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/