빅데이터 기반의 철도차량 분야 사회동향 분석

백승현*, 김춘수*, 이호규*, 박대식*, 진영권**
*한국철도기술연구원 기획조정본부
**국가과학기술연구회 감사기획부
e-mail:baeksh@krri.re.kr

An analysis of social trends based on big data in railway vehicle

Sunghyun Baek*, Choonsu Kim*, Hogyu Lee*, Daeshik Park*, Yeongkwon Jin**
*Planning and Coordination Division, Korea Railroad Research Institute
**Audit Planning Division, National Research Council of Science & Technology

요 약

이 연구는 공공재로서의 철도차량의 이용 고객인 국민의 시각이 반영된 포털사이트의 뉴스, SNS 텍스트를 활용하여 분석하고, 철도전문가의 시각이 반영된 학술논문 정보 분석결과를 함께 분석하여, 철도차량 분야에 대한 사회적인 관심과 동향을 도출하는데 그 목적이 있다. 포털사이트 분석결과 '교통', '지하철', '고속열차', '안전' 등 수요자 중심의 철도차량 검색어 빈도가 높게 나타났고, 학술논문정보 분석결과 '시스템', '안전', '교통' 등의 검색어 빈도가 높게 나타났다. '안전', '사고'에 대해서는 공통으로 높은 관심을 확인하였으며, 포털사이트에서는 '수소', '연결' 등 친환경 신기술 및 교통편의 제고를 위한 이슈에도 높은 관심이 확인되었고, 학술논문정보에서는 '데이터', '예측' 등에 대한 높은 관심이 확인되었다. 이를 통해 철도안전, 사고예방, 친환경 철도 신기술, 교통체계 연결, 데이터 중심의 예측기술 등에 대한 향후 정책·기술수요를 도출할 수 있었다. 담론분석과 토픽분석 결과, 포털사이트 정보에서는 철도차량 유형 및 수요를 통한 영향(역세권, 개통 등)을 중심으로 높은 관심이 나타났고, 학술논문 정보에서는 철도차량을 중심으로 하는 연구분야와 분석대상에 대한 관심이 확인되었다.

1. 서론

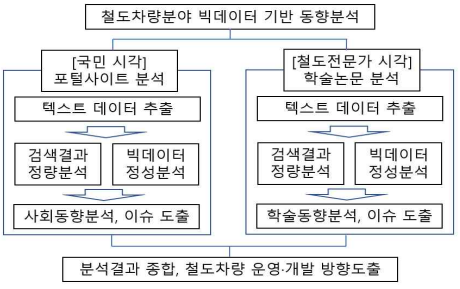
철도차량은 고속철도, 일반철도, 도시철도 등으로 국민에게 제공되는 교통수단으로, 정부 예산을 투입하여 공공재를 도입하고 운영하며 기술개발하는 하는 분야이다. 이 연구는 철도차량 분야를 대상으로 사회전반의 빅데이터에 기반한 국민체감 여론 동향분석, 학술논문 빅데이터에 기반한 기술동향 분석으로, 공공재로서의 철도차량 운영·개발의 전략을 수립하고 방향을 정립하는 데 기여하는 것을 목적으로 한다.

기존 특허, 논문 기반의 기술동향 분석[1]은 많이 수행되어 왔지만, 주요 포털사이트의 뉴스, SNS 등을 활용한 빅데이터 기반의 질적 분석이 철도분야에서 시행된 사례가 거의 없었다. 따라서, 이 연구를 통해 사회전반의 철도차량에 대한 시각과 요구사항, 학술논문에서의 기술동향 분석을 함께 분석하여, 일반 국민과 전문가의 시각이 균형있게 반영되는 철도차량 운영·개발 전략수립에 기여하고자 한다.

2. 분석의 틀 및 연구설계

2.1 분석의 틀

이 연구에서 분석하고자 하는 연구주제에 대한 분석의 틀은 [그림 1]과 같다.



[그림 1] 분석의 틀

이 연구는 공공재로서의 철도차량의 최종 서비스 수혜자인 국민의 시각을 포털사이트에서의 뉴스, SNS 텍스트를 활용하여 분석하고, 전문가 학술논문 정보 분석결과를 함께 반영하였다는 점에서 그 의의가 크다 하겠다. 이를 통해 국민과 전문가의 여론동향을 함께 반영하는 정책수립이 가능할 것으로 기대된다.

2.2 분석대상 및 자료의 수집

이 연구는 한국철도기술연구원 2022년 자체연구사업으로 수행하는 '지속가능 철도연구 운영을 위한 정책동향 및 성과체계 분석' 연구과제 내용 중 '빅데이터기반 철도과학기술정책 사회동향 조사분석[2]'에서, 철도차량 분야에 국한하여 자료를 분석하였다.

세부적으로, 철도차량 분야 중 '고속철도, 도시철도, 경전철, 트램, 철도동역학, 열차기술, 고속열차, 철도차량기술' 등을 주요 검색어로 사용하였고, 시범 검색결과를 토대로 중복성과 유의미성을 판단하여 핵심 키워드를 '고속열차, 도시철도, 경전철, 트램, 열차+기술'로 한정하였다.

자료의 수집을 위한 범위와 내용은 [표 1]과 같다. 수집 대상기간은 2019년 1월 1일부터 2021년 12월 31일까지 3년으로 하였다.

[표 1] 자료 수집방법

구 분	수집 채널		수집 내용
[국민시각 분석자료] 포털사이트 자료수집	네이버	뉴스	제목, 본문, URL, 날짜
		블로그	
		카페	
	다음	뉴스	
		블로그	
		카페	
철도전문가 시각 분석자료] 학술논문 분석	RISS	국내학술 논문	제목, 발행연도, 주제어, 국문초록

2.3 분석도구 및 분석방법

이 연구는 포털사이트와 학술논문정보의 비정형 텍스트의 분석을 수행함에 따라, 이를 위한 전문 분석도구를 활용하였고, [표2]와 같다.

[표 2] 분석도구

도구	활용
텍스톰 (TEXTOM)	- 웹 비정형 데이터 수집, 정제, 분석 - 워드클라우드, 네트워크, 담론분석, 토픽모델링 등 시각화 자료 생성

이 연구에서는 포털사이트와 학술정보의 검색자료를 정량적 정보분석과 정성적 질적분석으로 구분하여 분석하였고, 그 세부내용은 [표3]과 같다.

[표 3] 분석방법

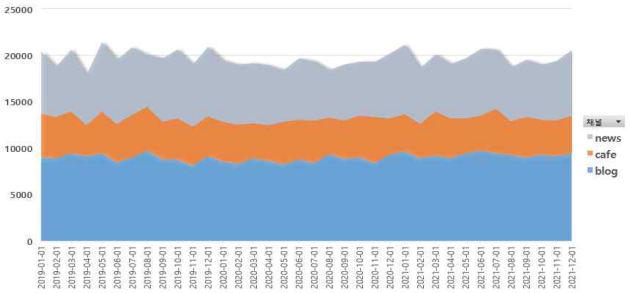
구분	분석내용
정량분석	- 정보량 및 검색량 분석을 통한 국민 인지도와 관심도 파악, 분야별·연도별 검색결과 분석
질적분석	- 텍스트마이닝 분석, 워드클라우드 시각화[3]
	- 담론분석 : 상관관계 분석을 반복적으로 실시하여 적정한 수준의 유사성 집단을 찾아내는 블록 모델링(blocking modeling) - 상위빈도 40개 단어를 토대로 분석
	- LDA 토픽모델링 : 데이터 마이닝 기법 중의 하나로, 비구조화된 텍스트 자료들의 뭉치로부터 의미있는 주제(토픽)들을 추출해주는 확률모델 알고리즘[4] - 상위빈도 40개 단어를 토대로 분석
	- N-GRAM 분석 : n개 단어의 연쇄를 의미함. 만일 특정 단어 A와 B의 N-GRAM 빈도가 높다는 것은 두 단어가 함께 등장하는(공출현) 빈도가 높음을 의미 - 공출현 빈도순 50개 단어를 토대로 분석

3. 분석결과

3.1 포털사이트 데이터 분석결과

3.1.1 정보량 및 검색량 정량 분석결과

철도차량 분야에 대해 2019~2021년간의 검색 데이터량은 총 706,517건이며, 채널별 데이터량 추이는 [그림2]와 같다. 블로그, 뉴스, 카페 순으로 검색 데이터량이 도출되었다.



[그림 2] 철도차량 분야 채널별 데이터량 추이(포털데이터)

철도차량 관련 데이터를 바탕으로 포털사이트에서 등장한 키워드의 빈도를 분석한 결과, '교통'이 86,198건으로 가장 많이 등장하였으며 '지하철'(60,888건), '버스'(51,539건), '고속열차'(50,839건), '역세권'(43,592건) 순으로 빈도가 높은 것으로 확인되었다. 연도별 상위빈도 순위는 [표4]와 같다.

[표 4] 철도차량 분야 연도별 키워드 분석(빈도 상위 10개)

2019년			2020년			2021년		
순위	단어	빈도	순위	단어	빈도	순위	단어	빈도
1	교통	27987	1	교통	28176	1	교통	30035
2	지하철	21489	2	지하철	21225	2	지하철	18174
3	버스	20232	3	버스	16532	3	고속열차	17962
4	고속열차	19356	4	역세권	15489	4	수소	16297
5	고속	13382	5	고속열차	13519	5	연결	15450
6	역세권	13146	6	연결	13010	6	역세권	14959
7	안전	12945	7	KTX	11352	7	버스	14775
8	연결	11994	8	수소	9851	8	KTX	12641
9	KTX	11433	9	고속	9828	9	고속	11896
10	대중교통	8884	10	안전	9268	10	안전	10468

'교통', '지하철', '고속열차', '안전'에 대한 국민적 관심은 꾸준히 상위빈도를 나타내며 지속되는 것으로 확인되었다. 2019~2021년 중 '수소', '연결'에 대한 관심이 지속 증가하는 것을 확인할 수 있으며, 이를 통해 철도차량에 대한 새로운 수소기술의 적용, 철도를 포함한 교통체계 연결에 대한

국민적 관심이 계속 증가하는 것을 확인할 수 있다.

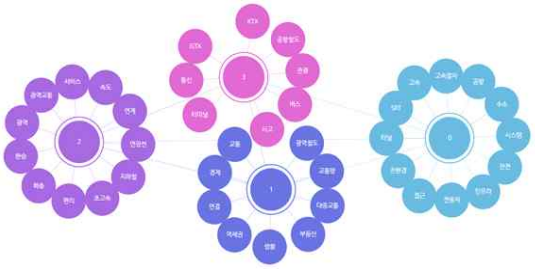
3.1.2 텍스트마이닝 분석결과

철도차량 분야에 대한 포털사이트 검색 데이터를 워드클라우드를 활용하여 시각화하면 [그림3]과 같다. 검색 상위빈도 키워드를 중심으로 분포하는 것을 확인할 수 있다.

[그림 3] 철도차량 분야 워드클라우드 (포털데이터)

3.1.3 담론분석 결과

철도차량 분야에 대한 포털사이트 검색 데이터를 담론 시각화하면 [그림4]와 같고, 4개의 그룹에 대한 주요 단어를 집약하면 [표5]와 같다.



[그림 4] 철도차량 분야 담론 시각화 (포털데이터)

[표 5] 철도차량 분야 담론 그룹 (포털데이터)

구분	그룹명	주요 단어
그룹(0)	기술 체계	고속열차, 시스템, 전동차, 수소, 친환경 등
그룹(1)	생활 편의	교통망, 대중교통, 부동산, 역세권, 생활, 경제 등
그룹(2)	교통망확충	광역교통, 환승, 확충, 지하철, 연장선, 서비스 등
그룹(3)	연계 교통	KTX, GTX, 공항철도, 버스, 터미널 등

3.1.4 토필모델링 분석 결과

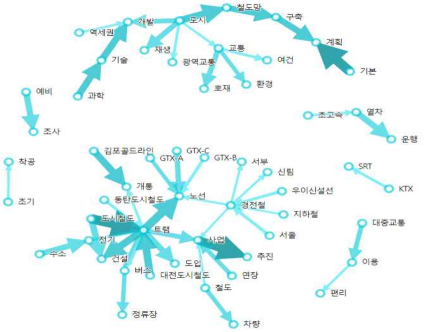
철도차량 분야에 대한 포털사이트 검색 데이터를 주요 토픽으로 집약하면 [표6]과 같다.

[표 6] 철도차량 분야 주요 토픽 (포털데이터)

토픽		주요 키워드
1	부동산 관련 토픽 (λ = 0.5 / 28%)	예정, 개통, 분양, 노선, 아파트 등
2	트램 이용 토픽 (λ = 0.5 / 22.3%)	트램, 여행, 버스, 도착, 부산 등
3	철도망 구축 사업 토픽 (λ = 0.5 / 32.9%)	사업, 도시철도, 철도, 도시, 건설 등
4	차량 기술 토픽 (λ = 0.5 / 16.8%)	기술, 열차, 고속열차, KTX, 고속 등

3.1.5 네트워크 분석(N-GRAM) 결과

철도차량 분야에 대한 포털사이트 검색 데이터를 주요 토픽으로 집약하면 [그림5]와 같다. 도시철도-트램, 트램-노선, 도시-철도망, 트램-건설, 대전도시철도-트램 등과 같이 상호 연계된 네트워크로 선후관계가 연계됨을 확인할 수 있다.



[그림 5] 철도차량 분야 N-GRAM (포털데이터)

3.2 학술정보 데이터 분석결과

3.2.1 정보량 및 검색량 정량 분석결과

철도차량 분야에 대해 2019~2021년간의 국내 학술논문 RISS 검색 데이터량은 총 916건이며, '시스템', '안전', '교통' 등의 키워드가 상위에 랭크되어 있었다. 포털사이트와 유사하지만, '사고', '데이터', '예측', '속도' 등 학술적 연구주제와 연계된 단어가 상위 10개에 포함된 것이 특징적이다.

[표 7] 철도차량 분야 빈도순 상위 10개 단어 (학술논문정보)

순위	단어	빈도	TF-IDF(중요도)
1	시스템	451	755.7
2	안전	387	745.1
3	교통	370	705.3
4	경제	262	594.3
5	궤도	256	707.6
6	서비스	229	584.0
7	사고	195	563.8
8	데이터	184	483.0
9	예측	163	434.4
10	속도	159	419.5

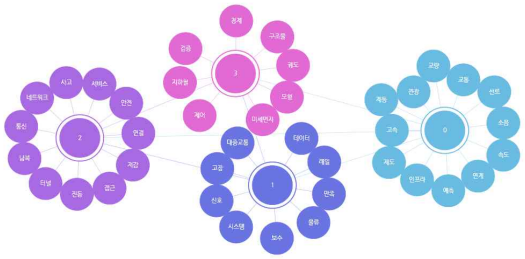
3.2.2 텍스트마이닝 분석결과

철도차량 분야에 대한 학술논문정보 검색 데이터를 워드클라우드를 활용하여 시각화하면 [그림6]과 같다. 검색 상위빈도 키워드를 중심으로 분포하는 것을 확인할 수 있다.

[그림 6] 철도차량 분야 워드클라우드 (학술논문정보)

3.2.3 담론분석 결과

철도차량 분야에 대한 학술논문정보 검색 데이터를 담론 시각화하면 [그림7]와 같고, 4개의 그룹에 대한 주요 단어를 집약하면 [표8]과 같다.



[그림 7] 철도차량 분야 담론 시각화 (학술논문정보)

[표 8] 철도차량 분야 담론 그룹 (학술논문정보)

구분	그룹명	주요 단어
그룹(0)	총체적 기술 접근	속도, 고속, 소음, 인프라, 선로 등
그룹(1)	데이터 활용	대중교통, 데이터, 시스템, 고장, 신호 등
그룹(2)	차량 및 시설 안전	진동, 저감, 터널, 안전, 사고 등
그룹(3)	검증 및 제어 기술	검증, 제어, 미세먼지, 모형, 지하철 등

3.2.4 토픽모델링 분석 결과

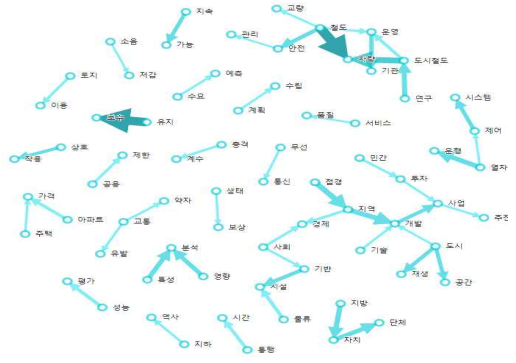
철도차량 분야에 대한 학술논문정보 검색 데이터를 주요 토픽으로 집약하면 [표9]와 같다.

[표 9] 철도차량 분야 주요 토픽 (학술논문정보)

토픽		주요 키워드
1	공간 기반 토픽 (λ = 0.5 / 15.3%)	도시, 시스템, 철도, 공간, 구축 등
2	지역간 연결 토픽 (λ = 0.5 / 23.5%)	지역, 철도, 물류, 개발, 유지 등
3	영향 분석 토픽 (λ = 0.5 / 31%)	연구, 영향, 궤도, 도시철도, 속도 등
4	평가 검증 분석 토픽 (λ = 0.5 / 30.3%)	평가, 환경, 안전, 철도, 도시 등

3.2.5 네트워크 분석(N-GRAM) 결과

철도차량 분야에 대한 학술논문정보 검색 데이터를 주요 토픽으로 집약하면 [그림8]과 같다. 유지-보수, 철도-차량, 열차-운행 등과 같이 상호 연계된 네트워크로 검색어가 선후관계로 연계되어 있음을 확인할 수 있다.



[그림 8] 철도차량 분야 N-GRAM (학술논문정보)

4. 결론

이 연구는 공공재로서의 철도차량의 최종 서비스 수혜자인 국민의 시각을 포털사이트에서의 뉴스, SNS 텍스트를 활용하여 분석하고, 전문가 학술논문 정보 분석결과를 함께 분석하였다. 주요 검색어의 검색빈도를 중심으로 정량적 분석과 함께 텍스트마이닝 등의 정성적 분석을 병행하여 방법론적 다원성을 확보하였다.

포털사이트 분석결과 '교통', '지하철', '고속열차', '안전' 등 수요자 중심의 철도차량 검색어의 빈도가 높게 나타났고, 학술논문정보 분석결과 '시스템', '안전', '교통' 등의 검색어 빈도가 높게 나타났다. '안전', '사고'에 대해서는 공통으로 높은 관심을 확인하였으며, 포털사이트에서는 '수소', '연결' 등 친환경 신기술 및 교통편의 제고를 위한 이슈에도 높은 관심이 확인되었고, 학술논문정보에서는 '데이터', '예측' 등에 대한 높은 관심이 확인되었다. 이를 통해 철도안전, 사고예방, 친환경 철도 신기술, 교통체계 연결, 데이터 중심의 예측기술 등에 대한 향후 정책·기술수요를 도출할 수 있다.

담론그룹과 토픽분석 결과, 포털사이트 정보에서는 철도차량 유형 및 수요를 통한 영향(역세권, 개통 등)을 중심으로 높은 관심이 나타났고, 학술논문 정보에서는 철도차량을 중심으로 하는 연구분야와 분석대상에 대한 관심이 확인되었다.

일반 국민의 시각이 반영된 포털사이트 정보분석과 철도전문가의 시각이 반영된 학술논문 정보분석은 각각 접근 시각과 관심영역의 차이에 기반하여 다소 다른 관심이 확인되었지만, 이에 대한 상호 보완적 접근을 통해 양측을 함께 고려할 수 있는 정책·기술수요 반영 노력이 요구된다.

참고문헌

[1] 백승현, 이윤주, 한국, 중국, 일본 철도연구기관 특허 및 논문실적 비교분석, 한국산학기술학회지, 21(6), pp. 455-460, 2020년

[2] 김용희, 김민영, 빅데이터 기반 철도과학 기술정책 사회동향 조사분석, 더아이엠씨, 2022년

[3] 강욱건, 고의석, 이학래, 김재능, 빅데이터 분석을 통한 패키징에 대한 소비자의 주요 인식 조사, 한국융합학회논문지, pp. 9(4), 15-22, 2018년

[4] 남춘호, 일기자료 연구에서 토픽모델링 기법의 활용가능성 검토, 비교문화연구, 22(1), pp. 89-135, 2016년