https://doi.org/10.5392/JKCA.2020.20.05.118

사회과학을 위한 양적 텍스트 마이닝: 이주, 이민 키워드 논문 및 언론기사 분석

Quantitative Text Mining for Social Science: Analysis of Immigrant in the Articles

이수정*, 최두영**
한국외국어대학교 아랍어통번역과*, 한국외국어대학교 중동아프리카학과**

Soo-Jeong Yi(sooislam86@gmail.com)*, Doo-Young Choi(choidooyoungwicks@gmail.com)**

요약

본 연구는 최근 사회과학에서 실시되고 있는 양적 텍스트 분석의 흐름과 분석을 실시함에 있어 주의해야 할 사례를 포함하여 기술 하였다. 특히, 2017년부터 2019년까지 3년간 학술지와 언론에서 사용된 "이주", "이민" 키워드를 기반으로 사례연구를 실시하였다. 이를 위해 최근 사회과학분야에서 주목 받는 자연어 처리 기술(NLP)를 이용한 양적 텍스트 분석 (Quantitate text analysis)을 사용하였다. 양적 텍스트 분석은 문서를 구조적 데이터로 변환하여, 가설의 발견 및 검증을 실시하는 데이터 과학의 영역으로, 데이터의 모델링 및 가시화 등이 가능하고, 특히 비구조화 된 데이터를 구조화할 수 있다는 점에서 사회과학 분야에 많이 도입하였다. 따라서 본 연구는 양적 텍스트 분석을 통해 "이주", "이민"을 키워드로 한 연구 및 언론 기사에 대한 통계 분석을 실시하고 도출된 결론에 대한 해석을 실시하였다.

■ 중심어 : | 데이터 마이닝 | 양적분석 | R 텍스트 마이닝 | 이민 | 이주 |

Abstract

The paper introduces trends and methodological challenges of quantitative Korean text analysis by using the case studies of academic and news media articles on "migration" and "immigration" within the periods of 2017-2019. The quantitative text analysis based on natural language processing technology (NLP) and this became an essential tool for social science. It is a part of data science that converts documents into structured data and performs hypothesis discovery and verification as the data and visualize data. Furthermore, we examed the commonly applied social scientific statistical models of quantitative text analysis by using Natural Language Processing (NLP) with R programming and Quanteda.

■ keyword : | Data Mining | Quantitative Analysis | R Text Mining | Immigrant | Migration |

I. 서 론

최근 빅 데이터가 이슈화 되고 있으며, 빅 데이터라는 말 속에는 첨단이라는 이미지가 강하다. 하지만 다양한 정보를 통계적으로 변환한 후 사회 현상을 분석하는 기법은 언어학, 통계학, 계산 공학 분야에서 오래전부터 사용되던 개념이며 1959년 콘텐츠 분석의 창시자인 풀(Pool)의 "저자의 말은 곧 의미이다"[1]는 현대 양적 텍스트 분석의 근간으로 자주 인용된다.

본 논문은 텍스트를 자연어처리 기법을 이용, 통계적

접수일자 : 2020년 02월 28일
수정일자 : 2020년 03월 23일
심사완료일 : 2020년 03월 30일
교신저자 : 이수정, e-mail : sooislam86@gmail.com

으로 변환하여 사회현상을 분석하는 기법을 설명하고 분석 사례를 제시하고자 한다. 텍스트는 사전 변수를 가지지 않는 비구조형 데이터(unstructured data)인 동시에 각각의 단어가 변수가 되는 다차원 데이터(high dimensional data)가 될 수도 있다[2].

가장 대표적인 비구조형 데이터란 문서 정보가 있으며, 영상, 음원, 이미지 등도 이 범주에 포함된다. 기업에서 생산되는 80% 이상의 데이터는 비구조형 데이터이다[3]. 또한 다차원 데이터는 주로 회귀 분석 가능한 데이터를 말한다[4]. 따라서 텍스트를 기반으로 사회과학적 의미를 가지는 정보를 추출하는 작업이란 비구조형 데이터의 다차원 데이터 분석이라고 할 수 있다. 문서 데이터에서 고도의 신뢰성을 가진 정보를 얻기 위해서는 많은 양의 문서를 사람이 직접 수기로 처리해야 했다. 이와 같은 처리 방식은 많은 노동력과 시간을 할애해야 하는 기술이었다. 그러나 최근 ICT기술 발전과 더불어 컴퓨터를 사용한 자연어 처리가 가능해졌다. 이 기술을 사용하면 연구에 필요한 노동력을 최소화 하고, 시간적 비용을 획기적으로 단축할 수 있었다.

본 논문은 최근 ICT 기술, 특히 최신 양적 텍스트 분석 기법을 기반으로 연구를 진행하였다. 사회 과학에 필요한 양적 텍스트 분석의 개념과 함께 적용에 필요한 과정을 설명하고자 하였다. 아울러, 양적 텍스트 분석에 사용되는 최신 통계분석 모델을 실제 사례에 적용하여 설명하고자 하였다.

II. 분석 목적과 방법

본 논문은 비구조형 데이터를 가진 텍스트에서 단어와 단어를 구성하는 구절이 어떠한 연관성을 가지는지 분석을 하고 이것을 다차원 데이터로 변환하여 통계적 유의미성을 찾고자 하는데 있다. 분석을 위해서 통계계산 위한 프로그래밍 언어이자 소프트웨어 환경인 R[5]을 이용하였으며, 양적 텍스트 분석을 위해서는 2012년부터 EU의 지원을 받아 영국 런던 정경대 (LSE)에서 개발한 사회과학을 위한 양적 텍스트 분석 R 패키지인 Quanteda[6]를 이용하였다. 특히 본 논문에서는 공기어 분석(Co-occurrence Analysis)을 이용하였다. 공기어 네트워크 분석에 있어 구절과 단어의 공기성(공통적으로 언급되며 발생하는 상관관계)이 높을 수록 선이 굵어지며, 단어의 출현 빈도가 높을 수록 텍스트가 커진다. 여기서 R패키지를 쓴 이유로, 범용적인 재현성이 가장 큰 이유다. 따라서 본 논문에서 분석을 위해 사용한 모든 데이터는 https://github.com/Deuy76/Korean_Quantative_Text_Minind에서 확인할 수 있다. 또한 Quanteda의 경우 사회과학을 목적으로 한 패키지인 동시에 기존의 일본의 Mecab[7]등을 이용한 형태소 기반의 방식을 쓰지 않기 때문에 사전 없이 분석 가능한 장점이 있다.

III. 사회과학을 위한 양적 텍스트 분석

양적 텍스트 분석은 다양한 연구에서 사용된다. 본 논문은 이주와 난민이라는 키워드로 사례를 제시함으로써, 연구 방법을 보여주는 것이다. 이를 토대로 다른 사회과학 연구에서 본 연구 방법을 적용하여 다양한 양적 분석을 진행할 수 있을 것이다. 양적 텍스트 마이닝을 통해 다다를 수 있는 목적으로 1) 내용의 검증 2) 문서가 가지는 영향력 검증 3) 문서의 정보를 이용한 현상의 측정이다.

1) 문서 내용의 검증이란 예를 들어 국회 및 의회 등에서의 발언 내용이 어떠한 내용을 내포하는지 확인하거나, 언론사가 특정 종교에 대해 어떠한 관점으로 바라보고 있는지 문서가 내포하는 특정 주제의 경향을 분석을 말한다. 2)영향력의 검증이란 문서에 포함된 내용이 어떠한 사회적 영향력이 있는지 분석하는 것이다. 예를 들어 NGO 등의 정치 이해관계자가 작성한 문서가 유럽의회에 입법과정에서 어떠한 영향을 주는 지 또는 입법되는 법률 내용과 지역사회의 사립도가 얼마나 연관이 있는지 상관 분석을 핵심으로 한다. 마지막으로 3)문서의 정보를 이용한 측정이란 문서를 통해 직접 관찰하지 못하는 사회현상을 계량화 하는 것이라 할 수 있다. 예를 들어 경제 기사가 사람들에게 얼마나 영향을 주고 있으며, 실제 경제와 얼마나 상관이 있는지 등의 연구라고 할 수 있다. 최근의 연구 트랜드는 잠재변수[8] (Latent Variable) 문서를 기반으로 잠재 디리클레 할당 (Latent Dirichlet Allocation:LDA)[9] 수

집된 텍스트에 포함된 단어를 추출하고 문서의 계통적 분석을 통한 추정을 하는 모델이다.

Ⅳ. 선행 연구 사례

본 연구와 관련된 선행 연구 내용은 아래와 같다.

표 1. 유형별 선행연구 사례

문서 내용의 검증	문서의 영향력 검증	문서의 정보를 이용한 현상의 측정
의회 발언에 대한 검증 (Quinn 외 2010) [10]	정치 이해관계자 생성 문서가 유럽의회에 끼치는 영향 (Klüver 2010) [12]	정책 문서를 통해 바라본 정책 우선성 연구 (Grimmer 2010) [15]
영국 언론의 이슬람교도에 보도 공평서 검증 (Baker 외 2012) [11]	미국 하원 제출 법안과 가결 법안 간의 상관성 연구 (Wilkerson 외 2015) [13]	러시아 국영 미디어 언론과 대통령 지지도의 상관 관계 (Rozenas and Stukal 2018) [16]
	미국 주립 의회의 법안 내용과 주 자립도간의 상관 연구 (Jansa 외 2018) [14]	미국의 경제언론 기사와 경제 불안감에 대한 상관 (Baker 외 2016) [17]

문서 내용을 검증한 사례를 보면, 미의회에서 진행된 의원 발언에 대한 내용을 분석하고 핵심 내용을 검증하는 연구 및 언론이 특정 종교를 바라보는 관점을 분석하는 등 토픽 분석을 주로 이용하고 있다.

문서의 영향력을 검증하는 경우, 앞에도 간단히 설명하였지만 마이닝 된 문서와 사회적 지표를 가진 설문조사결과 또는 경제지표 등을 이용, 함께 분석하는 방법이다. 그 예로는 의회에서 생성된 법안의 내용과 해당지역의 재정자립도를 비교하여 법안의 내용이 그 시의 예산 자립도에 얼마나 영향을 주는지 연구 할 수 있으며, 유사한 내용으로는 NGO 및 로비스트 등 정치 이해관계자가 만들어낸 보고서가 유럽의회 정책 수립과정에 미치는 영향 등을 분석할 수 있다.

마지막으로 문서의 정보를 이용한 현상의 측정인데, 미국의 경제 신문기사를 장기간 시계열 분석함으로서 미국의 경제 정책에 불안감과 해당 시기의 토픽과 얼마나 상관이 있는지 등이 분석 가능하며, 정부의 정책 문서를 분석함으로서 그 정부가 어떠한 정책에 우선성을 주는지, TV 및 미디어의 언론 기사가 선거에 어떠한 영향을 주는 지 등 다양한 분석이 가능하다.

Ⅴ. 양적 텍스트 분석 방법

영어권 국가에서는 Python의 NLTK 및 Gensim 또는 R의 TM, Tidytext, 및 Quanteda 등의 패키지를 주로 쓰고 있으며, 우리나라에서는 Python 또는 R 을 바탕으로 Mecab 등의 조합이 많이 사용된다. 프로그램에 능숙한 개발자의 경우 개별적인 소프트웨어를 개발하여 사용할 수 있으나 연구의 재현성 측면에서는 가장 보편적인 기술을 사용하는 것이 타당하다. 본 논문은 Rstudio를 개발 환경으로 하였으며, LSE가 제공하는 문헌 양적 분석 도구인 Quanteda를 사용하였다. 따라서 만약 본 연구를 재현하고자 하는 경우 R 과 Quanteda를 통해 재현가능하며 관련 소스 및 재료는 Github를 통해 확인 가능하다.

양적 텍스트 마이닝을 위해서 위해서는 디지털화 된 텍스트 데이터가 필요하다. 통상 온라인 상에서 다양한 형태의 Text, PDF, HTML, PPT 등 디지털 자료를 구하는 것이 가능하다. 통상 언론사의 경우 언론사 웹페이지를 통해 접근이 가능하며 학술 자료의 경우 대학의 도서관 또는 EBSCO, JSTOR, RISS 등 다양한 학술 데이터베이스에서 정보 취득이 가능하다. 이러한 과정을 통해 컨텐츠가 수집되면 다음과 같은 과정을 실시한다.

1) 데이터의 수집
2) 텍스트의 전처리
3) 문서행렬의 작성
4) 통계분석의 실시
5) 결과 해석

1. 데이터의 수집

크게는 위에 5가지 프로세스로 이루어지는데 2에서 4의 작업은 통계 프로그램을 통해 실시되며 1번의 경우도 Python의 Beautiful Soup, R의 Rcrawler 등을 이용한 크롤링을 실시 할 수 있다. 크롤링(crawling)은 웹페이지를 그대로 가져와서 거기서 데이터를 추출해 내는 행위다. 크롤링하는 소프트웨어는 크롤러(crawler)라고 부른다. 통상 수집된 텍스트의 집합을 코퍼스(corpus)라고 한다. 코퍼스를 구축함에 있어 수집된 데이터가 표본으로 적합한지 검토하는 과정을 거쳐야 한다. 가령, 인스타그램 또는 페이스북 등의 SNS 텍스트

를 수집후, 이것이 모든 연령의 의견이 반영 되었다고 판단 하거나, 진보 언론 또는 보수 언론사만 가지고 모든 민의가 반영 되었다고 보면 곤란하다. 따라서 다양한 표본을 가치판단 없이 수집하거나, 연구에 맞도록 제한 범주를 두고 데이터를 수집하는 것이 중요하다.

본 논문은 2017년에서 2019년까지의 국내 관련 이민 및 이주를 제목과 키워드로 포함하는 KCI 논문 46편과 조선일보, 중앙일보, 동아일보 및 한겨레신문의 기사 16편을 수집하였으며, 논문의 경우 초록을 데이터로 활용하였으며, 신문기사는 전문을 이용하였다.

2. 텍스트의 전처리

텍스트 데이터를 컴퓨터가 효율적으로 처리하기 위해서는 자연언어에서 문장의 최소 단위가 되는 문자 또는 문자열로 분해 하는 토큰화(tokenization)라는 작업을 해야 한다. 즉, 문단의 단어, 숫자 및 기호 요소를 분류하고 데이터를 단순화 하기위해 기호 및 조사 및 접속사 등을 불용어(Stop-words)처리를 실시한다. 본고에서 사용된 불용어는 1,744단어이며, 김호현의 텍스트마이닝을 위한 한국어 불용어 목록 연구[18] 등에서 제시한 293단어를 참고하였다.

R을 이용할 경우 앞서 수집된 데이터를 본인이 원하는 파일 형태 *.docs, *.txt, *.pdf, 및 *.xlsx로 가공이 가능하다. 본 연구에서는 날자 (date), 제목(title), 본문(text), 및 출처(sources)의 형태로 엑셀 작업하였으며, Quantada의 corpus함수를 이용 문서에서 단어를 분할하는 토큰화를 실시했다.

본 논문의 경우 공기어 분석(Co-occurrence analysis)을 실시하기 때문에 토큰을 문서 별로 집계하고 빈도가 낮은 경우 무시하는 방식으로 처리한다. 특히 이렇게 하면 통계 분석에 필요한 계산량을 줄일 수 있기 때문에 특정한 단어만 선택해서 통계 작업을 하는 특징 선택(feature selection)을 한다. 다만 특징 선택의 경우 단순화로 인한 통계의 오류가 발생될 수 있기 때문에 검증을 요한다.

3. 문서행렬의 작성

텍스트의 전처리가 완료되고 코퍼스 처리가 되면, 단어를 리스트로 전환할 수 있다. 아래의 그림에서 쉽게 알 수 있듯 문서의 단어 빈도수를 행렬로 변환하였다고 생각하면 쉽다.

표 2. 문서행렬

	한국인	체류자격을	대법원	베트남	출신	여성
text1	4	0	1	3	0	2
text2	3	0	0	3	2	0
text3	8	6	2	1	0	0
text4	0	0	0	0	0	0
text5	1	0	0	2	0	0
text6	0	0	0	0	0	0
text7	1	0	0	5	0	3
text8	0	0	0	0	0	0
text9	0	0	0	0	0	0

텍스트 문서의 통계 분석을 위해서는 문장의 특징화할 수 있는 문서행렬 (Matrix)의 작성이 필요하다. 여기에는 Quanteda를 이용함으로 Document-feature matrix (DFM)방식의 문서 행렬을 작성해야 한다. 여기서의 행은 document 즉 문서가 들어가며, 열은 문서의 단어의 특징인 feature가 들어간다. 통상의 경우 document-term matrix로 처리되는 방식으로 단어, 단어의 집합, N-gram 및 품사 등 다양한 성격을 가질 수 있다.

4. 통계분석의 실시

4.1 빈도분석

단어의 빈도를 문서 별로 비교하여 전체 문서에서 어떠한 빈도로 단어가 노출 되어있는지 보여준다. 사실 여기까지는 비구조형 데이터 분석이라고 할 수 있으며, 분석을 위한 전초 단계라고 할 수 있다.

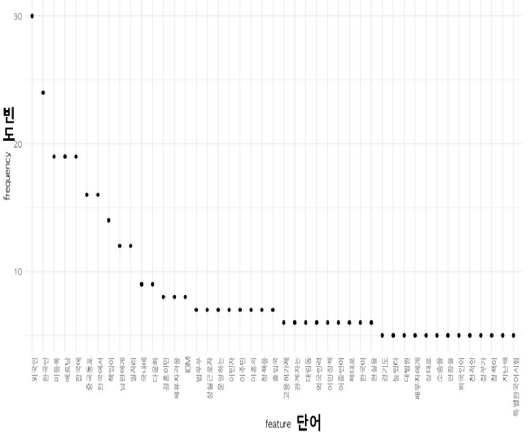

그림 1. 본 논문 자료의 빈도 분석

4.2 워드 클라우드

문서의 출현 빈도가 높은 단어를 복수 선택하여, 빈도에 맞춰 글의 크기를 바꿔가며, 글을 도식화는 기법이다. 통계적인 유의미성보다는 단어를 한눈에 볼 수 있다는 장점이 있어 손쉽게 사용된다.

그림 2. 워드 클라우드로 본 논문의 자료를 분석함

4.3 상대빈도분석

단어의 빈도를 문서 별로 비교하여, 통계적으로 상관을 보여준다. 본 사례는 언론사가 바라보는 2017-19년도 국내 외국인 이주민 또는 이민에 대한 언론 기사로 통계적인 수치를 볼 때 지속적인 관점의 연관성은 없는 것으로 나타나다.

표 3.상대빈도분석표

Feature	chi2	p	n target	n reference
대법원	9.0252293	0.0026628	2	3
남편에게	2.3711718	0.1235942	2	10
베트남	0.8113892	0.3677099	2	17
상대로	0.4855192	0.4859329	1	4
소송을	0.4855192	0.4859329	1	4
책임이	0.3861731	0.5343179	1	13
한국인	0.3516744	0.5531668	2	22
출입국	0.1896820	0.6631814	1	6
이혼의	0.1896820	0.6631814	1	6
체류자격을	0.1124884	0.7373287	1	7
결혼이민	0.1124884	0.7373287	1	7
한국에서	-0.0305679	0.8612079	0	16
중국동포	-0.0305679	0.8612079	0	16
한국에	-0.0925584	0.7609496	0	19
미등록	-0.0925584	0.7609496	0	19
외국인이	-0.2087693	0.6477337	0	5
정책이	-0.2087693	0.6477337	0	5

4.4 토픽분석

모여진 문서에서 추상적인 무언가를 발견하기 위한 통계 모델로, 문서에서 숨겨진 의미구조를 찾기 위해 사용되는 마이닝 기법 중 하나로 Quanteda 패키지와 topicmodels 패키지[19]를 혼합하여 사용하는 것이 가능하다.

표 4. 본 논문 자료의 상대빈도분석표

Topic									
1	2	3	4	5	6	7	8	9	10
일자리	다문화	남편에게	외국인	미등록	베트남	성실근로자	한국인	외국인	중국동포
IOM	이중언어	책임이	한국에서	한국에	한국인	한국에	책임이	이민정책	한국에서
이주민	한국인	한국인	정책이	출입국	한국에	국내에	체류자격	외국인력	대림동
이민자	이주민	베트남	이주민	정부가	결혼이민	고용허가제	결혼이민	이민자	운영하는
정책을	대림동	이혼의	미등록	한국에서	남편에게	특별한국어시험	배우자에게	정책을	외국인
정부가	늘었다	소송을	한국인	법무부	상대로	외국인	이혼의	법무부	늘었다
미등록	관계자는	대법원	현실을	제대로	소송을	경기도	외국인	외국인이	외국인
한국에서	제대로	국내	외국인이	현실을	운영하는	관계자는	대법원	일자리	출입국
외국인	정책을	상대	일자리	한국인	현실을	베트남	남편에게	다문화	법무부
법무부	운영하는	연장	베트남	IOM	외국인이	법무부	출입국	늘었다	고용

4.5 공기어 분석

공기어 분석이란 출현 패턴이 비슷한 단어 특히 동시에 발생되는 비슷한 단어를 네트워크로 연결하고 단어와 단어 간의 상관을 보는 방법이다. 특히 단어와 단어가 선으로 연결 되어 있음으로 다차원 척도 구성법보다 이해가 쉽다는 장점이 있다[20]. 본 논문에서는 공기어 분석을 통해 도출된 자료를 해석하고자 한다.

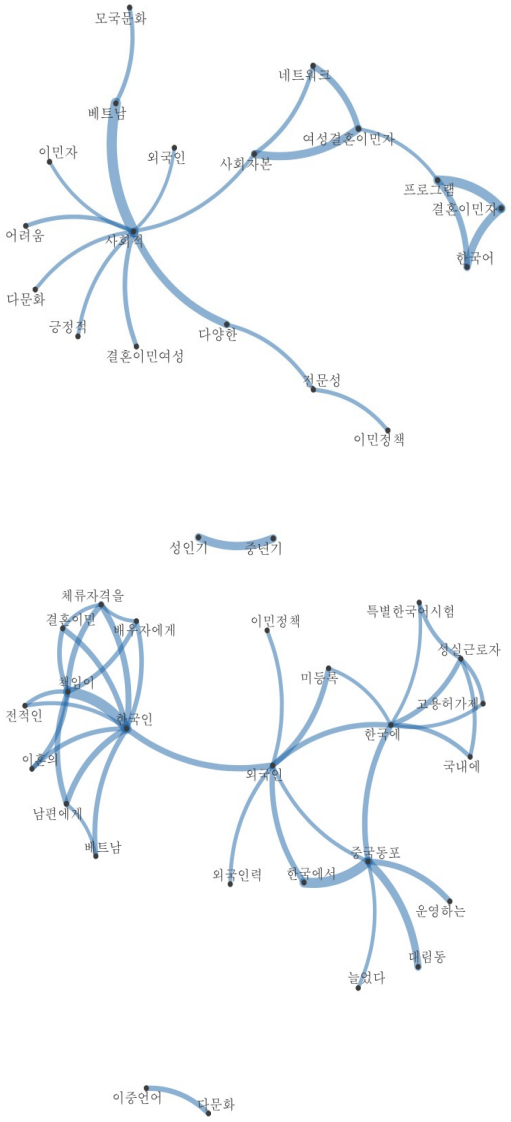

그림 3. 2017-2019년 "이민 및 이주에 관한" KCI 논문 (위) 및 주요 언론사(아래)의 공기어 네트워크

결과를 해석할 때, 텍스트 데이터의 다차원성을 항상 인식해야 한다. 처음부터 단순하게 공기어 분석만을 실시할 경우, 높은 상관관계를 가진 네트워크가 구성될 가능성이 있다. 따라서 우연에 기반한 결과를 견제해야 한다. 특히 사회 과학적 문서의 경우 문서 자체에 대한 사회적 현상의 이해가 없이는 생성된 결과에 대한 이해가 어려울 수도 있다. 따라서 연구에 앞서 충분한 선행 연구의 필요성을 요한다.

4.6 분석 대상

본 논문에서 사례 적용을 위해 선정한 키워드는 이주와 이민이다. 인구 절벽에 직면한 우리나라가 이주와 이민에 대해서 어떻게 바라보고 있는지 학계와 언론의 연구 자료 및 기사문을 분석하였다. 이주와 이민의 경우 다양한 관점에서 분석이 시도 되었고, 연구 범주의 폭 또한 광범위하기 때문에 본 논문에서 시행하고자 하는 연구 방법론을 적용할 수 있는 최적의 키워드 중 하나라고 할 수 있다. 본 연구 분석을 통하여 학계에서 출판된 논문과 신문에서 언급한 기사에서 우리나라가 이주와 이민에 대해 바라보는 관점은 무엇이며 주요 키워드와 핵심은 무엇인지 다양한 방식을 사용해 보여주고자 하였다. 본 사례는 연구 방법을 보여주는 것으로 다른 사회과학적 연구에 대해서도 본 방법을 적용할 경우 다양한 양적 분석 진행이 가능 할 것으로 판단된다.

4.7 결과 해석

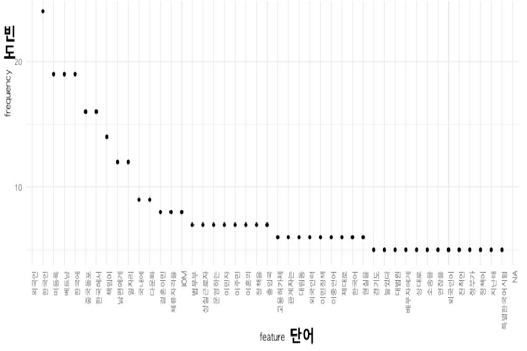

그림 4. 주요 언론사 '이주, 이민' 키워드 빈도분석

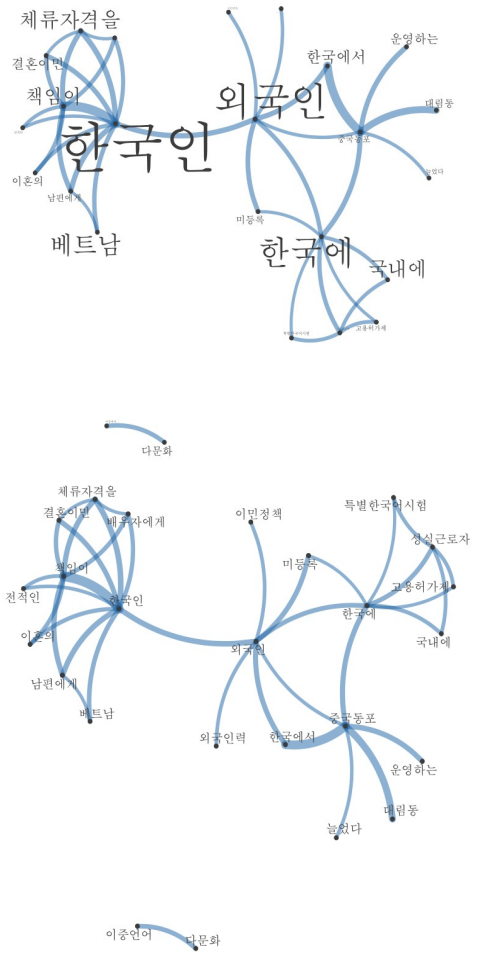

그림 5. 2017-2019년 주요언론사 "이민 및 이주에 관한 기사" 공기어 네트워크

2017년에서 2019년 동안 주요 언론사에서 '이주'와 '이민'을 주제로 집필된 기사는 총 17개였다. 이들 기사를 공기어 네트워크로 분석한 결과 위의 [그림 5]와 같은 결과가 나왔다.

결과를 보면 한국으로 이주한 이민자에 대한 기사문을 분석했음에도 불구하고 한국인이라는 키워드가 중심적으로 많이 사용되었음을 확인할 수 있다. 이는 외국인에 대한 기술이 주를 이루었음에도 불구하고 한국인 및 한국에 미친 영향에 초점이 맞춰져 있었을 것이라는 해석 및 추측이 가능하다.

나아가 결혼이민, 이혼, 남편, 체류자격, 베트남 등과 같은 키워드가 상관관계가 있는 것으로 나타났다. 이를 토대로 국내 이주, 이민은 결혼이주여성과 관계된 내용으로 많이 다루어지고 있다는 것을 알 수 있다. 아울러 중국동포, 고용허가제 등의 단어가 상관관계를 나타낸 것으로 나타났다. 이는 외국인 노동자에 대해 다루고 있는 기사가 주된 기사였다는 사실을 나타내고 있다.

본 논문은 공기어 네트워크를 통한 연구 방법론을 다루고 있기 때문에 분석을 통해 나타난 상관관계를 가치 판단 없이 표면적으로 분석하는 것을 목표로 한다. 본 논문을 토대로 분석한 내용에 인류학적인 접근이나, 인문학적인 접근법을 활용한 분석을 덧붙인다면 '이주' 및 '이민'이라는 사회 현상이 갖는 다양한 의미를 분석하는 것이 가능 할 것이라 사료된다.

본 논문은 KCI에 등재되어 있는 논문 중에서 '이주' 및 '이민'을 제목에 포함하거나 키워드로 하는 논문의 초록을 정리, 분석하였다. 아래 [그림 6]은 해당 자료의 단어 빈도분석을 나타낸 것이다.

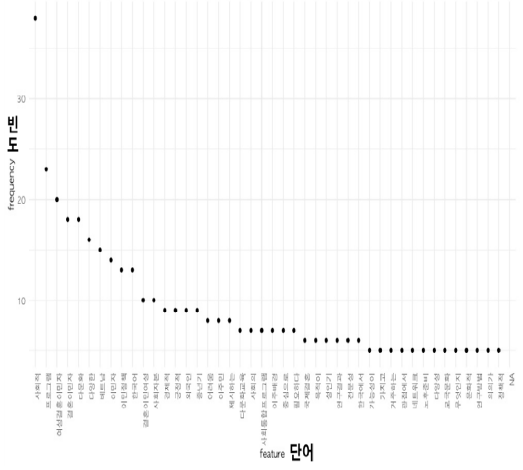

그림 6. 2017-2019년 KCI "이민 및 이주에 관한 논문 초록" 주요 단어 및 빈도

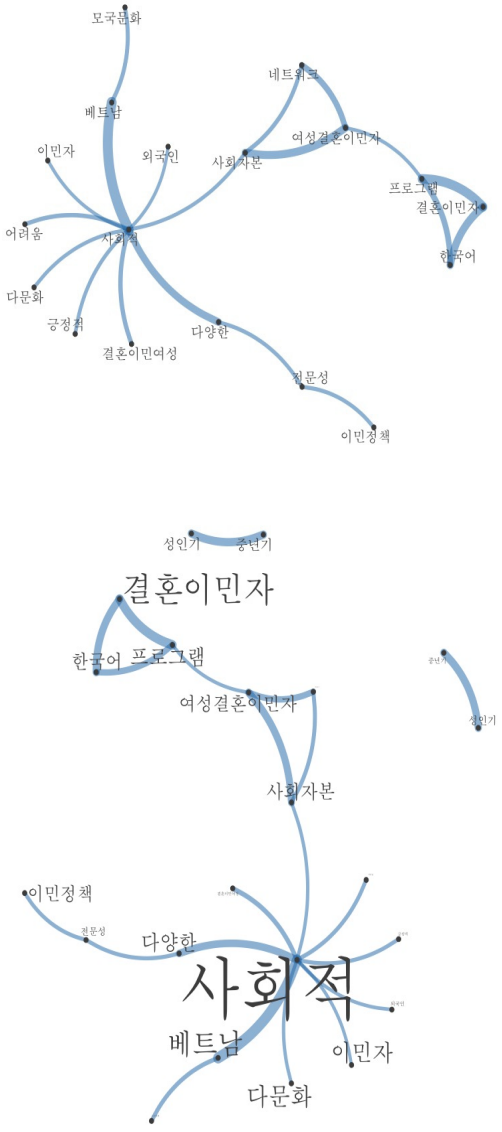

그림 7. 2017-2019년 KCI "이민 및 이주에 관한 논문 초록" 공기어 네트워크

2017년에서 2019년 사이에 작성된 논문은 '사회적'이라는 단어가 중심 키워드로 작용하였다. 이를 바탕으로 다문화, 이민자, 베트남 등의 단어가 상관관계를 보였으며 이민정책, 사회 자본 등도 상관관계가 있는 것으로 나타났다. 이를 바탕으로 추정해 볼 때, 국내에서 연구된 '이주' 및 '이민' 관련 논문은 사회에 어떤 영향을 미치고 있는지에 대해 논하고 있다고 해석할 수 있을 것이다. 아울러 결혼 이민자가 중요한 상관관계를 갖고 있는 것으로 나타났다. 즉, 국내 연구는 결혼 이주자를 중심으로 연구가 진행 된 것으로 보인다.

Ⅵ. 결 론

본 연구는 어떻게 하면 분석 샘플로 적용된 연구논문 및 언론 기사에 대한 객관성을 유지 할지를 주안점을 두었다. 연구 논문 및 언론 기사 모두 글에는 의도가 포함 될 수 밖에 없다. 이를 다시 분석하는 과정은 텍스트라는 비구조형 데이터를 다차원 데이터로 전환함으로서, 문헌 연구 자료인 논문과 신문기사가 어떤 논조를 보이고, 어떤 영역에 집중하였는지 밝히고자 하였다. 이를 위해 단순한 단어 빈도 분석 뿐 만 아니라, 공기어 네트워크 분석을 통해 상관 분석을 시도하였다. 본 연구의 가장 큰 특징은 형태소를 이용한 기존의 방식을 탈피하여 사전의 필요성을 피했다는 점이다. 또한 본 논문에서 시도하지 않았으나, 감정분석을 위한 감정사전을 추가 할 경우 문헌자료의 중립성 등을 추가로 분석 할 수 있다. 이번 연구 논문의 분석을 통해 얻은 결과로, 동일 기간의 동일 주제임에도 불구하고, 학계와 언론사 간의 관점의 차이를 명확히 보여줬다는 점에서 의미가 있다. 또한 이러한 방식을 통해 전체적인 경향을 비교적 쉽게 도식화 가능하기에 객관적 데이터가 없는 사회과학 연구 분야에서도 다차원 데이터를 이용한 연구 방식으로 연구 확장이 가능하다는 점에서 본 논문이 설명하고 있는 양적 텍스트 분석이 방법이 다른 사회과학의 연구에 기여했으면 기대하는 바이다.

참 고 문 헌

[1] Ithiel de Sola Pool, *Trends in Content Analysis*, University of Illinois Press, 1959.

[2] Kulkarni, Parag, Sarang Joshi, and Meta S. Brown, Big data analytics, PHI Learning Pvt. Ltd., 2016.

[3] W. H. Inmon, Daniel Linst, and Mary Levins,

Data Architecture: A Primer for the Data Scientist, London: Academic Press, 2019.
[4] A. Frigessi, P. Bühlmann, I. Glad, M. Langaas, S. Richardson, and M. E. Vannucci, "Statistical Analysis for High-Dimensional Data," Springer, 2016.
[5] Team R Core, "R: A language and environment for statistical computing," 2013, http://www.R-project.org
[6] K. Benoit, K. Watanabe, H. Wang, P. Nulty, A. Obeng, S. Müller, and A. Matsuo, "Quanteda: An R Package for the Quantitative Analysis of Textual Data," Journal of Open Source Software, Vol.3, No.30, p.774, 2018.
[7] Taku Kudo, "MeCab," Source Forge: http://sourceforge.net/projects/mecab, 2008.
[8] Borsboom, Denny, Gideon J. Mellenbergh, and Jaap Van Heerden, "The theoretical status of latent variables," Psychological review, Vol.110, No.2, p.203, 2003
[9] A. Frigessi, P. Bühlmann, I. Glad, M. Langaas, S. Richardson, and M. E. Vannucci, "Statistical Analysis for High-Dimensional Data," Springer, 2016
[10] K. M. Quinn, B. L. Monroe, M. Colaresi, H. M. Crespin, and D. R. Radev, "How to analyze political attention with minimal assumptions and costs," American Journal of Political Science, Vol. 54, No.1, pp.209-228, 2010.
[11] Baker, Paul, Costas Gabrielatos, and Tony McEnery, "Sketching Muslims: A corpus driven analysis of representations around the word 'Muslim'in the British press 1998-2009," Applied linguistics, Vol.34, No.3, pp.255-278, 2013.
[12] H. Klüver, "Europeanization of lobbying activities: When national interest groups spill over to the European level," European Integration, Vol.32, No.2, pp.175-191, 2010.
[13] Wilkerson, John, David Smith, and Nicholas Stramp, "Trac-ing the Flow of Policy Ideas in Legislatures: A Text ReuseApproach," American Journal of Political Science, Vol.59, No.4, pp.943-956, 2015.
[14] Jansa, Joshua M., Eric R. Hansen, and Virginia H. Gray, "Copy and Paste Lawmaking: LegislativeProfessionalism and Policy Reinvention in the States," forthcoming, American Politics Research, published onlineMay, 31, 2018.
[15] J. Grimmer, "A Bayesian Hierarchical Topic Modelfor Political Texts: Measuring Expressed Agendas in Sen-ate Press Releases," Political Analysis, Vol.18, No.1, pp.1-35, 2010.
[16] Rozenas, Arturas and Denis Stukal, "How Autocrats Manipulate Economic News: Evidence from Russia'sState-Controlled Television," forthcoming, Journal of Politics, Vol.81, No.3, pp.982-996, 2018.
[17] S. R. Baker, "Measuring Eco-nomic Policy Uncertainty," The Quarterly Journal ofEconomics, Vol.131, No.4, pp.1593-1636, 2016.
[18] 길호현, "텍스트마이닝을 위한 한국어 불용어 목록 연구," 우리말글, Vol.78, pp.1-25, 2018.
[19] B. Grun and K. Hornik, "topicmodels: an R package for fitting topic models," Journal of Statistical Software, Vol.40, No.13, pp.1-30, 2011.
[20] Higuchi Koichi, *社会調査のための計量テキスト分析*, ナカニシヤ出版, 2014.

저 자 소 개

이 수 정(Soo-Jeong Yi) 정회원

- 2012년 2월 : 한국외국어대학교 중동아프리카학과(문화인류학 석사)
- 2015년 8월 : 한국외국어대학교 국제관계학과(중동아프리카학 박사)
- 2019년 9월 ~ 현재 : 한국외국어대학교 아랍어통번역학과 강사

〈관심분야〉 : 국내 이주 무슬림, 이슬람, 모스크, 중동지역 연구, 양적 데이터 마이닝

최 두 영(Doo-Young Choi) **정회원**

- 2012년 2월 : 한국외국어대학교 중동아프리카학과(정치학 석사)
- 2019년 10월 : 한국외국어대학교 중동아프리카학과(경제학 박사과정)
- 2019년 3월 ~ 현재 : ㈜ 더워크 코리아 개발이사

〈관심분야〉 : 자연어처리, 양적 데이터 마이닝, 중동아프리카 경제, 언론분석