텍스트 마이닝을 활용한 국가 R&D과제 동향 분석: ICT 분야와 스마트시티 중심으로

김성순, 양명석
한국과학기술정보연구원
{seongkim, msyang}@kisti.re.kr

A Study on the Analysis of ICT R&D using Text Mining Method: Focused on ICT Field and Smart City

Seong-soon Kim and Myung-seok Yang
Korea Institute of Science and Technology Information

요 약

본 연구는 최근 ICT분야 R&D 동향을 파악하기 위하여 NTIS에서 제공하는 국가연구개발사업 과제 정보를 텍스트 마이닝 기법을 통해 분석하였다. 2017년부터 2020까지의 과제 정보에서 키워드를 추출하고 연결 관계 마이닝을 통해 키워드 네트워크를 시각화하였다. 분석 결과는 다음과 같다. 첫째, 정보통신 각 분야에서 핵심 연구주제가 기술의 발전에 따라 변화하고 있음을 관찰하였다. 둘째, 키워드 네트워크 상에서 허브 역할을 하는 키워드를 통해 분야 간 융합의 매개 기술을 파악할 수 있었다. 마지막으로, 연도별 키워드 네트워크를 비교·분석함으로써 새롭게 등장하거나 연결 상태의 변화를 보이는 이머징(Emerging) 키워드를 통해 미래 유망 기술이나 최신 연구 방향성을 감지할 수 있음을 보였다.

1. 서론

최근 디지털 기반 지능화 기술을 바탕으로 하는 4차 산업혁명이 국가 당면 위기인 저성장 극복 및 미래형 사회로의 대전환을 가능케 하는 새로운 동력으로 주목받음에 따라, 정부를 중심으로 국가 R&D 체계를 포함한 경제·산업 전반에 걸쳐 디지털 지능화 프로젝트가 다각도로 추진되고 있다. 국가 R&D 사업 중 연구개발과제의 상당수는 중장기 투자 계획과 연계되어 국가 과학기술 경쟁력 제고에 필요한 요소 기술을 중심으로 기획 및 선정되고 있다. 중장기적 관점에서 많은 비용과 인력이 투자되는 R&D 정책이 성공적으로 추진되기 위해서는 연구 트렌드 파악을 통한 국가 정책목표와의 일치 여부, 기술 로드맵과의 정합성 등을 모니터링 하는 것이 중요하다. 특히, 기존 대비 빠른 주기로 광범위하게 나타나는 ICT분야 신기술의 발현 특성과 분야 간 융합 현상을 적시에 발견하고 해당 정보를 R&D 정책 수립 등에 적극적으로 활용하는 것이 관건이다.

본 연구에서는, 국가 R&D 과제 정보를 바탕으로 텍스트 마이닝 방법을 적용하여 ICT 분야 국가 R&D의 최근 연구 동향을 파악하고 새롭게 부상하는 기술의 탐색 방법을 제안한다. 구체적으로는, 연관규칙 기반 동시출현단어 분석 기법(Co-occurrence Word Analysis, 이하 CWA)을 통해 ICT 분야 전체, 그리고 하위 분야인 스마트시티 관련 R&D과제의 키워드 네트워크를 구축하고 주요 연구 주제 및 연구 분야 간의 지식 관계를 추출한다. 그 다음으로, 구축된 네트워크의 중심성 분석을 통해 주요 연구 분야의 추세 변화를 계량화한다. 또한, 연도별 네트워크 구조의 비교 분석을 통해 새롭게 출현하는 이머징 트렌드(Emerging Trend)를 포착한다.

본 논문의 구성은 다음과 같다. 2장 관련 연구에서는 국가 R&D 자료 바탕의 텍스트 분석 방법론을 살펴본다. 3장에서는 본 연구에서 사용한 데이터와 분석 방법을 소개하고 4장에서 분석 결과를 요약하며 5장에서 향후 계획 및 결론을 맺는다.

2. 관련 연구

텍스트 마이닝 기반 트렌드 분석. 서원철[1]은 국가 R&D 특허 정보를 활용하여 키워드 네트워크를 구축하고 시각화하였다. 분석 대상 연도 별 네트

워크의 양상을 비교함으로써 산업 분야 별 연구 트렌드의 변화를 규명하였다. 김지훈[2]은 농업 분야의 R&D 트렌드를 파악하기 위하여 국가과학지식정보서비스(National science and Technology Information Service, 이하 NTIS)에서 획득한 과제 정보에 텍스트 마이닝 기법을 적용하고 향후 농업 연구의 방향성을 전망하였다.

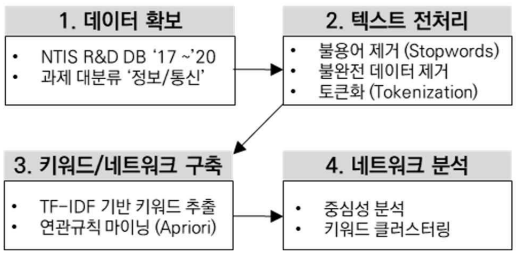
R&D과제정보 기반 키워드 마이닝. 소재현[3]은 국내·외 연구사업 보고서에서 텍스트 마이닝을 통해 추출한 핵심 키워드를 바탕으로 스마트시티의 개념 정립을 시도하였다. 최현홍[4]은 국내 ICT 분야의 보고서·간행물 자료에 텍스트 분석을 수행하고 잠재적 토픽을 도출하여 ICT융합 관련 기술 흐름을 식별하였다. 또한, 연도별 토픽의 상승 및 하락 추세 분석을 통해 차세대 융복합 분야를 추정하였다. 김태현[5]은 국가 R&D 과제의 특성을 효율적으로 파악하기 위하여 국가과학기술표준분류 체계를 활용한 R&D 용어 구축 방법을 제안하였다.

관련연구 조사 결과, 국가 R&D 정보나 보고서 등 텍스트 분석 기반의 기술 동향 파악 연구에서는 주로 TF-IDF[6] 계산 기반 키워드 추출, 연관 관계 추출 및 키워드 네트워크 구축, 빈도 분석, 토픽 모델링 등의 작업이 공통적으로 수행되는 것으로 파악되었다. 본 연구에서도 기존 텍스트 기반 분석 방법의 기본적인 틀을 유지하면서 정보통신 분야, 그 중에서도 4차 산업혁명의 핵심 융합 사례로 꼽히는 스마트 시티 관련 R&D 과제를 중점적으로 살펴본다.

3. 분석 방법

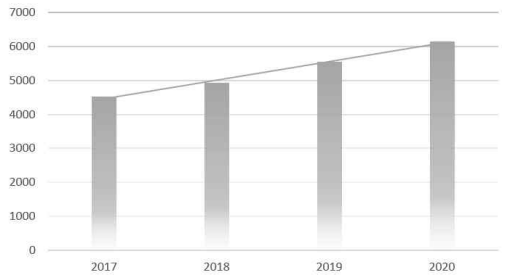
본 연구의 분석 절차는 그림 1과 같이 크게 데이터 수집, 텍스트 전처리, 키워드 추출 및 네트워크 구축, 네트워크 분석 단계로 구성된다.

(그림 1) 분석 절차



1. 데이터 확보. 국가 R&D 과제 정보를 수집하기 위하여 2017년부터 2020년까지 NTIS (https://www.ntis.go.kr) 시스템에 등록된 연구과제 중 대분류 코드가 '정보/통신'인 과제들을 분석 대상으로 하였으며, 연도별 과제 수는 그림 2와 같다. 2017년도를 기점으로 하여 과학기술 투자 확대 정책의 영향으로 ICT분야의 국가 R&D 과제의 양이 꾸준히 증가함을 알 수 있다.

(그림 2) 연도별 정보/통신 분야 과제 검색량



2. 텍스트 전처리. 과제를 구성하는 핵심 텍스트를 추출하기 위해 '과제명', '연구 요약', '키워드' 필드를 활용하였으며, 과제명이 중복되거나, 필수 값이 누락된 데이터는 전처리 과정에서 제외하였다. 선별된 텍스트는 KoNLPy와 Mecab 패키지의 형태소 분석기를 통해 토큰화 하였다.

3. 키워드 추출 및 연관 분석. 분석 대상 과제의 특성을 잘 나타내는 키워드를 추출하기 위해 과제명, 연구요약, 키워드 정보를 핵심 요약 정보로 간주하고 해당 필드를 대상으로 명사, 명사구를 추출하였다. 이 과정에서 키워드의 TF-IDF값을 계산하여 영향력이 크지 않은 단어들은 불용어 처리 하였다. TF(Term Frequency)는 문서 내에 특정 단어의 출현 빈도이며, DF(Document Frequency)는 전체 문서에서 특정 단어의 출현 빈도를 의미하는 값으로, DF의 역수를 취한 IDF(Inverse Document Frequency)값과 TF값의 결합을 통해 특정 단어의 문서 내 중요도를 파악할 수 있다.

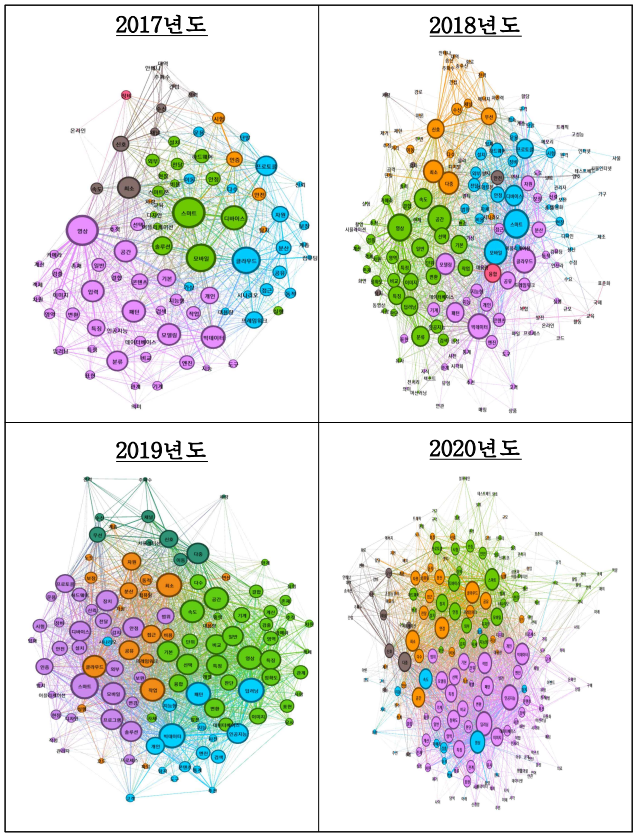
동시출현단어 분석 기법(CWA)는 문서 내에서 두 단어가 동시에 자주 출현하였을 때, 두 단어간의 관련성이 높다고 판단하는 것이다[7]. 연관 규칙의 유효성을 나타내는 지표는 지지도(Support), 신뢰도(Confidence), 향상도(Lift)가 있다. 지지도는 규칙 $A \rightarrow B$에 대하여 사건 A와B 가 동시에 일어날 확률 $P(A \cap B)$로 정의된다. 신뢰도는 A가 주어졌을 때 사건 B 가 일어날 조건부 확률 $P(B|A)$로 표현된다. 향상도는 신뢰도를 기대 확률로 나눈 값으로, 1이상 값을 가질 때 향상도가 높다고 판단한다[8, 9]. 본 연구에서는 신뢰도(Confidence) 값을 이용하여 과제 정보에서 동시에 출현하는 단어 간의 통계적 유효성을 판단하고 의미 있는 관계를 필터링 하였다.

4. 네트워크 분석. 연관규칙 마이닝을 통해 추출된 키워드 쌍 정보를 바탕으로 네트워크를 구축한 다음, 각 키워드 간의 상호작용 및 연결 상태의 구조적 속성을 판단하기 위해 중심성 분석 및 노드 클러스터링을 수행하였다. 중심성 지표로는 연결된 노드의 개수뿐만 아니라 연결된 다른 노드의 중요도를 함께 고려하는 아이겐벡터 중심성(Eigenvector Centrality)를 채택하였다. 또한, 서브 네트워크를 추출하기 위하여 커뮤니티 탐지 알고리즘(Community Detection)을 적용하였다.

4. 분석 결과

본 장에서는 연관규칙 마이닝을 통해 추출된 연관 키워드 쌍으로 네트워크를 구축하고 중심성 분석을 수행한 결과를 서술한다. 그림 3은 그래프 분석 툴인 Gephi를 활용하여 키워드 네트워크를 연도별로 시각화 한 것이다.

(그림 3) 정보통신 분야 국가 R&D 과제 키워드 네트워크



노드의 크기는 해당 키워드의 위세 중심성을 나타내며, 엣지의 굵기는 연결된 노드 사이의 연결 강도를 표시한다. 각각 다른 색으로 표시된 노드 집합은 모듈 추출을 통해 자동적으로 분리된 결과이다.

1. 정보통신 R&D 키워드 중심성 변화. 표 1은 중심성 상위 키워드를 순위화 하여 나타낸 것이다. 키워드의 중심성 변화를 연대기적으로 살펴봄으로써 시간의 흐름에 따른 주요 키워드의 상승·하락 추세를 파악하는 것이 가능하다. 2017년도에는 '스마트', '모바일', '클라우드', '빅데이터' 등의 키워드가 상위권에 위치하고 있다. 주목할 점은, 2018년 본격적인 지능화 세대에 접어들면서 '딥러닝', '융합' 이라는 키워드가 등장하기 시작했다. 이것은 2017년 말, 정부가 4차 산업혁명 대응 계획(I-Korea 4.0)을 발표하고, 인공지능 및 차세대 신기술 관련 투자를 대폭 확대한 것과 무관하지 않은 것으로 보인다. 2019년 이후부터는 '딥러닝', '인공지능', '빅데이터'로 대표되는 지능정보기술 분야 핵심 키워드가 최상위권에 자리 잡고 있다.

<표 1> 연도별 중심성 상위 키워드 변화

Rank	2017	2018	2019	2020
1	스마트	스마트	**딥러닝**	**딥러닝**
2	**영상**	**영상**	**영상**	**인공지능**
3	모바일	공간	카메라	**영상**
4	클라우드	모바일	**인공지능**	**빅데이터**
5	디바이스	**빅데이터**	디바이스	카메라
6	**빅데이터**	클라우드	스마트	이미지
7	패턴	모델링	**빅데이터**	특징
8	모델링	디바이스	지능형	검출
9	공간	패턴	지능	객체
10	프로토콜	프로토콜	검출	디바이스
11	프레임워크	신호	이미지	스마트
12	콘텐츠	**딥러닝**	패턴	지능형
13	분산	융합	모바일	지능
14	지능형	무선	주파수	클라우드
15	데이터베이스	이미지	클라우드	컴퓨팅

2. 세부 키워드 변화. 키워드 중심성 추세를 통해 거시적인 수준에서 기술 동향을 살펴볼 수 있다면, 중심 주제와 연결된 세부 키워드의 연결 상태를 통해 해당 분야에서의 세부 연구 트렌드를 파악 가능하다. 대표적으로 '엣지 컴퓨팅'의 예를 들 수 있다. '엣지 컴퓨팅'이란, 중앙 집중형 '클라우드'의 한계점을 보완하기 위해 등장한 개념으로, 기하급수적으로 늘어나는 데이터 소스의 (말단 기기) 근거리에서 데이터와 연산을 수행하는 분산 컴퓨팅 패러다임의 일종이다. '엣지'라는 키워드는 2019년까지는 네트워크 상에 등장하지 않다가 2020년부터 유의미한 수준으로 언급되기 시작하였다. 해당 키워드는 '모바일', '컴퓨팅', '클라우드'와 같은 상위개념 키워드와 관계를 형성하며 '엣지 모바일', '엣지 클라우드'와 같은 신종 기술용어를 파생시키고 있다. 실제로, NTIS시스템에 '엣지 컴퓨팅'으로 검색 시, 결과로 반환되는 과제 개수가 2017년 31개에서 2020년 274개로 약 9배 이상 증가한 것으로 나타났다. 위의 결과를 종합

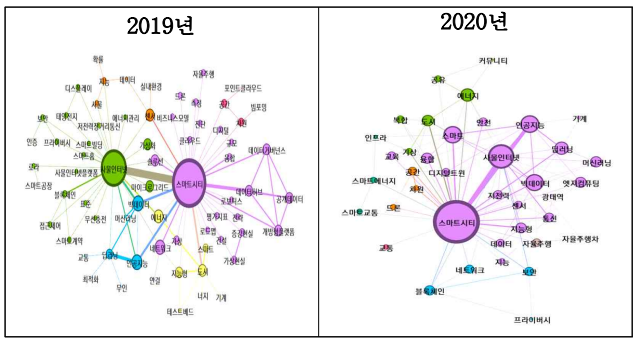
해 볼 때, 네트워크 상에서 새롭게 등장하거나 연결 상태가 변화하는 키워드는 유망 기술일 확률이 높다. 유망기술 후보를 추가적으로 추출하기 위하여 키워드 네트워크를 대조하여 전년도에 등장하지 않은 신규 키워드를 조사하였으며, 결과는 아래 표 2와 같다.

<표 2> 연도별 신규 키워드 목록

	2018-2017	2019-2018	2020-2019
신규 키워드	자율주행	거래 인증	데이터셋
	증강현실	프라이버시	영상분석
	신경망	스트리밍	컨텐츠
	블록체인	엣지	드론

3. Case study: 스마트 시티. 상기 서술한 분석 방법을 스마트시티 분야에 적용한 결과는 다음과 같다. 스마트시티는 도시 인프라의 디지털 전환이 핵심으로, 다양한 센서에서 얻어지는 데이터를 기반으로 지능형 도시를 구현하는 것이 목적이다. 이러한 개념이 네트워크 관점에서도 명확하게 확인 된다 (그림 4). '스마트시티' 키워드가 '사물인터넷'과 가장 강하게 연결되어 있으며, '빅데이터'를 융합의 매개로 하여 '딥러닝', '인공지능'과 같은 지능형 시스템 관련 토픽과 자연스럽게 연결된다.

(그림 4) '스마트시티' 분야 키워드 네트워크



2020년도에는 '사물인터넷'과 '인공지능' 사이의 결합이 더욱 강해지는 것을 볼 수 있다. 새롭게 등장한 '엣지 컴퓨팅' 키워드는 IoT기기에서 발생하는 수많은 데이터를 효율적으로 처리하기 위한 최신 연구 트렌드가 반영된 결과라고 할 수 있다. 네트워크 형성 결과를 바탕으로, 향후 스마트시티 관련 연구 분야에서 '엣지 컴퓨팅', '스마트 + X (에너지, 홈, 교통 등)', '자율주행', '블록체인'과 같은 최신 유망 기술들이 활발히 연구될 것이라고 추측할 수 있다.

5. 결론 및 향후 계획

본 연구에서는 국가 R&D과제 정보에 텍스트 마이닝 및 네트워크 분석 기법을 적용하여 정보통신 분야와 스마트시티의 최근 연구 동향을 살펴보았다. 분석 결과, 최근 ICT 연구는 '빅데이터', '딥러닝', '인공지능' 키워드를 중심으로, '드론', '엣지 컴퓨팅'과 같은 융합 신기술이 새로운 트렌드로 부상하고 있음을 알 수 있었다. 후속 연구로는 네트워크 상에서 주어진 시그널을 바탕으로 보다 정교한 기술수요 예측 및 미래기술 전망을 진행할 계획이다.

"본 연구는 과학기술정보통신부 한국과학기술정보연구원 (K-21-L01-C05-S01)의 지원을 받아 수행되었습니다."

참고문헌

[1] W. Seo, H. Park, and J. Yoon, "단어동시출현분석을 통한 한국의 국가 R&D 연구동향에 관한 탐색적 연구," Journal of Information Technology Applications and Management, vol. 19, no. 4, pp. 1-18, 2012.

[2] 김지훈, 김성섭 "텍스트마이닝을 활용한 농업 R&D 키워드 분석," Journal of the Korea Academia-Industrial cooperation Society v.22 no.2 , pp.721-732, 2021

[3] SO, Jaehyun et al., "A Study on the Concept of Smart City and Smart City Transport," Journal of Korean Society of Transportation, vol. 37, no. 2, pp. 79-91, 2019

[4] 최현홍, 심동녘. "텍스트마이닝을 적용한 ICT융합 트렌드 분석" 혁신학회지, vol. 15, no. 3, pp 257-281, 2020

[5] T.-H. Kim, M.-S. Yang, and K.-N. Choi, "국가R&D정보 활용을 위한 전문용어사전 구축," 한국콘텐츠학회논문지, vol. 19, no. 10, pp. 217 - 225, Oct. 2019.

[6] 이성직, 김한준, "TF-IDF의 변형을 이용한 전자뉴스에서의 키워드 추출 기법," 한국전자거래학회지, vol. 14, no. 4, pp. 59-73, 2009

[7] 김하진, 송민 "동시출현단어 분석을 통한 국내외 정보학 학회지 연구동향 파악," 정보관리학회지, vol. 31, no. 1, pp. 99 - 118, Mar. 2014.

[8] 전익진, 이학연 "연관규칙 기반 동시출현단어 분석을 활용한 기술경영 연구 주제 네트워크 분석," 기술혁신연구, vol. 24, no. 4, pp. 101 - 126, 2016.

[9] Agrawal, R., Imieliński, T. and Swami, A., "Mining Association Rules between Sets of Items in Large Databases", Acm Sigmod Record, Vol. 22, No. 2, pp. 207-216. 1993