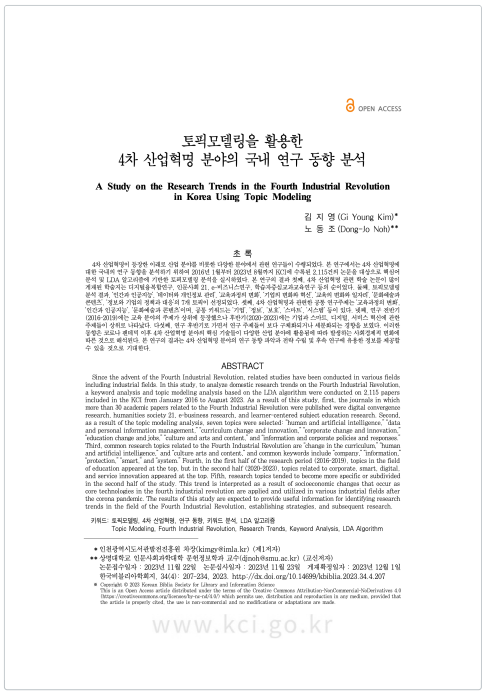
- 코드 분석
- 방법론 분석(LDA)
- 문제 정의를 생각해볼 것
  - 세계 무역량의 85% 이상이 해상운송
  - 선박의 대형화 (규모의 경제 달성을 위해)
  - 스마트 항만의 개발(로테르담, 상하이, 칭다오 항만 무인 완전 자동화 시스템 도입.)
  - 글로벌 컨테이너 항만 물동량의 구조적 변화에 관한 연구(2003~'19) - 세계 100대 컨테이너 항만을 대상으로* 이충배·**이영신***·유염봉****
- 관련 논문 3개 이상 보기(설명해줄 수 있도록 이해하기)
  - [토필모델링을 활용한 4차 산업혁명 분야의 국내 연구 동향 분석]



OPEN ACCESS

토픽모델링을 활용한
4차 산업혁명 분야의 국내 연구 동향 분석

A Study on the Research Trends in the Fourth Industrial Revolution in Korea Using Topic Modeling

김 지 영 (Gi Young Kim)*
노 동 조 (Dong-Jo Noh)**

초 록

4차 산업혁명이 등장한 이래로 산업 분야를 비롯한 다양한 분야에서 관련 연구들이 수행되었다. 본 연구에서는 4차 산업혁명에 대한 국내외 연구 동향을 분석하기 위하여 2016년 1월부터 2023년 8월까지 KCI에 수록된 2,115건의 논문을 대상으로 핵심어 분석 및 LDA 알고리즘에 기반한 토픽모델링 분석을 실시하였다. 본 연구의 결과 첫째, 4차 산업혁명 관련 학술 논문이 많이 게재된 학술지는 디지털융복합연구, 인문사회 21, e-비즈니스연구, 학습자중심교과교육연구 등의 순이었다. 둘째, 토픽모델링 분석 결과, '인간과 인공지능', '데이터와 개인정보 관련', '교육과정의 변화', '기업의 변화와 혁신', '교육의 변화와 일자리', '문화예술과 커뮤니티', '정보과 기업의 경제과 생산과 기술 트레이 선정되었다. 셋째, 4차 산업혁명과 관련한 많은 연구주제는 '교육과정의 변화', '인간과 인공지능', '문화예술과 콘텐츠'이며, 공통 키워드는 '기업', '정보', '보호', '스마트', '시스템' 등이 있다. 넷째, 연구 전반기(2016~2019)에는 교육 분야의 주제가 상위에 등장했으나 후반기(2020~2023)에는 '기업과 스마트, 디지털, 서비스 혁신에 관한 주제들이 상위로 나타났다. 다섯째, 연구 후반기로 가면서 연구 주제들이 보다 구체화되거나 세분화되는 경향을 보였다. 이러한 일련은 코로나 팬데믹 이후 4차 산업혁명 분야의 핵심 기술들이 다양한 산업 분야에 활용됨에 따라 발생되는 사회경제적 변화에 따른 것으로 해석된다. 본 연구의 결과는 4차 산업혁명 분야의 연구 동향 파악과 관련 주제 및 후속 연구에 유용한 정보를 제공할 수 있을 것으로 기대한다.

ABSTRACT

Since the advent of the Fourth Industrial Revolution, related studies have been conducted in various fields including industrial fields. In this study, to analyze domestic research trends on the Fourth Industrial Revolution, a keyword analysis and topic modeling analysis based on the LDA algorithm were conducted on 2,115 papers included in the KCI from January 2016 to August 2023. As a result of this study, first, the journals in which more than 30 academic papers related to the Fourth Industrial Revolution were published were digital convergence research, humanities society 21, e-business research, and learner-centered subject education research. Second, as a result of the topic modeling analysis, seven topics were selected: "human and artificial intelligence," "data and personal information management," "curriculum change and innovation," "corporate change and innovation," "education change and jobs," "culture and arts and content," and "information and corporate policies and responses." Third, common research topics related to the Fourth Industrial Revolution are "change in the curriculum," "human and artificial intelligence," and "culture arts and content," and common keywords include "company," "information," "protection," "smart," and "system." Fourth, in the first half of the research period (2016~2019), topics in the field of education appeared at the top, but in the second half (2020~2023), topics related to corporate, smart, digital, and service innovation appeared at the top. Fifth, research topics tended to become more specific or subdivided in the second half of the study. This trend is interpreted as a result of socioeconomic changes that occur as core technologies in the fourth industrial revolution are applied and utilized in various industrial fields after the corona pandemic. The results of this study are expected to provide useful information for identifying research trends in the field of the Fourth Industrial Revolution, establishing strategies, and subsequent research.

키워드: 토픽모델링, 4차 산업혁명, 연구 동향, 키워드 분석, LDA 알고리즘
Topic Modeling, Fourth Industrial Revolution, Research Trends, Keyword Analysis, LDA Algorithm

* 인천광역시도서관정보진흥원 차장(kjimgy@imla.kr) (제1저자)
** 상명대학교 인문사회과학대학 문헌정보학과 교수(djnoh@smu.ac.kr) (교신저자)
논문접수일자 : 2023년 11월 22일 논문심사일자 : 2023년 11월 23일 게재확정일자 : 2023년 12월 1일
한국비블리아학회지, 34(4) : 207-234, 2023. http://dx.doi.org/10.14699/kbiblia.2023.34.4.207

- 토픽모델링의 결과를 평가하는 방법 - 일관성 혼잡도
- 선행연구분석 우리는 특허 텍스트마이닝 / 항만 텍스트 마이닝 이 두 개 찾은거 넣으면 될듯

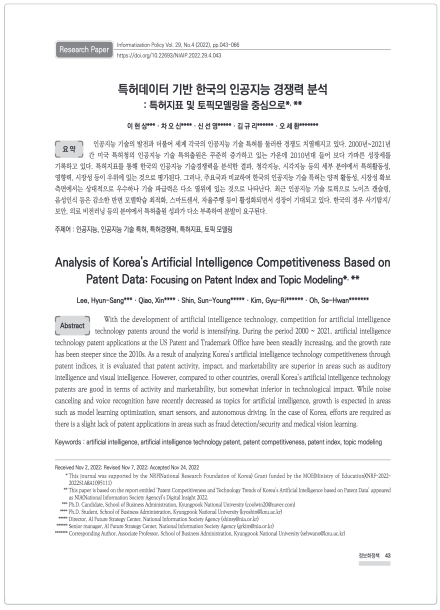
〈그림 1〉 연구 과정

〈표 1〉 데이터 전처리 과정

단 계	방 법	처 리 내 용
1단계	자료수집	• NetMiner Import로 '4차 산업혁명' 자료 수집
2단계	표준화	• 유사어 통일(제4차 산업혁명, 4차 산업혁명) • 제외어 정리(한 글자, 조사, 국가명 등) • 불용어 처리(반복적인 단어 등)
3단계	필터링	• 토픽 추출 검색 키워드 제외
4단계	관련 논문 추출	• 키워드: '4차 산업혁명' • 분석범위: 2016년 1월 ~ 2023년 8월까지 관련 논문 추출

-
  - 이거보니깐 생각난건데, 여기의 '4차 산업혁명' 키워드 사용한 것처럼 우리도 특정 키워드가 하나 있어야 할 것 같다. 3만 5천개가 다 제대로된 애들인지 모르니깐 키워드 넣고 뽑아야 할지도,,?
  - 아니면 특허 추출한 과정을 정확히 알 수 있나? 추출하는 과정이 신뢰성있을지도 모르니깐
- '최적의 토픽 수를 결정하기 위하여 NetMiner의 토픽모델링 평가(Evaluation of Topic Models) 기능을 활 용하였다. 토픽모델링 결과를 평가하는 토픽 일관성 지표의 c_v 점수를 사용하여 최적의 토 픽 개수를 선정하였다.'
  - 최적 토픽 수 어떻게 정할지 중요할 듯.
- 얘네꺼 LDA 이후 토픽 합친건 좀 별로인듯 그냥 더한게 다임.

- [한국의 인공지능 경쟁력 분석]



Research Paper Informatization Policy Vol. 29, No.4 (2022), pp.043-066
https://doi.org/10.22693/NIAIP.2022.29.4.043

특허데이터 기반 한국의 인공지능 경쟁력 분석 : 특허지표 및 토픽모델링을 중심으로*, **

이 현 상*** · 차 오 신**** · 신 선 영***** · 김 규 리****** · 오 세 환*******

요 약 인공지능 기술의 발전과 더불어 세계 각국의 인공지능 기술 특허를 둘러싼 경쟁도 치열해지고 있다. 2000년~2021년 간 미국 특허청의 인공지능 기술 특허출원은 꾸준히 증가하고 있는 가운데 2010년대 들어 보다 가파른 성장세를 기록하고 있다. 특허지표를 통해 한국의 인공지능 기술경쟁력을 분석한 결과, 청각지능, 시각지능 등의 세부 분야에서 특허활동성, 영향력, 시장성 등이 우위에 있는 것으로 평가된다. 그러나, 주요국과 비교해서 한국의 인공지능 기술 특허는 양적 활동성, 시장성 확보 측면에서는 상대적으로 우수하나 기술 파급력은 다소 떨어져 있는 것으로 나타난다. 최근 인공지능 기술 토픽으로 노이즈 캔슬링, 음성인식 등은 감소한 반면 모델학습 최적화, 스마트센서, 자율주행 등이 활성화되면서 점점이 기대되고 있다. 한국의 경우 사이language/보안, 의료 비전러닝 등의 분야에서 특허출원 성과가 다소 부족하여 분발이 요구된다.

주제어 : 인공지능, 인공지능 기술 특허, 특허경쟁력, 특허지표, 토픽 모델링

Analysis of Korea's Artificial Intelligence Competitiveness Based on Patent Data: Focusing on Patent Index and Topic Modeling*, **

Lee, Hyun-Sang*** · Qiao, Xin**** · Shin, Sun-Young***** · Kim, Gyu-Ri****** · Oh, Se-Hwan*******

Abstract With the development of artificial intelligence technology, competition for artificial intelligence technology patents around the world is intensifying. During the period 2000 ~ 2021, artificial intelligence technology patent applications at the US Patent and Trademark Office have been steadily increasing, and the growth rate has been steeper since the 2010s. As a result of analyzing Korea's artificial intelligence technology competitiveness through patent indices, it is evaluated that patent activity, impact, and marketability are superior in areas such as auditory intelligence and visual intelligence. However, compared to other countries, overall Korea's artificial intelligence technology patents are good in terms of activity and marketability, but somewhat inferior in technological impact. While noise canceling and voice recognition have recently decreased as topics for artificial intelligence, growth is expected in areas such as model learning optimization, smart sensors, and autonomous driving. In the case of Korea, efforts are required as there is a slight lack of patent applications in areas such as fraud detection/security and medical vision learning.

Keywords : artificial intelligence, artificial intelligence technology patent, patent competitiveness, patent index, topic modeling

Received Nov 2, 2022; Revised Nov 7, 2022; Accepted Nov 24, 2022

* This Journal was supported by the NRF(National Research Foundation of Korea)-Grant funded by the MOE(Ministry of Education)(NRF-2022-S22DLAB109511)

** This paper is based on the report entitled 'Patent Competitiveness and Technology Trends of Korea's Artificial Intelligence based on Patent Data' appeared as NIA(National Information Society Agency)'s Digital Insight 2022.

*** Ph.D. Candidate, School of Business Administration, Kyungpook National University (coolvin20@naver.com)

**** Ph.D. Student, School of Business Administration, Kyungpook National University (kyookim@knu.ac.kr)

***** Director, AI Future Strategy Center, National Information Society Agency (shinsy@nia.or.kr)

****** Senior manager, AI Future Strategy Center, National Information Society Agency (grkim@nia.or.kr)

******* Corresponding Author, Associate Professor, School of Business Administration, Kyungpook National University (sehwano@knu.ac.kr)

- 얘네도 특허 데이터를 사용했음
- 근데 특허로 미국과 한국에서 각각 어떤 토픽의 특허 출원이 많이 일어나고 경쟁력이 어떤지 분석한 논문임.
- 인공지능 경쟁이 치열한데, 국가별 기술경쟁력 평가가 필요하지 않겠냐 이런 느낌

- 얘네의 연구질문이
    - 최근 인공지능 기술 관련 특허출원 동향 및 추이는 어떠한가?
    - 한국의 인공지능 세부 기술별 경쟁력은 어떠한가?
    - 한국을 포함한 주요국의 인공지능 기술경쟁력은 어떠한가?
    - 인공지능 기술의 주요 토픽 변화 양상은 무엇인가?

  이렇게 연구질문 정해놓는거 좋은거 같음.
- 토픽 모델링한 부분만 보면, 나라별 토픽의 등장확률을 집계하여 한국의 부족한 부분이 어느분야인지 분석하네.

Journal of the Korea Academia-Industrial cooperation Society
Vol. 17, No. 11 pp. 670-681, 2016

http://dx.doi.org/10.5762/KAIS.2016.17.11.670
ISSN 1975-4701 / eISSN 2288-4688

토픽 모델링을 이용한 핀테크 기술 동향 분석

김태경, 최희련, 이홍철*
고려대학교 산업경영공학과

A Study on the Research Trends in Fintech using Topic Modeling

TaeKyung Kim, HoeRyeon Choi, HongChul Lee*
Department of Industrial and Management Engineering, Korea University

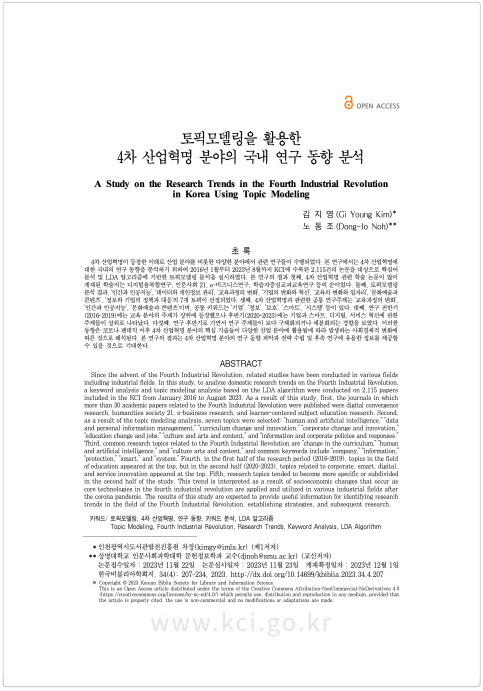
요 약 최근 인터넷과 모바일 환경을 기반으로 금융과 IT가 융합된 핀테크(Fintech) 산업이 급속히 성장하고 있으며 간편성, 편리성 등으로 무장한 핀테크 서비스는 모든 금융서비스의 온라인 모바일 화를 주도하고 있다. 그러나 핀테크 산업의 급격한 성장에도 불구하고, 핀테크 기술에 대한 세부기술 분류와 주요 시장국의 기술개발 동향을 분석하고 기술기획을 지원하기 위한 연구는 매우 미흡한 실정이다. 이에 본 연구는 핀테크 기술의 비정형 데이터 형태의 특허 데이터를 이용하여 토픽모델링 기법을 통해, 핀테크 세부 기술을 추출하고 정의한다. 도출된 핀테크 세부 기술에 대해 **Hot&Cold topic** 을 파악하여 핀테크 기술의 트렌드를 파악한다. 또한 핀테크 산업의 주요 기술에 대한 주요 시장국인 미국, 한국, 중국의 기술개발 동향을 각각 분석한다. 마지막으로 핀테크 세부 기술 간 네트워크 분석을 통해 기술 간의 연계 관계를 살펴본다. 본 연구를 통해 파악된 핀테크 산업 기술 동향은 핀테크 관련분야의 정책 수립과 핀테크 관련 기업의 기술 전략 수립에 효과적으로 활용될 수 있을 것으로 기대된다.

Abstract Recently, based on Internet and mobile environments, the Fintech industry that fuses finance and IT together has been rapidly growing and Fintech services armed with simplicity and convenience have been leading the conversion of all financial services into online and mobile services. However, despite the rapid growth of the Fintech industry, few studies have classified Fintech technologies into detailed technologies, analyzed the technology development trends of major market countries, and supported technology planning. In this respect, using Fintech technological data in the form of unstructured data, the present study extracts and defines detailed Fintech technologies through the topic modeling technique. Thereafter, hot and cold topics of the derived detailed Fintech technologies are identified to determine the trend of Fintech technologies. In addition, the trends of technology development in the USA, South Korea, and China, which are major market countries for major Fintech industrial technologies, are analyzed. Finally, through the analyses of networks between detailed Fintech technologies, linkages between the technologies are examined. The trends of Fintech industrial technologies identified in the present study are expected to be effectively utilized for the establishment of policies in the area of the Fintech industry and Fintech related enterprises' establishment of technology strategies.

Keywords : Big Data, Fintech, Patent analysis, Text mining, Topic modeling

본 논문은 교육부 및 한국연구재단의 BK21 플러스 사업(고려대학교 세종융합 문제해석의 빅데이터 분석 사업팀)으로 지원된 연구임.
*Corresponding Author : Hong-Chul Lee(Korea Univ.)
Tel: +82-2-3290-3767 email: holee@korea.ac.kr
Received October 21, 2016 Revised (1st November 4, 2016, 2nd November 9, 2016)
Accepted November 10, 2016 Published November 30, 2016

- 고려대에서는 논문을 다 이렇게 쓰는건지 모르겠는데, 그냥 교수님꺼랑 주제만 다르고 나머지는 100% 똑같음
-



토픽모델링을 활용한 4차 산업혁명 분야의 국내 연구 동향 분석

A Study on the Research Trends in the Fourth Industrial Revolution in Korea Using Topic Modeling

김 지 영 (Gi Young Kim)*
노 동 조 (Dong-Jo Noh)**

초 록

4차 산업혁명이 등장함 이래로 산업 분야를 비롯한 다양한 분야에서 관련 연구들이 수행되었다. 본 연구에서는 4차 산업혁명에 대한 국내의 연구 동향을 분석하기 위하여 2016년 1월부터 2023년 8월까지 KCI에 수록된 2,115건의 논문을 대상으로 확장어 분석 및 LDA 알고리즘에 기반한 토픽모델링 분석을 실시하였다. 본 연구의 결과 첫째, 4차 산업혁명 관련 학술 논문이 많이 게재된 학술지는 디지털융복합연구, 인문사회 21, e-비즈니스연구, 학습자중심교과교육연구 등의 순이었다. 둘째, 토픽모델링 분석 결과, '인간과 인공지능', '데이터와 개인정보 관리', '교육과정의 변화', '기업의 변화와 혁신', '교육의 변화와 융복합', '융합예술과 콘텐츠', '정보화 기업의 정책과 대응의 7개 토픽이 선정되었다. 셋째, 4차 산업혁명과 관련한 공통 연구주제는 '교육과정의 변화', '인간과 인공지능', '콘텐츠와 콘텐츠기반, 공통 키워드는 '기업', '정보', '교육', '스마트' 등이 있다. 넷째, 연구 분야가 [2016-2019]에는 교육 분야의 주제가 상위에 등장했으나 후반기[2020-2023]에는 기업과 스마트, 디지털, 서비스 혁신에 관한 주제들의 상위로 나타났다. 다섯째, 연구 후반기로 가면서 연구 간 네트워크 보다 구체화되거나 세분화되는 경향을 보였다. 이러한 동향은 코로나 팬데믹 이후 4차 산업혁명 분야의 핵심 기술들이 다양한 산업 분야에 활용됨에 따른 발생하는 사회경제적 변화에 따른 것으로 해석된다. 본 연구의 결과는 4차 산업혁명 분야의 연구 동향 파악과 전략 수립 및 후속 연구에 유용한 정보를 제공할 수 있을 것으로 기대한다.

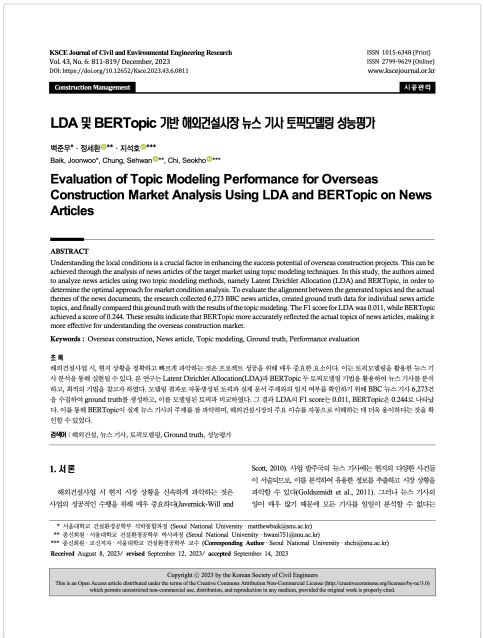
ABSTRACT

Since the advent of the Fourth Industrial Revolution, related studies have been conducted in various fields including industrial fields. In this study, to analyze domestic research trends on the Fourth Industrial Revolution, a keyword analysis and topic modeling analysis based on the LDA algorithm were conducted on 2,115 papers included in the KCI from January 2016 to August 2023. As a result of this study, first, the journals in which more than 30 academic papers related to the Fourth Industrial Revolution were published were digital convergence research, humanities society 21, e-business research, and learner-centered subject education research. Second, as a result of the topic modeling analysis, seven topics were selected: 'human and artificial intelligence,' 'data and personal information management,' 'curriculum change and innovation,' 'corporate change and innovation,' 'education change and jobs,' 'culture and arts and content,' and 'information and corporate policies and responses.' Third, common research topics related to the Fourth Industrial Revolution are 'change in the curriculum,' 'human and artificial intelligence,' and 'culture arts and content,' and common keywords include 'company,' 'information,' 'protection,' 'smart,' and 'system.' Fourth, in the first half of the research period (2016-2019), topics in the field of education appeared at the top, but in the second half (2020-2023), topics related to corporate, smart, digital, and service innovation appeared at the top. Fifth, research topics tended to become more specific or subdivided in the second half of the study. This trend is interpreted as a result of socioeconomic changes that occur as core technologies in the fourth industrial revolution are applied and utilized in various industrial fields after the corona pandemic. The results of this study are expected to provide useful information for identifying research trends in the field of the Fourth Industrial Revolution, establishing strategies, and subsequent research.

키워드: 토픽모델링, 4차 산업혁명, 연구 동향, 키워드 분석, LDA 알고리즘
Topic Modeling, Fourth Industrial Revolution, Research Trends, Keyword Analysis, LDA Algorithm

* 인천광역시도서관발전동행 차장(kimgy@imla.kr) (제1저자)
** 상명대학교 인문사회과학대학 문헌정보학과 교수(djnoh@smu.ac.kr) (교신저자)
논문접수일자 : 2023년 11월 22일 논문심사일자 : 2023년 11월 23일 게재확정일자 : 2023년 12월 1일
한국비블리아학회지, 34(4) : 207-234, 2023. http://dx.doi.org/10.14699/kbiblia.2023.34.4.207

요건 1번이랑 주제만 다르고 복붙 논문



KSCE Journal of Civil and Environmental Engineering Research
Vol. 43, No. 6: 811-819/ December, 2023
DOI: https://doi.org/10.12652/Ksce.2023.43.6.0811

ISSN 1015-6348 (Print)
ISSN 2799-9629 (Online)
www.kscejournal.or.kr

Construction Management
시공관리

LDA 및 BERTopic 기반 해외건설시장 뉴스 기사 토픽모델링 성능평가

백준우* · 정세환** · 지석호***
Baik, Joonwoo*, Chung, Sehwan**, Chi, Seokho***

Evaluation of Topic Modeling Performance for Overseas Construction Market Analysis Using LDA and BERTopic on News Articles

ABSTRACT

Understanding the local conditions is a crucial factor in enhancing the success potential of overseas construction projects. This can be achieved through the analysis of news articles of the target market using topic modeling techniques. In this study, the authors aimed to analyze news articles using two topic modeling methods, namely Latent Dirichlet Allocation (LDA) and BERTopic, in order to determine the optimal approach for market condition analysis. To evaluate the alignment between the generated topics and the actual themes of the news documents, the research collected 6,273 BBC news articles, created ground truth data for individual news article topics, and finally compared this ground truth with the results of the topic modeling. The F1 score for LDA was 0.011, while BERTopic achieved a score of 0.244. These results indicate that BERTopic more accurately reflected the actual topics of news articles, making it more effective for understanding the overseas construction market.

Keywords : Overseas construction, News article, Topic modeling, Ground truth, Performance evaluation

초 록

해외건설사업 시, 현지 상황을 정확하고 빠르게 파악하는 것은 프로젝트 성공을 위해 매우 중요한 요소이다. 이는 토픽모델링을 활용한 뉴스 기사 분석을 통해 실현될 수 있다. 본 연구는 Latent Dirichlet Allocation(LDA)과 BERTopic 두 모피모델링 기법을 활용하여 뉴스 기사를 분석하고, 최적의 기법을 찾고자 하였다. 모델링 결과로 자동생성된 토피과 실제 문서 주제라의 일치 여부를 확인하기 위해 BBC 뉴스 기사 6,273건을 수집하여 ground truth를 생성하고, 이를 모델링된 토피과 비교하였다. 그 결과 LDA의 F1 score는 0.011, BERTopic은 0.244로 나타났다. 이를 통해 BERTopic이 실제 뉴스 기사의 주제를 잘 파악하며, 해외건설시장의 주요 이슈를 자동으로 이해하는 데 더욱 유리하다는 것을 확인할 수 있었다.

검색어 : 해외건설, 뉴스 기사, 토피모델링, Ground truth, 성능평가

1. 서 론

해외건설사업 시 현지 시장 상황을 신속하게 파악하는 것은 사업의 성공적인 수행을 위해 매우 중요하다(Javernick-Will and Scott, 2010). 시장 밖주국의 뉴스 기사에는 현지의 다양한 사건들이 서술되므로, 이를 분석하여 유용한 정보를 추출하고 시장 상황을 파악할 수 있다(Goldszmidt et al., 2011). 그러나 뉴스 기사의 양이 매우 많기 때문에 모든 기사를 일일이 분석할 수 없다는

* 서울대학교 건설환경공학부 · 석사통합과정 (Seoul National University · matthewbaik@snu.ac.kr)
** 종신회원 · 서울대학교 건설환경공학부 · 박사과정 (Seoul National University · hwani751@snu.ac.kr)
*** 종신회원 · 교신저자 · 서울대학교 건설환경공학부 · 교수 (**Corresponding Author** · Seoul National University · shchi@snu.ac.kr)
Received August 8, 2023/ **revised** September 12, 2023/ **accepted** September 14, 2023

BERTopic
- HDBSCAN 알고리즘 적용 시, 파라미터를 조정함으로써 토픽 개수를 특정 개수로 설정하 거나, 알고리즘에 의해 최적의 토픽 개수를 자동으로 찾아낼 수 있다.
- 뉴스 기사 대상으로 LDA보다 훨씬 낫다는 것을 보여주는 논문.

BERTopic

자연어 처리 기술의 발달로, 22년 BERTopic이라는 방법이 새로 등장
LDA 방법의 우수성과 유영성은 다양한 분야에서 입증되었지만 통계기반으로 토픽을 산출하기 때문데 글의 맥락을 고려하지 못한다는 단점이 있다.

하지만, BERTopic은 임베딩 방식을 통해 조금 더 정확하게 토픽이 산출될 수 있을 것으로 기되됨. 그러나 BERTopic의 경우 LDA에 비해 적용 사례가 많이 않으므로 좀 더 다양한 분야에서 성능을 검증할 필요가 있다.

따라서 본 연구에서는 LDA와 BERTopic 간의 성능을 비교하고 BERTopic을 통해 인간-로봇 상호작용 연구의 토픽을 살펴봄으로써 현재까지의 연구 동향을 파악한다.

〈표 1〉 LDA와 BERTopic의 비교

Topic Model	Advantage	Disadvantage
LDA	- 비교적 간단하고 계산 비용이 저렴한 알고리즘임 - 워드 임베딩 기반 접근법에 비해 적은 수의 주제를 생성함 - 해석이 용이함 - 다수의 연구에서 검증된 방법임	- 단어주머니 방식을 사용하여 문맥을 고려하지 않음 - 대용량의 데이터셋에서의 성능이 보장되지 않음 - 너무 일반적인 주제나 단어를 추출하는 경향이 있음
BERTopic	- 워드 임베딩 방식을 사용하여 문맥을 고려하여 주제를 추출함 - 대용량의 데이터셋에 적합한 알고리즘임 - 단어의 희소성을 고려한 방법임	- 비교적 적은 용량의 데이터셋에서의 검증이 이루어지지 않음 - 해석이 용이하지 않음

이후 결과, (둘의 성능 비교, 토픽이 잘 형성되었는지 확인하기 위해 일관성과 혼잡도를 비교하는 방식.) 보편적인 토픽 모델의 평가 방식인가봄.(주제 추출)

우수하다고 나왔음. (이런 논문이 좀 많음.)

> BERTopic을 활용한 인간-로봇 상호작용 동향 연구

최근까지 LDA 가 가장 일반적으로 활용되는 토픽 모델링 방법론이었으나, 머신 러닝이 토픽모델링에 응용되기 시작하면서, 기존 방법의 단점을 보완하여 정확도를 높이려는 시도들이 이루어지고 있다. 대표적으로 2021년, BERT기반 임베딩 처리와 클래스 기반 TF-IDF등을 활용하여 일관된 토픽을 생성하는 BERTopic 기술이 제안되었는데, 기존 통계기반 토픽모델링 방법론들에 대비하여 높은 주제 일관성과 다양성을 보이는 것으로 확인되었다.

> BERTopic을 활용한 텍스트마이닝 기반 인공지능 반도체 기술 및 연구동향 분석 -특허와 논문 빅데이터를 중심으로

회의할 내용.

최근에 경영과학회, 산업공학회 토픽모델링 관련 주제로 갱장히 핫한게 BERTopic
기존의 LDA(Latent Dirichlet Allocation)와 같은 주제 모델링 기법에 비해 더 정확하고 세분화된 주제 분류를 가능하게 합니다.

1. 항만 특허 주제로 어떤 연구 질문을 선정할지. (= 연구목적을 분명히 하자.)
    1. 항만 특허 기술 관련 특허출원 동향 및 추이는 어떠한가?
    2. 항만 기술의 토픽 변화 양상?
    3. 아니면, 한국의 항만 특허도 수집해서 미국과 한국의 경쟁력을 비교해볼 수 도 있을듯
        1. 위 논문중에 KCI 논문 크롤링하는 프로그램을 사용해서 금방 뚝딱 크롤링 하는 것 같음.
2. BERTopic의 등장. LDA보다 우월하고, 최근 토픽모델링 연구 이걸로 많이 하는 듯.

- 항만 디지털화 특허 출원 동향 및 추이
    - HOT 토픽, COLD 토픽
    - GTM 활용 공백기술 뭔지?
- 항만 디지털화 기술의 토픽 변화 양상
-

연구 방법론에 대한 스케치 / 개념도

항만 관련 특허 데이터(약 3만5천여개)

- 목차

국가	출원번호	출원일자	공개번호	공개일자	공고번호	공고일자	등록번호	등록일자	발명의 명칭	IPC분류	출원인	발명자/고안자	대리인	요약	청구항	UPC	CPC 분류

1. 특허 데이터의 요약 및 청구항에 우리의 주제가 되는 '디지털화'관련 특허를 찾기 위해서
2. 키워드 선발기준 (- 찾아보고)
3. 3만 5천개의 특허에서 선발한 키워드를 통해 유효 특허를 분류한다.
4. 데이터 전처리 - 소문자화, 특수문자, 불용어, 숫자 등을 제거하는 작업.
5. 이후 LDA, BERTopic 각각 해보고,
6. 이후 둘 중 토픽 모델링 성능 평가
7. BERTopic 우수하다
8. 여기서 나온 토픽 사용해서
9. 시계열 분석 or GTM? or 을 통해 결론