정보통신정책연구 제 28 권 제 1 호 (2021. 3) pp.27-50
http://dx.doi.org/10.37793/ITPR.28.1.2

딥러닝과 LDA 모델링을 통한 AI 분야 장기특허 예측

The Prediction of Long-Term Survival of Artificial Intelligence Patents Based on Deep-Learning and Latent Dirichlet Allocation Modeling

하 태 현 (Taehyun Ha)*
이 재 민 (Jae-Min Lee)**
이 창 환 (Chang-Hoan Lee)***
고 병 열 (Byoung-Youl Coh)****

국문초록

본 논문에서는 최근 10년간 등록된 인공지능과 기계학습 분야 특허들을 대상으로 향후 20년간 특허 권리가 유지될 장기특허를 판별, 예측하고 이들의 내용을 LDA 모델링으로 분석하여 인공지능 분야의 기술정책 방향을 제시한다. 딥러닝 모델을 통해 약 16만 건의 미국 특허청 등록 특허의 장기특허 여부를 학습하였으며, 학습된 모델을 3,281개의 인공지능과 기계학습 분야 특허들에 적용하여 장기특허로 예측되는 2,004개의 특허를 판별하였다. 도출된 2004개의 장기특허에 대한 LDA 모델링을 수행하였으며, 장기전략적으로 중요해질 6개의 주요 토픽들을 확인하였다. 또한, 기술통계치와 통계 분석을 통해 인공지능 분야 내 장기특허와 단기특허 간 차이점에 대해 알아보았으며, LDA 토픽 모델링을 통해 도출된 결과와 함께, 향후 인공지능 분야에서 고려되어야 할 정책적 함의들에 대해 종합적으로 논의하였다.

주제어: 인공지능, 기계학습, 장기특허, 딥러닝 모델, 토픽 모델링

※ 논문접수일: 2020. 11. 6, 수정일: 2021. 1. 26, 게재확정일: 2021. 2. 1
* 한국과학기술정보연구원 미래기술분석센터 선임연구원, 제1저자, E-mail: taehyunha@kisti.re.kr
** 한국과학기술정보연구원 미래기술분석센터 책임연구원, 교신저자, E-mail: jmlee@kisti.re.kr
*** 한국과학기술정보연구원 미래기술분석센터 책임연구원, E-mail: cheree1@kisti.re.kr
**** 한국과학기술정보연구원 미래기술분석센터 책임연구원, E-mail: cohby@kisti.re.kr

ABSTRACT

This study predicts the long-term continuance of patents and analyzes their content based on deep-learning and latent Dirichlet allocation modeling. To predict the long-term continuance of patents, we develop a deep-learning model based on 160 thousand patents submitted to the United States Patent and Trademark Office. The model is applied to 3,281 patents for artificial intelligence of which 2,004 are predicted to remain registered long-term. The long-term patents are analyzed using the latent Dirichlet allocation modeling, and are contrasted with short-term patents. The analysis leads to the discovery of six major topics associated with log-term patents. Several policy implications are drawn.

Key words: Artificial intelligence, Machine learning, Long-term patents, Deep-learning model, Topic modeling

Ⅰ. 서론

4차 산업혁명은 디지털 정보를 효율적으로 관리할 수 있는 기법과 이를 뒷받침해 줄 수 있는 장비들의 개발을 통해 산업 전 분야에 큰 변화를 가져오고 있다. 그중에서도 특히 인공지능 관련 기술은 여러 영역에서 해결되지 않았던 난제들을 해결하면서 급속도로 발전해왔다. 인공지능 기술은 운전자의 도움 없이 스스로 주행하는 자율주행자동차, 제조/검수와 같이 공정 전반에 걸친 과정을 시스템 스스로 수행할 수 있는 스마트팩토리 등과 같은 산업 분야 기술부터, 체스와 바둑, 스타크래프트와 같은 게임 영역까지, 광범위한 분야에 걸쳐 새로운 변화를 이끌어왔으며, 4차 산업혁명 분야 내에서는 빅데이터, 로봇, 3D 프린팅, 양자 컴퓨팅, 재료과학, 나노기술, 바이오기술 등과 함께 더 큰 변화들을 끌어낼 것으로 기대되고 있다 (Schwab, 2016).

이러한 변화에 대응하기 위해, 미국은 2019년 국가 인공지능 연구 개발 전략 계획(The National Artificial Intelligence Research and Development Strategic Plan)을 통해 국가 주도의 인공지능 연구 개발 관련 기술 측정/평가 및 전문 인력 확충의 중요성을 언급하였으며, 유럽 연합은 데이터 보호규칙(General Data Protection Regulation), 신뢰할 수 있는 인공지능 윤리 가이드라인(Ethics Guidelines for Trustworthy AI)등을 발표하여 사용자의 권리를 보장하고, 안전한 인공지능 사용을 위한 가이드라인 등을 제공하였다. 이러한 정책들은 인공지능 기술이 양적으로 성장하는 동시에, 질적으로도 성장할 수 있도록 하는 방안을 제공하는 데 초점이 맞춰져 있다.

실제로 많은 사례들은 인공지능 관련 제품과 서비스의 질적인 성장이 양적인 성장과 비례하지 않을 수 있음을 보여주고 있다. IBM에서 자연어 처리를 위해 제작한 Watson은 의학 진단 분야에서 의사의 역량과 비슷하거나 이를 뛰어넘는 성능을 낼 것으로 기대를 모았으나, 국내 도입 이후에는 이렇다 할 성과를 내지 못했고, 상당히 저조한 수준의 암 판단 정확도 수치를 나타내는 것으로 드러났다 (박선재, 2019). 또한, 미국의 컨설팅 기관 IDC(International Data Corporation)의 조사 결과에서는 많은 수의 회사가 인공지능 기술에 대해 과해석하고 있으며, 숙련된 기술과 이해 부족으로 인해 절반 넘는 프로젝트가 실패로 돌아가는 것으로 나타났다 (IDC, 2019). 이와 같은 사례들은, 성공적인 인공지능 도입을 위해서는 인

공지능 기술에 대한 올바른 이해와 더불어 꾸준히 성장할 수 있는 장기전략적 기술을 파악하는 것이 중요하다는 것을 보여주고 있다. 그러나 이와 같은 필요성에도 불구하고, 아직까지 정량적인 방법을 통해 인공지능 관련 장기전략기술을 분석하고 이를 정책적인 관점에서 해석하고자 한 노력은 잘 이뤄지지 않았다.

특허 분석을 통해 미래에 유망해질 기술들에 대해 파악하려고 하는 시도는 과학계량학(Scientometrics) 분야를 중심으로 이뤄져 왔으며, 주로 특허의 인용 네트워크 분석, 특허 속성을 이용한 클러스터링 분석, 특허 제목과 초록에 대한 자연어 처리 기법 적용을 통한 분석 등이 이뤄져 왔다 (Aristodemou & Tietze, 2018). 본 연구에서는 자연어 처리 기법을 활용하여 특허의 내용을 분석하는 한편, 단순히 인공지능과 관련된 특허들을 분석하는 것에서 나아가, 장기간 특허권을 유지할 것으로 예측되는 장기특허들에 대해서만 분석을 시행하여, 인공지능 관련 유망 기술들의 예측 정확도를 더 높이고자 하였다. 또한, 본 연구진들은 도출된 분석 결과를 토대로, 인공지능 연구 분야에서 전략적으로 유망하게 떠오르고 있는 기술들에 대해 논의하고, 이를 바탕으로 향후 효과적인 기술정책 수립에 활용될 수 있는 함의점을 제시하고자 하였다.

Ⅱ. 배경연구

1. 특허 분석

특허 분석은 특정 기술의 현재 상태를 이해하고, 향후 유망해질 것으로 나타나는 기술을 파악하기 위한 용도로 많이 활용됐다. 기술적인 내용을 축약적으로 담고 있는 특허는 속지주의적 원칙에 따라서 등록된 국가 내에서만 효력을 발휘할 수 있는데, 일반적으로 가장 많은 특허가 등록된 미국 특허청(USPTO, United States Patent and Trademark Office) 데이터와 유럽 특허청(EPO, European Patent Office) 데이터를 많이 활용하고 있다. 등록된 특허들은 특허를 통해 보호받고자 하는 기술적 내용을 잘 표현하는 제목과 초록을 가지고 있으며, 기술적 특징을 잘 대표하는 IPC(International Patent Classification) 분류 코드와 CPC(Cooperative Patent Classification) 분류 코드를 가지고 있다. 각 분류 코드 체계는 계층적으로 구성되어 있으며, 섹션, 클래스, 서브클래스, 메인그룹, 서브그룹으로 구성된 분류체계를

사용하고 있다. IPC 분류 코드의 경우, 1968년에 제정되었으며 약 7만개의 코드로 구성되어 있고, CPC 분류 코드의 경우, 2013년에 제정되어 약 26만개의 코드로 구성되어 있다.

이처럼 특허 서지 정보들은 체계화된 형식을 갖추고 있으므로, 특허의 분류 코드, 제목, 초록 등을 분석하면 특정 분야에서 이뤄지고 있는 기술적 변화를 파악하기에 용이하다. 지금까지 인공신경망, 클러스터링, 의사결정나무, 확률 모델링, 텍스트 마이닝 기법 등 다양한 방법들이 활용됐으며, 지식 관리, 기술 관리, 지식 재산권의 경제적 가치 평가 등의 목적으로 활용되어 왔다 (Aristodemou & Tietze, 2018). Kim & Bae (2017)는 건강 관리 분야에서 유망한 기술을 파악하기 위해 특허 데이터를 수집하여 클러스터링 분석을 수행한 뒤, 특허의 피인용지수(forward citations), 삼극 특허 패밀리(triadic patent families), 독립 청구항(independent claims) 등의 지표를 활용해 군집화된 특허들의 유망 여부를 판별하였다. Cho, Lim, Lee, Cho, & Kang (2018)는 건축 기술 분야에서 유망하게 떠오를 기술을 파악하기 위해, 건축 분야 특허 데이터를 수집한 뒤, 트랜드 분석과 포트폴리오 분석을 통해 기술시장 내에서 해당 특허들의 중요도를 분석하고, 특정 기술과 관련된 특허들의 수 변화 추세 관찰을 통해 유망 기술을 예측하였다. 이 밖에도, 증강현실 (Evangelista et al., 2020), 무선전력송신 (Kim, Han, Lee, Cho, & Lee, 2019), 물류 (Choi & Song, 2018), 태양광 (Shubbak, 2019) 등, 여러 영역에서 특허 분석을 통한 시장 조사와 기술 흐름 예측 연구들이 이뤄져 왔다.

2. 장기특허

인공지능과 관련된 특허들을 분석하고 이들의 특징을 살펴보는 것은 인공지능 정책 수립을 위해 필요한 기초적인 분석 결과를 제공해줄 수 있다. 그러나, 더욱 실질적이고 효과적인 정책 수립을 위해서는 인공지능 분야에서 등록된 특허 중 유망할 것으로 예측되는 특허와 그렇지 않은 특허를 구분하고 각각의 특징들을 분석하는 과정이 필요하다. 과거 여러 연구가 유망 특허를 정의하는 과정에서, 특허가 얼마나 많이 인용되고(피인용 수), 얼마나 여러 나라에서 권리를 행사하는가(패밀리 특허의 수)의 정도를 활용해왔지만 (예: Kim & Bae, 2017), 이러한 방식은 여러 지표를 종합적으로 고려해야 하고, 유망과 비 유망을 나누기 위한 임계값 설정이 주관적일 수 있다는 한계점이 있었다. 이러한 측면에서 볼 때, 특허의 유지

기간은 등록인이 해당 특허를 장기적이고 전략적으로 활용하고자 하는 의지를 보여주는 직접적인 척도가 될 수 있으며, 유망 특허를 나타내는 지표로 활용될 수 있다 (이재민·고병열·윤장혁, 2019). 특허를 장기간 유지하는 것은 큰 비용이 소요되는 만큼, 특허 유지 기간은 등록인이 해당 특허를 전략적으로 운용하고자 하는 의지와 특허의 부가가치를 나타내는 것으로 생각할 수 있다 (Guellec & de la Potterie, 2000; 이재민 외, 2019; 최재웅 외, 2018).

일반적으로 특허를 소유한 사람은 특허의 권리를 출원일로부터 최대 20년 동안 행사할 수 있다. 특허청은 등록된 특허에 대해 처음 4년의 권리를 보장하며, 이후 8년, 12년이 될 때마다 권리 연장 여부를 신청받게 된다. 이와 같은 연장을 모두 신청하게 되면 최장기간인 20년 동안 특허 권리를 보장받게 된다. 미국 특허청에 등록된 특허의 경우 등록 특허 한 건을 최대 수명인 20년 동안 연장하기 위해서는 약 1557만 원이 필요한 것으로 알려져 있으며, 기타 옵션에 따라서 상당한 비용이 추가될 수 있다. 이처럼 높은 비용을 들여 특허의 권리를 주장하고자 하는 것은, 등록인이 특허가 가지고 있는 높은 기술적 가치를 강력하게 주장하고 있다는 것이며, 장기적이고 전략적으로 해당 특허를 운용하고자 하는 의지를 대변한다고 볼 수 있다 (이재민 외, 2019).

3. 토픽 모델링

특허 서지가 가지고 있는 분류 코드를 활용하면, 특정 분야에서 등록되고 있는 기술 특허들의 범위를 한정 지을 수 있지만, 특허들이 내포하고 있는 의미적 특징을 주제별로 파악하기에는 어려움이 있다. 이와 같은 어려움을 해결하기 위해 과거 연구자들은 토픽 모델링 기법을 통해 특정 분야 내에서 이뤄지고 있는 등록 특허들의 내용을 효과적으로 파악하기 위한 연구들을 진행해 왔다. 토픽 모델링은 초기, 문서-단어 행렬(Document-Term Matrix)과 같은 단순 통계 분석부터 시작해서, 차원 축소를 통해 문서가 가진 함축적 의미를 도출해내고자 한 LSA(Latent Semantic Analysis), 문서 내 단어들의 등장 횟수를 확률적으로 처리하여 분석한 pLSA (probabilistic LSA)를 거쳐, 등장 확률이 Dirichlet 분포를 따른다고 가정하고, 확률 추정을 베이지안 방식으로 해결하고자 한 LDA(Latent Dirichlet Allocation) 기법으로 발전해왔다. LDA는 Blei, Ng, & Jordan (2003)에 의해 제시된 방법으로, 문서 내 단어와 문서 집합 내 토픽의 등장 확률을 Dirichlet 분포로 모델링한 다음, 베이

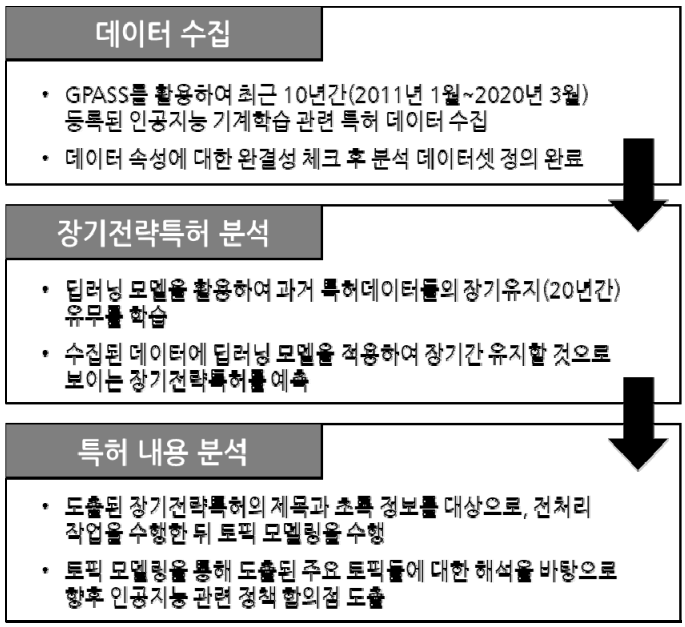
지안 방식을 통해 각 확률 분포의 모수를 추정해 나가는 과정을 거친다. 초기 LDA 방법은 입력된 데이터를 토대로 한 번의 추정을 통해 모수를 추정하는 방식이었으나, Hoffman, Bach, & Blei (2010)에 의해 데이터가 갱신될 때마다 실시간으로 모수 추정을 하는 online LDA 방식이 제시되었다.

LDA는 여러 연구자에 의해 특허를 분석하는 용도로 활용됐다. Choi & Song (2018)는 물류 분야에서 이뤄지고 있는 기술들의 특징을 파악하고 이들의 추세를 분석하기 위해, LDA를 통해 토픽 모델링을 수행한 뒤, 주요 토픽들에 속한 특허들을 연평균 성장률과 점유율 측면에서 구분하여 유망 특허들을 도출해내는 연구를 수행하였다. Kang, Lee, Jang, & Park (2019)는 전기 자동차 기술과 관련된 특허를 대상으로 LDA 모델링을 수행하여 주요 토픽들을 선정한 뒤, 각 토픽에 할당된 특허들의 정보를 토대로 기업과 학교간 협력 관계를 분석한 연구 결과를 내놓았다. Wang, Yang, Wang, Xia, & Wang (2020)는 ICVs(Intelligent Connected Vehicles) 관련 특허들을 대상으로 LDA를 수행한 뒤, 주요 토픽들을 선정하고 각 토픽에 속한 특허들을 대상으로 기술 다양성 지표와 기술특화 지표를 계산한 뒤, 이를 바탕으로 기업의 기술 경쟁력을 평가하는 연구를 수행하였다.

Ⅲ. 분석 방법

본 연구에서는 인공지능에 관련된 특허들을 대상으로, 특허 수명이 장기간 유지가 될 것으로 보이는 장기특허들을 파악하고, 이들의 내용을 파악하여 장기특허에 대한 종합적인 이해를 도출하고자 하였다. <그림 1>은 본 연구에서 수행한 분석 방법에 대한 대략적인 개요를 보여준다.

<그림 1> 분석 방법 개요



1. 데이터 수집

인공지능과 관련된 기술 특허들을 분석하기 위해, 미국 특허청(USPTO, United States Patent and Trademark Office)에 등록된 인공지능기술 관련 특허 중, 2011년 1월부터 2020년 3월까지 등록된 특허들을 수집하였으며, 기계학습과 관련된 IPC 코드인 G06N003, G06N005, G06N007에 해당하는 특허만을 대상으로 분석을 시행하였다. 특허 정보는 한국과학기술정보연구원(KISTI, Korean Institute of Science and Technology Information)이 운영하는 특허 데이터베이스 GPASS(Global Patent Analysis Service System)를 이용 하였으며, 총 4862개의 등록 특허들을 수집하였으나, 이 중 뒤에 나올 장기특허 분석 모델에 활용될 27개의 속성값 중 하나라도 없는 1581개의 특허들(약 32.52%)을 제외한 뒤, 최종적으로 3281개(특허 등록인 기준 3409개)의 특허를 분석에 활용하였다. <표 1>, <표 2>는 수집된 특허들의 속성에 대한 대략적인 정보를 보여준다.

<표 1> 등록 특허 수 기준 상위 등록인 10인

등록인	특허 개수	백분율
IBM	468	13.73
Google	145	4.25
Microsoft Technology Licensing	108	3.17
Amazon Technologies	86	2.52
Qualcomm	61	1.79
Samsung Electronics	40	1.17
Xerox	37	1.09
HRL Laboratories	37	1.09
EMC	35	1.03
Fujitsu	34	1.00

* 두 명 이상의 등록인을 갖는 특허는 각 등록인의 등록 특허로 계산하였음. 백분율은 전체 특허 수 대비 해당 등록인으로 등록된 특허 수의 비중을 나타냄

<표 2> 특허 인용 수 기준 상위 특허 10개

등록인	특허 제목	인용 횟수
Blanding Hovenweep	Adaptive pattern recognition based controller apparatus and method and human-interface therefore	7783
Microsoft Technology Licensing	Electronic form user interfaces	1271
LivePerson	Method and system for providing targeted content to a surfer	1222
Consumerinfo.com	Systems and methods for data verification	840
Newvaluexchange	Apparatuses, methods and systems for a digital conversation management platform	769
Apple	Method and apparatus for building an intelligent automated assistant	741
The Nielsen Company	System and method for gathering and analyzing biometric user feedback for use in social media and advertising applications	741
State farm Mutual Automobile Insurance Company	Autonomous operation suitability assessment and mapping	736
Pegasystems	Methods and apparatus for user interface optimization	697
Palantir Technologies	Malicious software detection in a computing system	678

2. 장기특허 분석

과거 연구들은 특허의 분류 코드, 인용수, 수명 등을 예측하기 위해 회귀분석, 지지벡터머신(Support Vector Machine), 인공신경망 등의 기법을 활용해 왔으나, 각자 활용한 특허 데이터의 성격이 서로 달랐기 때문에 기법들의 직접적인 성능 비교는 어렵다. 그러나, 일반적으로 높은 차원의 입력 변수를 활용하여 특정 변수를 예측해야 하는 문제의 경우, 많은 파라미터를 활용하여 복잡성을 효과적으로 학습할 수 있는 깊은 신경망(Deep Neural Network)이 기존의 방법들과 비교하여 나은 성능을 나타낸다고 알려져 있고, 최근에는 딥러닝 학습을 통해 특허의 장기보유 여부를 예측하는 모델이 연구되어 그 효과성을 입증한 바 있다(Choi, Jeong, Yoon, Coh, & Lee, 2020), 본 연구에서는 이에 근거하여, 깊은 신경망을 활용하여 장기특허를 예측하기 위한 모델을 수립하고자 하였다.

앞서 배경연구에서 언급되었듯, 특허의 권리는 등록 직후 4년까지 보장되지만, 이후 갱신을 거쳐 8년, 12년, 20년까지 연장될 수 있다. 본 연구에서 정의하는 장기특허는 특허가 보장받을 수 있는 최대 기간인 20년을 신청한 특허들로, 출원일로부터 12년이 지난 후 권리 연장을 신청한 특허들에 해당한다. 즉, 2020년을 기준으로 12년 전인 2008년에 출원된 특허들까지만 장기특허 여부를 판별할 수 있다. 본 연구에서는 2008년으로부터 최근 3년간인 2005년부터 2008년까지 미국 특허청에 등록된 특허 중에서, 연차료 납부 정보가 있고 특허권 유지 기간이 확정된 특허를 학습 대상으로 하였다. <표 3>은 확보한 특허들의 특허 연장 정보를 보여준다.

<표 3> 학습 특허의 권리 연장 기간별 비중

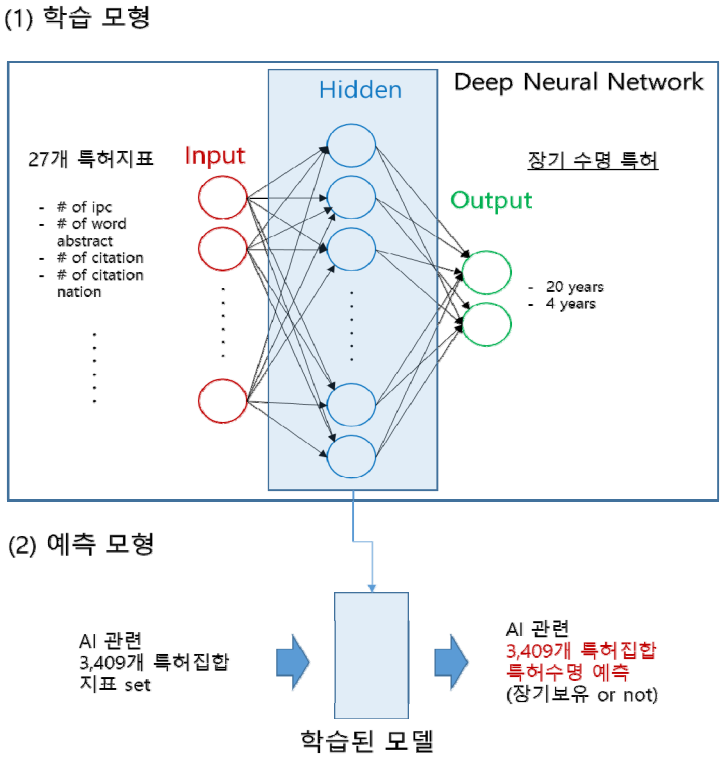
권리 연장 기간	특허수	비중
20년(최장 기간 연장)	211,398	42.11%
12년(2차 기간 연장)	88,535	17.64%
8년(1차 기간 연장)	121,464	24.19%
4년(미연장)	80,641	16.06%

일반적으로 분류모델의 성능은 분류해야 할 종류(class)가 많을수록 성능이 급격히 저하되기 때문에, 많은 경우 다중 분류 문제를 이진 분류 문제로 간소화하여 해결하려 노력한다. 이에 따라 본 연구에서도 4년, 8년, 12년, 20년 연장 특허를

각각 분류해내는 모델을 학습하는 대신, 미연장 특허(4년)와 최장 연장 특허(20년)를 분류하는 모델을 학습하였다. 확보한 특허 중, 20년 연장 특허 211,398건과 미연장 특허 80,641건을 학습에 활용하였으며, 총 290,640건을 가지고 다층 구조를 갖는 깊은 신경망 모델을 학습하였다.

모델의 입력값으로는 특허 제목, 초록 글자 수 등으로 구성된 특허지표 27개를 사용하였으며, 출력값으로는 특허 수명이 20년이 되는지 아닌지를 나타내는 이진 분류 값을 사용하였다. 모델의 출력단에 있는 소프트맥스 함수의 임계값은 0.5로 설정하였으며, 임곗값 이상을 나타내는 특허를 장기특허로, 이하를 나타내는 특허를 단기특허로 판별하였다. 데이터의 3/4은 학습 데이터로 1/4은 검증 데이터로 활용하였다. <그림 2>는 본 연구에서 활용한 유망 특허 예측 모델의 개념도를 보여주며, <표 4>는 모델의 학습에 활용한 데이터들의 입력 변수들과 기술 통계치를 보여준다.

<그림 2> 장기특허 분석 모델 개념도



<표 4> 학습 데이터 입력 변수와 통계치

모델 변수명	설명	평균값	표준편차	최솟값	최댓값
number_ipc	특허의 ipc의 수	1.602	1.106	1	29
number_word_abstract	특허의 초록 단어 수	112.364	45.514	3	470
number_citation	특허의 인용특허 수	21.395	42.468	0	1966
number_citation_nation	특허의 인용특허 국가 수	2.138	1.399	0	19
number_priority	특허의 우선권 수	0.462	0.788	0	56
number_priority_nation	특허의 우선권 국가 수	0.371	0.486	0	4
number_claim	특허의 청구항 수	18.807	15.002	1	803
number_claim_indep	특허의 독립 청구항 수	3.106	2.565	0	81
number_claim_dep	특허의 종속 청구항 수	15.701	13.859	0	779
number_applicant	특허의 출원인 수	2.530	1.803	1	32
number_foreign_applicant	특허의 해외 출원인 수	1.201	1.728	0	25
number_applicant_nation	특허의 출원인 소속 국가 수	1.057	0.252	1	7
number_assignee	특허의 권리인 수	2.530	1.803	1	32
number_avgword_indep	독립 청구항 당 평균 단어 수	70.858	42.980	0	222
average_gap_citation	특허와 인용 문헌 사이의 평균 시간 차이	11.059	8.139	0	299
number_family	패밀리 특허 수	23.772	386.364	0	9889
number_foreign_family	해외 패밀리 특허 수	23.772	386.364	0	9889
number_family_nation	패밀리 특허의 국가 수	3.648	3.741	0	54
delivery_time	특허 심사 기간(단위: 개월)	3.134	1.432	0.150	26.33
ipc_A	특허의 A섹션 IPC 수	0.199	0.622	0	29
ipc_B	특허의 B섹션 IPC 수	0.220	0.592	0	15
ipc_C	특허의 C섹션 IPC 수	0.197	0.711	0	21
ipc_D	특허의 D섹션 IPC 수	0.010	0.138	0	8
ipc_E	특허의 E섹션 IPC 수	0.031	0.216	0	10
ipc_F	특허의 F섹션 IPC 수	0.101	0.395	0	13
ipc_G	특허의 G섹션 IPC 수	0.453	0.792	0	18
ipc_H	특허의 H섹션 IPC 수	0.391	0.773	0	13

* 하나의 특허는 다른 섹션에 속하는 여러 개의 IPC 분류 코드를 할당받을 수 있음.

모델의 학습과 예측 성능 평가는 알고리즘 성능 평가에 일반적으로 많이 활용되는 Precision, Recall, F_2 척도 측면에서 이뤄졌다(<표 5> 참조). Precision은 참이라고 판단한 것 중 실제 참인 것의 비율을 나타내며, Recall은 실제 참과 거짓 중

찾아낸 참의 비율을 나타낸다. 두 척도는 각각 모델의 정확도를 나타내지만 서로 반비례하는 성격을 갖는데, 모델이 참을 판단하는데 엄격하여 참이라고 판단하는 비율을 낮추게 되면 Precision이 높아지지만(분모의 감소), Recall은 낮아지는 경향이 있고(분자의 감소), 반대로 모델이 참을 판단하는데 관대하여 참이라고 판단하는 비율을 높이게 되면 Precision은 낮아지고(분모의 증가) Recall은 높아지는 경향이 있다(분자의 증가). 이에 따라 두 지표를 적절한 비율로 합쳐서 모델의 성능을 평가하고자 하는 F_β 척도가 고안되었는데, 두 지표를 어떠한 비중으로 합칠 것인지에 따라서 β를 결정하게 된다. 과거 연구에서는 일반적으로 F_2와 $F_{0.5}$ 척도를 활용해왔는데, 본 연구에서는 장기특허 발굴이라는 목적에 맞춰 Recall에 좀 더 비중을 둔 F_2 척도를 활용하여 모델의 성능을 측정하였다. 모델 학습 결과, 학습 데이터에 대한 Precision, Recall, F_2 값은 각각 0.66, 0.68, 0.68로 나타났고, 검증 데이터에 대한 Precision, Recall, F_2 값은 각각 0.65, 0.7, 0.69로 나타났다.

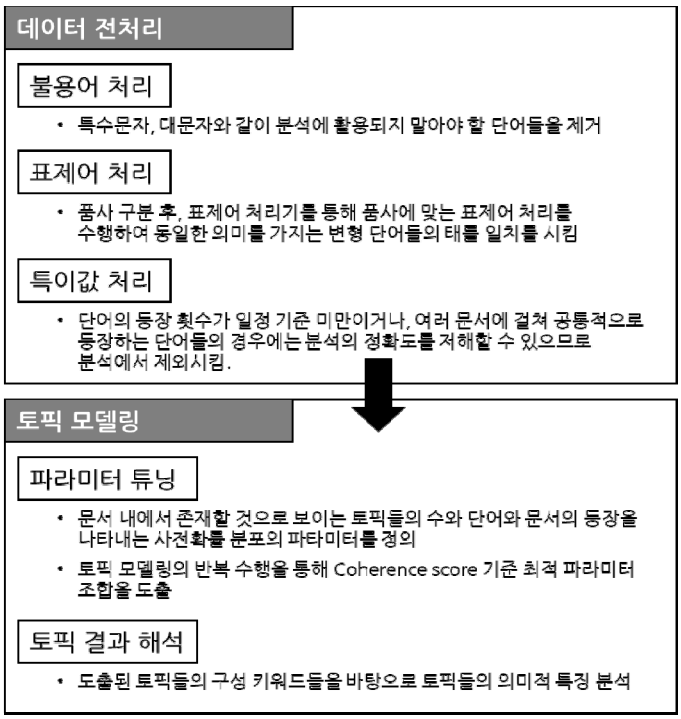
<표 5> 평가지표 설명과 수식

지표명	설명	수식
Precision	참이라고 판단한 것 중 실제 참인 것의 비율	$Precision = \frac{True\,positive}{True\,positive + False\,positive}$
Recall	실제 참과 거짓 중 찾아낸 참의 비율	$Recall = \frac{True\,positive}{True\,positive + False\,negative}$
F_β	Precision과 Recall의 조화평균	$F_\beta = \frac{(\beta^2+1) \times Precision \times Recall}{\beta^2 \times Precision + Recall}$

3. 특허 내용 분석

장기 유지할 것으로 예상되는 인공지능 특허들이 파악된 이후에는 특허 내용을 분석하기 위해 특허 제목과 초록을 대상으로 전처리를 수행하고 LDA 모델링을 수행하였다. 전처리의 경우 Python 자연어 처리 패키지 Gensim(https://radimrehurek.com/gensim)을 활용하였으며, 불용어(Stopwords) 제거와 품사(POS, Part Of Speech) 구분 후, 각 품사에 맞는 표제어 처리(Lemmatization)를 수행하여, 분석에 쓰일 단어들을 정제하였다. 또한, 3번 미만으로 등장한 단어들과 5% 이상의 문서에 공통으로 등장하는 단어들은 정확도를 위해 분석에서 제외하였다.

<그림 3> 특허 내용 분석 개념도



LDA 모델 학습을 위해 사전에 설정해줘야 하는 파라미터로는 1) Dirichlet 분포 파라미터(alpha, beta)와 2) 추정하고자 하는 토픽의 수가 있다. LDA 모델은 여러 개의 문서가 각 토픽에 차지하는 비중과 여러 개의 키워드가 각 토픽에 차지하는 비중이 각각 Dirichlet 분포를 따른다고 가정하는데, Alpha와 Beta 파라미터는 이들 Dirichlet 분포의 형태를 규정하는 역할을 한다. Alpha가 커질수록 문서들에 비슷한 비중을 부여하여 토픽들의 분포는 비슷해지고, Beta가 커질수록 단어들에 비슷한 비중을 부여하여 토픽들의 분포들은 비슷해진다. Alpha와 Beta 파라미터를 추정하는 방법에는 유전자 알고리즘 등의 휴리스틱 최적화 방법을 활용하는 방법(Agrawal, Fu, & Menzies, 2018; Pathik & Shukla, 2020) 등이 있으나, 일반적으로는 모델의 반복 수립과 평가에 따른 비용이 많이 들기 때문에, 각각의 토픽과 단어가 같은 확률로 등장한다는 가정을 가지고 대칭적(symmetric) Dirichlet 사전확률 분포를 사용하는 방법과 미리 정의된 비대칭적(asymmetric) Dirichlet 사전확률 분

포를 사용하는 방법, 그리고 학습 과정에서 스텝마다 사전확률을 갱신하는 자동화된 기법(auto)을 주로 사용한다. 이러한 세 가지 방법들은 자연어 처리 Python 패키지 Gensim에서 제공하고 있으며, 본 연구에서도 가장 적합한 파라미터를 추정하기 위해, 세 가지 방법을 모두 활용하였다. 적합한 토픽의 수를 판단하는 기준은 토픽들의 의미적 응집도를 나타내는 Coherence score를 주로 활용하고 있으며, 여러 유형의 Coherence score 중 c_v 유형의 Coherence score가 가장 좋은 성능을 나타낸다고 알려져 있다 (Roder, Both, & Hinneburg, 2015). 본 연구에서도 이에 따라 c_v Coherence score를 기준으로 적합한 모델을 선정하였다. <그림 3>은 특허 내용 분석에 대한 절차를 보여준다.

Ⅳ. 분석 결과

1. 장기특허 분석 결과

앞서 사전에 학습한 장기특허 모델을 인공지능 기계학습 분야 특허들에 적용한 결과, 분석에 활용한 3281개(특허 등록인 기준 3409개)의 특허 중 2004개(특허 등록인 기준 2084개)의 특허가 20년 동안 유지가 될 것으로 보이는 장기특허로 예상되었고, 1277개(특허 등록인 기준 1325개)의 특허는 짧은 시간 동안만 유지될 것으로 보이는 단기특허로 예상되었으며, 둘의 비율은 1.570:1인 것으로 나타났다. 학습에 활용되었던 장기특허(211,398개)와 단기특허(80,641개)의 비율이 2.621:1이었다는 점을 고려해보면, 인공지능 분야에서는 일반적인 특허들과 비교해보았을 때 장기특허가 차지하는 비중이 다소 낮게 나타남을 알 수 있다. <표 6>, <표 7>, <표 8>, <표 9>는 장기, 단기특허들의 속성들에 대한 대략적인 정보를 보여준다.

<표 6> 전체, 장기, 단기특허 속성에 대한 통계치

발행 연도	특허 개수			평균 피인용 회수		
	전체	장기특허	단기특허	전체	장기특허	단기특허
2011	1	1	0	31.00	31.00	0
2012	18	16	2	49.61	42.00	110.50
2013	75	62	13	19.16	18.34	23.08
2014	321	251	70	42.73	18.48	129.70
2015	668	484	184	24.04	23.57	25.27
2016	865	554	311	32.37	26.81	42.27
2017	949	523	426	21.59	17.86	26.18
2018	116	25	91	33.70	67.00	24.55
2019	268	88	180	27.63	36.15	23.47

<표 7> 단기특허 속성에 대한 통계치

발행 연도	특허 개수	평균 피인용 회수
2012	2	110.50
2013	13	23.08
2014	70	129.70
2015	184	25.27
2016	311	42.27
2017	426	26.18
2018	91	24.55
2019	180	23.47

<표 8> 장기특허 등록인 상위 10개

등록인	특허 개수	백분율
IBM	274	13.15
Google	109	5.23
Microsoft Technology Licensing	72	3.45
Qualcomm	49	2.35
Amazon Technologies	39	1.87
Brain	28	1.34
Xerox	26	1.25
Microsoft	24	1.15
HRL Laboratories	24	1.15
Sap	24	1.15

<표 9> 특허 인용 수 기준 상위 장기특허 10개

등록인	특허 제목	인용 횟수
Cisco Technology	Conditional policies	560
DGS Global Systems	Systems, methods, and devices for automatic signal detection with temporal feature extraction within a spectrum	353
GE Intelligent Platforms	Method of sequential kernel regression modeling for forecasting and prognostics	291
Primal Fusion	Systems and methods for semantic concept definition and semantic concept relationship synthesis utilizing existing domain definitions	279
Google	Evaluating quality based on neighbor features	267
IBM	Measuring and displaying facets in context-based conformed dimensional data gravity wells	259
Nara Logics	Apparatus and method for providing harmonized recommendations based on an integrated user profile	259
Amazon Technologies	Machine learning based content delivery	255
Siemens Aktiengesellschaft	Actuation of a technical system based on solutions of relaxed abduction	251
Good Start Genetics	Variant database	246

2. 특허 내용 분석 결과

장기특허로 판별된 3281개의 특허를 대상으로 내용 분석을 수행하기 위해, 해당 특허들의 제목과 초록을 대상으로 텍스트에 대한 전처리를 거쳐 LDA 모델링을 실시하였다. 최적의 파라미터를 갖는 모델을 찾기 위해, 토픽의 개수를 1에서 31까지 범위에서 조정하며 가장 높은 Coherence score를 나타내는 모델을 선정하였으며, 최종적으로 6개의 토픽을 갖고 alpha와 beta값은 auto 설정으로 학습된 모델이 선정되었다(Coherence score = 0.501). 도출된 6개의 토픽에 대한 세부 내용은 다음과 같다.

1) 이미지/자연어 분류, 질의응답, 인식 기술

첫 번째 주제에서는, 이미지(image), 지도(map), 벡터(vector), 화자(agent), 범주

(category), 대답(answer), 순서(sequence), 부호화(encoding), 목소리(voice), 식별자(identifier), 의미(semantic), 언어(language)등의 단어가 출현하였으며, 키워드를 바탕으로 토픽을 해석하면, 이미지처리와 자연어 처리 분야에서 다뤄져 왔던 전통적인 주제인, 분류, 질의응답, 인식, 이해 등과 관련된 특허 내용을 담고 있는 것으로 볼 수 있다.

2) 자율주행 자동차 인터페이스 기술

두 번째 주제에서는, 이미지(image), 감정(sentiment), 물음(query), 인공지능(AI), 화면(display), 활동(active), 자동차(vehicle), 영역(region), 관심(interest), 파악(configuration), 작동(operate), 스마트(smart), 추천(recommend) 등의 단어가 출현하였으며, 키워드를 바탕으로 토픽을 해석하면, 자율주행 자동차에서 내/외부 상황에 대한 정보를 제공하고, 사용자 정보(예: 감정)를 활용하여 맞춤형 추천 서비스를 제공하는 화면 인터페이스 관련 특허 내용을 담고 있는 것으로 볼 수 있다.

3) 개인화된 헬스케어 맞춤 솔루션 기술

세 번째 토픽에서는, 문서(document), 개인(individual), 탐색(navigation), 최적화(optimization), 목록(list), 속성(attribute), 해결책(solution), 프로필(profile), 유전자(gene), 무선(wireless), 피트니스(fitness), 유전자형(genotype) 등의 단어가 출현하였다. 키워드를 바탕으로 토픽을 해석하면, 개인 정보를 활용하여 사용자의 일상과 관련된 정보(특히 무선 스마트 디바이스 등을 활용한 피트니스 정보)를 제공하는 기술과 관련된 특허 내용을 담고 있는 것으로 볼 수 있다.

4) 자원의 군집화 및 할당/배치 기술

네 번째 토픽에서는, 군집(cluster), 파악(configuration), 속성(attribute), 백엔드(backend), 영역(zone), 학습된(trained), 배달(delivery), 학습(learning), 해답(solution), 자식(child), 충돌(conflict), 거래(trade) 등의 단어가 출현하였다. 키워드를 바탕으로 토픽을 해석하면, 분산된 자원들의 속성을 토대로 군집화 분석을 수행하고, 트리 구조 형태로 되어 있는 부모-자식 노드에 자원을 효과적으로 할당하는 문제를 해결하기 위한 기술 특허 내용을 담고 있는 것으로 볼 수 있다.

5) 이미지처리용 신경망 구조 개발 기술

다섯 번째 토픽에서는, 회로(circuit), 층(layer), 문(gate), 이미지(image), 플로트(float), 컨볼루션(convolutional), 시냅틱(synaptic), 스파이크(spike), 세포(cell), 부호화(encode), 반복(iteration), 전극(electronode) 등의 단어가 출현하였다. 키워드를 바탕으로 토픽을 해석하면, 이미지처리에 효과적일 수 있는 신경망 구조를 개발하기 위한 기술 특허 내용들을 담고 있는 것으로 볼 수 있다.

6) 서버 자원 및 통신 최적화 기술

여섯 번째 토픽에서는, 지도(map), 모니터(monitor), 자원(resource), 엔진(engine), 이동(travel), 서버(server), 클라이언트(client), 프라이버시(privacy), 할당(allocate), 버짓(budget) 등의 단어가 출현하였다. 키워드를 바탕으로 토픽을 해석하면, 많은 지원이 연결된 서버 환경에서 지원을 관찰하고 통신을 원활히 하기 위한 서버 최적화 관련 특허 내용을 담고 있는 것으로 볼 수 있다.

Ⅴ. 토의

수집된 인공지능 기계학습 관련 특허에 대한 통계 분석 수행 결과, 상위 10개의 기업이 전체 특허의 약 32%를 차지하는 쏠림 현상이 두드러지게 나타났으며, 장기특허 분석에서도 비슷한 결과를 확인할 수 있었다. 이는 소수의 대기업이 대부분의 인공지능 기술 관련 특허를 소유하고 있음을 의미하며, 새로운 인공지능 관련 기술을 개발하기 위해서는 많은 자원과 지원이 필요함을 보여준다고 할 수 있다. 한편, 등록인별 특허 수 통계 분석과는 달리, 많은 인용 수 특허를 가지고 있는 등록 특허의 등록인을 분석한 결과에서는 Blanding Hovenweep, Newvaluexchange, Primal Fusion, Nara Logics, Good Start Genetics 등과 같이 잘 알려지지는 않았지만 높은 기술력을 가진 회사들이 등장하기도 하였다. 이는 상대적으로 자원과 지원이 떨어지는 소규모 기업들은 특허 등록 수 측면에서는 뒤떨어질 수 있지만, 가치가 높은 소수 특허를 출원할 수 있다는 것을 나타낸다. 다시 말해, 인공지능 기술 선점을 위해서는 많은 자본을 바탕으로 한 기업 주도의 투자와 동시에, 높은 기술력을 바탕으로 한 스타트업 회사에 대한 투자가 함께 이뤄져야 함을 시사한다.

최근 10년 동안의 특허 등록 수를 살펴보면, 2011년 단 하나의 특허만이 등록된 것에서 시작해 지속해서 성장세를 나타내다가 2017년을 기준으로 급격한 내림세를 나타내는 것을 확인할 수 있었다. 특히, 장기특허로 분류된 특허들의 등록 수는 2016년에 가장 높은 숫자를 나타냈다가 2017년 다소 감소한 뒤, 2018년부터 급격한 내림세를 나타냈는데, 이는 인공지능에 관한 관심은 지속해서 높아져 온 것에 반해 실질적으로 기술과 관련된 특허 등록은 2015년과 2017년 사이에 대부분 이뤄졌으며, 인공지능 관련 기술의 상당 부분이 이미 선점 상태에 놓여 있다는 것을 의미한다. 이에 따라 향후 인공지능 관련 기술에 대한 특허 등록은 점차 어려워질 것이며, 새로운 기술 개발을 위해서 들여야 할 자원이 더 커질 것으로 보인다.

딥러닝 모델을 통해 도출한 장기특허와 단기특허 사이에는 피인용 수 측면에서 유의한 차이를 확인할 수 있었다. 장기특허들의 피인용 수 평균과 표준편차는 각각 23.42, 40.38, 단기특허들의 피인용 수 평균과 표준편차는 각각 35.25, 236.50을 나타냈으며, 두 집단의 차이는 95% 수준에서 통계적으로 유의한 차이를 나타냈다 ($t = 2.189$, $p\text{-value} = 0.029$). 이는 장기특허가 단기특허들보다 상대적으로 낮은 피인용 수를 나타내지만, 특허 등록인이 특허 권리를 오랫동안 유지할 것으로 예상된다는 것을 의미한다. 이러한 결과는, 피인용 수와 패밀리 특허 수와 같이 전통적으로 활용됐던 유망 특허 지표와는 달리, 실제 특허 등록인이 오랫동안 권리를 유지하고 싶은 특허는 따로 있을 수 있음을 시사한다. 다시 말해, 유망 특허에 대한 판단 기준이 다양한 측면에서 이뤄져야 할 필요성이 있음을 보여준다고 할 수 있다.

장기특허와 단기특허의 인용수 관계를 보다 구체적으로 살펴보면, 2012년부터 2017년까지의 특허는 단기특허의 피인용 횟수가 장기특허의 피인용 횟수보다 높은 수치를 나타냈었고, 그 이후인 2018년, 2019년 특허에서는 장기특허의 피인용 횟수가 단기특허의 피인용 횟수를 앞서는 것으로 나타났다. 이는 단기적 관점에서 유용할 것으로 보이는 특허들을 주로 인용하던 방식이 점차 장기적 관점에서 유용할 것으로 보이는 특허들을 주로 인용하는 방식으로 변하고 있다는 것을 보여준다. 즉, 기업들의 인공지능 기술 개발 방향이 점차 장기적 관점에서 이뤄지기 시작했다는 점을 보여주며, 이에 따라 인공지능 관련 정책적 흐름 역시 단기적 기술 수준 향상을 넘어 지속적이고 장기적인 기술 수준 향상으로 이어질 수 있는 방향으로 제시되어야 함을 시사한다.

분석에 활용한 2011-2017 특허 3281개(특허 등록인 기준 3409개) 중 2004개(특허 등록인 기준 2084개)의 특허가 장기특허로 예상되었고, 1277개(특허 등록인 기

준 1325개)의 특허가 단기특허로 예상되었다. 이 둘의 비율은 1.570:1인 것으로 나타났으며, 학습에 활용되었던 장기특허(211,398개)와 단기특허(80,641개)의 비율(2.621:1)보다 다소 낮은 수치를 나타내었다. 이는 인공지능 분야에서는 일반적인 특허들과 비교해보았을 때 장기특허가 차지하는 비중이 다소 낮게 나타나고 있음을 의미하며, 인공지능 분야에서의 기술 성장이 질적인 측면보다 양적인 측면에 초점이 맞춰져 급격히 이뤄졌음을 의미한다. 이러한 결과는 향후 인공지능 관련 정책이 인공지능 전반에 걸쳐 이뤄지기보다는 실질적이고 활용성이 높은 기술에 대해 초점을 맞춰 선별적으로 실행되는 것이 효과적일 수 있음을 나타낸다.

토픽 모델링을 통해 도출된 유망 특허들의 내용은 기계학습 분야에서 꾸준히 논의되고 있는 주제뿐만 아니라 공학 분야에서 전통적으로 관심을 가져왔던 최적화, 네트워크 문제들을 아우르고 있었다. 구체적으로는, 이미지, 음성과 관련된 딥러닝 기술을 포함해 자율주행, 헬스케어와 같은 응용 분야와 관련된 기술들이 나타났으며, 네트워크 최적화와 같은 자원 할당 문제에 관한 내용도 다루고 있는 것으로 나타났다. 이와 같은 결과는 기업이 소유하고 있는 인공지능 관련 기술 특허 중에서 장기적이고 전략적으로 유지하기를 원하는 기술들은 인공지능에 특화된 기술뿐만 아니라 전통적인 정보통신기술을 포함하고 있다는 것을 의미하며, 전통적인 정보통신기술 위에서 유망한 인공지능 관련 기술을 접목해 나가며 전문성 역량을 키워나가고자 하는 시도를 하고 있다고 해석할 수 있다.

그러나 기초 기술의 중요성을 너무 강조한 나머지, 인공지능 기술에 대한 투자가 한쪽으로 편향되어서는 안 될 것이다. 실제로, 특허 등록, 인용 측면에서 상위에 노출된 기업들의 사례들을 살펴보면, 특정 분야에 편중된 기술 투자는 인공지능 분야에서의 기술 경쟁력을 저하할 수 있음을 알 수 있다. Google과 Microsoft, Amazon과 같은 기업들은 전통적인 정보통신기술 분야뿐만 아니라 이미지/자연어 처리, 헬스케어, 자율주행과 같은 인공지능 관련 분야에 적극적인 투자를 수행해 왔던 기업들로, 지금까지도 관련 분야에서 선도적인 위치를 차지하고 있다는 평가를 받고 있다. 반면 IBM과 Cisco와 같은 기업들은 서버 자원 할당, 관리 등과 관련된 전통적인 정보통신기술 분야에 대한 투자에 집중하고, 인공지능 분야에 대한 투자를 상대적으로 늦게 시작한 결과, 앞서 언급된 기업들에 비해 상대적으로 인공지능 기술이 뒤처졌다는 평가를 받고 있다. 이러한 인식은 최근 10년간 나타난 각 기업의 시가총액 변화를 비교해보면 더욱 명확해진다. 이와 같은 사례는 향후 인공지능 관련 분야에 대한 정책적 방향은 인공지능의 근간이 되는 정보통신기술

과 이를 응용한 인공지능 기술에 대한 연구개발을 함께 장려하는 방향으로 수립되어야 한다는 점을 다시 한번 강조하고 있다.

Ⅵ. 결론

본 연구에서는 특허를 통해 인공지능 기계학습 분야에서 최근 10년간 이뤄지고 있는 연구들의 특징에 대해 살펴보고, 이 중 장기특허를 추출하여 장기적으로 유지될 것으로 보이는 특허들의 특징과 내용을 살펴보았다. 이를 위해 미국 특허청에 등록된 인공지능 기계학습 관련 등록 특허 데이터를 수집하는 한편, 전체 특허들을 대상으로 장기간 유지되었던 특허들을 학습하고, 향후 장기간 유지될 것으로 보이는 장기특허를 예측하는 딥러닝 모델을 활용하였다. 딥러닝 모델을 통해 예측된 장기특허들의 내용은 토픽 모델링을 통해 구체적으로 분석되었으며, 기초 통계치와 함께 종합적인 측면에서 논의되었다. 분석을 통해 본 연구에서는, (1) 인공지능 관련 기술에 대한 투자는 대기업뿐만 아니라 기술력을 바탕으로 이뤄진 일부 기업들에 대한 지원이 함께 이뤄져야 한다는 점, (2) 인공지능 관련 기술 개발이 점차 어려워지는 만큼, 추가적인 자원과 지원이 필요할 수 있다는 점, (3) 피인용수 측면에서는 낮아 보이는 특허들이 실제로는 권리 행사가 장기간 유지되는 전략적 특허일 수 있으므로, 유망 특허 발굴에 대한 여러 각도에서의 접근이 필요하다는 점, (4) 인공지능 관련 기술정책의 방향은 점차 단기에서 장기적 방향으로 나아가야 한다는 점, (5) 인공지능 관련 기술 특허들은 양적 성장을 중심으로 급격히 성장해온 만큼, 유망 기술을 중심으로 한 선별적 투자가 이뤄져야 한다는 점, (6) 인공지능 분야의 장기전략적 기술들에 대한 투자는 전반적인 분야에 걸쳐 이뤄져야 한다는 점 등을 제시하였다.

도출된 연구 결과의 정확도를 높이기 위해서는 향후 기계학습 분야뿐만 아니라 보다 다양한 범위에서의 인공지능 관련 특허 연구가 이뤄질 필요가 있다. 또한, 특허 등록인의 명칭 문제(동일 회사이면서 다른 이름을 쓰는 경우, 모회사/자회사인 경우, 특허 관리를 위한 별도 회사인 경우 등)와 장기특허 모델의 정확도 향상 문제(학습 모델의 개선 및 특징 벡터 변화 등)에 대한 해결을 통해 더욱 높은 정확도를 갖는 장기특허 예측 기술을 마련할 수 있을 것으로 보인다. 또한, 활용할 수 있는 등록 특허의 양과 온전한 형태의 특허 제목, 초록이 확보된다면, 워드 임베딩

과 같이 더욱 정교한 형태의 텍스트 마이닝 기법들을 통해 깊이 있는 특허 내용 분석이 이뤄질 수 있을 것으로 보인다.

참고문헌

박선재 (2019. 1. 3). 꽃길 걷던 IBM 인공지능, 가시밭길 접어든 이유는? 『Medical Observer』, http://www.monews.co.kr/news/articleView.html?idxno=123818.

이재민·고병열·윤장혁 (2019). 딥러닝 기반 글로벌 특허기술 장기전략 예측. 『KISTI Data Insight』, 10, 1-21.

최재웅·정병기·윤장혁·권오진·고병열·이재민 (2018).『특허수명예측장치 및 그 동작 방법』(등록번호 10-1877235-0000). 대한민국특허청.

Agrawal, A., Fu, W., & Menzies, T. (2018). What is wrong with topic modeling? And how to fix it using search-based software engineering. *Information and Software Technology*, *98*, 74-88.

Aristodemou, L., & Tietze, F. (2018). The state-of-the-art on intellectual property analytics (IPA): A literature review on artificial intelligence, machine learning and deep learning methods for analysing intellectual property (IP) data. *World Patent Information*, *55*, 37-51.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Cho, H. P., Lim, H., Lee, D., Cho, H., & Kang, K.-I. (2018). Patent analysis for forecasting promising technology in high-rise building construction. *Technological Forecasting and Social Change*, *128*, 144-153.

Choi, J., Jeong, B., Yoon, J., Coh, B., & Lee, J.-M. (2020). A novel approach to evaluating the business potential of intellectual properties: A machine learning-based predictive analysis of patent lifetime. *Computer & Industrial Engineering*, *145*, 106544.

Choi, D., & Song, B. (2018). Exploring technological trends in logistics: Topic modeling-based patent analysis. *Sustainability*, *10(8)*, 2810.

Evangelista, A., Ardito, L., Boccaccio, A., Fiorentino, M., Petruzzelli, A. M., & Uva, A. E. (2020). Unveiling the technological trends of augmented reality: A patent analysis. *Computers in Industry*, *118*, 103221.

Guellec, D., & de la Potterie, B. v. P. (2000). Applications, grants and the value of patent. *Economics Letters*, *69(1)*, 109-114.

Hoffman, M., Bach, F. R., & Blei, D. M. (2010). Online learning for latent dirichlet allocation. *Proceedings of the 23rd International Conference on Neural Information Processing Systems*, 856-864.

IDC. (2019). IDC survey finds artificial intelligence to be a priority for organizations but few have implemented an enterprise-wide strategy, https://www.idc.com/getdoc.jsp?containerId=prUS45344519.

Kang, J., Lee, J., Jang, D., & Park, S. (2019). A methodology of partner selection for sustainable industry-university cooperation based on LDA topic model. *Sustainability*, *11(12)*, 3478.

Kim, G., & Bae, J. (2017). A novel approach to forecast promising technology through patent analysis. *Technological Forecasting and Social Change*, *117*, 228-237.

Kim, K. H., Han, Y. J., Lee, S., Cho, S. W., & Lee, C. (2019). Text mining for patent analysis to forecast emerging technologies in wireless power transfer. *Sustainability*, *11(22)*, 6240.

Pathik, N., & Shukla, P. (2020). Simulated annealing based algorithm for tuning LDA hyper parameters, In M. Pant, T. K. Sharma, R. Arya, B. C. Sahana, H. Zolfagharinia (Eds.), *Soft Computing: Theories and Applications* (pp. 515-521). Singapore: Springer.

Roder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures, *Proceedings of the 8th ACM International Conference on Web Search and Data Mining*, 399-408.

Schwab, K. (2016). Re: The fourth Industrial revolution: What it means, how to respond, https://www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/.

Shubbak, M. H. (2019). Advances in solar photovoltaics: Technology review and patent trends. *Renewable and Sustainable Energy Reviews*, *115*, 109383.

Wang, X., Yang, X., Wang, X., Xia, M., & Wang, J. (2020). Evaluating the competitiveness of enterprise's technology based on LDA topic model. *Technology Analysis & Strategic Management*, *32(2)*, 208-222.