Journal of the Korea Convergence Society
Vol. 11. No. 9, pp. 21-29, 2020

ISSN 2233-4890 / eISSN 2713-6353
https://doi.org/10.15207/JKCS.2020.11.9.021

다이나믹 토픽 모델을 활용한 D(Data) · N(Network) · A(A.I) 중심의 연구동향 분석

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

Investigation of Research Trends in the D(Data) · N(Network) · A(A.I) Field Using the Dynamic Topic Model

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

요 약 최근 디지털 사회의 도래로 다양한 데이터가 폭발적으로 증가하고, 그중 문헌 내 주제어를 도출하는 토픽 모델링에 관한 연구가 활발히 진행되고 있다. 본 논문의 연구목표는 토픽 모델링 방법 중 하나인 DTM(Dynamic Topic Model) 모델을 적용해 D.N.A.(Data, Network, A.I) 분야에 대한 연구동향을 탐색하는데 있다. 실험 데이터는 최근 6년간(2015~2020) ICT(Information and Communication Technology) 분야 중 기술대분류가 SW · AI에 해당하는 연구과제 1,519개 사업에 대해 DTM 모델을 적용하였다. 실험결과로, D.N.A. 분야의 기술 키워드 Big data, Cloud, Artificial Intelligence와 확장된 의미의 기술 키워드 Unstructured, Edge Computing, Learning, Recognition 등이 매년 연구에 표출되었으며, 해당 키워드 들이 특정 연구과제에 종속되지 않고 다른 연구과제에서도 포괄적으로 연구되고 있음을 확인하였다. 끝으로 본 논문의 연구결과는 향후 D.N.A. 분야에 대한 정책기획 · 과제기획 등 연구개발 기획과정과 기업의 기술 확보전략 · 마케팅 전략 등 다양한 곳에 활용될 수 있을 것으로 기대한다.

주제어 : 다이나믹 토픽 모델링, 연구 트렌드 분석, 시계열 분석, 텍스트 마이닝, 데이터 마이닝

Abstract The Topic Modeling research, the methodology for deduction keyword within literature, has become active with the explosion of data from digital society transition. The research objective is to investigate research trends in D.N.A.(Data, Network, Artificial Intelligence) field using DTM(Dynamic Topic Model). DTM model was applied to the 1,519 of research projects with SW · A.I technology classifications among ICT(Information and Communication Technology) field projects between 6 years(2015~2020). As a result, technology keyword for D.N.A. field; Big data, Cloud, Artificial Intelligence, extended keyword; Unstructured, Edge Computing, Learning, Recognition was appeared every year, and accordingly that the above technology is being researched inclusively from other projects can be inferred. Finally, it is expected that the result from this paper become useful for future policy · R&D planning and corporation's technology · marketing strategy.

Key Words : Dynamic Topic Model, Research Trend Analysis, Time Series Analysis, Text Mining, Data Mining

*Corresponding Author : Jong Yun Lee(jongyun@chungbuk.ac.kr)
Received July 24, 2020 Revised August 28, 2020
Accepted September 20, 2020 Published September 28, 2020

1. 서론

최근 디지털 사회의 도래로 다양한 데이터가 폭발적으로 증가하면서 문헌 내 주제어를 도출하는 연구방법 중 하나인 토픽 모델링에 관한 연구가 활발해지고 있다. 시장조사 전문기관인 Gartner와 IDC에서는 ICT (Information and Communication Technology) 기술의 메가트렌드 중 하나로 빅데이터와 인공지능을, MIT에서는 추론과 학습 및 의사소통이 가능한 인공지능을 혁신기술로 선정하였고[1] 사람과 사람을 연결시키던 전통적인 네트워크 기술은 더 이상 사람과 사람이 아닌 Cloud, Machine to Machine, Internet of Things 등 사람과 사물, 사물과 사물을 대상으로 하는 초연결 사회로 나아가기 위해 변화하고 있다[2].

Apple, Google, Microsoft 등 글로벌 기업들은 인공지능을 핵심 R&D(Research and Development) 분야로 삼고 API(Application Programming Interface) 및 클라우드 서비스 등을 통해 기술의 대중화에 힘쓰고 있지만, 국내의 경우 시장참여자 전체를 견인할 수 있는 선순환 구조의 생태계가 미비함이 현실이다[3]. 이러한 우려 속에서 D.N.A. 분야의 기술경쟁력 확보와 선진국과의 기술격차를 줄이기 위해서는 실패의 가능성이 높은 기술에 대한 진취적인 도전적 연구와 함께 기존의 기술개발 연구동향을 분석해 해당 영역의 생태계 속에서 경쟁력을 가질 수 있는 기술을 개발하는 것이 기술격차를 줄이는 방법 중 하나가 될 수 있다.

위와 관련되어 연구동향을 파악하기 위한 데이터 분석 방법은 통계적 분석, 정량적 분석, 전문가 토의 및 설문조사 등 연구목적과 방법에 따라 다양하게 선택할 수 있지만 본 연구에서는 전체 문서집합을 대상으로 텍스트 분석과 통계적 분석을 적용한 토픽 모델링(Topic Modeling) 방법으로 분석하고자 하였다. 토픽 모델링은 개별 문서에서 내포하고 있는 단어의 확률적인 계산을 통해 전체 문서가 의미하는 토픽을 발견하는 데이터마이닝 기법 중 하나로, 대표적인 LDA(Latent Dirichlet Allocation) 모델[4] 외에 DTM(Dynamic Topic Model) 모델[5], ATM(Author Topic Model) 모델[6], PAM(Pachinko Allocation Model) 모델[7], TOT(Topics Over Time) 모델[8] 등 수십여 개의 파생된 모델이 제시되었다[9].

하지만 토픽 모델링에 대한 대부분의 연구는 모델의 응용보다 알고리즘 개선에 집중되어있고 일부 논문, 특허, 뉴스 등 데이터를 기반으로 연구동향 탐색에 관한 연구가 진행되었지만 이러한 연구 또한 LDA 모델을 중심으로 진행되어 DTM 모델과 같은 파생된 모델을 적용한 사례는 찾을 수 없었다[10-15]. 또한, 연구동향 파악을 위해서는 시계열 분석이 가능해야 하지만 시계열 분석이 가능한 데이터가 있더라도 LDA 모델의 경우 알고리즘의 특징상 단일 시점으로 계산하기 때문에 시계열 분석에 적용하기에는 어려운 구조적인 한계점이 존재한다.

따라서 본 논문에서는 D.N.A. 분야에 대한 연구동향 파악을 위해 국내 연구개발 활동을 지원하는 ICT분야 국가연구개발사업 과제정보 데이터에 시계열 분석이 가능한 DTM 모델을 적용하는데 연구목표를 둔다. 세부적인 연구방법은 다음과 같다. 첫째, 실험 데이터로 과기정통부 ICT분야 국가연구개발사업을 전담하고 있는 정보통신기획평가원에서 최근 6년간(2015~2020) 수행된 전체 과제정보 데이터를 확보하였다. 둘째, 확보된 데이터에 대해 ICT기술대분류를 기준으로 D.N.A. 분야에 적합하다고 판단되는 SW · AI 분야를 선택하였고, 특수문자 제거 및 불용어 제거 등 데이터 전처리를 수행하였다. 셋째, 토픽 모델링 방법 중 시계열 분석이 가능한 Dynamic Topic Model 방법을 적용해 D.N.A. 분야에 대한 연구동향을 도출하였다. 넷째, DTM 모델 결과에 대한 해석과 함께 LDA 모델에도 동일한 데이터셋을 적용해 DTM 모델과 LDA 모델간의 차이를 비교하였다.

마지막으로 본 논문을 통해 도출된 기술의 연구동향과 토픽 모델링을 적용한 연구동향 도출 방법론은 향후 D.N.A. 분야에 대한 정책기획 · 과제기획 등 연구개발 기획 과정과 기업의 기술 확보전략 · 마케팅 전략 등 다양한 곳에 활용될 수 있을 것으로 기대한다.

본 논문의 구성은 다음과 같다. 먼저 2장은 기존의 토픽 모델링 및 시계열분석 등 연구동향 파악을 위해 수행된 연구사례를 분석한 다음, 3장에서 토픽 모델링의 기본 이론이 되는 LDA 모델과 그로부터 확장된 모델 중 하나인 DTM 모델을 요약하고 D.N.A. 분야 연구과제 데이터를 분석하기 위한 연구방법과 실험 데이터를 기술한다. 그리고 4장에서는 DTM 모델과 LDA 모델을 통해 얻은 실험 결과를 분석하고 토의하며 마지막 5장에서 본 연구가 주는 시사점과 향후 연구내용을 요약하며 결론을 맺는다.

2. 관련 연구

김태종(2020)은 코로나19에 관한 언론동향 탐색을 위해 2019년 12월~2020년 3월 기간 동안 '코로나19' 에

관한 기사 47,816건을 수집하였고 토픽 모델링 방법을 적용해 국제적 불안감 고조, 중국 춘절 시기 확산 우려, 국내 확진자 발생, 정부 대응조치 강화, 크루즈선 집단감염 각국 대응현황, 우한 교민 귀국 및 격리, 대구 · 경북지역 확진/사망자 증가, 시민들의 피해 극복 노력 등 토픽을 도출하였다[10].

이현출(2020)은 제20대 국회 보건복지위원회의 회의록을 중심으로 토픽 모델링 방법을 적용해 16개의 토픽을 도출하며 이를 통해 정당별, 주제별, 시기별, 의원 개인의 특성에 따른 대결 구도가 명확하였음을 확인하였고, 정책 및 입법 현안에 대한 관심 분야에서도 각 집단에 따라 차이가 있음을 확인하였다[11].

김갑철(2019)은 국내 지역지리교육 연구동향 분석을 위해 1993년~2019년 기간 동안 '지역, 지역지리, 교육, 학습' 등을 키워드로 입력한 논문 226편을 수집해 토픽 모델링 방법을 적용해 전통지역지리, 평가주의, 수업기술, 가치교육, 교수 방법론 이라는 토픽을 도출하였다[12].

이왕재(2019)는 인공지능의 기술구조 파악을 위해 2007년~2017년 기간 동안 미국 특허청에 등록된 2,589개의 인공지능 특허를 바탕으로 토픽 모델링 방법을 적용해 20개의 토픽을 도출한 후 6개의 분야로 재그룹 함으로써 언어이해, 음성기술, 컴퓨터비전, 데이터 분석, 모션제어, 기계학습으로 구분하였다[13].

조경원(2019)은 4차 산업에 관한 논문의 세부 연구주제 파악을 위해 2016년~2019년 기간 동안 한국학술지인용색인을 통해 '4차 산업' 키워드에 관한 논문 총 685건을 검색/수집하였고 토픽 모델링 방법을 적용해 인공지능, 빅데이터, 사물인터넷, 디지털, 네트워크 등 주요 기술 키워드와 산업, 정부, 교육현장, 일자리 등 4차 산업과 관련된 변화 키워드를 도출하였다[14].

황서이(2019)는 국내 인공지능분야 연구동향 분석을 위해 1985년~2018년 기간 동안 총 1,691편의 학술논문 데이터(서명, 주제어, 초록 등)를 수집해 토픽 모델링 방법을 적용함으로써 자율주행, 인공지능과 인문학, 인공지능 교육, 자연언어생성, 인공지능과 저작권, 인공지능과 인식, 인공지능과 윤리, 인공지능분야 일반, 딥러닝 플랫폼 등 토픽을 도출하였다[15].

3. 연구방법 및 실험 데이터

이 장에서는 실험 데이터 수집과 전처리 방법, 그리고 실질적인 실험 방법에 대해 기술한다.

3.1 데이터 수집

D.N.A. 분야의 연구동향 분석을 위해 과기정통부의 ICT분야 국가연구개발사업을 전담하고 있는 정보통신기획평가원에서 최근 6년간(2015~2020) 지원한 연구과제 데이터 6,576개를 확보하였다. 수집된 데이터는 ICT 분야 기술분류체계에 따라 6가지(통신 · 전파, SW · AI, 방송 · 콘텐츠, 차세대 보안, 디바이스, 블록체인 · 융합)로 분류되어 있으며, 이중 'ICT기술대분류'를 기준으로 과제 및 내용의 주된 성격이 D.N.A. 분야에 적합하다고 판단되는 SW · AI 분야에 해당하는 1,519개 데이터를 선택하였다(Table 1 참고). 수집한 데이터는 124개의 속성을 제공하고 있지만, 연구동향 분석에 불필요하다고 판단되는 속성인 '과학기술표준분야 대 · 중 · 소 · %', 'TRL 시작단계', '주관기관명', '참여기관명', '사업비 지급유형' 등의 속성을 제외한 후, '사업년도', '한글과제명', '영문과제명', '영문키워드', 'ICT기술대분류', '최종목표' 속성을 최종적으로 선택하였다.

Table 1. Number of National ICT R&D from IITP

Field	'15	'16	'17	'18	'19	'20	Total
①	171	185	204	216	220	229	1,225
②	163	270	249	231	291	315	1,519
③	166	172	180	182	160	141	1,001
④	197	121	149	139	124	115	845
⑤	152	104	128	126	100	104	714
⑥	144	218	212	217	249	232	1,272
Total	993	1,070	1,122	1,111	1,144	1,136	6,576

① Telecommunications & Radio
② Software & Artificial Intelligence
③ Broadcasting & Contents
④ Cyber Security
⑤ Device
⑥ Blockchain & Convergence

3.2 데이터 전처리

데이터 전처리는 토픽 모델링의 입력 데이터로 사용되는 '영문키워드'를 중심으로 진행하였다. 첫째, 영문키워드가 영어가 아닌 한글 또는 기호 '-', 'β' 등으로 되어있는 경우 해당 키워드를 제거하였다. 둘째, 기술 중심의 키워드가 아닌 'ICT R&D', 'Development', 'Consulting' 등 연구동향 분석에 영향을 미치지 않는 키워드는 불용어로 판단해 제거하였다. 셋째, 축약어 및 띄워쓰기로 인해 다른 의미로 인식되는 단어 'App/Application', 'AI/Artificial Intelligence', 'Bigdata/Big data' 등은

동일한 의미를 가질 수 있게 변환하였으며, 마지막으로 연구과제 당 키워드 개수가 상이함으로 인해 5개를 초과하거나 크게 미달하는 연구과제는 모두 제거하였다. 위의 전처리 과정을 거쳐 최종 943개의 연구과제에 대해 4,634개의 영문키워드를 확보하였으며, 이를 중심으로 DTM 모델을 적용해 D.N.A 분야 연구동향 분석을 수행하였다(Fig. 1 참고).

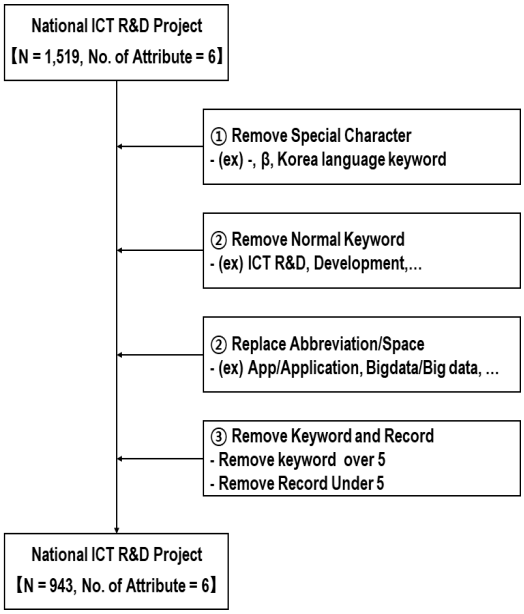


Fig. 1. Data Preprocessing Steps

3.3 토픽 모델링

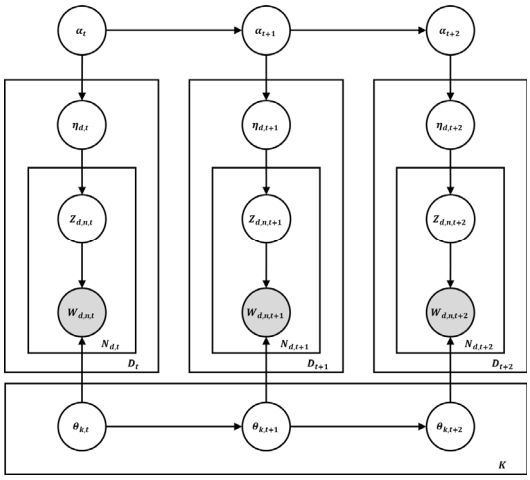
토픽 모델링이란 문서에서 내포하고 있는 단어의 확률적인 계산을 통해 숨겨진 토픽을 발견하는 텍스트마이닝 기법 중 하나로, 최초의 토픽 모델은 1990년 문서-단어 행렬에 대한 SVD(Singular Value Decomposition)를 기본 개념으로 출발해 {문서-단어} 하나의 행렬을 {문서-토픽}과 {토픽-단어} 두 개의 행렬로 분해하는 LSI(Latent Semantic Indexing) 모델[16] 이다. 이후 단어의 출현 빈도를 확률로 대체하기 위한 pLSI (probabilistic Latent Sematic Indexing) 모델[17]과 불안전한 확률을 갖는 pLSI 모델을 보완한 LDA(Latent Dirichlet Allocation) 모델이 등장하면서 현재 토픽 모델링의 표준으로 자리 잡게 되었다. LDA 모델은 문서-토픽별 분포 값과 토픽-단어별 분포 값을 미리 결정한 후 단어들을 관측하고 반복함으로써 단어가 어떤 토픽에 속해야 하는지 확률적으로 추론한다. 2003년 LDA가 발표된 이후 본격적으로 토픽 모델링 분야에 대한 연구가 시작되었으며 현재까지 DTM(Dynamic Topic Model) 모델을 포함해 ATM(Author Topic Model), PAM(Pachinko Allocation Model), TOT(Topics Over Time) 등 수십개의 파생된 모델이 제시되고 있다(Table 2 참고).

Table 2. Some impressive articles based on LDA

Year	Model	Year	Model
2003	corr_LDA	2011	MM LDA
2004	Author Topic Model	2011	Sentence LDA
2005	Author Recipient Topic	2011	PLDA+
2006	Dynamic Topic Model	2011	Dirichlet class language model
2006	Pachinko Allocation Model	2012	Locally Discriminative Topic Model
2006	Topics Over Time	2012	ET-LDA
2007	GWN LDA	2012	iLHA
2008	on line LDA	2012	Mr. LDA
2008	MG LDA	2012	FB LDA
2009	Labeled LDA	2012	Factorial LDA
2009	HDP LDA	2012	SSh LDA
2009	Disc LDA	2013	Utopian
2010	GeoFolk	2013	LDA-GA
2010	Joint LDA	2013	MSA-COSRs
2010	Topic Aspect Model	2013	TopicSpam
2010	Topic XP	2013	WT-LDA
2010	Constrained LDA	2014	Emotion-LDA
2010	EDCoW	2014	Red-LDA
2010	Dependency Sentiment LDA	2014	Biterm Topic Modeling

3.4 다이나믹 토픽 모델링

DTM 모델은 LDA 모델을 기반으로 하지만 시간이라는 변수를 추가함으로써 시간의 흐름에 따라 토픽이 서로 독립적이지 않고 연결될 수 있다고 제시하였다[5,18]. 이로 인해 DTM 모델을 사용한다면 시간에 따라 토픽과 토픽에 속하는 단어가 어떻게 변화하는지 파악할 수 있다. DTM 모델의 데이터 처리 절차는 Fig. 2와 같다. 먼저 ① t개의 연도에 대해 연도-문서의 토픽별 편향도 벡터인 α_t와, ② t개의 연도 내 K개의 토픽에 대해 연도-토픽의 단어 분포 벡터 $\theta_{k,t}$를 결정한다. ③ 그리고 t개의 연도 내에 있는 전체 문서 D_t에서 연도-문서의 토픽 분포 벡터 $\eta_{d,t}$를 선택한다. 다음으로 ④ 토픽 번호인 $z_{d,n}$을 계산한 후 실제 관측하는 단어 ⑤ $w_{d,n,t}$를 계산, ④~⑤ 과정을 반복하며 관측하는 단어 $w_{d,n,t}$가 어떤 토픽번호 $z_{d,n}$에 속해야 하는지 추론한다.



α_t	:	연도-문서의 토픽별 편향도 벡터
$\eta_{d,t}$	:	연도-문서의 토픽 분포 벡터
$z_{d,n}$	:	연도-문서-단어의 토픽 번호
$w_{d,n,t}$	:	연도-문서-단어의 실제 단어
$\theta_{k,t}$	:	연도-토픽의 단어 분포 벡터
$N_{d,t}$	:	문서의 단어 개수
D_t	:	문서의 개수
K	:	토픽의 수
t	:	연도의 수

Fig. 2. Generative Process of Dynamic Topic Model (for three time slices)

3.5 실험환경 및 분석 프로세스

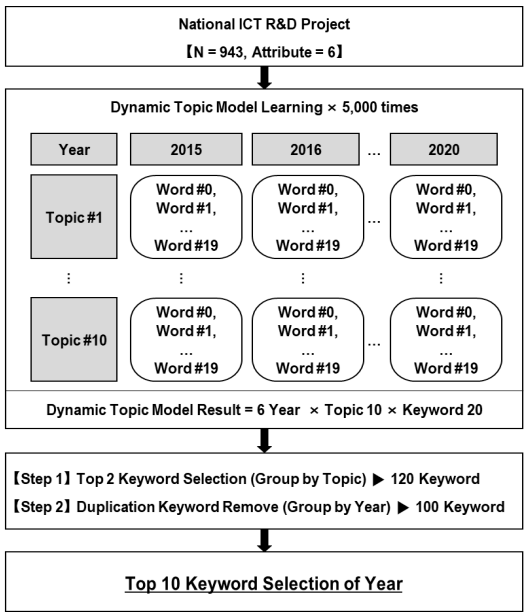


Fig. 3. Analysis Processing Steps

앞서 전처리가 완료된 943개의 ICT분야 국가연구개발사업 연구과제 데이터에 대해 토픽 모델링 방법 중 하나인 DTM 모델을 적용해 5,000번의 학습을 통해 2015년~2020년 각 연도별 연구토픽 10개와 토픽별 상위 20개 키워드 총 1,200개 키워드를 도출하였고, 토픽별 상위 2개 키워드 선택 및 중복키워드 제거를 통해 최종 연도별 상위 10개의 키워드를 선별해 연구동향을 탐색하였다(Fig. 3 참고).

4. 실험결과 및 토의

4.1 실험결과

위에서 서술한 분석 프로세스에 따라 실험한 결과, ICT분야 국가연구개발로 수행된 SW · AI 연구과제는 대부분 D.N.A. 분야의 기술로 확인되었다(Table 3 참고).

구체적인 결과의 해석을 위해 확장된 의미에서 D.N.A. 분야에 포함된다고 판단한 Unstructured, Edge Computing, Learning, Recognition 등 키워드는 음영처리를 하였으며, D.N.A. 분야에 대한 키워드가 정확하게 일치하는 Big data, Cloud, Artificial Intelligence 키워드에는 음영처리와 함께 추가로 표기(★) 하였다.

Table 3 과 같이 연도별로 정확하게 일치하거나 확장된 기술적 의미와 키워드를 포함하고 있는 각각 과제에 대한 최종목표 등을 검토하였을 때 2015년도 D.N.A. 분야에 관한 키워드는 총 5개로, Data 분야에 관련된 키워드로 Dashboard, Mapping, Monitoring이 빅데이터 기반의 분석도구 및 시스템 등을 개발하는 과제로 확인되었고, Network 분야에는 Cloud, RealTime OS가 클라우드 보안 및 클라우드 기반의 개방형 플랫폼 등을 개발하는 과제임을 확인하였다.

2016년도 D.N.A. 분야에 관한 키워드는 총 7개로, Data 분야에 관련된 키워드로 Big data가, Network 분야에는 Cloud, Cloud Computing, Cloud Platform, Digital Extinction, Internet of Things, RealTime IoT가 클라우드 기반의 플랫폼 및 SaaS(Software as a Service) 등을 개발하는 과제임을 확인하였다.

2017년도 D.N.A. 분야에 관한 키워드는 총 6개로, Data 분야에 관련된 키워드로 Big data가, Network 분야에는 Controlling as a Service, Virtual Desktop Infrastructure가 Fog Computing 및 IaaS(Infrastructure as a Service) 등을 개발하는 과제, 그리고 A.I 분야에는

Table 3. Dynamic Topic Model Results of National ICT R&D between 2015 and 2020

Keyword#	2015	2016	2017	2018	2019	2020
0	Application	Big data★	Big data★	Artificial Intelligence★	Artificial Intelligence★	Artificial Intelligence★
1	Application Classification	Cloud★	Calendar Photo Album	Automotive Display	Natural Language Processing	Big data★
2	Cloud★	Cloud Computing	Controlling as a Service	Big data★	Cloud Computing	Data Augmentation
3	Dashboard	Cloud Platform	Emotion Recognition	Data Preprocessing	Deep Learning	Database
4	Device API	Digital content	Image Recognition	Deep Learning	Human AI Interaction	Deep Learning
5	E-Book	Digital Extinction	Pattern Recognition	Intelligence and Information	Internet of Things	Edge Computing
6	Light Weight Encryption	Digital rights management	Response Generation	Learning and Inference	Long Time Tracking	Emergency Alert System
7	Mapping	Internet of Things	Software Prosumer	Natural Language Processing	Machine Learning	Framework
8	Monitoring	Platform	Verification	Operating System	Platform	Interactive Learning
9	RealTime OS	RealTime IoT	Virtual Desktop Infrastructure	Video	Unstructured	Machine Learning

Emotion Recognition, Image Recognition, Pattern Recognition이 통합 인지 플랫폼, 사용자 감정 추론 등을 개발하는 과제임을 확인하였다.

2018년도 D.N.A. 분야에 관한 키워드는 총 6개로, Data 분야에 관련된 키워드로 Big data, Data Preprocessing이 데이터 품질에 관한 평가방법 등을 개발하는 과제로 확인되었고, A.I 분야에는 Artificial Intelligence, Deep Learning, Learning and Inference, Natural Language Processing이 감염병 대응 시스템, 교통 관련 솔루션 등을 개발하는 과제임을 확인하였다.

2019년도 D.N.A. 분야에 관한 키워드는 총 8개로, Data 분야에 관련된 키워드로 Unstructured가 비정형 데이터 처리 및 이와 관련된 서비스 등을 개발하는 과제로 확인되었고, Network 분야에는 Cloud Computing, Internet of Things가 Edge Cloud, PaaS(Platform as a Service) 및 Public Cloud 등을 개발하는 과제, 그리고 A.I 분야에는 Artificial Intelligence, Natural Language Processing, Deep Learning, Human AI Interaction, Machine Learning이 Chatbot, BCI(Brain Computer Interface) 등을 개발하는 과제임을 확인하였다.

마지막으로 2020년도 D.N.A. 분야에 관한 키워드는 총 8개로, Data 분야에 관련된 키워드로 Big data, Data Augmentation, Database가 데이터 변형 및 증식과 In-memory 기반의 DBMS 구축 등을 개발하는 과제로 확인되었고, Network 분야에는 Edge Computing이 초저지연/지능형 클라우드 및 엣지 클라우드 가속화 기술 등을 개발하는 과제, 그리고 A.I 분야에는 Artificial Intelligence, Deep Learning, Interactive Learning, Machine Learning이 의사결정 지원 시스템, 멀티 모달 데이터 분석 및 시맨틱 분석과 추론 등을 개발하는 과제임을 확인하였다.

4.2 토의

본 연구에서 활용한 SW · AI 연구과제 데이터는 기술분류체계에 따라 D.N.A. 분야에 해당하는 빅데이터, 클라우드 컴퓨팅, 인공지능 기술 외에도 응용기반 및 특화에 관한 응용SW, 운영체제와 미들웨어에 관한 시스템SW, 그리고 차세대 컴퓨팅 및 기반 컴퓨팅을 아우르는 컴퓨팅시스템 등 다른 기술까지 폭넓게 포함하고 있다. 연구과제가 기술분류체계 별로 균등하게 지원되지 않는다는 점과 실험을 통해 확인한 최근 6년간의 연구동향이 D.N.A. 분야에 집중되어 있음을 볼 때 D.N.A. 분야의 기술이 다른 기술 분야의 연구과제에도 포괄적으로 연구되고 있음을 확인하였고, 해당 키워드를 포함하고 있는 각각의 연구과제에 대한 최종목표를 함께 확인한 결과 기술개발뿐만 아니라 응용SW와 서비스 등도 활발하게 개발되고 있음을 확인할 수 있었다. 특히, 인공지능 기술의 경우 Machine Learning, Deep Learning, Pattern Recognition 및 A.I Accelerator 등 시간이 지날수록 기술의 고도화와 함께 다양한 플랫폼과 서비스 등이 지속적으로 확산되고 있음을 확인할 수 있었다.

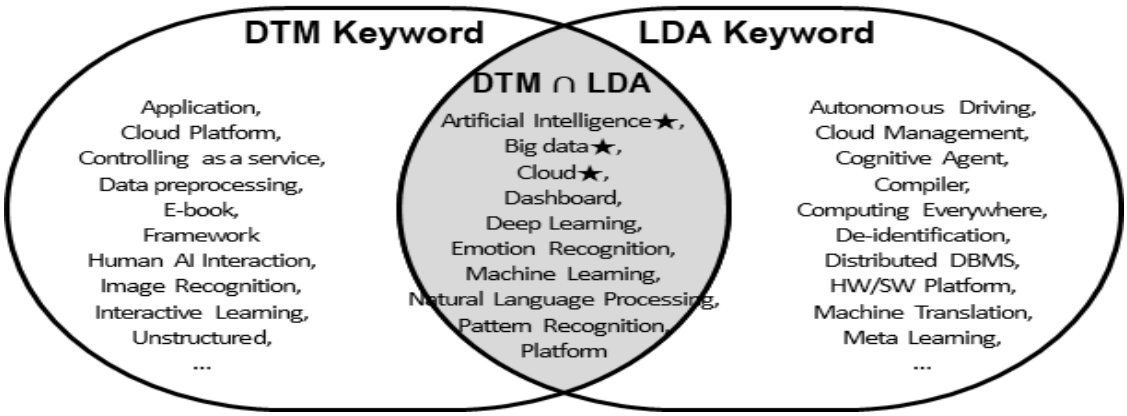


Fig. 4. Comparison of DTM Keywords with LDA Keywords between 2015 and 2020

더불어 위 실험과 동일한 데이터 및 전처리 방법을 적용해 LDA 모델과의 비교를 위한 추가 실험을 진행하였다. LDA 모델의 토픽 수는 위 실험 모델의 기간(2015~2020)과 동일하게 6개로 설정하였으며 키워드 또한 기존 실험과 비교 될 수 있게 상위 10개의 키워드를 선택하였다. 위 과정을 통해 도출한 LDA 모델의 키워드와 기존 실험을 통해 얻은 DTM 모델의 키워드를 중심으로 Fig. 4와 같이 키워드간 상호비교를 진행하였다.

비교 결과 D.N.A. 분야에 해당하는 핵심 키워드인 Big data, Cloud, Artificial Intelligence 는 두 모델에서 모두 등장하였으며, 이 외도 Dashboard, Deep Learning, Emotion Recognition, Machine Learning, Natural Language Processing, Pattern Recognition, Platform 키워드도 공통적으로 등장하였다. 하지만, DTM 모델에서 등장한 Cloud Platform, Human AI Interaction 등 키워드는 LDA 모델에서 나타나지 않았고, LDA 모델에서 등장한 Compiler, Distributed DBMS 등 키워드 또한 DTM 모델에서 나타나지 않았다.

LDA 모델과 DTM 모델은 동일한 데이터와 절차를 통해 D.N.A. 분야의 핵심 키워드와 확장된 의미의 키워드 일부를 공통적으로 도출하였는데, 이는 DTM 모델의 기본 알고리즘이 LDA 모델을 근간으로 출발하였기 때문에 유사한 결과가 나타날 수 있었음을 확인할 수 있었다. 하지만, LDA 모델은 전체 데이터를 하나의 시점으로 관측해 주제를 도출하는 반면 DTM 모델의 경우 전체 데이터를 각각의 시점으로 분석하기 때문에 전혀 다른 키워드가 등장할 수 있었다. 실제로 LDA 모델에서 도출된 키워드가 속한 과제들을 추적한 결과 연도의 구분이 없었지만, DTM 모델의 경우 연도별로 수행된 과제 내에서만 키워드가 도출 되었다.

이는 20년, 30년 등 더 긴 시점을 갖는 데이터를 DTM 모델과 LDA 모델로 분석하였을 때 시간의 흐름에 따른 토픽의 변화로 인해 두 모델이 완전히 다른 키워드를 반환할 수 있다는 점을 시사하고 있으며, 본 연구는 연구목적에 맞는 D.N.A. 분야의 연구동향 분석을 위해 DTM 모델이 적합하다고 판단하였다.

5. 결론

본 논문은 최근 6년간 수행된 SW · AI 분야의 연구과제 데이터 분석을 통해 D.N.A. 분야의 연구동향을 도출하였다. 실험결과 2015년과 2016년은 데이터와 네트워크 분야의 연구가 활발하게 진행되었으며, 2017년 이후부터는 인공지능에 관한 연구가 활발하게 진행되었음을 확인하였다.

본 연구가 주는 시사점은 다음과 같다. 첫째, D.N.A. 분야의 기술이 다른 기술 분야의 연구과제 등에도 포괄적으로 적용되어 연구되고 있음을 확인하였고, 둘째, D.N.A. 분야의 기술개발 및 고도화뿐만 아니라 관련된 응용SW 및 서비스 등도 활발하게 개발되고 있음을 확인하였다. 셋째, 전체 문서집합에 대한 연구주제를 파악하는 방법으로 대표적인 토픽 모델링 방법인 LDA 모델을 사용 할 수 있지만, 연구주제와 연구동향을 함께 파악하기 위해서는 DTM 모델이 더 적합함을 확인하였다. 끝으로 본 연구를 통해 도출한 D.N.A. 분야의 연구동향과 토픽 모델링을 적용한 연구동향 도출 방법론은 향후

D.N.A. 분야에 대한 정책기획·과제기획 등 연구개발 기획 과정과 기업의 기술 확보전략·마케팅 전략 등 다양한 곳에 활용될 수 있을 것으로 기대한다.

본 연구의 한계점은 다음과 같다. DTM 모델의 결과값이 전체 연도에 해당하는 연도별 토픽 K개와 그에 해당하는 키워드 N개를($Year \times K \times N$) 반환하기 때문에 토픽을 해석하는 방법에 따라 결과가 크게 변할 수 있음을 밝힌다. 본 연구에서는 일정 개수의 토픽을 유지하고 상위 키워드를 선택하는 방법으로 해석하였지만, 연구방법에 따라 토픽의 개수를 최소로 줄이거나 직관적으로 키워드를 제거하는 등 다양한 방법으로 해석할 수 있다. 이러한 점은 모든 토픽 모델링의 결과가 토픽과 키워드로 반환되기 때문에 사용자 관점의 2차 해석이 필요하다는 특징을 갖는다.

따라서 향후 토픽 모델링 결과 자체의 정확도를 높이기 위해 연구과제 데이터와 기술 키워드에 대한 리스트 등이 추가로 확보될 필요가 있으며, 토픽 모델링 결과에 대한 사용자 관점의 2차 해석을 지원하기 위해 추가적인 모델 또는 프레임워크 등 대한 연구가 진행될 필요가 있다.

REFERENCES

[1] H. J. Lee, P. M. Ryu, S. J. Lim, M. K. Jang & H. K. Ki. (2014). Technology Trends of AI for Big Data Knowledge Processing. *Electronics and Telecommunications Trends, 29(4),* 30-38.
DOI : 10.22648/ETRI.2014.J.290404

[2] T. K. Kang, Y. H. Kang, Y. C. Ryoo & T. S. Cheung. (2019). Research Trend in Ultra-Low Latency Networking for Fourth Industrial Revolution. *Electronics and Telecommunications Trends, 34(6),* 108-122.
DOI : 10.22648/ETRI.2019.J.340610

[3] J. Y. Lee & B. S. Cho. (2016). Suggestions for Nurturing Ecosystem to Spur Artificial Intelligence Industry. *Electronics and Telecommunications Trends, 31(2),* 51-62.
DOI : 10.22648/ETRI.2016.J.310206

[4] 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

[5] David M. Blei & John D. Lafferty. (2006). Dynamic Topic Models. *Proceedings of the 23rd international conference on Machine learning,* 113-120.
DOI : 10.1145/1143844.1143859

[6] Michal Rosen Zvi, Thomas Griffiths, Mark Steyvers & Padhraic Smyth. (2004). The Author-Topic Model for Authors and Documents. *Proceedings of the 20th conference on Uncertainty in artificial intelligence,* 487-494.

[7] Wei Li & Andrew McCallum. (2006). Pachinko Allocation : DAG-Structured Mixture Models of Topic Correlations. *Proceedings of the 23rd international conference on Machine learning,* 577-584.
DOI : 10.1145/1143844.1143917

[8] Xuerui Wang & Andrew McCallum. (2006). Topics over time: a non-Markov continuous-time model of topical trends. *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining,* 424-433.
DOI : 10.1145/1150402.1150450

[9] Hamed Jelodar, Yongli Wang, Chi Yuan, Xia Feng, Xiahui Jiang, Yanchao Li & Liang Zhao. (2018). Latent Dirichlet Allocation (LDA) and Topic modeling: models, applications, a survey. *Multimedia Tools and Applications, 78,* 15169-15211.
DOI : 10.1007/s11042-018-6894-4

[10] T. J. Kim. (2020). COVID-19 News Analysis Using News Big Data : Focusing on Topic Modeling Analysis. *The Journal of the Korea Contents Association, 20(5),* 457-466.
DOI : 10.5392/JKCA.2020.20.05.457.

[11] H. C. Lee, J. H. Jang & K. T. Kim. (2020). A Study on the Conflict Structure of the Standing Committee through Topic A[4nalysis of the National Assembly Minutes : Health and Welfare Committee in the First Half of the 20th National Assembly. *Korean Party Studies Review, 19(2),* 131-167.
DOI : 10.30992/KPSR.2020.06.19.2.131

[12] G. C. Kim & H. J. Noh. (2019). Research Trends of Regional Geography Education Using Topic Modeling. *Social Studies Education, 58(4),* 49-67.
DOI : 10.37561/sse.2019.12.58.4.49

[13] W. J. Lee & H. Y. Lee. (2019). A Technology Landscape of Artificial Intelligence: Technological Structure and Firms' Competitive Advantages. *Journal of Korea technology innovation society, 22(3),* 340-361.
DOI : 10.35978/jktis.2019.06.22.3.340

[14] K. W. Cho & Y. W. Woo. (2019). Topic Modeling on Research Trends of Industry 4.0 Using Text Mining. *Journal of the Korea Institute of Information and Communication Engineering, 23(7),* 764-770.
DOI : 10.6109/jkiice.2019.23.7.764

[15] S. I. Hwang & M. K. Kim. (2019). An Analysis of Artificial Intelligence(A.I.) : related Studies' Trends in Korea Focused on Topic Modeling and Semantic Network Analysis. *Journal of Digital Contents Society, 20(9),* 1847-1855.
DOI : 10.9728/dcs.2019.20.9.1847

[16] 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

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

[18] Arnab Bhadury, Jianfei Chen, Jun Zhu & Shixia Liu. (2016). Scaling up Dynamic Topic Models. *WWW '16 : Proceedings of the 25th International Conference on World Wide Web,* 381-390.
DOI : 10.1145/2872427.2883046

[19] 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

[20] Derek Greene & James P. Cross. (2017). Exploring the Political Agenda of the European Parliament Using a Dynamic Topic Modeling Approach. *Political Analysis, 25(1),* 77-94.
DOI : 10.1017/pan.2016.7

[21] Allen H. Huang, Reuven Lehavy, Amy Y. Zang & Rong Zheng. (2017). Analyst Information Discovery and Interpretation Roles: A Topic Modeling Approach. *MANAGEMENT SCIENCE, 64(6),* 2473-2972.
DOI : 10.1287/mnsc.2017.2751

[22] Andra-Selina Pietsch & Stefan Lessmannb. (2018). Topic modeling for analyzing open-ended survey responses, *JOURNAL OF BUSINESS ANALYTICS, 1(2),* 93-116.
DOI : 10.1080/2573234X.2019.1590131

[23] Paulo Bicalho, Marcelo Pita, Gabriel Pedrosa, Anisio Lacerda & Gisele L. Papp. (2017). A general framework to expand short text for topic modeling. *Information Sciences, 393,* 66-81.
DOI : 10.1016/j.ins.2017.02.007

[24] Pooja Kherwa & Poonam Bansal. (2019). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems, 7(24),* 1-16.
DOI : 10.4108/eai.13-7-2018.159623

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