Journal of the Korea Convergence Society
Vol. 11. No. 7, pp. 9-18, 2020

ISSN 2233-4890 / eISSN 2713-6353
https://doi.org/10.15207/JKCS.2020.11.7.009

LDA 토픽모델링을 통한 ICT분야 국가연구개발사업의 주요 연구토픽 및 동향 탐색

우창우[1,3], 이종연[2*]
[1]충북대학교 컴퓨터과학과 박사수료, [2]충북대학교 소프트웨어학과 교수, [3]정보통신기획평가원 SW클라우드기획팀 책임

Investigation of Research Topic and Trends of National ICT Research-Development Using the LDA Model

Chang Woo Woo[1,3], Jong Yun Lee[2*]
[1]Ph. D. Candidate, Department of Computer Science, Chungbuk National University
[2]Professor, Department of Computer Science, Chungbuk National University
[3]Manager, SW & Cloud Planning Team, Institute of Information & Communications Technology Planning & Evaluation

요 약 본 논문의 연구목표는 LDA(Latent Dirichlet Allocation) 모델을 적용하여 국가연구개발사업을 통해 수행되고 있는 ICT(Information and Communication Technology) 분야의 연구과제에 대한 주요 연구 토픽과 동향을 탐색하는데 있다. 연구방법에는 NTIS(National Science and Technology Information Service)로부터 최근 5년간 국가연구개발사업의 전체 연구과제 정보를 다운로드받고 이를 정보통신기획평가원(IITP)의 EZone 시스템과 매칭하여 ICT분야 연구과제 5,200건을 확보하고, 토픽모델링 기법중 하나인 LDA 모델을 적용하여 연구토픽과 연구동향을 조사하였다. 실험결과로, ICT분야 연구과제에 대한 연구토픽은 인공지능, 빅데이터, 사물인터넷(Internet of Things)과 같은 지능정보기술로 확인되었고 연구동향에는 초실감미디어에 관한 연구가 활발히 진행되고 있음을 확인하였다. 끝으로 본 논문에서 진행된 국가연구개발사업에 대한 토픽모델링 결과는 향후 ICT분야 연구개발 계획 및 전략수립, 정책, 과제기획 등 중요한 정보로 활용될 수 있을 것이다.

주제어 : LDA(Latent Dirichlet Allocation), 토픽모델링, 연구토픽, 연구동향, ICT분야 국가연구개발사업

Abstract The research objectives investigates main research topics and trends in the information and communication technology(ICT) field, Korea using LDA(Latent Dirichlet Allocation), one of the topic modeling techniques. The experimental dataset of ICT research and development(R&D) project of 5,200 was acquired through matching with the EZone system of IITP after downloading R&D project dataset from NTIS(National Science and Technology Information Service) during recent five years. Consequently, our finding was that the majority research topics were found as intelligent information technologies such as AI, big data, and IoT, and the main research trends was hyper realistic media. Finally, it is expected that the research results of topic modeling on the national R&D foundation dataset become the powerful information about establishment of planning and strategy of future's research and development in the ICT field.

Key Words : LDA(Latent Dirichlet Allocation), Topic Modeling, Research Topic, Research Trends, National ICT R&D

*Corresponding Author : Jong Yun Lee(jongyun@chungbuk.ac.kr)
Received May 7, 2020 Revised June 19, 2020
Accepted July 20, 2020 Published July 28, 2020

1. 서론

2016년 다보스에서 개최된 세계경제포럼에서 클라우스 슈밥 회장은 '제4차산업혁명' 키워드를 언급하며 "제4차산업혁명 시대가 도래하면 우리가 '하는 일'이 아닌 '우리'가 변화될 것"이라고 연설하였고[1], 이후 각국 정부는 제4차산업혁명을 준비하기 위해 다양한 전략과 정책을 발표하며 인공지능 · 빅데이터 · 초연결 등 혁신을 만드는 지능정보기술에 전폭적인 R&D(Research and Development) 예산을 투자하겠다고 하였다[2]. 국내에서는 2017년 11월 '대통령 직속-4차 산업혁명위원회'를 출범 향후 5년간 지능정보기술에 국비 2.2조원을 투자해 제4차산업혁명에 대응하는 국가 미래 기술경쟁력을 확보하겠다고 하였고[3], 이듬해 1월 과학기술정보통신부는 'I-Korea 4.0 : ICT R&D 혁신전략'을 발표하며 기술개발 실패의 가능성이 크고 도전적이지만 성공할 경우 파급효과가 높은 고위험 · 도전형 R&D에 적극적으로 투자하겠다고 발표하였다[4].

정부에서는 매년 국가 성장 잠재력 확보와 핵심원천기술 개발을 위해 고위험 · 도전형 R&D를 추진하고 있지만, ICT분야는 보건의료 · 우주 · 항공 등 다른 분야와 다르게 타 분야의 인프라 성격을 갖는 기반기술로 융합되어 있어 기존기술의 경쟁력 향상을 위한 추격형 R&D도 계속 추진되어야 한다. 특히, 전 세계적으로 폭발적인 R&D가 증가한 지능정보기술(인공지능 · 빅데이터 · 초연결)의 경우 국내 기술 수준을 고려한다면 고위험 · 도전형 R&D뿐만 아니라 기존 기술의 간극을 좁히기 위해 추격형 R&D도 지속되어야 한다. 특히 국내 ICT분야의 전체 기술격차는 미국을 기준으로 1.4년의 기술격차가 있고, 인공정보기술의 경우 미국 대비 인공지능 2.0년의 격차가 존재한다[5].

제4차 산업혁명에 따른 지능정보기술의 수요 증가와 고위험 · 도전형 및 추격형 R&D를 함께 해야 하는 ICT분야의 기술적인 특성, 마지막으로 지능정보기술을 이루는 요소기술에 대한 기술수준을 함께 본다면 인공지능 · 빅데이터 · 사물인터넷은 ICT분야의 다른 기술들보다 더 중점적으로 투자되어야 할 필요가 있다. 특히 토픽모델링 기법중 하나인 LDA(Latent Dirichlet Allocation) 모델을 적용해서 연구토픽과 동향을 탐색하는 연구는 과거에도 있었지만 대부분 논문데이터(제목, 초록, 키워드 등) 또는 특허데이터(특허등록 및 출원) 등을 중심으로 주요 토픽을 도출하였다[8-12]. 하지만 지금까지의 토픽모델링에 대한 연구동향 조사 연구는 주로 논문데이터, 특허데이터를 활용한 주요 토픽과 연구동향을 탐색에 중점을 두었고, 국가연구개발사업 데이터를 활용한 ICT분야의 연구개발사업 주요 연구토픽 및 연구동향을 탐색한 사례는 거의 찾을 수 없다. 또한, 국가연구개발사업은 주로 기업 · 대학 · 연구소 등 다양한 곳에서 참여하고 있고 사업의 결과물로 다수의 논문 · 특허 · 기술이전 등이 나타나고 있다. 더불어 기초과학이 아닌 응용과학의 경우 기술개발이 선행된 후 논문으로 정리 · 발표되기 때문에 연구자들의 주요토픽과 동향을 탐색하기 위해서는 논문이나 특허 데이터 분석보다는 좀 더 빠른 방법의 연구동향 파악이 요구된다.

따라서 본 논문의 연구목표는 국내외적으로 빠르게 변화하는 ICT 흐름을 파악하기 위해 LDA 토픽모델링 기법을 적용하여 국내 연구자들이 수행하고 있는 ICT분야 국가연구개발사업 과제정보에 대한 주요 연구토픽과 연구동향을 탐색하는데 있다. 아울러 세부적인 연구방법은 다음과 같다. 첫째, 실험 데이터로 최근 5년간(2015~2019) 국가연구개발사업을 통해 수행된 과제 중 ICT분야로 분류된 과제정보를 수집하였다. 이를 위해 한국과학기술정보연구원에서 운영 중인 NTIS(National Science and Technology Information Service) 시스템[6]에서 국가연구개발사업 전체 데이터를 우선 수집하였고, ICT분야만 추출하기 위해 과기정통부의 ICT분야 전담기관인 정보통신기획평가원의 EZone 시스템[7]과 매칭하여 ICT분야에 대한 과제데이터를 확보하였다. 둘째, 확보된 ICT분야의 데이터에 대해 특수문자 제거, 약어 풀이 등 분석을 위한 데이터 전처리를 진행하였다. 셋째, 토픽모델링 기법중 하나인 LDA 모델을 적용하여 분석 결과를 연도별 주요토픽-키워드와 연도별 키워드 히트맵으로 표현해 ICT분야 국가연구개발사업으로 투자 · 연구되고 있는 과제정보에 대한 주요 연구토픽과 연구동향을 분석하였다. 끝으로 본 논문의 연구결과는 향후 ICT 분야 전략기획 수립 · 정책기획 · 과제기획 등 연구개발 기획 과정에 사용될 수 있을 것으로 기대한다.

본 논문의 구성은 다음과 같다. 먼저 2장은 기존의 논문정보 및 특허정보를 통해 토픽모델링을 수행한 연구사례를 분석한다. 그리고 3장은 토픽모델링의 기법중 하나인 LDA 모델을 요약하고, 이를 적용하여 ICT분야 국가연구개발사업 데이터를 분석하기 위한 연구방법과 실험 데이터를 기술한다. 그리고 4장에서는 실험결과를 분석하고, 끝으로 5장은 본 연구가 주는 시사점과 향후 연구내용을 요약한다.

2. 관련연구

2.1 ICT의 국내외 수준

국내 ICT분야의 전체 기술격차는 세계 최고 기술수준 보유국인 미국을 기준으로 유럽 0.7년, 일본 1.1년, 중국 1.2년 순으로 미국에 비해 1.4년의 기술격차가 존재하고, 지능정보기술의 경우 미국 대비 인공지능 2.0년, 빅데이터 1.9년, 사물인터넷 1.2년으로 타 국가 대비 가장 많은 격차가 벌어져있다[5](Table 1 참고).

Table 1. Comparison of technological development level gap between USA and other countries

Field	Relative technological development level(Year)				
	S.Korea	U.S	Japan	China	EU
ICT	1.4	0.0	1.1	1.2	0.7
AI	2.0	0.0	1.8	1.5	1.4
Bigdata	1.9	0.0	1.4	1.1	0.8
IoT	1.2	0.0	0.9	1.0	0.5

2.2 연구 사례

김태경(2016)은 핀테크 기술의 동향 분석을 위해 특허데이터를 활용하여 1990년 1월~2016년 7월 까지 출원된 미국 · 한국 · 중국 특허 4,681건을 수집하여 핀테크 산업의 주요 기술에 대한 개발 동향을 각각 분석하였고, 분석 결과 미국은 모바일 결제, 금융 데이터 분석, 온라인 상거래, 한국은 인증/보안, 모바일결제, 중국은 인증/보안, NFC(Near Field Communication) 기술 등을 연구동향으로 도출하였다[8].

김창식(2017)은 2002년~2016년 정보시스템분야 연구동향 분석을 위해 APJIS(Asia Pacific Journal of Information Systems), ISR(Information Systems Review, JIS(The Journal of Information Systems) 저널에 발표된 1,245편의 논문초록을 대상으로 토픽모델링과 시계열회귀분석 기법을 활용하여 '시스템 구축', '혁신 역량', '고객 충성도' 등을 주요 연구토픽으로 도출하였다[9].

박주섭(2018)은 연구동향과 과학기술의 동향을 파악하고자 키워드 네트워크 분석을 통해 미국의 2002년~2016년 특허중 'Artificial Intelligence(AI)' 특허의 초록 13,618개를 대상으로 키워드 네트워크 분석을 활용하여 데이터를 분석, 시간이 지날수록 AI 응용 분야의 방법에 관련된 핵심어들이 부각되었음을 확인하였다[10].

김용환(2019)은 DBpia 학술DB 내 다수의 국내저널에 분포되어있는 헬스케어 관련 논문데이터 수집을 통해 토픽모델링을 적용하여 국내 헬스케어 연구를 '국가정책', '보건의료 및 체육', '연구디자인', 'ICT 기술', '생애주기', '환경', '여성' 등 40개의 토픽으로 분석하였다[11].

조혜인(2019)은 블록체인의 학술연구 동향 분석을 위해 Web of Science 데이터베이스에 누적된 논문초록을 활용하여 '블록체인(Blockchain)' 및 유사 키워드 '비트코인(Bitcoin)', '암호화폐(Cryptocurrency)' 등이 들어간 키워드를 중심으로 699건의 학술논문을 수집하여 미국은 '경제 · 금융', 중국은 '기술', 한국은 '정책 · 규제' 등의 토픽을 도출함으로써 국가별 연구동향의 차이를 확인하였다[12].

3. 연구방법 및 실험데이터

3.1 데이터 수집

분석을 위해 한국정보기술연구원에서 운영하고 있는 NTIS 시스템에서[6] 최근 5년간 국가연구개발사업으로 수행된 과제 데이터 수십만 건을 다운로드 받았고, 해당 데이터의 과제제목 및 과제번호 등의 속성을 활용하여 정보통신기획평가원의 EZone 시스템[7] 매칭을 통해 ICT분야로 연구된 국가연구개발사업 과제 데이터 약 5,200건을 추출하였다. NTIS에서 제공되는 데이터 속성은 '부처명', '사업기간', '사업예산', '과제명' 등으로 관련 규정에 따라 총 164개의 속성을 제공하고 있으며[13], 불필요한 속성을 제외한 후 '과제연도', '계속과제여부', '영문키워드' 3개 속성과, 분석결과의 해석을 위해 '과제명', '국문키워드', '최종목표', '주요내용', '기대효과' 등 5개 속성을 추가로 선택하였다(Fig. 1 참고).

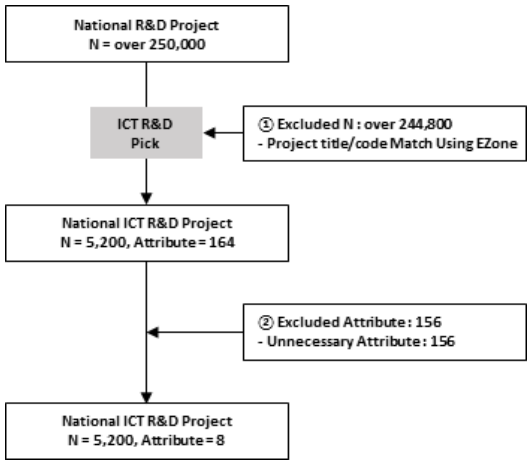
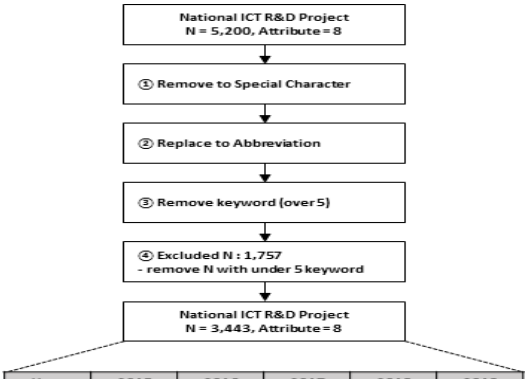


Fig. 1. Data extraction process

3.2 데이터 전처리

데이터 전처리는 입력데이터로 사용될 영문키워드 속성을 중심으로 진행하였다. 첫째, 데이터에 포함되어 있는 특수문자 및 영어가 아닌 한글과 기호 등을 제거하였다. 둘째, 'IoT', 'AR' 등 일반적으로 쓰이는 약어는 'Internet of Things', 'Augmented Reality' 등 동일한 단어로 표준형 변환을 수행하였다. 셋째, 'SW', 'ICT', 'Development' 등 빈번하게 등장하거나 분석에 영향을 미치지 않는다고 판단되는 키워드 'Manpower', 'Seoul Accord', 'University' 등은 불용어로 처리하였다. 넷째, 과제별 키워드가 5개 미만인 경우 과제의 내용을 충분히 설명하기 어렵다고 판단하여 데이터 자체를 삭제하였으며, 과제별 키워드가 5개를 초과하는 경우 후순위 키워드일수록 우선순위가 낮다고 판단하여 초과되는 키워드는 모두 삭제하였다. 위의 과정을 통해 최종 3,443건의 과제정보 데이터를 확보하였으며, 총 17,215건의 키워드를 기반으로 분석을 진행하였다(Fig. 2 참고).



Year	2015	2016	2017	2018	2019
N	352	624	793	817	857
Attribute	8	8	8	8	8
keyword	1,760	3,120	3,965	4,085	4,285

Fig. 2. Data preprocessing steps for experimental data

3.3 LDA 분석 모델

데이터 분석은 토픽모델링 방법 중 대표적으로 활용되는 LDA(Latent Dirichlet Allocation) 모델을 사용하였다[14]. 토픽모델링이란 문서 집합의 추상적인 '토픽'을 발견하기 위한 텍스트마이닝 기법으로 하나의 문서가 특정 단어들의 집합이라고 가정하고 문서에 나오는 단어를 확률적으로 계산하여 본문의 숨겨진 의미구조를 발견하는 통계적 모델 중 하나이다. 1990년 Deerwester et al.[15]을 통해 등장한 최초의 토픽모델인 LSI(Latent Semantic Indexing)는 문서-단어 행렬을 문서-토픽 행렬과 토픽-단어 행렬로 분해하는 과정을 기술하였고, 2001년 Hofmann et al.[16]은 LSI를 기반으로 단어의 출현 빈도를 확률로 대체하는 모형인 pLSI(probabilistic Latent Semantic Indexing) 모델을 기술하였다. pLSI 모델은 확률을 적용하지만 생성된 모델에 새로운 문서가 입력될 경우 기존 모델을 적용할 수 없는 오버피팅(Overfitting) 문제가 존재하였고, 이후 2003년 Blei et al.은 pLSI의 불완전한 확률모델을 보완해 LDA 모델을 발표하면서 현재의 토픽모델링 분야를 정착시켰다.

LDA 모델은 문서나 단어 등 관찰된 변수를 통해 문서 구조처럼 보이지 않는 변수를 추론하는 것을 목적으로 하며 이를 통해 전체 문서 집합의 토픽, 각 문서별 토픽 비율, 각 토픽에 포함된 단어의 분포 등을 도출 할 수 있다. LDA 모델의 데이터 처리과정은 Fig. 3과 같다. 사전에 문서-토픽별 분포값(θ)과 토픽의 단어 분포값(β) 및 토픽의 개수(k)를 하이퍼 파라미터(Hyper parameter)로 미리 입력하여야 하며 입력된 값을 통해 단어(w)를 관측해 단어마다 적절한 토픽번호(z)를 정해준다. 그리고 모델을 반복하며 θ, β 값을 갱신, 모든 z값 중 가장 높은 z값을 찾아 문서에 있는 각각의 단어들이 어디에 속해야 하는지 추론한다.

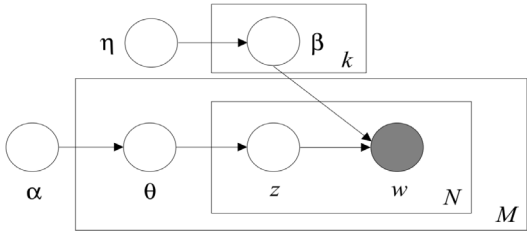


Fig. 3. Graphical model representation of the smoothed LDA model

- M : 문서의 개수
- N : 문서에 속한 단어의 개수
- W : 단어
- Z : 해당 단어가 속한 토픽번호
- k : 토픽의 개수 (Hyper parameter)
- α : 문서-토픽별 θ 분포값 (Hyper parameter)
- η : 토픽-단어별 β 분포값 (Hyper parameter)
- θ : 문서별 토픽의 분포
- β : 토픽의 단어 분포

3.4 실험환경 및 분석 프로세스

앞서 전처리가 완료된 3,443건의 ICT분야 국가연구개발사업 과제데이터를 각각 연도별로 LDA 모델에 적용하여 1,000번의 학습을 통해 주요 연구토픽 10개를 도출하였고, 각 연구토픽에 도출된 키워드를 연도별로 재배치하여 분포값을 기준으로 연구동향을 탐색하였다. 실험은 Python 3.7.7 버전을 사용하였고, LDA 모델에 대한 하이퍼 파라미터 값은 k=10, α=0.1, n=0.01로 진행하였다(Fig. 4 참고).

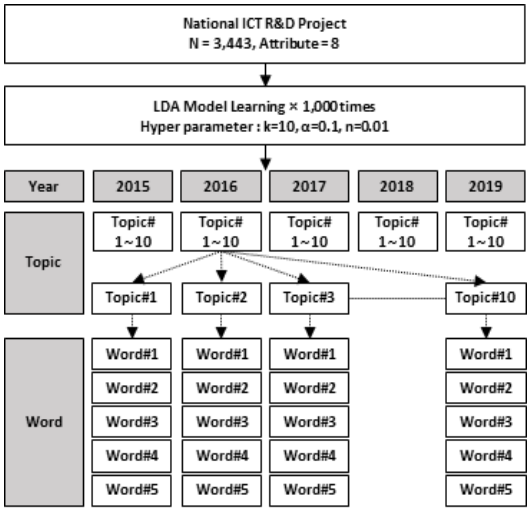


Fig. 4. Topic modeling processes using LDA

4. 실험결과 및 토의

4.1 연구토픽 탐색결과

실험 결과, 2015년~2019년 국가연구개발사업을 통해 수행된 ICT분야 연구과제 대부분은 제4차산업혁명의 핵심 동인으로 언급되는 지능정보기술인 인공지능, 빅데이터, 사물인터넷 관련 연구토픽으로 확인되었다. 연도별 토픽모델링 결과표에 지능정보기술에 해당하는 'artificial intelligence(AI)', 'big data/big data analysis', 'internet of things(IoT)'의 키워드에 밑줄과 이탤릭체로 표기하였지만 실제로는 인공지능(AI)의 범위를 'machine learning', 'deep learning', 그리고 사물인터넷(IoT) 범위를 'cloud'처럼 연관성 높은 키워드와 함께 본다면 대부분의 연구토픽이 지능정보기술을 기반으로 이루어진다고 해석할 수 있다.

과거 ICT분야에 등장한 몇 가지 중요 이슈를 토픽과 연관시켜 보면 2016년 3월 구글에서 개발한 인공지능 프로그램(AlphaGo)과 바둑프로기사 이세돌의 챌린지 매치이다[17]. 당시 인공지능과 세계 최고의 바둑프로기사 대결로 주목을 받으며 진행된 경기는 4대 1로 알파고가 승리하며 전 세계에 엄청난 충격을 안겨주었다. 향후 '알파고 쇼크'라고 불리게 된 이 사건은 지금도 우리가 지능정보기술에 왜 투자해야 하는지 대중들에게 명확하게 설명할 수 있는 대표적인 사례로 사용되고 있다. 알파고 쇼크 이후 인공지능과 관련된 키워드인 'artificial intelligence', 'machine learning', 'deep learning', 'big data' 등은 2017년 10개 토픽 중 5개, 2018년 7개, 2019년 7개로 2015년 4개, 2016년 3개에 비해 약 2배 이상 증가하였다.

또 다른 이슈는 2017년 하반기부터 등장한 비트코인이다. 비트코인은 블록체인을 기반으로 만들어진 온라인 암호화폐로[18] 금융시장에 엄청난 이슈를 만들어내며 기존 법률의 개정과 특별법 발의까지 이끌어냈다[19]. 비트코인의 핵심기술로 알려져 있는 'block chain' 과 'smart contract'는 2017년 과제기획을 통해 2018년 토픽#9에 등장하였지만, 기대와는 다르게 큰 파동을 만들어내지 못하고 2019년 토픽#2(자율주행차 관련), 토픽#9(스마트팩토리 관련) 등 다른 토픽의 인프라 성격을 갖는 요소기술로 자리매김하였다.

실질적으로 연도별 토픽 모델링 결과는 다음과 같다. 첫째, 2015년도에는 10개의 주요토픽 중 총 6개(#2, #3, #5, #6, #7, #10의 토픽에서 지능정보기술 키워드가 등장하였고, #4에서 등장한 platform, smart contents, gamification 키워드와 관련된 연구토픽은 세부 과제정보 확인결과 주로 생산성 향상을 위한 게임화 서비스 플랫폼 연구로 확인되었다(Table 2 참고).

Table 2. Topic modeling results of ICT R&D in 2015

Topic#	Keyword of 2015
#1	security, emotional closed caption, uav, lidar, management
#2	cloud, *big data*, curation, sns, information security
#3	autonomous vehicle, *big data analysis*, html5, interconnection of internet network, input device
#4	platform, database, social network, smart contents, gamification
#5	*internet of things*, *big data*, cloud computing, access control, platform
#6	machine learning, image processing, *artificial intelligence*, network, auction
#7	*internet of things*, wearable, security assessment, software platform, embedded software
#8	virtual reality, monitoring, 3d, augmented reality, dashboard
#9	optical transceiver, uhd, quantum cryptography, wearable, distributed streaming
#10	*internet of things*, sdn, security, 3d printing, nfv

둘째, 2016년도에는 10개의 주요토픽 중 총 3개(#1, #3, #8)의 토픽에서 지능정보기술 키워드가 등장하였고, #2에서 등장한 cloud computing, cloud, digital signage 키워드와 관련된 연구토픽은 세부 과제정보 확인결과 주로 디지털 사이니지와 관련된 온라인/클라우드 지능형 미디어 구축으로 확인되었다(Table 3 참고).

Table 3. Topic modeling results of ICT R&D in 2016

Topic#	Keyword of 2016
#1	machine learning, nfv, deep learning, bluetooth, *big data analysis*
#2	cloud computing, cloud, safety, digital signage, education
#3	wearable device, *internet of things,* visualization, gateway, smart band
#4	auction, multimedia, quantum cryptography, html5, wireless power transfer
#5	soc, wearable, smart device, image processing, smart toy
#6	uhd, optical transceiver, 3d printing, unmanned aerial vehicle, lidar
#7	malware, black box, deep learning, health questionnaire, health enhancement
#8	*internet of things,* big data, platform, cloud, fintech
#9	virtual reality, augmented reality, 3d, simulator, sdn
#10	software, security, autonomous vehicle, management, verification

셋째, 2017년도에는 10개의 주요토픽 중 총 5개(#1, #3, #4, #5, #8)의 토픽에서 지능정보기술 키워드가 등장하였고, #6에서 등장한 privacy, digital signature, function encryption, ransomware 키워드와 관련된 연구토픽은 세부 과제정보 확인결과 주로 보안 및 암호화에 관련된 연구로 확인되었다(Table 4 참고).

Table 4. Topic modeling results of ICT R&D in 2017

Topic#	Keyword of 2017
#1	3d printing, application, cloud, 4th industrial revolution, access control
#2	platform, machine learning, *internet of things,* cloud, *artificial intelligence*
#3	wireless power transfer, 5g, *artificial intelligence,* standardization, ps-lte
#4	software, cloud platform, *big data analysis,* security, deep learning
#5	*internet of things,* block chain, information security, drone, smart factory
#6	privacy, digital signature, functional encryption, reservation, ransomware
#7	uhd, satellite communication, self-driving car, social media, distributed streaming
#8	cloud computing, millimeter wave, *artificial intelligence,* beamforming, image processing
#9	virtual reality, augmented reality, 360 immersive video, mixed reality, algorithm
#10	smart device, deep learning, wearable device, optical transceiver, machine learning

넷째, 2018년도에는 10개의 주요토픽 중 총 7개(#1, #2, #5, #6, #7, #9, #10)의 토픽에서 지능정보기술 키워드가 등장하였고, #4에서 등장한 object detection, smart car, connected car 키워드와 관련된 연구토픽은 세부 과제정보 확인결과 주로 영상정보 분석을 위한 객체탐지와 관련된 연구로 확인되었다(Table 5 참고).

Table 5. Topic modeling results of ICT R&D in 2018

Topic#	Keyword of 2018
#1	lpwa, wearable device, ict convergence, deep learning, *internet of things*
#2	security, cloud, *internet of things,* 3d printing, context awareness
#3	5g, augmented reality, ransomware, 10gbps, mobility
#4	deep learning, object detection, fintech, smart car, connected car
#5	virtual reality, augmented reality, mixed reality, *internet of things, big data*
#6	*internet of things, big data, artificial intelligence,* wireless power transfer, quantum cryptography
#7	*big data,* power amplifier, uhd, platform, cloud
#8	optical transceiver, cloud computing, cloud platform, optical transmission, terahertz
#9	block chain, privacy, smart contract, *big data analysis,* platform
#10	360 immersive video, *artificial intelligence,* machine learning, drone, smart factory

다섯째, 2019년도에는 10개의 주요토픽 중 총 6개(#1, #2, #4, #5, #6, #9)의 토픽에서 지능정보기술 키워드가 등장하였고, #10에서 등장한 wireless power transfer, single frequency network, massive mino 키워드와 관련된 연구토픽은 세부 과제정보 확인결과 주로 안테나, 전파위성에 관한 연구로 확인되었다(Table 6 참고).

Table 6. Topic modeling results of ICT R&D in 2019

Topic#	Keyword of 2019
#1	augmented reality, *internet of things,* machine learning, *big data, artificial intelligence*
#2	block chain, platform, cloud platform, *big data,* connected car
#3	360 immersive video, video coding, social media, deep learning, standardization
#4	deep learning, object detection, *artificial intelligence,* autonomous vehicle, machine learning
#5	virtual reality, augmented reality, beamforming, *internet of things, artificial intelligence*
#6	*artificial intelligence, big data analysis,* natural language processing, machine learning, deep learning
#7	millimeter wave, smart city, 5g, digital signage, mobile communication
#8	cloud, cloud computing, cmos, drone, security
#9	smart factory, block chain, *artificial intelligence,* 4th industrial revolution, *big data*
#10	wireless power transfer, single frequency network, massive mimo, artificial neural network, iot security

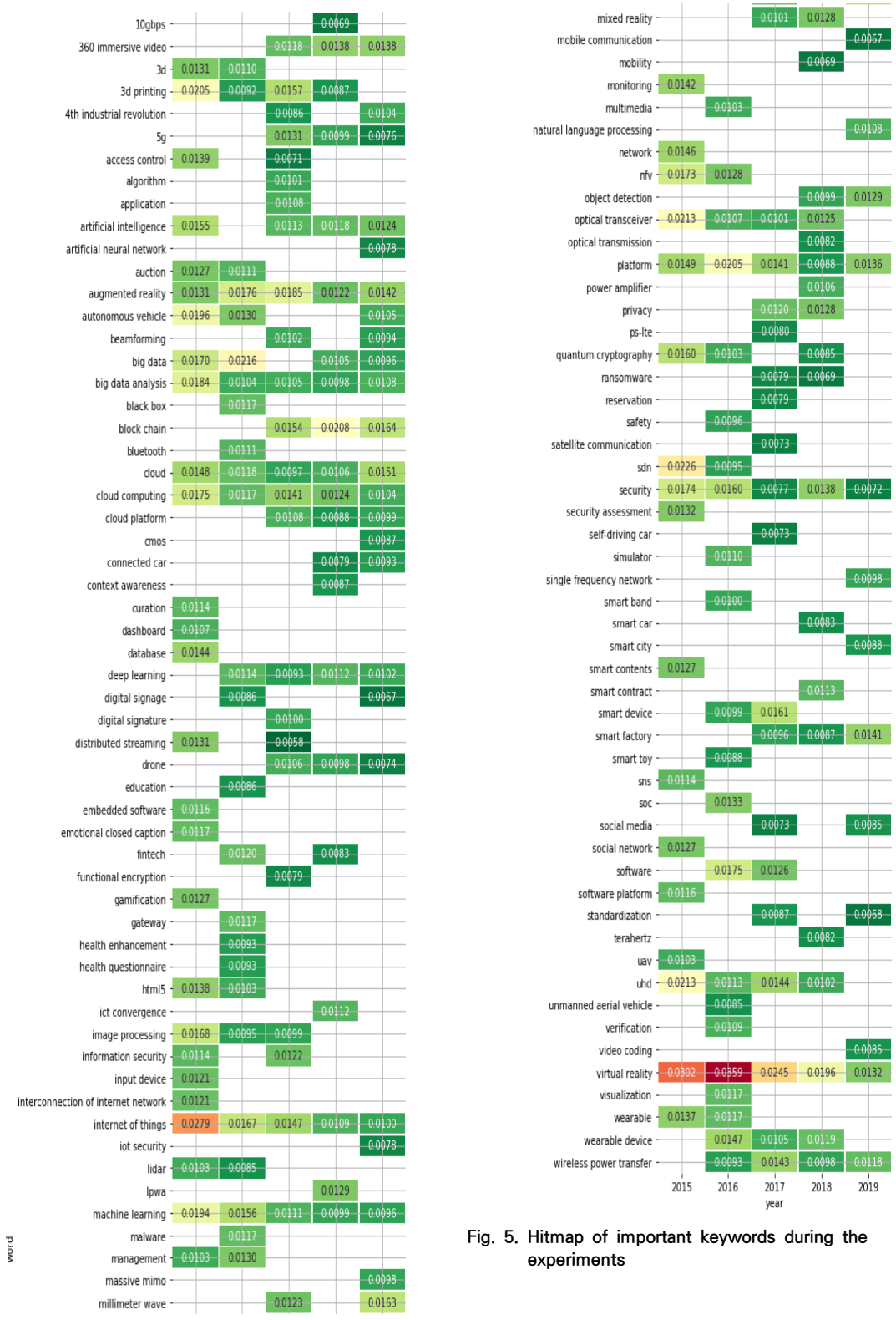


Fig. 5. Hitmap of important keywords during the experiments

4.2 연구동향 탐색결과

연구동향 탐색은 위 실험을 통해 얻은 연도별-키워드 확률값을 히트맵(Hitmap)으로 표현하였다(Fig. 5 참고). 확률값이 높게 표기된 키워드는 해당 연도에 연구가 활발했다는 의미로 볼 수 있고, 매년 키워드가 등장한 경우는 연구의 연속성이 높은 분야로 해석 할 수 있다. 따라서 키워드의 확률값과 연속성을 함께 본다면 연구자들이 수행하고 있는 연구동향의 파악이 가능하다. 위 결과에 따라 연구의 연속성과 확률값이 가장 높은 분야는 Virtual Reality로 확인되었다(Table 7 참고). Virtual Reality 키워드는 Augmented Reality, Mixed Reality, 360 Immersive Video와 함께 초실감미디어 분야의 대표적인 키워드라고 할 수 있다.

Table 7. Key importance and high probability in topic modeling results between 2015 and 2019

Year	Keyword#1	Keyword#2	Keyword#3
2015	virtual reality (0.0302)	internet of things (0.0279)	sdn (0.0226)
2016	virtual reality (0.0359)	big data (0.0216)	platform (0.0205)
2017	virtual reality (0.0245)	augmented reality (0.0185)	smart device (0.0161)
2018	block chain (0.0208)	virtual reality (0.0196)	360 immersive video (0.0138)
2019	block chain (0.0164)	millimeter wave (0.0163)	cloud (0.0151)

※ A real number in the parenthesis denotes the probability of key importance

4.3 토의

실험결과 Table. 2~6 같이 국내 연구자들이 수행하는 주요 연구토픽으로 'artificial intelligence', 'big data', 'internet of things' 등 지능정보기술에 대한 고도화 및 응용연구를 수행하고 있음을 확인하였다. 이를 통해 자연스럽게 ICT분야가 갖는 특성을 반영하여 고위험 · 도전형 R&D와 함께 추격형 R&D를 수행하고 있음을 확인할 수 있었고, Table. 7에 표현된 히트맵을 통해 연도별로 등장하는 'virtual reality(Avg. 0.0247)' 키워드를 도출함으로써 연관 키워드인 'augmented reality(Avg. 0.01512)', '360 immersive video(Avg. 0.0131)' 등 초실감미디어에 관한 연구동향을 탐색할 수 있었다. 지능정보기술과 초실감미디어 관련 키워드 이외에 Fig. 5를 통해 표현한 키워드를 다른 키워드와 함께 해석하면, 3D printing 분야는 'wearable(Avg. 0.0127)', 'wearable device (Avg. 0.0124)'와 함께 연구되고 있으며, 자율주행차는 'autonomous vehicle(Avg. 0.0078)', 'connected car(Avg. 0.0086)', 'self-driving car(0.0073)', 'smart car(0.0083)', 'object detection(Avg. 0.0114)' 키워드를 통해 많은 연구가 이루어지고 있음을 확인하였다. 또한, 무선통신 및 안테나에 관한 키워드로 '5G(Avg. 0.0102)', 'beamforming(Avg. 0.0098)', 'lidar(Avg. 0.0094)', 'millimeter wave(Avg. 0.0143)', 'massive mimo(0.0098)', 'terahertz(0.0082)' 키워드를 통해 연구를 진행하고 있었고, 이 외에도 'drone(Avg. 0.0093)', 'fintech(Avg. 0.0101)', 'malware/ransomware (0.0117/Avg. 0.0074)', 'smart factory(Avg. 0.0108)' 등 ICT분야 국가연구개발사업을 통해 다양한 연구가 지속되고 있음을 확인할 수 있었다.

5. 결론

본 논문은 ICT분야 국가연구개발사업 과제정보 데이터 분석을 통해 주요토픽의 결과로 지능정보기술(인공지능 · 빅데이터 · 사물인터넷)에 관한 연구가 주로 이루어지고 있으며, 구체적인 연구동향으로 초실감 미디어 분야가 활발히 연구되고 있음을 확인하였다. 본 연구가 주는 시사점은 다음과 같다. 첫째, 제4차 산업혁명 이후 촉발된 지능정보기술에 대한 소리 없는 전쟁에 대해 국내에서도 꾸준한 R&D를 통해 참여하고 있다는 것을 확인할 수 있었다. 둘째, 논문 · 특허 · SNS 등의 데이터 분석을 통해 연구토픽을 추론하는 것보다 국가연구개발사업 과제정보 데이터 분석을 통해 연구토픽을 추론하는 것이 더 빠른 연구트렌드를 파악할 수 있음을 확인하였다.

본 연구를 통해 도출한 연구주제 및 연구동향의 결과는 향후 ICT분야 연구개발 계획 및 전략수립, 정책 및 과제기획 등 다양한 과정에 중요한 정보로 활용될 수 있을 것으로 기대하며, 해당 연구의 한계점은 LDA 토픽모델링이 비지도학습의 기계학습 모델로 알고리즘을 통해 주제를 반환하지만, 분석에 사용되는 데이터 키워드에 대한 각각의 의미를 기계가 정확히 인식할 수 없음과 동시에 현재의 알고리즘이 문서 내에 등장한 키워드 확률을 기반으로 추론하기 때문에 해당 토픽에 나타난 키워드가 사람이 인지하는 의미와 정확할 수 없다는 것을 밝힌다. 따라서 이러한 문제점을 개선하기 위해 토픽과 토픽간의 관계나 다른 속성들을 배경정보로 활용하여 사람이 인지하는 의미에 더 가까운 결론을 도출하는 연구가 진행될

필요가 있으며, 더 정확한 연구동향을 파악하기 위해 시간의 흐름에 따라 연구주제 및 동향의 변화를 파악하기 위해 시계열 분석이 가능하도록 시간에 관한 속성을 추가하거나 최근의 시간에 가중치를 더 부여하는 등 기존 모델을 개선하는 연구가 진행될 필요가 있다.

REFERENCES

[1] Klaus Schwab. (2016). *The Fourth Industrial Revolution: what it means, hot to respond.* World Economic Forum Agenda. World Economic Forum.

[2] Presidential Committee on the Fourth Industrial Revolution. (2016). *Comprehensive Measures for Intelligence Information Society.* Seoul : Presidential Committee on the Fourth Industrial Revolution.

[3] Presidential Committee on the Fourth Industrial Revolution. (2017). *4th Industrial Revolution Response Plan.* Seoul : Presidential Committee on the Fourth Industrial Revolution.

[4] Ministry of Science and ICT. (2018). *I-KOREA 4.0 : ICT R&D Innovation Strategy.* Sejong : Ministry of Science and ICT Publishing.

[5] Institute of Information & Communications Technology Planning & Evaluation. (2019). *2018 ICT Technical Level Survey.* Daejeon : Institute of Information & Communications Technology Planning & Evaluation.

[6] NTIS. *National R&D Management System.* https://www.ntis.go.kr

[7] EZone. *National ICT R&D Management System.* https://https://ezone.iitp.kr

[8] T. K. Kim, H. R. Choi & H. C. Lee. (2016). A Study on the Research Trends in Fintech using Topic Modeling. *Journal of the Korea Academia Industrial cooperation Society, 17(11),* 670-681.
DOI : 10.5762/KAIS.2016.17.11.670

[9] C. S. Kim, S. J. Choi & K. Y. Kwahk. (2017). Investigation of Research Trends in Information Systems Domain Using Topic Modeling and Time Series Regression Analysis. *Journal of Digital Contents Society, 18(6),* 1143-1150.
DOI : 10.9728/dcs.2017.18.6.1143

[10] J. S. Park, N. R. Kim & E. J. Han. (2018). Analysis of Trends in Science and Technology using Keyword Network Analysis. *Journal of the Korea Industrial Information Systems Research, 23(2),* 63-73.
DOI : 10.9723/jksiis.2018.23.2.063

[11] H. Y. Kim & Y. S. Kim. (2019) Trend Analysis of Healthcare Research in Korea using Topic Modeling. *Journal of Wellness, 14(1),* 253-262.
DOI : 10.21097/ksw.2019.02.14.1.253

[12] H. I. Jo, J. W. Kim & B. K. Lee. (2019). A Study on Research Trends of Blockchain Using LDA Topic Modeling : Focusing on United States, China, and South Korea. *Journal of Digital Contents Society, 20(7),* 1453-1460.
DOI : 10.9728/dcs.2019.20.7.1453

[13] Ministry of Science and ICT. (2019). *Administrative Rules(2019-79) National research and development information standard.* Seoul : Ministry of Science and ICT.

[14] David M. Blei, Andrew Y. Ng & Michael I. Jordan. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3,* 993-1022.
DOI : 10.1162/jmlr.2003.3.4-5.993

[15] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer & Richard Harshman. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41(6),* 391-407.
DOI : 10.1002/(SICI)1097-4571(199009)41:6<391

[16] Thomas Hofmann. (2001). Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning, 42(1-2),* 177-196.
DOI : 10.1023/A:1007617005950

[17] AlphaGo versus Lee Sedol, AlphaGo versus Lee Sedol, https://en.wikipedia.org/wiki/AlphaGo_versus_Lee_Sedol

[18] Bitcoin, Bitcoin, https://en.wikipedia.org/wiki/Bitcoin

[19] Ministry of Government Legislation. (2019). *Current Status of Proposals Related to Cryptocurrency.* Seoul : Ministry of Government Legislation.

[20] Organization for Economic Cooperation and Development. (2019). *Artificial Intelligence in Society.* France: Organization for Economic Cooperation and Development.
DOI : 10.1787/eedfee77-en

[21] Lin Liu, Lin Tang, Wen Dong, Shaowen Yao & Wei Zhou. (2016). An Overview of Topic Modeling and its current applications in bioinformatics. *Springerplus, 5(1),* 1608-1630.
DOI : 10.1186/s40064-016-3252-8

우 창 우(Chang Woo Woo) [정회원]

· 2013년 2월 : 충북대학교 컴퓨터공학부(공학사)
· 2015년 2월 : 충북대학교 대학원 의생명과학경영융합대학원(공학석사)
· 2019년 2월 : 충북대학교 대학원 컴퓨터과학(박사수료)
· 2018년 11월 ~ 현재 : 정보통신기획평가원 책임

· 관심분야 : 토픽모델링, 머신러닝, 텍스트마이닝
· E-Mail : cwwoo@iitp.kr

이 종 연(Jong Yun Lee) [종신회원]

· 1985년 2월 : 충북대학교 컴퓨터공학과(공학사)
· 1987년 2월 : 충북대학교 대학원 컴퓨터공학과(공학석사)
· 1999년 2월 : 충북대학교 대학원 전자계산학과(이학박사)
· 1990년 2월 ~ 1996년 5월 : 현대전자산업(주) SW연구소 및 현대정보기술(주) CIM사업부 근무 (책임연구원)
· 1999년 3월 ~ 2003년 2월 : 강원대학교 삼척캠퍼스 정보통신공학과 조교수
· 2003년 3월 ~ 현재 : 충북대학교 소프트웨어학과 교수
· 2010년 3월 ~ 현재 : 한국컴퓨터교육학회 이사
· 2010년 5월 ~ 2017년 12월: 한국융합학회장 역임
· 2018년 9월 ~ 현재 : 충북대학교 전산정보원장
· 2020년 1월 ~ 현재 : 한국융합학회 논문지편집위원장
· 관심분야 : Database System, Medical Informatics, 평가방법론
· E-Mail : jongyun@chungbuk.ac.kr