NTIS 데이터를 이용한 국내 자율주행 연구 동향 분석에 관한 연구

A Study of the Trend Analysis of National Automated Vehicle Research Using NTIS Data

정 인 석* · 강 지 원** · 이 종 덕*** · 박 상 민****

* 주저자 : 한국교통연구원 도로교통연구본부 연구원
** 공저자 : 한국교통연구원 도로교통연구본부 전문연구원
*** 공저자 : 한국교통연구원 도로교통연구본부 책임전문원
**** 교신저자 : 한국교통연구원 도로교통연구본부 부연구위원

In-Seok Jeong* · Jiwon Kang* · Jongdeok Lee* · Sangmin Park*

* Dept. of Road Transport.Research, Korea Transport Institute

† Corresponding author : Sangmin Park, psm@koti.re.kr

Vol. 22 No.2(2023)
April, 2023
pp.147~163

pISSN 1738-0774
eISSN 2384-1729
https://doi.org/10.12815/kits.2023.22.2.147

Received 24 November 2022
Revised 22 December 2022
Accepted 7 February 2023

요 약

최근 전 세계적으로 첨단 이동 수단인 자율주행자동차에 대한 연구가 활발하다. 국내에서도 첨단 이동 수단 기술을 12대 국가 전략기술로 선정하였으며, 자율주행자동차와 관련된 국가 R&D 사업을 통해 연구가 꾸준히 진행되고 있다. 자율주행자동차 기술의 경우 다양한 분야의 기술이 집합된 결과물로 다양한 방향성을 보이고 있다. 그렇기에 자율주행 연구의 현 위치를 파악하고 향후 방향성을 정립하는 것이 필요하다. 본 연구에서는 국가과학기술지식정보서비스(National Science and Technology Information Service, NTIS)에서 제공하는 국가 R&D 사업에 등록된 성과 정보 중 논문 초록을 활용하여 연구 동향을 분석하는 방법론을 제시하였다. 또한, 제시된 방법론을 이용하여 주요 키워드 및 주요 토픽을 도출하여 개발된 연구 동향 방법론의 유효성을 검토하였다. 본 연구에서 개발된 방법론은 향후 자율주행자동차 연구 동향 파악 및 분석에 활용될 수 있을 것으로 기대된다.

핵심어 : 자율주행, 연구 동향 분석, 토픽 모델링, NTIS 데이터

ABSTRACT

Recently, there has been an increase in the research and development of automated vehicles worldwide. Research focused on automated vehicles in Korea is steadily progressing as a national R&D project. Since automated driving technology comprises diverse technology fields, it is necessary to identify the current position of the research. In this study, we propose a methodology for analyzing research trends using the NTIS data. In addition, we review the effectiveness of the currently developed research trend methodology by deriving primary keywords and major topics using the proposed method. We expect that the methodology developed in this study can be applied to identify and analyze future automated vehicle research trends.

Key words : Automated vehicle, Research trend analysis, Topic modeling, NTIS data

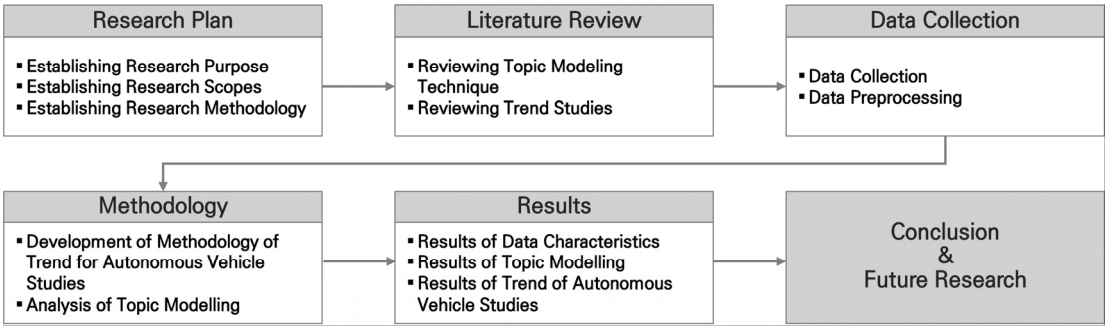
Ⅰ. 서 론

1. 연구의 배경 및 목적

2022년 10월 우리나라 정부는 '국가 전략기술 육성방안'을 발표하면서, 12대 국가 전략기술의 하나로 첨단 이동 수단을 선정하였다. 향후 우리나라가 전략적으로 육성하고자 하는 기술 분야로 선정된 대표적 첨단 이동 수단 중 하나인 자율주행자동차는 볼보, BMW, 벤츠, GM, 도요타 등 해외 완성차 제조업체뿐만 아니라, Google 등 다양한 기관에서도 개발 중에 있다(Park et al., 2019). 국내에서도 민간과 공공에서 지속적인 연구 및 개발을 수행하고 있으며, 최근에는 자율주행과 관련 있는 정부 부처들이 자율주행 SAE Level 4 기술 개발을 목적으로 자율주행기술개발혁신사업이 시행되는 등 국가 R&D도 꾸준히 진행하고 있는 추세이다. 자율주행 Level 4 수준은 일부 지역으로 ODD를 한정하지만, 사실상 완전한 자율주행 기술이 구현된 것으로 볼 수 있다. 따라서 본 연구에서는 Level 4 자율주행 기술을 개발하고자 하는 지금 시점에서, 현재까지 진행되어 온 자율주행 관련 연구개발 결과물이 실제로 어떤 분야를 중심으로 연구되어 왔는지 살펴보는 것이 필요하다. 따라서 본 연구에서는 국내 자율주행 관련 연구들의 동향을 살펴보기 위해 국내 자율주행 연구 동향 분석을 위한 방법론을 개발하였다. 이를 위해 국가과학기술지식정보서비스에서 제공 중인 연구개발 성과물을 분석하여 주요 연구 분야를 파악하였다. 또한, 향후 우리나라의 자율주행 연구개발사업 추진 시 강화시킬 강점이 무엇이며, 현재 어떤 분야의 연구가 필요한지를 분석하여 향후 자율주행 시대를 대비하기 위한 정책적 시사점을 제시하고자 한다.

2. 연구의 범위 및 방법

자율주행과 관련된 연구 동향을 도출하기 위해 연구범위를 설정하였다. 본 연구의 시간적 범위는 2017년~2021년까지 수행된 국가 연구개발 사업의 성과로 제출된 학술 논문 데이터를 이용하였다. 또한, 국내에서 수행된 자율주행자동차 국가 R&D의 연구성과 중 학술 논문 데이터를 이용하여 연구를 수행하였다. 이를 위해 2017년부터 2021년까지 수행된 국가연구개발사업의 성과로 제시된 KCI 및 SCI의 논문 데이터를 수집하고, 기초 데이터 분석 등을 통해 수집된 자료의 특성을 파악하였다. 또한, 텍스트 마이닝 기법 중 하나인 토픽 모델링 기반의 자율주행 연구 동향 분석 방법론을 개발하였다. 개발한 연구 동향 분석 방법론을 적용하여 주요 연구 분야들을 도출하였다. 마지막으로 결론 및 향후 연구과제를 도출하였다. 전체적인 연구 과정은 <Fig. 1>과 같다.



<Fig. 1> Procedure of the research

Ⅱ. 관련 연구 고찰

1. 텍스트 마이닝

1) 개요

기술의 발전에 따라 이전에 처리할 수 없었던 많은 양의 자료에 대한 분석이 가능해짐에 따라 정형 빅데이터뿐만 아니라 비정형 빅데이터를 활용한 연구 수행이 가능하게 되었다. 기존 정형 데이터와는 다르게 비정형 빅데이터는 텍스트, 이미지, 영상, 등 구조를 갖추지 않은 빅데이터로 비정형 빅데이터를 활용하면, 기존 정형 데이터에서는 추출하기 힘든 다양한 인사이트(Insight)를 추출하는 것이 가능하다. 특히, 비정형 빅데이터 중 하나인 텍스트 빅데이터는 텍스트 마이닝(Text Mining)을 통해 유의미한 정보를 추출하는 것 가능하다. 일반적으로 텍스트 마이닝은 비정형 데이터 수집, 데이터 전처리, 정보 추출, 정보 분석의 절차를 따르고 있다(Oh et al., 2016). 국내에서는 1998년부터 관련 연구가 시작되었고, 2011년 이후 빅데이터 기반의 연구가 활발해지면서 관련 연구가 증가하였다(Im et al., 2017). 텍스트 마이닝의 주요 기법인 토픽 모델링(Topic Modeling), 감성 분석 (Opinion Mining), 시맨틱웹(Semantic Web), 온톨로지 기법(Ontology) 등을 활용하여 많은 연구에서는 유의미한 정보를 도출하고 있다.

2) 텍스트 마이닝을 이용한 연구사례

Im et al.(2017)은 '자율주행' 키워드를 중심으로 국내 언론사 웹사이트에서 키워드가 포함된 기사, 해당 기사의 댓글 등을 대상으로 감성 평가를 활용하여 분석하였다. 한글 콘텐츠의 각 음절과 어절을 벡터 좌표상에 재배열하여 수학적 방법으로 풀이하고 그 값에 대한 의미를 분석하는 방법으로 해당 연구에서는 정확도를 높이기 위하여 형태소의 Label Propagation 기법과 형태소 간 k-NN 기법을 활용하였다. 해당 연구는 2015년 1월~2017년 8개월 동안 국내 언론사 164개 웹사이트의 기사 및 댓글과 포털 뉴스 섹션의 기사 및 댓글 인터넷 뉴스 기사와 해당 기사 댓글을 분석하여 자율주행자동차에 대한 시민 의식이 신기술에 대한 우려 등이 높게 나타남에 따라 사회적 수용성의 향상이 중요한 사항이라 판단하였다.

Kim et al.(2018)은 미세먼지 감축을 위한 정책의 반응에 대한 분석에 활용하였다. 해당 연구는 크게 두 부분으로 산업과 미세먼지 발생과의 연관성을 회귀분석으로 분석하고, 미세먼지 관련 교통정책인 '대중교통 무료'와 '차량 2부제'의 관련 기사의 댓글을 텍스트 마이닝을 통해 분석하여 그 결과를 워드 클라우드를 생성하고 정형 및 비정형 자료 분석 결과를 통합하여 미세먼지 비상저감조치에 대한 향후 방향을 제시하고자 하였다.

Lim et al.(2014)은 국가연구개발사업 보고서 및 논문을 중심으로 텍스트 마이닝 기술을 이용하여 공간 정보 분야의 연구 동향을 분석하였다. 해당 연구에서는 1996년부터 2013년까지 총 4,000개 논문과 1994년부터 2013년까지 총 225개의 국가연구개발보고서를 국가과학기술전자도서관(NDSL)에서 추출하여 사용하였다.

Park et al.(2021)은 도시부 간선도로에서 발생한 교통사고 데이터를 이용하여 자율주행자동차의 평가 시나리오를 발굴하는 연구를 수행하였다. 특히, 도시부 간선도로를 공간적 범위로 하여 주간 및 맑은 날을 자율주행자동차의 운행가능영역(Operational Domain Design, ODD)로 정의하고 텍스트 마이닝 기법을 이용하여 Target Object, Provoking Event, Maneuver 카테고리에 해당하는 주요 Feature를 추출하였다. 추출된 Feature들과 Functional Scenario 도출체계를 이용하여 자율주행자동차 평가 시나리오를 개발하였다.

2. 토픽 모델링

1) 개요

토픽 모델링은 큰 의미의 확률론적 모델링 중 하나로 그 중 LDA(Latent Dirichlet Allocation)는 자연어 처리 및 기계학습 분야에 큰 영향을 미쳤으며, 기계학습에서 가장 인기 있는 확률론적 토픽 모델링 기술이다(Kim et al., 2017). LDA 토픽 모델링은 각 문서에 어떤 토픽들이 존재하는지에 대한 확률 생성 모형이며, 문서들은 임의의 주제로 표현이 가능하고 임의의 주제는 단어의 분포로 나타낼 수 있다(Blei et al., 2003). LDA 토픽 모델링에서 사용하는 확률 분포 식은 (1)과 같다(Blei, 2012).

$$p(\beta_{1:K}, \theta_{1:D}, z_{1:D}, w_{1:D}) = \prod_{i=1}^{K} p(\beta_i) \prod_{d=1}^{D} p(\theta_d) \left(\prod_{n=1}^{N} p(z_{d,n} \mid \theta_d) p(w_{d,n} \mid \beta_{1:K}, z_{d,n}) \right) \quad \cdots\cdots (1)$$

여기서,

β_k : $kth\ topic\ in\ document$
θ_d : $Topic\ proportions\ on\ dth\ document$
$z_{d,n}$: $Topic\ assignment\ nth\ words\ on\ dth\ documnet$
$w_{d,n}$: $Observed\ nth\ word\ on\ dth\ documnet$

2) LDA 토픽 모델링 연구사례

Oh et al(2016)은 경계성 이론과 텍스트 마이닝을 이용하여 정책 이슈를 탐색하는 기법을 정립하고 이를 도로 부분 ITS를 대상으로 적용하였다. LDA 모형을 기반으로 하는 비대칭-대칭 혼합 어휘소 기반 LDA를 응용하여 2012년 1월부터 2014년 12월까지 2개의 포털사이트 및 205개 지방언론 매체를 소스로 사용하여 도시교통 정보 시스템의 교통정보 수집률 저조, 첨단교통관리 시스템과 중복, 디지털운행기록계의 주행거리 조작 등을 주요 이슈로 도출했다.

3. 동향 분석

1) 관련 연구

Park et al.(2017) LDA 토픽 모델링을 활용하여 과학기술 동향 및 예측을 위한 분석 방법론을 제시하고 있다. 2000년~2016 미국 특허 초록 14,187개를 대상으로 20개의 AI(Artificial Intelligence) 핵심 기술을 도출하여 Hot/Cold AI 세부 기술을 도출하였다. Jang and Jung(2021)에서는 최근 도시 분야 연구 동향을 분석하기 위하여 토픽 모델링을 사용하였다. 2002년부터 2019년 사이에 게재된 한국학술지인용색인(KCI)에 등재된 논문의 초록을 분석한 결과 도시재생 분야 연구가 지속적으로 증가하고 있는 반면 성장/개발과 에너지/환경과 같은 주제는 정체기에 들어간 것으로 분석되었다. Na et al.(2016)은 시뮬레이션 활용 연구 분야의 핵심 토픽을 도출하기 위해 텍스트 마이닝 기반의 트렌드 분석에 대한 활용 가능성을 제시하기 위하여 LDA 모델을 활용하였다. KCI DB(Korea Citation Index Database)에 저장된 논문을 기준으로 1992년부터 2015년까지 11,895건의 논문의 제목, 초록, 키워드, 게재 연도를 분석한 결과 공학 분야에서는 통신 및 전기분야에서, 사회과학 분야는 교육 및 오락 분야에서 시뮬레이션을 많이 활용한다고 분석하였다.

2) 국가과학기술지식정보서비스 데이터베이스

국가과학기술지식정보서비스(National Science & Technology Information Service, NTIS)는 국가 연구개발 사업에 대한 정보(사업, 과제, 연구자, 성과 등)를 서비스하는 국가 R&D 지식 정보 포털로, 부처별로 개별 관리되고 있는 국가 R&D 사업 관련 정보와 과학기술 정보를 공유하고 공동 활용해 국가 R&D 투자 효율성을 높이고 연구 생산성 향상에 기여하는 것을 주목적으로 하는 시스템이다(Yang et al., 2021). 특히. NTIS는 국가연구개발사업에 대한 과제정보뿐만 아니라, R&D 과제를 수행하면서 발생된 논문, 특허 등 연구성과에 대한 정보를 수집하여 다양한 서비스를 함께 제공하고 있다. NTIS는 「국가연구개발혁신법」재정 이후 공공데이터에 대한 개방 요구 정책에 맞춰 국가개발정보에 대한 개방 대상과 항목, 이용 범위 등을 명확히 하고 개방 서비스를 대폭 확대하여 제공하고 있다(Yang et al., 2021). 특히, 국가연구개발정보를 누구나 쉽게 접근하여 활용할 수 있도록 서비스를 제공하고 있으며, 검색 결과를 직접 다운로드하여 활용할 수 있도록 하고 있으며, 키워드에 따른 국가연구개발 정보를 분석 및 시각화할 수 있는 서비스를 지속적으로 개발하여 제공하고 있다(Yang et al., 2021).

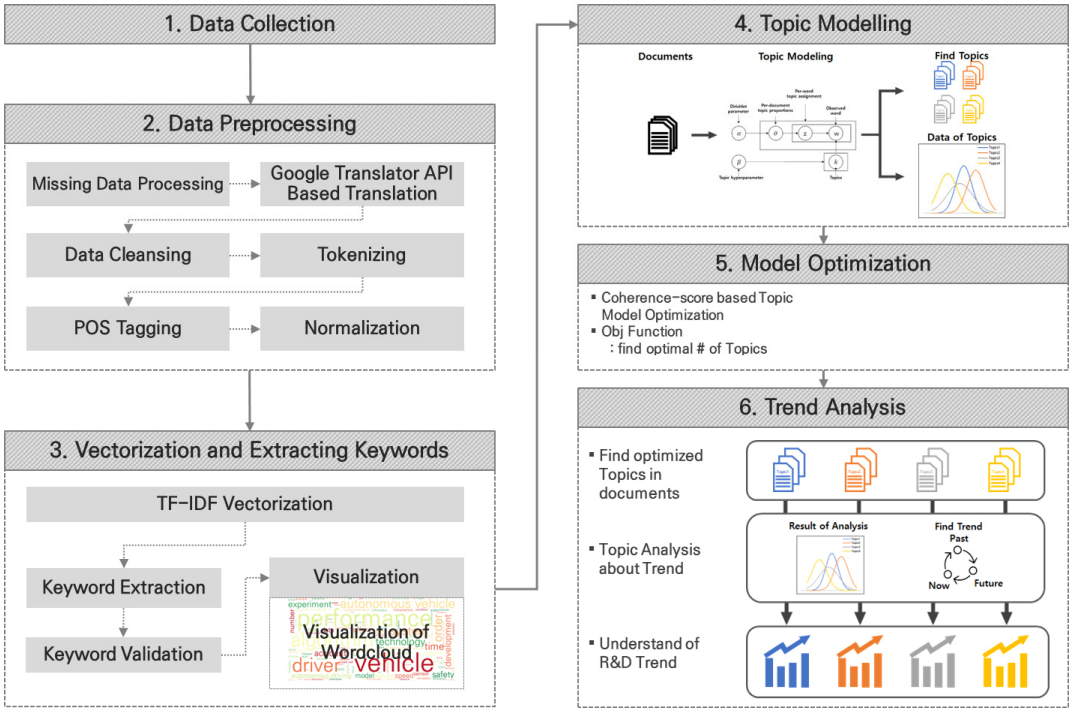
3) 자율주행 연구 동향 분석

Faisal et al.(2021)은 1998년부터 2017년까지 20년 동안 4,645건의 자율주행 논문을 통하여 경향과 패턴, 상호연결성에 관해 연구하였다. 자율주행 연구의 87.7%는 교육기관에서 수행되었고, 유럽은 35.9%의 출판물을 가졌으며, 북아메리카는 41.1% 인용을 받아 영향력이 높았다. 지난 3년 동안 연구의 50% 이상이 수행되었으며 도시와 사회학적 분야에서 자율주행 연구는 초기 단계고 학계와 산업계의 협업은 제한적이라는 결론을 도출하였다. Gandia et al.(2019)은 CiteSpace 소프트웨어를 활용하여 자율주행차 연구의 주요 특성과 잠재적 동향 등 파악을 위해 10,580개의 논문을 분석하였다. 자율주행차의 연구는 점점 증가하며 산업뿐만 아니라 경제 및 관리 문제와 같은 측면에서 이해의 필요성을 제시하였다. 다학제간의 연구가 다원제성 연구로 변화하고 있으며 또한 용어 표준화가 다양한 분야에서 진행되는 자율주행 연구의 다원화를 기여할 수 있다고 주장했다. Kim(2021)은 국내외 자율주행자동차 연구개발에 이슈 파악을 위하여 텍스트 마이닝 분석을 진행하였다. 사용된 데이터는 자율주행자동차 연구개발 관련 영문 뉴스 1,870개를 수집하고 텍스트 마이닝 분석을 통해 자율주행자동차 연구개발사업 논리 모형을 제시하였다. 분석 결과를 통하여 빠르게 변화하는 자율주행 기술 개발에 대비할 기초 자료로 제시하였다. Hacohen et al.(2022)은 자율주행 기술이 광범위하게 미칠 영향에 Google Scholar와 Web of Science(WoS)를 기반의 문헌으로 추가적 연구가 필요한 분야를 도출하였다. Google Trends 검색 결과의 수 분석을 기반으로 자율주행 연구 동향의 결과를 제시하였다. 분석 결과로 Vehicle-to-Vehicle(V2V)와 Vehicle-to-Cloud(V2C) 분야의 기술 격차와 표준화 프로세스에 관심이 높으며 또한 자율주행에 정보화 기술 적용으로 인한 보안에 대한 중요성의 증대됨에 따라 정보화 및 보안에 대한 연구의 수는 점점 증가하고 있다는 결론을 도출하였다.

Ⅲ. 연구 방법론

1. 개요

본 연구에서는 국내 자율주행자동차 관련 연구 동향 분석을 위해 데이터 수집, 데이터 전처리, 주요 특징 벡터화 및 키워드 추출, 토픽 모델링, 모델 최적화, 연구 동향 분석의 6단계로 구성된 방법론을 제시하였다. 다음 <Fig. 2>는 본 연구의 방법론을 체계적으로 도식화한 그림이다.



<Fig. 2> The Proposed Methodology

2. 데이터 수집

본 연구에서는 자율주행 연구 동향 분석 방법론을 개발하기 위해 국가과학기술지식정보서비스(NTIS)에서 제공하고 있는 국가연구개발 성과로 등록된 논문 데이터를 수집하였다. 수집된 논문 데이터는 학술지명, 논문명, ISSN_ISBN, SCI 구분, 초록, 저자, 과제명(국문), 과제 고유번호 등 정보를 포함하고 있다. 특히 국가연구개발의 주요 성과인 논문의 초록을 포함하고 있어 연구 성과의 내용을 파악할 수 있는 장점이 있다. 따라서, 본 연구에서는 국내 자율주행 연구 동향 분석 방법론 개발을 위해 2017년부터 2021년까지 5년간 “자율주행” 키워드로 등록된 과제의 성과로 등록된 625건의 데이터를 수집하였다. 수집된 데이터 중 동향 분석을 이용하기 위해 학술 논문 성과 데이터의 초록을 주요하게 활용하였다.

3. 데이터 전처리

국과가학기술지식정보서비스(NTIS)에 등록된 성과 정보 중 논문 초록을 활용하기 위해서는 텍스트 분석에 적절한 데이터 전처리가 필요하다. 특히 텍스트 데이터의 전처리는 기존 정형 데이터의 전처리와는 다르기 때문에 본 연구에서는 텍스트 데이터 전처리를 위해 6단계의 전처리 과정을 이용하였다.

1) 결측 데이터 처리

수집된 학술 논문 성과의 초록 데이터 수를 확보하기 위해 결측 데이터 처리를 진행하였다. 본 연구에서는 결측 데이터를 추가적으로 수집하는 것이 가능하다. 따라서 NTIS를 통하여 수집된 데이터 중 초록이 등

록되지 않은 경우 논문명을 이용하여 논문 초록을 추가적으로 수집하여 결측치를 보완하였다.

2) 데이터 언어 통일

성과로 등록된 논문 초록 데이터들은 국문 초록과 영문 초록이 혼용되어 데이터가 수집되었다. 본 연구에서는 데이터의 언어를 영어로 통일하였으며, 영문 텍스트 분석을 진행하기 위하여 국문 초록 데이터의 경우는 영문으로 번역을 진행하였다. 번역을 위해서 Python의 번역 라이브러리 중 하나인 Googletrans 라이브러리를 사용하였다. Googletrans 라이브러리는 구글 번역(Google Translate) API를 사용하여 번역을 자동으로 수행해 주며, 언어의 번역, 언어의 종류 감지 기능 등을 활용할 수 있어 대량의 데이터를 빠르게 번역하는 것이 가능하다.

3) 데이터 정제

번역된 데이터에서 품질을 낮출 수 있는 문자를 확인하고 분석의 정확도를 높이기 위해 데이터 정제 단계가 필요하다. 특수 문자, 공백 기호, 문장 기호 등 의미가 없는 불용어를 제거하는 것이 반드시 필요하며, 이를 위해 파이썬의 정규 표현식을 사용하여 제거하였다. 정규 표현식은 정해진 규칙을 통해 문자를 표현하는 것으로 지정한 조건을 통해 대량의 텍스트의 처리를 효율적으로 처리하는 것이 가능하다.

4) 데이터 토큰화

정제된 데이터를 통하여 분석용 데이터를 구성하기 위하여 토큰화를 수행하였다. 토큰화란 문장으로부터 정해진 조건에 따라 요소로 분리하는 작업을 뜻한다. 데이터의 토큰화 방식 중에 문장부호나, 어절을 기준으로 진행하는 방식을 활용하여 문장 데이터의 토큰화를 진행하였다. 데이터 토큰화의 진행은 Python 3.9 버전과 NLTK(Natural Language Toolkit) 라이브러리 3.6.7 버전을 사용하여 토큰화를 진행하였다.

5) POS 태깅

POS(Part of Speech) 태깅은 토큰화된 주요 특징에 품사를 태그하는 전처리 기법 중 하나이다. 토큰화된 데이터에 품사 태그를 통해 더욱 정밀한 분석이 가능하다. 영문의 경우 단어사전을 통하여 품사를 태그 할 수가 있다. 특히 텍스트 마이닝에서는 문장내 주요 특징(Feature)을 추출하기 위해 주어, 목적어 등으로 사용되는 명사를 주로 사용한다. 본 연구에서는 품사 태그를 통해 명사와 합성 명사들을 추출하였다.

6) 데이터 정규화

명사형 기반으로 추출된 단어로 정규화를 진행하였다. 정규화란 데이터를 특정 표준으로 일치하는 것을 뜻한다. 정규화는 복수명사인 경우 단수 명사로, 대문자는 소문자로, 명사형의 표제어로 바꿔주는 모든 과정을 포함했다. 텍스트의 같은 의미의 복수형과 단수형을 다르게 분류하기에 정규화를 통해서 같은 뜻을 지닌 단어의 경우 통일을 진행하였다.

4. 벡터화 및 키워드 추출

전처리 된 데이터 중 주요 특징을 추출하기 위해 벡터화를 진행하였다. 벡터화의 경우 텍스트 형식의 데이터를 분석에 용이하도록 변화하는 과정을 의미한다. 벡터화에 사용된 방법으로는 TF-IDF(Term Frequency-

Inverse Document Frequency) 가중치를 사용하였다. TF-IDF는 여러 문서에서 특정 단어가 중요한 정도를 나타내는 수치로 단순히 빈도로 중요도를 측정하는 것이 아니라 문서 전체에 단어의 분포도 고려하여 가중치를 도출한다. TF-IDF 가중치를 사용하여 단순히 높은 빈도의 키워드보다 중요성이 높은 키워드 도출을 위한 방법으로 사용하였다. 도출한 가중치를 통해 워드클라우드 시각화로 중요한 키워드를 파악하고자 하였다.

5. 토픽 모델링

연구 동향을 도출하기 위한 기법으로 토픽 모델링을 선정하였다. 토픽 모델링은 다양한 텍스트 마이닝 기법 중 고차원 텍스트 데이터로 구성된 다량의 문서를 차원 축소(Dimension Reduction)를 통하여 연관성 있는 집합으로 분류가 가능한 기법이다. 본 연구에서는 토픽 모델링 모형 중 성능이 높으며, 많이 사용되고 있는 LDA(Latent Dirichlet Allocation) 모델을 사용하였다. LDA 모델은 지정된 토픽 수를 기반으로 계산되는 확률적 모형으로 문서에 단어 분포를 통하여 어떤 토픽으로 구성되는지 예측할 수 있다. 본 연구에서는 LDA를 활용한 토픽 모델링을 수행하기 위해 Python과 Gensim 라이브러리를 사용하였다.

6. 모델 최적화

LDA 토픽 모델링을 수행하기 위해서는 최적의 토픽의 수를 선정하는 것이 필요하다. 본 연구에서는 데이터 속에서 의미 있는 정보를 도출하고자 Coherence 점수를 통해 토픽 수 결정하여 모델 최적화를 진행하였다. Coherence 점수는 토픽 내 단어의 유사도를 계산하여 해당 토픽이 의미론적으로 일치하는 단어로 구성되어 있는지를 파악할 수 있고, 높을수록 의미론적 일관성이 높다고 할 수 있다(Yu, 2017). 본 연구에서는 Coherence 점수를 도출하기 위해 NPMI(Normalized Pointwise mutual information) 코사인 유사도를 이용하여 최적화를 진행하였다. NPMI는 PMI를 정규화한 것으로 단어 집합에서 단어 간 짝 유사도(pairwise similarity)의 평균이 높을수록 토픽의 응집성이 높다고 가정한다(Yu, 2017). NPMI는 낮은 빈도에서 보이는 편향을 줄일 수 있다(Bouma, 2009). Coherence 점수 계산 식은 (2)로 다음과 같다(Röder, 2015).

$$Coherence(w_i,w_j) = NPMI(w_i,w_j)^{\gamma} = \left(\frac{\log \frac{P(w_i,w_j)+\epsilon}{P(w_i)P(w_j)}}{-\log(P(w_i,w_j)+\epsilon)} \right)^{\gamma} \quad \cdots\cdots (2)$$

여기서,

γ : $weight\ of\ NPMI$
w_i : $word\ i$
$P(w_i,w_j)$: $prob\ of\ words\ w_i$ and w_j $in\ same\ doc$
$P(w_i)$: $prob\ of\ word\ w_i$
ϵ : $using$ for $avoid$ log$arithm\ of\ zero$

7. 연구 동향 도출

최적화된 토픽 모델을 통해 도출된 토픽들로 연구 세부 주제를 도출하였다. 본 연구에서는 토픽 모델링을 통해 도출된 토픽의 주요 특징들을 분석하여 세부 주제를 도출하였다. 또한 토픽 모델의 결과를 각각의 문서에 적용하여 국내 자율주행자동차의 연구 동향을 도출하였다.

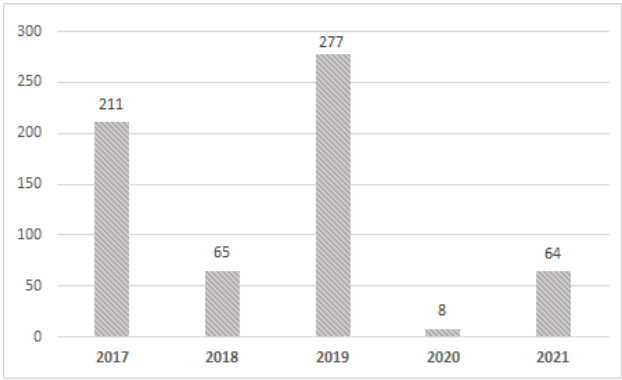
Ⅳ. 자율주행 연구 동향 분석 결과

1. 기초 데이터 분석

수집된 데이터를 분석하기에 앞서, 기본적인 정보를 파악하기 위해 기초 데이터 분석을 수행하였다. 데이터는 2017년, 2018년, 2019년, 2020년, 2021년으로 구분할 수 있으며, 발행지에 따라 SCI 논문과 KCI 논문으로 구분할 수 있다.

1) 연도별 빈도 분석

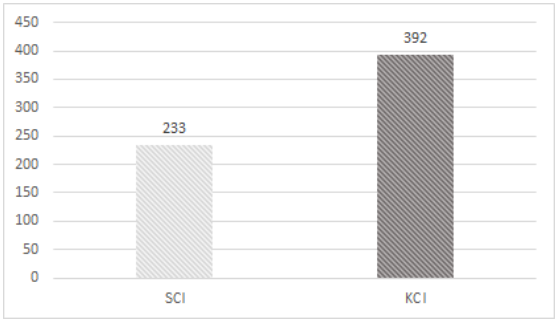
발행연도에 따른 연구 동향을 분석하기 위해 연도별 빈도 분석을 진행하였다. 625건의 과제정보 중 2017년은 211건, 2018년은 65건, 2019년은 277건, 2020년은 8건, 2021년은 64건으로 나타나, 연도별로 많은 차이를 보이는 것으로 나타났다. <Fig. 3>은 빈도 그래프이다.



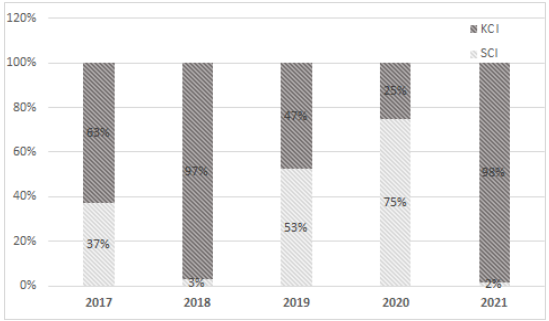
<Fig. 3> Frequency Analysis Results by Year

2) 연도별 등재지 분석

등재지에 따른 차이를 파악하기 위해 빈도 분석을 수행하였다. 625건 중 SCI와 KCI는 233건, 392건으로 KCI 논문이 더 많았다. 연도별로 비율을 비교하면 2017년, 2018년, 2021년도에는 KCI의 비율이 높았으며, 2019년, 2020년도에는 SCI 논문의 비율이 높았다. 특히 2018년에는 KCI의 비율이 97%, 2021년에는 98%로 높게 나타났다. <Fig. 4>는 학술 논문 성과 중 KCI와 SCI의 비율을 나타낸 그림이며 <Fig. 5>는 연도별 SCI와



<Fig. 4> Frequencies of SCI and KCI



<Fig. 5> Frequencies of SCI and KCI by Year

KCI의 비율을 나타낸 그림이다.

2. 키워드 도출을 통한 자율주행 연구 동향 분석 결과

1) 연도별 주요 키워드 분석 결과

자율주행 분야에서 연도별 연구 동향 변화를 알아보기 위해 연도별 키워드 분석을 진행하였다. 2020년 초록의 경우 충분한 문서의 개수 확보를 위해 2021년과 통합하여 진행하였다. 우선, 전체 데이터를 분석한 결과 “Control“, “Sensor”, “Safety”, “Driver”, “Simulation”의 키워드가 중요도가 높은 것을 확인하였다. 그중 가중치가 높은 키워드인 “Control”, “Sensor”, “Safety”를 통하여 자율주행 기술의 안전을 목표로 한 연구가 중심이 되었음을 유추할 수 있다.

<Fig. 6> Wordcloud of Total Data

연도별로 키워드 분석 진행 결과 연구 동향에서 세부적인 주제에 대한 차이가 있음을 확인할 수 있었다. 특히나 2018년도에는 “Security”의 TF-IDF 기반 가중치가 높은 것을 통해 통신 보안과 관련한 연구가 진행이 활발함을 유추할 수 있다. 2020년도에는 “Scenario”, “Case”가 새롭게 도출된 키워드로 보아 자율주행자동차의 평가 시나리오와 관련된 연구가 활발함을 유추할 수 있었다. 전체 데이터와 연도별 데이터의 키워드 분석 결과를 통해 자율주행 연구에서 안전에 관한 연구가 활발하게 진행됨을 확인하였다.

<Table 1> Keyword and TF-IDF Weight by Year

No	Overall Period		2017		2018		2019		2020-2021	
	Keyword	TF-IDF	Keyword	TF-IDF	Keyword	TF-IDF	Keyword	TF-IDF	Keyword	TF-IDF
1	Control	0.055	Control	0.064	Security	0.081	Control	0.072	Driver	0.108
2	Sensor	0.043	Sensor	0.052	Service	0.079	Sensor	0.053	Control	0.084
3	Safety	0.041	Driver	0.044	Control	0.071	Safety	0.047	Safety	0.08
4	Driver	0.037	Safety	0.041	Safety	0.071	Signal	0.044	Service	0.068
5	Simulation	0.031	Simulation	0.037	Situation	0.064	Simulation	0.044	Traffic	0.068
6	Service	0.028	Network	0.036	Sensor	0.058	Controller	0.042	Situation	0.066
7	Traffic	0.027	Traffic	0.032	Driver	0.056	Driver	0.039	Sensor	0.06

No	Overall Period		2017		2018		2019		2020-2021	
	Keyword	TF-IDF	Keyword	TF-IDF	Keyword	TF-IDF	Keyword	TF-IDF	Keyword	TF-IDF
8	Signal	0.027	Signal	0.031	Simulation	0.055	Network	0.031	Simulation	0.054
9	Network	0.025	Service	0.027	Accident	0.054	Robot	0.03	Scenario	0.051
10	Controller	0.024	Distance	0.025	Traffic	0.051	Service	0.026	Case	0.051

2) 등재지 별 주요 키워드 분석 결과

자율주행 분야에서 등재지 별 연구 동향 차이를 알아보기 위해 키워드 분석을 진행하였다. 분석 결과 등재지에 따른 연구 동향의 차이는 존재했다. 국외에 등재된 SCI 논문의 경우 “Control”, “Simulation”, “Controller”의 키워드를 통해 자율주행자동차 제어에 관한 연구가 중점으로 진행됨을 확인하였다. 국내에 등재된 KCI 논문의 경우 “Safety”의 키워드를 통해 안전에 관한 연구를 중점으로 진행됨을 확인하였다.

<Fig. 7> Wordcloud of SCI Paper

<Fig. 8> Wordcloud of KCI Paper

<Table 2> Keyword and TF-IDF Weight by SCI and KCI

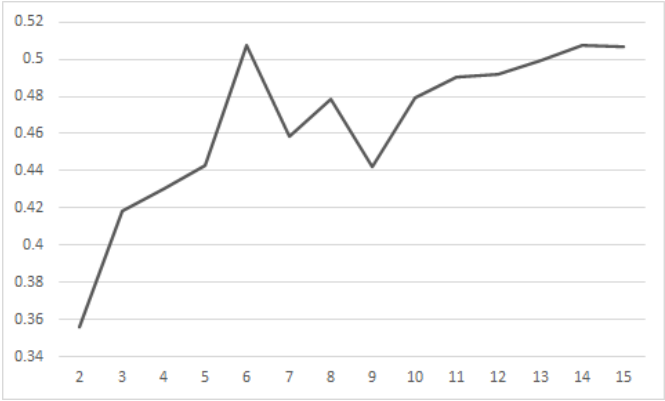
No	SCI		KCI	
	Keyword	TF-IDF	Keyword	TF-IDF
1	Control	0.074	Control	0.056
2	Simulation	0.048	Safety	0.054
3	Controller	0.043	Sensor	0.052
4	Sensor	0.042	Driver	0.052
5	Signal	0.041	Service	0.039
6	Network	0.037	Traffic	0.037
7	Stability	0.031	Simulation	0.033
8	Safety	0.026	Situation	0.031
9	Radar	0.023	Scenario	0.029
10	Process	0.022	Robot	0.028

3. 토픽 모델링을 통한 자율주행 연구 동향 분석 결과

1) 모델 최적화 결과

자율주행 키워드로 등록된 논문 성과의 동향 분석을 위해 LDA 알고리즘을 이용하여 토픽 모델링을 수행

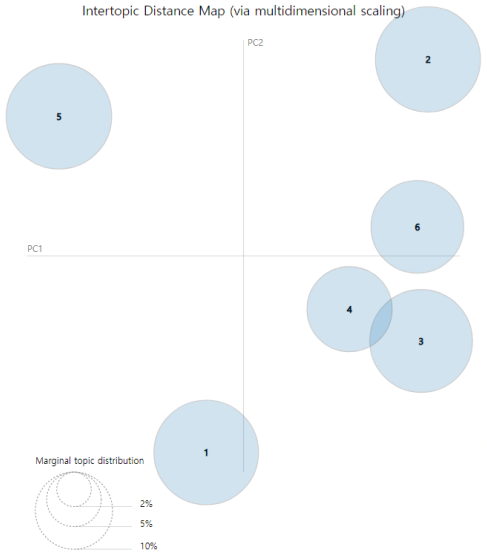
하였다. 최적의 토픽 수를 찾기 위해 본 연구에서는 Coherence 값을 이용하였으며, Coherence 값 중 NPMI를 이용하여 토픽 모델을 최적화하였다. 토픽 개수의 범위는 최소 2개와 최대 15개까지의 Coherence 값을 구하여 가장 높은 값을 최적의 토픽 수로 지정하여 분석하였다. 토픽 모델링 최적화 결과 Topic의 개수가 6개인 경우가 최적의 토픽 모델로 선정되었다. <Fig. 9>는 토픽 개수에 따른 Coherence 값의 변화 그래프이다.



<Fig. 9> Coherence Value by Number of Topics

2) 토픽 모델링 분석 결과

최적의 토픽 개수를 도출하여 토픽 모델링을 수행하였다. <Fig. 10>는 2차원 공간에 토픽 모델링 분석 결과를 시각화한 그림으로 Intertopic Distance Map이라고 부른다. Intertopic Distance Map에서 원의 면적은 각 토픽에 해당하는 단어의 양에 비례하며, 원의 거리가 가까울수록 토픽 간 공통점이 많다(Sivert and Shirley, 2014). 본 연구 결과에서는 3번과 4번을 제외한 다른 토픽 간의 원이 겹치지 않고 떨어져 있으며 독립된 토픽으로 도출되어 최적의 모델로 선정하였다.



<Fig. 10> Intertopic Distance Map of This Research

토픽 모델링 분석 결과로 각 토픽에 영향력이 높은 5개의 키워드를 도출하였다. 토픽 1의 경우 “Lane”, “Law”, “Passenger”, “Regulation”, “Control System”의 키워드가 도출되었으며, 토픽 2의 경우 “Obstacle”, “Lidar”, “Selection”, “Security”, “Guideline”의 키워드가 도출되었으며, 토픽 3의 경우 “Navigation”, “Velocity”, “Acceleration”, “Sector”, “liability”의 키워드가 도출되었으며, 토픽 4의 경우 “Radar”, “Stabilization”, “Object Recognition”, “Radar System”, “Amplification”의 키워드가 도출되었으며, 토픽 5의 경우 “Security”, “Standardization”, “Component”, “Emergency”, “Facility”의 키워드가 도출되었으며, 토픽 6의 경우 “Segmentation”, “Fusion”, “Perception”, “Functional Safety”, “Sensor Data”의 키워드가 도출되었다. <Table 3>은 토픽별 주요 키워드를 나타낸 표이다.

<Table 3> Results of the Topic Model

Keyword	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6
1	Lane	Obstacle	Navigation	Radar	Security	Segmentation
2	Law	Lidar	Velocity	Stabilization	Standardization	Fusion
3	Passenger	Selection	Acceleration	Object Recognition	Component	Perception
4	Regulation	Security	Sector	Radar Systems	Emergency	Functional Safety
5	Control System	Guideline	Liability	Amplification	Facility	Sensor Data

3) 토픽 라벨링 결과

각 토픽에 해당하는 키워드를 통하여 R&D 연구의 세부 주제를 도출하였다. 세부 주제의 도출은 분석 결과와 데이터를 도출하였다. 토픽 1의 경우 자율주행 관련 법률 연구로 자율주행의 등장과 함께 발생 될 사고에 법 제정에 관한 연구들로 도출되었다. 토픽 2의 경우 라이다 센서를 통한 장애물 종류 판단 알고리즘, 라이다 센서 기반 Point Cloud 등 라이다 센서 관련 연구로 도출되었다. 토픽 3의 경우 자율주행에서 속도와 가속도에 변화에 따른 안전한 항법 장치 관련 연구에 관련한 연구라 분석된다. 토픽 4의 경우 레이더 시스템 관련 연구로 레이더의 경우 파장의 종류, 기상 상태에 따라 정확도가 달라지기 때문에 안정화 부분에 관한 연구로 분석된다. 토픽 5의 경우 자율주행자동차 보안 및 표준화 관련 연구로 분석되었다. 토픽 6의 경우 센서 기반 객체 분류 및 인지 관련 연구로 분석되었다. 토픽별 정의한 세부 주제는 <Table 4>와 같다.

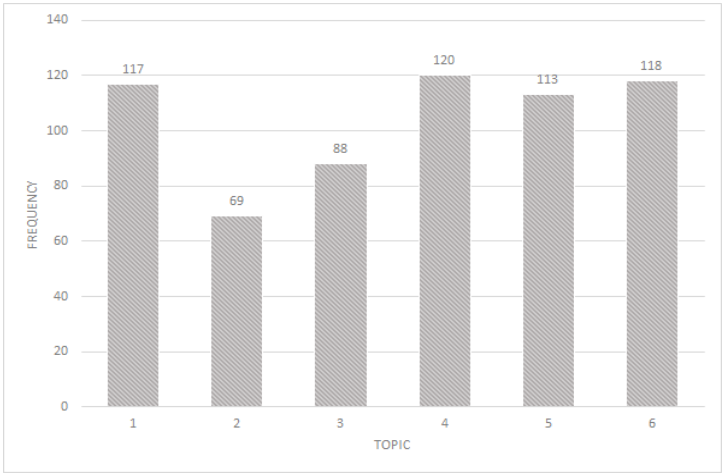
<Table 4> Results of Derived Topic Labelling

No	Topic Labeling
1	Study about Law for Automated Driving
2	Study about Lidar Sensor
3	Study of Navigation System
4	Study of Radar Sensor
5	Study of Security and Standardization
6	Study of Object Classification and Perception based on Sensors

4) 토픽 모델의 적용 결과

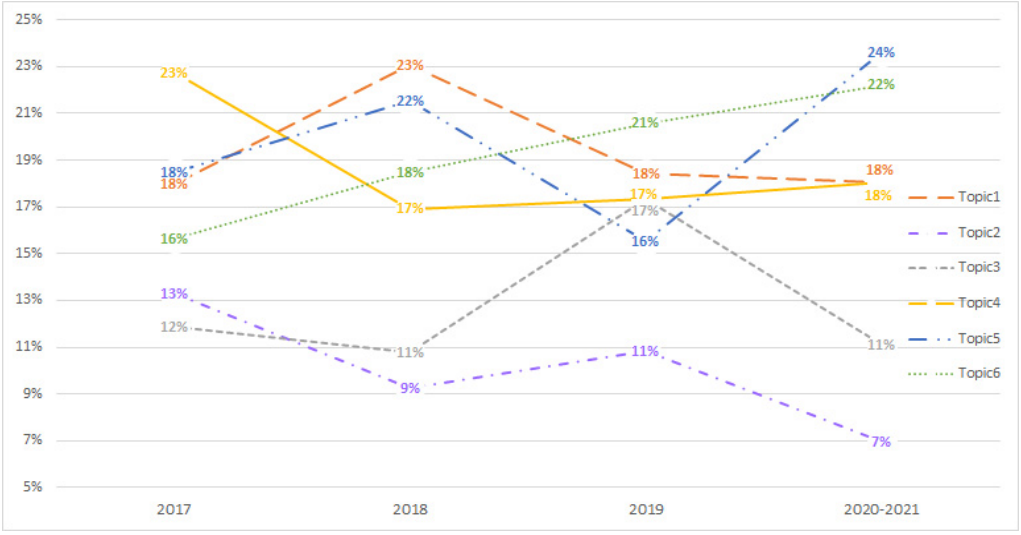
토픽 모델링으로 구성한 모델을 데이터에 적용하여 논문마다 확률이 높은 토픽을 선정하였다. 각 토픽에 해당하는 빈도는 <Fig. 11>와 같다. 토픽 1의 경우 117건, 토픽 2의 경우 69건, 토픽 3의 경우 88건, 토픽 4의 경우 120건, 토픽 5의 경우 113건, 토픽 6의 경우 118건으로 도출되었다. 세부 주제인 “라이다 센서 관련 연

구"라고 도출한 토픽 2의 건수가 가장 적었으며, 토픽 4인 "레이더 시스템 관련 연구"가 가장 많은 것으로 분석되었다.



<Fig. 11> Frequency Analysis Results by Topics

"자율주행 관련 법률 연구"에 해당하는 토픽 1의 경우 2017년에서 2018년에는 증가하였지만 2019년부터는 증가한 만큼 감소하였다. "라이다 센서에 관한 연구"에 해당하는 토픽 2의 경우 비율의 증감을 반복하다가 전체적으로 감소하였다. "안전한 항법 장치 관련 연구"에 해당하는 토픽 3의 경우 2017년에서 2018년까지는 감소하였다가 2019년에 급격히 증가하였다. 2020년도 이후에는 2017년에 비해 비율이 감소하였다. "레이더 시스템 관련 연구"에 해당하는 토픽 4의 경우 2017년에서 2018년까지는 급격한 감소를 하였다. 2018년도 이후에는 2021년도까지 소폭 증가하였다. "자율주행자동차 보안 및 표준화 관련 연구"에 해당하는 토픽 5의 경우 2017년도부터 2019년도까지 비율의 증가와 감소하였다가 급격한 증가가 되었다. "센서 기반 객체 분류



<Fig. 12> Trend Analysis Result by Topic

및 인지 관련 연구"에 해당하는 토픽 6의 경우 분석을 진행한 기간 동안 점진적으로 증가하였다. <Fig. 12>는 도출된 토픽의 연도별 점유율 변화를 나타낸 그래프이다.

Ⅴ. 결론 및 향후 연구과제

1. 결론

우리나라 정부는 '국가전략 기술 육성방안'의 12대 국가 전략기술의 하나로 첨단 이동 수단을 선정하였으며, 이 중 하나인 자율주행자동차에 관한 연구가 국내외에서 증가하고 있다. 이에 자율주행자동차와 관련된 연구의 동향을 살펴보는 것이 필요하다. 또한, 주요 연구 동향을 분석하여 자율주행 연구개발사업 추진 시 강화할 강점과 연구가 필요한 부분을 도출하는 것이 필요하다. 이를 위해 본 연구는 자율주행자동차의 연구 동향을 분석하기 위한 토픽 모델링 기반 연구 동향 분석 방법론을 개발하였다. 연구 동향 분석을 위해 국가과학기술지식정보서비스에서 제공 중인 연구개발 성과물 중 학술 논문을 수집하여 본 연구에서 개발한 토픽 모델링 기반 연구 동향 분석 방법론을 적용하여 분석하였다. 주요 특징 추출 결과, "Control", "Sensor", "Safety", "Driver", "Simulation"의 키워드가 중요도가 높은 것으로 도출되었으며, 연도별로 다른 주요 키워드가 도출되어, 연구의 동향이 변하는 것을 확인하였다. 다음으로 토픽 모델링을 통해 6개의 토픽을 도출하였고, 토픽별 주요 연구 주제들을 도출하였다. 분석 결과, 자율주행 관련 법률 연구, 라이더 센서 관련 연구, 항법 장치 연구, 레이더 시스템에 관련된 연구, 자율주행자동차 보안 및 표준 관련 연구, 센서 기반 객체 분류 및 인지 관련 연구들이 도출되어 자율주행과 관련된 기술들의 연구들이 많았음을 발견할 수 있었다. 다만, 새로운 기술들이 폭발적으로 증가하고 있는 시점에서 신기술의 사회에 적용하는 문제에 고민이 필요해 보인다. 자율주행 차량은 점진적으로 점유율이 증가할 것으로 생각되며 이러한 이유로 혼재기 상황에서의 연구와 운영에 관한 연구도 필요할 것으로 판단된다.

2. 향후 연구과제

본 연구는 자율주행자동차의 연구 동향 분석을 위해 토픽 모델링 기반 연구 동향 분석 방법론을 개발하여 자율주행 연구들을 분석하였으나, 몇 가지 연구의 한계가 존재한다. 우선, 국내 자율주행 관련 국가 주도형 R&D의 성과로 등록된 학술 논문 데이터만 사용하였다는 점이다. 성과로 등록된 학술 논문 이외에도 국내외 자율주행 관련 특허 및 보고서 민간의 자율주행 연구 내용 등 다양한 데이터를 추가로 이용한다면, 국가 주도 및 민간의 자율주행 연구의 비교와 국내외 자율주행 연구의 비교 등 본 연구에서 발견하지 못한 추가적인 연구 주제들을 발견할 수 있을 것으로 판단된다.

두 번째로, 본 연구는 특정 연구 영역과 분야를 지정하지 않았다는 점이다. 자율주행 기술의 경우 전기, 전자, 기계공학, 컴퓨터 공학, 자동차 공학, 교통 공학 등 다양한 분야에서 연구가 필요하다. 특정 연구 분야에 초점을 두어 동향 분석을 진행한다면 특정 분야에 대한 정밀한 동향을 도출할 수 있을 것으로 판단된다.

세 번째로, 본 연구에서는 TF-IDF 기반의 주요 특징을 추출하여 주제를 발견하였다는 점이다. 텍스트 데이터의 특성상 단어뿐만 아니라, 합성어 및 단어의 관계를 벡터화하면 보다 유의미한 결과를 도출하는 것이 가능할 것으로 판단된다. 추후 보다 고도화된 기법인 Word2Vec 등으로 주요 특징들을 벡터화하여 분석한다면, 주요 연구 주제 및 동향들을 추출할 수 있을 것으로 판단된다.

마지막으로, 2021년부터 시작한 자율주행기술개발혁신사업의 연구성과들을 고려하지 못하였다는 점이다. 자율주행기술개발혁신사업은 다부처에서 발주하여 자율주행과 관련된 다양한 연구들을 수행하고 있어 이를 고려하여 분석을 수행한다면, 본 연구에서 발견하지 못한 연구 주제들을 발견할 수 있을 것으로 판단되며, 이를 통해 더욱 정밀하게 연구 동향을 도출할 수 있을 것으로 판단된다.

ACKNOWLEDGEMENTS

본 연구는 국토교통부/국토교통과학기술진흥원의 지원으로 수행되었음(과제번호 RS-2022-00141102).

REFERENCES

Blei, D. M.(2012), “Probabilistic topic models”, *Communications of the ACM(Association for Computing Machinery)*, vol. 55, no. 4, pp.77-84.

Blei, D. M., Ng, A. Y. and Jordan, M. I.(2003), “Latent dirichlet allocation”, *Journal of Machine Learning Research*, vol. 3, no. Jan, pp.993-1022.

Bouma, G.(2009), “Normalized (pointwise) mutual information in collocation extraction”, *Proceedings of GSCL(German Society for Computational Linguistics)*, vol. 30, pp.31-40.

Faisal, A., Yigitcanlar, T., Kamruzzaman, M. and Paz, A.(2021), “Mapping two decades of autonomous vehicle research: A systematic scientometric analysis”, *Journal of Urban Technology*, vol. 28, no. 3-4, pp.45-74.

Gandia, R. M., Antonialli, F., Cavazza, B. H., Neto, A. M., Lima, D. A. D., Sugano, J. Y., Nicolai, I. and Zambalde, A. L.(2019), “Autonomous vehicles: Scientometric and bibliometric review”, *Transport Reviews*, vol. 39, no. 1, pp.9-28.

Hacohen, S., Medina, O. and Shoval, S.(2022), “Autonomous Driving: A Survey of Technological Gaps Using Google Scholar and Web of Science Trend Analysis”, *IEEE Transactions on Intelligent Transportation Systems*, vol. 23, no. 11, pp.21241-21258.

Im, I., Song, J. I., Lee, J. Y. and Hwang, K. Y.(2017), “Analysis of the perception of autonomous vehicles using text mining technique”, *The Journal of the Korea Institute of Intelligent Transport Systems*, vol. 16, no. 6, pp.231-243.

Jang, S. Y. and Jung, S. H.(2021), “An Analysis of the Research Trends for Urban Study using Topic Modeling”, *Journal of the Korea Academia-Industrial Cooperation Society*, vol. 22, no. 3, pp.661-670.

Kim, A., Jeong, S. H., Choi, H. B. and Kim, H. H.(2018), “Analysis of response to transportation policy for particulate matter reduction using regression analysis and text mining”, *Korea Information Processing Society, The KIPS Fall Conference*, pp.277-280.

Kim, G. L.(2021), “A Study on the Analysis of R&D Trends and the Development of Logic Models for Autonomous Vehicles”, *Journal of Digital Convergence*, vol. 19, no. 5, pp.31-39.

Kim, N., Lee, D., Choi, H. and Wong, W. X. S.(2017), “Investigations on techniques and applications

of text analytics", *The Journal of Korean Institute of Communications and Information Sciences*, vol. 42, no. 2, pp.471-492.

Lim, S. Y., Yi, M. S., Jin, G. H. and Shin, D. B.(2014), "A study on the research trends in the area of geospatial-information using text-mining technique focused on national R&D reports and theses", *Spatial Information Research*, vol. 22, no. 4, pp.11-20.

Na, S. T., Kim, J. H., Jung, M. H. and Ahn, J. E.(2016), "Trend Analysis using Topic Modeling for Simulation Studies", *Journal of the Korea Society for Simulation*, vol. 25, no. 3, pp.107-116.

National Science & Technology Information Service, https://www.ntis.go.kr/ThAbout.do, 2022.11.10.

Oh, C. S., Lee, Y. T. and Ko, M.(2016), "Establishment of ITS Policy Issues Investigation Method in the Road Section applied Textmining", *The Journal of the Korea Institute of Intelligent Transport Systems*, vol. 15, no. 6, pp.10-23.

Park, J. S., Hong, S. G. and Kim, J. W.(2017), "A study on science technology trend and prediction using topic modeling", *Journal of the Korea Industrial Information Systems Research*, vol. 22, no. 4, pp.19-28.

Park, S., Park, S., Jeong, H., Yun, I. and So, J.(2021), "Scenario-mining for level 4 automated vehicle safety assessment from real accident situations in urban areas using a natural language process", *Sensors*, vol. 21, no. 20, p.6929.

Park, S., So, J. J., Ko, H., Jeong, H. and Yun, I.(2019), "Development of Safety Evaluation Scenarios for Autonomous Vehicle Tests Using 5-Layer Format (Case of the Community Road)", *The Journal of the Korea Institute of Intelligent Transport Systems*, vol. 18, no. 2, pp.114-128.

Röder, M., Both, A. and Hinneburg, A.(2015), "Exploring the space of topic coherence measures", *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining,* February, pp.399-408.

Sievert, C. and Shirley, K.(2014), "LDAvis: A method for visualizing and interpreting topics", *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces,* June, pp.63-70.

Yang, M., Lee, S., Park, K., Choi, K. and Kim, T.(2021), "A Study on Analysis of national R&D research trends for Artificial Intelligence using LDA topic modeling", *Journal of Internet Computing and Services*, vol. 22, no. 5, pp.47-55.

Yu, Y. L.(2017), *Analysis of media coverage on 2015 revised curriculum policy using big data analysis*, Unpublished Doctoral Dissertation, Department of Education, Graduate School of Seoul National University.