Journal of Korean Society of Industrial and Systems Engineering
https://doi.org/10.11627/jksie.2023.46.2023.001

Vol. 46, No. 2023, Special Issue : 1-8, October 2023
ISSN : 2005-0461 / EISSN : 2287-7975

Text-Mining Analysis of Korea Government R&D Trends in Construction Machinery Domains

Bom Yun · Joonsoo Bae†

Department of Industrial and Information Systems Engineering, Jeonbuk National University

텍스트 마이닝을 통한 건설기계분야 국내 정부 R&D 연구동향 분석

윤 봄 · 배준수†

전북대학교 산업정보시스템공학과

To investigate the national science and technology policy direction in the field of construction machinery, an analysis was conducted on projects selected as national research and development (R&D) initiatives by the government. Assuming that the project titles contain key keywords, text mining was employed to substantiate this assumption. Project information data spanning nine years from 2014 to 2022 was collected through the National Science & Technology Information Service (NTIS). To observe changes over time, the years were divided into three-year sections. To analyze research trends efficiently, keywords were categorized into groups: ‘equipment,’ ‘smart,’ and ‘eco-friendly.’ Based on the collected data, keyword frequency analysis, N-gram analysis, and topic modeling were performed. The research findings indicate that domestic government R&D in the construction machinery field primarily focuses on smart-related research and development. Specifically, investments in monitoring systems and autonomous operation technologies are increasing. This study holds significance in analyzing objective research trends through the utilization of big data analysis techniques and is expected to contribute to future research and development planning, strategic formulation, and project management.

Keywords : Construction Equipment, Government R&D, Trend Analysis, Text Mining, Big Data

1. 서 론

건설기계는 건설공사에 사용되는 기계로 국내 「건설기계관리법 시행령」에서 굴삭기, 지게차 등 총 27종으로 분류하고 있다. 일반적으로 건설기계 산업은 경기변동에 민감하며 특히 국가별 건설 투자 및 인프라 정책 등에 따라 시장규모가 확대되는 특징이 있다. 현재 글로벌 건설시장은 각국의 코로나 펜데믹 이후 중단된 건설 인프라 개발이 재개되어 투자가 점진적으로 늘어날 것으로 전망되며, 주요국의 경기부양 정책, 원자재 채굴 수요 증가 등으로 연평균 5%대 성장이 기대된다. 한편 세계적으로 저출산, 고령화 등 인구구조 문제와 친환경화, 소형기종 중심 수요 확대 등 다양한 산업 트렌드에 맞춰 캐터필러, 고마츠 등 주요 선진사에서는 포트폴리오 다각화 및 작업자동화, 원격관리 등 ICT 기술 접목을 통해 제품군을 다양화하고 있다[2]. 정부에서는 국내 건설기계 산업 고도화를 위해 국가연구개발사업을 통해 연구개발을 지원하고 있다. 이에 본 연구에서는 향후 건설기계 분야 정부 R&D를 지원하고자 하는 사람들에게 의미 있는 정보를 제시하고자 한다.

Received 14 July 2023; Finally Revised 25 July 2023;
Accepted 25 July 2023
† Corresponding Author : jsbae@jbnu.ac.kr

이를 위해 건설기계 분야 국가연구개발사업으로 선정된 과제의 제목 데이터를 수집하여 빅데이터 분석 방법의 일환인 텍스트 마이닝을 활용하였다. 이 기법은 다량의 텍스트 데이터를 계량적으로 분석할 수 있다는 장점이 있어 다양한 분야에서 동향분석을 가능하게 한다. 키워드를 활용할 경우 미래 연구 트렌드를 파악하는데 연구 내용과 기술적인 측면을 잘 나타낼 수 있다는 장점이 있다[2, 3]. 본 연구에서는 제목에 핵심 키워드가 포함되어 있으며 이를 바탕으로 국가과학기술 정책 방향성을 파악할 수 있다고 가정한다.

건설기계 분야 국내 정부 R&D로 선정된 과제 데이터를 얻기 위해 '국가과학기술지식정보서비스(National Science & Technology Information Service, 이하 NTIS)'를 활용하였다. 해당 서비스를 통해 2014년부터 2022년까지 9년 동안 데이터를 수집하였고, 과제 시작연도와 제목 데이터를 바탕으로 텍스트 마이닝을 수행하였다. 분석에는 명사만 활용하였고, 시간의 흐름에 따른 변화를 확인하기 위해 3년 단위, 3개 구간으로 연도를 구분하였다. 연구 트렌드를 효율적으로 분석하기 위해 키워드를 '장비', '스마트', '친환경' 그룹으로 구분하였다. 연도 구간별 키워드 그룹의 빈도 수 및 관계 형성 변화 등을 확인하여 건설기계 분야 정부 R&D의 정책 동향을 분석하고자 한다. 본 연구에서는 분석을 위해 키워드 빈도분석, N-gram 분석, 토픽모델링 분석을 수행하였다. 키워드 빈도분석은 시간 흐름에 따른 키워드 그룹의 빈도 수 변화를 확인하기 위해 수행하였다. N-gram 분석을 통해 키워드 네트워크의 중심 척도를 바탕으로 키워드 간 상호 영향력을 파악하였다. 마지막으로 토픽모델링 분석을 통해 세부 연구토픽을 도출하고 각 토픽별 정부투자 추세를 확인하였다.

2. 이론적 배경 및 선행연구

2.1 텍스트 마이닝

Chakraborty et al.[5]에 따르면 세상에 존재하는 데이터의 80% 이상이 비정형 데이터로 추산된다고 한다. 텍스트 데이터는 대표적인 비정형 데이터이다. 텍스트 마이닝은 데이터 마이닝의 한 종류로 텍스트에 자연어 처리 방식을 더하여 다양한 정보를 얻을 수 있는 방법이다. 언어, 통계 등 여러 학문에 적용하여 목적에 맞는 유의미한 정보를 획득할 수 있다[1, 8]. 텍스트 데이터는 통계적으로 다양한 차원이 존재하므로 처리가 어렵다는 의견이 있으나, 정형화된 방법으로 주요 정보를 얻기 위해 수십 년간 관련 연구가 지속되어 왔다. 현재 형태소 분석, 의미연결망 분석, 토픽모델링 분석 등이 대표적으로 활용되고 있다[9].

2.2 동향분석 선행연구 사례

텍스트 마이닝은 공학, 금융, 사회과학 등 다양한 분야에서 연구 트렌드 파악을 위해 활용되고 있다. Deepak Sharma et al.[6]은 기계분야 상위 6개 저널에서 30년 동안 출판된 논문을 바탕으로 텍스트 마이닝 기법을 활용하여 머신러닝 분야 트렌드를 분석하였다[10]. Kim[9]은 금융 관련 시장흐름을 파악하기 위해 8년간 '핀테크'가 포함된 기사와 트위터 메시지 자료를 수집하였으며, 키워드 빈도분석, 토픽모델링, 감성분석 등을 수행하여 국내 핀테크를 비롯한 디지털 금융 서비스 트렌드를 파악하였다[11]. Lee[14]은 노인일자리 사업과 관련된 사회구조를 파악하기 위해 언론 기사를 수집하여 LDA 토픽모델링 분석을 수행함으로써 해당 분야의 질적 효과 및 고용 효과 등을 증진시킬 수 있는 방안을 제시하였다[12]. 텍스트 마이닝은 텍스트로 구성된 데이터라면 논문, 뉴스, SNS 등 다양한 자료에 적용할 수 있고, 데이터 양이 많더라도 객관적인 분석이 가능하다는 장점이 있어 연구 트렌드 또는 기술 동향 분석 등에 널리 활용되고 있다.

3. 데이터 분석방법

3.1 데이터 수집 및 전처리

데이터는 NTIS에서 건설기계분야 정부 R&D로 선정된 과제 정보를 취득하였다. 검색어를 선정하기 위해 「건설기계관리법 시행령」에서 정의한 27종의 명칭을 활용하였다. 그러나 일부 기종의 경우 다양한 명칭으로 불리거나 특정 조건이 만족되는 경우에만 건설기계로 분류되는 경우가 있어 일부 기종 명칭을 <Table 1>과 같이 추가 또는 수정하였다.

<Table 1> Search Word for Data Acquisition

불도저	굴착기	굴삭기
로더	지게차	스크레이퍼
덤프트럭	기중기	크레인
모터그레이더	건설기계롤러	노상안정기
콘크리트 뱃칭플랜트	콘크리트 피니셔	콘크리트 살포기
콘크리트 믹서트럭	콘크리트펌프	아스팔트 믹싱플랜트
아스팔트 피니셔	아스팔트 살포기	골재살포기
쇄석기	크러셔	건설공기 압축기
건설천공기	유압브레이커	항타기
항발기	건설자갈 채취기	준설선
준설장비	특수건설기계	타워크레인
건설기계		

전처리 작업을 위해 데이터 정제, 토큰화, 불용어 및 유의어 제거 3가지 단계를 수행하였다. 데이터 정제를 위해 실제 분석에 사용할 자료 이외 중복데이터를 제거하였다. 국가과학기술표준분류를 확인했을 때 '건강', '교육 서비스업' 등 건설기계와 거리가 먼 17개 내역을 삭제하여 3,428개 데이터를 확보하였다. 다음으로 토큰화를 진행하였는데, 이는 말뭉치(Corpus)에서 토큰(Token)이라 하는 단위로 나누는 작업이며, 토큰은 문법적으로 더 이상 나눌 수 없는 언어요소를 의미한다[13]. 한국어의 경우 형태소 기준으로 토큰화하는 것이 보편적이다[14]. 본 논문에서는 토큰화를 위해 대표적인 통계프로그램인 R의 KoNLP 패키지를 이용하여 명사와 외국어를 추출하였다. 이후 자주 등장하지만, 텍스트 데이터 분석을 하는 데 큰 도움이 되지 않는 '연구', '기술', 'bar', 'kg' 등 연구개발을 의미하거나 단위를 나타내는 상투적인 단어들을 불용어로 선정하여 삭제하였다. 유의어 처리는 'data', '데이터', '데이타' 등 같은 의미이지만 다른 단어로 표현된 경우 같은 단어로 변경하는 것 위주로 수행하였다. 전처리 완료된 데이터를 바탕으로 2차 토큰화를 수행하였고, 한 글자 단어와 빈도수 5 이하 단어를 제거하였다. 최종적으로 명사 토큰 5,953개를 확보하였다.

	Year	Number	Subject	word
1	2018	1425118988	SAND PLANT 용 하이망간주조합금 크러셔부품 내마모성 및 ...	내마모
2	2021	1711151389	톤급 파렛트형 물류 자동화용 병렬케이블로봇 개발	물류
3	2020	1425146084	(kg/hr, 기압)급 액화수소이송펌프 국산화 개발	국산화
4	2019	1425134059	"친환경 정밀제어(~ ppm) 방식의 이산화염소수기" 개발 및 ...	상용
5	2021	1415177136	. / kV 고전압 ℃ 내열 전기차용 평각(Rectangular) 케이블의 ...	케이블
6	2020	1425144836	급 중 안전장치, 중 내접기어 방식의 회전 퀵커플러	안전
7	2020	1425144836	급 중 안전장치, 중 내접기어 방식의 회전 퀵커플러	회전
8	2017	1415151808	. ~ Ton급 전동지게차의 에너지 효율 % 증가를 위한 kw급 PA...	전동지게차
9	2017	1415151808	. ~ Ton급 전동지게차의 에너지 효율 % 증가를 위한 kw급 PA...	에너지
10	2017	1415151808	. ~ Ton급 전동지게차의 에너지 효율 % 증가를 위한 kw급 PA...	모터

Showing 1 to 10 of 5,953 entries, 4 total columns

<Figure 1> Preprocessed Data (partial)

3.2 키워드 빈도분석

키워드 빈도 분석은 텍스트 내의 키워드 빈도 수(Term Frequency, TF)를 도출하는 방법으로 시간에 따라 정리한다면 손쉽게 트렌드를 파악할 수 있는 기법이다[11]. 본 연구에서는 토큰화 된 데이터를 3년 단위로 1구간(2014~2016), 2구간(2017~2019), 3구간(2020~2022)으로 나누어 분석을 수행하였다. 이후 각 연도 구간별 상위 15개 키워드를 추출하여 '장비그룹', '스마트그룹', '친환경그룹'에 각각 할당하고 그룹별 시간 흐름에 따른 변화를 확인하였다.

3.3 N-gram 분석

N-gram 분석은 특정 단어가 문서 내에서 어떤 단어와 함께 쓰였는지 파악하여 해당 문서 내 단어 조합을 확인할 수 있어 네트워크 분석에 많이 활용되고 있는 방법이다[15]. 키워드를 서로 연결하여 네트워크를 구성하였을 때 어떤 키워드가 영향력이 있는지 알 수 있기 때문이다. 네트워크는 노드(node)와 링크(link)의 조합으로 표현한다. 이때 노드는 키워드를 의미하고 링크는 키워드 간 연결선을 의미한다. 네트워크 내 노드의 영향력 정도를 표현하는 다양한 방법이 존재하는데, 본 논문에서는 연결중심성(Degree Centrality)과 매개중심성(Betweenness Centrality)을 사용하였고, 다음 식 (1), (2)를 통해 수치화 할 수 있다[16, 23, 24].

$$C_D(v_i) = \frac{1}{N-1} d_i \tag{1}$$

$$C_B(v_i) = \sum_{j<k} \frac{g_{jk(v_i)}}{g_{jk}} \tag{2}$$

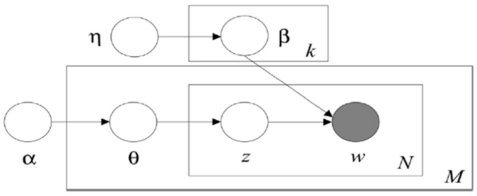
위 식에서 N은 노드의 수, d_i는 i번째 노드의 링크 수, g_{jk}는 두 노드 j와 k간의 최단경로 개수, $g_{jk}(v_i)$는 두 노드 j와 k간의 최단경로 중 노드i를 포함하는 경로 개수를 의미한다. 연결중심성은 연결이 가능한 최대 연결선 수에 대한 비율로 계산되며, 매개중심성은 특정 키워드가 다른 키워드를 경유하는 최단거리 경로 수 비율로 계산된다. 연결중심성을 통해 어떤 키워드가 지역 네트워크 수준에서의 영향력을 행사하는지 확인할 수 있다. 그러나 다른 그룹에까지 그 영향이 미친다고 해석하기엔 어려움이 있어 이를 보완하기 위해 매개중심성을 활용하였다. 매개중심성을 통해 소수의 연결만으로 다른 키워드의 매개체 역할을 수행하는 키워드를 찾을 수 있다[16, 17]. 본 연구에서는 각 연도 구간의 네트워크마다 연결중심성 및 매개중심성 상위 10개 키워드를 도출하였고, 해당 키워드를 3가지 그룹으로 분류하여 변화를 확인하였다.

3.4 토픽모델링 분석

토픽모델링 분석에는 잠재 디리클레 할당이라 불리는 LDA 모델을 사용하였다. 이 방법은 문서가 다양한 토픽들로 구성되어 있고 토픽은 단어들로 도출할 수 있다는 가정을 통해 텍스트 주제를 파악할 수 있는 확률모형이다[7, 18].

LDA 모델 아키텍처는 <Figure 2>와 같다. 사용자가 사전에 문서-토픽별 분포값(Θ)과 토픽 단어 분포값(β), 토픽 개수(k)를 하이퍼 파라미터로 입력하면 이를 바탕으로 단어(w)를 관측해 단어별로 적절한 토픽 번호(z)를 정하는

과정을 반복한다. 모든 z값 중 가장 높은 값에 키워드를 할당하는 방식으로 토픽별 단어 집합이 결정된다[19].

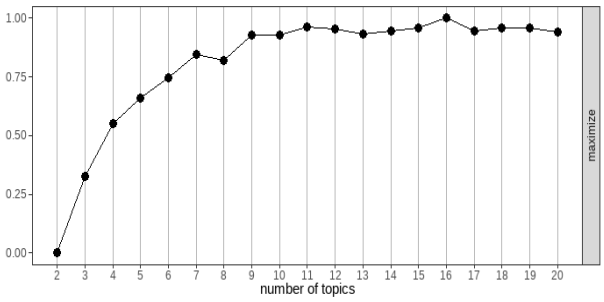


- M: The number of documents
- N: The number of words in the document
- W: Word
- Z: Topic number to which the word belongs
- k: The number of topics(Hyper parameter)
- α: θ value by document-topic
- η: β value by document-topic
- θ: Distribution of topics by document
- β: Distribution of words in a topic

<Figure 2> LDA Model Architecture

본 연구에서는 Θ와 β는 topicmodels 패키지의 lda() 함수에서 제공하는 기본값을 사용하였고, k값은 Griffiths2004 지수 알고리즘을 활용하여 결정하였다. 이 알고리즘은 텍스트 데이터를 바탕으로 토픽 수 k값을 2부터 20까지 변경하면서 해당 LDA 모델이 몇 개의 토픽을 가질 때 적합도가 높은지 그래프로 보여주며[25], 결과는 <Figure 3>과 같다.

Griffiths2004 알고리즘은 지수 알고리즘으로 값이 클수록 적합도가 높다는 것을 의미한다. 토픽 수 k가 9 이상일 때 적합도가 비슷한 경향이 있어 최적의 토픽 수 k는 9로 결정하였다. 이후 LDA 모델을 실행하여 각 토픽별 상위 5개 키워드를 도출하였다. 그러나 LDA 모델은 혼합모형(admixture model)으로 특정 키워드가 여러 토픽에 등장하는 경우가 있어 중복 키워드가 발생할 경우 확률값이 가장 높은 토픽에만 할당하였다[12]. 한편, LDA 모델을 통해 어떤 토픽에 어떤 문서가 할당되었는지도 함께 확인할 수 있다. 이를 바탕으로 연도 구간별 각 토픽의 문서 수 변화에 따라 해당 토픽의 추세를 파악하였다.



<Figure 3> Optimal Number of Topics

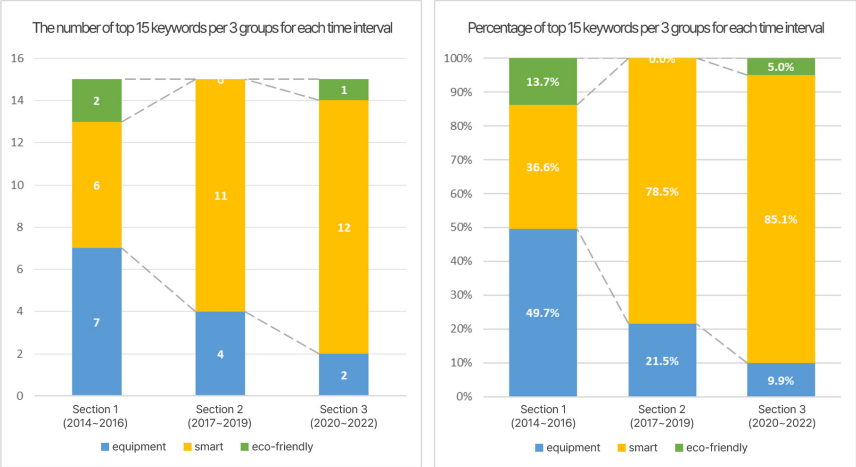
4. 연구 결과

4.1 키워드 빈도 분석

키워드 빈도 분석 결과 각 연도 구간별 상위 15개 키워드는 <Table 2>와 같으며, 시각화 결과는 <Figure 4>와 같다. 1구간에서 스마트 그룹 키워드의 빈도 수 합계 비율은 36.6%이나, 3구간에서는 85.1%로 시간이 흐를수록 폭발적으로 증가하고 있다. 반면에 장비그룹과 친환경 그룹 키워드의 빈도 수 합계는 1그룹에서 각각 49.7%, 13.7%이나, 3그룹에서는 9.9%, 5.0%로 감소하였다.

<Table 2> Top 15 Keywords Based on Term Frequency about 3 Time Sections

Rank	Section 1 (2014-2016)		Section 2 (2017-2019)		Section 3 (2020-2022)	
	Keyword(TF)	Group	Keyword(TF)	Group	Keyword(TF)	Group
1	굴착기(33)	equipment	스마트(61)	smart	스마트(104)	smart
2	국산화(27)	equipment	안전(37)	smart	안전(85)	smart
3	고효율(25)	equipment	IoT(36)	smart	인공지능(84)	smart
4	스마트(25)	smart	굴착기(33)	equipment	모니터링(43)	smart
5	안전(25)	smart	인공지능(32)	smart	굴착기(42)	equipment
6	친환경(24)	eco-friendly	최적(31)	smart	지능형(42)	smart
7	유압(20)	equipment	실시간(25)	smart	모델(41)	smart
8	크레인(20)	equipment	모니터링(23)	smart	로봇(40)	smart
9	에너지(19)	eco-friendly	모델(23)	smart	데이터(37)	smart
10	지능형(18)	smart	지능형(22)	smart	IoT(35)	smart
11	IoT(17)	smart	데이터(20)	smart	친환경(35)	eco-friendly
12	최적(17)	smart	크레인(20)	equipment	최적(32)	smart
13	신뢰성(16)	equipment	고효율(19)	equipment	실시간(29)	smart
14	제품(15)	equipment	로봇(19)	smart	예측(29)	smart
15	작업(13)	smart	센서(18)	equipment	센서(28)	equipment



<Figure 4> Trends in the Top 15 Keywords per 3 Groups for Each Time Section

4.2 N-gram 분석

연도 구간별 연결중심성 및 매개중심성 결과는 <Table 3>과 같다. 1구간 연결중심성 상위 키워드의 장비그룹, 스마트그룹, 친환경그룹 분포 수는 각각 4개, 3개, 3개이나 3구간에서는 1개, 9개, 0개로 스마트그룹의 영향력이 커지

<Table 3> Top 10 keywords in Degree Centrality and Betweenness Centrality

Time	Rank	Degree Centrality			Betweenness Centrality		
		keyword	Group	value	keyword	Group	value
Section 1	1	에너지	eco-friendly	8	IoT	smart	15.0000
	2	굴착기	equipment	6	안전	smart	11.0000
	3	IoT	smart	6	융합	smart	11.0000
	4	안전	smart	6	에너지	eco-friendly	9.0000
	5	크레인	equipment	6	고효율	equipment	4.0000
	6	융합	smart	6	굴착기	equipment	3.0000
	7	고효율	equipment	4	크레인	equipment	3.0000
	8	친환경	eco-friendly	2			
	9	어태치먼트	equipment	2			
	10	절감	eco-friendly	2			
Section 2	1	실시간	smart	14	IoT	smart	136.3333
	2	최적	smart	10	스마트	smart	126.8333
	3	IoT	smart	10	최적	smart	108.83333
	4	스마트	smart	10	인공지능	smart	108.3333
	5	인공지능	smart	10	실시간	smart	84.3333
	6	굴착기	equipment	8	센서	equipment	66.0000
	7	모니터링	smart	8	모니터링	smart	52.3333
	8	해상	equipment	6	융합	smart	46.0000
	9	저소음	eco-friendly	6	에너지	eco-friendly	24.0000
	10	에너지	eco-friendly	4	안전	smart	24.0000
Section 3	1	인공지능	smart	36	인공지능	smart	562.9601
	2	스마트	smart	34	스마트	smart	401.5918
	3	안전	smart	30	로봇	smart	328.3453
	4	모니터링	smart	22	안전	smart	277.5211
	5	로봇	smart	22	모니터링	smart	173.8971
	6	모델	smart	16	예측	smart	125.6310
	7	데이터	smart	16	모델	smart	110.7571
	8	예측	smart	14	데이터	smart	106.5380
	9	IoT	smart	12	디지털	smart	103.0000
	10	센서	equipment	10	최적	smart	75.4079

고 있다. 각 그룹별 매개중심성 상위 10개 키워드 변화에서도 1구간 장비그룹 3개, 스마트그룹 3개, 친환경그룹 1개에서 3구간에서는 10개 모두 스마트그룹의 키워드로 구성되어 있다. 1구간에서 연결중심성 1~3위 키워드는 각각 '에너지', '굴착기', 'IoT'이나 매개중심성에서는 'IoT', '안전', '융합'이다. 1구간의 지역 네트워크에서는 친환경그룹, 장비그룹 키워드의 영향력이 높으나 지역 네트워크를 연결하는 매개역할 최상위권 키워드는 모두 스마트그룹 키워드이다. 이는 1구간의 개별적인 지역 네트워크에서는 스마트그룹의 영향력이 두드러지지는 않으나 각 네트워크를 연결하는 주요 매개체 역할을 수행한다는 의미로 해석할 수 있다.

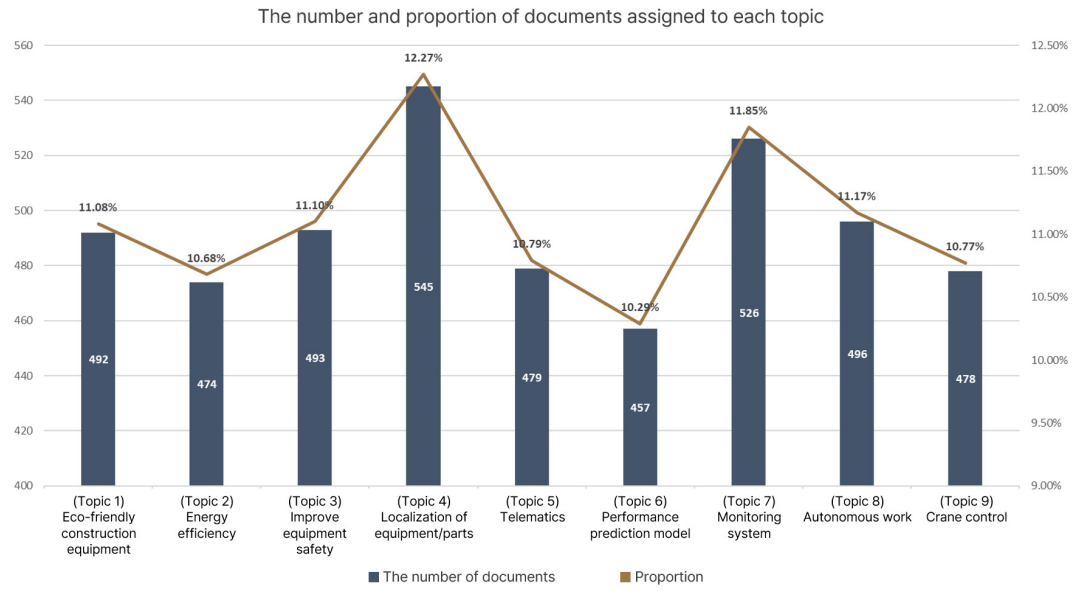
4.3 토픽모델링 분석

토픽모델링 분석을 위해 토픽 수 k는 9로 결정하여 이를 바탕으로 LDA 모델을 실행하였다. 확률값 beta에 따라 각 토픽별 상위 5개 키워드를 할당하고 토픽 제목을 부여하였으며 상세결과는 <Table 4>와 같다. LDA 모델은 토픽별 키워드를 할당할 뿐만 아니라 gamma 값을 통해 원본 문서도 토픽별로 분류할 수 있다. 토픽별 할당된 문서 수를 바탕으로 국가연구개발사업으로 어떤 토픽이 많은 투자를 받았는지 알 수 있다. 또한 시간 흐름에 따라 토픽별 문서 수를 비교한다면, 해당 토픽의 R&D 투자 동향을 파악할 수 있다. <Figure 5>는 토픽별 할당된 문서 수를 나타낸 그래프이며, <Figure 6>은 토픽별 연도 구간에 따른 문서 수 변화를 나타낸 그래프이다.

문서 수가 가장 많은 토픽은 Topic4(장비/부품 국산화)이며, 두 번째는 Topic7(모니터링 시스템), 세 번째는 Topic8(자율작업)이다. 최근 9년간 건설기계 분야 국가연구개발사업에서 '장비/부품 국산화', '모니터링 시스템', '자율작업'이 가장 많은 비율을 차지한다. 반면에 문서 수

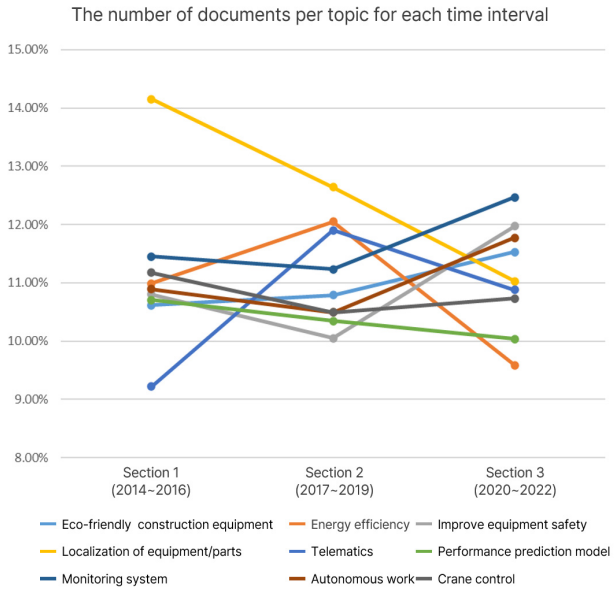
<Table 4> Results of LDA Analysis and the top 5 Keywords per Topic

Topic	Title	Keyword(beta)
Topic1	Eco-friendly construction equipment	친환경(0.0779), 데이터(0.0549), 환경(0.0405), 선박(0.0304), 고도화(0.0290)
Topic2	Energy efficiency	지능형(0.1029), 고효율(0.0836), 에너지(0.0642), 저감(0.0582), 최적(0.0404)
Topic3	Improve equipment safety	안전(0.1476), 차세대(0.0313), 대형(0.0299), 플랜트(0.0285), 진동(0.0271)
Topic4	Localization of equipment/parts	굴착기(0.1392), 국산화(0.0855), 유압(0.0690), 현장(0.0332), 경량(0.0332)
Topic5	Telematics	IoT(0.0683), 실시간(0.0564), 하이브리드(0.0461), 센서(0.0431), 건축물(0.0416)
Topic6	Performance prediction model	예측(0.0598), 로봇(0.0553), 제조(0.0445), 생산성(0.0430), 모델(0.0400)
Topic7	Monitoring system	스마트(0.1540), 모니터링(0.0951), 건설(0.0416), 빅데이터(0.0389), 솔루션(0.0282)
Topic8	Autonomous work	인공지능(0.1096), 융합(0.0727), 작업(0.0542), 신뢰성(0.0400), 자율주행(0.0385)
Topic9	Crane control	크레인(0.0366), 자동화(0.0366), 서비스(0.0337), 건설현장(0.0322), 제어(0.0293)



<Figure 5> The Number and Proportion of Documents Assigned to Each Topic

하위 3개 토픽은 Topic6(성능 예측 모델), Topic2(에너지 효율), Topic9(크레인 제어)로 정부 투자가 다소 적다. 토픽별 연도구간에 따른 추세를 살펴보면, Topic4는 1구간에서 1위를 차지하나, 3구간에서는 5위에 위치한다. '장비/부품 국산화'는 가장 많은 투자를 받은 토픽이나, 점차 감소하는 추세이다. Topic7의 경우 1구간에서 2위, 3구간에서는 1위를 차지하였다. Topic8은 1구간 5위에서 3구간 3위로 순위가 상승하였다. Topic3(장비 안전성 향상)의 경우 1~2구간에서 하위권에 위치하나, 3구간에서 2위를 차지하고 있다. 3구간에서 1~3위에 위치한 토픽으로 미루어보아 최근 건설기계분야 국가연구개발사업은 '모니터링 시스템', '장비 안전성 향상', '자율작업'을 위주로 수행되고 있으며, 이는 개별장비 단위의 스마트 기술과 관련된 연구개발 토픽이라 할 수 있다.



<Figure 6> The Number of Documents per Topic for Each Time Section

5. 결 론

본 연구는 건설기계 분야 국가연구개발사업의 연구 동향을 객관적으로 분석하기 위해 텍스트 마이닝을 활용하였다. 연구동향 분석을 위한 기존 방식은 대부분 전문가 지식을 기반으로 한 정성적인 평가가 대부분이었다. 이 방법은 평가자에 따라 주관적인 의견이 반영되므로 객관성을 확보하기에 어려움이 있으며, 일반 사용자의 접근성이 떨어질 수 있다는 단점이 있다. 이를 보완하기 위해 본 연구에서는 빅데이터 분석 방법 중 하나인 텍스트 마이닝 기법을 활용하였고, 누구나 접근할 수 있는 정보를 기반으로 객관적인 연구 트렌드 분석 결과를 도출했다는 데에 의의가 있다. 아울러 본 연구는 전략기획 등 기술정책 수립, 연구동향 조사에 활용될 수 있을 것으로 기대한다. 또한 과학적인 방법을 사용하여 트렌드를 분석하였으므로 다른 분야 데이터를 적용하여도 유의미한 정보를 얻을 수 있을 것이다. 연구 수행을 위한 자료 수집을 위해 NTIS를 통해 최근 9년간 건설기계 분야의 국가연구개발사업 제목 데이터를 취득하여 명사를 추출하였고, 키워드 빈도분석, N-gram 분석, 토픽모델링 분석을 수행하였다. 키워드 빈도 분석 및 N-gram 분석 결과를 바탕으로 시간이 흐를수록 정부에서는 스마트 관련 기술개발에 집중 투자를 하고 있음을 확인하였다. 토픽모델링 분석 결과 얼마 전까지는 국산화 기술개발에 많은 투자를 하고 있었지만, 최근에는 '모니터링 시스템', '장비 안전성 향상', '자율작업' 관련 연구개발을 주로 수행하고 있으며, 글로벌 환경 이슈에 따라 친환경 건설기계와 관련된 연구개발도 꾸준히 진행 중임을 확인하였다. 특히 모니터링, 안전성 향상, 자율작업 기술은 대표적인 건설기계 지능화 기술로 평가받고 있다. 이 기술에 대한 지속적인 연구개발이 이뤄진다면, BIM, Fleet Management 등 건설현장 무인화와 관련된 기술개발이 가능할 것으로 사료된다. 이를 이룩하기 위해 국토교통부, 산업통상자원부 등 정부 부처에서는 생산성 및 안전성 향상, Fleet Management, 디지털 트윈 등 건설기계 지능화와 무인화를 위한 로드맵 수립을 통해 연구개발 투자 계획을 마련하였다[2, 20]. 해외 주요 선진업체에서는 원격 모니터링 및 제어 솔루션, 건설장비 유지관리 서비스 시스템, 자동측정 및 자율작업 등이 가능한 건설기계를 출시하고 있다[21, 22]. 향후 세계시장에서 제품 경쟁력을 확보할 수 있도록 정부의 적극적인 지원이 필요하다.

References

[1] Adeva, J.J.G. and Calvo, R.A., Mining Text with Piminto, *IEEE Internet Computing*, 2006, Vol.10, No.4, pp. 27-35.

[2] Ahn, S.J. and Yoo, W.J., Introduction to Natural Language Processing Using Deep Learning. WikiDocs (Webbook). https://wikidocs.net/21698. (Accessed: Jul 9, 2023.)

[3] Blei, D.M., Ng, A.Y., and Jordan, M.I., Latent Dirichlet Allocation, *Journal of Machine Learning Research*, 2003, Vol. 3, pp. 993-1022.

[4] Centrality, https://bigdata-analyst.tistory.com/319 (Accessed: Jul 12, 2023.).

[5] Chakraborty, G., Pagolu, M., and Garla, S., Text Mining and Analysis: Practical Methods, Examples, and Case Studies Using SAS, SAS Institute, 2014.

[6] Deepak Sharma, Bijendra Kumar, Satish Chand. 2018. “Trend Analysis in Machine Learning Research Using Text Mining.” International Confernece on Advances in Computing, Comminication Control and Networking (ICACCCN2018).

[7] Gokhberg, L., Kuzminov, I., Khabirova, E., and Thurner, T., Advanced text-mining for trend analysis of Russia’s Extractive Industries, Futures, 2020, Vol.115.

[8] Griffiths, T.L. and Steyvers, M., Finding scientific topics, *Proceedings of the National Academy of Sciences*, 2004, Vol. 101, pp. 5228-5235.

[9] Kim, D.H., Analysis of FinTech and Digital Financial Service Trends in the New Normal Era through Text Mining. [dissertation], [Seoul, Koea]: Sookmyung Women's University, 2021

[10] Kim, D.H., How text is analyzed, Statistics Korea, Statistics Training Institute, 2017.

[11] Kim, H.W. and Song, J.W., Analysis of Quantum Mechanics Related Articles Using Topic Modeling: A Comparison of LDA, CTM, and STM Models, *New Physics: Sae Mulli*, 2020, Vol. 70, No. 5, pp. 493-502.

[12] Kim, J.H. and Kim, S.S., A Study on the Analysis of Agricultural R&D Keywords Using Textmining Method, *Journal of the Korea Academia-Industrial Cooperation Society*, 2021, Vol.22, No.2, pp. 721-732.

[13] Kim, Y.H., Social Network Analysis, Park Youngsa, 2016.

[14] Kim, Y.H., Understanding and Application of Social Network Analysis Techniques: Network Structure, Clustering, and QAP, *Korea Institute of Public Administration*, 2020, Vol. 34, pp. 58-68.

[15] Lee, D.W., New Horizons for the Construction Machinery Industry, *Auto Journal*, 2021, pp. 39-42.

[16] Lee, S.J., Topic Analysis of Newspaper Articles on ‘Elderly Employment’ Using Latent Dirichlet Allocation, *Journal of Digital Convergence*, 2020, Vol.18, No.10, pp. 537-546.

[17] Lee, S.S., Yoo, I.H., and Kim, J.H., Social Perception of AI Education through Big Data Analysis: Focus on News Articles and Twitter, *Korean Society for Digital Policy and Management*, 2020, Vol. 18, No. 6, pp. 9-16.

[18] Lee, Y.S., A Study on Serviceization Research Topics Using Ontology and Text Mining. [dissertation], [Seoul, Koea]: Konkuk University, 2016.

[19] Lim, S.Y. and Kim, S., Analysis of Research Trends in Construction Automation Based on Text Mining Techniques, *Korean Journal of Construction Engineering and Management*, 2016, Vol. 17, No. 6, pp.13-23.

[20] Ministry of Land, Infrastructure and Transport, Basic Plan for Construction Technology Promotion, MOLIT, 2017.

[21] Ministry of Trade Industry and Energy, Korea Institute of Machinery & Materials, Machine industry development plan, MOTIE, 2022.

[22] Network Centrality Value – Degree, Betweenness, Eigenvector, Closeness. https://m.blog.naver.com/PostView.naver?blogId=applwoods&logNo=222318849314&categoryNo=60&proxyReferer= (Accessed: Jul 12, 2023.).

[23] Oh, C.S., Lee, Y.T., and Ko, M.S., Establishment of ITS Policy Issues Investigation Method in the Road Section applied Textmining, *The Journal of The Korea Institute of Intelligent Transportation Systems*, 2016, Vol.15, No.6, pp.10-23.

[24] Oh, S.H., Current Status and Future Directions of the Construction Machinery Industry, *KEIT Industry Economy*, 2023, pp.67-74.

[25] Woo, C.W. and Lee, J.Y., Exploring the Key Research Topics and Trends in the ICT Field through LDA Topic Modeling, *Journal of the Korea Convergence Society*, 2020, Vol. 11, No. 7, pp. 9-18.

ORCID

Bom Yun | http://orcid.org/0009-0002-8828-4361

Joonsoo Bae | http://orcid.org/0000-0001-8872-5169