토픽 모델링을 활용한 다문화 연구의 이슈 추적 연구*

A Study on Issue Tracking on Multi-cultural Studies Using Topic Modeling

박 종 도 (Jong Do Park)**

목 차

초 록

본 논문은 국내 다문화 관련 분야의 연구동향을 규명하기 위하여 다문화와 관련한 국내 학술 문헌을 수집하여 LDA (Latent Dirichlet Allocation) 기반의 토픽 모델링을 통해 토픽을 분석하였다. 이를 통해 국내 다문화 관련 연구에서의 중심 연구 토픽을 시기별로 추적하여 그 변화의 양상을 관찰하였고, 그 결과 핫 토픽으로는 '다문화 사회통합'과 '학교 다문화 교육'이 관찰되었으며 콜드 토픽으로는 '문화정체성과 민족주의' 관련 토픽이 관찰되었다.

ABSTRACT

The goal of this study is to analyze topics discussed in academic papers on multiculture in Korea to figure out research trends in the field. In order to do topic analysis, LDA (Latent Dirichlet Allocation)-based topic modeling methods are employed. Through the analysis, it is possible to track topic changes in the field and it is found that topics related to 'social integration' and 'multicultural education in schools' are hot topics, and topics related to 'cultural identity and nationalism' are cold topics among top five topics in the field.

키워드: 다문화, 이슈 추적, 토픽 모델링
Multiculture, Issue Track, Topic Modeling

---

* 이 논문은 2018년도 인천대학교 교내연구비의 지원을 받아 연구되었음.
** 인천대학교 문헌정보학과 조교수; 인천대학교 사회과학연구원 연구위원
(jdp23@inu.ac.kr / ISNI 0000 0004 7358 748X)
논문접수일자: 2019년 7월 16일 최초심사일자: 2019년 8월 17일 게재확정일자: 2019년 8월 20일
한국문헌정보학회지, 53(3): 273-289, 2019. [http://dx.doi.org/10.4275/KSLIS.2019.53.3.273]

1. 서 론

1.1 연구배경

한국사회는 2000년대 들어서면서 이주노동자와 결혼이민자가 증가함에 따라 다문화사회로 변모하고 있다. 다문화사회란 민족이나 인종, 문화적으로 다원화되어 있는 사회로 한 국가나 사회 속에 여러 다른 생활양식이 존재한다는 것을 의미한다(국가기록원 2019). 국제사회에서 단일민족 단일국가 체제의 순수한 이념적 유형을 대표하는 나라 중 하나인 한국이 다인종, 다민족, 다문화 사회로 진입하고 있다. 행정안전부의 보도자료 따르면 2017년 11월 1일 인구주택총조사 기준으로 우리나라에 거주하는 장기체류 외국인・귀화자・외국인주민 자녀는 모두 186만 1,084명인 것으로 조사되어 우리나라 총인구(51,422,507명) 대비 3.6%를 차지하고 있다. 이렇게 한국사회가 다문화사회가 되어 가면서 다문화에 대한 사회적, 정책적, 학문적 관심이 증가하기 시작했다. 그 결과 다문화와 관련한 다양한 학회가 설립되고 다양한 학술지를 통해 많은 학자들의 학술활동이 활발하게 진행되어 왔다.

한편, 국내 다문화 관련 연구 활동이 활발해짐에 따라 연구자들은 다문화관련 연구가 어떤 주제를 중심으로 발전하고 있으며 어떻게 변화하고 있으며 학문적 유행은 어떤 양상을 보이는지 등 그 현상을 조사해 볼 필요가 있다. 이에 본 연구에서는 다문화와 관련한 국내 연구자들의 연구주제를 분석하여 보고 국내 연구자들의 연구 관심의 흐름을 추적해보고자 한다.

1.2 연구목표와 기대효과

본 연구에서는 다문화 관련 국내 주요 학술문헌에서 다루어지는 주요 토픽을 분석하여 다문화 관련 연구에서 나타나는 주요 이슈들이 시간의 흐름에 따라 어떻게 변하고 있는지를 살펴보고 이를 통해 다문화 관련 국내의 연구 동향을 살펴보고자 한다. 특히 다문화 관련 국내 연구 동향의 시간 흐름에 따른 토픽의 변화를 살펴보기 위해 토픽 모델링 기법을 사용하고자 한다.

토픽 모델링 기법은 텍스트 마이닝 기법을 활용한 통계 추론 모델로서 이 방법을 활용하면 연구자의 주관에 영향을 받을 수 있는 연구주제에 대한 평가를 문헌에 나타나는 텍스트를 통계적으로 분석하여 더욱 객관적이고 명확하게 연구 주제를 분석하여 낼 수 있는 장점이 있다. 이를 통해 다문화 관련 국내 학술 연구에서의 연구주제의 변화와 그 양상을 종합적으로 분석하고자 한다.

2. 이론적 배경 및 선행연구

2.1 Latent Dirichlet Allocation (LDA) 기반 토픽 모델링

토픽 모델링은 주로 비구조화된 텍스트 문서에서 새로운 정보를 추출하기 위한 연구에 많이 활용되고 있으며, 특정 분야의 주제 및 동향을 파악하는데 효과적인 수단으로 제안되고 있다(이금실, 이인주, 이영경 2018). 토픽 모델링은 전체 문서 집합에서 나타나는 의미론적 구조를

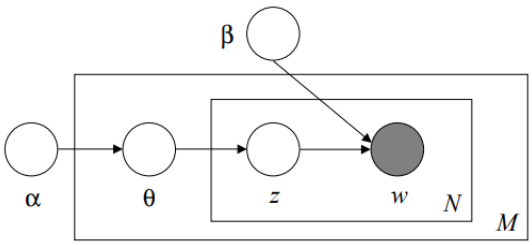
탐색하기 위해 사용되는 텍스트 마이닝 기법으로 문서 집합 내에 숨겨져 있는 토픽을 다양한 통계적 방법을 활용하여 표현하는 것을 말한다. 이러한 토픽 모델링은 어떤 문서의 토픽이 여러 개의 추상적인 토픽으로 구성되어 있으며, 각각의 토픽은 여러 단어의 집합으로 구성되어 있다는 것을 전제하고 있다.

LDA는 문헌 내의 잠재된 토픽을 추정하는 것으로 어떤 단어들을 어떠한 토픽에서 선택하여 배치할 것인지 각각의 파라미터(parameter: α, β)로 모델링하는 기법이다. 문헌, 단어 등 관찰된 변수(w)를 통해 문헌의 구조 같은 잠재 변수를 추론하는 것을 목적으로 하며, 궁극적으로 전체 문서 집합의 주제들과 문서별 주제 비율 및 각 주제에 포함될 단어들의 분포를 도출할 수 있다.

LDA는 우선 디리클레 확률 분포(Dirichlet probability distribution)를 활용하여 문서의 잠재적인 확률을 설정한 다음 깁스 샘플링(Gibbs sampling) 알고리듬을 활용하여 주어진 문서의 토픽과 단어에 대한 확률 값을 추정한다. 이러한 LDA의 장점은 어떤 토픽과 단어에 대한 사전 정보가 없는 비지도 생성 모델(unsupervised generative model)이기 때문에 각 토픽과 관련한 모든 단어를 쉽게 찾을 수 있다는 점이다(Blei et al. 2003). 반면, 데이터의 양이 적거나 정규분포의 환경에는 적합하지 않다(Liu, Zhang, Chang and Sun 2011; Wang and Blei 2013).

〈그림 1〉에서 박스로 표현된 M과 N은 판(plate)으로서 반복되는 것을 의미하는데, M은 개별 문서를 나타내고 N은 해당 문서에 대한 토픽과 단어를 반복해서 선택하는 것을 의미한다. α는 각 문서별 토픽의 분포에 대한 디리클레 분포의 설정값을 의미하고, β는 각 토픽의 단어 분포에 대한 디리클레 분포의 설정값을 의미한다. θ는 개별 문서에 대한 토픽 분포를 의미하고, z는 특정 문서 내의 단어에 대한 토픽을 의미하며 w는 특정 단어를 의미한다. 〈그림 1〉에서 w만 짙은 색으로 표현되었는데, 이는 w만 관찰 가능하고 나머지 것들은 잠재되어 있어 관찰이 불가함을 나타내고 있다.

여기에서 α는 문서와 토픽 간의 밀도를 나타내는데, α의 값이 커질수록 문서의 집합이 더 많은 수의 토픽으로 구성되어 있음을 의미하고, α의 값이 낮아질수록 문서의 집합은 더 적은 수의 토픽으로 구성되어 있음을 의미한다. β는 토픽과 단어 사이의 밀도를 의미하는데, 이 값이 커질수록 토픽이 더 많은 단어로 구성되어 있음



〈그림 1〉 Graphical model representation of LDA from "Latent Dirichlet Allocation," by Blei, D. Ng, A. and Jordan, M. 2003. *Journal of Machine Learning Research*, 3: 997.

을 의미하고 이 값이 낮을수록 더 적은 수의 단어로 토픽이 구성되어 있음을 의미한다. 따라서, LDA를 활용하는 경우 주어진 문서 집합에 대해 토픽의 수나 토픽에 포함되는 단어의 수를 사용자가 임의로 결정할 수 있다는 장점이 있다. 하지만 LDA는 토픽 내 단어들 간의 관계를 알기 어렵다는 단점이 있다(안성주, 양정진 2018).

LDA 기법을 활용하여 토픽 모델링을 하는 경우 토픽 모델링의 결과의 품질은 설정된 α와 β의 값에 많은 영향을 받는다. 따라서, LDA 기법을 활용한 토픽 모델링의 경우 최적의 α와 β의 값을 찾아 내는 것이 매우 중요하다.

본 연구에서는 다문화 관련 국내 학술잡지에 실린 논문을 대상으로 LDA 분석의 장점을 최대한 활용하기 위해 논문의 제목, 저자 키워드, 초록 모두를 대상으로 최적의 토픽 수를 찾고 이를 통해 다문화 관련 연구의 토픽을 분석하고 연구 동향을 분석하고자 한다.

2.2 선행연구

2.2.1 다문화 관련 연구 동향 분석

다문화와 관련한 연구는 일반적인 학문과는 달리 다양한 학문에 두루 걸치는 학제적 성격이 강하다. 다문화와 관련한 주제는 교육학, 언어학, 문화인류학, 정치학 등 다양한 학문 분야에서 다루어지고 있다. 이러한 특성으로 인해 다문화와 관련한 연구 동향 분석은 주로 다문화라는 거시적인 개념을 대상으로 학문 전반에 걸친 연구 동향을 분석하기보다 특정 학문 분야 내에서 미시적 개념으로서 다문화와 관련한 연구 동향 분석하는 연구들이 주로 진행되었다. 상대적으로 다문화와 관련한 연구 동향을 거시적으로 분석한 연구는 적은 실정이다.

다문화와 관련하여 특정 세부분야가 아닌 일반 분야의 연구를 대상으로 다문화와 관련한 연구의 동향을 밝히고자 한 연구는 김세현(2018)과 장임숙, 장덕현, 이수상(2011)의 연구가 있다.

김세현(2018)은 국내에서 진행된 다문화 연구를 대상으로 주요 연구 주제를 추출하여 다문화 영역의 지적구조의 특성과 변화를 파악하기 위해 한양대학교 SSK 다문화연구센터에서 제공하는 다문화 아카이브(CSMR Archive)에 수록된 다문화 관련 논문정보 중 국문 초록 정보를 활용하여 텍스트마이닝 기법을 활용하여 텍스트를 정련한 후 Latent Dirichlet Allocation (LDA)분석과 텍스트 연결망 분석을 통해 토픽을 분석하였다.

장임숙, 장덕현, 이수상(2011)은 한국 다문화 지식체계의 구조를 분석하기 위하여 2005년부터 2010년까지 발행된 한국연구재단 등재 및 등재 후보 학술지를 대상으로 저자가 부여한 키워드를 활용한 키워드 동시출현을 통한 연결망을 분석하여 다문화 연구가 활성화된 학문분야의 분야별 핵심 주제와 다문화 지식구조의 특성을 분석하였다. 이상의 두 연구는 다문화 관련 연구의 연구 주제와 연구 동향을 분석하기 위하여 각각 초록 또는 저자 키워드를 대상으로 분석을 실시하였다.

한편 다문화와 관련한 연구 동향 분석 연구에 있어서 특정 학문 영역에 한정하여 연구 동향을 분석하는 연구가 많이 진행되고 있다. 다문화의 주제는 교육학, 언어학, 지역학 등에서부터 건축학, 의학에 이르기까지 매우 광범위한 분야에서 다루어지고 있어서 모든 학문 분야를 포괄하여 연구 동향을 분석하기보다는 특

정한 학문 영역에 한정하여 연구 동향을 분석하는 연구가 많이 진행되고 있다.

문화진(2019)은 2010년부터 2019년까지 한국연구재단의 등재지 및 등재후보지에 게재된 연구 중 한국 대학생 대상의 다문화교육과 관련한 논문을 대상으로 연구의 특징 및 경향을 내용분석(content analysis)을 통해 분석하였다.

장은영과 이정아(2018)는 국내 다문화교육 연구 중 이중언어와 관련한 연구 동향을 한국연구재단 등재(후보)학술지에 실린 관련 연구를 토대로 연구방법, 연구주제 등을 분석하였다. 이러한 연구들의 특징은 다문화와 관련한 특정 학문영역 또는 주제에 관한 논문을 대상으로 내용분석을 통해 연구 동향을 분석하는 것이다.

위에서 살펴본 바와 같이 특정 분야의 연구 동향을 분석하기 위해서는 분석에 활용하는 자료를 대상으로 텍스트마이닝과 계량서지학 도구 및 방법을 활용하여 분석하는 방법과 내용분석을 통한 질적 분석을 통해 연구 동향을 파악하는 방법으로 구분할 수 있다. 내용분석을 통한 연구 동향의 파악 연구에서는 주로 연구자가 처리할 수 있는 역량 이내의 자료를 대상으로 분석하게 되므로 분석의 대상으로 선별된 자료의 양이 적은 특징을 보이는 반면, 텍스트마이닝이나 계량서지학 도구를 활용하는 방법에서는 컴퓨터의 기술을 활용하여 많은 양의 자료도 처리하는 것이 가능하므로 주로 많은 자료를 선별하여 분석하는 경향이 있다.

2.2.2 토픽 모델링을 활용한 연구

토픽 모델링을 활용한 국내 연구는 다양한 분야에서 진행되어 왔는데, 특히 토픽 모델링을 이용한 연구로는 강범일, 송민, 조화순(2013), 박자현, 송민(2013), 진설아, 송민(2016), 박준형, 오효정(2017), 이상연, 이건명(2014), 배정환, 한남기, 송민(2014), 이금실, 이인주, 이영경(2018), 김세현(2018) 등의 연구가 있다.

강범일, 송민, 조화순(2013)은 웹에서 대선 후보와 관련한 국내 신문자료의 기사를 수집하여 해당 기사에 나타난 토픽을 분석하기 위해 LDA 기반 토픽 모델링 알고리즘을 사용하여 주제를 분석하였다. 이를 통해 18대 대선 후보들의 기사에서 형성되는 주제들을 추출하고, 이러한 주제들이 매체별로 보이는 차이, 주제를 구성하는 단어들의 내용의 차이, 시기별 주제 분포의 차이를 살펴보았다.

박자현, 송민(2013)은 국내 문헌정보학 분야의 연구동향을 분석하기 위하여 문헌정보학 분야의 주요 학술지 4종에 실린 논문의 초록을 수집하여 LDA 기반의 토픽 모델링을 실시하였다. 이를 통해 문헌정보학자들의 주요 연구주제를 규명하고 문헌정보학의 주요 연구주제의 연도별 추이를 분석하고 새롭게 활발한 연구가 진행되는 연구주제(hot topics)와 연구의 인기가 줄어들고 있는 연구주제(cold topics)를 밝히고자 하였다.

진설아, 송민(2016)은 정보학 분야 학술지의 학제성을 측정하는 연구에서 정보학 분야의 학술지 20종을 대상으로 논문의 제목과 초록을 대상으로 LDA를 활용한 토픽 모델링을 실행하여 정보학 분야 학술지의 전체 토픽과 개별 학술지의 토픽 분포를 파악하고, 이를 활용하여 정보학 분야 학술지의 학제적 특성을 분석하였다.

박준형, 오효정(2017)은 국내 기록관리학 분야의 연구동향을 분석하기 위해 LDA 기반의 토

픽 모델링 기법과 HDP(Hierarchical Dirichlet Process) 기반의 토픽 모델링 기법을 적용하고 그 결과를 바탕으로 두 토픽 모델링 기법의 차이를 비교 분석하였다. 이를 통해 LDA 토픽 모델링은 빈도수가 높은 키워드에 영향을 많이 받으며 일반적인 키워드가 많아 각 토픽의 특징을 파악하기 어려운 반면, HDP 토픽 모델링은 고빈도 키워드에 상대적으로 적은 영향을 받으며 각 토픽마다 유일하게 등장하는 키워드가 많이 포함되어 있어 각 토픽의 특징을 뚜렷하게 구분할 수 있음을 밝혔다.

이상연, 이건명(2014)과 배정환, 한남기, 송민(2014)은 SNS상에서의 토픽 트랜드를 분석하기 위해 트위터로부터 데이터를 활용하여 트윗에 나타난 토픽을 분석하기 위해 LDA 토픽 모델링 기법을 사용하였다. 이를 통해 LDA와 같은 토픽 모델링의 방법이 트위터와 같은 SNS상의 데이터에도 적용할 수 있음을 밝히고 나아가 SNS상의 이슈 변화를 추적할 수 있음을 보여주었다.

이금실, 이인주, 이영경(2018)은 관광분야의 저널을 대상으로 2009년부터 2018년까지 10년 동안의 가상현실과 관련한 논문의 연구동향을 분석하기 위해서 LDA 기반의 토픽 모델링 기법을 사용하였고 이를 통해 10년 동안 가상현실과 관련한 관광분야에서 주목받는 주제(hot topics)과 소멸되는 주제(cold topics)를 도출함으로써 해당 분야의 연구 동향을 살펴보았다.

김세현(2018)은 1993년 이후부터 2016년까지 국내에서 발간된 3,242편의 다문화와 관련한 논문 중 국문 초록 정보가 확인된 2,334편의 논문을 중심으로 국문 초록에 대하여 LDA 기반 토픽 모델링 분석과 연결망 분석을 진행하였다. 이를 통해 한국의 다문화 연구가 교육, 이주, 정책 토픽을 중심으로 논의가 전개되었으며 시기에 따른 세부 주제 및 연구대상, 연구방향의 변화가 있음을 관찰하였다.

이상에서 살펴본 선행 연구의 특징을 요약하면 다음과 같다. LDA 기반의 토픽 모델링을 적용한 연구 동향의 분석이 문헌정보학, 정보학뿐만 아니라 관광분야, 다문화 영역에 이르기까지 다양하게 수행되었다. LDA 기반의 토픽 모델링 기법을 적용함으로써 해당 분야의 주제를 분석하고 또한 이러한 주제를 다양한 시기별로 분석함으로써 주제의 변화를 추적하여 연구 동향을 분석하였다. 이러한 방법을 적용하는 주제분석 및 연구 동향 분석은 SNS 분야에도 적용되었으며 트윗과 같은 데이터를 활용하여 SNS상의 텍스트에 나타난 주제 분석과 이슈의 변화를 추적하는 연구가 진행되었다.

다문화와 관련한 연구동향 분석은 주로 다문화 영역 내의 세부분야와 관련한 논문을 대상으로 내용분석에 기반한 주제 분석을 통한 연구 동향을 분석하는 연구가 주로 진행되었으며 다문화와 관련한 모든 영역에 대한 주제 분석 및 연구 동향 분석에 대한 연구는 매우 드물게 수행되었다.

특정 학문 분야의 학술지 논문을 대상으로 LDA 기반의 토픽 모델링 기법을 적용한 연구는 토픽 분석을 위해 주로 논문의 초록을 활용하였으며 논문에 나타난 텍스트 전체를 대상으로 한 연구는 찾기 어려웠는데 이는 LDA 기반의 토픽 모델링을 적용한 주제분석에서 텍스트의 양이 많아지면 그만큼 토픽 분석에 소요되는 시간이 늘어남으로써 연구 수행에 제약이

많아지게 되기 때문에 상대적으로 분석이 용이한 초록을 분석의 대상으로 선택한 결과로 나타난 현상으로 생각된다.

LDA 기반의 토픽 모델링 기법을 다문화 주제 분야에 적용하는 것은 토픽 모델링을 문헌정보학과 같은 특정한 단위 학문 영역에 적용하는 것과는 차이가 있다. 이는 다문화 관련 연구가 특정 단위학문 분야 내에서 연구되기보다 훨씬 다양한 분야에서 광범위하게 연구가 진행되고 있어 특정한 하나의 학문 영역을 다문화와 관련한 학문 영역이라고 규정하기가 매우 어렵기 때문이다. 이러한 의미에서 토픽 모델링을 다문화 주제 분야에 적용하여 해당 주제 분야의 이슈 변화를 살펴보는 것이 광범위한 학문영역에서 다루어지는 특정 주제(다문화)와 관련한 이슈를 추적하므로 연구의 동향을 밝히는 것은 여전히 의미가 있다. 본 연구에서는 다양한 학문 분야에서 진행된 다문화와 관련한 국내 연구를 대상으로 초록이 아닌 논문 전체의 내용을 대상으로 토픽 모델링을 실시하여 다문화와 관련한 국내 연구의 주요 토픽을 규명함으로써 다문화 관련 연구의 동향을 분석하고자 한다.

3. 연구의 내용, 범위 및 방법

3.1 연구 질의

본 연구를 위한 연구 질의는 다음과 같다.

(1) 다문화 관련 국내 학술 문헌에 나타나는 토픽은 시간의 흐름에 따라 어떻게 변화하고 있는가?
(2) 최근 5년 동안 새롭게 떠오른 주요 토픽에는 어떤 것이 있는가?
(3) 최근 10년 동안 꾸준히 관심을 보이는 핵심 토픽에는 어떤 것이 있는가?

3.2 데이터의 수집 범위

다문화 관련 국내 학술 연구의 연구주제 변화를 파악하기 위하여 한국연구재단에 등재 및 등재 후보 학술지를 대상으로 '다문화'라는 키워드를 포함한 논문 약 8,000건 중 텍스트로 원문 입수가 가능한 논문을 한정하였다. 논문의 발행 시기는 앞에서 수집한 논문의 가장 빠른 발행년도인 2000년 이후 발행된 학술 논문을 대상으로 하였다.

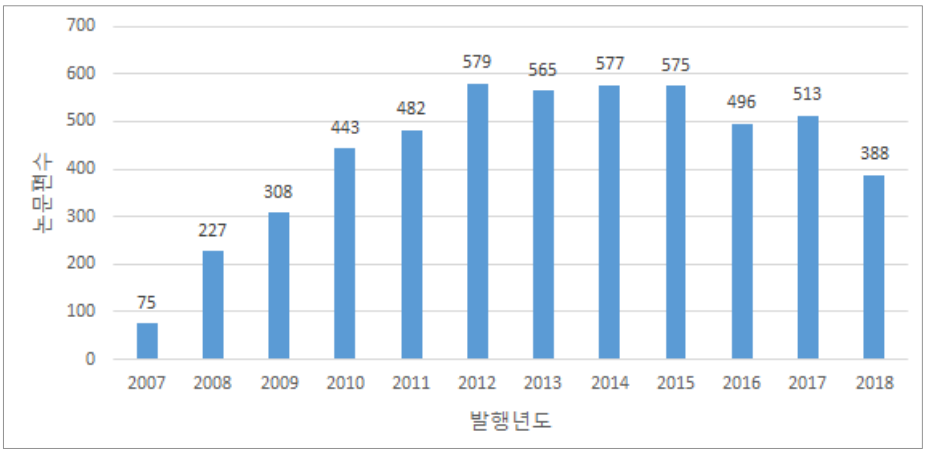
3.3 방법

3.3.1 텍스트의 수집 및 전처리

본 실험 연구에 사용될 텍스트를 수집하기 위해 한국학술지인용색인(https://www.kci.go.kr)에서 키워드로 "다문화"라는 단어를 포함한 논문을 망라적으로 수집하였다. 이렇게 수집된 문헌의 텍스트를 제목, 키워드, 초록 등으로 구조화하였다. 수집된 텍스트를 활용하여 토픽 분석을 하기 위해서 각 논문의 텍스트를 코모란 한글형태소를 분석기를 활용하여 형태소 분석을 한 다음 각 논문의 텍스트 중 명사만을 추출하여 해당 명사 단어군을 분석의 대상으로 삼았다. 한글형태소 분석기를 통해 추출된 단어를 살펴보고 한글형태소분석기의 성능 제약으로 인해 잘못 분석한 단어들은 별도로 형태소분석기를 위한 이용자사전에 해당 단어를 추가하여 한글형태소분석기의 성능을 보완하였다.

〈표 1〉 코퍼스 기본 정보

수집 경로	https://www.kci.go.kr/kciportal/po/search/poArtiSear.kci
코퍼스 크기	논문 5,345개
수록 기간	2007년 ~ 2018년

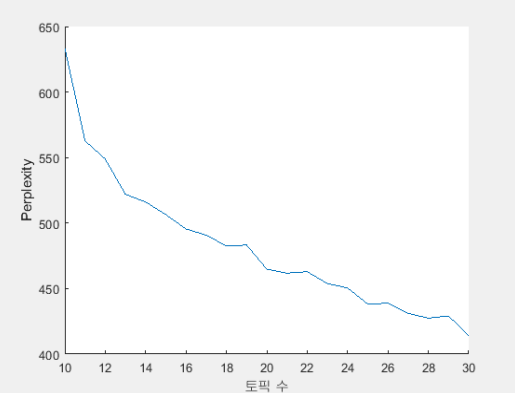


〈그림 2〉 국내 다문화 관련 학술 논문 발행 추이

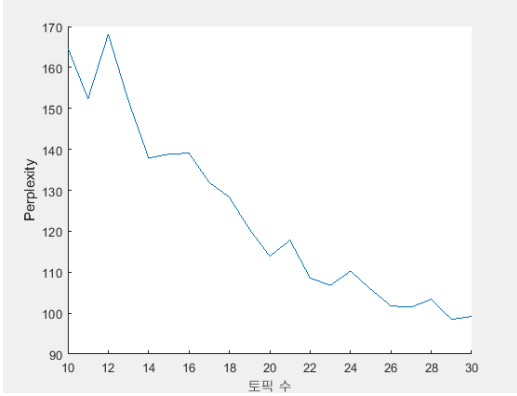
3.3.2 토픽 분석 방법

전처리 과정을 통해 정제된 텍스트를 활용하여 토픽분석을 하기 위하여 본 연구에서는 Matlab의 Text Analytics Toolbox를 활용하여 LDA 기반의 토픽 모델링을 한 후 이를 활용하여 토픽 분석하였다. LDA 토픽 분석의 경우, LDA 토픽 분석 결과의 품질 향상을 위해서는 문서와 토픽 간의 밀도를 나타내는 α의 값을 최적의 값으로 설정하는 것이 중요하다. 즉 해당 코퍼스 내에 존재하는 토픽의 개수(a)를 몇 개로 설정하느냐에 따라 LDA 분석결과로 나타나는 토픽의 품질이 결정되므로 최적의 토픽 개수를 찾는 작업이 매우 중요하다. 그러나, LDA 기반 토픽 모델링에서는 식별가능한 단어(w)이외의 나머지 변수들은 모두 감추어져(latent) 있기 때문에 가장 이상적인 토픽의 개수를 명확히 정의할 수 없는 한계가 있다. 본 연구에서는 이러한 문제를 해결하기 위한 한 방법으로 LDA 모델에 적용할 최적의 토픽의 개수를 선정하기 위하여 토픽의 개수가 수집된 전체 코퍼스 내에 최소 10개에서 최대 30개의 숨겨진 토픽이 있음을 가정하고 최적의 토픽 수를 찾기 위하여 각 토픽 모델을 시뮬레이션하였다.

〈그림 3〉은 수집된 각 논문에 표현된 제목, 저자 키워드, 초록 모두를 하나의 문서에 포함된 텍스트로 간주하여 하나로 묶고 LDA 토픽 모델을 작성하여 각 모델별로 혼란도(perplexity) 값의 변화를 추적한 그림이다. 〈그림 4〉는 수집된 각 논문에 포함된 저자 키워드만을 대상으로 하여 LDA 토픽 모델을 작성하여 각 모델별로 혼란도 값의 변화를 추적한 그림이다. 〈그림 3〉을 〈그림 4〉와 비교하여 볼 때 〈그림 3〉은



〈그림 3〉 토픽 수와 혼란도의 상관관계 - 제목, 키워드, 초록

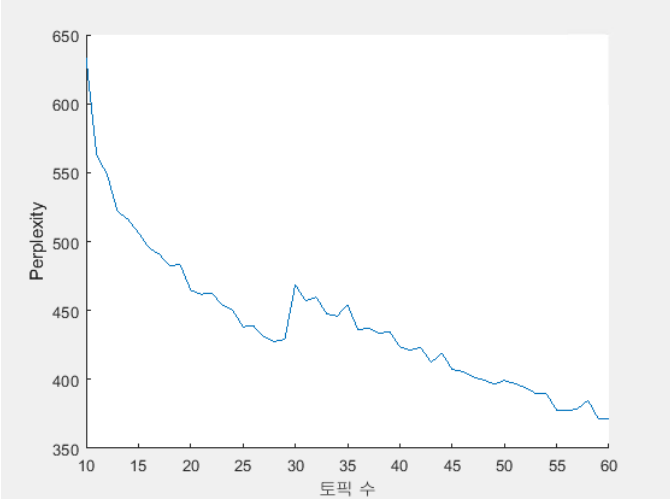


〈그림 4〉 토픽 수와 혼란도의 상관관계 - 키워드

비교적 완만하고 매끄럽게 혼란도의 값이 토픽 수와 반비례하는 상태로 나타나는 반면, 〈그림 4〉의 경우 중간중간 혼란도의 값이 감소하다가 크게 돌출하는 경우(토픽수 12, 16, 21, 24, 28, 30)가 여러 번 발생함을 볼 수 있다. 이는 〈그림 4〉로 표현된 실험에서 저자 키워드만을 대상으로 토픽 분석을 하였기 때문에 〈그림 3〉의 실험과 비교하여 볼 때 제공되는 텍스트의 양이 상대적으로 매우 적은 결과로 인한 것으로 해석할 수 있다. 〈그림 3〉의 경우에도 중간중간에 미세하게 두드러지는 부분(토픽수 19, 22, 26, 29)이 식별되는데, 이러한 지점에서 토픽 분석의 정확성을 담보하는 통계적 확률이 변화하는 지점이라 할 수 있다. 이러한 사실을 바탕으로 토픽 모델 생성을 위해서 키워드만을 이용하는 대신 코퍼스 내의 각 논문을 표현하는 제목, 저자키워드, 초록 모두를 포함하는 텍스트를 활용하여 토픽 모델을 생성하였다.

본 연구에서 LDA 토픽 분석에서 토픽과 단어 사이의 밀도를 설정하는 β의 값으로는 50을 기본으로 설정하였다. 이는 하나의 토픽을 표현하는 연관 단어의 수가 50개임을 의미한다. β의 값이 다르게 변화함에 따라 결과로 제시되는 토픽을 구성하는 단어의 종류와 수가 달라지게되므로 α의 설정 값과 마찬가지로 토픽의 품질에 영향을 미치나 본 연구에서는 이를 50으로 설정하여 실험을 진행하였다.

각 논문의 텍스트를 모두 활용하는 방법을 통해 토픽 수를 60개까지 더 늘려 실험한 결과를 보면 토픽 수 29개까지의 혼란도의 값이 조금씩 감소하다가 토픽 수 30개에서 그 값이 갑자기 증가하는 것을 〈그림 5〉와 같이 관찰하였다. 이를 통해 볼 때, 전체 코퍼스 내에 존재하는 잠재적인 토픽의 수를 29개로 보는 것이 실험 결과에 비추어 볼 때 적합하다고 판단하여 이후 토픽 수를 29개로 설정하여 실험을 진행하였다. 물론 토픽의 수를 지속적으로 증가시켜 실험을 진행하면 혼란도의 값이 최소가 되는 지점을 찾을 수도 있겠지만 그 경우에는 토픽의 수가 매우 많아지게 되어 본 연구의 주제인 다문화 관련 연구분야의 주요 연구동향을 추적하는데 분석해야 하는 토픽의 수가 너무



〈그림 5〉 토픽 수와 혼란도의 상관관계 - 제목, 키워드, 초록 (토픽 수 10~60개)

많아 오히려 부정적인 효과를 나타낼 수도 있을 것이다. 여기서 복잡도의 값이 작다는 것은 해당 토픽이 실제 문헌의 사례를 기계학습을 통해 잘 반영한다는 의미이지 사람이 잘 이해하고 해석하기 좋다는 의미는 아니다.

토픽의 수를 설정하고 토픽을 분석한 후, 2002년도부터 2018년도까지 시계열별로 다문화 관련 연구 토픽 트랜드의 변화를 살펴보기 위하여 앞서 설정한 29개의 토픽에 대한 트랜드의 변화를 추적하였다.

4. 다문화 관련 토픽 분석

4.1 연도별 연구 토픽 트랜드

연도별 연구 토픽의 트랜드를 분석하기 위해서 각 연도에 발행된 모든 논문의 제목, 저자 키워드, 초록에 나타난 텍스트를 각 연도별 하나의 문서로 간주하여 모으고 이를 대상으로 앞에서 분석한 토픽 모델을 적용하여 각 연도별 문헌에 나타나는 토픽의 분석을 시도하였다.

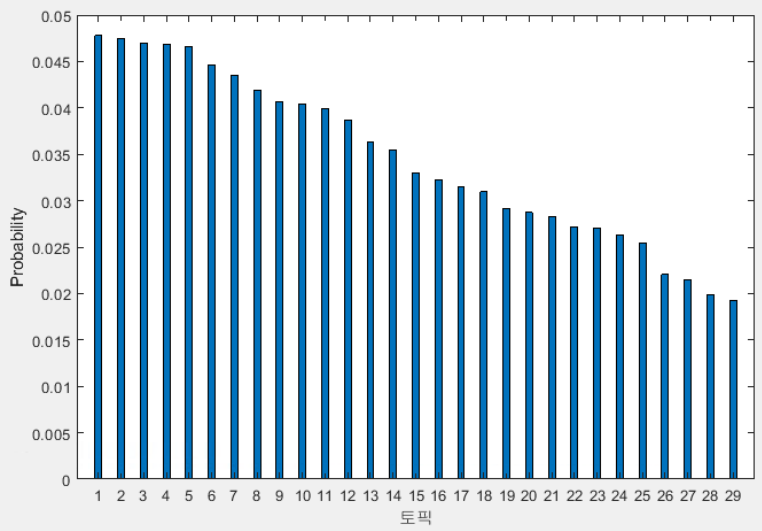
토픽 모델링을 통해 분석한 29개의 토픽의 확률 분포는 〈그림 6〉과 같다. 다문화 관련 국내 연구와 관련하여 연도별 연구 토픽 트랜드의 변화를 추적하기 위하여 전체 29개의 토픽 중 토픽 확률 분포의 상위 5개의 토픽(설정 값 p〉0.045)을 선정하고 이를 추적하였다.

이를 통해 선정한 상위 5개의 토픽을 구성하는 상위 단어는 〈표 2〉와 같다.

4.1.1 토픽별 주제

(1) 토픽 1. 학교 다문화 교육

토픽 1을 구성하는 단어의 군집을 활용하여 토픽 1이 다루고 있는 토픽을 좀 더 이해하기 쉽도록 〈그림 7〉과 같이 워드 클라우드로 정보를 시각화하였는데, 이를 통해 나타난 토픽 1의 주제를 살펴보면 학교를 중심으로 한 다문화



〈그림 6〉 각 토픽별 확률 분포

〈표 2〉 토픽별 주요 단어(상위 30개)

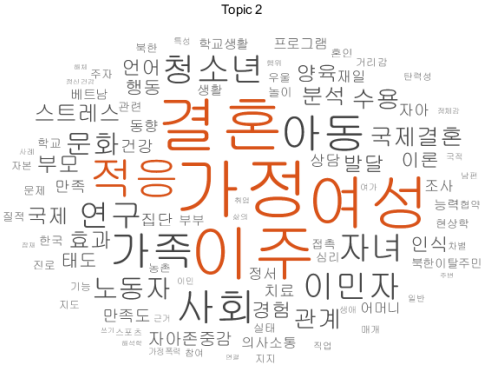
토픽1		토픽2		토픽3		토픽4		토픽5	
단어	Score	단어	Score	단어	Score	단어	Score	단어	Score
교육	0.225	가정	0.068	문화	0.062	문화	0.059	사회	0.095
교사	0.031	이주	0.059	교육	0.059	정체성	0.036	정책	0.071
문화	0.029	결혼	0.058	주의	0.054	주의	0.029	외국인	0.052
한국어	0.026	여성	0.056	사회	0.044	한국	0.026	가족	0.040
과정	0.025	적응	0.031	언어	0.034	민족	0.023	통합	0.038
역량	0.024	아동	0.028	상호	0.020	시민	0.021	지원	0.033
유아	0.020	가족	0.028	문학	0.018	국가	0.016	이민	0.022
학습	0.018	사회	0.026	다양성	0.017	사회	0.015	주민	0.022
인식	0.015	청소년	0.019	통합	0.016	종교	0.013	지역	0.021
프로그램	0.015	자녀	0.018	인종	0.015	공동체	0.013	복지	0.017
가정	0.012	이민자	0.018	동화	0.015	통일	0.013	청소년	0.013
연구	0.012	연구	0.017	정체성	0.014	갈등	0.011	주의	0.012
분석	0.012	문화	0.013	이중	0.013	디아스포라	0.011	이주	0.012
효능	0.011	노동자	0.012	차별	0.012	국민	0.010	범죄	0.012
교수	0.011	수용	0.011	타자	0.012	미디어	0.010	인권	0.012
음악	0.011	관계	0.010	공간	0.010	기독교	0.009	센터	0.011
능력	0.011	국제결혼	0.010	정책	0.009	중국	0.009	서비스	0.011
학생	0.011	분석	0.009	윤리	0.009	의식	0.009	요인	0.010
미술	0.010	스트레스	0.009	시민	0.008	역사	0.009	입국	0.008
교과서	0.010	효과	0.009	소수자	0.007	도시	0.008	활동	0.008
태도	0.010	부모	0.008	국제	0.007	민족주의	0.008	중도	0.008
학교	0.009	양육	0.008	차이	0.007	전통	0.008	제도	0.008
감수성	0.008	국제	0.008	이해	0.007	선교	0.008	경찰	0.007
모형	0.007	발달	0.008	공동체	0.007	주민	0.007	프로그램	0.007
내용	0.007	경험	0.008	인권	0.006	세계	0.007	체계	0.007
예비	0.007	언어	0.007	정치	0.006	교회	0.007	보호	0.006
국어	0.007	이론	0.007	소설	0.005	읽기	0.007	학교	0.006
수용	0.007	인식	0.007	소통	0.005	네트워크	0.007	국제	0.006
활동	0.007	태도	0.006	관용	0.005	시민권	0.007	이민자	0.006
경험	0.006	행동	0.006	미국	0.005	도서관	0.007	지방	0.006

관련 교육에 관한 것임을 파악할 수 있다. 이를 바탕으로 토픽 1의 잠정적인 주제를 “학교 다문화 교육”으로 명명하였다.

〈그림 7〉 토픽 1의 워드 클라우드 - 학교 다문화 교육

(2) 토픽 2. 결혼 이주 여성

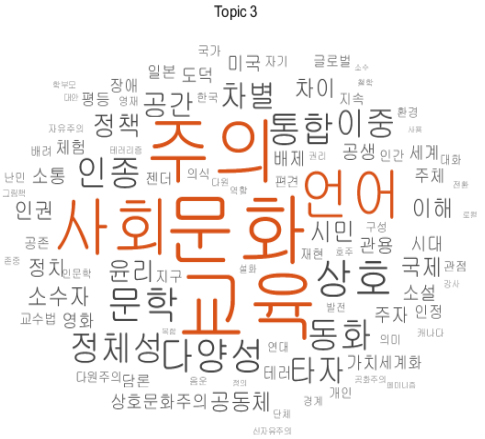
〈그림 8〉을 살펴보면 두 번째 토픽 모델의 중심 단어군이 다문화와 관련한 가정, 결혼, 여성, 이주 등으로 나타남을 알 수 있고, 이를 바탕으로 토픽 2의 잠정적인 주제를 “결혼 이주 여성”으로 명명하였다.



〈그림 8〉 토픽 2의 워드 클라우드 - 결혼 이주 여성

(3) 토픽 3. 문화 교육 및 다양성

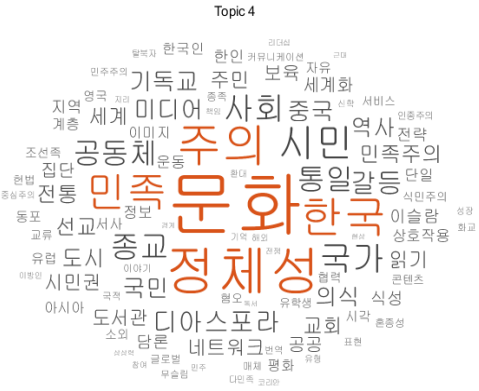
토픽 3은 〈그림 9〉와 같이 문화, 교육, 주의, 사회, 언어, 상호, 문학, 다양성 등의 중심어로 구성되어 있다. 이 토픽에 포함된 단어에 상호문화주의, 다원주의 등의 단어가 등장하는 것으로 미루어 보아 상호문화주의를 해당 토픽의 주제 중 하나로 볼 수 있겠다. 또한 다른 토픽의 상위어에 포함되지 않은 문학이 포함되어 있는 것으로 미루어 보아 문학과 관련한 다문화 연구도 이 토픽에서 다루어지고 있음을 추측할 수 있겠다. 토픽 3의 잠정적인 주제를 “문화 교육 및 다양성”으로 명명하였다.



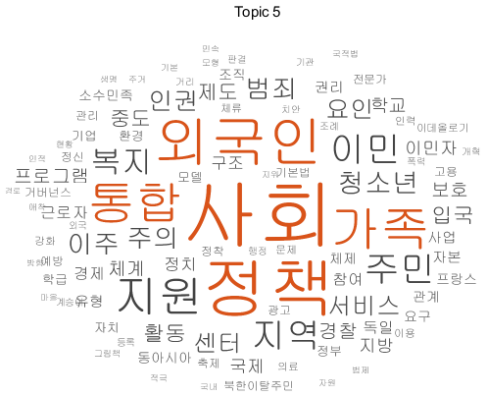
〈그림 9〉 토픽 3의 워드 클라우드 - 문화 교육 및 다양성

(4) 토픽 4. 문화 정체성과 민족주의

네 번째 토픽의 중심어는 〈그림 10〉과 같이 문화 정체성, 민족, 주의 등으로 표현되고 있다. 이를 바탕으로 토픽 4의 잠정적인 주제를 “문화정체성과 민족주의”로 명명하였다.



〈그림 10〉 토픽 4의 워드 클라우드 - 문화 정체성과 민족주의



〈그림 11〉 토픽 5의 워드 클라우드 - 외국인 사회 통합 정책

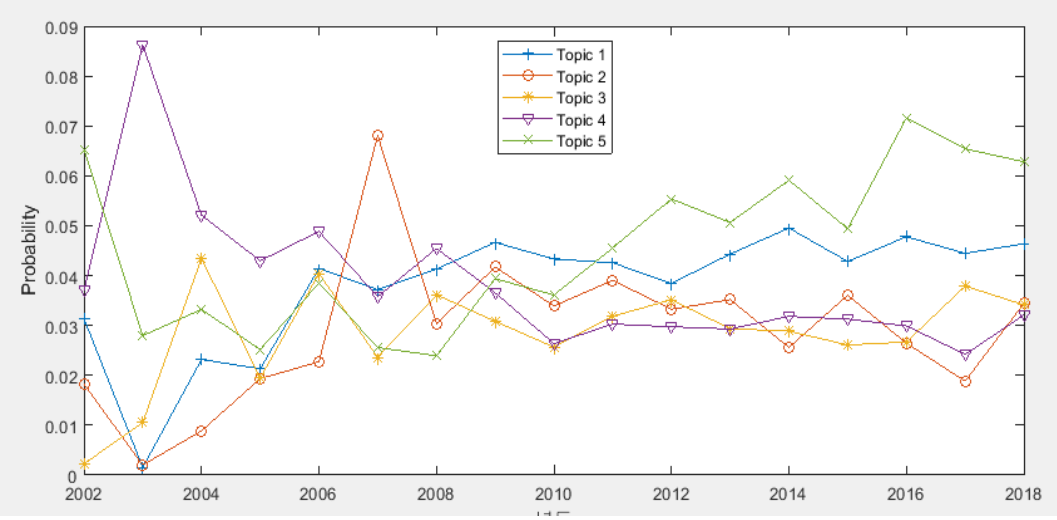
(5) 토픽 5. 외국인 사회 통합 정책

다섯 번째 토픽의 중심어는 〈그림 11〉과 같이 사회, 정책, 외국인, 가족, 통합 등으로 표현되고 있다. 이를 바탕으로 토픽 5의 잠정적인 주제를 "외국인 사회 통합 정책"으로 명명하였다.

4.1.2 토픽 트랜드 분석

앞에서 살펴본 토픽들이 2002년 이후의 년도별 트랜드 변화를 그래프로 나타내면 〈그림 12〉와 같다.

국내 다문화와 관련한 연구가 2000년대 초반에는 토픽 4(문화정체성과 민족주의)와 관련한 연구가 활발하였던 반면 시간이 지나면서



〈그림 12〉 연도별 토픽 트랜드의 변화

해당 토픽에 대한 연구가 전에 비해 줄어든 것을 알 수 있다. 한편, 토픽 5(외국인 사회 통합 정책)의 경우에는 해당 토픽과 관련한 연구가 시간이 지나면서 점점 더 증가하고 중요한 이슈로 부각되고 있음이 드러난다. 토픽 1(학교 다문화 교육)의 경우 2006년 이후부터 꾸준하게 연구가 진행되고 있음을 알 수 있다. 토픽 2(결혼 이주 여성)와 관련한 연구는 2007년도에 정점을 이루었고 그 이후에 다소 감소하였으나 최근까지도 지속적으로 연구되고 있음을 알 수 있다.

5. 결 론

이 논문은 국내에서 진행된 다문화 연구를 대상으로 중요 연구 토픽을 추출하여 다문화 영역의 토픽의 흐름을 파악함으로써 해당 주제 분야의 연구 동향을 파악하고자 하였다. 기존의 선행 연구에서는 다문화와 관련한 연구자료의 수집 범위를 작은 하위 분야로 제한하여 연구자가 직접 내용 분석을 행하는 연구들이 많이 진행되었다. 한편 텍스트마이닝을 활용한 최근의 연구에서는 기관 아카이브 내에 수집된 자료들을 대상으로 국내 다문화 관련 논문의 토픽을 분석하였다. 본 연구에서는 한국연구재단의 인용색인에 등재된 국내 학술지를 대상으로 특정 주제분야를 제한하지 않고 다문화와 관련된 논문의 제목, 저자 키워드, 초록을 모두 포함한 텍스트를 대상으로 토픽 모델링 기법을 적용하여 포괄적으로 다문화 관련 연구 토픽의 흐름을 분석하고자 하였다.

이 연구에서 연구 토픽의 흐름을 분석하기 위해 LDA 분석을 시행하였는데, 최적의 LDA 모델을 선정하기 위해 각 모델의 토픽 수를 달리하여 실험한 후 해당 모델의 혼란도를 비교하여 가장 안정된 모델을 선정하였다. 그 결과 본 연구에서 사용한 코퍼스를 대상으로 하였을 경우 29개의 토픽 수가 가장 적합한 것으로 파악되었다. 이렇게 선택된 LDA 모델을 토대로 각 연도별 연구 문헌의 토픽을 비교하고 그 흐름을 추적하였다. 이를 통해 수집된 코퍼스의 수록 기간 전반기에는 문화, 정체성, 주의, 한국, 민족, 시민 등으로 대표되는 “문화정체성과 민족주의”의 주제(토픽4)가 활발하게 진행된 반면, 근래에 와서는 이에 관한 관심이 상대적으로 줄어든 것을 알 수 있었다. 전반적으로 상위군에 포함된 주요 토픽들은 2008년 이후 지속해서 연구 토픽으로 논의되고 있으며 토픽 1(학교 다문화 교육)과 토픽 5(외국인 사회 통합 정책)는 최근 5년 동안 다른 토픽에 비해 두드러지게 주목받고 있음을 파악할 수 있었다. 이를 통해 다문화 관련 연구에서의 최근 중요 관심사를 식별할 수 있었다.

본 연구에서 살펴본 다문화 관련 연구의 토픽의 흐름을 통해 최근 5년간 지속적으로 다문화와 관련한 사회 통합 문제가 주요 이슈였음을 볼 때 앞으로도 이 토픽에 대한 관심이 지속될 것으로 예측할 수 있다. 또한 학교 교육에서의 다문화와 관련한 이슈들도 앞으로도 계속 지속될 것으로 예상된다.

참 고 문 헌

[1] 강범일, 송민, 조화순. 2013. 토픽 모델링을 이용한 신문 자료의 오피니언 마이닝에 대한 연구.『한국문헌정보학회지』, 47(4): 315-334.

[2] 국가기록원. 2019. 기록으로 만나는 대한민국: 다문화사회. [online] [cited 2019. 8. 4.] 〈http://theme.archives.go.kr/next/koreaOfRecord/MultiSociety.do〉

[3] 안성주, 양정진. 2018. LDA와 Word2Vec을 결합한 생물정보 토픽 모델 연구.『2018 한국컴퓨터종합학술대회 논문집』, 2065-2067.

[4] 김세현. 2018. 비정형자료분석을 통해 살펴본 한국의 다문화 연구.『한국인구학』, 41(1): 1-27.

[5] 문화진. 2019. 한국 대학생 대상의 다문화교육 연구 동향.『다문화사회연구』, 12(2): 181-215.

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

[7] 박준형, 오효정. 2017. 국내 기록관리학 연구동향 분석을 위한 토픽모델링 기법 비교.『한국도서관·정보학회지』, 48(4): 235-258.

[8] 배정환, 한남기, 송민. 2014. 토픽 모델링을 이용한 트위터 이슈 트레킹 시스템.『한국지능정보시스템학회 2014년 춘계학술대회논문집』, 305-312.

[9] 음수민, 이수길, Xiangyu Meng, 조성원, 이철웅. 2019. LDA 기반의 토픽모델링을 이용한 철도차량용 무선급전시스템 연구 동향 분석.『대한산업공학회지』, 45(4): 284-301.

[10] 이금실, 이인주, 이영경. 2018. LDA(Latent Dirichlet Allocation) 기반 토픽 모델링 기법을 활용한 관광분야의 가상현실(Virtual Reality) 연구동향 분석.『2018년도 제50차 한국관광레저학회 학술발표대회 논문집』, 425-432.

[11] 이상연, 이건명. 2014. 댓글 그래프 기반 토픽 모델을 사용한 트렌드 추출.『한국지능시스템학회 학술발표 논문집』, 24(2): 99-100.

[12] 장은영, 이정아. 2018. 국내 다문화교육 연구 동향 분석.『교육문화연구』, 24(3): 501-521.

[13] 장임숙, 장덕현, 이수상. 2011. 다문화연구의 지식구조에 관한 네트워크 분석.『한국도서관·정보학회지』, 42(4): 353-374.

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

[15] Blei, David M., Ng, Andrew Y. and Jordan, Michael I. 2003. "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, 3: 993-1022.

[16] Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K. and Harshman, R. 1990. "Indexing by latent semantic analysis." *Journal of the American society for information*

science, 41(6): 391-407.

[17] Hofmann, T. 1999. Probabilistic latent semantic analysis. In Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence (UAI'99), Kathryn B. Laskey and Henri Prade (Eds.). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 289-296.

[18] Liu, Z., Zhang, Y., Chang, E. Y. and Sun, M. 2011. Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing. ACM Transactions on Intelligent Systems and Technology (TIST), 2(3): 26.

[19] Wang, C. and Blei, D. M. 2013. "Variational inference in nonconjugate models." *Journal of Machine Learning Research*, 14(1): 1005-1031.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] Kang, Beomil, Song, Min and Jho, Whasun. 2013. "A Study on Opinion Mining of Newspaper Texts based on Topic Modeling." *Journal of the Korean Society for Library and Information Science*, 47(4): 315-334.

[2] National Archives of Korea. 2019. "Korea of Record: Multicultural Society." [online] [cited 2019. 8. 4.] 〈http://theme.archives.go.kr/next/koreaOfRecord/MultiSociety.do〉

[3] Ahn, Sung-Joo and Yang, Jung-Jin. 2018. "A Study on Topic Models using LDA and Word2Vec in Bioinformatics." *Proceedings of Korea Computer Congress 2018*, 2065-2067.

[4] Kim, Sehyun. 2018. "A Study of Korea's Multicultural Research Trends Using Unstructured Data Analysis." *Korea Journal of Population Studies*, 41(1): 1-27.

[5] Moon, Hwa-Jin. 2019. "Trends in Research on Multicultural Education of University Students in Korea." *The Journal of Multicultural Society*, 12(2): 181-215.

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

[7] Park, JunHyeong and Oh, Hyo-Jung. 2017. "Comparison of Topic Modeling Methods for Analyzing Research Trends of Archives Management in Korea: focused on LDA and HDP." *Journal of Korean Library and Information Science Society*, 48(4): 235-258.

[8] Bae, Jung-hwan, Han, Nam-gi and Song, Min. 2014. "Twitter Issue Tracking System by Topic Modeling Techniques." *Proceedings of the Korea Intelligent Information Systems Society*, 305-312.

[9] Eum, Soomin, Lee, Sugil, Meng, Xiangyu, Cho, Sung Won and Lee, Chulung. 2019. "Analysis of Research Trends of Wireless Power Transfer System for Locomotives Using Topic Modeling Based on LDA Algorithm." *Journal of the Korean Institute of Industrial Engineers*, 45(4): 284-301.

[10] Lee, K. S., Lee, Injoo and Lee, Young K. 2018. "Research Trends Analysis on Virtual Reality in Tourism using LDA(Latent Dirichlet Allocation) topic modeling." *Proceedings of the Korea Academic Society of Tourism and Leisure*, 425-432.

[11] Lee, Sang Yeon and Lee, Keon Myung. 2014. "Trend Extraction using Topic Model Based on Reply Graph." *Proceedings of the Korean Institute of Intelligent Systems Conference*, 24(2): 99-100.

[12] Jang, Eun-Young and Lee, Jeong-Ah. 2018. "Trends in multicultural studies published in Korea: An analysis of the studies focusing on bilingualism and/or bilingual education." *Journal of Education & Culture*, 24(3): 501-521.

[13] Jang, Im Sook, Chang, Durk-Hyun and Lee, Soosang. 2011. "The Knowledge Structure of Multicultural Research Papers in Korea." *Journal of Korean Library and Information Science Society*, 42(4): 353-374.

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