

‘인공지능’, ‘기계학습’, ‘딥 러닝’ 분야의 국내 논문 동향 분석

박홍진*

Trend Analysis of Korea Papers in the Fields of ‘Artificial Intelligence’, ‘Machine Learning’ and ‘Deep Learning’

Hong-Jin Park*

요약 4차 산업혁명의 대표적인 이미지 중 하나인 인공지능은 2016년 알파고 이후에 인공지능 인식이 매우 높아져 있다. 본 논문은 한국교육학술정보원에서 제공하는 국내 논문 중 ‘인공지능’, ‘기계학습’, ‘딥 러닝’으로 검색된 국내 발표 논문에 대해서 분석하였다. 검색된 논문은 약 1만여건이며 논문 동향을 파악하기 위해 빈도분석과 토픽 모델링, 의미 연결망을 이용하였다. 추출된 논문을 분석한 결과, 2015년에 비해 2016년에는 인공지능 분야는 600%, 기계학습은 176%, 딥 러닝 분야는 316% 증가하여 알파고 이후에 인공지능 분야의 연구가 활발히 진행됨을 확인할 수 있었다. 또한, 2018년 부터는 기계학습보다 딥 러닝 분야가 더 많이 연구 발표되고 있다. 기계학습에서는 서포트 벡터 머신 모델이, 딥 러닝에서는 텐서플로우를 이용한 컨볼루션 신경망이 많이 활용되고 있음을 알 수 있었다. 본 논문은 ‘인공지능’, ‘기계학습’, ‘딥 러닝’ 분야의 향후 연구 방향을 설정하는 도움을 제공할 수 있다.

Abstract Artificial intelligence, which is one of the representative images of the 4th industrial revolution, has been highly recognized since 2016. This paper analyzed domestic paper trends for ‘Artificial Intelligence’, ‘Machine Learning’, and ‘Deep Learning’ among the domestic papers provided by the Korea Academic Education and Information Service. There are approximately 10,000 searched papers, and word count analysis, topic modeling and semantic network is used to analyze paper's trends. As a result of analyzing the extracted papers, compared to 2015, in 2016, it increased 600% in the field of artificial intelligence, 176% in machine learning, and 316% in the field of deep learning. In machine learning, a support vector machine model has been studied, and in deep learning, convolutional neural networks using TensorFlow are widely used in deep learning. This paper can provide help in setting future research directions in the fields of ‘artificial intelligence’, ‘machine learning’, and ‘deep learning’.

Key Words : Artificial Intelligence, Deep Learning, Machine Learning, Paper trend, Semantic Network, Topic Modeling, Word Count Analysis

1. 서론

4차 산업혁명은 빅데이터를 기반으로 하여 인공지능, 기계학습과 딥 러닝 발달, 그리고 사물인터넷(IoT), 클라우드를 통해 인간뿐만 아니라 모든 기기가 연결되는 초연결 시대를 의미한다. 인공지능 기술의 발

달로 인해 인간의 지식 노동까지도 기계가 대체하는 초지능화 시대, 홀로그램등의 기술 발달로 인한 초현실 시대가 도래함을 의미한다. [1] 연구조사에 의하면 4차 산업혁명의 대표 이미지로 인공지능(28.6%) 분야가 가장 높았으며, 그 다음 사물인터넷(22.3%), 빅데이터(20.5), 클라우드(5.7%) 순이었다. 인공지능 발전이 미

This Paper was supported by research Fund of Sangji University in 2018.

*Dep. of Computer Engineering, Sangji University(hjpark1@sangji.ac.kr)

Received August 11, 2020

Revised August 14, 2020

Accepted August 14, 2020

래 삶의 질 향상에 대한 기대가 높고 전문적이고 정확한 서비스 등 경제적인 문제보다 생활 개선에 대한 기대감이 높은 것을 파악되었다. 4차 산업혁명 기술의 핵심 기술 중 인공지능을 가장 대표적인 이슈로 파악하고 있음을 알 수 있다.

4차 산업혁명의 핵심 기술 중 하나인 인공지능은 2016년 3월 구글의 알파고(AlphaGo)와 이세돌 9단의 바둑 대결 이후로 인공지능 인식이 매우 높아져 있다. 알파고는 16만 건의 기보로 배우는 지도 학습과 함께 바둑 지능을 학습시켰다. 알파고는 기계학습 중 딥 러닝을 기반으로 설계되었다. 일반적으로 바둑을 계산해야 할 경우의 수는 250의 150제곱이다. 몬테카를로 시뮬레이션을 사용하는 알파고는 딥 러닝 방식 없이 바둑을 둔다면 이 모든 경우의 수를 계산해야 하지만 이는 현실적으로 불가능하다. 알파고는 딥 러닝 기반으로 학습시켜서, 경우의 수를 줄이는 것이다[2]. 이에 따라 알파고에서 이용된 인공지능 분야 뿐만 아니라, 기계학습, 딥 러닝 기술의 논문 연구 동향 파악은 필요한 시점이다. 현재 이와 관련하여 기존 논문들을 분석해보면 인공지능 논문 연구 동향 파악에 집중되어 있고 기계학습이나 딥 러닝에 대한 논문 연구 동향 파악은 매우 미비하다.

본 논문은 한국교육학술정보원에서 제공하는 학술연구정보서비스(www.riiss.kr)에서 ‘인공지능’, ‘기계학습’, ‘딥 러닝’으로 검색된 국내 논문 데이터를 분석하여 인공지능 뿐만 아니라 기계학습, 딥 러닝의 논문 연구 동향을 파악 하고자 한다. 학술연구정보서비스는 1988년 개통 이후로 2019년 12월 기준 평균 20만 명의 연구자들이 학술연구정보 서비스를 이용하고 있으며, 일 평균 약 8만 1천명이 이용하고 있는 국내 최대 논문 제공 사이트이다[3][4]. 본 논문에서는 인공지능, 기계학습, 딥 러닝의 논문 연구 동향을 파악하고자, 2019년 12월 까지 검색된 논문을 대상으로 하였다. 검색된 논문은 총 11,473개의 논문 중 중복이나 오류를 제외한 9,888개의 논문을 대상으로 하였다. 대략 1만편의 논문의 처리하기 위해 텍스트 마이닝(Text mining) 기술을 이용하였다.

본 논문의 목적은 다음과 같다. 첫 번째, 2016년 알파고 등장으로 인한 인공지능, 기계학습, 딥 러닝의 논

문의 양적인 증가는 어느 정도인가? 이는 알파고 이후 인공지능, 기계학습, 딥 러닝 논문 연구가 어느 정도 증가 하였는가를 알아볼 수 있을 것이다. 두 번째는 텍스트 마이닝의 기본적인 분석 기법으로 빈도 분석을 이용하여 인공지능, 기계학습, 딥 러닝 분야에서 많이 연구되고 있는 연구 동향을 살펴본다. 세 번째는 토픽 모델링과 의미 연결망 분석을 통해서 연구되는 분야는 어떻게 서로 의미 구조로 연계되었는지와 각 주제에 대한 토픽을 추출한다.

2. 이론 배경과 관련 논문

2.1 인공지능과 텍스트 마이닝

인공지능은 2016년 알파고에 등장으로 다시 한번 활발한 연구가 진행되고 있다. 알파고에서 사용된 기술은 체협적 탐색 알고리즘인 몬테카를로 시뮬레이션과 기계학습 기술 중 딥 러닝 기술 등이 이용되었다. 기계학습은 인공지능 연구 분야 중 하나로, 인간의 학습 능력과 같은 기능을 컴퓨터에서 실현하고자 하는 기술 및 기법이다. 기계학습을 구현 모델로 의사결정나무, 앙상블학습, 신경망, 회귀분석, 군집분석등 다양한 모델이 있다. 딥 러닝은 기계학습의 한 모델로, 다양한 분야에서 기존 기계학습 방법보다 뛰어난 성능을 보이고 있다[5].

텍스트 마이닝은 언어학, 통계학, 기계 학습 등을 기반으로 한 자연언어 처리 기술을 활용하여 반정형 및 비정형 텍스트 데이터를 정형화하고, 특징을 추출하기 위한 기술과 추출된 특징으로부터 의미 있는 정보를 발견할 수 있도록 하는 기술이다. 즉, 텍스트 데이터에서 가치와 의미가 있는 정보를 찾아내는 기술이다. 텍스트 마이닝 과정은 먼저 크롤링(Crawling)이나 스크래핑(Scrapping)을 통해 데이터를 수집한다. 수집된 데이터를 분석에 적합하도록 토큰화, 정제화, 정규화, 불용어 처리등과 같은 작업을 수행시킨다. 그다음 텍스트 데이터의 특성과 유형을 추출하여 의미 있는 지식을 추출하기 위해 빈도 분석, 군집 분석, 토픽 모델링 등 다양한 분석 방법을 적용하여 텍스트를 분석한다. 마지막 단계에서는 분석된 텍스트를 평가하여 적용시킨다. 텍스트 마이닝의 주요 응용 분야로는 문서 분류,

문서 군집, 정보 추출, 문서 요약, 감정 분석등이 있다 [6-8]. 토픽 모델링은 문서를 이루는 핵심어 바탕으로 문서에서 토픽, 즉 주제를 추출해 주는 확률모델 알고리즘이다. 단어의 빈도 수만으로는 단어의 의미를 정확히 고려하지 못하기 때문에 문서의 주제(잠재 의미)를 찾아내기가 어렵다[9-11]. 이에 대한 대안이 바로 문서의 잠재된 의미를 찾아내는 잠재 의미 분석(Latent Semantic Analysis, LSA)나 단어가 특정 토픽에 존재할 확률과 문서에 특정 토픽이 존재할 확률을 결합 확률로 추정하여 토픽을 추출하는 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 알고리즘등을 이용한다[12]. 본 논문은 잠재 디리클레 할당 알고리즘을 이용한다.

2.2 관련 논문

[13] 논문은 인공지능과 관련된 주요 오픈 소스 소프트웨어 프로젝트들을 분석하여, 인공지능 기술 개발 현황에 대한 보다 실질적인 동향을 파악하였다. 이를 위하여 깃허브(Github) 상에서 2000년부터 2018년 7월까지 생성된 인공지능과 관련된 주요 프로젝트들의 목록을 검색 및 수집하였으며, 수집된 프로젝트들의 특징과 기술 분야를 의미하는 토픽 정보들을 대상으로 텍스트 마이닝 기법을 적용하여 주요 기술들의 개발 동향을 연도별로 상세하게 확인하였다. [14] 논문은 국내 인공지능 분야에 연구 동향을 분석하였다. 인공지능 키워드에 1985년부터 2018년까지 총 1,691편의 학술 논문에서 토픽 모델링과 의미연결망 분석을 중심으로 연구하였다. 분석 프로그램은 오픈 소프트웨어인 R을

이용하였으며, 인공지능과 인간, 인공지능과 기술이라는 주요 토픽을 추출하였다. [15] 논문은 논문 데이터 분석을 통한 인공지능 분야의 주요 연구 동향 분석으로 'Web of Science'에서 한국인 저자가 게재한 SCI(E)저널의 논문 중 인공지능 관련 논문을 수집하여 빈도 분석과 키워드 네트워크 분석하였다. [16] 논문은 특허 분석을 통해 인공지능 기술 분야의 연구 동향을 파악하였다. 인공지능 기술 분야에서 WIPSON 특허 데이터베이스를 사용하여 2017년 3월까지의 한국, 미국, 일본, 유럽, 중국 특허청에서 등록된 공개된 출원일 기준으로 분석하였다. [17] 논문은 2013년부터 2017년까지 5년간 KCI(Korea citation index) 등재지에 게재된 경영학 분야 논문을 웹 크롤러를 활용하여 논문을 수집하고, 데이터 클리닝 및 전처리 과정을 거쳐 4,210편에 대한 논문의 주제어 분석하였다.

기존 [13-16] 논문을 보면, 주로 인공지능 연구 동향 분석이 이루어지고 있음을 알 수 있다. 깃허브 상에서 생성된 인공지능 관련 주요 프로젝트를 분석하든지 [13], 인공지능과 인간, 인공지능과 기술의 토픽을 추출하였다[14]. [15]는 SCI(E) 저널 논문을 분석하였고, [16]은 특허 분석을 통해 인공지능 기술 분야의 연구 동향을 파악하였다. [17] 논문은 KCI의 등재된 경영학 분야의 논문을 분석하였다.

본 논문은 기존 논문과의 차이점은 인공지능 분야 뿐만 아니라, 기계학습, 딥 러닝 분야에서 논문 연구 동향을 같이 살펴보았다. 또한, 본 논문에서 파악된 논문의 양이 약 1만개 정도의 대용량의 논문 데이터이다.

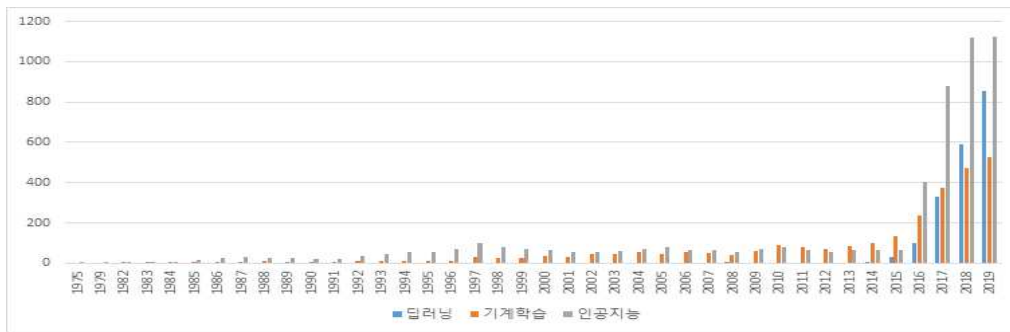


그림 1. 인공지능, 기계학습, 딥 러닝 논문 편수(1975-2019)
 Fig. 1. Numbers of AI, ML, Deep Learning Papers(1975-2019)

3. 국내 논문 동향 분석

본 논문은 학술연구정보서비스(RISS)에서 검색어를 '인공지능', '기계학습', '딥 러닝'으로 검색된 데이터는 인공지능은 6,101편 중 중복을 제외한 5,521편, 기계 학습은 3,117중 중복을 제외한 2,939편, 딥 러닝은 2,255편에서 중복을 제외한 2,112편에서 논문을 분석 대상으로 하였다.

3.1 인공지능, 기계학습, 딥 러닝 논문 편수 분석

1975년부터 2019년까지 인공지능, 기계학습, 딥 러닝과 관련하여 학술연구정보서비스에 년도별 논문 편수는 [그림 1]과 같다. [그림 1]을 보면, 논문 편수는 전체적으로 인공지능이 가장 많았으며, 그 다음 기계 학습과 딥 러닝 순으로 나타나 있다. 또한, 인공지능, 기계 학습, 딥 러닝 모두 해마다 발표되는 논문 편수가 지속적으로 증가하는 추세에 있음을 알 수 있다.

표 1. 인공지능, 기계학습, 딥 러닝 논문 편수(2014-2019)
Table 1. Numbers of AI, ML, Deep Learning Papers(2014-2019)

Years	'인공지능' Numbers	'기계학습' Numbers	'딥 러닝' Numbers
2014	66	98	4
2015	67	134	31
2016	406	237	98
2017	879	374	331
2018	1,119	471	591
2019	1,127	512	856

2016년 3월은 알파고가 등장한 해로써 알파고와 이세돌의 바둑 대전으로 인공지능의 관심이 어느 해보다도 급증되었던 해였다. [표 1]는 2016년을 기준으로 2014년부터 2019년까지 인공지능, 기계 학습, 딥 러닝 발표 논문 편수를 따로 추출해 보았다. [표 1]를 보면 크게 2가지 특징이 나타나고 있다. 첫째, 2016년을 기점으로 인공지능 논문은 2015년에 비해서 대략 600% 증가, 기계 학습은 176% 증가, 딥 러닝은 316%가 증가하였다. 이러한 발표된 국내 논문 편수를 보면 2016년 알파고 등장으로 인공지능 분야가 제2의 전성기를 맞이하고 있다고 볼 수 있다. 둘째는, 기계 학습과

딥 러닝을 보면, 2014년에 딥 러닝이 4편이 발표, 기계 학습이 98편으로 기계 학습이 월등히 많았으나, 2017년에는 딥 러닝과 기계 학습과 비슷하게 발표되었고, 2018년부터는 기계 학습 보다 딥 러닝이 분야가 더 많이 논문이 발표되고 있으며 2019년에서 그 격차가 더욱 커져 가고 있음을 알 수 있다.

3.2 핵심어 분석

핵심어(Keyword)란 텍스트 자료의 중요한 내용을 압축적으로 나타내는 단어나 문구를 의미한다. 핵심어 분석이란 노이즈 데이터를 제거하고, 중복된 단어를 정규화하여 단어의 등장 빈도를 분석함으로써 핵심어를 추출하는 과정으로 텍스트 마이닝 처리에서 가장 기본적인 분석 방법이다. 이는 특정 텍스트에서 많이 나타나는 단어가 그 텍스트의 주제가 될 가능성이 높다. 텍스트의 핵심어를 분석함으로써 텍스트의 주제가 무엇인지 짐작할 수 있고, 텍스트가 서로 어느 정보 비슷한지도 파악할 수도 있다. 텍스트 마이닝에서 가장 흔하게 사용하는 핵심어 분석 방법으로 단어의 단순 빈도를 제시하는 방법과 단어 구름의 빈도 분석 결과를 나타내는 워드 클라우드(Word cloud) 방법이 있다.

본 논문에서는 논문의 동향을 파악하기 위해 두 가지 빈도 분석을 수행하였다. 첫째, 논문 제목에서의 단어 빈도를 추출하여 논문 제목에서의 주요 연구 부분이 어느 부분인지를 알아보고자 하였다. 두 번째로, 논문에 나타나 있는 키워드의 단어의 빈도를 분석하여 논문 제목보다 세밀한 논문 분야 키워드에 대해서 알아보고자 하였다.

먼저, 논문의 제목에서의 단어 빈도를 추출하였으며 결과는 [표 2]와 같다. 학술연구정보서비스에서 '인공지능'으로 검색하여 논문 제목으로 단어의 빈도를 분석한 결과 인공지능 분야에서 연구 > 이용 > 분석 > 활용 > 개발 > 설계 > 구현 순으로 빈도수가 높았다. 인공지능 분야에서는 연구 논문이 개발논문에 대해 대략 3배, 설계에 비해 5배로, 개발이나 설계 분야보다 연구 및 이용 분야 논문이 많음을 알 수 있다.

표 2. 논문 제목 기반의 단어 빈도 분석
Table 2. Word Count Analysis based Paper Title

Rank	'인공지능'		'기계학습'		'딥 러닝'	
	word	count	word	count	word	count
1	지능	2,526	기계학습	691	러닝	1,382
2	인공	2,405	연구	524	연구	389
3	연구	1,190	예측	385	활용	283
4	이용	734	분석	359	시스템	275
5	시스템	632	시스템	295	예측	270
6	기술	506	모델	270	기술	226
7	분석	422	분류	263	모델	225
8	산업혁명	393	활용	250	분석	213
9	활용	383	데이터	224	인식	208
10	개발	381	알고리즘	218	영상	202
11	교육	308	개발	217	분류	177
12	로봇	236	기술	178	데이터	173
13	설계	235	방법	173	개발	157
14	알고리즘	228	인공	163	지능	136
15	예측	217	지능	149	알고리즘	135
16	미래	202	신경망	144	신경망	126
17	학습	194	교육	134	인공	124
18	모델	192	러닝	124	학습	106
19	동향	189	적용	122	이미지	103
20	적용	186	인식	117	방법	101
21	신경망	183	자동	115	탐지	98
22	서비스	180	성능	113	자동	89
23	인식	176	탐지	104	적용	83
24	산업	168	비교	101	성능	82
25	정보	167	설계	99	동향	80
26	인간	164	평가	97	객체	79
27	방법	148	한국어	85	설계	67
28	데이터	146	모형	82	추정	61
29	구현	140	효과	78	향상	60
30	사례	136	향상	75	환경	60

'기계학습'으로 검색하여 논문 제목을 단어 빈도를 분석한 결과 연구 > 분석 > 활용 > 개발 > 설계 순으로 빈도수가 높았다. 기계학습 논문도 연구 논문이 개발 논문 보다 2.4배가 많았다. 기계학습은 훈련 데이터를 통해 알려진 속성을 기반으로 예측에 초점을 분석하는 분야로 발표된 논문에서 추출된 논문 제목을 보면 '예측'과 '분석'이란 단어가 월등히 많이 포함되어 있음을 알 수 있다. '딥 러닝'으로 검색한 논문 제목의 단어 빈도를 비교해 보면 연구 > 활용 > 예측 > 분석 > 개발 > 설계 순으로 빈도수가 높았다. 딥 러닝 분야도 개발 논문 보다 연구 논문이 2.4배 높았으며, 딥 러닝 분야에서는 분석보다 활용이 많이 발표되어 있으며, 영상 인식 관련 분야가 많은 논문이 발표됨을 알 수 있다. 두 번째로는 논문의 작성된 키워드를 가지고 빈도

분석하였다. 텍스트를 분석하기 전에 토큰화와 정제 및 정규화가 매우 중요하다. 주어진 코퍼스(Corpus)에서 토큰이라는 불리는 단위로 나누는 작업을 토큰화(Tokenization)라고 한다. 키워드는 단어로 구성 되어 있기 때문에 토큰화를 수행 하지 않아도 된다. 정제(Cleaning)는 갖고 있는 코퍼스로부터 노이즈 데이터를 제거하는 작업을 의미한다. 정제를 수행하기 위해서 먼저, 영문자의 대소문자를 통합해야 된다. 영어권에서는 대부분의 글이 소문자로 작성되기 때문에 일반적으로 대문자를 소문자로 변환하는 소문자 변환작업을 진행하게 된다. 두 번째를 불필요한 단어를 제거해야 한다. 분석하고자 하는 목적에 맞지 않은 불필요 단어들의 노이즈 데이터라고 하는데 정제 작업에서 노이즈 데이터를 제거된다. 검색어 '인공지능'에서 추출된 논문의 키워드를 분석한 결과 [표 3]과 같은 노이즈 데이터를 알아 낼 수 있었다. 키워드란 논문의 주제에 맞는 핵심어이지만, '학술저널', 'KCI등재', '300'등 같은 단어가 논문의 핵심어가 될 수 없다. 데이터를 분석하기 전에 이러한 노이즈 데이터는 제거되었다. 노이즈 데이터 중 '학술저널'은 1,148회, 'KCI등재'는 1,071회 나와 1위와 2위를 나타내고 있었다.

표 3. 키워드 기반의 노이즈 데이터 분석
Table 3. Noise Data Analysis based Keyword

'학술저널', 'KCI등재','300','SCOPUS', '3', '4', '28', 'KCI우수등재', 'KCI등재 후보', '530', '40', '325', '560', '569', '453'
--

키워드를 분석한 결과 같은 의미의 단어들어 영어와 한글이 혼재되어 있었으며, 영어도 대문자 소문자가 중복되어 있었다. 또한, 같은 의미의 한글도 여러 가지로 표현되어 있었다. 정규화(Normalization)는 표현 방법이 다른 단어들을 통합시켜서 같은 단어로 만들어 주는 작업을 의미한다. 예를 들어, 'AI'와 '인공지능', 'Artificial Intelligence' 등은 모든 표현 방법이 다르지만 '인공지능'을 의미한다. 본 논문에서는 같은 의미의 영어와 한글이 표현되어 있으면, 한글로 정규화를 수행하였다. 예를들어 'Artificial Intelligence'와 '인공 지능'이 있으면, '인공지능'으로 정규화를 수행하였

다. 본 논문에서 수행된 정규화는 [표 4]와 같다.

표 4. 키워드 기반의 정규화 분석
Table 4. Normalization Analysis based Keyword

The same meaning words with different expressions	Normalized Word
"인공지능", "artificial intelligence", "artificial intelligence(artificial intelligence)", "인공지능 (인공지능)"	인공지능
"the fourth industrial revolution", "제4차 산업혁명", "fourth industrial revolution", "4th industrial revolution", "the 4차 산업혁명"	4차 산업혁명
"인공 신경망", "인공신경망", "neural network", "artificial 신경망", "신경망s", "신경망 (신경망)"	신경망
"mchine learning", "기계 학습(기계 학습)", "machine learning(기계 학습)", "기계 학습 (machine learning)", "machine learning (기계 학습)", "기계 학습 (기계 학습)", "머신러닝"	기계 학습
"iot", "internet of things", "사물 인터넷(사물 인터넷)", "사물인터넷"	사물 인터넷
"support vector machine", "svm", "지지벡터 기계", "서포트 벡터 머신(서포트 벡터 머신)", "서포트 벡터 머신(서포트 벡터 머신)"	서포트 벡터 머신
"deep learning", "deep-learning", "딥러닝", "딥 러닝(딥 러닝)", "딥 러닝 (딥 러닝)", "deep 신경망", "deep 신경망(심층 신경망)", "심층 신경망(심층 신경망)"	딥 러닝
"텍스트마이닝", "data mining", "데이터마이닝", "데이터 마이닝"	텍스트 마이닝
"convolutional 신경망", "합성곱 신경망", "cnn(cnn)", "cnn", "convolution 신경망", "cnn (cnn)", "컨볼루션 신경망 (컨볼루션 신경망)", "컨볼루션 신경망(컨볼루션 신경망)"	컨볼루션 신경망

[표 5]에 나타난 것처럼 '인공지능'을 검색어를 가지고 논문의 키워드를 분석한 결과, 신경망, 기계학습, 빅 데이터, 딥 러닝, 사물인터넷 중심으로 논문이 많이 발표되어서 4차 산업혁명에서 인공지능은 매우 핵심 기술이며, 활발하게 연구되고 있으며 논문으로 발표됨을 알 수 있다. 챗봇, 블록체인 등도 인공지능과 연계된 논문이 발표되고 있음을 알 수 있다. 기계학습은 지도 학습, 비지도 학습, 강화 학습으로 분류된다. '기계학

습'을 검색어를 가지고 논문의 키워드를 분석한 결과 기계학습의 다양한 모델 중에서 서포트 벡터 머신 (SVM : Support Vector Machine) 모델이 가장 많은 논문 발표되었다.

서포트 벡터 머신은 패턴 인식, 자료 분석을 위한 지도 학습 모델 중 하나이며, 주로 분류와 회귀 분석을 위해 사용한다. 기계학습에서 신경망과 딥 러닝 모델도 많은 논문이 발표되었다. 딥 러닝 모델 기법을 제외하면, 램덤 포레스트 모델, 의사결정나무 모델이 많이 발표되어 기계학습 분야에서는 지도 학습 분야가 많이 발표되고 있음을 알 수 있다. 딥 러닝은 기계학습 모델 중 하나이며, '딥 러닝'을 검색어를 가지고 논문의 키워드를 분석한 결과 딥 러닝의 대표적인 모델인 컨볼루션 신경망(CNN : Convolutional Neural Network)과 순환 신경망(RNN : Recurrent Neural Network) 모델이 가장 많은 논문 발표가 됨을 알 수 있다.

순환신경망의 일종인 LSTM(Long Short Term Memory)도 많이 발표되어 있다. 딥 러닝 분야에서 객체 탐색이나 이미지 분류가 많이 나와 있으며, 복잡성이 심화된 딥 러닝 기술을 보다 쉽게 구축할 수 있는 다양한 프레임워크가 있다. 텐서플로우(TenforFlow), 토치(Torch), 케라스(Keras), 카페(Caffe)등 다양한 프레임워크 중 텐서플로우가 가장 많이 이용하고 있음을 알 수 있다. 물론 텐서플로우는 케라스와 같이 사용될 수 있다. 객체 검출 기법은 R-CNN (Region with Convolutional Neural Network), Faster R-CNN, YOLO(You Only Look Once)등이 있으나, YOLO가 가장 많이 이용하여 논문이 발표됨을 알 수 있다. YOLO는 실시간 객체 탐지 기법으로 객체 탐지 기법에 있어 다른 기법에 비해 속도 빠르다고 알려져 있다.

표 5. 키워드 기반의 단어 빈도 분석
Table 5. Word Count Analysis based Keyword

Rank	'인공지능'	Count	'기계학습'	Count	'딥 러닝'	Count
1	인공지능	1,558	기계 학습	1,032	딥 러닝	865
2	4차 산업혁명	433	인공지능	126	컨볼루션 신경망	177
3	신경망	209	서포트 벡터 머신	119	인공지능	122
4	기계 학습	184	신경망	118	기계 학습	109
5	빅데이터	173	딥 러닝	114	신경망	64
6	딥 러닝	155	텍스트 마이닝	94	순환 신경망	44
7	사물 인터넷	118	빅데이터	65	lstm	35
8	알고리즘	67	분류	44	빅데이터	26
9	텍스트 마이닝	65	컨볼루션 신경망	41	object detection	23
10	로봇	51	자연어 처리	38	이미지 분류	23
11	인공지능 스피커	40	랜덤 포레스트	33	자연어 처리	18
12	드론	37	순환 신경망	27	사물 인터넷	15
13	융합	36	의사결정나무	26	예측	14
14	유전자 알고리즘	35	예측	25	텍스트 마이닝	14
15	컨볼루션 신경망	35	사물 인터넷	22	4차 산업혁명	13
16	챗봇	33	클러스터링	22	텐서플로우	12
17	포스트휴머니즘	32	feature selection	21	분류	12
18	서포트 벡터 머신	28	4차 산업혁명	19	서포트 벡터 머신	11
19	포스트휴먼	28	알고리즘	18	image recognition	11
20	자율주행자동차	27	심층 신경망	17	드론	11
21	개인정보	27	sentiment analysis	16	transfer learning	10
22	블록체인	20	lstm	15	data augmentation	10
23	트랜스휴머니즘	18	로봇	15	segmentation	8
24	순환 신경망	17	강화 학습	11	yolo	7
25	강화 학습	17	기계 학습 알고리즘	11	semantic segmentation	7
26	의사결정나무	17	텐서플로우	10	image processing	7
27	텐서플로우	15	유전자 알고리즘	9	감성분석	7
28	가상현실	15	semi-supervised learning	9	알고리즘	7
29	지능정보사회	15	unsupervised learning	8	융합	7
30	저작권	14	machine translation	8	face recognition	6

3.3 토픽 모델링과 의미 연결망

주어진 문서에 대해 각 문서에 어떤 주제들이 존재하는지를 서술하는 대한 확률적 토픽 모델 기법 중 하나인 LDA(잠재 디리클레 할당) 모델링은 LSA(잠재 의미 분석) 모델링을 개선한 토픽 모델링 알고리즘이다.

본 논문에서 LDA 모델링을 이용하여 토픽 모델링을 수행한다. LDA에서는 문서의 집합으로부터 어떤 토픽이 존재하는지 알아내기 위한 모델링 알고리즘이다. LDA는 토픽의 개수를 미리 정해야 하며, 모든 단어를 미리 정해진 토픽에 중 하나에 할당한다. 그 후에 모든 문서의 모든 단어에 대해서 문서 단어들 중 토픽에 해당하는 단어의 비율과 문서에 특정 토픽이 존재할 확률을 결합 확률로 추정하여 토픽을 추출한다. 본 논문에서는 국내 논문 검색어 인공지능, 기계학습, 딥러닝에서 토픽의 개수를 10로 정했다.

토픽의 개수를 10로 한 토픽 모델링 결과는 [표 6]

과 같다. 학술연구정보서비스에서 검색어 인공지능을 10개로 토픽 모델링한 결과 인공지능 분야의 학술 논문은 연구, 분류, 기술, 시스템 등이 중요한 토픽으로 분류되었다. 검색어 기계학습을 10개로 토픽 모델링한 결과 기계학습 분야의 학술 논문은 지능, 기술, 교육, 시스템, 이용, 예측 등이 중요한 토픽으로 분류되었다. 검색어 딥러닝을 10개로 토픽 모델링한 결과 딥러닝 분야의 학술 논문은 영상, 지능, 러닝, 이용, 시스템 등이 중요한 토픽으로 분류되었다.

텍스트 내 단어의 관계에 적용하여 텍스트의 의미구조를 파악하려는 분석기법이 의미연결망 분석(Semantic Network Analysis) 기법이다. 의미연결망 분석에서는 일정한 범위 내에서 어휘가 동시에 등장하면 이를 서로 연결된 것으로 간주하여 의미연결망이라고 분석한다.

표 6. 인공지능, 기계학습, 딥 러닝 토픽 모델링
Table 6. AI, ML, Deep Learning Topic Modeling

AI Topic Modeling	(0, '0.055*연구' + 0.037*지능' + 0.033*영향' + 0.030*인공") (1, '0.032*분류' + 0.031*기반' + 0.028*교육' + 0.027*활용") (2, '0.027*연구' + 0.019*분석' + 0.015*게임' + 0.014*지능") (3, '0.061*지능' + 0.060*인공' + 0.026*연구' + 0.024*대환") (4, '0.053*기술' + 0.047*지능' + 0.043*인공' + 0.043*연구") (5, '0.038*중심' + 0.026*휴먼' + 0.024*포스트' + 0.020*분석") (6, '0.026*지능' + 0.025*이용' + 0.023*인공' + 0.023*알고리즘") (7, '0.073*시대' + 0.060*지능' + 0.056*인공' + 0.056*산업혁명") (8, '0.105*시스템' + 0.070*이용' + 0.063*인공' + 0.063*지능") (9, '0.031*모형' + 0.030*구조' + 0.030*연구' + 0.023*인공")
ML Topic Modeling	(0, '0.024*기술' + 0.022*연구' + 0.018*탐색' + 0.018*적용") (1, '0.037*교육' + 0.030*학습' + 0.027*기반' + 0.023*이용") (2, '0.037*시스템' + 0.033*기반' + 0.025*문서' + 0.020*이용") (3, '0.061*학습' + 0.058*이용' + 0.042*기계' + 0.025*분류") (4, '0.050*지능' + 0.042*인공' + 0.026*연구' + 0.021*기반") (5, '0.041*이용' + 0.039*예측' + 0.033*신경망' + 0.025*모델") (6, '0.051*기반' + 0.043*이용' + 0.041*기계학습' + 0.034*기법") (7, '0.038*기술' + 0.034*연구' + 0.027*분석' + 0.025*지능") (8, '0.043*이용' + 0.041*기반' + 0.034*영상' + 0.021*알고리즘") (9, '0.040*예측' + 0.037*이용' + 0.032*기계학습' + 0.025*연구")
Deep Learning Topic Modeling	(0, '0.051*영상' + 0.038*이용' + 0.036*기계' + 0.034*번역") (1, '0.092*지능' + 0.091*인공' + 0.069*연구' + 0.030*중심") (2, '0.089*러닝' + 0.050*이용' + 0.041*모델' + 0.035*기반") (3, '0.059*기반' + 0.055*러닝' + 0.052*인식' + 0.024*활용") (4, '0.041*이용' + 0.038*기법' + 0.037*분석' + 0.037*기반") (5, '0.071*분류' + 0.041*연구' + 0.030*기반' + 0.028*러닝") (6, '0.077*이용' + 0.042*신경망' + 0.034*태양광' + 0.033*예측") (7, '0.061*기반' + 0.040*기법' + 0.035*영상' + 0.028*인간") (8, '0.051*시스템' + 0.036*연구' + 0.036*기반' + 0.030*러닝") (9, '0.076*기반' + 0.064*러닝' + 0.038*연구' + 0.028*이용")

[그림 2]과 같이 인공지능의 의미연결망 분석을 보면 인공지능 중심에 연구와 이용이 개발이나 구현보다 가까이 있음을 알 수 있다. [그림 3]을 보면 기계학습의 의미 연결망에서는 연구, 이용, 분류, 기법등이 기계학습에 밀접하게 연결되어 있으며, [그림 4]의 딥 러닝의 의미 연결망을 보면 이용과 예측이 딥러닝에 가까이 연결되어 있음을 알 수 있다.

4. 결론

4차 산업혁명의 대표적인 이미지인 인공지능이 2016년 3월 구글의 알파고 등장으로 인해 인공지능 인식이 매우 높아지고 이로 인해 연구가 활발히 진행 되었다.

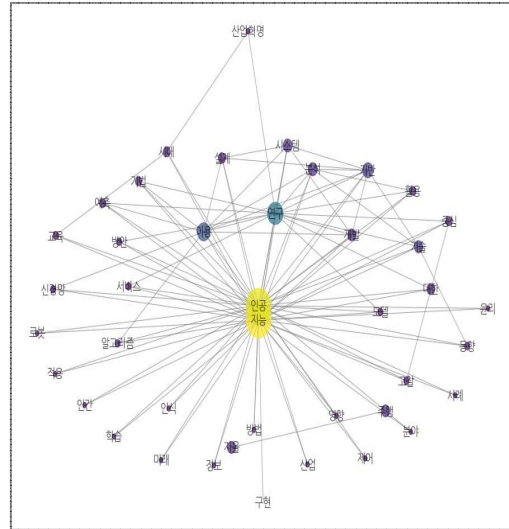


그림 2. 인공지능의 의미 연결망
Fig. 2. Semantic Network Analysis in AI

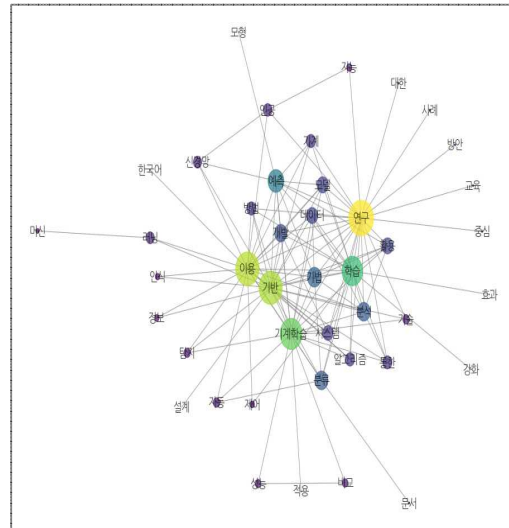


그림 3. 기계학습의 의미 연결망
Fig. 3. Semantic Network Analysis in Machine Learning

- urvey", Multimedia Tools and Applications, Nov. 2018.
- [13] J. J. Sun, D. S. Kim, "A Study on the Development Trend of Artificial Intelligence Using Text Mining Technique: Focused on Open Source Software Projects on Github", Journal of Intelligence and Information Systems, pp. 1-19, Vol. 225, No. 1, Mar. 2019.
- [14] S. Hwang, M. K. Kim, "Analysis of research trends in the field of artificial intelligence in Korea-focusing on topic modeling and semantic network analysis-", Journal of Digital Contents Society, pp. 1847-1855, Vol. 20, No. 9, Sep. 2019.
- [15] J. M. Seok, S. H. Park, "Analysis of major research trends in artificial intelligence through analysis of thesis data", Journal of Digital Convergence, pp. 225-233, Vol. 15, No. 5. Jun. 2018.
- [16] S. M. Rho, "A Study of Topic Modeling Methods", Journal of Digital Contents Society, pp. 423-428, Vol. 18, No. 2, Feb. 2017.
- [17] K. S. Kang, S. M. Min, "Keyword Analysis of KCI Journals on Business Administration using Web Crawling and Machine Learning", Korea Journal of Business Administration, pp. 597-615, Vol. 32, No. 4, Apr. 2019.

저자약력

박 홍 진(Hong-Jin Park)

[중신회원]



〈관심분야〉

- 1995년 8월 : 중앙대학교 일반대학교 컴퓨터공학과(공학석사)
- 2001년 8월 : 중앙대학교 일반대학교 컴퓨터공학과(공학박사)
- 2001년 9월 ~ 현재 : 상지대학교 컴퓨터공학과 교수

시스템, 지능형 서버 관리, 빅데이터 분석