ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2019 March: 25(1): 1~19 http://dx.doi.org/10.13088/jiis.2019.25.1.001

텍스트 마이닝 기법을 활용한 인공지능 기술개발 동향 분석 연구: 깃허브 상의 오픈 소스 소프트웨어 프로젝트를 대상으로*

정지선
한양대학교 일반대학원 경영학과
(*skyhee84@hanyang.ac.kr*)

김동성
한양대학교 일반대학원 경영학과
(*paulus82@hanyang.ac.kr*)

이홍주
가톨릭대학교 경영학과
(*hongjoo@catholic.ac.kr*)

김종우
한양대학교 경영대학 경영학부
(*kjw@hanyang.ac.kr*)

제4차 산업혁명을 이끄는 주요 원동력 중 하나인 인공지능 기술은 이미지와 음성 인식 등 여러 분야에서 사람과 유사하거나 더 뛰어난 능력을 보이며, 사회 전반에 미치게 될 다양한 영향력으로 인하여 높은 주목을 받고 있다. 특히, 인공지능 기술은 의료, 금융, 제조, 서비스, 교육 등 광범위한 분야에서 활용이 가능하기 때문에, 현재의 기술 동향을 파악하고 발전 방향을 분석하기 위한 노력들 또한 활발히 이루어지고 있다. 한편, 이러한 인공지능 기술의 급속한 발전 배경에는 학습, 추론, 인식 등의 복잡한 인공지능 알고리즘을 개발할 수 있는 주요 플랫폼들이 오픈 소스로 공개되면서, 이를 활용한 기술과 서비스들의 개발이 비약적으로 증가하고 있는 것이 주요 요인 중 하나로 확인된다. 또한, 주요 글로벌 기업들이 개발한 자연어 인식, 음성 인식, 이미지 인식 기능 등의 인공지능 소프트웨어들이 오픈 소스 소프트웨어(OSS: Open Sources Software)로 무료로 공개되면서 기술 확산에 크게 기여하고 있다. 이에 따라, 본 연구에서는 온라인상에서 다수의 협업을 통하여 개발이 이루어지고 있는 인공지능과 관련된 주요 오픈 소스 소프트웨어 프로젝트들을 분석하여, 인공지능 기술 개발 현황에 대한 보다 실질적인 동향을 파악하고자 한다. 이를 위하여 깃허브(Github) 상에서 2000년부터 2018년 7월까지 생성된 인공지능과 관련된 주요 프로젝트들의 목록을 검색 및 수집하였으며, 수집 된 프로젝트들의 특징과 기술 분야를 의미하는 토픽 정보들을 대상으로 텍스트 마이닝 기법을 적용하여 주요 기술들의 개발 동향을 연도별로 상세하게 확인하였다. 분석 결과, 인공지능과 관련된 오픈 소스 소프트웨어들은 2016년을 기준으로 급격하게 증가하는 추세이며, 토픽들의 관계 분석을 통하여 주요 기술 동향이 '알고리즘', '프로그래밍 언어', '응용분야', '개발 도구'의 범주로 구분하는 것이 가능함을 확인하였다. 이러한 분석 결과를 바탕으로, 향후 다양한 분야에서의 활용을 위해 개발되고 있는 인공지능 관련 기술들을 보다 상세하게 구분하여 확인하는 것이 가능할 것이며, 효과적인 발전 방향 모색과 변화 추이 분석에 활용이 가능할 것이다.

주제어 : 인공지능, 기술 동향, 오픈 소스 소프트웨어, 깃허브, 텍스트 마이닝

논문접수일 : 2019년 1월 18일 논문수정일 : 2019년 3월 8일 게재확정일 : 2019년 3월 11일
원고유형 : 학술대회(급행) 교신저자 : 김종우

* 이 논문 또는 저서는 2017년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2017S1A3A2066740)

1. 서론

인공지능(AI: Artificial Intelligence) 기술의 발전으로 인하여, 다양한 분야에서 이를 기반으로 한 서비스와 응용 기술의 개발이 활발히 이루어지고 있다. 가트너(Gartner)의 2018년 10대 전략 기술 보고서에 따르면, 향후 5년 동안 혁신적 잠재력을 갖고 있는 기술로 인공지능 기반의 3가지 기술 분야들을 우선으로 언급하였다[1]. 인공지능 기술은 의료, 금융, 제조, 서비스, 교육을 비롯하여 법률 분야까지 광범위하게 활용이 가능하며, 기술의 파급력 또한 크기 때문에 향후 발전 가능한 분야들을 분석하고 예측하기 위한 연구들이 다수 이루어지고 있다(Bae et al., 2017; Chung et al., 2017; Chung et al., 2018). 인공지능 관련 기술 동향의 분석은 현재 급부상하고 있는 기술들을 활용하여 새로운 가치 창출의 기회를 모색하는 것에서 나아가, 기술의 영향 범위가 개인, 기업, 산업, 경제 및 법·제도 등 사회 전반에 걸쳐 있기 때문에 미래사회의 변화를 예측하기 위한 중요 자료로도 활용이 가능하다.

인공지능 기술의 급속한 발전 배경에는 학습, 추론, 인식 등의 복잡한 인공지능 알고리즘을 개발할 수 있는 주요 플랫폼들이 오픈 소스로 공개되면서, 이를 활용한 기술과 서비스들의 개발이 비약적으로 증가하고 있는 것이 주요 요인 중 하나로 확인된다(Nam, 2016). 또한, 주요 글로벌 기업들이 개발한 자연어 인식, 음성 인식, 이미지 인식 등의 인공지능 기반 소프트웨어들이 오픈 소스 소프트웨어(OSS: Open Sources Software)로 공개되면서 기술 확산에 크게 기여하고 있다.

한편, 기술개발 동향에 대한 분석과 관련하여 고려해야 할 주요 사항 중 하나로는, 동향을 파악하고자 하는 기술의 특성을 고려한 적절한 분석 대상 데이터의 선정이 필요하다. 인공지능 관련 기술의 경우, 그 발전 속도가 빠르고 분야 또한 다양함에 따라, 외부에 공개되기까지는 시간적 지연이 다소 존재하는 논문이나 특허 정보를 활용하는 것 보다, 기술 개발의 계획부터 배포, 지속적 업데이트까지 확인이 가능한 오픈 소스 소프트웨어 프로젝트를 분석하는 것이 보다 실증적인 분석 결과의 도출이 가능할 것이다. 또한, 인공지능 관련 기술의 특성상 다수의 기술들이 소프트웨어의 형태로 개발되고 있는 것도 본 연구에서 분석 대상 데이터의 선정 시 고려한 중요 사항 중 하나이다.

이러한 배경에서, 본 연구는 깃허브(github) 상의 인공지능과 관련된 소프트웨어 개발 프로젝트들을 분석하고, 보다 실증적인 인공지능 기술 개발의 동향 파악을 꾀하였다. 깃허브는 온라인 상에서 다수의 개발자와 참여자들의 소스코드 기여를 통하여 소프트웨어 개발이 이루어지는 대표적인 소셜 코딩(social coding) 플랫폼이며, 구글, 페이스북, 마이크로소프트 등 다수의 글로벌 선도 기업들이 오픈 소스로 인공지능 관련 주요 기술들을 공개하고 있다. 또한, 인공지능과 관련된 주요 연구 성과들도 논문 발표와 함께, 연구의 재현 및 후속 연구가 가능하도록 깃허브 상에 공개하고 있는 추세이다.

이에 따라, 본 연구에서는 2000년부터 2018년 7월까지 깃허브에서 생성된 소프트웨어 개발 프로젝트들 중에서, 소프트웨어의 주요 특징과 기술 분야를 의미하는 토픽 정보를 활용하여 인공지능과 관련된 오픈 소스 소프트웨어 프로젝트

1) https://www.gartner.com/smarterwithgartner/gartner-top-10-strategic-technology-trends-for-2018

들을 선별 및 주요 정보들을 수집하였다. 실제 기술 동향의 분석은 프로젝트들의 토픽 정보를 대상으로 텍스트 마이닝 기법을 활용하였으며, 소프트웨어 개발 프로젝트들에서 함께 사용되는 토픽들의 관계를 바탕으로 토픽 네트워크를 구성하였다. 기술과 관련된 용어들의 네트워크 관계 분석은 해당 기술과 관련된 주요 동향을 파악하고, 연도별로 네트워크를 구성함으로써 시간에 따른 기술의 발전 추이를 확인할 수 있다(Kho et al., 2013).

이후 본 논문의 구성은 다음과 같다. 2장에서는 인공지능 관련 기술 동향 연구, 깃허브 오픈 소스 소프트웨어 현황에 대하여 검토한다. 3장에서는 깃허브 상의 오픈 소스 소프트웨어 정보를 활용한 인공지능 기술개발 동향의 분석 방안을 제시하며, 4장에서는 실증적 분석을 기반으로 한 주요 인공지능 기술들의 개발 현황을 연도별로 확인한다. 마지막 5장에서는 결론과 추후 연구 방안에 대하여 제시한다.

2. 관련 연구

2.1 인공지능 기술 동향 분석

급격하게 변화하는 과학기술의 발달로 인하여 새로이 등장하는 다양한 기술에 대한 수준이나 동향을 파악하기 위해 많은 연구들이 지속적으로 수행 되어왔다. 특히, 기술 동향 및 변화에 대한 연구는 텍스트 마이닝 기법과 네트워크 분석 등을 토대로 하여 특허 관련 데이터나 문헌 및 서지 정보를 활용한 분석 방안이 여러학문에서 제시되고 있다(Choi et al., 2011; Kho et al., 2013; Kim and Kim, 2014). 이에, 인공지능이 각 산업 분야의 지능화와 가속화를 심화하면서 앞서 설명한 방안들을 통한 인공지능 기술에 대한 연구 동향 분석 또한 상당히 진행되고 있다.

우선 특허데이터는 출원 및 등록 날짜, 등록자, 특허 제목, 기술 요약, 인용 정보, 상세 기술, 도면, 절차도 등 다양한 정보를 포함하고 있고(Jun, 2013) 분석 방안에 따라 기술 동향이나 관련 시장 동향 등의 전반적인 흐름을 볼 수 있기 때문에(Tseng, Y. H et al., 2007) 그 활용가치가 높게 평가되며 기술 동향 분석에 빈번히 활용되고 있다.

이와 관련된 연구로 웹스(WIPSON) 특허 데이터베이스를 기반으로 인공지능 기술 관련 특허 정보를 수집하여 한국, 미국, 일본, 유럽, 중국의 인공지능 동향 분석을 수행한 연구가 있다. 인공지능 기술을 5개의 핵심기술과 15개의 세부 기술로 분류하여 국가별 기술 수준을 분석하였으며, 전 세계적으로는 언어 이해 기술과 시각 기술 등이 우세한 것을 확인하고, 국내에서는 행동 인식 기술과 음성 처리 기술, 시각 기술 등의 특허가 다수 발생하고 있는 것을 확인하였다(Rho, 2017).

유사한 연구로 국내외의 인공지능 기술 수준에 대한 비교분석을 통하여 국내의 향후 발전 가능성이 높은 세부적 기술을 도출하고, 이를 토대로 발전 방향성을 제시한 연구가 있다. 국내외의 특허 데이터 중 ‘인공지능’ 키워드 검색으로 도출된 데이터를 기반으로 키워드 네트워크 분석 및 국제특허분류(IPC: International Patent Classification)를 기준으로 공백 기술 분석을 수행하여 인공지능 분야의 기술 동향을 파악하였다. 이를 통해 국내 인공지능 관련 기술 개발 건수는 미국, 유럽 등 선진국 대비 1.2% 수준이었으며, 주요 개발 분야의 경우 데이터 인식 기술,

디지털 정보 전송 기술 등에서 상대적으로 부족한 것으로 나타났다(Chung et al., 2018).

이 외에도 인공 지능 기술 발전의 결정 요인을 명확히 하기 위해 미국, 일본, 중국, 유럽 및 특허협력 조약(PCT: Patent Cooperation Treaty)의 특허 데이터를 활용하여 (1) 생물학적 기반 모델, (2) 지식 기반 모델, (3) 특정 수학적 모델, 그리고 (4) 기타 인공 지능 기술 모델 등 4가지 기술 유형에 대한 인공 지능 기술의 추세와 우선 순위 변화를 분해 프레임워크(decomposition framework)를 적용하여 분석한 연구가 있다. 그 결과로, 생물학과 지식 기반 모델에서 특정 수학적 모델과 기타 인공 지능 기술로 특허 발명의 우선 순위가 이동하고 있음을 확인하였으며, 인공지능 기술 특허의 특징은 국가별, 기업별로 다름을 발견하였다(Fujii and Managi, 2018). 같은 일환으로 문헌에 수록되어 있는 정보를 통해 인공지능 기술 동향을 파악한 연구도 다수 진행되었다.

또한, 'Web of Science'에서 한국인 저자가 게재한 SCIE(Science Citation Index Expanded) 학술지의 논문들 중 인공지능과 관련된 논문을 수집 및 분석하여, 국내 연구자들이 기 수행한 인공지능 관련 주요 연구 분야 및 동향을 도출한 연구도 존재한다(Chung et al. 2017). 이를 통해 이론적 연구가 하향세를 보이며, 기술적 연구가 상향세를 보인다는 것을 확인하였다. 또한 산업 간 융복합이 활발하게 이루어지고 있음을 확인하였다(Chung et al., 2017).

또 다른 연구에서는 공간 분석(spatial analysis)과 사회 네트워크 분석(social network analysis)을 활용하여 SCIE와 CPCI-S(Conference Proceedings Citation Index-Science)의 데이터를 분석함으로써 인공지능의 최근 이슈 및 기술 동향에 대해 확인하였다. 또한 키워드 분석을 통해 연구 선호도를 파악하였고, 최근 몇 년 동안 새로운 모델 및 응용 분야를 식별하는 데 도움이 되는 공존 빈도가 높은 관련 키워드를 확인하였다(Niu et al., 2016).

2.2 인공지능 기술과 오픈 소스 소프트웨어

인공지능이 미래의 최대 성장동력으로 인식되면서 국내외 주요 기업들이 인공지능 관련 기술

〈Table 1〉 Review of research related to AI technology trend analysis using patent and literature data

Authors (year)	Research overview
Choi, J. H and S. H. Jun, (2018)	Analysis of artificial intelligence technology trend using bayesian inference-based statistical analysis with patent data of artificial intelligence technology.
Chung, M. S and J. Y. Lee, (2018)	Suggestion of core technology and possible growth research related to artificial intelligence using text mining and topic modeling by collecting papers related to artificial intelligence in SCIE journal.
Park, J. Y., (2018)	Suggestion of directions of artificial intelligence technology research and trend analysis of core artificial Intelligent technology using quantitative analysis of patent information.
Park, J. S. et al., (2017)	Research on the core technologies of artificial intelligence by using the LDA topic modeling for artificial intelligence abstracts on US patent documents.
Niu, J et al., (2016)	Research on recent issues and technology trends in artificial intelligence by analyzing data from SCIE and CPCI-S using spatial analysis and social network analysis.

개발에 대거 참여하고 있다. 이에 따라, 인공지능 적용 분야가 의료기술 향상, 신약 개발, 금융거래, 유전자 분석 등 다양한 방면으로 빠르게 확대되고 있으며, 글로벌 기업들은 인공지능 생태계를 만들어 선도하겠다는 공통된 전략을 가지고 있다. 이를 위해 더 많은 개발자 우군을 확보하고 인공지능 생태계 진화를 앞당기기 위하여 공통적으로 인공지능 소프트웨어 기술을 오픈 소스로 공개하고 있다.

오픈 소스 소프트웨어란 소프트웨어의 저작권자가 해당 소스코드를 공개해 이를 사용, 복제, 수정, 배포할 수 있는 권한을 부여한 소프트웨어를 의미한다. 이러한 오픈 소스 소프트웨어는 소스 코드 공개로 인해 신기술이나 핵심 기술을 보다 쉽게 접근 및 습득할 기회가 높아지며 다양한 개발에 참여하면서 개인 역량 향상의 기회가 제공되기도 한다. 이외에도 시스템 개발 기간 단축, 비용 절감, 관련 정보 획득 용이 등의 장점을 지닌다(Bonaccorsi and Rossi, 2003).

이처럼 오픈 소스 소프트웨어는 개방적 협업을 통해 경제적 효율성, 시장 경쟁 촉진, 기술혁신 가속화 및 인력 양성 등의 주요한 가치를 지닌다. 또한 기업 입장에서 오픈 소스 소프트웨어는 자사의 소프트웨어 저변을 확대하는 데도 유용하게 쓰일 수 있다. 실제로 해외 조사에 따르면 상용 소프트웨어의 96%는 오픈 소스 소프트웨어를 기반으로 개발되고 있으며, 국내 경우는 기업의 95%가 오픈 소스 소프트웨어를 활용하고 있다(Synopsys, 2019; Kim 2018).

이러한 추세에 따라, 본 연구에서는 온라인상에서 다수의 협업을 통하여 개발이 이루어지고 있는 인공지능과 관련된 주요 오픈 소스 소프트웨어 프로젝트들을 수집 및 분석하여, 인공지능 기술 개발 현황에 대한 보다 실질적 동향을 파악하고자 하였다. 이를 기반으로, 다양한 분야에서 활용을 위해 개발되고 있는 인공지능 관련 기술들을 보다 상세하게 구분하여 확인하는 것이 가능하며, 효과적인 발전 방향 모색과 변화 추이 분석에 활용하는 것이 가능할 것이다.

3. 분석 방안

3.1 분석 데이터

본 연구의 분석 데이터는 깃허브에서 제공하는 API(Application Programming Interface)[2]를 활용하여 인공지능 기술과 관련된 소프트웨어 개발 프로젝트들을 검색하여 수집하였다. 깃허브 API는 소프트웨어 개발 프로젝트의 소스 코드 변경, 참여자 활동 내용, 기타 주요 변경 사항 등과 같이 프로젝트의 주요 기초 정보들에 대한 검색과 수집이 가능하다. 본 연구에서는 인공지능 관련 기술들의 급속한 발전 동향을 고려하여, 연구 데이터의 검색 및 수집 시점을 2018년 8월 이전까지의 주요 데이터로 선정하였다.

이에 따라, 2000년부터 2018년 7월까지 깃허브 상에서 생성된 프로젝트를 대상으로, 프로젝트의 주요 특징을 나타내는 토픽 정보들 중에서 인공지능과 관련한 주요 키워드들이 포함된 프로젝트들을 검색 및 수집하였다. 또한, 실제 소프트웨어 개발 프로젝트만을 분석 대상으로 수집하기 위하여, 다음과 같이 소프트웨어 개발을 위해 사용하는 프로그래밍 언어를 검색 기준으로 함께 활용하였다. 프로젝트 검색 키워드로 사

2) https://developer.github.com/v3/

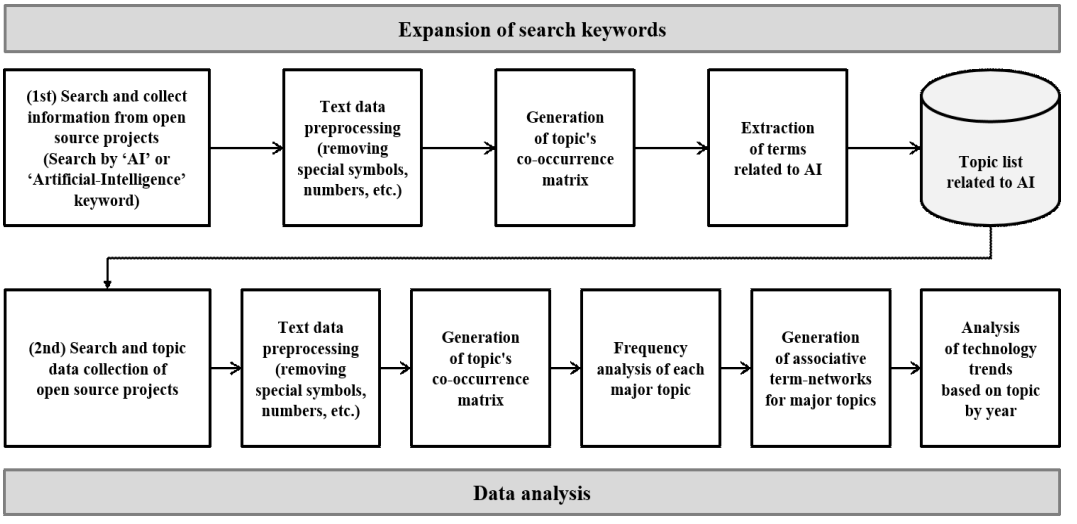
용된 프로그래밍 언어는 ‘Python’, ‘JavaScript’, ‘Java’, ‘C++’, ‘C#’, ‘Jupyter-Notebook’, ‘C’, ‘Swift’, ‘Go’, ‘R’, ‘Perl’, ‘Matlab’, ‘Assembly’, ‘Objective-C++’, ‘Cuda’ 로 총 15개이다. 이를 바탕으로 검색 및 수집된 프로젝트는 키워드별로 중복 검색된 프로젝트들을 제외한 뒤, 총 26,505개의 프로젝트 목록과 관련 데이터들이 수집되었다.

수집된 주요 데이터 현황은 프로젝트 이름, 프로젝트 소유자(생성자), 소유자 구분(개인/조직), 프로젝트 생성일, 프로젝트의 소프트웨어 개발 프로그래밍 언어, 프로젝트에 저장된 소스 코드의 파일 크기, 프로젝트의 기여자 수, 프로젝트의 인기도를 의미하는 스타(star) 수, 프로젝트의 소스 코드를 수정하거나 사용하기 위해 복사한 횟수를 의미하는 포크(fork) 수, 프로젝트의 주요 변경 사항 확인과 알림을 받기 위해 사용자들이 즐겨 찾기로 등록한 횟수를 의미하는 왓치(watch) 수 등으로 구성되어 있다.

3.2 분석 절차

본 연구는 다음과 같은 분석 절차를 바탕으로 수행되었다(<Figure 1> 참조). 본 연구에서 소프트웨어 개발 프로젝트의 검색은 개발 중인 소프트웨어의 특징을 나타내는 토픽 정보를 대상으로 하였으며, 인공지능을 의미하는 ‘AI’와 ‘Artificial Intelligence’ 외에 이와 높은 연관성을 띄는 토픽을 사용한 경우도 검색 및 수집하기 위해 검색 키워드의 확장을 진행하였다.

검색 키워드 확장은 ‘AI’와 ‘Artificial Intelligence’를 토픽으로 사용한 프로젝트들을 검색하고, 해당 프로젝트들이 사용한 토픽들을 재검토하여 대소문자 통일, 불필요한 기호 및 숫자 제거 등의 텍스트 정규화를 바탕으로 토픽들의 동시출현 행렬을 작성하였다. 이를 기반으로 ‘AI’ 및 ‘Artificial Intelligence’와 함께 사용되는 상위 토픽들을 인공지능 기술과 연관된 토픽 목록으로 정의하고 소프트웨어 개발 프로젝트들의 검색과 수집을 위한 최종 키워드로 활용하였다. 이에 해



〈Figure 1〉 Research Procedure

당하는 주요 키워드는 'Machine Learning', 'Neural Network', 'Deep Learning', 'Tensorflow' 등이 있다.

확장된 인공지능 연관 키워드를 바탕으로 검색 및 수집된 공개 소프트웨어 개발 프로젝트들을 대상으로 실제 인공지능 연구 동향 분석을 위한 토픽들을 최종 추출하였다. 추출된 토픽들은 토픽 간의 네트워크 생성 시 동일한 토픽이나 대소문자 구분 또는 특수기호 등으로 인하여 다른 개체로 인식되는 문제를 해결하기 위하여 텍스트 전처리를 진행하였으며, 이를 바탕으로 주요 토픽들의 출현 빈도와 토픽 네트워크 구성을 연도별로 구분하여 분석하였다. 본 연구에서의 데이터 분석은 파이썬 프로그램 환경에서 수행되었으며, 텍스트 분석은 nltk[3]를 활용하였고 네트워크 분석에는 networkx[4] 패키지를 활용하였다.

3.3 분석 데이터 선정

인공지능 관련 주요 키워드들을 활용하여 수집된 공개 소프트웨어 개발 프로젝트들 중에서도 실제 개발이 진행되고 있는 프로젝트만을 선별하기 위하여, 프로젝트 소스 코드의 크기 정보를 참고하였다. 이에 따라, 프로젝트 소스 코드의 크기가 '0'인 2,455개의 프로젝트들은 잠정적으로 개발 활동이 없는 것으로 보고 분석에서 제외하였다. 최종적으로 24,050개의 프로젝트를 본 연구의 분석 대상 프로젝트로 선정하였으며, 이들 프로젝트들의 토픽 정보에서 추출된 15,752개의 용어들을 인공지능 기술 동향 분석에 활용하였다.

3.4 인공지능 토픽 네트워크 구성

본 연구에서의 토픽 간 네트워크 구성은 인공지능 소프트웨어 개발 프로젝트들의 토픽 정보들을 바탕으로, 동일한 프로젝트 내에서 사용된 토픽들의 관계를 동시출현 행렬로 작성하였다. 즉, 인공지능 기술과 관련된 주요 토픽들이 사회 네트워크에서의 노드(node)이며, 특정 프로젝트의 토픽으로 함께 사용된 경우를 엣지(edge)로 나타내어 네트워크를 구성하였다. 이를 기반으로 각각의 토픽에 대하여 사회 네트워크 분석의 연결 중심성(degree centrality) 척도를 산출하였다.

3.5 연도별 기술 동향 분석 방안

본 연구에서는 인공지능 관련 기술의 동향을 연도별로 구분한 뒤, 기술 개발이 지속적으로 활발히 이루어지는 분야 또는 급격한 성장 추이를 보이는 분야 등의 변화 추이를 확인하기 위하여, 기존 기술 동향 분석과 관련된 연구에서의 분석 방안을 참조하였다(Han et al., 2009; Kim and Kim, 2014). 기존 선행 연구에서는 특정 기술 분야와 관련된 주요 용어들에 대하여, 용어의 출현 빈도와 연결 중심성을 기준으로 해당 기술의 발전과 융합 정도를 확인하였다(Kim and Kim, 2014).

이에 따라, 본 연구에서는 주요 토픽의 빈도수와 연결중심성을 산출하였으며, 토픽별로 X축, Y축의 좌표 평면에 산출된 값을 나타내어 분석하였다. 이때, 연도별로 산출된 토픽들의 출현 빈도와 연결 중심성 값의 최댓값과 최솟값

3) https://pypi.org/project/nltk/
4) https://pypi.org/project/networkx/

이 상이한 점을 고려하여, 다음과 같이 표준화하여 연도별 기술 동향 비교분석에 활용하였다(<Formula 1- Formula 2> 참조).

$$\boldsymbol{A_x}(t) = \frac{DC(A,t) - \mu(DC(t))}{\sigma(DC(t))} \quad (1)$$

$$A_y(t) = \frac{Freq(A,t) - \mu(Freq(t))}{\sigma(Freq(t))} \quad (2)$$

DC(A, t) : Degree centrality of topic(t) by year(A)
Freq(A, t) : Frequency of topic(t) by year(A)

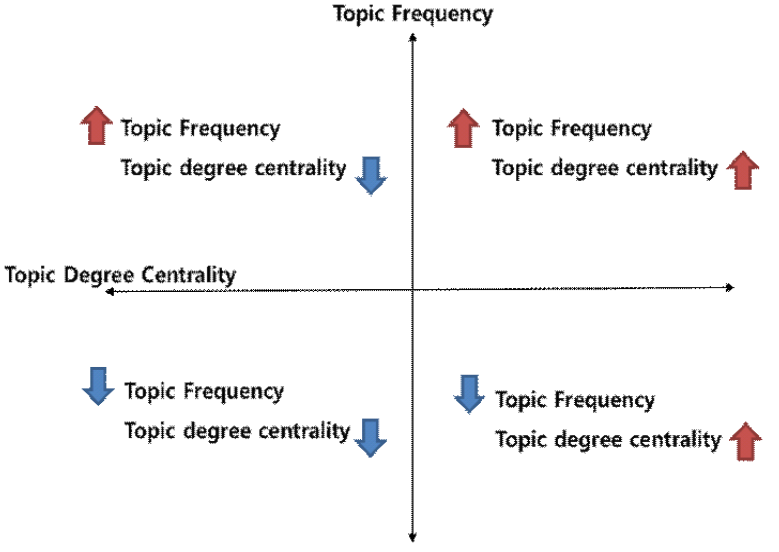
표준화된 값을 기반으로 좌표 평면 상에 표기되는 토픽들의 주요 의미는 출현 빈도의 높고 낮음은 해당 토픽과 관련된 기술 개발의 증감을 나타낸다. 또한, 연결 중심성의 높고 낮음의 의미는 다른 용어들과 함께 사용되는 정도의 증감을 의미한다(<Figure 2> 참조).

이를 기반으로 출현 빈도와 연결 중심성 모두가 높은 토픽은 기술 개발이 다수 이루어지고 있으며, 여러 기술 분야에서 활용되어 성장하고 있는 기술로 확인할 수 있다. 이와 반대로, 출현 빈도와 연결 중심성 모두가 낮은 토픽은 관련 기술의 개발이 적으며, 다른 기술 분야에서도 언급되지 않는 발전 정도가 낮은 기술이라고 할 수 있다.

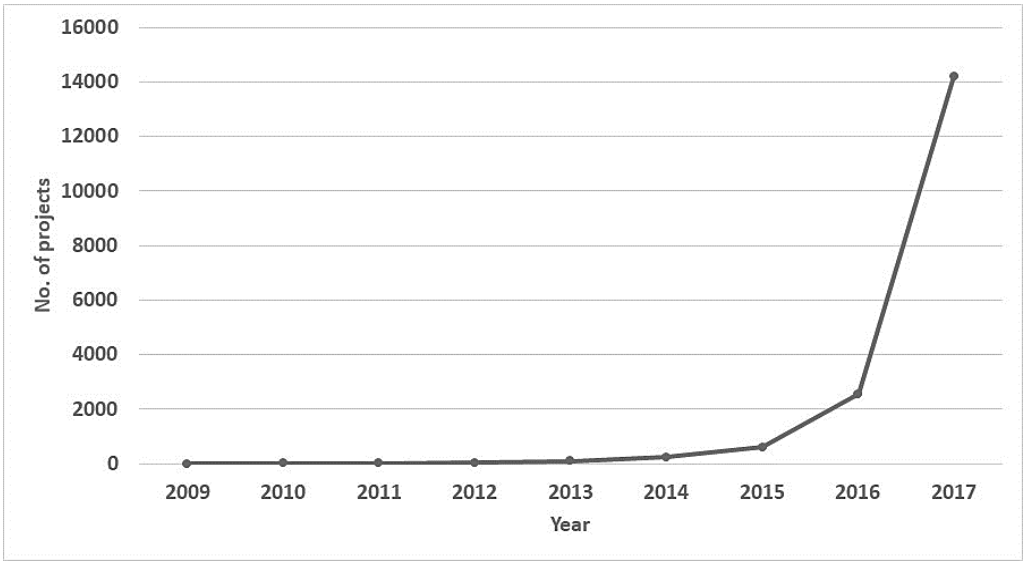
4. 분석 결과

4.1 인공지능 오픈 소스 소프트웨어 프로젝트 주요 현황

본 연구에서 분석 대상으로 수집된 인공지능 기술과 관련한 오픈 소스 소프트웨어 프로젝트들의 2009년부터 2017년까지 연도별 생성 현황은 다음과 같다(<Figure 3> 참조).



〈Figure 2〉 Technique Trend Matrix



〈Figure 3〉 Number of projects by year

우선 연도별로 생성된 프로젝트 수에 대하여 확인한 결과 최초로 프로젝트가 생성된 연도는 2009년으로 총 4개의 프로젝트가 있으며, 해당 프로젝트들은 최근까지도 주요 정보의 갱신과 소스 코드의 수정이 꾸준히 이루어지고 있다.

2013년까지 연도별로 100개 미만의 프로젝트가 생성되었으며, 2014년 229개, 2015년 597개의 프로젝트가 생성되었다. 2016년에는 인공지능 관련 오픈 소스 프로젝트의 수가 급격히 증가하였으며, 2,559개의 관련된 오픈 소스 소프트웨어 프로젝트들이 생성되었다. 2016년은 전 세계적으로 주목을 받았던 인공지능 알파고와 이세돌의 바둑 대국이 진행되었던 시기로, 인공지능 알파고가 이세돌을 승리하면서 인공지능, 딥러닝(deep learning), 기계 학습(machine learning) 등에 대한 관심과 기술 개발이 급격히 증가하기 시작한 시기이다. 2017년에 생성된 프로젝트의 수는 14,213개로 확인되었으며, 이는 2009년부터 2016년까지 생성된 프로젝트 수의 총합인 3,555개의 약 4배에 가까운 수치이다. 2018년 1월부터 7월까지 생성된 프로젝트의 수는 8,737개로 확인된다.

4.2 연도별 인공지능 오픈 소스 소프트웨어 동향

4.2.1 인공지능 관련 토픽 빈도 분석

인공지능 관련 토픽들의 빈도 분석은 프로젝트가 최초 생성된 연도인 2009년을 기준으로 2009년~2012년, 2013년~2015년, 2016년~2018년 7월까지 3개의 기간으로 구분하였으며, 각각의 기간별 출현 빈도 상위 10개의 토픽은 다음과 같다(<Table 2> 참조). 2009년부터 2012년까지 출현 빈도 상위의 토픽들은 기계 학습, 인공지능이 각각 57개, 26개이며, 소프트웨어 개발 프로그래밍 언어인 파이썬(python)이 16개로 상위 출현

〈Table 2〉 Frequency of Artificial Intelligence Related Topics Top 10

2009～2012		2013～2015		2016～2018.7	
Topic	Freq.	Topic	Freq.	Topic	Freq.
Machine Learning	57	Machine Learning	652	Machine Learning	15,842
Artificial Intelligence	26	Python	211	Deep Learning	8.299
Python	16	Artificial Intelligence	193	Python	5.174
Neural Network	15	Neural Network	149	Neural Network	4.523
Natural Language Processing	12	Deep Learning	144	Tensorflow	3.644
Java	11	Java	95	Artificial Intelligence	2.555
Genetic Algorithm	7	Natural Language Processing	78	Keras	1.766
C++	7	Data Science	57	Reinforcement Learning	1.493
Java Script	5	C++	43	Data Science	1.339
Deep Learning	5	Computer Vision	38	Convolutional Neural Network	1.278

빈도 용어로 나타난다. 이외 자연어 처리(natural language processing), 유전자 알고리즘(genetic algorithm), 딥 러닝(deep learning)이 각각 12개, 7개, 5개의 출현 빈도를 보이는 것을 확인할 수 있다.

2013년부터 2015년까지 출현 빈도 상위 토픽은 기계 학습, 파이썬, 인공지능이 각각 652개, 211개, 193개이며, 2009년부터 2012년 기간과 동일한 토픽이 상위인 것으로 나타난다. 딥 러닝이 144개의 출현 빈도를 보이며, 이전 기간과 비교하여 급격하게 증가한 것을 확인할 수 있다. 이외, 인공지능 응용 기술 분야를 의미하는 컴퓨터 비전(computer vision) 토픽이 38개의 출현 빈도를 보이며, 자연어 처리와 함께 상위 10개의 토픽으로 포함된 것으로 나타난다. 2016년부터 2018년까지 출현 빈도 상위 토픽들은 이전 기간들과 비교하여 급격한 출현 빈도 수의 증가 추이를 보이며, 딥 러닝 토픽은 출현 빈도가 8,299개로 인공지능 토픽의 출현 빈도 2,555개 보다 높은 것으로 나타났다.

이전 분석 기간에 확인되지 않았던 새로운 토픽으로는 텐서플로우(tensorflow), 케라스(keras)와 같이 알고리즘 개발 플랫폼이 각각 3,644개, 1,766개로 나타났다. 이외 강화 학습(reinforcement learning), 합성곱 신경망(convolutional neural network) 토픽이 각각 1,493개, 1,278개로 상위 10개의 토픽으로 포함되었다.

4.2.2 인공지능 관련 토픽 네트워크 분석

인공지능 관련 토픽 네트워크의 분석 결과, 각각의 기간별 연결 중심성 상위 10개의 토픽들은 다음과 같다(<Table 3> 참조). 2009년부터 2012년까지 연결 중심성 상위 토픽은 기계 학습, 인공지능, 인공 신경망(neural network)이 각각

〈Table 3〉 Degree Centrality of Artificial Intelligence Related Topics Top 10

2009~2012		2013~2015		2016~2018.7	
Topic	Cent.	Topic	Cent.	Topic	Cent.
Machine Learning	0.748	Machine Learning	0.706	Machine Learning	0.664
Artificial Intelligence	0.276	Python	0.324	Deep Learning	0.418
Neural Network	0.276	Artificial Intelligence	0.275	Python	0.319
Python	0.272	Neural Network	0.236	Neural Network	0.280
C++	0.154	Deep Learning	0.216	Artificial Intelligence	0.238
Natural Language Processing	0.126	Java	0.166	Tensorflow	0.235
Java	0.122	Data Science	0.122	Keras	0.140
Visualization	0.110	C++	0.092	Data Science	0.120
C#	0.106	Natural Language Processing	0.085	Convolutional Neural Network	0.111
Medical Imaging	0.106	Java Script	0.080	Reinforcement Learning	0.107

0.748, 0.276, 0.276으로 확인되었다. 전반적으로는, 출현 빈도 분석의 동일 기간과 유사한 것으로 확인되며, 출현 빈도 상위에서 확인되지 않은 토픽으로는 시각화(visualization)와 의학 화상(medical imaging) 토픽의 연결 중심성이 각각 0.110, 0.106으로 상위 10개의 토픽으로 포함되었다.

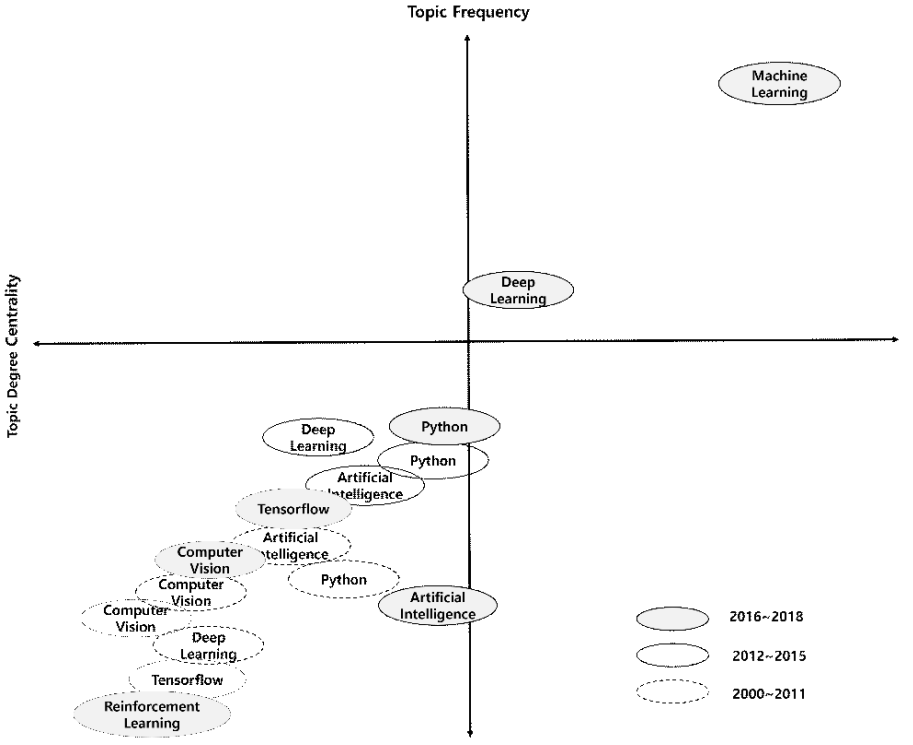
2013년부터 2015년까지 연결 중심성 상위 토픽들은 이전 분석 기간과 유사하게, 기계 학습, 파이썬, 인공지능, 인공 신경망이 상위로 확인되었다. 이외, 딥 러닝, 데이터 과학(data science) 토픽의 연결 중심성이 각각 0.216, 0.122로 새롭게 상위 10개의 토픽에 포함되었다.

2016년부터 2018년까지 연결 중심성 상위 토픽들은, 출현 빈도 상위 10개와 동일한 토픽들이 모두 확인되었다.

4.2.3 인공지능 관련 기술 동향 분석

인공지능 기술 토픽들의 기간별 출현 빈도와 연결 중심성 표준화 값을 바탕으로, 다음과 같이 좌표 평면 상에 주요 토픽들을 나타내어 연도별 기술 동향의 변화 추이를 확인하였다(<Figure 4> 참조). 기계 학습의 경우 모든 연도에서 가장 높은 출현 빈도와 연결 중심성을 보이는 것으로 나타났다. 다음으로 딥 러닝 토픽의 성장 추이가 급격히 높은 것을 확인할 수 있으며, 2009년~2012년 기간에는 출현 빈도와 연결 중심성이 모두 낮은 영역에 위치하였으나, 2013년~2015년부터 급격히 증가하여 최근에는 출현 빈도와 연결 중심성 모두 높은 기술로 확인할 수 있다.

그 다음으로는 프로그래밍 언어인 파이썬이 다른 토픽들과 비교하여 상대적으로 높은 출현 빈도와 연결 중심성을 보였으며, 2012년 이후 큰 증가 폭을 보였던 것으로 좌표 평면 상에서 확인



〈Figure 4〉 Technique Trend Matrix of Artificial Intelligence

이 가능하다. 2013년~2015년 기간에 처음 등장한 텐서플로우는 2016년~2018년에 출현 빈도와 연결 중심성이 급격히 증가하여, 딥 러닝 파이썬 다음으로 상위 영역에 위치하는 것을 확인 할 수 있다. 이외, 컴퓨터 비전, 강화 학습 등은 급격한 증감을 보이지는 않으나, 앞서 언급한 토픽들과 비교하여 상대적으로 낮은 출현 빈도와 연결 중심성을 갖는 것을 확인할 수 있다.

4.3 인공지능 오픈 소스 소프트웨어 동향 분석 결과 정리

인공지능 오픈 소스 소프트웨어들의 기술 동향 분석 결과, 2016년을 기준으로 인공지능 기술에 대한 관심 증가와 함께 오픈 소스 소프트웨어 개발을 위한 프로젝트들의 생성이 급격히 증가한 것으로 확인 된다. 2016년부터 최근 3년까지 생성된 인공지능 소프트웨어 개발 프로젝트의 수는 25,509개이며, 이는 분석 대상 전체 기간 내에 생성된 26,505개 프로젝트의 약 96%에 해당하는 수치이다.

인공지능 관련 오픈 소스 소프트웨어 프로젝트들의 특징을 나타내는 토픽들의 출현 빈도 분석 결과, 자연어 처리 기술이 지속적으로 상위에 확인되며 관련 오픈 소스 소프트웨어 개발이 꾸준히 이루어진 것을 확인할 수 있다. 2015년 까지 출현 빈도 상위 10개의 토픽에는 파이썬,

C++, 자바와 같은 프로그래밍 언어도 포함되었으나, 2016년 이후에는 파이썬을 제외한 다른 프로그래밍 언어들은 상위 10개의 토픽에서 제외되었다. 이를 대신하여 텐서플로우, 케라스와 같이 인공지능 알고리즘의 구현을 보다 손쉽게 할 수 있도록 지원하는 플랫폼이 출현 빈도 상위인 것을 확인할 수 있다. 이 밖에도 강화 학습 알고리즘, 합성곱 신경망과 같이 최근 여러 분야에서 활용되고 있는 기법들도 출현 빈도 상위 토픽으로 나타났다.

오픈 소스 소프트웨어 프로젝트에서 개발 중인 여러 토픽들 간의 관계에 기반한 네트워크 분석 결과, 연결 중심성 지표 상위의 토픽들은 출현 빈도 상위 토픽들과 유사한 것으로 나타난다. 주요 차이점으로는 2009년부터 2012년까지 출현 빈도 상위에서는 확인되지 않았던 시각화와 의학 화상 토픽이 나타났으며, 의료 분야에서 인공지능 기술의 활용을 위한 오픈 소스 소프트웨어 개발이 이루어졌음을 확인할 수 있다. 또한, 컴퓨터 비전이 2013년부터 2015년까지의 출현 빈도 상위 10개 내에서는 확인된 것과는 다르게, 연결 중심성 상위 10개에는 포함되지 않은 것으로 나타났다. 2016년부터 2018년까지의 연결 중심성 상위 토픽들은 출현 빈도 상위와 유사하였으며, 합성곱 신경망과 강화 학습의 순위가 근소한 차이로 바뀌었음을 확인할 수 있다.

본 연구에서 도출된 주요 토픽들을 기반으로 그 특성에 따라 인공지능 관련 오픈소스 소프트웨어의 개발 동향은 '알고리즘', '프로그래밍 언어', '응용분야', '개발 도구'의 4가지 유형으로 구분이 가능하다. 이러한 유형 구분을 기준으로 토픽들의 연도별 변화 추이를 확인하면, '알고리즘'의 경우 2009년~2012년 사이에는 유전자 알고리즘과 인공 신경망이 상위 토픽이었으나, 2013년~2015년에는 딥 러닝과 인공 신경망이 상위 토픽으로 확인된다. 2016년~2018년 7월까지의 토픽으로는 합성곱 신경망과 강화 학습이 상위에 나타나며, 인공지능 기술과 관련된 알고리즘이 지속적으로 발전해 온 것을 확인할 수 있다.

'프로그래밍 언어'의 경우 자바, C++, C# 은 2015년까지만 상위 토픽에 나타나지만, 파이썬은 전체 기간에서 상위 토픽으로 확인되면서 인공지능 기술 개발에 사용되는 주요 프로그래밍 언어임을 확인할 수 있다. '응용분야' 유형에 해당하는 토픽의 경우는 자연어 처리와 시각화가 상위 토픽에 나타나며, 연도별로 특별한 동향 변화는 없는 것으로 확인할 수 있다. 마지막으로 소프트웨어와 알고리즘 개발을 지원하는 '개발 도구' 유형에 해당하는 토픽들은 2015년 이후 기간에 처음으로 확인이 가능하다. 이를 통하여 2015년 이후에 인공지능 관련 오픈소스 소프트웨어 개발 프로젝트 수의 급격한 증가는 소프트웨어 개발을 보다 원활히 수행할 수 있는 '개발 도구'의 발전이 주요 요인들 중 하나임을 고려할 수 있다.

5. 결론

본 연구에서는 오픈 소스 소프트웨어 개발 프로젝트들을 대상으로 인공지능과 관련된 주요 기술들의 개발 동향을 분석하였다. 논문이나 특허 정보를 활용한 기존의 기술 동향 분석 연구들과는 다르게, 실제 개발이 이루어지고 있는 소프트웨어들을 분석 대상으로 하였으며, 특정 분야를 제한하지 않은 기술 동향의 확인을 시도하였다.

연구 논문들의 경우 일련의 연구 수행 과정부터 논문 발표를 통하여 지식이 확산되는데 까지 시간적 차이가 발생하며, 특허의 경우도 심사를 통하여 외부에 공개되는 시점까지 일반적으로 많은 시간이 소요된다. 기술에 대한 발전 속도가 빠르고, 활용 분야도 다양한 인공지능 관련 기술의 특성을 고려하면, 연구 논문의 서지 정보나 특허를 대상으로 한 동향 분석에는 시간적인 측면에서 한계점이 존재하며, 이를 보완하기 위한 방안으로 오픈 소스 소프트웨어를 분석 대상으로 선정하였다는 점에서 본 연구가 기존 연구들과 비교하여 갖는 주요 학문적 차별점이라고 할 수 있다.

또한, 실제 오픈 소스 소프트웨어 개발 프로젝트들은 인공지능과 관련된 기술뿐만 아니라, 다양한 분야들과 관련된 소프트웨어들이 공유와 협업을 통하여 개발이 이루어지고 있다. 이를 고려하면 기술 개발 동향 분석을 위해 논문의 서지 정보 및 특허 데이터에서 나아가, 오픈 소스 데이터의 활용 가능성을 제시하였다는 측면에서 학술적 의의를 갖는다고 할 수 있다.

실무적 측면에서 본 연구가 갖는 주요 시사점으로는, 우선 인공지능 관련 오픈 소스 소프트웨어의 동향을 크게 '알고리즘', '프로그래밍 언어', '응용분야', '개발 도구'의 4가지 유형으로 구분할 수 있는 것을 확인하였다. 이러한 거시적 측면에서의 유형 구분을 기준으로, 각각의 유형 내에서 어떠한 토픽들이 발생하는 지에 대하여 미시적인 측면에서 분석하고 지속적으로 변화 추이를 확인한다면 기술 개발 동향에 대하여 보다 효과적으로 파악하는 것이 가능할 것이다.

또한, 인공지능 기술과 관련된 오픈 소스 소프트웨어 프로젝트의 토픽별 출현 빈도와 함께, 프로젝트에서 함께 언급되는 토픽 간의 관계에 대한 네트워크 분석 결과를 활용하여 주요 응용 기술들의 도출을 시도하였다. 이러한 분석 결과를 바탕으로, 현재 인공지능 기술과 관련하여 활발히 개발이 이루어지고 있는 분야들을 확인하는 것이 가능하며, 향후 융합 가능한 기술 동향 등에 대하여 보다 실증적인 분석을 위한 기초자료로 활용하는 것이 가능할 것이다.

이러한 연구 노력에도 불구하고 본 연구가 갖는 주요 한계점은 다음과 같다. 우선, 본 연구에서는 오픈 소스 소프트웨어 프로젝트들을 생성 연도를 기준으로 구분하였으나, 일부 프로젝트의 경우는 기존에 개발을 진행하던 오픈 소스 소프트웨어를 최근에 인공지능 관련 기술로 변경한 경우도 존재한다. 이러한 경우 해당 프로젝트의 생성일과 실제 인공지능 관련 오픈 소스 소프트웨어 개발 시작 시점이 일치하지 않아 정확한 기술 동향 분석에 어려움이 존재하였다. 또한, 실제 소프트웨어 개발 프로젝트들을 대상으로 하기 위하여 분석 대상 프로젝트들의 검색 시 프로그래밍 언어를 검색 질의어로 함께 활용하였으나, 일부 프로젝트의 경우는 관련 연구 논문과 같이 문서화 된 자료들의 저장소로만 활용되는 경우도 존재하였다. 이에 따라, 향후 연구에서는 오픈 소스 소프트웨어 개발 시에 소스 코드를 기여하는 참여자 수를 기준으로 프로젝트의 활성화 정도를 측정하고, 프로젝트의 업데이트 내역, 소스 코드의 수정 및 업로드 기록 등을 확인하여 보다 정교한 데이터 선별 과정이 필요하다. 이와 함께, 인공지능 기술과 관련된 특허 및 연구 논문 동향에 대한 분석 결과를 반영하여, 소프트웨어 개발 동향과의 비교 분석 연구, 국내외 인공지능 기술 동향의 차이점을 확인하는 연구 등이 진행되어야 할 것이다.

참고문헌(References)

Bae, Y. I. and H. R. Shin, "A Study on Convergence Patterns of Artificial Intelligence Technology using Patent Network Analysis," *GRI Review,* Vol.19, No.1(2017), 113~133.

Bonaccorsi, A. and C. Rossi, "Why Open Source Software Can Succeed," *Research policy*," Vol.32, No.7(2003), 1243~1258.

Choi, J. H., H. S. Kim, and N. G. Im, "Keyword Network Analysis for Technology Forecasting," *Journal of Intelligence and Information Systems*, Vol.17, No.4(2011), 227~240.

Choi, J. H. and S. H. Jun, "Bayesian Inference for Technology Analysis of Artificial Intelligence," *Journal of Korean Institute of Intelligent Systems*, Vol.28, No.4(2018), 411~416.

Chung, M. S. and J. Y. Lee, "Systemic Analysis of Research Activities and Trends Related to Artificial Intelligence(A.I.) Technology Based on Latent Dirichlet Allocation (LDA) Model," *Journal of the Korea Industrial Information Systems Research*, Vol.23, No.3 (2018), 87~95.

Chung, M. S., S. H. Park, B. H. Chae, and J. Y. Lee, "Analysis of Major Research Trends in Artificial Intelligence through Analysis of Thesis Data," *Journal of Digital Convergence*, Von.15, No.5(2017), 225~233.

Chung, M. S., S. H. Jeong, and J. Y. Lee, "Analysis of Major Research Trends in Artificial Intelligence based on Domestic/ International Patent Data," *Journal of Digital Convergence*, Vol.16, No.6(2018), 187~195.

Fujii, H. and S. Managi, "Trends and Priority Shifts in Artificial Intelligence Technology Invention: A Global Patent Analysis," *Economic Analysis and Policy*, Vol.58(2018), 60~69.

Han, M. U., S. H. LEE, W. H. Lee, and M. H. Lee, "A study on the IT R&D Emerging Technology Detection through Information Analysis Method -Focus on Next Generation Computing Field-," *proceeding of The Korean Operations Research and Management Science Society*, (2009), 1066~1073.

Jun, S. H., "A Big Data Learning for Patent Analysis," *Journal of Korean Institute of Intelligent Systems*, Vol.23, No.5(2013), 406-411.

Kim, D. H., "4th Industrial Revolution, Development of Technology for Open SW Innovation[written in Korean] ," *NIPA Issue Report*, No.33(2018).

Kim, D. S. and J. W. Kim, "Research Trend Analysis Using Bibliographic Information and Citations of Cloud Computing Articles: Application of Social Network Analysis," *Journal of Intelligence and Information Systems,* Vol.20, No.1(2014), 195~211.

Kho, J. C., K. T. Cho, and Y. H. Cho, "A Study on Recent Research Trend in Management of Technology Using Keywords Network Analysis," *Journal of Intelligence and Information Systems,* Vol.19, No.2(2013), 101~123.

Nam, C. H., "Open Source AI - Artificial Intelligence Ecosystem and Open Innovation," *KISDI Premium Report,* (2016), 4~22.

Niu, J., W. Tang, F. Xu, X. Zhou, and Y. Song, "Global Research on Artificial Intelligence

from 1990–2014: Spatially-Explicit Bibliometric Analysis," *ISPRS International Journal of Geo-Information*, Vol.5, No.5(2016), 66~84.

Park, J. S., S. G. Hong, and J. W. Kim, "A Study on Science Technology Trend and Prediction Using Topic Modeling," *Journal of the Korea Industrial Information Systems Research*, Vo.22, No.4(2017), 19-28.

Park, J. Y., "Trend Analysis of Artificial Intelligence Technology Using Patent Information," *Journal of the Korea Society of Computer and Information*, Vol.23, No.4(2018), 9~16.

Rho, S., "Artificial Intelligence Technology R&D Trend by Patent Analysis," *The Journal of Digital Contents Society*, Vol.18, No.2(2017), 423~428.

Synopsys, "*2018 Open Source Security and Risk Analysis*", 2019. Available at https://www.synopsys.com/software-integrity/resources/analyst-reports/open-source-security-risk-analysis-2018.html(Downloaded 8 March 2019).

Tseng, Y. H., C. J. Lin, and Y. I. Lin, "Text Mining Techniques for Patent Analysis," *Information Processing & Management*, Vol.43, No.5(2007), 1216~1247.

Abstract

A Study on the Development Trend of Artificial Intelligence Using Text Mining Technique: Focused on Open Source Software Projects on Github*

JiSeon Chong** · Dongsung Kim** · Hong Joo Lee*** · Jong Woo Kim****

Artificial intelligence (AI) is one of the main driving forces leading the Fourth Industrial Revolution. The technologies associated with AI have already shown superior abilities that are equal to or better than people in many fields including image and speech recognition. Particularly, many efforts have been actively given to identify the current technology trends and analyze development directions of it, because AI technologies can be utilized in a wide range of fields including medical, financial, manufacturing, service, and education fields. Major platforms that can develop complex AI algorithms for learning, reasoning, and recognition have been open to the public as open source projects. As a result, technologies and services that utilize them have increased rapidly. It has been confirmed as one of the major reasons for the fast development of AI technologies. Additionally, the spread of the technology is greatly in debt to open source software, developed by major global companies, supporting natural language recognition, speech recognition, and image recognition. Therefore, this study aimed to identify the practical trend of AI technology development by analyzing OSS projects associated with AI, which have been developed by the online collaboration of many parties. This study searched and collected a list of major projects related to AI, which were generated from 2000 to July 2018 on Github. This study confirmed the development trends of major technologies in detail by applying text mining technique targeting topic information, which indicates the characteristics of the collected projects and technical fields.

The results of the analysis showed that the number of software development projects by year was less than 100 projects per year until 2013. However, it increased to 229 projects in 2014 and 597 projects in 2015. Particularly, the number of open source projects related to AI increased rapidly in 2016 (2,559

* This work was supported by the Ministry of Education of Korea and the National Research Foundation of Korea in 2017(NRF-2017S1A3A2066740).

** Department of Business Administration, Graduate School, Hanyang University
*** Department of Business Administration, The Catholic University of Korea
**** Corresponding Author: Jong Woo Kim
School of Business, Hanyang University
222 Wangsimni-ro, Seongdong-gu, Seoul 133-791, Korea
Tel: +82-2-2220-1067, Fax: +82-2-2220-1169, E-mail: kjw@hanyang.ac.kr

OSS projects). It was confirmed that the number of projects initiated in 2017 was 14,213, which is almost four-folds of the number of total projects generated from 2009 to 2016 (3,555 projects). The number of projects initiated from Jan to Jul 2018 was 8,737.

The development trend of AI-related technologies was evaluated by dividing the study period into three phases. The appearance frequency of topics indicate the technology trends of AI-related OSS projects. The results showed that the natural language processing technology has continued to be at the top in all years. It implied that OSS had been developed continuously. Until 2015, Python, C ++, and Java, programming languages, were listed as the top ten frequently appeared topics. However, after 2016, programming languages other than Python disappeared from the top ten topics. Instead of them, platforms supporting the development of AI algorithms, such as TensorFlow and Keras, are showing high appearance frequency. Additionally, reinforcement learning algorithms and convolutional neural networks, which have been used in various fields, were frequently appeared topics.

The results of topic network analysis showed that the most important topics of degree centrality were similar to those of appearance frequency. The main difference was that visualization and medical imaging topics were found at the top of the list, although they were not in the top of the list from 2009 to 2012. The results indicated that OSS was developed in the medical field in order to utilize the AI technology. Moreover, although the computer vision was in the top 10 of the appearance frequency list from 2013 to 2015, they were not in the top 10 of the degree centrality. The topics at the top of the degree centrality list were similar to those at the top of the appearance frequency list. It was found that the ranks of the composite neural network and reinforcement learning were changed slightly.

The trend of technology development was examined using the appearance frequency of topics and degree centrality. The results showed that machine learning revealed the highest frequency and the highest degree centrality in all years. Moreover, it is noteworthy that, although the deep learning topic showed a low frequency and a low degree centrality between 2009 and 2012, their ranks abruptly increased between 2013 and 2015. It was confirmed that in recent years both technologies had high appearance frequency and degree centrality. TensorFlow first appeared during the phase of 2013-2015, and the appearance frequency and degree centrality of it soared between 2016 and 2018 to be at the top of the lists after deep learning, python. Computer vision and reinforcement learning did not show an abrupt increase or decrease, and they had relatively low appearance frequency and degree centrality compared with the above-mentioned topics.

Based on these analysis results, it is possible to identify the fields in which AI technologies are actively developed. The results of this study can be used as a baseline dataset for more empirical analysis on future technology trends that can be converged.

Key Words : Artificial Intelligence, Technology Trends, Open Source Software, Github, Text Mining

Received : January 18, 2019 Revised : March 8, 2019 Accepted : March 11, 2019
Publication Type : Conference(Fast-track) Corresponding Author : Jong Woo Kim

저 자 소 개

정 지 선

현재 한양대학교 일반대학원 경영학과에서 경영정보시스템 전공으로 박사과정에 재학 중이다. 홍익대학교 컴퓨터 정보통신 공학과에서 학사를 마쳤으며, 한양대학교에서 경영정보시스템 전공으로 석사학위를 취득하였다. 주요 연구 관심분야는 데이터 마이닝, 상품 추천, 사회 네트워크 분석, 클라우드 컴퓨팅 서비스 등이다.

김 동 성

현재 한양대학교 일반대학원 경영학과에서 경영정보시스템 전공으로 박사과정에 재학 중이며, 동 대학원에서 경영정보시스템 전공으로 석사학위를 취득하였다. 주요 연구 관심분야는 데이터 마이닝 기법과 응용, 기계 학습, 오피니언 마이닝, 사회 네트워크 분석, 딥러닝 기법의 활용 등이다.

이 홍 주

현재 가톨릭대학교 경영학전공 교수로 재직 중이다. KAIST 산업경영학과를 졸업하고 KAIST 테크노경영대학원에서 석사 및 박사학위를 취득하였다. 주요 관심분야는 데이터 분석, 지능형 정보시스템, 온라인 사용자들의 상호작용 등이다.

김 종 우

현재 한양대학교 경영대학 경영학부 교수로 재직 중이다. 서울대학교 수학과에서 학사를 마쳤으며, 한국과학기술원에서 경영과학으로 석사학위를, 산업경영학으로 박사학위를 취득하였다. 주요 연구 관심분야는 데이터마이닝 기법과 응용, 기계학습과 딥러닝, 오피니언 마이닝, 상품추천기술, 지능형 정보시스템, 집단지성, 사회 네트워크 분석, 클라우드 컴퓨팅 서비스 등이다.