중남미연구 제42권 1호
2023년 2월 28일 95~130쪽
DOI http://dx.doi.org/10.17855/jlas.2023.2.42.1.95

한국의 라틴아메리카 연구동향 분석(방법)의 제안: 토픽모델링, 연관분석 그리고 워드클라우드 분석을 중심으로*

이 태 혁 · 강 지 훈
(부산외국어대학교)

* 이 논문은 2018년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2018S1A3A02081030).
이 논문은 2022년 12월 17일에 개최한 한국라틴아메리카 학회 동계 학술대회의 발표 내용을 수정·보완한 것입니다. 졸고에 대해 학회 발표 시 토론을 해주신 선생님과 이 논문을 심사하신 익명의 심사위원분들에게 감사의 말씀을 전합니다.

본 연구는 방법론적 다양성을 통해 한국의 라틴아메리카 연구 동향을 분석하는 것이 목적이다. 다시 말해 본 연구 주제와 관련한 선도적 연구를 바탕으로 본고는 기계학습(machine learning) 기반의 데이터 분석을 실시함으로써 기존의 정성적 또는 기초 통계의 방법론적 차별성을 견지한다. 이를 위해 첫째, 본고는 국내 논문 제공 사이트 한국학술정보(RISS)를 통해 데이터를 수집한다. 본 연구의 연구 대상과 범위는 기 작성된 연구물의 제목에서는 '중남미' 그리고 초록의 키워드는 '라틴아메리카'로 설정하며 R 소프트웨어 프로그램에서의 크롤링(crawling) 정의에 따라 수집한다. 그리고 둘째, 크롤링 정의로 추출된 자료를 본고의 연구 방법인 토픽모델링, 연관분석 그리고 워드클라우드 분석을 위해 데이터 전처리를 진행하며 이후 각 분석에 맞는 알고리즘으로 데이터를 처리한다. 셋째, 본 연구의 분석단위가 세대(시대)별 구분이다. 따라서 이렇게 분류된 시대별 데이터는 상기 3가지 빅데이터 분석을 통해 산출된 결과를 바탕으로 한국의 라틴아메리카 연구 동향을 고찰한다. 지난 68년 이래 현재까지 꾸준히 진행된 한국의 라틴아메리카 주된 연구 주제는 경제(무역)이지만 세대(시대)에 따라 라틴아메리카 상황 또는 글로벌 현상을 반영하며 다양한 연구가 진행되었음을 확인할 수 있다.

주제어

라틴아메리카, 한국, 토픽모델링, 연관분석, 워드클라우드

I. 들어가며

21세기 셋째 순(旬)에 접어든 현재, 한국과 라틴아메리카는 지난 60년간의 외교관계를 통한 역사를 공유했다. 특히 2022년을 기준으로 한국은 멕시코, 콜롬비아, 칠레 등 중남미 15개 국가와 외교관계를 기념하며 각종 기념행사를 통해 '회갑'의 우의를 다졌다. 이처럼 한국과 라틴아메리카 60년간의 동행으로 이순(耳順) 즉 "서로 듣기만 해도 이해할 수 있을 정도"로 양측이 긴밀한 파트너십을 구가하고 있다(최종욱, 2022, p. 23). 이와 같이 양측의 협력과 교류의 역사 가운데 상대방에 대한 이해정도는 어떻게 진전 또는 변화 되었을까? 다시 말해, 양측의 이해도를 증진하기 위해서 한국은 라틴아메리카에 대해 그리고 동시에 라틴아메리카는 한국에 대한 어떠한 연구가 얼마만큼 개진되었을까? 본 연구의 범위(scope)와 수준(level)에 따라 본고는 한국의 對라틴아메리카 연구 동향을 중심으로 살펴보고자 한다.

상기 질문에 대한 기존의 연구물에 따르면, 외부적 환경 변화 그리고 국내 라틴아메리카 연구자들의 수적 증대와 함께 연구주제의 확장성에 따라 시기별로 구분할 수 있다. 이러한 맥락에서 이성형(2009)의 분류에 따르면, 제1세대는 해방~1980년대, 제2세대는 1990년대 그리고 제3세대는 2000년대 시기로 구분하며 각 세대별 라틴아메리카 연구 동향을 파악할 수 있다. 본고는 이와 같은 세대별로 구분된 분류에 비판적 접근을 견지하며, 실질적으로 각 분류 기간 동안 연구의 경향을 빅데이터 분석을 통해 살펴보는 것을 목적으로 한다. 이를 위해 본 연구는 중남미 또는 라틴아메리카라는 단어를 제목 또는 초록의 키워드에 작성한 문헌을 대상으로 텍스트 마이닝 방법 가운데 토픽모델링, 연관분석 그리고 워드클라우드 분석을 활용하여 국내 라틴아메리카 연구 동향을 파악한다.

본고는 다음과 같은 순서로 작성한다. 첫째, 라틴아메리카 지역연구 동향은 기존의 문헌을 통해서 확인해 본다. 그리고 둘째, 상기 주제에 대한 기존의 선도적 연구물들의 연장선상에서 본고는 다양한 빅데이터 분석을 통해 연구방법론적 확장을 시도한다. 즉, 지난 60여년 이상의 국내 라틴아메리카 연구물은 기 제시한 중남미 또는 라틴아메리카라는 단어를 포함한 자료만

천 여건이 넘는다. 따라서 양산된 연구물이 방대한 규모인 만큼 빅데이터 분석 기법을 활용하여 국내 라틴아메리카 연구동향을 파악하고자 한다. 특히, 텍스트 마이닝 방법을 통해 연구대상, 자료수집, 자료 분석 등을 진행하도록 한다. 셋째, 본고는 본 연구물을 통해 확인된 결과물에 대한 함의를 재고찰함과 동시에 연구의 한계점을 명시한다. 이를 바탕으로 후속연구 과제를 제시하며 마무리 하도록 한다.

II. 선행연구 고찰

1. 국내 라틴아메리카 연구 동향에 대한 연구물

국내 라틴아메리카 연구 동향에 대한 선행연구는 연구물 기준 시대별(세대별), 분과학문별 그리고 연구 분석 접근 방법으로 크게 3가지로 분류할 수 있다. 연구물 발간 기준은 김달용(1989)의 『라틴아메리카 지역연구에 관한 일고』 제하의 논문이 선구적이다. 본고는 비단 한국의 라틴아메리카 연구 동향뿐만 아니라 미국, 일본, 소련, 동독, 중국, 인도 등 주요 국가의 대 라틴아메리카 연구동향을 고찰하며 우리나라의 70-80년대 라틴아메리카 연구 현상을 분석하며 방향성을 제시했다. 2000년대에 들어서며 라틴아메리카 동향에 관한 일련의 연구물이 등장한다. 곽재성(2002)은 1990년대 김영삼 정부의 국제화의 정책적 드라이브로 해외지역연구가 활성화되는 전기를 마련하게 되었다고 피력했다. 즉 1990년대는 이전 시대와 다른 라틴아메리카지역 등을 위시로 하는 지역연구에 대한 새로운 '토양'이 제공된 것이다. 그는 라틴아메리카를 대상으로 한 학술연구 경향에 관한 정량적 분석을 통해 연구의 특징은 무엇이며, 학문적 심화와 지역연구의 활성화를 위해 학계가 풀어야할 과제는 무엇인지에 관해 논의를 개진했다.

이러한 관점을 견지한 최윤국(2003)은 국내 라틴아메리카 관련 간행물, 단행본, 학위논문 총 4162건(건국 이래 2002.12까지)을 유형 및 주제별로 분류

하며 분석하며 지역전문가 양성을 위한 학계와 정부의 협업을 제안하기도 했다. 이와 같은 연구물별 기준을 이성형(2009)은 한국의 라틴아메리카 동향에 대해 시대별로 구분했다. 그는 1980년대 이전, 1990년대 그리고 2000년대로 시기별로 각각 구분하며 시대별 연구동향 및 특성을 분석했다. 특히 1990년대는 정부의 국제정책과 맞물려 당시 신진학자의 등장으로 이전 시대에 비해 산출된 연구물이 양적 및 질적으로 확장되었다고 피력했다.

한편, 국내 라틴아메리카 동향을 연구한 홍욱헌(2012)은 1971-2012년 간 발표된 연구물(논문 및 정부기관 발행물)을 DBpia 논문 검색사이트를 활용하여 조사했다. 그는 라틴아메리카 지역 및 국별 제목과 주제어를 바탕으로 분류하며 한국의 라틴아메리카 연구의 양적성장 정도를 통계수치로 제시했다. 차경미(2011)는 2000-2010년을 분석기간으로 하여 정량적 분석하며 이 기간 동안 라틴아메리카 지역 연구 동향을 고찰했다. 그녀는 분석기간 동안의 라틴아메리카 지역연구는 연구사례가 확장되었지만 지역 내의 일부 국가를 제외한 나머지 국가들에 대한 연구는 희박하거나 부재하다고 분석했다.

앞서 살펴본 이성형(2009)은 한국의 라틴아메리카 정치학 분과를 중심으로 연구동향을 파악했다면 최권준(2021)은 한국의 중남미문학 연구 동향을 연구한 바 있다. 그는 우리나라 최초의 중남미 관련 전문학술지인 한국외대 중남미 연구소의 『중남미연구』가 발행된 1976년부터 2020년까지 45년을 기준으로 국내 6개 전문학술지별, 연도별, 연구자별, 작가별로 정리하며 비교분석했다. 국내 라틴아메리카 연구 동향의 기존의 정성적 또는 기초 통계 분석에 대비해 임두빈(2020), 정호윤(2021) 그리고 이태혁(2022)은 연구대상을 라틴아메리카로 하여 각각 주제에 따라 빅데이터 분석을 실시했다. 임두빈은 국내 포르투갈어권 연구 동향을 2004년부터 2020년까지 모두 126편의 논문 초록을 대상으로 텍스트 마이닝의 의미 네트워크(semantic network)분석 등을 통해 논문 간 존재하는 연결성과 중심성에 주목하며 연구동향을 파악했다. 정호윤(2021)은 국내 포털뉴스(2018-2019)에서 보도된 라틴아메리카 국가 관련 텍스트를 빅데이터 분석방법을 활용하여 연구했다. 이태혁(2022)은 R소프트웨어 프로그램을 활용하여 빅데이터 분석 가운데 워드 클라우드 기법을 활용하며 중남미지역의 한류 패턴(2011-2022)을 제시했다. 한국의 라틴아메

리카 대상 또는 주제 연구에 대한 연구방법론적 접근은 기 살펴본 바와 같이 정성적 또는 기초 통계수준에서 접근하고 있다. 또한 최근 일부 연구가 빅데이터 분석을 위해 Textom 또는 빅카인즈(BIGKinds) 등 뉴스 분석서비스 제공업체를 통해 진행되고 있다. 이와 같은 방법론적 다양한 접근 가운데 본고는 R 소프트웨어 프로그램을 통해 크롤링(crawling)기법으로 웹사이트 기반 라틴아메리카 연구를 전수 조사했다. 이는 기존의 라틴아메리카 지역 연구 동향에 대한 방법론적 차별성으로 피력될 수 있다. 아래 절에서는 기존의 빅데이터 분석 동향을 바탕으로 본고가 연구방법으로 진행하고자 하는 토픽모델링, 연관분석 그리고 워드클라우드 분석 방법 등의 연구방법 동향을 고찰한다.

2. 방법론적 선행연구

최근 기계학습(machine learning) 기반의 데이터 분석 연구가 활발히 수행되고 있다. 특히 학술분야에서 특정 주제의 연구동향 파악을 위해 관련 논문을 수집하여 분석하고 의미 있는 패턴을 생성하는 일련의 과정이 진행된다. 토픽모델링, 연관분석, 워드클라우드는 연구동향 분석을 위해 대표적으로 사용되는 분석 기법이다(이유빈 외 2020; 박진희 외 2021; 신서영 · 이범준 2021). 기존의 연구동향 분석 연구는 단일 기법 혹은 두 가지 기법을 병행하는 경우는 있으나 세 가지 기법을 모두 병행하며 적용한 사례는 드물다. 본 연구에서는 한국의 라틴아메리카 연구동향을 보다 계량적 그리고 총체적으로 분석함으로 신뢰도를 높이기 위해 위 세 가지 분석 기법을 활용하고자 한다.

2.1 토픽모델링(Topic Modeling)

토픽모델링은 주어진 문서에서 단어의 분포를 통계적으로 분석하는 확률 토픽 모델 기법으로 데이터로 활용된 문서에 대한 잠재적 주제를 자동으로 추출하는 방법이다. 토픽모델링은 문서 전체를 대상으로 하므로 연구동향 분

석, 프레임 분석 등에 사용된다(강지훈 · 조치영, 2022). 잠재 디리클레 할당(Latent dirichlet allocation, LDA)은 토픽모델링을 위한 대표적인 알고리즘으로 주어진 문서와 문장, 그리고 단어들을 대상으로 베이지안(bayesian) 확률모형 기반의 정량분석을 통해 키워드들을 군집화 한다(이유빈, 2020). 그리고 이와 같이 군집화 된 키워드 집합을 통해 대량의 문서집합에서 잠재된 주제를 추출할 수 있다. 이처럼 LDA알고리즘은 텍스트에 내포되어 있는 주제를 도출하며 동시에 질적 수준에 준하는 의미와 해석을 준행하게끔 하는 장점이 있음에 학계의 다양한 분과학문에서 본 연구방법론이 진행되고 있다(Jacobs & Tschötschel, 2019).

예를 들어, 강지훈 · 조치영(2022)은 텍스트 마이닝 기반 국내 예루살렘 연구동향 분석 연구에서 토픽모델링과 연관분석을 통해 국내에서 수행된 예루살렘 지역 연구동향을 분석하고 특정 주제에 편중된 연구 주제의 다각적인 확장을 위한 연구 분야를 제시하였다(강지훈 · 조치영, 2022). 장성희 · 이창원(2022)은 토픽모델링을 이용한 기독교 연구 동향 분석 연구를 통해 기독교 연구 분야의 토픽을 11개로 정의하고 향후 기독교 연구에 대한 통찰력과 발전을 도모하였다(장성희 · 이창원, 2022). 노설현(2020)은 토픽모델링을 활용한 인공지능 관련 이슈 분석 연구를 통해 인공지능과 관련된 기사들을 LDA 알고리즘 기반의 토픽모델링 기법으로 분석하여 인공지능 관련 주요 이슈를 분석함으로써 인공지능 기술을 다양한 분야와 융합하기 위한 의미 있는 정보를 생산하고자 하였다.

2.2 연관분석(Association analysis)

연관분석은 텍스트 마이닝 기반의 비지도 기계학습 기법으로 대량의 키워드 집합 혹은 단어들 간 상호연관성을 추출한다[1]. 특히 단어들 간 동시출현

1) 머신러닝에는 모두 4가지 학습이 있음. 첫째, 지도형 학습은 예시를 통해 머신을 훈련함. 둘째 비지도 학습은 머신이 입력 데이터를 학습한 다음 관련성 있고 액세스 가능한 데이터를 모두 사용해 패턴과 상관관계를 인식함. 셋째, 준지도 학습은 대량의 원시 비정형 데이터를 처리하는 경우 활용함. 넷째, 강화 학습은 지도형 학습처럼 머신에 정답 키를 제공해 주는 대신 일련의 허용 가능한 행동, 규칙, 잠재적 최종 상태가 입력됨. 이를 바탕으로 경험과 보상을 통해 학습함

빈도를 분석하여 단어들 간의 연관 관계를 추출하고 이를 통해 특정 문서에 대한 주제 혹은 문서 내용의 흐름을 파악하는데 활용될 수 있다(김한준 · 장재영, 2011). Apriori 알고리즘은 연관분석을 위한 대표적인 알고리즘으로 동시출현 단어들 간의 지지도(support), 신뢰도(confidence), 향상도(lift)를 기반으로 연관규칙을 생성한다. 이와 같은 연관분석의 방법론적 차별성은 경영학에서는 장바구니 분석(Market Basket Analysis)으로 알려지는 등 다제학에서 본 연구 방법론이 활용되고 있다 (류기동 외 2016).

이러한 맥락에서 최출헌 · 장필식(2019)은 글로벌 디자인 연구동향에 대한 키워드 네트워크 분석 연구(1999～2018)에서 특정 기간 내에 발간된 디자인 관련 연구 논문들을 대상으로 키워드 네트워크 분석을 시행함으로써, 최근 20년간의 디자인분야 글로벌 연구의 특성과 동향을 파악하였다. 김은경 · 조대연(2022)은 「토픽 모델링과 키워드 네트워크 분석을 활용한 국내 성인 대상 교육훈련 요구분석 연구동향: 2012-2021년 학술지 게재 논문을 중심으로」 제하의 연구에서 국내 성인 대상 교육훈련 요구분석 연구의 동향을 파악하기 위해 키워드 네트워크 분석을 실시하였다. 이를 통해 관련 연구 동향에 대한 핵심 단어를 추출함으로써 연구의 발전적 방향 모색과 이에 따른 시사점을 제시하였다.

2.3 워드클라우드 분석(Word cloud analysis)

워드클라우드는 문서에서 단어의 출현 빈도를 기반으로 핵심 단어 혹은 키워드를 직관적으로 파악할 수 있도록 시각화하는 분석기법이다(Heimerl 외 2014). 대량의 문서에서 출현빈도가 높은 단어나 키워드를 직관적으로 표현함으로써 해당 문서의 내용이나 주제의 특징을 유추할 수 있다. 특히, 결과의 표현이 단어(키워드)가 출현한 빈도가 높으면 크게, 상대적으로 낮으면 작게 나타내는 특징이 있음에 관련 내용을 직관적으로 알 수 있게 안내 해 신속하지만 비교적 정확한 정보를 제공해 준다(노형남, 2014). 따라서, 트랜드 변화를 주요 연구 대상으로 삼고 있는 광고, 경영 그리고 교육 분야 등 까지 다양

(www.sap.com).

한 학문 분과에서 본 분석법을 활용한 연구물을 확인할 수 있다.

박애스더 · 김정민(2021)은 워드클라우드 기법과 KJ법을 통한 교양교육 연구 동향 연구에서 최근 대학의 교양교육 동향을 파악하기 위해 정량적인 워드클라우드 기법과 KJ법을 활용한 정성적 분석을 시행하고 향후 교양교육 연구를 위한 현안 및 시사점을 제공하고자 하였다(박애스더 · 김정민, 2021). 김준환 · 문형진 · 이항(2021)은 워드클라우드 기법을 이용한 국내 융복합 학술연구 트랜드 분석 연구에서 워드클라우드 기법을 적용하여 최근 10년 간 융복합 연구동향을 파악하고 분석된 결과를 융복합 분야의 추후 연구를 위한 기초자료로 활용하고자 하였다(김준환 · 문형진 · 이항, 2021). 김민경(2020)은 글로벌 공유재로 활용되고 있는 에어비앤비 숙소의 트랜드와 인기 정도를 워드클라우드 기법을 활용하여 분석 및 결과를 제시했다(김민경, 2020). 이태혁(2021)은 한류 발전관련 선행 연구를 기반으로 라틴아메리카 지역의 한류의 경향 정도를 파악했다. 이를 위해 그는 K-콘텐츠 관련기사를 활용하며 워드클라우드 기법으로 연구 결과의 의미를 시각화 했다.

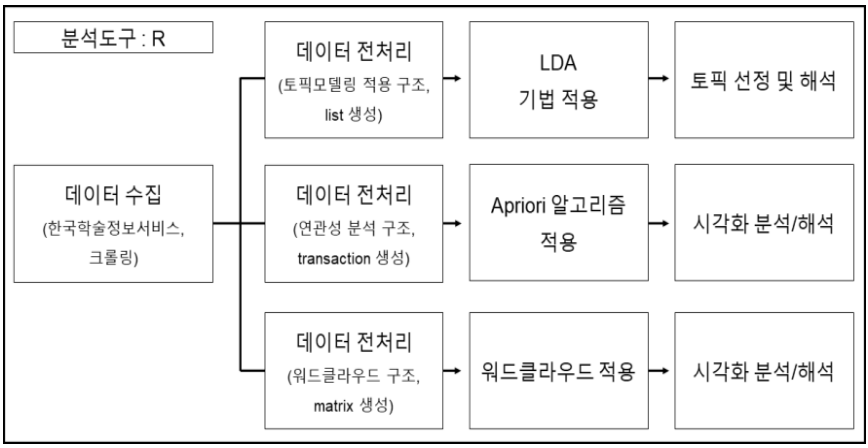
III. 연구방법

1. 분석 절차 및 방법

본 연구는 기 논의한 바 토픽모델링 분석과 더불어 연관분석, 워드클라우드 분석을 수행한다. 세 가지 분석을 병행함으로써 본고는 연구 결과물에 대한 신뢰도를 높이고자 한다. 먼저 R 소프트웨어를 통해 데이터 수집을 위한 크롤링 정의, 데이터 분석, 시각화를 진행한다. 이에 따른 연구방법 및 절차는 다음의 그림과 같다. 크롤링으로 수집된 데이터를 기반으로 세 가지 분석 기법에 적합한 데이터 구조를 생성하고 불용어(stopwords)를 처리하는 순서로 데이터를 전처리한다. 이 과정에서 필요시 탐색적 분석을 수행하여 1차적으로 의미 있는 정보를 추출한다. 이후 토픽모델링, 연관분석, 워드클라우드

분석을 위한 알고리즘 및 함수를 적용함으로써 분석 결과물을 생성한다.

그림 1. 연구방법 및 절차



출처: 저자 작성

2. 데이터수집

본고는 데이터 수집을 위해 국내 논문 제공 사이트 한국학술정보(RISS)를 활용했다. 국내에서는 논문 작성자에 따라 중남미 또는 라틴아메리카 용어로 작성하는 만큼 상기 사이트로부터 논문 제목에 '중남미' 그리고 국문초록 키워드에 '라틴아메리카'가 포함된 국내 등재학술지를 포함한 국내 학술 논문 전체를 대상으로 검색한 결과 1968년부터 2022년 12월 현재 발행된 논문 총 1071편이 검색되었다. 그리고 이 논문들에 대한 한글 요약문(의 주제어)이 수집 대상이다. 본고는 이 과정에서 웹사이트에서 제공하는 방대한 분량에 대한 데이터의 자동화된 수집을 위해 사용자 크롤링(Crawling)을 정의했다. 연구동향을 분석하는 기존 연구는 논문 키워드를 수집하여 분석하는 사례(박진희 외, 2021; 전은수 외, 2022)를 확인 할 수 있다. 본 연구는 논문에 대한 내용을 비교적 자세히 축약하고 있는 요약문을 대상으로 분석을 수행함으로써 보다 구체적인 연구동향 분석을 진행한다.

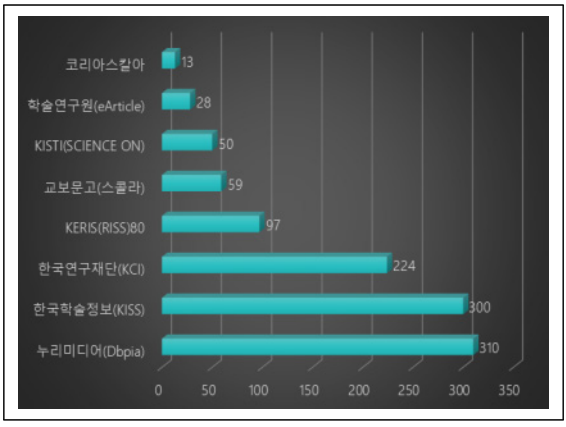
표 1. 데이터 수집 대상 및 범위

수집기관	발행 연도	수집대상	수집방법
한국학술 정보서비스 (RISS)	1968 ~ 2022	- 논문 제목에 '중남미' 단어 포함 - 논문 국문 초록 키워드에 '라틴아메리카' 단어 포함 - 국내등재학술지 (후보지 포함) 및 간행물 - 국문요약문	R에서의 크롤링 정의

출처: 저자 작성

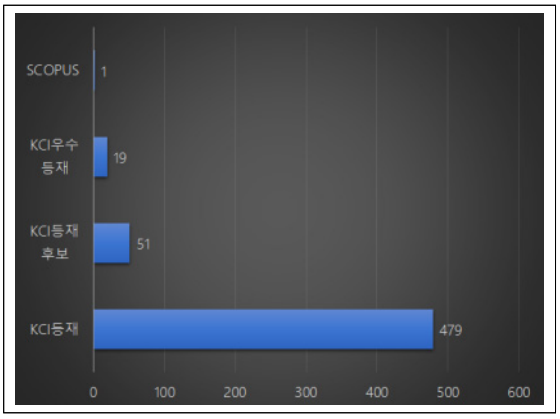
한국학술정보 서비스(RISS) 웹사이트를 통해 수집 대상을 산정하며 추출한 데이터의 기초 정보는 다음과 같다.

그림 2. 원문제공처



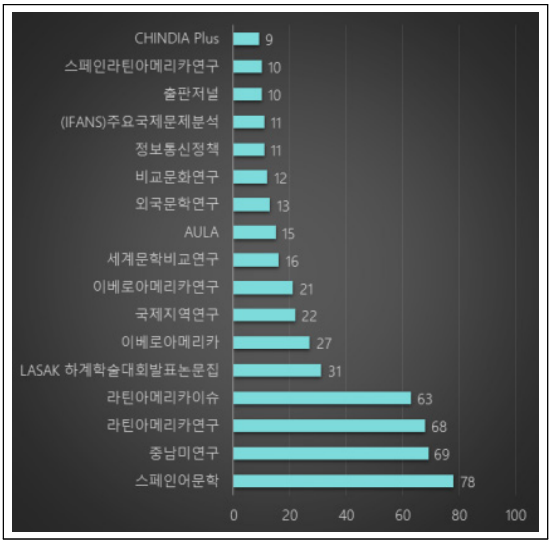
출처: RISS 활용, 저자 작성

그림 3. 등재정보



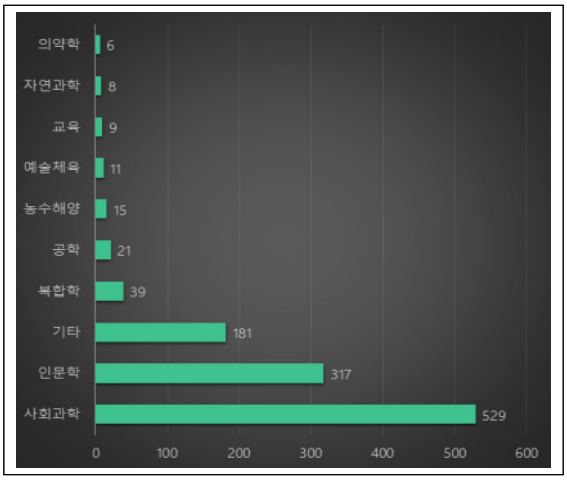
출처: RISS 활용, 저자 작성

그림 4. 학술지명 (간행물 포함)



출처: RISS 활용, 저자 작성

그림 5. 주제별 분류



출처: RISS 활용, 저자 작성

본 연구의 조건에 따라 산출된 모두 1071건의 데이터의 기초자료는 원문 제공처, 등재정보, 학술지명 그리고 주제별 분류 등을 통해 수집된 자료의 특성을 일부 파악할 수 있다. 특히 추후 논의할 바 이지만, 본 연구는 개별 국가 가령 브라질 등을 산정하지 않고 '중남미' 또는 '라틴아메리카' 를 제목 또는 주제어로 설정하여 데이터 추출 한 바 학술지명에는 포-브라질 학회지 등을 확인할 수 없다. 이와 같이 RISS에서 제공하는 데이터를 바탕으로 기초정보는 R 프로그램을 활용 크롤링(crawling) 정의에 따라 다음과 데이터 전처리로 진행된다.

3. 데이터 전처리

본 연구는 앞서 설명한 바, 연구결과물에 대한 신뢰도를 높이기 위해 세 가지 분석을 병행 수행했다. 먼저 토픽모델링 분석을 위해 LDA(Latent Dirichlet Allocation) 기법을 적용한다. 먼저 두 글자 이상의 명사 추출 후 LDA기법을 적용하기 위한 말뭉치(Corpus)를 생성하기 위해 어휘화(lexicalize) 함수를 사용했다. lexicalize() 함수는 LDA패키지에 정의된 추론 절차에 적합

한 형태의 말뭉치(Corpus)와 어휘를 추출한다(강지훈 · 조치영 2022).

동일한 방법으로 연관분석을 위해 두 글자 이상의 명사를 추출 후 트랜잭션(Transaction) 객체를 생성한다. 연관분석에는 Apriori알고리즘이 사용되는데 Apriori알고리즘은 트랜잭션(Transaction) 데이터의 객체를 기반으로 수행되므로 추출된 단어를 트랜잭션(Transaction)객체로 변환하는 과정이 필요하다. 마지막으로 워드클라우드 분석을 위해 말뭉치(Corpus)를 생성한다. 이후 빈도분석이 용이하도록 데이터 구조를 매트릭스(Matrix)로 변환하는 과정을 거쳤다. 세 가지 분석 방법 모두 공통적으로 R프로그램을 사용하여 불용어(stopwords)를 제거하는 과정을 거쳤다. 불용어는 특수기호 등의 문자를 포함해 출현빈도는 높으나 분석에 특별한 의미를 가지지 않는다고 판단된 단어를 제거했다. 가령 '논문', '연구'. '분석', '중남미'. '라틴아메리카' 등의 단어 등은 본고의 연구대상에 빈번히 들어가는 단어들이므로 이와 같은 단어는 불용어로 처리했다.

표 2. 분석방법

내용	분석방법		
	1. 토픽모델링 분석	2. 연관분석	3. 워드클라우드 분석
데이터 전처리	Corpus 생성 및 어휘 추출, 불용어 처리	Transaction 데이터 생성, 불용어 처리	Corpus 생성, 불용어 처리
분석	LDA(잠재 디리클레 할당) 분석	연관성 분석 및 시각화	워드클라우드 시각화

출처: 저자 작성

IV. 연구 결과 및 토론

1. 토픽모델링 분석과 결과

기 수행된 전처리 절차를 거쳐 LDA 분석을 위해 깁스 샘플링(Collapsed Gibbs Sampling)을 진행했다. 깁스 샘플링은 N차 자료를 대상으로 확률 분포를 계산하기 위한 방법이다. 다시 말해, 특정 문서에 포함된 단어가 지정된 N개의 주제에 포함될 확률을 구한 후 각 문서에서 추출된 단어들이 어떤 주제에 포함될지 주제집단을 생성 한다(강지훈 · 조치훈 2022). 이와 같은 정의에 따라 본고는 연구 동향의 양태를 구현 및 분석하기 위해 이성형(2009)이 구분한 시대(세대)간 분류를 반영하며 다음과 같이 시대별로 나눠서 토픽모델링을 진행했다. 아울러, 본 연구는 4~6개의 토픽을 추출하고 각각의 모집단에 각 토픽과 연관성이 높은 상위 10~20의 단어를 추출하는 조건으로 R 프로그램 분석을 진행했다. 이와 같은 방법으로 구현된 토픽은 각 세대(시대)별 4개의 주제별 토픽을 제시할 수 있었다. 그리고 각각의 토픽별 10개의 단어를 추출하며 토픽명을 정했다(한성수 · 양정우 2017, p. 6).

표 3. 한국의 라틴아메리카 연구동향의 주제그룹(토픽) 및 단어 들 (1968~1989)[2)]

안데스	자원	경제	무역
국가	-	-	가능성
페루	분류	통산성	자원에너지청
주요	봉착	수입원	지역
베네수엘라	일반적	자원에너지청	현지조사
뿌리	반면	지역	수입
정착	모색	생산	베네수엘라
양국	하는	수출	중심
민족주의	통산성	가능성등	각국

2) 표 3 이하 '-' 로 표시된 단어는 에러 텀(error term)임.

사상	자원에너지청	현지조사	생산
안데스	지역	실시	수출

<표 3>은 상위 4개의 주제그룹에 따른 단어들이다. 토픽명은 단어들을 보고 유추한 것이다. 본 토픽모델링에 따르면 본 1세대(1968~1989)는 안데스가 주요 대상 지역이며, 자원, 무역, 경제 등의 주요 토픽명 하에 연구가 수행된 것으로 해석할 수 있다.

표 4. 한국의 라틴아메리카 연구동향의 주제그룹(토픽) 및 단어 들 (1990~1999)

인터넷	정체성	철학	서구
-	정체성	인간	종교
경제	소설	노력	국가
인터넷	문학	정치적	부흥
브라질	현실	정신적	주소
제공	역사적	철학	서구
검색	언어	문화	구조
엔진	미학적	독립	여부
목표	정립	확립	존재
기본	진정한	본질	경우
법칙	작가들	낭만주의	지역

<표 4> 역시 상위 4개의 주제 그룹에 따른 10개의 단어들이다. 토픽 모델링에 따른 이 시기의 한국의 라틴아메리카 연구동향은 90년대의 시대상을 반영하여 인터넷이라는 주제그룹이 등장했다. 아울러, 이전 세대와 달리 정체성, 철학, 서구 등으로 형성될 수 있는 주제그룹이 등장하는 등 연구의 지평이 확장된 시기였다.

표 5. 한국의 라틴아메리카 연구동향의 주제그룹(토픽) 및 단어 들 (2000~2010)

정체성	민중주의	경제	문학
정체성	-	사회	-
구분	게릴라	청각화	사회주의
문화	운동	지역	리얼리즘
단어	콜롬비아	이후	시인
본질적	민중주의	경제적	문학
기준	자연	브라질	네루다
논의	통합주의	경제성장	촉진
대통령	정권	국가들	스페인
의미	수출	확대	공산주의
실질적	상품	정책	대부분

<표 5>는 21세기 첫 10년의 한국의 라틴아메리카 연구 동향을 담고 있다. 역시 4개의 주제 그룹에 각각 다른 10개의 단어들의 조합으로 볼 때 라틴아메리카의 주요 담론인 정체성 부분이 지속적으로 부각되고 있다. 이러한 주제의 연장선상에서 또 다른 연구 주제그룹은 민중주의로 담아낼 수 있는 게릴라 운동 등도 확인된다.

표 6. 한국의 라틴아메리카 연구동향의 주제그룹(토픽) 및 단어 들 (2011~2019)

여성	종교	경제	시장
-	사회	결과	지역
여성	논의	영향	경제
사랑	개신교	정책	국가들
등장	필요	증가	확대
바로크적	형성	변화	강화
의류업	관점	미국	진출
나르키소스	신학	출신	고찰
아름다움	문화적	아르헨티나	기업
세계	역사	프로그램	시장
과정	참여	미술	현황

<표 6>은 이전 10년의 연구의 동향과 다른 주제 그룹들도 등장한다. 여성 등으로 토픽모델명을 담아낼 수 있는 일련의 단어들이 등장하며 종교가 주요 연구 동향으로 부각된다. 아울러 경제와 시장 등은 지속적으로 한국의 라틴아메리카 연구의 주요 주제 그룹임이 확인된다.

표 7. 한국의 라틴아메리카 연구동향의 주제그룹(토픽) 및 단어 들 (2020~2022)

스타트업	글로벌 사우스	자연권	발전
브라질	아프리카	-	정책
영향	지역	경제적	지역
결과	경우	자연권	발전
칠레	소장품	주장	국가
생태계	아시아	관계	다양한
변화	라틴	이민	수준
콜롬비아	아메리카	시작	중심
스타트업	무역	인정	정도
대표적	도시	관점	문제
한인사회	가톨릭	문학	사례

<표 7>에 따르면 이전 세대와는 사뭇 다른 연구동향이 이 시기에 확인된다. 주제어 및 이를 담고 있는 주제 그룹이 이를 반영한다. 스타트업 관련 토픽과 자연권의 등장 그리고 한국의 라티아메리카 연구에서 아시아-아프리카 연구 관련 주제가 제시되었다. 그리고 라틴아메리카 연구에서 꾸준히 주제로 다뤄지고 있는 발전을 주제그룹을 통해서 확인할 수 있다.

한국학술정보서비스(RISS)를 통해 확보된 데이터를 R 프로그램의 크롤링 정의에 따라 데이터전처리 이후 토픽모델링을 구동한 결과 한국의 라틴아메리카의 연구는 다음과 같이 정리할 수 있다. 1세대로 구분지은 68년-89년 사이의 한국의 라틴아메리카 연구 동향은 자원경제 그리고 무역과 관련된 단어로 구성된 주제어를 통해서 확인할 수 있다. 이는 한국을 필두로 하는 아시아의 발전주의국가론의 담론 하에 한국의 對 라틴아메리카 특성을 자원경제에 집중을 두고 연구를 진행한 것으로 분석할 수 있다. 20세기 마지막 10

년은 신자유주의 물결하의 글로벌 차원의 연구주제가 다변화 되었음을 확인할 수 있다. 일례로 인터넷을 필두로 하는 주제그룹이 등장하며 라틴아메리카의 정체성에 대한 연구가 진행되었기 때문이다. 또한 21세기의 첫 10년은 정체성 연구의 연장선상에서 민주주의 등 라틴아메리카적 연구를 한국학계에서 심도 있게 진행한 것으로 판단할 수 있다. 이후의 10년 간은 여성, 종교 등 라틴아메리카에서 소외되었던 연구 주제가 등장한 시기이다. 또한 본고에서는 5세대로 명명한 코로나 발생시기의 지난 3년간의 연구는 라틴아메리카의 가능성과 시대상이 반영된 연구가 진행되었다 분석할 수 있다. 스타트업 연구와 자연권의 주제그룹의 등장했기 때문이다. 라틴아메리카가 글로벌 다국적 기업의 전초기지(생산기지)의 역할도 하고 있지만 동시에 라틴아메리카 자체의 자생적 기업(스타트업)의 등장을 반영한 연구가 진행되고 있다. 더욱이 코로나19로 더욱 그 의미가 부각되는 현 인류 발전 모델인 자본주의체제의 대척점이자 대안으로 자연권에 대한 연구가 진행되고 있다.

2. 연관분석의 분석과 결과

연관분석은 다양한 지표를 통해 데이터 간의 연관성을 추적하고, 추출된 단어의 좌항(left hand side, lhs)과 우항(right hand side, rhs)의 연관성 여부를 분석하여 표현할 수 있다. 즉, lhs 단어가 출현 했을 때 rhs 단어의 동시 출현 확률을 의미한다. 반대로 rhs 출현 시 lhs 출현 확률도 분석 가능하다. 앞서 서술한 것처럼 연관분석을 위해 지지도(Support), 신뢰도(Confidence), 향상도(Lift)의 세 가지 변수가 활용된다. 일반적으로 지지도(Support)가 높을수록 연관성이 높다고 본다. 또한 신뢰도(Confidence) 값이 1에 가까울수록 연관성이 높은 것으로 해석할 수 있다. 향상도(Lift)는 수치 1을 기준으로, lhs와 rhs가 관련성이 없다면 1이 출력되고 향상도가 1보다 크면 향상도 값이 클수록 관련성이 높다고 본다. 그러나 향상도가 1보다 작으면 lhs가 출현했을 때 rhs는 출현하지 않는 것으로 해석한다. 이와 같은 연관성 분석에 대한 해석을 바탕으로 한국의 라틴아메리카 연구 동향을 시대(세대)별로 구분했으며 다음과 같은 결과가 도출되었다.

표 8. 한국의 라틴아메리카 연구동향의 연관성 분석결과 (1968~1989)

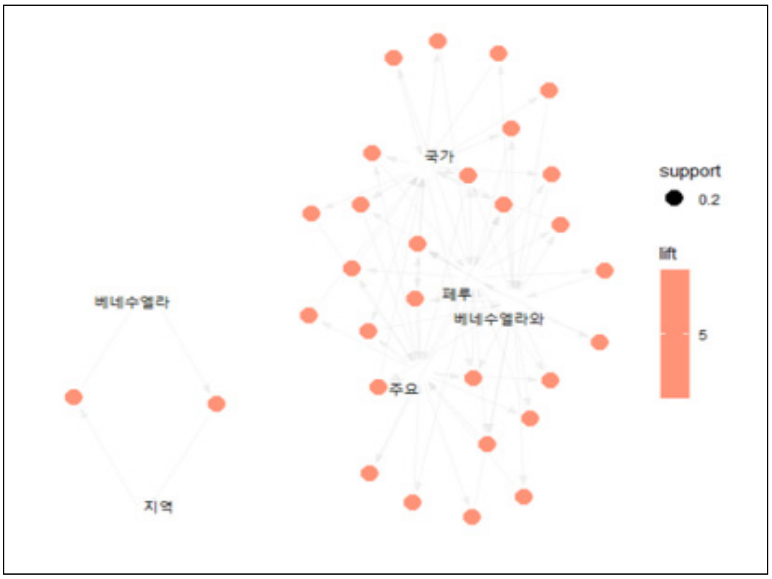
lhs	rhs	support	confidence	lift
{페루} =>	{주요}	0.2	1	5
{주요} =>	{페루}	0.2	1	5
{페루} =>	{베네수엘라와}	0.2	1	5
{베네수엘라와} =>	{페루}	0.2	1	5
{페루} =>	{국가}	0.2	1	5
{국가} =>	{페루}	0.2	1	5
{주요} =>	{베네수엘라와}	0.2	1	5
{베네수엘라와} =>	{주요}	0.2	1	5
{주요} =>	{국가}	0.2	1	5
{국가} =>	{주요}	0.2	1	5
{베네수엘라와} =>	{국가}	0.2	1	5
{국가} =>	{베네수엘라와}	0.2	1	5
{지역} =>	{베네수엘라}	0.2	1	5
{베네수엘라} =>	{지역}	0.2	1	5
{주요,페루} =>	{베네수엘라와}	0.2	1	5
{베네수엘라와,페루} =>	{주요}	0.2	1	5
{베네수엘라와,주요} =>	{페루}	0.2	1	5
{주요,페루} =>	{국가}	0.2	1	5
{국가,페루} =>	{주요}	0.2	1	5
{국가,주요} =>	{페루}	0.2	1	5

본고의 시대(세대)별 구분에 따라 분류한 1세대는 다른 세대에 비해 중남미 또는 라틴아메리카로 제목 또는 연구 초록의 키워드로 명시한 연구 산출물이 상대적으로 미비하다.[3] 이와 같이 연구 주제 및 범위의 제한성 가운데 <표 9>에 따르면 예를 들어, '페루' 출현 시 '베네수엘라'가 동시에 출현 할

3) 본고의 연구 범위에 기준으로 RISS 데이터 베이스 확인한 결과 1968-89년 사이 모두 105편이 등재됨. 이는 년 간 평균적으로 5편 정도의 논문이 발간 됨. 본고에서는 구체적으로 다루지는 않았지만 RISS 데이터에 따르면, 1세대 가운데 80년대 초반 논문이 집중적으로 양산됨. 예를 들어, 68년 1편, 71, 72년 각 1편인 반면 82년 14편, 83년 20편 임.

확률이 신뢰도(confidence)가 1이므로 100%이고, 동일한 조건하에 향상도(lift) 수치가 5로 1이상임으로 상호 연관성이 높다고 해석할 수 있다. 이에 대해 <그림 6>는 lift 와 support를 중심으로 키워드 네트워크 시각화 한 것이다.

그림 6. 한국의 라틴아메리카 연구동향의 연관분석 시각화(1968~1989)[4)]



이전 시대(세대)와 동일한 조건하에 제2세대는 모두 (200여) 건의 논문을 발간했으며, 이를 바탕으로 <표 9>와 같은 분석결과가 도출되었다.

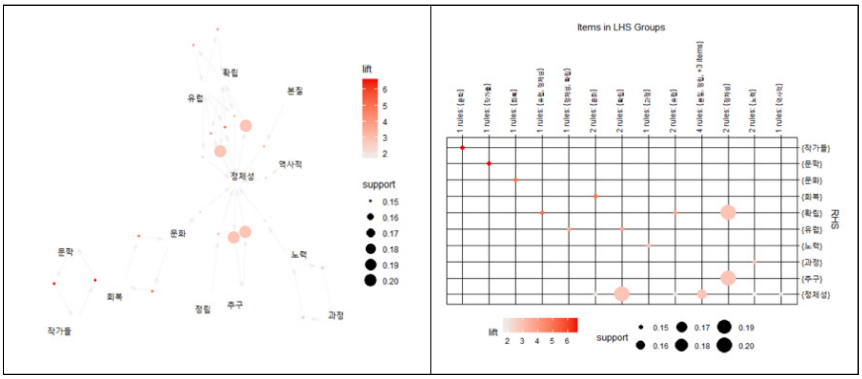
4) <그림 7>과 같이 <그림 6>이 좌항과 우항의 연관성을 시각화한 격자 모양의 그림이 부재한 것은 앞서 설명한 바, 데이터가 소량으로 R 프로그램 상에서 구현이 되지 않음.

표 9. 한국의 라틴아메리카 연구동향의 연관성 분석결과 (1990~1999)

lhs	rhs	support	confidence	lift
{회복} =>	{문화}	0.15	1	5
{문화} =>	{회복}	0.15	0.75	5
{본질} =>	{정체성}	0.15	1	2.857143
{문화} =>	{정체성}	0.15	0.75	2.142857
{추구} =>	{정체성}	0.2	1	2.857143
{정체성} =>	{추구}	0.2	0.571429	2.857143
{확립} =>	{유럽}	0.15	0.75	3.75
{유럽} =>	{확립}	0.15	0.75	3.75
{확립} =>	{정체성}	0.2	1	2.857143
{정체성} =>	{확립}	0.2	0.571429	2.857143
{문학} =>	{작가들}	0.15	1	6.666667
{작가들} =>	{문학}	0.15	1	6.666667
{유럽} =>	{정체성}	0.15	0.75	2.142857
{과정} =>	{노력}	0.15	0.75	3
{노력} =>	{과정}	0.15	0.6	3
{정립} =>	{정체성}	0.15	1	2.857143
{역사적} =>	{정체성}	0.15	0.75	2.142857
{노력} =>	{정체성}	0.15	0.6	1.714286
{유럽,확립} =>	{정체성}	0.15	1	2.857143
{정체성,확립} =>	{유럽}	0.15	0.75	3.75

R 프로그램으로 도출된 상위 20개 연관성 분석 결과 예를 들어 '정체성' 단어 출현 시 '확립' 단어가 동시에 출현할 확률은 57%이다. 그리고 '문화' 단어 출현 시 '회복' 단어의 출현 확률은 75% 이다. 또한 '정체성'과 '확립' 단어 대한 lift값은 각각 2.857143이고 '문화'와 '회복' 단어의 5.000000으로 나왔기 때문에 향상도가 1이상이므로 연관성이 높다고 해석할 수 있다. 그리고 <그림 3>은 이와 같은 연관성의 결과를 시각화 한 것이다.

그림 7. 한국의 라틴아메리카 연구동향의 연관분석 시각화 (1990~1999)



<그림 7>의 좌측은 향상도(lift)와 지지도(support)를 중심으로 키워드별 네트워크 시각화 한 것이다. 우측의 그림은 좌항(lhs)과 우항(rhs)에 있는 단어간 상호 연관성을 시각적으로 확인하도록 한다. 동일한 조건하에 3세대인 2000-2010년간의 한국의 라틴아메리카 연구동향의 연관성 분석결과는 다음과 같다.

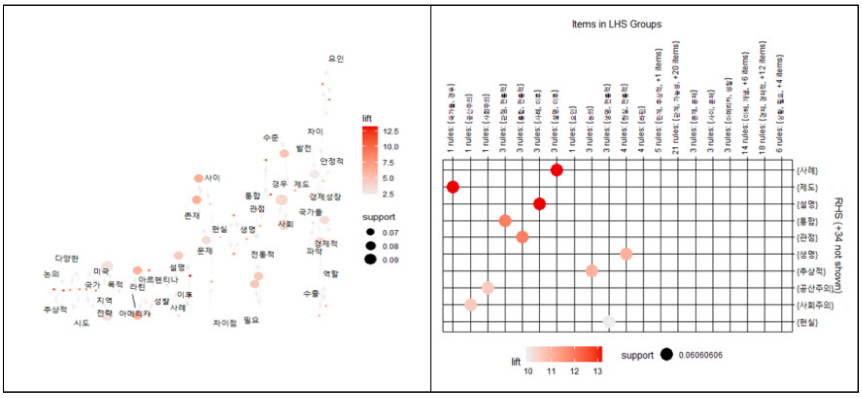
표 10. 한국의 라틴아메리카 연구동향의 연관성 분석결과 (2000~2010)

lhs	rhs	support	confidence	lift
{지역} =>	{국가}	0.090909	0.545455	3
{국가} =>	{지역}	0.090909	0.5	3
{사회} =>	{경우}	0.090909	0.6	3.6
{경우} =>	{사회}	0.090909	0.545455	3.6
{설명} =>	{이후}	0.075758	1	4.4
{수출} =>	{국가들}	0.075758	0.714286	3.367347
{사이} =>	{존재}	0.075758	0.833333	6.111111
{존재} =>	{사이}	0.075758	0.555556	6.111111
{존재} =>	{문제}	0.075758	0.555556	3.333333
{수준} =>	{경우}	0.075758	0.833333	5
{아메리카} =>	{라틴}	0.075758	0.625	5.892857

{라틴} =>	{아메리카}	0.075758	0.714286	5.892857
{발전} =>	{국가들}	0.075758	0.555556	2.619048
{경제적} =>	{국가들}	0.075758	0.714286	3.367347
{전통적} =>	{필요}	0.075758	0.555556	4.583333
{필요} =>	{전통적}	0.075758	0.625	4.583333
{아르헨티나} =>	{이후}	0.060606	0.8	3.52
{경제성장} =>	{국가들}	0.060606	0.8	3.771429
{관점} =>	{통합}	0.060606	0.8	10.56
{통합} =>	{관점}	0.060606	0.8	10.56

21세기 첫 10년은 경제, 특히 아르헨티나, 발전, 통합 등 을 주요 단어로 확인할 수 있다. 예를 들어 '아르헨티나' 단어 출현 시 '이후' 단어가 동시에 출현할 확률은 80%이다. 또한, '관점' 단어 출현 시 '통합' 단어의 출현 확률도 80% 이다. 그리고 '아르헨티나'와 '이후' 단어 대한 lift값은 3.52이고 '관점'과 '통합' 단어는 10.56으로 연관성이 높다고 해석할 수 있다. 그리고 <그림 8>은 이와 같은 연관성의 결과를 시각화 한 것이다.

그림 8. 한국의 라틴아메리카 연구동향의 연관분석 시각화 (2000~2010)



<그림 8>에 따르면 '아르헨티나' 단어를 중심으로 '성찰', '이후' 등의 주요 단어가 네트워크로 연결되어 있다. 이는 21세기 초 아르헨티나 디폴트 상황

에 대한 한국의 라틴아메리카 연구의 정향을 파악할 수 있다. 한편, 본고의 분류 기준에 따른 4세대 연구 시대의 주요 키워드 및 이 단어들 간 연관성 분석은 다음과 같다.

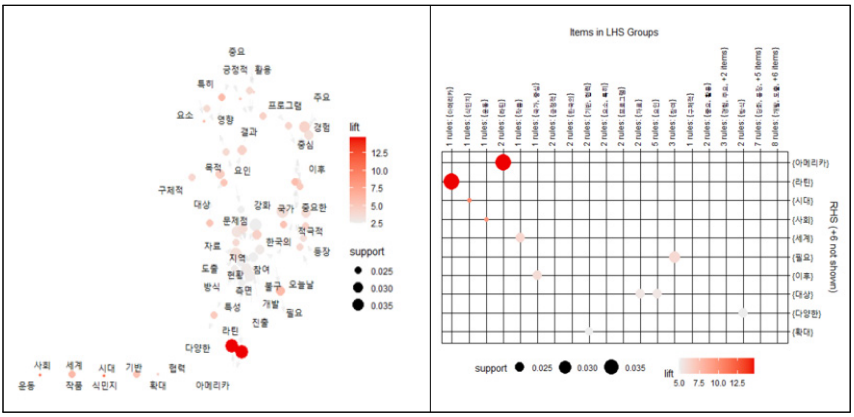
표 11. 한국의 라틴아메리카 연구동향의 연관성 분석결과 (2011~2019)

lhs	rhs	support	confidence	lift
{라틴} =>	{아메리카}	0.039146	0.785714	14.71905
{아메리카} =>	{라틴}	0.039146	0.733333	14.71905
{현황} =>	{지역}	0.035587	0.714286	3.521303
{강화} =>	{국가}	0.035587	0.5	3.426829
{강화} =>	{지역}	0.035587	0.5	2.464912
{진출} =>	{지역}	0.032028	0.529412	2.609907
{특성} =>	{지역}	0.032028	0.5625	2.773026
{자료} =>	{지역}	0.032028	0.642857	3.169173
{경험} =>	{중심}	0.032028	0.6	4.215
{한국의} =>	{지역}	0.02847	0.888889	4.382066
{문제점} =>	{지역}	0.02847	0.727273	3.585327
{요인} =>	{목적}	0.02847	0.571429	4.587755
{요인} =>	{결과}	0.02847	0.571429	4.339768
{중요한} =>	{국가}	0.02847	0.533333	3.655285
{참여} =>	{필요}	0.02847	0.571429	5.94709
{참여} =>	{지역}	0.02847	0.571429	2.817043
{불구} =>	{지역}	0.02847	0.5	2.464912
{주요} =>	{중심}	0.02847	0.5	3.5125
{긍정적} =>	{영향}	0.024911	0.777778	5.906907
{도출} =>	{지역}	0.024911	0.777778	3.834308

<표 11>에 따르면 이전 시기와 다르게 '한국'이라는 단어가 등장하며 '한국' 단어 출현 시 '지역' 단어가 동시에 출현할 확률은 88%이다. 또한, '문제점' 단어 출현 시 '지역' 단어의 출현 확률도 72% 이다. 그리고 '한국'과 '지

역’ 단어 대한 lift값은 4.38이고 ‘문제점’과 ‘지역’ 단어는 3.58로 연관성이 높다고 해석할 수 있다. 그리고 <그림9>은 이와 같은 연관성의 결과를 시각화한 것이다.

그림 9. 한국의 라틴아메리카 연구동향의 연관분석 시각화 (2011~2019)



특히 <그림 9>의 좌측 현황에 따르면 ‘한국’ 이라는 단어를 중심으로 ‘적극적’ ‘등장’, ‘주요한’ 이라는 단어들이 네트워크 즉 상호 연결됨을 확인 할 수 있다. 이는 한국의 라틴아메리카 연구에서 담긴 한국과 라틴아메리카 간의 상호 관계성을 짐작할 수 있다. 한편, 본고는 10년 단위의 시대 구분과 더불어 코로나 19의 원년이 2020년 인바 이를 반영해서 또 다른 시대의 구분으로 설정했다. <표 12>는 이와 같이 구분한 바에 따른 한국의 라틴아메리카 연구동향의 연관성 분석을 담고 있다.

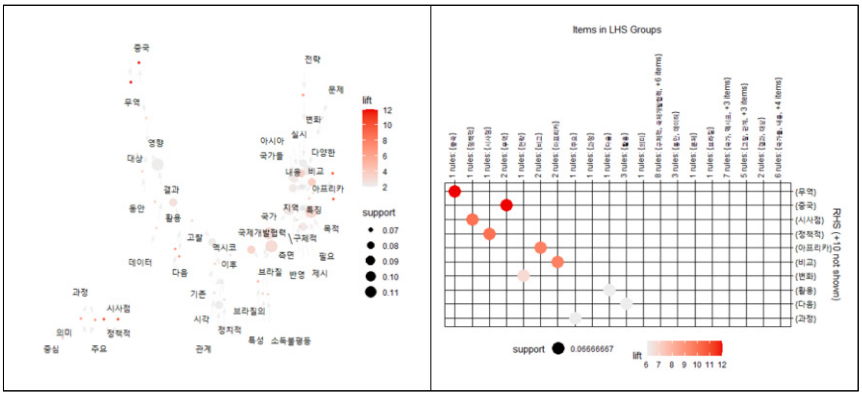
표 12. 한국의 라틴아메리카 연구동향의 연관성 분석결과 (2020~2022)

lhs	rhs	support	confidence	lift
{목적} =>	{지역}	0.116667	0.875	3.088235
{결과} =>	{영향}	0.116667	0.5	2
{국가} =>	{브라질}	0.116667	0.636364	2.937063
{브라질} =>	{국가}	0.116667	0.538462	2.937063

{국가} =>	{지역}	0.116667	0.636364	2.245989
{다양한} =>	{지역}	0.1	0.545455	1.925134
{비교} =>	{지역}	0.083333	1	3.529412
{멕시코} =>	{브라질}	0.083333	0.714286	3.296703
{실시} =>	{지역}	0.083333	1	3.529412
{활용} =>	{결과}	0.083333	0.625	2.678571
{국가들} =>	{지역}	0.083333	0.625	2.205882
{내용} =>	{지역}	0.083333	0.625	2.205882
{제시} =>	{지역}	0.083333	0.625	2.205882
{고찰} =>	{이후}	0.083333	0.5	2
{관계} =>	{이후}	0.083333	0.5	2
{변화} =>	{지역}	0.083333	0.555556	1.960784
{소득불평등} =>	{브라질}	0.066667	0.666667	3.076923
{정책적} =>	{시사점}	0.066667	1	10
{시사점} =>	{정책적}	0.066667	0.666667	10
{국제개발협력} =>	{지역}	0.066667	1	3.529412
{데이터} =>	{결과}	0.066667	0.666667	2.857143
{구체적} =>	{지역}	0.066667	1	3.529412
{특징} =>	{지역}	0.066667	1	3.529412
{중국} =>	{무역}	0.066667	1	12
{무역} =>	{중국}	0.066667	0.8	12
{아시아} =>	{지역}	0.066667	0.8	2.823529
{비교} =>	{아프리카}	0.066667	0.8	9.6

<표 12>에 따르면 1세대에 등장한 바 있는 '페루'와 '베네수엘라'와 같이 5세대는 '멕시코'와 '브라질'이 동시 출현함으로 상호 연관 단어로 등장한다. 이와 더불어 '소득불평등'과 '브라질', '국제개발협력'과 '지역', '아시아'와 '지역', '중국'과 '무역' 그리고 '아프리카'와 '비교' 등 단어 간의 신뢰도와 연관성이 모두 높다. <그림 10>은 이를 시각화 한 것이다.

그림 10. 한국의 라틴아메리카 연구동향의 연관분석 시각화 (2020~2022)



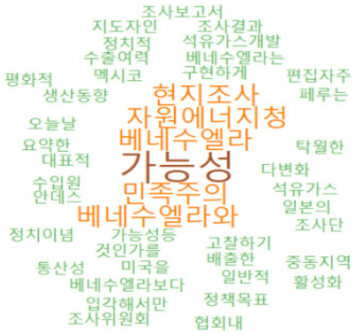
시각화된 <그림 10>에서도 확인할 수 있듯이 연관성 분석에 따른 한국의 라틴아메리카 연구 동향에서는 코로나 19와 관련된 직접적 연구는 확인할 수 없다. 앞서 설명했듯이 본고의 연구 범위는 중남미와 라틴아메리카를 각각 연구물의 제목 또는 초록의 키워드로 작성된 것으로 산정한바 한국의 라틴아메리카 모든 연구물을 오롯이 담아내는 데는 한계가 있다[5].

3. 워드클라우드 분석과 결과

앞서 수행된 연관분석은 동시출현단어에 대한 연관성을 분석하는 알고리즘이다. 즉 연관분석은 특정 단어의 출현빈도(단독)가 높다하더라도 일정 횟수 이상 동시에 출현하는 단어가 없으면 추출되지 않을 가능성을 배제할 수 없다. 따라서 이와 같은 연구의 맹점을 보완하고자 단순히 출현빈도가 높은 단어의 확인을 위해 워드클라우드 분석을 병행했다. 다시 말해, 특정 문서를 해석함에 있어 단어의 동시출현 빈도와는 별개로 문서의 맥락과 흐름을 결정하는 핵심 단어를 추출하며 그 의미를 고찰하기 위함이다.

5) 본고가 한국의 라틴아메리카 코로나 19관련 논문을 전수 조사 한 바 권기수 「브라질에서 코로나 19 치명률의 사회경제적 결정요인에 관한 연구」 등 모두 6편임.

그림 11. 한국의 라틴아메리카 연구동향의 워드클라우드 분석(시대별)

<table>
<tr><th>(1968~1989)</th><th>(1990~1999)</th></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td>(2000~2010)</td><td>(2011~2019)</td></tr>
<tr><td> 
</td><td> 
</td></tr>
<tr><td colspan="2">2020–2022</td></tr>
<tr><td colspan="2"> 
</td></tr>
</table>

워드클라우드 분석의 특성은 시각화이다. 다시 말해 앞서 설명한 바, 워드클라우드 기법은 결과물에 대해 직관적으로 파악할 수 있도록 한다. 문서상에 가장 많이 출현하는 즉 빈도수가 높은 단어는 시각적으로 크게 상대적으로 빈도수가 낮으면 작게 구현되기 때문이다. 한국의 라틴아메리카 연구동향을 세대별로 구분한 것을 '한 눈(taking a quick look)'으로 확인할 수 있다. 1세대는 '가능성', '자원에너지청', 2세대는 '정체성' '인터넷', 3세대는 '정체성', '게릴라', 4세대는 '다양한', '국가들', 5세대는 '브라질' '생태계' 등의 키워드가 두드러진다. 1세대는 라틴아메리카의 '가능성', '자원'을 통한 경제성장 가능성과 관련된 연구가 주된 주제로 분석할 수 있다. 2세대와 3세대는 '정체성' 그리고 '게릴라' 등의 키워드 등장이 신자유주의에 대한 대척점적 연구가 반영된 것으로 분석할 수 있다. 4세대는 '다양성'이 주요 연구의 소재였고 5세대는 작금의 코로나 19로 부각된 자본주의 등에 대안적 연구로 '생태계'가 등장했다고 설명할 수 있다.

V. 나가며

본고는 한국의 라틴아메리카 지역 연구 동향 파악이 목적이다. 이를 위해 기존의 연구물의 선도성을 바탕으로 방법론적 차별성을 구가한 연구를 진행했다. 한국의 라틴아메리카 지역연구 현황과 관련된 방법론적 선도성 이외에도 기존의 빅데이터 분석을 활용한 다른 분과 학문 또는 연구주제와 견주어도 방법론적 의미가 있다. 텍스트 마이닝 기반의 기계학습 시 본고는 토픽모델링, 연관분석 그리고 워드클라우드 세 가지 기법을 동일한 데이터에 적용하며 한국의 라틴아메리카 연구 동향을 분석했기 때문이다.

이와 같은 방법론적 차별성과 선도성에도 불구하고 한국의 라틴아메리카 연구의 동향을 온전히 담아내는 데는 한계가 있다. 본고가 빅데이터 분석을 위해 활용한 RISS 데이터베이스를 통해 기초 데이터 확보 시 한국에서 발행한 모든 라틴아메리카 관련 연구물을 크롤링, 즉 데이터를 전처리함에 제약

이 있기 때문이다. 다시 말해, 본 RISS 사이트에 제목과 키워드 등을 입력할 수 있는 검색 창이 최대 5개가 까지 제공됨에 따라 라틴아메리카 33개국 모두를 입력할 수 는 없다. 하지만, 라틴아메리카 주요 국가 또는 4-5개 국 단위로 구분한 이후 이를 수 차례 R 프로그램의 크롤링 기법으로 진행한다면 좀 더 광범위한 데이터를 바탕으로 한국의 라틴아메리카 연구 동향에 대해 기계학습을 통한 결과가 산출되며 연구자가 이를 분석할 수 있기 때문이다.

앞서 설명한 바, 본 연구의 방법론적 차별성과 한계성과는 별도로 본 연구의 학술적 의미는 다음과 같이 세 가지로 정리할 수 있다. 첫째, 한국의 라틴아메리카 연구 동향을 60년대부터 현재까지 본 연구가 설정한 범주 내에서 전수 조사 한 것이다. 둘째, 세대별로 구분하며 시대별 연구의 정향을 고찰한 기존 연구의 선도성을 바탕으로 본고는 5세대로 구분 지으며 각 세대별 연구의 특성을 파악 및 분석했다. 본 연구의 분석 단위인 시대(세대)별 한국의 라틴아메리카의 연구의 공통점은 경제(무역)이 주요 연구 대상이며 90년대 이후 다양한 연구 주제가 등장했다. 특히 글로벌 차원의 문제성에 대한 대안 및 선도성으로 라틴아메리카의 생태 연구 등이 진행되고 있음을 확인할 수 있었다. 셋째, 본 연구방법론을 통해 중미 또는 남미의 소지역 단위의 연구 동향 파악 뿐만 아니라, jstor 등 국제 학술지 데이터베이스 사이트를 통해 크롤링 정의하고 글로벌 차원의 라틴아메리카 연구 동향을 분석할 수 있는 선행 연구로의 의미를 가진다.

참고문헌

강지훈 · 조치영 (2022). 텍스트마이닝 기반 국내 예루살렘 연구동향 분석 연구. *차세대 융합기술학회 논문지*, *6*(5), 799-809.

곽재성 (2002). 라틴아메리카 지역연구의 동향과 발전과제. *국제지역연구*, *6*(2), 3-27.

김나연 · 이상엽 (2022). 신문사의 정치 성향에 따른 코로나 19 보도 내용 분석. *한국정보사회학회·한국미디어경영학회*, *23*(1), 69-105.

김달용 (1989). 라틴아메리카 지역연구에 관한 일고. *라틴아메리카 연구*, *2*(1), 154-174.

김민경 (2020). 워드 클라우드 분석으로 본 에어비앤비 숙소의 인기 키워드 연구. *호텔외식관광경영학회*, *29*(4), 347-363.

김은경 · 조대연 (2022). 토픽 모델링과 키워드 네트워크 분석을 활용한 국내 성인 대상 교육훈련 요구분석 연구동향: 2012-2021년 학술지 게재 논문을 중심으로. *HRD연구*, *24*(1), 397-421.

김준환 · 문형진 · 이항 (2021). 워드 클라우드 기법을 이용한 국내 융복합 학술연구 트렌드 분석. *디지털융복합연구*, *19*(2), 33-38.

김한준 · 장재영 (2011). 연관규칙 마이닝을 활용한 뉴스기사 키워드의 연관성 탐사. *한국인터넷방송통신학회 논문지*, *11*(6), 63-71.

노설현 (2020). 토픽모델링을 활용한 인공지능 관련 이슈 분석. *디지털융복합연구*, *18*(5), 75-87.

노형남 (2014). 워드 클라우드에 의한 환대 경영 전략. *관광연구*, *29*(4), 335-353

류기동 · 김종명 · 금영정 · 강필성 · 김우제 (2016). 연관 규칙 분석을 활용한 ARS 추천 메뉴 시스템 연구. *Journal of KITT*, *14*(3), 127-136.

박애스더 · 김정민 (2021). 워드 클라우드 기법과 KJ법을 통한 교양교육 연구동향. *학습자중심교과교육연구*, *21*(4), 1209-1231.

박진희 · 전미선 · 배선형 · 김희준 (2021). 암생존자 삶의 질 영향요인에 대한 연구동향: 텍스트 네트워크 분석과 토픽모델링. *대한종양간호학회*,

21(4), 231-240.

신서영 · 이범준(2021). 코로나 19 확산에 따른 외식에 대한 소비자 인식 분석: 토픽모델링 및 네트워크 분석의 활용. *호텔경영학연구*, *30*(8). 71-90.

이성형 (2009). 한국 라틴아메리카 연구의 회고와 전망: 정치학 분야를 중심으로. *트랜스라틴, 7*(5), 33-49.
http://translatin.snu.ac.kr (접속일 2022.12.01.)

이유빈 · 이영호 · 성정창 · 애나 스타네스쿠 · 지상훈 · 황철수 (2020). 계량적 모델을 통한 지리학 연구의 최신동향 및 토픽 분석. *대한지리학회지, 55*(6), 589-599.

이태혁 (2022). 워드 클라우드 기법을 이용한 K-콘텐츠 관련기사(2011-2022) 분석과 함의. *한국과 세계, 4*(11), 149-167.

임두빈 · 임병학 (2020). 텍스트 마이닝 분석을 활용한 국내 포르투갈어권 지역 연구동향 분석. *중남미연구, 37*(4), 179-214.

장성희 · 이창원 (2022). 토픽모델링을 이용한 기독교 연구 동향 분석. *로고스경영연구, 20*(4), 93~112.

정호윤 (2021). 빅데이터 분석을 통한 한국 포털사이트의 라틴아메리카 관련 뉴스 보도 행태 연구. *포르투갈-브라질 연구, 18*(1), 125-151.

차경미 (2021). 라틴아메리카 지역연구동향 및 개별국가연구. *비교문화연구*, *22*, 203-221.

최권준 (2021). 한국의 중남미 문학 연구 동향 연구 -중남미 전문 학술지들을 중심으로. *중남미연구, 40*(3), 199-244.

최윤국 (2003). 중남미문학 한국의 스페인·중남미 연구 동향 및 향후 과제 - 대학 교육 개선을 중심으로. *한국스페인어문학회*, *28*, 733-753.

최종욱 외 (2022). *한-중남미 수교 60주년*. 외교부.

최출헌 · 장필식 (2019). 글로벌 디자인 연구동향에 대한 키워드 네트워크 분석 연구(1999~2018). *융합정보논문지, 9*(2), 7-16.

한성수 · 양동우 (2017). 텍스트마이닝을 이용한 창업 관련 연구 동향 분석. *벤처창업연구, 12*(5), 1-12.

홍욱헌 (2012). 한국의 라틴아메리카 지역연구 동향. *아시아리뷰, 2*(2), 149-169.

Heimerl, F., Lohmann, S., Lange, S. & Ertl, T. (2014). Word Cloud Explorer: Text Analytics Based on Word Clouds. *Proceedings of the 47th Hawaii International Conference on System Science* (HICSS 2014), January 6-9, 2014 IEEE Computer Society.

Jacobs, T., & Tschötschel, R. (2019). Topic models meet discourse analysis: A quantitative tool for a qualitative approach. *International Journal of Social Research Methodology*, *22*(5), 469-485.

■ **이태혁**

영국 York대학교에서 국제정치학박사학위를 취득하였다. 주요 연구 분야는 글로벌 거버넌스, 지역주의(기구), 아시아-라틴아메리카 관계, 국제개발협력, 정치경제 그리고 질적과 양적의 혼합 연구방법론 등이다. 현재 부산외국어대학교 중남미지역원에서 HK연구교수로 재직 중이다.

(E-mail: gwheok@bufs.ac.kr)

■ **강지훈**

부산외국어대학교에서 공학박사(데이터베이스)학위를 취득하였다. 주요 연구 분야는 디지털인문학, 인문ICT융합, 연구방법론 분야이며 현재 부산외국어대학교 지중해지역원에서 HK연구교수로 재직 중이다.

(E-mail: mooswon@bufs.ac.kr)

Abstract

Suggestion of the Analysis of Latin American Research Trends in Korea: Focusing on topic modeling, association analysis and word cloud analysis

Lee, Taeheok · Kang, Jihoon
(Busan University of Foreign Studies)

The purpose of this study is to analyze the Latin American research trends in Korea through methodological diversity. In other words, based on the leading research related to this research topic, this paper maintains the methodological differentiation of existing qualitative or basic statistics by conducting machine learning-based data analysis. To this end, first, this paper collects data through the Research Information Sharing Service (RISS). The subject and scope of this study is to set to 'Latin America (Jungnammi)' as the title of the paper and and 'Latin America' in the keyword of the abstract, and subsequently are collected according to the definition of crawling in the R software program. Secondly, the data extracted by the definition of crawling is preprocessed for topic modeling, association analysis, and word cloud analysis, which are the research methods of this paper, and then the data is processed with algorithms suitable for each analysis. Third, the unit of analysis of this study is the classification by generation (years). Therefore, based on the results calculated through the analysis of the three big data, the data classified in this way examines trends in Latin American research in Korea. The main topic of research in Latin America, which has been steadily conducted since 1968, is economy (trade)-related topics. and Also other research themes have been studied based on Latin American region-related or global issue.

Key Words

Latin America, Korea, Topic Modeling, Association Analysis, Word Cloud

논문 접수일 : 2023년 2월 01일
심사 완료일 : 2023년 2월 15일
게재 확정일 : 2023년 2월 20일