들은 지속적으로 '자동문헌분석 및 검색 그리고 그 응용' 분야를 주요 연구 관심사로 두고 연구해 왔음을 알 수 있었고, 정보학 연구 영역이 점차 넓어져 가는 추세를 확인하였다. 결론적으로 대부분의 정보학자들은 계속해서 '도서관 관련 정보학' 연구를 주로 수행하고 있지만, 정보학 및 도서관학 분야 이외의 주제영역(특히 전산학 분야)의 이론 및 기법을 더 많이 이용함에 따라 점점 '컴퓨터 관련 정보학' 연구가 부각되고 있음을 발견하였다.

정진식(2001)은 한국문헌정보학분야의 연구동향을 규명하고 학문의 연구영역과 타 학문과의 주제 의존도 및 문헌정보학의 지적구조변화를 조사하였다. 인용패턴을 이용한 계량적 방법을 이용, 연구의 동향과 변화를 대응분석 방법을 적용하였다. 조사결과 문헌정보학을 비롯한 정보학분야에 대한 의존도가 크게 나타났으며, 학문의 이론 및 응용기법의 활용도가 높은 컴퓨터 관련 전산학이 부상될 것으로 예상하였다.

손정표(2003)는 문헌정보학 관련 7개 학술지,

16개 대학 문헌정보학과 창립 기념논문집, 3개 전문기관지에 수록된 논문을 분석하여 1957~2002년간의 연구동향을 제시하였다. 그 결과, 주제 영역별 연평균 발표량의 순위는 정보학, 서지학, 도서관 경영, 자료조직, 공공봉사, 문헌정보학 기초, 도서 및 도서관사, 장서개발(관기) 순으로, 각 영역의 주제별로는 문헌정보학교육, 도서관 경영론 전반, 장서개발정책・방침과 장서평가, 참고・정보봉사, 분석서지학, 도서관사, 정보검색 분야가 각 영역의 타주제보다 높게 나타났다.

오세훈(2005)은 1946년부터 2004년까지의 전문기관지 및 학술지논문 2,571편과 이에 인용된 문헌 30,418편을 대상으로 문헌과 피인용 외국 문헌정보학 문헌의 주제를 분석하기 위하여 주제영역을 설정하고, 간행시기와 인용시기 등 연관성을 5개의 가설을 설정하여 통계적으로 검증하였다. 그 결과 논문의 주제와 간행시기 사이에 상관성이 있음을 입증하였다.

서은경(2010)은 정보관리학회지를 대상으로 발표 시점을 25년 단위로 구분하여, 각 논문의 주제 분포와 주제영역별 논문의 증감 등을 살펴보았다. 주요 대주제 영역은 정보서비스, 정보조직, 정보시스템으로, 소주제 영역은 도서관서비스, 이용자연구, 자동문헌처리, 도서관통합시스템, 시소러스/온톨로지, 디지털도서관으로 나타났다고 보고하고 있다.

정재영과 박진희(2011)는 문헌정보학 학술지의 최근 10년간 발표된 연구논문 2,165건에 대한 내용분석을 통해 현장연구의 현황 즉, 양적 변화, 주제, 연구자, 연구비 지원여부 등을 분석하였다. 연구결과, 전체 논문 중 현장연구는 691건으로 31.9%를 차지하는 것으로 조사되었다. 이들 현장연구의 주제를 도서관 유형별로 구분한 결과 공공도서관이 246편(35.6%)으로 가장 많았으며, 대학도서관이 238편(34.4%), 학교도서관 141편(20.4%), 그리고 기타 66편(9.6%)으로 조사되었다. 문헌정보학 관련 주제별로 구분한 결과는 도서관운영 관련 연구가 328편(47.5%)으로 가장 많았으며, 정보봉사, 정보시스템 및 전자도서관, 이용자교육, 독서교육 순으로 조사되었다.

선행연구를 조사한 결과, 이들의 연구 중 다수는 연구자가 직접 내용을 분석하고 판단해내는 작업을 수행하였음을 알 수 있다. 이러한 방법은 연구자의 가치와 개인적인 의견이 반영될 위험성이 있으며, 많은 분량의 데이터를 소화하기 어렵다는 단점이 있다.

1.2.2 텍스트 마이닝 기법 활용 관련 연구

텍스트 마이닝 기법을 활용한다면, 방대한 텍스트 데이터를 대상으로 객관적인 분석이 가능해지고 각 분류에 따라 연구의 주제와 동향을 다각도로 조명해볼 수 있다. 연구동향 분석에 텍스트 마이닝 기법을 적용한 국외연구는 다음과 같다.

Griffiths와 Steyvers(2004)는 PNAS(Proceedings of the National Academy of Sciences of the United States of America, 미국국립과학원회보)의 초록을 대상으로 토픽모델링 기법을 적용하고 분석하였다. 그 결과 PNAS 초록에 대한 의미 있는 토픽을 발견함과 동시에 PNAS에서 근래에 가장 활발하게 연구되고 있는 연구주제(hot topics)와 점차 연구되지 않는 연구주제(cold topics)를 밝혔다.

Wang과 McCallum(2006)은 LDA 토픽모

델 중 시간의 흐름에 따라 토픽이 어떻게 변화하는지 살펴볼 수 있는 TOT(Topics Over Time) 모델을 소개하였다. 또한 개인이메일, 1987년부터 2003년도까지의 NIPS(Neural Information Processing Systems) 컨퍼런스 프로시딩, 200년 동안의 미연방 대통령 연설문을 대상으로 TOT 모델을 적용함으로써 각 텍스트에 대한 토픽을 발견하였으며, 그 시계열적 추이를 분석하였다.

Mimno와 McCallum(2006)은 Rexa 자동 인용색인 시스템에서 수집한 컴퓨터 공학 분야의 300,000편의 논문을 대상으로 구절(phrase) 기반의 토픽 도출모델인 Topical N-Grams (TNG)를 적용함으로써 영향력 측정에 토픽 모델이 유용하게 적용될 수 있음을 증명하고자 하였다.

Blei(2012)는 'Science' 저널의 17,000편의 논문과 'Yale Law' 저널을 대상으로 LDA 모델을 활용하여 토픽 모델을 도출한 사례를 소개하면서 토픽모델이 정치학, 심리학뿐만 아니라 계량서지분석에도 활용될 수 있음을 언급하였다.

국내 문헌정보학자가 텍스트 마이닝의 주요 기법을 적용한 연구는 국내 기록관리학 분야를 분석한 이재윤, 문주영, 김희정(2007), 문헌정보학 아카이브 관련 연구 영역의 지적구조를 파악한 Kim과 Lee(2009), 네트워크 텍스트 분석을 통하여 문헌정보학 최근 연구 경향을 분석한 조재인(2011), LDA 등의 텍스트 마이닝 기법을 활용하여 생물정보학을 분석한 Song과 Kim(Scientometrics, in press) 등의 연구가 있다.

이재윤, 문주영, 김희정(2007)은 텍스트 마이닝의 주요 기법인 문헌 클러스터링과 문헌 유사도 네트워크 분석을 적용하여 기록관리학 연구의 지적구조를 분석하였다. 군집단위 지적구조 분석 결과, 국내에서 수행된 기록관리학 영역의 핵심적인 주제 영역은 '전자기록관리・디지털 보존', '기록관리정책・제도', '기록물 기술/목록', '기록관리학 영역・교육'이었으며, 문헌단위 지적구조 분석을 통하여서는 '디지털 아카이빙' 주제 영역이 중심을 이루고 있음을 확인할 수 있었다.

Kim과 Lee(2009)는 문헌정보학 분야에서 아카이브 관련 연구 영역의 지적구조를 프로파일링 분석 방법을 활용하여 분석하였다. 그 결과, 아카이브 관련 여섯 개의 대표적인 저널을 확인할 수 있었으며, 디지털 도서관을 가장 핵심적인 주제 영역으로, 전자 매체를 가장 핵심적인 객체로 도출하였다.

조재인(2011)은 최근 7년간 문헌정보학분야에 게재된 논문 1,752건을 대상으로 빈도분석과 네트워크 텍스트 분석을 실시하여 다양한 주제 개념의 분포와 그 관계성을 도출하였다. 그 결과, 최근 7년간 문헌정보학 분야는 "공공도서관"과 "대학도서관" 개념을 중심으로 하는 연구가 가장 높은 출현 빈도를 보였으며, "평가", "교육", "웹"은 가장 높은 연결 중심성을 나타내었다.

Song과 Kim(2012)은 생물정보학 분야의 지적구조를 분석하기 위하여 LDA(Latent Dirichlet Allocation)기반 토픽모델링 등의 텍스트 마이닝 기법을 활용하였다. 그 결과, 생물정보학 분야에서 발표논문의 수가 2003년을 기점으로 높은 상승세를 보이고 있으며, 미국과 유럽국가가 높은 연구생산성을 보이고 있음을

확인하였다. 또한 토픽 모델링과 단어 동시발생빈도 분석을 통해 생물정보학의 계산적 측면(computational aspects)보다 생물학적 측면(biological aspects)이 더 부각되고 있음을 밝혔다. 반면, 페이지 랭크를 통한 분석에서 상위에 해당하는 논문 중 생물정보학의 계산적 측면에 해당하는 논문이 생물학적 측면의 논문보다 많았다.

그러나 아직 국내 연구에서 문헌정보학의 학문영역 전반을 대상으로 LDA(Latent Dirichlet Allocation)기반의 토픽 모델링 기법을 활용함으로써 문헌정보학의 연구동향을 파악하는 연구는 이루어지지 않았다. 따라서 본 연구는 토픽 모델링 기법을 통해 문헌정보학자들이 공통적으로 관심을 갖는 연구주제가 무엇인지를 규명하고 연도별 연구주제의 변화를 파악하며, 문헌정보학 주요 학술지별 연구주제를 비교하고자 한다.

1.3 연구문제

본 연구를 통해 알아보고자 하는 연구문제는 다음과 같다.

연구문제 1. 문헌정보학자들이 관심을 갖는 주요 연구주제가 무엇인지를 규명하고 가장 많은 연구주제가 발견된 학문영역은 무엇인지 밝힌다.

연구문제 2. 문헌정보학 주요 연구주제의 연도별 추이를 분석하고 근래에 활발하게 연구되고 있는 연구주제(hot topics)와 점차 연구되지 않고 있는 연구주제(cold topics)를 밝힌다.

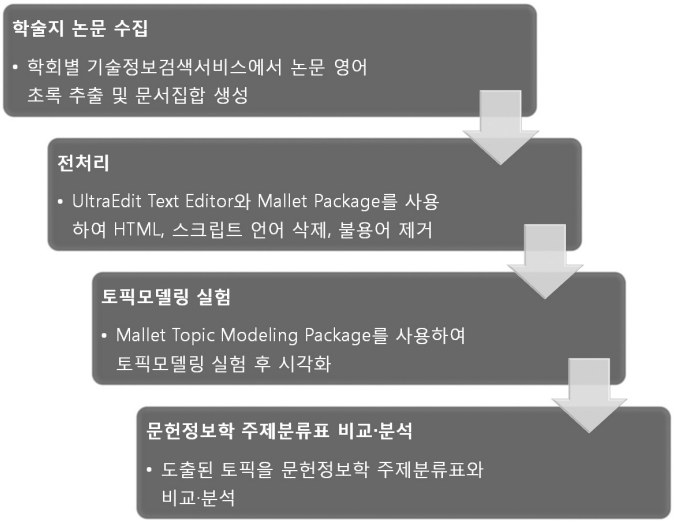
연구문제 3. 문헌정보학 학술지별 주요 연구주제를 밝히고 학술지마다 관심을 갖는 주요 연구주제가 어떤 공통점과 차이점을 갖는지 비교·분석한다.

2. 연구 설계

토픽모델링 기법을 활용하여 문헌정보학 연구동향 분석을 위해 다음과 같은 과정을 거쳤다. 우선 학술지 논문의 수집은 학회별 기술정보검색서비스[1]의 논문 상세정보 페이지를 자동으로 추출하여 문서집합을 생성한 후, 전처리 과정을 거쳤다. 전처리는 UltraEdit Text Editor를 사용하여 분석 시에 불필요한 데이터를 삭제하고 데이터의 형식을 통일했다. 즉 HTML, 스크립트 언어 등을 삭제하고 영어초록 텍스트만을 추출하였다(〈그림 1〉 참조).

또한 JAVA로 구현된 Mallet Package를 활용하여 문자열을 자르고, 불용어를 제거하였다. 논문 초록의 분석 역시 Mallet Topic Modeling Package를 사용하여 토픽모델링을 수행한 후 각각의 결과를 그래프 등으로 시각화하였다. 마지막으로 결과를 의미 있게 해석하기 위하여 도출된 연구주제를 문헌정보학 주제분류표와 비교·분석하는 작업을 수행하였다.

1) 학회별 기술정보검색서비스: http://ocean.kisti.re.kr



〈그림 1〉 연구설계 개요

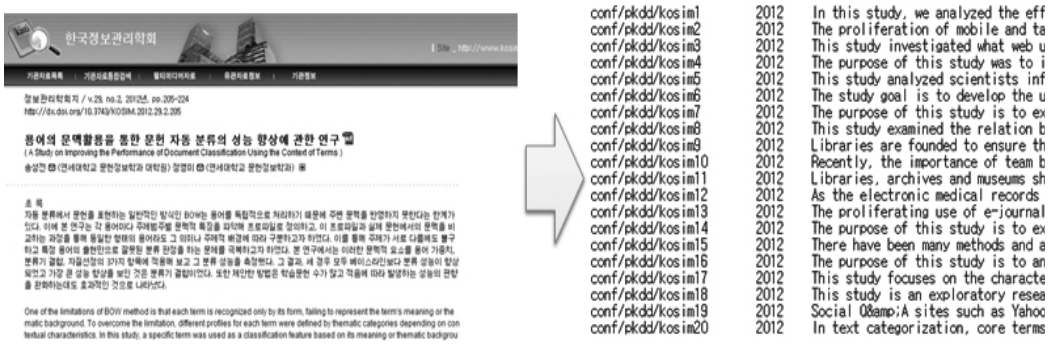
2.1 논문 초록 수집 및 전처리

학술지논문의 수집은 학회별 기술정보검색서비스의 웹 페이지를 크로울링(crawling)한 후 초록 정보만 자동 추출하였다. 수집대상은 한국연구재단에 등재된 문헌정보학 주요 학술지인, 정보관리학회지, 한국문헌정보학회지, 한국도서관·정보학회지, 한국비블리아학회지를 대상으로 하였고, 1970년부터 2012년도까지 총 3,834편의 영문초록을 수집하였다. 분석대상 학술지논문은 〈표 1〉과 같다.

연구대상 논문은 선정된 학술지에 발표된 논문 중 저자가 1명 이상 명시되고, 편집인에 의해 '논문'(article)이라고 명시된 것을 대상으로 하였으며, 학회 회칙, 논문심사 규정, 편집위원회 규정 등은 제외하였다. 또한 학회의 학술대회 논문집 역시 제외하였다. 학술지논문 중 영문초록이 없는 논문을 포함하여 특수기호 등의 원인으로 시스템 오류를 유발하는 논문은 수집과정에서 제외하였다. 수집된 데이터는 UltraEdit Text Editor를 사용하여 분석 시에 불필요한 데이터를 삭제하고 데이터의 형식을 통일했다.

〈표 1〉 분석대상

학술지명	권호	논문 수 (편)
정보관리학회지	1권 1호 ~ 29권 3호	945
한국문헌정보학회지	1권 ~ 46권 3호	1,140
한국도서관·정보학회지	1권 ~ 44권 3호	1,245
한국비블리아학회지	1권 1호 ~ 23권 3호	504
합계	총 40권 156호	3,834



〈그림 2〉 학회별 기술정보검색서비스

즉 HTML, 스크립트 언어 등을 삭제하고 영어 초록 텍스트만을 추출하였다. 이런 과정을 통해 인터넷상의 게시되어 있는 논문의 상세서지 내용 중 영문초록의 텍스트와 연도정보에 대한 메타데이터를 추출하여 분석에 활용할 수 있었다.

〈그림 2〉와 같이 JAVA crawler에 의해 수집한 학회별 기술정보검색서비스의 웹 정보를 텍스트 파일로 처리하여 분석에 활용하였다. 또한 기본적인 전처리 작업으로 일반적인 불용어(stop words) 외에 논문 초록에 자주 등장하는 불용어들을 추가하여 빈도 높은 기능어와 주제어로서 가치 없는 기타 고빈도어들을 제거하였다. 전처리가 완료된 모든 텍스트는 JAVA 프로그램을 활용하여 실험을 수행하였다.

2.2 분석방법

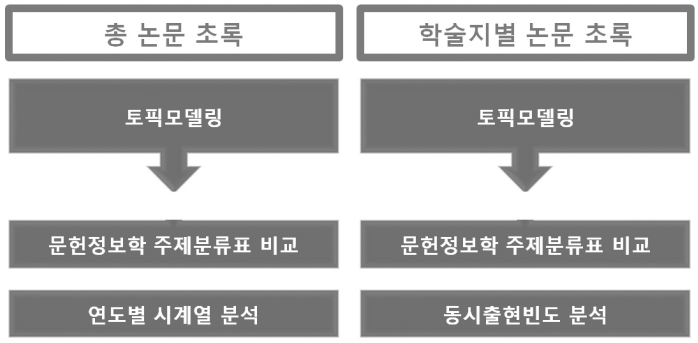
문헌정보학의 연구동향을 파악하기 위해, 실험을 수행하고 분석한 과정은 다음과 같다.

우선, 문헌정보학자들이 공통적으로 관심을 갖는 연구주제가 무엇인지를 규명하고자 수집된 전체 텍스트를 대상으로 토픽모델링을 수행하였다. 토픽모델링 결과를 해석하기 위하여 도출된 토픽들을 문헌정보학 주제분류표와 비교·분석한 후 연구주제가 발견된 학문영역을 분석하였다. 또한 위의 토픽모델링 실험에서 도출된 토픽마다 연도별 추이 그래프를 그림으로써 문헌정보학 주요 연구주제의 연도별 추이를 분석하고 문헌정보학에서 근래에 활발하게 연구되고 있는 연구주제(hot topics)와 점차 연구되지 않고 있는 연구주제(cold topics)를 밝혔다.

다음은 학술지마다 등장하는 연구주제의 공통점과 차이점을 밝히기 위하여 학술지별 텍스트를 대상으로 토픽모델링을 수행하였다. 마찬가지로 도출된 토픽들을 문헌정보학 주제분류표와 비교·분석하여 연구주제가 발견된 학문영역을 분석하였다. 또한 공통되는 연구주제를 발견하기 위하여 동시출현단어(word co-occurrence)의 출현 빈도수를 기준으로 각 학술지별 연구주제를 연결하였다(〈그림 3〉 참조).

2.2.1 토픽 모델링

토픽 모델링 기법은 각 문헌을 토픽의 확률적 혼합체로 표현하고 각 토픽을 단어의 분포로 표현함으로써 문헌의 구조를 예측하는 문헌 분석

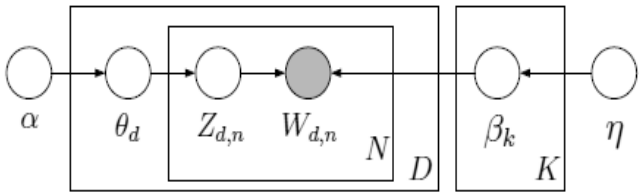


〈그림 3〉 분석방법

모델이다. 즉 문헌에서 숨겨져 있는 주제들을 찾아내기 위해 개발된 통계 추론 모델이다. 이 기법은 기존의 pLSI 기반 토픽 표현기법의 단점을 보완한, 의미론적 표현법의 새로운 패러다임으로 주목받고 있다(Griffiths & Steyvers, 2004).

이러한 토픽 모델링 기법 중 가장 대표적인 방법으로 LDA(Latent Dirichlet Allocation) 알고리즘이 있다. LDA 알고리즘은 생성모델로서 문헌 내의 숨겨져 있는 주제들을 찾아내는 알고리즘이다. 생성모델은 실제 문헌을 작성하는 과정으로 보고 문헌을 작성하기 위해 각 문헌에 어떤 주제들을 포함시킬 것인지, 또 그에 따라 어떤 단어들을 어떤 주제에서 선택하여 배치할 것인지를 각각의 파라미터로 모델링한다. 따라서 문헌, 단어 등 관찰된 변수(observed variable)를 통해 문헌의 구조와 같은 보이지 않는 변수(hidden variable)를 추론하는 것을 목적으로 하며 결과적으로 전체 문헌 집합의 주제들과 각 문헌별 주제 비율, 각 단어들이 각 주제에 포함될 확률들을 알아낼 수 있다. 단어가 독립적인 베이지언 기법(bayesian methods)과 달리 토픽 모델링은 단어가 독립적이지 않다는 가정에서 단어를 생성하는 조건에 따라 사후확률을 추론한다. 이러한 관계는 〈그림 4〉와 같은 확률적 그래프 모델로 표현할 수 있다(Blei, 2012).

관찰된 변수(observed variable)인, 단어와 문헌(코퍼스)을 통해 보이지 않는 변수(hidden variable), 즉 문헌별 주제 비율(topic proportions, Θ), 단어별 주제 지정(per-word topic assignment, Z), 그리고 주제(topic, β)를 예측한다. 〈그림 4〉의 α와 η는 하이퍼 파라미터로서 전체 문헌 집합에 대하여 동일한 값을 갖는다. α는 각 문서가 어떠한 주제 비율로 구성될지를 나타내는 Θ값을 결정하는 파라미터이다. Θ는 Dirichlet 분포를 따르는 값이며, 따라서 α값에 따라 Θ가 분포하게 될 Dirichlet 분포의 형태가 결정된다. 마찬가지로 η도 각 단어가 어떠한 주제들의 비율로 구성될지를 나타내는 β값을 결정하는 파라미터로서 η값에 따라 β가 분포하게 될 Dirichlet 분포의 형태가 결정된다. 또한 Θ는 각 문헌에 대한 주제 비율 값으로서 Dirichlet 분포를 따르며 Θ값에 따라 문헌 내에 존재하는 단어들의 주제인, z가 결정된다. 각 단어의 주제들을 나타내는 z값과 각 단어에 대한 전체 주제



〈그림 4〉 LDA의 확률 그래프 모델 표현

The purpose of this study is to verify the **factors**[17] determining core **journals**[7] in a field based upon the results of the **citation**[7] analysis of the **journals**[7] in the field of **Korean**[6] History(KH). In order to verify the determinant **factors**[17], dividing the **articles**[7] of the KH **journals**[7] into their novelty and **author**[11]'s seniority, the following hypotheses were established.

〈그림 5〉 토픽모델링의 주제 지정(topic assignment)에 의한 태그

에 대한 비율 β값에 따라 단어 w가 결정된다. 〈그림 5〉는 정보관리학회지의 초록 텍스트를 토픽모델링의 단어별 주제 지정(per-word topic assignment)에 의하여 태깅한 예시이다.

본 연구는 저자, 연도 등의 메타데이터를 제3의 파라미터로 지정하여 결과를 도출하는 다중 토픽모델링 기법(Multinominal topic model)을 활용하였다. 추가되는 메타데이터를 연도로 설정함으로써 초록에서 도출되는 연구주제마다, 그 연도별 추이를 분석하고자 하였다.

2.2.2 문헌정보학 주제분류표

토픽모델링 실험의 결과는 주제에 속하는 단어의 집합으로 표현된다. 따라서 본 연구에서 결과를 의미 있게 해석하기 위하여 연구주제를 표현하는 단어의 집합이 어떤 학문영역에 속하는지 판단할 필요가 있다. 따라서 문헌정보학 주제 영역과 그 영역의 하위, 세부주제로 구성된 문헌정보학 주제분류표[2]를 〈표 2〉와 같이 설정하였다.

문헌정보학 주제분류표와 토픽모델링 실험의 연구주제를 비교・분석하기 위하여 실험에서 도출된 동시출현 단어를 검색어로 연구주제에 해당하는 논문을 검색한 후, 문헌정보학 주제분류표를 기준으로 분류하였다. 따라서 복수주제가 결합된 토픽이 발견된 경우, 하나의 토픽에 다수의 하위주제가 연결될 수 있다.

2) 오세훈. 2005. 우리나라 문헌정보학 학술지 논문 및 인용문헌 분석을 통한 연구동향 연구. 정보관리학회지, 22(3), 382에 수록한 〈표 1〉 문헌정보학 문헌의 주제 분석 도구를 참조함.

〈표 2〉 문헌정보학 주제분류표

영역	하위주제	세부주제
문헌정보학 기초		
	도서관 역사	
	도서관과 사회	
	법령/기준	
		도서관 기준
		도서관법/저작권법
	연구	
		비교 문헌정보학
		연구동향
		연구일반
		연구방법론
	이론 및 철학	
		문헌정보학 이론
		철학/사상
		학문의 지적구조
	전문성	
		교육
		사서직
		윤리/검열/지적자유
		전문단체
		전문성 일반
도서관 건물 및 설비		
도서관 체제		
	도서관 및 정보정책	
	도서관 사정/실태조사	
		도서관 사정 일반
		도서관 통계/연감
		실태 조사
	도서관 협동/자원공유	
도서관 경영		
	경영관리	
		경영 기법/전략
		도서관 경영/지식경영
	도서관 기획/활성화	
	도서관 평가	
	인사관리	
	자료의 유형	
		전자저널
		정부간행물
		학위논문
		연속간행물
		웹 자원
		자료유형 일반
	장서개발/관리	
		보존/제본/수리
		서고관리
		수서/등록/교환/납본
		자료의 유형별 관리
		장서개발/정책
		장서관리 일반
		장서점검
		장서평가
		폐기
도서관 경영(계속)		
	재정관리	
	조직관리	
		의사결정
		조직 일반
		조직 커뮤니케이션
		직무만족
	홍보/마케팅	
도서관 서비스		
	정보서비스 일반	
	도서관 유형별 서비스	
		공공/국가도서관
		대학도서관
		학교도서관
		전문도서관
		특수도서관
	독서교육/지도	
	이용자 교육/정보리터러시	
	서비스 평가	
	열람/대출 봉사	
자료조직		
	분류	
		고서분류
		도서기호
		분류법
		분류일반
		인터넷자원 분류
	서지통정	
	주제분석	
		색인/초록
		시소러스
		주제명표목표
		주제분석 일반
	메타데이터	
	편목	
		MARC
		고서편목
		목록규칙
		온라인열람용 목록(OPAC)
		웹 자원조직
		전거통제
		편목일반
정보학		
	계량정보학	
		계량정보학 일반
		인용분석/인용색인
		학술커뮤니케이션
	자동화	
		자동화 일반
		도서관업무별 자동화
		자동문헌처리/분류
		자동색인/초록
		자동화효율성 평가
정보학(계속)		
	디지털도서관	
	이용연구	
		이용연구 일반
		이용자 인식/요구
		정보이용행태
	인터넷	
		웹사이트 비교/평가
		웹사이트 설계/구축
		인터넷 일반
		콘텐츠 개발/관리
	전문가시스템	
		전문가시스템 일반
		전문가시스템 평가
		지식관리시스템
		참고전문가시스템
	정보검색	
		검색 기법/전략
		검색시스템
		검색어
		검색엔진
		시스템/검색 효율성 평가
		정보검색 일반
	정보기술	
		소프트웨어
		인공지능
		인터페이스
		정보기술일반
		정보보안
		컴퓨터 언어
		패턴/문자 인식
		하드웨어
	정보유통	
	정보이론	
	정보시스템	
		데이터 구조/설계
		데이터베이스 일반
		데이터베이스 평가
		서지데이터베이스
		원문데이터베이스
		이미지데이터베이스
		하이퍼텍스트
	컴퓨터네트워크	
서지학		
	서지학 일반	
	고서지	
	체계서지학(목록학)	
	형태서지학(판본학)	
출판		
	전자출판	
	도서/출판	
기록관리학		

3. 실험 결과

앞서 설명한 텍스트 마이닝 기법의 토픽모델링과 문헌정보학 주제분류표를 활용하여 다음 두 가지의 실험을 수행하였다. 첫째, 총 논문 초록을 대상으로 토픽모델링 실험을 수행하였고 토픽모델링 결과를 해석하기 위하여 연구주제들을 문헌정보학 주제분류표와 비교·분석한 후 연구주제마다 연도별 추이 그래프를 그렸다. 둘째, 학술지별 논문 초록을 대상으로 연구주제들을 문헌정보학 주제분류표와 비교·분석한 후 동시출현빈도를 활용하여 분석하였다.

3.1 총 논문 초록

3.1.1 토픽모델링 결과 분석

총 논문 초록을 대상으로 토픽모델링을 수행한 결과, 〈표 3〉과 같이 20건의 연구주제를 도

〈표 3〉 토픽모델링 결과: 총 논문 초록

토픽0	토픽1	토픽2	토픽3	토픽4
information	books	social	model	library
service	dynasty	communication	quality	public
technology	king	cultural	user	service
user	tripitaka	development	performance	user
resources	published	theory	suggested	community
environment	historical	process	developed	local
development	edition	society	develop	facilities
resource	made	culture	process	cultural
society	history	life	factor	people
토픽5	토픽6	토픽7	토픽8	토픽9
library	knowledge	system	digital	school
university	management	information	library	reading
collection	structure	record	electronic	education
academic	system	retrieval	copyright	library
materials	organization	user	access	children
user	business	text	resources	program
service	marketing	time	law	learning
collections	environment	developed	legal	literacy
number	strategy	full	open	instruction
토픽10	토픽11	토픽12	토픽13	토픽14
document	metadata	korea	library	classification
retrieval	element	national	librarian	subject
term	ontology	management	education	scheme
indexing	model	korean	science	ddc
performance	web	policy	subject	kdc
subject	semantic	record	professional	system
index	thesaurus	system	korea	edition
query	xml	development	organization	korean
clustering	core	government	educational	class
토픽15	토픽16	토픽17	토픽18	토픽19
cataloging	journal	user	factor	book
catalog	science	web	satisfaction	books
bibliographic	citation	search	group	korean
rules	korean	service	students	number
record	articles	internet	significant	publications
authority	medical	information	level	japanese
description	field	reference	variables	period
control	published	site	groups	materials
access	korea	searching	differences	literature

출할 수 있었다. 발견된 연구주제를 문헌정보학 주제분류표와 비교・분석한 결과를 종합하면 다음과 같다.

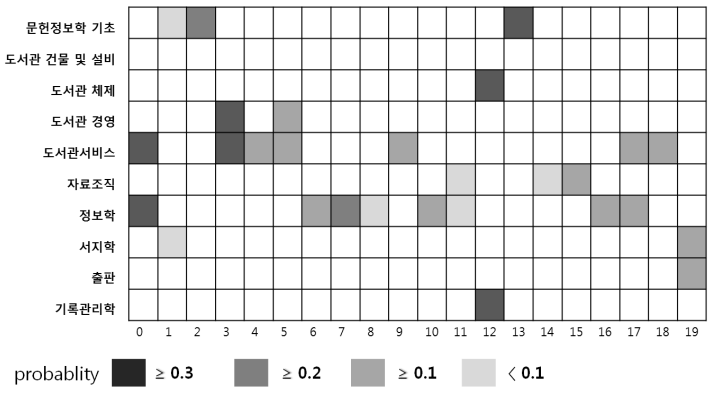
총 논문 초록에서 발견된 연구주제는 '문헌정보학 기초' 영역의 도서관과 사회, 전문성, '도서관 체제' 영역의 도서관 및 정보정책, '도서관 경영' 영역의 도서관 평가, 장서개발/관리, '도서관 서비스' 영역의 정보서비스, 도서관 유형별 서비스, 이용자 교육/정보리터러시, 서비스 평가, '자료조직' 영역의 분류, 편목, 메타데이터, '정보학' 영역의 계량정보학, 자동화, 디지털도서관, 이용연구, 인터넷, 전문가시스템, 정보검색, 정보시스템, '서지학' 영역의 고서지, '출판' 영역의 도서/출판, '기록관리학' 영역의 하위주제 등과 연결할 수 있었다.

〈그림 6〉은 토픽모델링에서 발견된 각각의 연구주제와 문헌정보학 주제분류표를 비교한 것이다. 가장 많은 연구주제가 발견된 학문영역은 정보학(28.6%), 도서관 서비스(25.0%), 문헌정보학 기초(10.7%), 자료조직(10.7%), 도서관 경영(7.1%), 서지학(7.1%), 도서관 체제(3.6%), 출판(3.6%), 기록관리학(3.6%) 순으로 나타났다. 도서관 건물 및 설비 영역에 해당되는 연구주제는 한 건도 발견되지 않았다. 이에 따라 문헌정보학의 주요 학술지논문 초록에서 정보학과 도서관 서비스에 관한 연구주제가 많이 발견되었음을 알 수 있다.

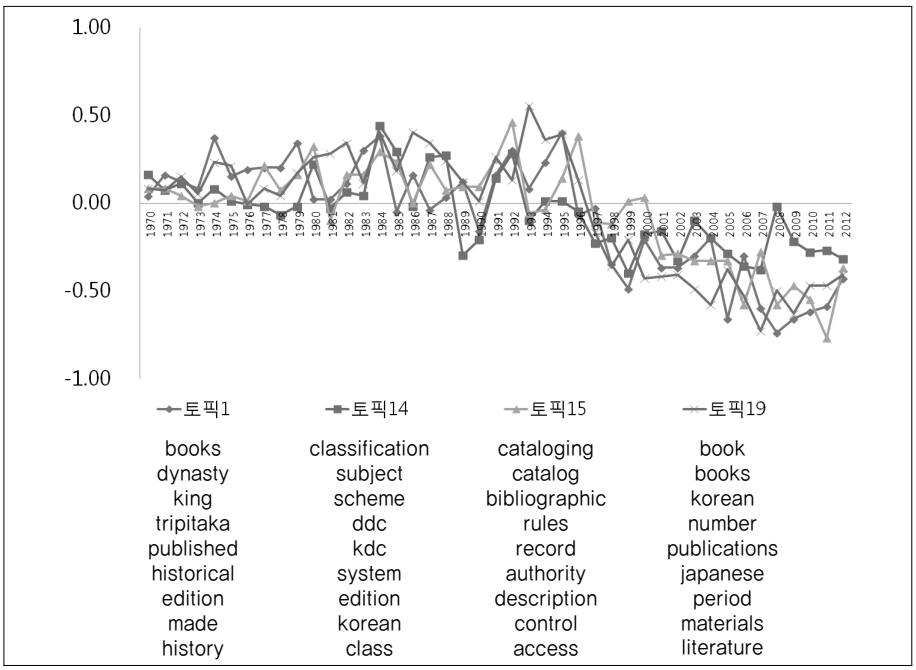
또한 단어들의 동시출현 확률(probability)의 경우, 토픽0, 토픽3, 토픽12, 토픽13이 0.3 이상의 확률로 가장 높았으며, 토픽2, 토픽7은 0.2 이상의 확률로 그 다음으로 높았다. 토픽4, 토픽5, 토픽6, 토픽9, 토픽10, 토픽15, 토픽16, 토픽17, 토픽18, 토픽19가 모두 0.1 이상의 확률을 보였으며, 토픽1, 토픽8, 토픽 11, 토픽14는 0.1 이하의 동시출현 확률을 가졌다.

3.1.2 연도별 분석

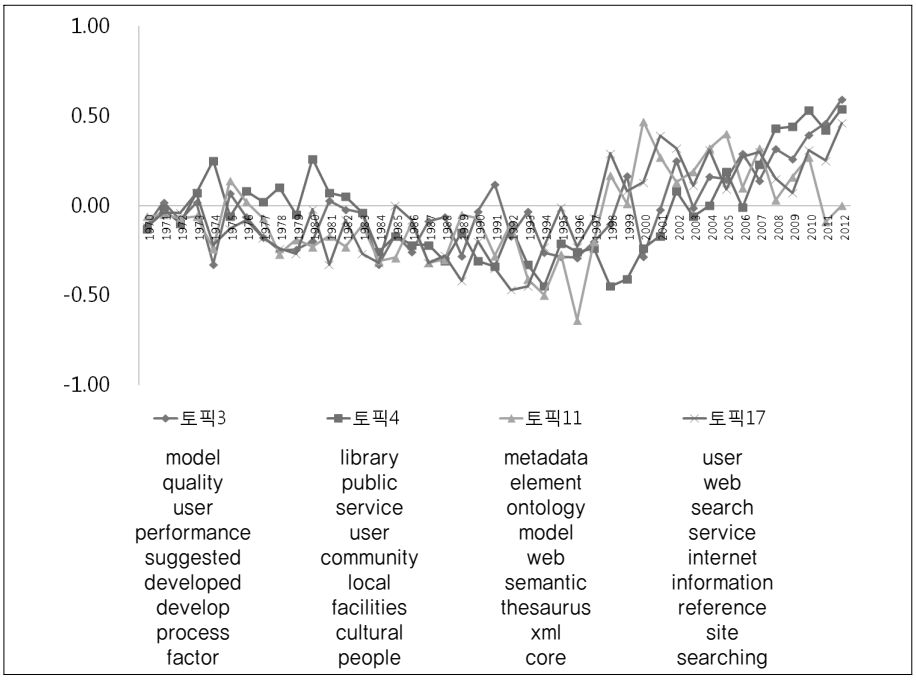
연구주제의 연도별 추이를 분석한 결과는 다음과 같다. 하강세를 보인 상위 4건의 연구주제(cold topics)는 토픽1, 토픽14, 토픽15, 토픽19이며, 연도별 추이는 〈그림 7〉과 같다. 상승세를 보인 상위 4건의 연구주제(hot topics)는 토픽3, 토픽4, 토픽11, 토픽17이며, 그 연도별 추이는 〈그림 8〉과 같다.



〈그림 6〉 토픽모델링 결과와 주제분류표 비교: 총 논문 초록



〈그림 7〉 문헌정보학의 cold topics



〈그림 8〉 문헌정보학의 hot topics

'정보학' 영역의 인터넷과 메타데이터에 관한 연구주제는 1990년대 중・후반부터 활발히 연구되었다가 2000년대 중반부터 약한 감소세를 보였다. '도서관 서비스' 영역의 도서관 유형별 서비스에 관한 연구주제와 서비스 평가 연구주제는 2000년대에 들어 활발하게 연구되고 있다. 반면 도서와 관련된 연구주제는 80, 90년대에 비해 2000년대에 적게 발견되는 것을 확인할 수 있었다. 또한 자료조직의 경우, 분류, 목록의 연구주제에서 점차 메타데이터, 시소러스 등 정보학과 접목한 연구주제로 확장되고 있는 것으로 보인다. 이에 따라 전통적인 분류학, 목록학을 대표하는 'DDC', 'KDC', 'cataloging' 등의 단어 출현빈도가 점점 낮아지고 있다. 서지학의 고서지에 해당하는 연구주제는 한국문헌정보학회지, 한국도서관・정보학회지, 한국비블리아학회지에서 그 비중이 낮아지고 있는 것으로 보인다.

3.2 학술지논문 초록

3.2.1 토픽모델링 결과 분석

정보관리학회지 논문 초록을 대상으로 토픽모델링을 수행한 결과, 〈표 4〉와 같이 20건의 연구주제를 도출할 수 있었다. 연구주제를 문헌정보학 주제분류표와 비교・분석한 결과를 종합하면 다음과 같다.

'도서관 체제' 영역의 도서관 및 정보정책, '도서관 경영' 영역의 도서관 평가, 장서개발/관리, '도서관 서비스' 영역의 정보서비스, 도서관 유형별 서비스, 이용자 교육/정보리터러시, 서비스 평가, '자료조직' 영역의 분류, 편목, 메타데이터, 주제분석, '정보학' 영역의 계량정보학, 자동화, 디지털도서관, 이용연구, 인터넷, 정보검색, 정보기술, 정보유통, 정보시스템, '기록관리학' 영역의 하위주제 등과 연결할 수 있었다.

〈그림 9〉는 토픽모델링에서 발견된 각각의 연구주제와 문헌정보학 주제분류표를 비교한 것이다. 가장 많은 연구주제가 발견된 학문영역은 정보학(51.7%), 도서관 서비스(20.7%), 자료조직(17.2%), 도서관 경영(6.9%), 기록관리학(3.4%) 순으로 나타났다. 문헌정보학 기초, 서지학, 도서관 체제, 출판에 대한 연구주제는 발견되지 않았다. 총 논문 초록의 결과와 비교하였을 때, 정보관리학회지는 정보학에 관한 연구주제가 활발히 연구되고 있으며, 도서관 서비스, 도서관 경영, 도서관 체제, 문헌정보학 기초 등의 연구주제에 대한 비중이 상대적으로 낮은 것을 알 수 있다.

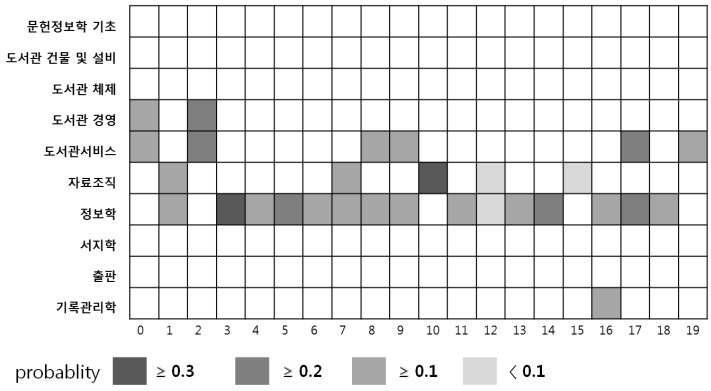
한국문헌정보학회지 논문 초록을 대상으로 토픽모델링을 수행한 결과, 〈표 5〉와 같이 20건의 연구주제를 도출할 수 있었다. 연구주제를 문헌정보학 주제분류표와 비교・분석한 결과를 종합하면 다음과 같다.

'문헌정보학 기초' 영역의 법령/기준, 전문성, '도서관 체제' 영역의 도서관 및 정보정책, '도서관 경영' 영역의 도서관 평가, 인사관리, 자료의 유형, 장서개발/관리, 조직관리, '도서관 서비스' 영역의 정보서비스, 도서관 유형별 서비스, 독서교육/지도, 이용자 교육/정보리터러시, 서비스 평가, '자료조직' 영역의 분류, 편목, 메타데이터, '정보학' 영역의 계량정보학, 자동화, 디지털도서관, 이용연구, 인터넷, 정보검색, 정보기술, 정보시스템, '서지학' 영역의 고서지, '출판' 영역의 도서/출판, '기록관리학' 영역의 하위주제 등과 연결할 수 있었다.

〈그림 10〉은 토픽모델링에서 발견된 각각의

〈표 4〉 토픽모델링 결과: 정보관리학회지

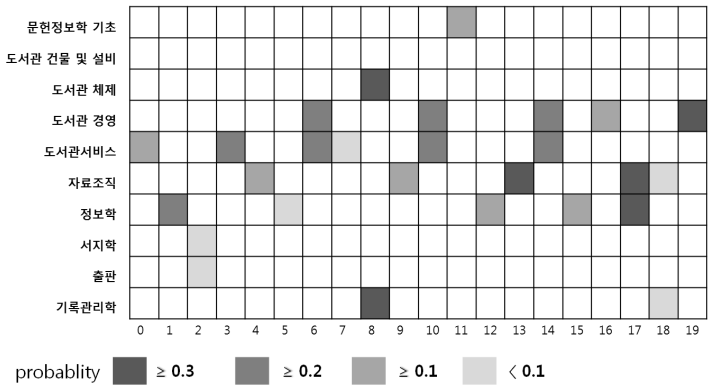
토픽0	토픽1	토픽2	토픽3	토픽4
quality	term	library	information	digital
factor	indexing	public	system	electronic
satisfaction	news	university	model	contents
management	sentences	librarian	management	resources
organizational	word	academic	record	element
service	number	book	order	materials
sharing	automatic	collection	process	content
performance	text	development	retrieval	structure
relationship	index	service	effective	resource
토픽5	토픽6	토픽7	토픽8	토픽9
subject	internet	metadata	user	service
structure	web	ontology	communication	information
author	user	semantic	behavior	user
network	meta	web	social	library
science	searching	resources	online	internet
areas	element	knowledge	important	reference
term	image	application	information	digital
field	search	tool	behaviors	technology
intellectual	query	concept	seeking	center
토픽10	토픽11	토픽12	토픽13	토픽14
cataloging	journal	thesaurus	information	document
system	science	term	effectiveness	retrieval
base	citation	relationships	indicators	performance
bibliographic	korean	phase	retrieval	term
retrieval	articles	topic	economic	set
xml	medical	descriptors	content	clustering
format	index	relation	comparison	relevance
developed	korea	concept	filtering	algorithm
frbr	cited	ontology	model	test
토픽15	토픽16	토픽17	토픽18	토픽19
classification	record	user	information	student
copyright	national	interface	patent	level
works	access	site	technical	literacy
law	korea	website	scientific	learning
ebook	management	criteria	organizations	librarian
ddc	open	survey	korean	group
scheme	meta	factor	analyzed	difference
rights	preservation	design	fields	education
materials	archives	usability	citation	instruction



〈그림 9〉 토픽모델링 결과와 주제분류표 비교: 정보관리학회지

〈표 5〉 토픽모델링 결과: 한국문헌정보학회지

토픽0	토픽1	토픽2	토픽3	토픽4
library	web	books	students	classification
school	internet	dynasty	group	system
education	content	chinese	variables	korean
librarian	site	book	significant	subject
science	approach	titles	differences	scheme
learning	paper	published	factor	class
curriculum	structure	characters	high	number
educational	techniques	history	findings	ddc
teacher	traditional	type	higher	materials
토픽5	토픽6	토픽7	토픽8	토픽9
document	library	reading	korea	cataloging
search	university	children	national	rules
searching	academic	books	development	bibliographic
term	quality	book	current	authority
retrieval	collection	program	system	entry
text	user	bibliotherapy	policy	access
performance	indicators	school	korean	reference
query	performance	programs	suggested	works
video	satisfaction	grade	major	industrial
토픽10	토픽11	토픽12	토픽13	토픽14
service	legal	system	subject	library
library	korea	user	author	public
user	system	search	important	law
librarian	korean	catalog	work	local
reference	materials	retrieval	present	korean
public	social	information	field	central
programs	copyright	base	time	japanese
people	cultural	design	theory	staff
provide	access	online	table	community
토픽15	토픽16	토픽17	토픽18	토픽19
journal	factor	information	bibliographic	library
articles	librarian	digital	archives	management
science	variables	knowledge	archival	system
citation	organizational	literacy	fields	organization
korean	job	metadata	edition	electronic
index	satisfaction	technology	items	environment
published	relationship	science	record	resources
field	type	element	description	model
literature	influence	resources	db	technical



〈그림 10〉 토픽모델링 결과와 주제분류표 비교: 한국문헌정보학회지

연구주제와 문헌정보학 주제분류표를 비교한 것이다. 가장 많은 연구주제가 발견된 학문영역은 도서관 서비스(22.2%)이며, 그 다음은 정보학(18.5%), 도서관 경영(18.5%), 자료조직(18.5%), 기록관리학(7.4%), 서지학(3.7%), 도서관 체제(3.7%), 문헌정보학 기초(3.7%), 출판(3.7%) 순으로 나타났다. 총 논문 초록과 비교하였을 때, 도서관 건물 및 설비 영역을 제외한 모든 영역에서 연구주제가 발견된 것이 동일하나, 도서관 서비스가 정보학보다 연구주제가 많이 발견되었다는 점에서 차이를 보인다.

한국도서관·정보학회지 논문 초록을 대상으로 토픽모델링을 수행한 결과, 〈표 6〉과 같이 20건의 연구주제를 도출할 수 있었다. 연구주제

〈표 6〉 토픽모델링 결과: 한국도서관·정보학회지

토픽0	토픽1	토픽2	토픽3	토픽4
cataloging	library	laws	information	library
bibliographic	korea	government	science	public
rules	system	law	field	service
control	korean	copyright	social	community
title	analyzed	publications	journal	local
authority	policy	movement	courses	facilities
description	major	legal	papers	culture
work	current	children	subjects	cultural
outsourcing	development	distribution	curriculum	city
토픽5	토픽6	토픽7	토픽8	토픽9
system	library	book	library	books
retrieval	university	books	staff	dynasty
term	materials	life	number	book
index	system	students	university	made
base	collection	reading	volumes	published
automatic	national	college	person	printed
indexing	subject	literature	factor	printing
language	development	review	expenditure	king
indexes	librarian	read	student	sa
토픽10	토픽11	토픽12	토픽13	토픽14
quality	library	school	search	service
factor	librarian	reading	user	user
satisfaction	academic	education	catalog	information
librarian	development	library	searching	reference
job	management	program	opac	library
students	society	librarian	web	web
level	role	teacher	document	site
performance	future	instruction	online	university
survey	environment	educational	bases	provide
토픽15	토픽16	토픽17	토픽18	토픽19
model	information	korean	classification	jing
knowledge	resources	period	kdc	classics
information	electronic	record	ddc	political
digital	internet	archives	edition	kyung
management	access	history	scheme	type
element	digital	japanese	table	gu
paper	technology	form	items	books
process	resource	time	class	busan
metadata	journal	published	schemes	civilization

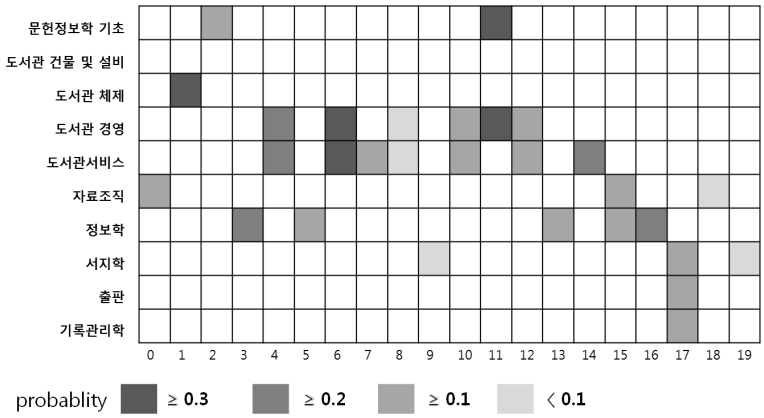
를 문헌정보학 주제분류표와 비교·분석한 결과를 종합하면 다음과 같다.

'문헌정보학 기초' 영역의 도서관과 사회, 전문성, '도서관 체제' 영역의 도서관 및 정보정책, '도서관 경영' 영역의 도서관 평가, 인사관리, 장서개발/관리, 조직관리, '도서관 서비스' 영역의 정보서비스, 도서관 유형별 서비스, 이용자 교육/정보리터러시, 서비스 평가, '자료조직' 영역의 분류, 편목, 메타데이터, '정보학' 영역의 계량정보학, 디지털도서관, 이용연구, 인터넷, 전문가시스템, 정보검색, 정보시스템, '서지학' 영역의 고서지, '출판' 영역의 도서/출판, '기록관리학' 영역의 하위주제 등과 연결할 수 있었다.

〈그림 11〉은 토픽모델링에서 발견된 각각의 연구주제와 문헌정보학 주제분류표를 비교한 것이다. 가장 많은 연구주제가 발견된 학문영역은 도서관 서비스(24.1%)이며, 그 다음은 도서관 경영(20.7%), 정보학(17.2%), 자료조직(10.3%), 서지학(10.3%), 문헌정보학 기초(6.9%), 도서관 체제(3.4%), 출판(3.4%), 기록관리학(3.4%) 순으로 나타났다. 총 논문 초록과 비교하였을 때, 도서관 건물 및 설비 영역을 제외한 모든 영역에서 연구주제가 발견된 것이 동일하나, 도서관 서비스와 도서관 경영이 정보학보다 연구주제가 많이 발견되었다는 점에서 차이를 보인다. 따라서 한국도서관·정보학회지는 도서관 서비스, 도서관 경영 등의 도서관 연구주제가 활발히 연구되고 있으며, 정보학에 대한 비중이 낮은 것을 알 수 있다.

한국비블리아학회지 논문 초록을 대상으로 토픽모델링을 수행한 결과, 〈표 7〉과 같이 20건의 연구주제를 도출할 수 있었다. 연구주제를 문헌정보학 주제분류표와 비교·분석한 결과를 종합하면 다음과 같다.

'문헌정보학 기초' 영역의 전문성, '도서관 체제' 영역의 도서관 및 정보정책, '도서관 경영' 영역의 도서관 평가, 자료의 유형, 장서개발/관리, '도서관 서비스' 영역의 도서관 유형별 서비스, 이용자 교육/정보리터러시, 서비스 평가, '자료조직' 영역의 분류, 편목, 주제분석, 메타데이터, '정보학' 영역의 계량정보학, 디지털도서관, 인터넷, 전문가시스템, 정보검색, 정보기술, 정



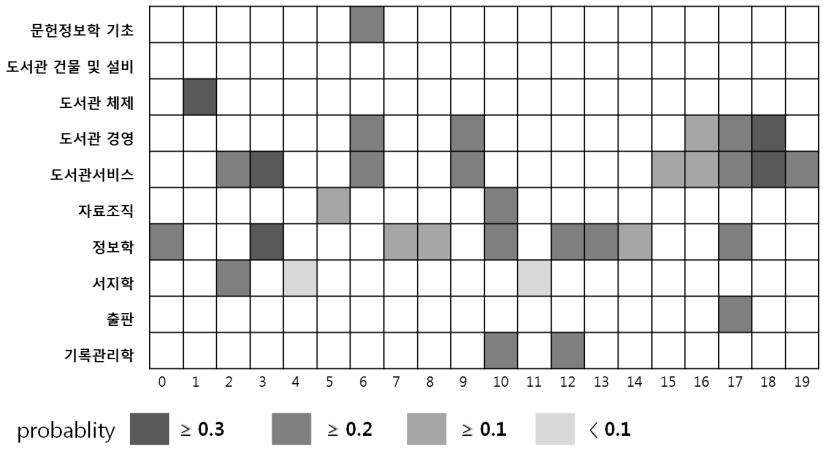
〈그림 11〉 토픽모델링 결과와 주제분류표 비교: 한국도서관·정보학회지

〈표 7〉 토픽모델링 결과: 한국비블리아학회지

토픽0	토픽1	토픽2	토픽3	토픽4
subject	library	book	information	government
number	korea	materials	knowledge	tripitaka
structure	collection	collections	paper	shelf
area	national	books	management	pages
document	development	social	service	lock
areas	policy	science	technology	buddhist
contents	resources	present	development	jang
order	plan	access	environment	dae
term	analyzed	factor	society	scriptures
토픽5	토픽6	토픽7	토픽8	토픽9
classification	library	web	digital	factor
system	librarian	user	resources	level
kdc	university	design	media	quality
scheme	survey	search	project	satisfaction
cataloging	professional	interface	copyright	survey
won	local	website	industry	academic
edition	job	information	material	conducted
korean	questionnaire	effective	protection	students
standard	reference	access	law	performance
토픽10	토픽11	토픽12	토픽13	토픽14
model	korean	record	system	journal
resources	japanese	cultural	user	science
preservation	books	management	time	field
metadata	dynasty	archives	retrieval	medical
archiving	king	electronic	search	literature
ontology	yi	element	cost	korean
suggested	hongmunkwan	statistics	model	korea
strategy	printed	current	unit	citation
proposed	chinese	paper	base	published
토픽15	토픽16	토픽17	토픽18	토픽19
school	education	communication	library	information
reading	activities	space	service	literacy
program	job	publishing	public	academic
programs	case	understand	user	korean
children	cooperation	existing	community	lis
education	college	improve	provide	education
students	ability	presented	internet	korea
learning	educational	social	space	american
instruction	confidence	order	academic	instruction

보시스템, '서지학' 영역의 고서지, '출판' 영역의 도서/출판, '기록관리학' 영역의 하위주제 등과 연결할 수 있었다.

〈그림 12〉는 토픽모델링에서 발견된 각각의 연구주제와 문헌정보학 주제분류표를 비교한 것이다. 가장 많은 연구주제가 발견된 학문영역은 도서관 서비스(27.3%)와 정보학(27.3%)이며, 그 다음은 도서관 경영(15.15%), 서지학(9.1%), 자료조직(6.1%), 기록관리학(6.1%), 문헌정보학 기초(3.0%), 도서관 체제(3.0%), 출판(3.0%) 순으로 나타났다. 총 논문 초록과 비교하였을 때, 도서관 건물 및 설비 영역을 제외한 모든 영역에서 연구주제가 발견된 것이 동일하나, 도서관 서비스와 정보학의 연구주제가



〈그림 12〉 토픽모델링 결과와 주제분류표 비교: 한국비블리아학회지

가장 많이 발견되었다는 점에서 차이를 보인다. 따라서 한국비블리아학회지는 전반적으로 도서관의 연구주제가 정보학의 연구주제보다 많이 나타났다.

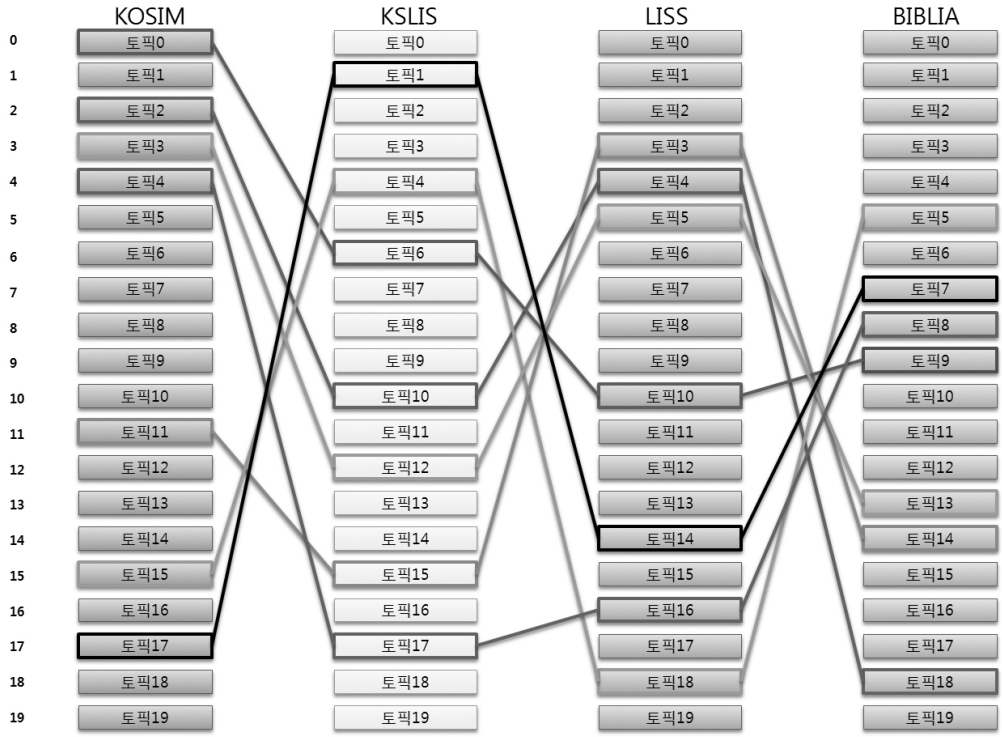
3.2.2 동시출현빈도 분석

학술지논문 초록의 실험 결과를 비교・분석하기 위하여 동시출현단어(word co-occurrence)의 출현 빈도수를 기준으로 각 연구주제를 연결하였다.

그 결과는 〈그림 13〉과 같다. '도서관 서비스' 영역의 도서관 유형별 서비스, 서비스 평가, '정보학' 영역의 정보검색과 정보시스템, 디지털도서관, 인터넷, 계량정보학, '자료조직' 영역의 분류 등에 해당하는 연구주제가 4개의 학술지에서 모두 출현한 것을 알 수 있다.

'도서관 서비스' 영역의 도서관 유형별 서비스에 관한 연구주제는 'library', 'service', 'public', 'university' 등의 단어가 정보관리학회지의 토픽2, 한국문헌정보학회지의 토픽10, 한국도서관・정보학회지의 토픽4, 한국비블리아학회지의 토픽18에서 동시에 출현하였으며, 서비스 평가와 관련하여 'quality', 'satisfaction', 'survey' 등의 단어가 정보관리학회지의 토픽0, 한국문헌정보학회지의 토픽6, 한국도서관・정보학회지의 토픽10, 한국비블리아학회지의 토픽9에서 동시에 출현하였다.

'정보학' 영역의 정보검색과 정보시스템 연구주제는 'information', 'system', 'retrieval' 또는 'system', 'retrieval' 등의 단어가 정보관리학회지의 토픽3, 한국문헌정보학회지의 토픽12, 한국도서관・정보학회지의 토픽5, 한국비블리아학회지의 토픽13에서 동시에 출현하였다. 디지털도서관에 관한 연구주제는 'digital', 'resources'와 'electronic' 등의 단어가 정보관리학회지의 토픽4, 한국문헌정보학회지의 토픽17, 한국도서관・정보학회지의 토픽16, 한국비블리아학회지의 토픽8에서 동시에 나타났으며, 인터넷에 관하여 'web', 'website', 'internet' 등의 단어가 정보관리학회지의 토픽17, 한국문헌정보학회지의



〈그림 13〉 학술지논문 초록의 토픽모델링 결과 비교·분석

토픽1, 한국도서관·정보학회지의 토픽14, 한국비블리아학회지의 토픽7에서 출현하였다. 계량정보학에 관한 연구주제는 'journal', 'science', 'citation' 등의 단어가 정보관리학회지의 토픽11, 한국문헌정보학회지의 토픽15, 한국도서관·정보학회지의 토픽3, 한국비블리아학회지의 토픽14에서 공통적으로 나타났다.

마지막으로 '자료조직' 영역의 분류에 관한 연구주제는 'classification', 'scheme', 'ddc', 'kdc' 등의 단어가 정보관리학회지의 토픽15, 한국문헌정보학회지의 토픽4, 한국도서관·정보학회지의 토픽18, 한국비블리아학회지의 토픽5에서 나타났다.

4. 결론 및 제언

본 연구는 국내 문헌정보학 분야의 연구동향을 규명하기 위해 문헌정보학 주요 학술지인, 정보관리학회지, 한국문헌정보학회지, 한국도서관·정보학회지, 한국비블리아학회지의 1970년대부터 2012년도까지 발표 논문 초록을 수집하여 LDA(Latent Dirichlet Allocation)기반의 토픽 모델링 실험을 수행하였다. 그 결과를 요약하면 다음과 같다.

첫째, 문헌정보학자들이 관심을 갖는 주요 연구주제가 무엇인지를 규명하고 가장 많은 연구주제가 발견된 학문영역은 무엇인지 밝히기 위해 토픽모델링 실험을 하였다.

실험 결과, '정보학' 영역의 디지털도서관, 이용연구, 인터넷, 전문가시스템, 계량정보학, 자동화, 정보검색, 정보시스템, '도서관 서비스' 영역의 정보서비스, 도서관 유형별 서비스, 이용자 교육/정보리터러시, 서비스 평가, '문헌정보학 기초' 영역의 도서관과 사회, 전문성, '자료조직' 영역의 분류, 편목, 메타데이터, '도서관 경영' 영역의 도서관 평가, 장서개발/관리, '서지학' 영역의 고서지, '도서관 체제' 영역의 도서관 및 정보정책, '출판' 영역의 도서/출판, '기록관리학' 영역의 하위주제 등과 연결할 수 있었다.

둘째, 문헌정보학 주요 연구주제의 연도별 추이를 분석하고 근래에 활발하게 연구되고 있는 연구주제(hot topics)와 점차 연구되지 않고 있는 연구주제(cold topics)를 밝히고자 하였다.

실험 결과, '정보학' 영역의 인터넷과 메타데이터에 관한 연구주제는 1990년대 중·후반부터 활발히 연구되었다가 2000년대 중반부터 약한 감소세를 보이고 있다. '도서관 서비스' 영역의 도서관 유형별 서비스 및 서비스 평가에 관한 연구주제는 2000년대에 높은 상승세를 보였다. 반면, 도서, 분류, 편목, 고서지 등의 연구주제는 90년대 중후반부터 하강세를 보였다.

셋째, 문헌정보학 학술지별 주요 연구주제를 밝히고 학술지마다 관심을 갖는 주요 연구주제가 어떤 공통점과 차이점을 갖는지 비교·분석하였다.

실험 결과, 정보관리학회지는 정보학의 연구주제가 집중적으로 출현하였으며, 도서관 서비스, 도서관 경영, 도서관 체제, 문헌정보학 기초 등의 연구주제에 대한 비중이 낮게 출현하였다. 반면, 한국문헌정보학회지와 한국도서관·정보학회지, 한국비블리아학회지는 도서관 서비스, 도서관 경영 등의 도서관 연구주제가 다양하게 출현하였으나, 정보학에 대한 연구주제는 정보관리학회지에 비해 그 비중이 낮은 것을 알 수 있다.

한편 동시출현단어(word co-occurrence)의 출현 빈도수를 기준으로 공통 연구주제를 도출한 결과, '도서관 서비스' 영역의 도서관 유형별 서비스, 서비스 평가, '정보학' 영역의 정보검색과 정보시스템, 디지털도서관, 인터넷, '자료조직' 영역의 분류 등에 해당하는 연구주제가 4개의 학술지에서 모두 출현한 것을 알 수 있다.

본 연구는 텍스트 마이닝 기법을 적용하여 연구동향을 분석한 결과가 기존의 연구자가 직접 내용을 판단하여 분석한 결과와 유사하게 도출되었다는 것에 그 의의가 있다. 또한 연구동향 분석뿐만 아니라, 계량정보학의 내용분석에 대하여 새로운 방법론을 제시하였다는 점에서 의미가 있다. 그러나 한국기록관리학회지, 기록학연구, 서지학 연구를 대상에서 제외함으로써 서지학과 기록관리학의 연구동향을 살필 수 없었다는 한계점을 갖는다. 향후 본 연구에서 활용한 텍스트 마이닝의 기법이 다른 학문영역의 연구동향을 분석하는 데 활용될 수 있을 것이라 기대한다. 또한 토픽모델링을 활용하여 연구주제의 결합과 분리에 대한 현상을 면밀하게 고찰할 수 있을 것이다.

참 고 문 헌

김판준, 이재윤 (2007). 연구 영역 분석을 위한 디스크립터 프로파일링에 관한 연구. 정보관리학회지, 24(4), 285-303. http://dx.doi.org/10.3743/KOSIM.2007.24.4.285

서은경 (1997). 정보학분야 연구동향 분석: 정보관리학회지와 JASIS의 비교분석을 중심으로. 정보관리학회지, 14(1), 269-291.

서은경 (2010). 정보관리학회지 연구의 동향분석. 정보관리학회지, 27(4), 7-31. http://dx.doi.org/10.3743/KOSIM.2010.27.4.007

손정표 (2003). 한국의 문헌정보학 분야 연구동향 분석 1957-2002. 한국도서관·정보학회지, 34(3), 9-21.

오세훈 (2005). 우리나라 문헌정보학 학술지 논문 및 인용문헌 분석을 통한 연구동향 연구. 정보관리학회지, 22(3), 379-408. http://dx.doi.org/10.3743/KOSIM.2005.22.3.379

오세훈, 이두영 (2005). 우리나라의 정보학 연구동향에 관한 연구. 정보관리학회지, 22(1), 167-189. http://dx.doi.org/10.3743/KOSIM.2005.22.1.167

이재윤, 문주영, 김희정 (2007). 텍스트마이닝을 이용한 국내 기록관리학 분야 지적구조 분석. 한국문헌정보학회지, 41(1), 345-372. http://dx.doi.org/10.4275/KSLIS.2007.41.1.345

정재영, 박진희 (2011). 한국 문헌정보학의 현장연구 현황 분석. 한국도서관·정보학회지, 42(2), 171-191.

정진식 (2001). 한국 문헌정보학 분야의 연구동향 분석, 1996-2000. 한국문헌정보학회지, 35(3), 55-78.

조재인 (2011). 네트워크 텍스트 분석을 통한 문헌정보학 최근 연구 경향 분석. 정보관리학회지, 28(4), 65-83. http://dx.doi.org/10.3743/KOSIM.2011.28.4.065

한상완, 조인숙 (1996). 문헌정보학분야 학회지의 논문분석. 도서관, 51(1), 114-139.

Blei, D. (2012). Probabilistic topic models. Communications of the ACM, 55(4), 77-84. http://dx.doi.org/10.1145/2133806.2133826

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics, PNAS, 1(101), 5228-5235. http://dx.doi.org/10.1073/pnas.0307752101

Kao, A. & Poteet, S. R. (Eds.) (2007). Natural language processing and text mining. London: Springer-Verlag.

Kim, Heejung, & Lee, Jae Yun (2009). Archiving research trends in LIS domain using profiling analysis. Scientometrics, 80(1), 75-90. http://dx.doi.org/10.1007/s11192-007-1998-z

Mimno, D., & McCallum, A. (2008). Topic models conditioned on arbitrary features with Dirichlet-multinomial regression. Proceedings of the 24th Conference in Uncertainty in Artificial Intelligence (UAI 2008), 411-418.

Song, Min, & Kim, Su Yeon (in press). Detecting the knowledge structure of bioinformatics by mining full-text collections. Scientometrics, in press. http://dx.doi.org/10.1007/s11192-012-0900-9

Wang, X., & McCallum, A. (2006). Topics over time: A non-Markov continuous-time model of topical trends. Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '06), 424-433.

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Cho, Jane (2011). A study for research area of library and information science by network text analysis. Journal of the Korean Society for Information Management, 28(4), 65-83. http://dx.doi.org/10.3743/KOSIM.2011.28.4.065

Chung, Jae-Young, & Park, Jin-Hee (2011). Analysis of the trends in the field studies of library and information science in Korea. Journal of Korean Library and Information Science Society, 42(2), 171-191.

Chung, Jin-Sik (2001). An analytical study on research trends of library and information science in Korea: 1996-2000. Journal of the Korean Society for Library and Information Science, 35(3), 55-78.

Kim, Panjun, & Lee, Jae Yun (2007). Descriptor profiling for research domain analysis. Journal of the Korean Society for Information Management, 24(4), 285-303. http://dx.doi.org/10.3743/KOSIM.2007.24.4.285

Lee, Jae Yun, Moon, Ju-Young, & Kim, Hee-Jung (2007). Examining the intellectual structure of records management & archival science in Korea with text mining. Journal of the Korean Society for Library and Information Science, 41(1), 345-372. http://dx.doi.org/10.4275/KSLIS.2007.41.1.345

Oh, Se-Hoon (2005). A study on the research trends of library & information science in Korea by analyzing journal articles and the cited literatures. Journal of the Korean Society for Information Management, 22(3), 379-408. http://dx.doi.org/10.3743/KOSIM.2005.22.3.379

Oh, Se-Hoon, & Lee, Too-Young (2005). Research trends of information science in Korea. Journal of the Korean Society for Information Management, 22(1), 167-189.

Seo, Eun-Gyoung (1997). An analytical study on research patterns in information science. Journal of the Korean Society for Information Management, 14(1), 269-291.

Seo, Eun-Gyoung (2010). Trends analysis on research articles in the Journal of Korean Society for Information Management. Journal of the Korean Society for Information Management, 27(4), 7-31. http://dx.doi.org/10.3743/KOSIM.2010.27.4.007

Sohn, Jung-Pyo (2003). An analytical study on research trends of library and information science in Korea: 1957~2002. Journal of Korean Library and Information Science Society, 34(3), 9-21.