토픽 모델링 기반 내용 분석을 통한 학제 간 융합기술 도출 방법

Discovering Interdisciplinary Convergence Technologies Using Content Analysis Technique Based on Topic Modeling

정도헌 (Do-Heon Jeong)*
주황수 (Hwang-Soo Joo)**

초 록

본 연구는 텍스트 마이닝 기법을 활용하여 대량의 데이터로부터 학제 간 융합 기술을 발굴하는 일련의 과정을 제시하는 것을 목표로 한다. 바이오공학 기술(BT) 분야와 정보통신 기술(ICT) 분야 간의 융합 연구를 위해 (1) BT 분야의 기술용어 목록을 작성하여 대량의 학술논문 메타데이터를 수집한 후 (2) 패스파인더 네트워크 척도 알고리즘을 이용해 유망 기술의 지식 구조를 생성하고 (3) 토픽 모델링 기법을 사용하여 BT분야 중심의 내용 분석을 수행하였다. 다음 단계인 BT-ICT 융합 기술 아이템 도출을 위해, (4) BT-ICT 관련 정보를 얻기 위해 BT 기술용어 목록을 상위 개념으로 확장한 후 (5) OpenAPI 서비스를 이용하여 두 분야가 관련된 학술 정보의 메타데이터를 자동 수집하여 (6) BT-ICT 토픽 모델의 내용 분석을 실시하였다. 연구를 통해 첫째, 융합 기술의 발굴을 위해서는 기술 용어 목록의 작성이 중요한 지식 베이스가 된다는 점과 둘째, 대량의 수집 문헌을 분석하기 위해서는 데이터의 차원을 줄여 분석을 용이하게 해주는 텍스트 마이닝 기법이 필요하다는 점을 확인하였다. 본 연구에서 제안한 데이터 처리 및 분석 과정이 학제 간 융합 연구의 가능성이 있는 기술 요소들을 발굴하는 데 효과적이었음을 확인할 수 있었다.

ABSTRACT

The objectives of this study is to present a discovering process of interdisciplinary convergence technology using text mining of big data. For the convergence research of biotechnology(BT) and information communications technology (ICT), the following processes were performed. (1) Collecting sufficient meta data of research articles based on BT terminology list. (2) Generating intellectual structure of emerging technologies by using a Pathfinder network scaling algorithm. (3) Analyzing contents with topic modeling. Next three steps were also used to derive items of BT-ICT convergence technology. (4) Expanding BT terminology list into superior concepts of technology to obtain ICT-related information from BT. (5) Automatically collecting meta data of research articles of two fields by using OpenAPI service. (6) Analyzing contents of BT-ICT topic models. Our study proclaims the following findings. Firstly, terminology list can be an important knowledge base for discovering convergence technologies. Secondly, the analysis of a large quantity of literature requires text mining that facilitates the analysis by reducing the dimension of the data. The methodology we suggest here to process and analyze data is efficient to discover technologies with high possibility of interdisciplinary convergence.

키워드: 융합기술, 유망기술, 지적구조, 토픽 모델링, 내용분석
convergence technology, emerging technology, intellectual structure, topic modeling, content analysis

* 덕성여자대학교 문헌정보학과 조교수(doheonjeong@duksung.ac.kr) (제1저자)
** 덕성여자대학교 바이오공학과 조교수(hwangsoojoo27@duksung.ac.kr) (교신저자)

■ 논문접수일자 : 2018년 8월 18일 ■ 최초심사일자 : 2018년 9월 5일 ■ 게재확정일자 : 2018년 9월 12일
■ 정보관리학회지, 35(3), 77-100, 2018. [http://dx.doi.org/10.3743/KOSIM.2018.35.3.077]

1. 서 론

최근 많은 산업 및 연구 분야에서 효율성과 생산성 향상에 초점을 맞춘 기술 중심의 관점이 더욱 강조됨에 따라 정부와 산업계, 학계 모두가 세상을 바꾸는 혁신적 기술, 즉 와해성 기술 발굴에 박차를 가하고 있다. 와해성 기술(disruptive technology) 또는 와해성 혁신(disruptive innovation)이란, 업계를 완전히 재편성하고 시장 대부분을 점유하게 될 기술, 제품, 또는 서비스를 의미한다(Wikipedia, 2018). 메사추세츠 공과대학(Massachusetts Institute of Technology; MIT)에서는 MIT Technology Review를 통해 매년 10대 혁신 기술(10 Breakthrough Technologies)을 선정하여 최신 동향을 발표하는 등 혁신을 통해 세상을 바꾸려는 이른바 와해성 기술 발굴에 적극적으로 가세하고 있다(MIT, 2015). 대한민국 미래창조과학부 역시 이러한 추세에 발맞추어 2017년 정부의 R&D 사업설명회를 통해 와해성 기술 개발과 같은 혁신적 변화를 모색하는 것을 정책 방향의 핵심으로 발표하였다(The Science Times, 2016). 주된 내용으로 연구자 중심의 연구 지원, 창의적 R&D, 인재 개발과 글로벌 협업 강화 등과 더불어 유망 핵심기술 확보를 강조하고 있다. 이처럼 혁신적 기술이 많은 이익을 창출할 수 있다는 기대와 당면 사회 문제를 해결할 수 있다는 사회적 요구에 의해, 최근 많은 연구 개발이 정부, 산업계, 학계를 중심으로 활발하게 진행되고 있다. 특히 다양한 당면 사회 문제를 해결하기 위해 기술들의 융합을 추진하는 경향도 증가하는 것이 특징적이다(과학기술정책연구원, 2011). 그러나 기술 분야의 빠른 변화와 폭발적 기술 증가 현상 속에서 융합 기술 아이템을 발굴하는 것은 여전히 쉽지 않은 일이며, 데이터의 증가로 인한 기술 분석의 어려움 역시 증가하고 있다. 이러한 이유로 유망 기술이나 유망 융합기술을 발굴하는 과정은 빅데이터 기반의 데이터 중심의 관점으로 변화되고 있으며 많은 연구개발에서 데이터 기반 분석 기법을 적극적으로 도입하고 있다(McKinsey Global Institute, 2011; 산업연구원, 2014).

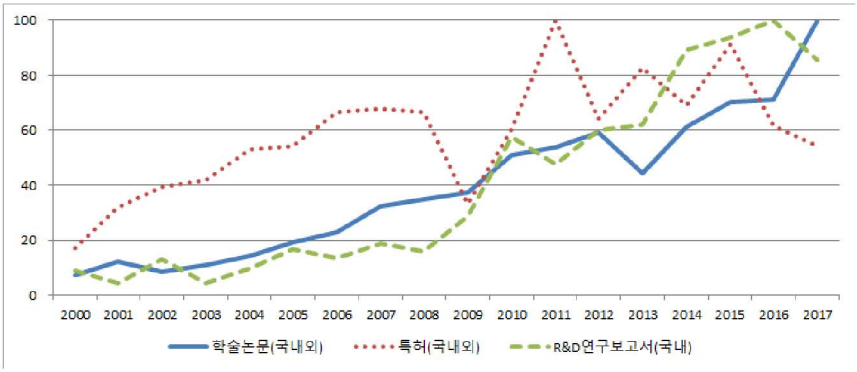
〈표 1〉과 〈그림 1〉은 국내외 융합 기술 관련 연구의 증가 추세를 파악하기 위해 빅데이터 기반 학술정보 시스템을 활용한 분석 결과이다. 국내외 학술논문, 특허, 국가 R&D 연구보고서 등 과학기술 분야의 지식정보 포털서비스를 운영하는 한국과학기술정보연구원(KISTI)의 NDSL 서비스(http://www.ndsl.kr/)를 이용해 'convergence technology'를 키워드로 검색한 통계치는 〈표 1〉에서 보는 바와 같다. 학술논문은 지난 18년간 2만 건 이상 발표되었고, 특허 기술도 상당량이 출원되었음을 알 수 있다. 특히 정부주도의 국가R&D 사업의 결과물인 연구보고서도 거의 1천 건에 달하는 수준이다. 〈표 1〉의 데이터를 기반으로 각 항목의 최대값을 100점으로 환산한 결과는 〈그림 1〉과 같다. 학술논문은 많은 양이 지속적으로 급증하고 있음을 잘 보여주고 있다. 기술 선점이 중요한 특허는 2011년부터 2015년 사이에 많은 양이 출원되었으며, R&D연구보고서는 가장 급격한 증가세를 보이고 있어 최근까지 융합기술에 대한 국가적 관심과 정책적 추진이 매우 활발함을 알 수 있다.

본 연구는 학제 간 연구를 통해 새로운 융합

〈표 1〉 융합 기술 관련 산학연 연구 동향 추세

유 형	합 계 (건)	2000	2001	2002	2003	2004	2005	2006	2007	2008
		2009	2010	2011	2012	2013	2014	2015	2016	2017
학술논문 (국내외)	22,921	229	390	279	353	461	621	736	1,044	1,129
		1,201	1,649	1,731	1,911	1,426	1,975	2,263	2,297	3,226
특허 (국내외)	856	14	26	32	34	43	44	54	55	54
		27	49	81	52	67	56	74	50	44
R&D연구보고서 (국내)	965	12	6	17	6	13	22	18	25	21
		38	76	63	79	82	118	124	132	113

(http://www.ndsl.kr/ 2018년7월31일 검색)



〈그림 1〉 융합 기술 관련 산학연 연구 동향 그래프

(http://www.ndsl.kr/ 2018년7월31일 검색)

기술을 발굴하기 위한 내용 분석 방법을 제안하고 사례를 제시하는 것을 목표로 한다. 특히 학제 간 융합 기술 아이템을 도출하고 내용을 해석하는 일련의 과정에서 발생하는 주요 이슈들을 제시하고자 하였다. 융합대상 학문 분야로는, 사회적 파급력이 큰 분야인 바이오공학 기술(BT) 분야와 정보통신 기술(ICT) 분야로 선정하였으며 전문적 성격이 강한 BT 분야를 중심으로, 활용 및 응용성이 강한 기술 중심의 ICT 분야를 접목하여 융합 가능 아이템의 발굴을 시도하고자 하였다.

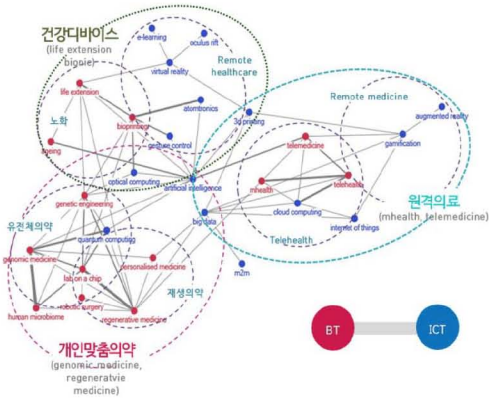
논문의 구성은 다음과 같다. 2장에서는 학제 간 융합기술 도출을 위한 관련 연구와 선행 연구를 제시한다. 유망 기술 융합에 대한 관련 연구 및 동향을 살펴본 후, 내용 분석을 위해 사용하는 잠재 색인 기법인 토픽 모델링에 대한 개념과 관련 연구들을 소개한다. 3장은 크게 2개의 부분으로 구성되어 있다. 1단계에서는 융합 기술 발굴의 시발점이 되는 바이오 과학 분야의 유망 기술을 중심으로 기술용어 목록을 작성한 후, 기술 네트워크를 시각화하여 텍스트 마이닝 기반의 토픽 모델링 결과를 전문가의 지식을 바탕으로 해석하고자 한다. 2단계에서는 대량의 과학기술 학술정보를 자동 연계 수집하여, 토픽 모델링을 기반으로 BT-ICT 두 학문 분야 간의 연관 데이터를 생성하고자 한

다. 4장에서는 데이터 차원이 축소된 토픽 데이터들의 내용 분석을 통해 융합 가능 후보 기술을 도출한 후 관련 연구의 추가 조사를 통해 도출된 기술의 타당성을 살펴본 후, 마지막 5장에서는 연구의 의의와 한계점, 향후 연구에 대해 언급한다.

2. 선행 연구 및 관련 연구

2.1 융합 기술 발굴 연구

과학기술정보통신부 산하 연구기관인 생명공학정책연구센터(Biotech Policy Research Center, http://www.bioin.or.kr/)에서는 2015 바이오 미래 유망 기술 발굴(BioINsay No.2)이라는 보고서 발간을 통해 생명공학 분야의 10대 미래 유망 기술을 발굴하는 기획 연구를 수행하였다. 특히 본 연구를 통해 이전에는 수행되지 않았던 정보통신 분야와의 융합을 통한 바이오헬스 관련 기술을 도출하고자 하였다(<그림 2> 참조). 이 연구 보고서를 통해 창의적 아이디어에 기반한 미래 유망 기술에 대한 R&D 투자는 국가의 미래 성장 동력 확보에 매우 중요하며, 국가 차원에서 지속적인 관심을 가지고 추진이 필요함을 주장하였다. 특히 다양한 분야의 융합을 통한 새로운 기술 창출의 기회(technological opportunity)를 확보해야 한다는 필요성을 강조하면서, BT-ICT 간의 융합 기술 도출의 결과로 '차세대 유전체분석 칩(NGS-on-a-chip)', '체내 이식형 스마트 바이오센서', '유전자 교정세포 3D 프린팅'을 비롯해 '인지/감각기능 증강용 가상현실' 등 10가지의 기술을 미래 유망 기술로 제안하였다. 본 연구는 융합기술을 도출하는 방법을 제시하는 데 있어, 빅데이터 분석 파트와 전문가의 집단지성 활용 파트의 2단계로 구분함으로써, 기존의 전문가 집단 중심의 정책 선정과정에 데이터 중심의 의사결정 지원 과정을 적극 도입하였다는데 큰 의의가 있다.



<그림 2> BT-ICT 융합 네트워크 구조
(생명공학정책연구센터, 2015)

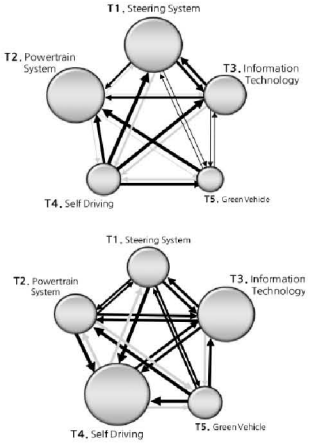
융합 기술 관련 연구들은 세부 기술 분야에서 융합 기술 개발의 사례를 다루는 연구가 대부분이며, 그에 비해 융합 기술을 발굴하는 방법론을 다루는 연구는 상대적으로 많이 부족한 상태이다. 특정 기술 영역에서의 기술 단위의 융합 연구 사례는 매우 많이 존재하며, 순환기 질환의 진단 기술 개발을 위해 X선 영상 기법을 혈류 계측에 적용하는 융합 연구(Jung, Ahn, Nam, Lee, & Lee, 2012), 위성항법시스템의 기존 문제를 해결하기 위한 비전시스템의 결합 방법에 대한 연구(박지호, 권순, 이충희, 정우영, 2011) 등이 선행 기술을 바탕으로 타 분야와 결합하는 기술 융합 연구의 대표적 유형이

라 할 수 있다. 리뷰 형식의 융합 관련 동향 정보를 다루는 연구 역시 다수이며, 시장에 등장한 융합 기술의 동향을 분석하는 연구 유형이다. 진설아와 송민(2016)은 융합과 유사한 학제성을 분석하는데 토픽모델링 기법을 활용하였으며, 특허 네트워크 분석을 통해 특정 분야의 국내외 융합 기술 트렌드를 분석하고자 하는 연구들도 다수 수행된 바 있다(강태규, 박성희, 장일순, 김인수, 한동원, 2009; 백현미, 김명숙, 2013; 산업연구원, 2014).

융합 기술을 발굴하는 방법론에 대한 연구 사례로, 최호창, 곽기영, 김남규(2018)는 이미 발생한 융합 현상을 규명하는 연구에 비해 특정 융합 기술의 향후 전망을 파악하기 위한 연구는 상대적으로 부족함을 언급하면서, 1만 3천여 건의 특허 기술 문서를 이용해 유망 융합 기술의 발굴 방법을 제안하였다. 내용 분석을 위해 LDA 토픽 모델링 기법을 적용하였고 〈그림 3〉에서 보는 바와 같이 토픽 간의 사회 네트워크 중심성(centrality) 분석 기법을 이용해 주요

Period A	In-degree Centrality
Topic1	0.631
Topic2	0.629
Topic3	0.569
Topic4	0.446
Topic5	0.338

Period B	In-degree Centrality
Topic1	0.616
Topic2	0.603
Topic3	0.714
Topic4	0.842
Topic5	0.557



〈그림 3〉 내향 연결정도 중심성(in-degree centrality)과 토픽 네트워크(최호창 외, 2015)

기술 간의 융합 정도를 측정하고자 하였다. 마지막으로 제안한 모형의 성능 평가를 통해 기존 방법론에 비해 우수한 예측 정확도를 보임을 제시하였다. 본 연구는 텍스트 마이닝과 네트워크 분석 기법 등을 사용하여 유망 기술의 융합 아이템을 도출하는 예측 모형을 제시하였다는 점에서 의의가 있다.

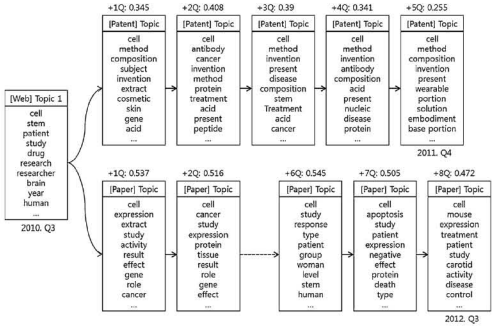
2.2 토픽 모델링 기법

정보 검색을 위한 색인 이론은 Salton & McGill(1983)에 의해 제안된 tf-idf 모델을 기반으로, Deerwester, Dumais, Landauer, Furnas, Harshman(1990)이 제안한 잠재 의미 색인(latent semantic indexing; LSI) 기법으로 발전하였다. 잠재 의미 색인은 특이치 분해(singular value decomposition; SVD) 기법을 바탕으로 용어와 문헌 간의 행렬을 축소된 차원 공간에 표현하고 유사한 문헌이 모이도록 함으로써 문헌의 잠재적 의미 구조를 표현하고자 하였다. 이후 확률적 잠재 의미 색인(probabilistic Latent Semantic Indexing; pLSI) 기법과 같은 확률적 생성 모델 기법들이 소개되었고(Hofmann, 1999), pLSI의 문제점을 보완한 생성 모델인 LDA(Latent Dirichlet allocation) 기법이 제안되었다(Blei, Ng, & Jordan, 2003). LDA는 널리 사용되는 토픽 모델링 기법으로 문서를 구성하는 워드의 집합체인 토픽, 그리고 토픽들의 혼합체(mixture)로 문서를 모델링하는 확률적 생성 모델이다.

Blei, Ng, Jordan(2003)은 LDA 기법을 주로 문서의 내용 분석을 위한 수학적 모델로 제안을 하였으나, 문헌 자체를 군집화 하는 것이

아니라 문헌에 출현한 워드들의 관계를 통해 개념을 모델링하는 특징이 있어 텍스트의 내용 분석을 비롯해 다양한 분야에서 잠재된 패턴을 찾아내고 분석하는 데 활용되고 있다. 육지희와 송민(2018)은 토픽모델링과 딥러닝 기법을 문헌 자동분류에 활용하였다. Farrahi, Gatica-Perez, Gatica-Perez(2012)는 모바일 환경에서 이용자들의 행동 패턴을 찾아내기 위해 토픽 모델링을 활용하였다. 그 밖에도 비디오 내의 얼굴 이미지를 찾아내 핑거프린팅을 하는 기법의 연구(Vretos, Nikolaidis, & Pitas, 2012), 지역 교통 카드 데이터로부터 얻은 버스 이동 경로 정보를 기반으로 탑승자의 이동 행위 패턴과 특징을 찾아내는 연구(조아, 이경희, 조완섭, 2015)도 있었다.

본 연구와 유사하게 계량적 분석에 토픽 모델링 기법을 사용한 연구로, Song과 Kim(2013)은 생물정보학 분야의 지적 구조와 내용 분석을 위해 링크 분석, 토픽 모델링, 페이지랭크, 동시발생용어 분석 등의 다양한 기법을 활용하였다. LDA 기반의 분석결과, 생물정보학 분야에서는 컴퓨터 관련 내용보다는 생물학적인 관점의 토픽이 다수를 차지하고 있음을 확인하였다. Jeong과 Song(2014)은 저널, 프로시딩, 특허, 웹 뉴스 기사와 같이 다양한 정보자원의 내용 분석을 위해 LDA를 사용함으로써 자원별 특징과 기술 용어들의 시간별 토픽 흐름을 파악하였으며, 문서의 내용 분석을 기반으로 각 자원의 생성 시간에 시차가 존재함을 파악하였다. 대량의 다종 자원을 수집하여 텍스트 마이닝 기법을 이용해 자동화된 방식으로 해석하는 시계열 분석 기법을 새롭게 제안하였다(〈그림 4〉 참조).

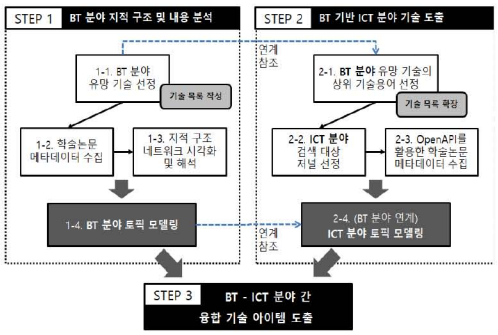


〈그림 4〉 의학 분야 특허-논문 간의 토픽 변화 시계열 분석(Jeong & Song, 2014)

3. BT-ICT 분야 간 융합 기술 도출 방법

본 연구에서 제안하는 BT-ICT 분야 간 융합기술 도출 과정은 〈그림 5〉와 같이 크게 3단계로 구성되어 있다. 첫 번째는 융합의 중심인 BT 분야의 지적 구조와 주요 토픽을 분석하는 단계이다. 우선 BT 분야의 최근 떠오르는 기술을 선정하고 이를 기반으로 기술용어 목록을 작성한다. 학술논문 메타데이터를 수집한 후 네트워크 분석 기법을 이용해 BT 분야의 지적 구조를 분석하고 마지막으로 LDA 토픽 모델링 기법을 이용해서 주요 토픽을 추출하여 내용을 분석한다. 두 번째는 기 작성된 BT 분야의 기술 용어 목록을 이용해 융합의 대상인 ICT 분야의 주요 토픽을 발굴하는 과정이다. 타 분야의 학술문헌을 검색하기 위해 기 작성된 기술용어 목록을 상위 개념으로 확장하는 작업을 수행한 후, 학술정보 포털 시스템이 제공하는 OpenAPI(application programming interface) 서비스를 이용해 ICT 분야의 저널로부터 검색된 데이터를 자동 수집한다. 2단계의 마지막 과

정 역시 LDA 기법을 이용해 ICT 분야의 주요 토픽들을 계산해 낸다. 첫 번째와 두 번째 단계는 본 장에서 서술하며, 마지막 세 번째 단계는 4장에서 다룬다. 1단계와 2단계를 통해 추출된 주요 기술에 대한 토픽들을 전문가의 도메인 지식(domain expertise)을 기반으로 분석하여 최종적으로 융합 기술 후보 아이템들을 도출하는 과정을 설명한다.



〈그림 5〉 BT-ICT 분야 간 융합기술 도출 과정 개요도

3.1 BT 분야 지적 구조 분석 및 토픽 기반 내용 분석

3.1.1 BT 분야 유망 기술 선정 및 기술용어 목록 작성

BT 분야의 지적 구조 분석(〈그림 5〉의 STEP 1)을 위한 첫 과정으로, 최근 떠오르는 BT 유망 기술을 분석 대상으로 선정하고 주요 키워드들을 수집하였다. 2015년부터 2018년까지의 MIT Technology Reviews로부터 BT 분야 기술(http://www.technologyreview.com/), Cell Press Selections의 전체 주제 분야(http://www.cell.com/selections), Science 저널 웹페이지의 Topics 섹션에서 BT 분야 기술, 생명공학정책연구센터에서 발표한 2017 바이오 미래 유망 기술(BioINsay No.14)과 2018 바이오 미래 유망 기술(BioINsay No.27) 등을 기술용어 목록 작성에 활용하였다. 특히, 해외 BT 유망 기술 현황, 주요 저널과 고인용 논문, 전문가/비전문가 설문조사 등 다각적인 방법으로 BT 유망 기술을 선정한 생명공학정책연구센터의 분석 결과를 주로 참고하였다. 생명공학정책연구센터의 2017년 자료는 바이오 뉴스 빅데이터 기반의 이슈대응형으로, 2018년 자료는 미래 파급력을 지닌 연구성과 모니터링을 통한 혁신발견형 기술들로 다른 기준으로 선정되어 주요 기술이 중복되는 경우가 거의 없었으며, 일부 기술들에 있어 연구 범위의 중복이 있기도 하였지만(예: in vivo genome editing과 genome editing therapy) 세부 기술의 적용 범위가 상이하므로 각각의 키워드를 그대로 유지하였다. 주요 기술용어를 확정한 후, 대량의 메타데이터를 수집하기 위해 각 용어마다 관련 어휘로 확장을 수행하였는데, 상기한 자료 외에도 구글의 관련검색(Google Searches related to) 기능과 Keyword Tool 사이트(http://keywordtool.io/google)를 이용하여 추가 확장 키워드를 작성하였다. 관련 검색어를 선정한 예로, 'next generation cancer vaccine'보다 구체적 기술을 다루는 'dendritic cell cancer vaccine'과 일반적인 의미를 담고 있어서 더 많은 수의 논문이 검색될 수 있는 'therapeutic cancer vaccine'를 모두 연관 검색어로 포함하여 확장 검색을 시도하였다. 확장된 유사 키워드 27종을 추가하여 최종 작성된 BT 분야의 유망 기술용어 목록은 〈표 2〉와 같다. 4장에서는 타 분야와의 상호 검색 재현율을 높이기 위해 본 기술용어 목록을 상위 개념

〈표 2〉 BT 분야 유망 기술용어 목록 작성 및 데이터 수집 통계

용어 코드	BT 유망 기술 30종 (대표어)	유사어 키워드 확장 (27종)	수집 문헌수
T01	Genetic remediation	safe genes / genetic baseline states	3,545
T02	Synthetic embryo	artificial embryonic stem cells / artificial embryo	2,091
T03	Single neuron analysis		3,110
T04	Metabolomic reprogramming	metabolic reprogramming	2,507
T05	Membrane proteome structure	membrane proteome map / membrane proteins database	3,420
T06	Glycomimetics		99
T07	in vivo Genome editing		872
T08	Next generation cancer vaccine	therapeutic cancer vaccines / dendritic cell cancer vaccine	7,638
T09	Biomimetics		3,511
T10	Neuroimaging for psychiatry	neuroimaging mental illness	1,128
T11	3D printing drug delivery	customized drug delivery system	426
T12	Open source drug discovery		311
T13	Quantitative traits engineering	quantitative traits genome editing	609
T14	Simultaneous detecting sensor	simultaneous detection sensor	1,029
T15	Plant based vaccine		1,075
T16	Artificial meat	cultured meat	1,176
T17	Carbon photosynthetic cell		981
T18	Artificial enzyme	nanozyme	4,823
T19	Xenobiotics ecosystem	biodegradation of xenobiotics	2,805
T20	Gene drive		172
T21	Clinico genomics	big data applications in genomics / big data human genome	1,706
T22	Single cell genomics	single cell genome sequencing	1,997
T23	Infoepidemiology	epidemiology bioinformatics / public health informatics	17,340
T24	Mobile artificial intelligence diagnosis	machine learning for healthcare applications	1,066
T25	Wearable health devices	personal health monitoring devices	1,766
T26	Genome editing therapy	therapeutic genome editing	3,310
T27	Circulating tumor DNA detection	circulating tumor DNA biomarker	2,863
T28	Continuous glucose monitoring		2,099
T29	in vivo direct reprogramming	in vivo reprogramming	1,574
T30	Epigenetic regulation of development	epigenetics in development	11,468
합 계			86,517

으로 확장한다. T01, T02 등과 같은 용어 코드는 확장된 상위 개념과 연계하기 위한 고유번호이다.

3.1.2 학술 논문 메타데이터 수집과 전처리

작성된 기술용어 목록을 이용해 PubMed 사이트(http://www.ncbi.nlm.nih.gov/pubmed/)로부터 최근 5년간(2013.1월-2018.4월 현재)의 논문을 검색하였으며, 30개 기술에 대해 8만 6천여 건의 논문을 수집하였다. 검색된 논문들은 MEDLINE 형식의 텍스트 파일로 저장하였으며 데이터 구조는 〈그림 6〉과 같다. 기술용어

PMID- 29605641
OWN - NLM
STAT- Publisher
LR - 20180401
IS - 1090-2414 (Electronic)
IS - 0147-6513 (Linking)
VI - 157
DP - 2018 Mar 29
TI - Biochemical mechanism of phytoremediation process of with Mucor circinelloides and Trichoderma asperellum.
PG - 21-28
LID - S0147-6513(18)30236-7 [pii]
LID - 10.1016/j.ecoenv.2018.03.047 [doi]
AB - This study focused on the bioremediation mechanisms 1000mgkg(-1)) and cadmium (0,10,50,100mgkg(-1)) conta indigenous fungi selected from mine tailings as the phyt

생략

PL - Netherlands
TA - Ecotoxicol Environ Saf
JT - Ecotoxicology and environmental safety
JID - 7805381
OTO - NOTNLM
OT - Combined-remediation
OT - Heavy metal
OT - High-throughput sequencing
OT - Microbial community
OT - Plant resistance
OT - Soil pollution
EDAT- 2018/04/02 06:00

〈그림 6〉 Pubmed 메타데이터 샘플 (medline 포맷)

[['epigenet', 'germ', 'cell', 'nuclear', 'factor', 'induc', 'pluripot', 'stem', 'cell', 'somat', 'cell', 'reprogram', 'stem', 'cell'],
['decidua', 'induc', 'pluripot', 'stem', 'cell', 'ipsc', 'mesenchym', 'cell', 'x', 'chromosom', 'inactiv'],
['autophagi', 'breast', 'cancer', 'cancer', 'prevent', 'carcinogen', 'cigarett', 'smoke', 'keton', 'bodi', 'lactat', 'microenviron', 'mitochondri', 'dysfunct', 'prematur', 'ag', 'senesc', 'tumor', 'growth'],
['cancer', 'metabol', 'embryon', 'transcript', 'factor', 'neoplast', 'progress', 'p53', 'lack', 'po2', 'cell', 'cycl', 'tumor', 'reprogram'],
['bioelectr', 'microenviron', 'normal', 'reprogram', 'rest', 'potenti', 'voltag', 'gradient'],
['epigenet', 'mesenchym', 'stem', 'cell', 'neural', 'stem', 'cell', 'plastic', 'regen', 'medicin', 'reprogram'],
['cardiomyocyt', 'cell', 'therapi', 'heart', 'failur', 'induc', 'pluripot', 'stem', 'cell']]

〈그림 7〉 전처리 과정을 거친 문헌 행렬 예시(모든 용어는 스테밍 처리됨)

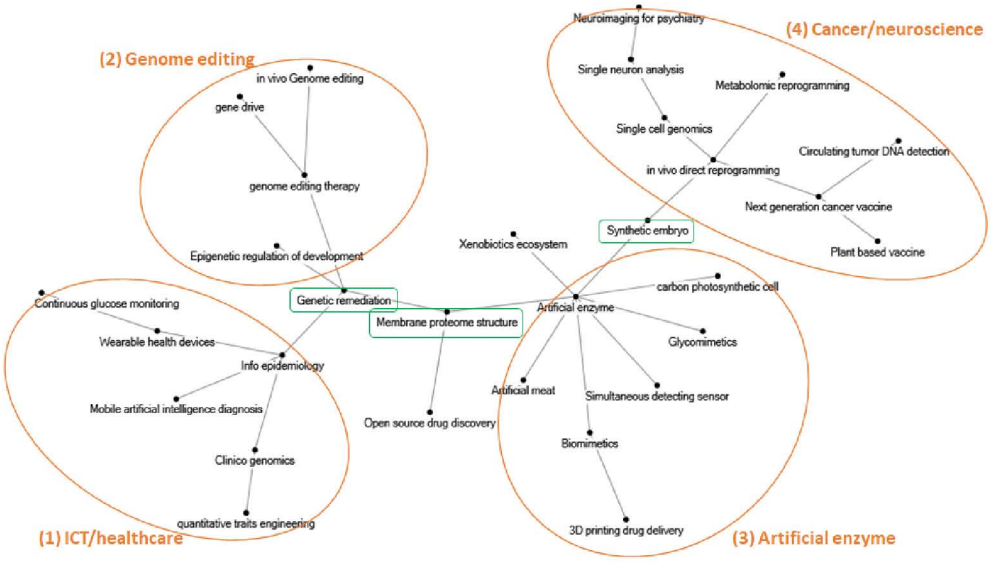
목록을 이용해 수집한 논문 데이터 셋은 일부 문서가 중복되므로 문서의 고유한 식별기호인 PMID로 우선 중복제거를 실시하였다. 각 문서에서 키워드 필드인 OT 항목을 추출하였으며, 하이픈('-') 등의 기호를 포함하여 어절단위로 나누어 전처리를 수행하였다. 〈그림 7〉은 처리된 문서의 최종 전처리 결과이다. 용어들은 텍스트 마이닝 기법을 적용하기 위해 사전 스테밍(stemming) 처리를 하였으며 모든 데이터 처리 과정은 필요한 라이브러리를 활용하여 파이썬(python) 언어로 프로그래밍하여 수행하였다.

3.1.3 BT 분야 유망 기술의 지적 구조 시각화와 해석

본 연구에서는 BT 유망 기술의 지적 구조를 분석하기 위한 기법으로 패스파인더 네트워크(PathFinder Network; PFNet) 척도 알고리즘을 사용하였다(Schvaneveldt, Durso, & Dearholt, 1989). 이 기법은 지적 구조를 세부적으로 표현하기 어려운 다차원 척도법(Multidimensional Scaling; MDS)을 보완하는 네트워크 척도 알고리즘 중 하나로 최근 계량정보학 분야에서 널리 활용되고 있다(정도헌, 2017; Quirin, Cordon, Guerrero-Bote, Vargas-Quesada, & Moya-Anegon, 2008). 본 연구를 수행하기 위해 Quirin 등(2008)에 의해 제안된 MST-PF 알고리즘을 기반으로 한 PFNet 생성 소프트웨어를 직접 개발하여 지적 구조를 생성하였으며, 문헌 간 유사도를 측정하기 위한 방법으로 Salton과 McGill(1983)이 제안한 코사인 유사계수(cosine coefficient)를 사용하였다. 지적 구조 분석을 위한 마지막 준비 단계인 네트워크 시각화를 위

해서는 마이크로소프트 엑셀의 플러그인 형태로 제공되는 NodeXL(http://nodexl.com/)을 이용하였다. 최종 작성된 네트워크 구조는 〈그림 8〉과 같다. 영역별로 지정된 주제 레이블은 바이오공학을 전공한 연구진이 직접 노드의 관계 및 내용 분석을 하여 그룹에 추가 부여한 것이다.

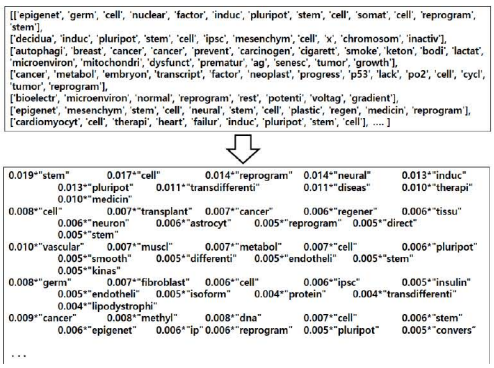
작성된 BT 미래 유망 기술 네트워크는 4개의 주제 그룹과 3개의 허브 기술로 구성되어 있다. 4개의 그룹은 각각 (1) ICT/healthcare 관련 기술, (2) genome editing 관련 기술, (3) 인공 효소 관련 기술, (4) 암/뇌과학 관련 기술로 그룹 (1)과 (2)는 genetic remediation 기술로 연결되고 이는 다시 membrane proteome structure 기술을 통해 그룹 (3)과 연결된다. 끝으로 그룹 (3)과 (4)는 synthetic embryo 기술을 통해 연결되고 있다. 흥미로운 점은 허브노드 역할을 하는 기술들이 모두 2018 바이오 미래 유망 기술(BioINsay No.27)에서는 코어바이오(기반 기술)로 분류되었던 기술들이라는 사실로, 네트워크 분석 결과가 기존에 수행된 전문가 평가 기반의 분석 내용과 부합하는 결과를 보여주었다는 점은 주목할 만하다. 네트워크 상의 특징적인 부분을 살펴보면, 먼저 (1) ICT/healthcare 그룹은 informatics/info epidemiology 기술을 중심으로 ICT 분야와 융합도가 높은 개인용 healthcare 장치 및 보건의료 빅데이터 관리기술로 구성됨을 볼 수 있다. quantitative traits (QT) engineering 기술은 기술적인 측면에서 볼 때는 오히려 그룹 (2) genome editing 기술에 포함되어야 하나, 본 연구에서는 타겟 유전자의 데이터 측면으로 해석되어 그룹 (1)로 분류된 것으로 보인다. 2018 바이오 미래 유망 기술(BioINsay No.27)에서 구분한 후보 기술군 중 저탄소, 친환경 중심의 화이트바이오 기술들은 인공 효소 기술을 중심으로 비교적 잘 모여 있는 것 또한 확인할 수 있었다.



〈그림 8〉 BT 분야 미래 유망 기술 지적 구조와 주제 그룹

3.1.4 BT 분야 토픽 모델링 기반 내용 분석

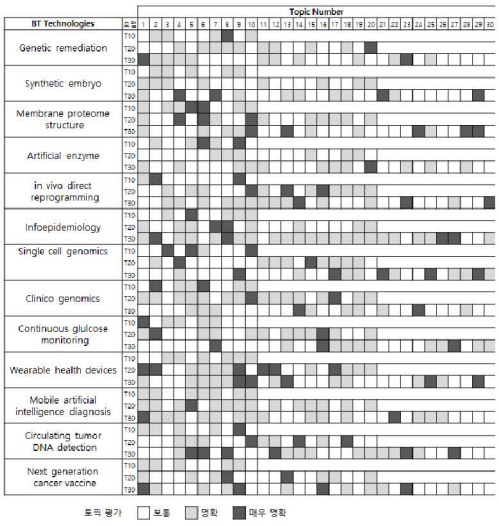
본 연구에서는 토픽 모델링 기법을 구현하기 위해 공개 라이브러리 소프트웨어인 Gensim(http://radimrehurek.com/gensim/)을 활용하였으며 전체적인 실험은 파이썬 프로그래밍 언어로 직접 소프트웨어를 작성하여 수행하였다. Gensim을 사용하기 위해서는 문헌 집단을 2차원의 배열 형태로 작성하여 입력하여야 한다. 〈그림 9〉는 토픽 모델링 과정의 이해를 돕기 위한 실제 입력 데이터와 출력 결과의 예시이다. 앞서 소개한 〈그림 7〉과 같은 키워드 벡터로 구성된 문헌 형식의 2차원 행렬을 입력 받은 후 LDA 기법을 적용하면 데이터 축소 및 추상화를 통해 〈그림 9〉의 하단과 같은 형태의 확률 모델이 생성된다. 그림 예시에서, *"0.019*"stem", 0.017*"cell", 0.014*"reprogram", 0.014*"neural", 0.013*"induc", 0.013*"pluripot", 0.011*"transdifferenti", 0.011*"diseas", 0.010*"therapi", 0.010*"medicin"* 한 줄이 하나의 토픽이며, 토픽은 내용 상 연관성이 높은 단어들로 구성된 하나의 주제로 이해할 수 있다. 토픽 모델링을 실행할 때, 생성하고자 하는 토픽의 수(K)와 토픽을 구성하는 상위 N개의 용어를 지정할 수 있다. 본 연구에서는 K=10, 20, 30, 50, 100으로 지정하여 다수의 토픽을 생성하였으나 실제 도메인 전문가가 직접 내용 해석을 실시하기에 K=50과 K=100 유형은 해석해야 할 토픽 수가 지나치게 많아 해석의 제한을 두었다. 최종적으로 K=10, 20, 30까지 실시하여 각 기술 용어별로 60개의 토픽에 대한 내용 분석을 실시하였다. 이후 K=10, 20, 30의 토픽모델 유형은 각각 T10, T20, T30과 같이 줄여 표기하였다. LDA 토픽 모델링을 위한 파라미터 추정 최적화 알고리즘은 관련 연구를 참고하여 기대치최대화(expectation-maximization; EM) 알고리즘을 사용하였다(Jeong & Song, 2014).



〈그림 9〉 문헌 행렬의 LDA 토픽 모델링 예시(용어는 스테밍 전처리됨)

모든 용어에 대해 토픽 모델링을 실시한 후 토픽 분석을 실시하였다. 우선 앞서 선정한 3가지 유형의 토픽 모델인 T10, T20, T30에 대해 주제의 명확성 여부를 평가하였다. 평가의 대상으로 〈그림 8〉의 BT 유망 기술들 중 4개의 그룹을 서로 연결해주는 코어 바이오 기술 3가지(genetic remediation, membrane proteome structure, synthetic embryo)와 이에 직접 연결된 기술 3가지(infoepidemiology, artificial enzyme, in vivo direct reprogramming)를 포함해 총 13개 유망 기술에 대한 토픽 분석을 수행하였다. 30개의 토픽을 추출한 T30 모델에서 명확히 해석되는 토픽들을 가장 많이 발견할 수 있었지만 몇 가지 중복 되는 경우를 제외하고 10개, 20개, 30개 모든 경우에서 의미가 있는 토픽들을 추출할 수 있었다(〈그림 10〉 참조). 토픽 평가를 통해, T10, T20, T30 유형의 토픽들을 종합적으로 고려하여 분석을 수행하는 것이

융합 후보 기술을 발굴하는 데 유용함을 알 수 있었고 이후의 토픽 생성과 해석 과정 역시 같은 방법으로 진행하였다.



〈그림 10〉 BT 유망 기술의 토픽별 내용 평가 결과(일부 예시)

중심축을 형성하는 6개의 기술들에 대한 상세한 토픽 분석을 수행한 결과, membrane proteome structure의 토픽들 중 질량분석 기술을 이용한 막단백질체 분석(〈그림 11〉의 T20_Topic4, T20_Topic10, T30_Topic13, 파란색 box 표시)은 이웃하는 artificial enzyme의 자가조립 막단백질이나 줄기세포 관련 기술을 통해 다른 중심 기술인 synthetic embryo로 자연스럽게 이어지는 것을 확인할 수 있었다. 또한 membrane proteome structure의 반대편에 위치한 genetic remediation으로는 두 기술에 공통적으로 자주 등장하는 제 2형 당뇨병(〈그림 11〉의 T10_Topic5, T20_Topic6, T30_Topic28, 빨간색 box 표시)을 키워드로 연결할 수 있었다. 이 외에도 제 2형 당뇨병과 직접 연관이 되는 continuous glucose monitoring 뿐만 아니라 in vivo direct reprogramming과 같은 또 다른 핵심 기술에서도 제 2형 당뇨병과 관련된 토픽을 두루 발견할 수 있어 BT 유망 기술에서 제2형 당뇨병이 이슈가 되고 있음을 파악할 수 있었다. 다음 장에서는 본 장에서 수행한 방법을 토대로, BT 분야에서 해석된 토픽들을 ICT 분야로부터 해석된 토픽과 상호 연결하는 방안에 대해 설명하고자 한다.

3.2 BT-IT 학제 간 융합 가능 기술 도출

3.2.1 BT 분야 상위어 선정을 통한 기술용어 목록 확장

BT-ICT 융합 기술 발굴을 위해서는 상호 연계 데이터 추출 가능성을 확인하기 위한 사전 실험이 필요하였다. 학술정보 포털 서비스인 NDSL 검색 시스템을 활용하여 BT의 기술용어를 ICT 저널을 대상으로 직접 검색을 수행하였다. 분석이 될 만큼의 충분한 검색 결과가 나올 것을 예상하였으나, 30개의 BT 유망 기술명을 검색어로 사용한 경우에는 ICT 분야의 논문이 거의 검색되지 않는 결과를 보여주었다. 특정 분야의 전문적이고 세부적인 기술명을 사용하여 타 분야를 검색한 것이 결과의 원인이었으므로, 융합 데이터를 확보하기 위해서는 BT 분야의 대표 유망 기술을 다시 일반적인 개념의 상위 용어로 그룹화하여 확장하였다.

BT 상위 기술들은 앞서 BT 분야의 기술용어 목록 작성을 위해 기 수집된 관련 자료를 활용하여 분야 전문가의 도메인 지식을 바탕으로 추가 작성하였다. 우선 총 26개의 BT 상위 기

Membrane proteome structure (T10)

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
cancer	diseas	cancer	cell	diabet	proteom	protein	cell	cancer	protein
tumor	alzheim	breast	receptor	type 2	protein	membran	receptor	receptor	ms
factor	cancer	cell	epitop	meta	membran	proteom	lung	cell	gene
protein	gene	protein	protein	analysi	cancer	cell	small	target	cancer
analysi	cognit	diseas	express	protein	diseas	analysi	cancer	analysi	hla
receptor	amyloid	polymorph	molecular	cell	surviv	prognosi	erlotinib	antibodi	therapi
growth	beta	gene	sequenc	polymorph	interact	outer	antimicrobi	breast	carcinoma

Membrane proteome structure (T20)

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
neck	proteom	protein	transport	gene	diabet	pharmacophor	diseas	cell	protein
head	protein	diseas	protein	proteom	type 2	virtual	gene	plant	membran
squamou	drug	profil	cardiovascular	express	receptor	screen	proteom	analysi	spectrometri
protein	analysi	proteom	specif	nephropathi	mellitu	molecular	restrict	cancer	mass
meta	cell	gene	prostat	biomark	glp	cancer	vaccinolog	meta	alzheim
carcinoma	stem	syndrom	spectrometri	profil	agonist	dock	cancer	disord	structur
review	system	antipsychot	mass	cell	sglt2	qsar	chromatographi	factor	proteom
Topic11	**Topic12**	**Topic13**	**Topic14**	**Topic15**	**Topic16**	**Topic17**	**Topic18**	**Topic19**	**Topic20**
singl	arthriti	therapi	gene	etanercept	cell	hla	cancer	protein	breast
nucleotid	factor	preterm	function	adalimumab	analysi	frequenc	receptor	antigen	cancer
polymorph	rheumatoid	prognosi	pathwai	analysi	alk	analysi	growth	structur	her2
snp	growth	birth	analysi	psoriasi	polymorph	allel	antibodi	membran	metastat
diseas	cancer	fetal	arteri	advers	meta	haplotyp	cell	mirna	surviv
receptor	protein	fibronectin	magnet	cancer	rearrang	cancer	epiderm	molecular	posit
analysi	membran	nmda	hypertens	event	lymphoma	mutat	egfr	mutat	neg

Membrane proteome structure (T30)

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	Topic7	Topic8	Topic9	Topic10
proteom	outer	effect	diseas	cell	virtual	immunoglobulin	cancer	singl	alzheim
carcinoma	breast	glp	mild	pharmacogenet	screen	cell	lung	nucleotid	diseas
protein	carcinoma	advers	protein	tumour	pharmacophor	killer	growth	polymorph	amyloid
subcellular	membran	protein	data	motif	dock	transduct	factor	carcinoma	snp
cancer	polymorph	agonist	analysi	epidemiolog	molecular	mitochondri	receptor	gene	beta
prognosi	cancer	studi	proteom	antagonist	epitop	signal	egfr	analysi	cognit
erythropoietin	cholesterol	therapi	gener	grave	qsar	zoonot	small	squamou	biomark
Topic11	**Topic12**	**Topic13**	**Topic14**	**Topic15**	**Topic16**	**Topic17**	**Topic18**	**Topic19**	**Topic20**
arthriti	vascular	breast	loop	meta	genet	tgfbeta	genom	tyrosin	receptor
anti	emiss	her2	antibodi	polymorph	mitochondrion	coupl	secret	prostat	motif
tnf	tomographi	posit	model	arthriti	hla	system	tube	receptor	calcium
psoriat	positron	cancer	engin	depress	trial	carcinoma	protein	kinas	surfac
treatment	protein	sequenc	structur	transport	featur	surviv	xanthin	transcript	gene
cytokin	endotheli	mass	stem	rheumatoid	inhibitor	receptor	electrophysiolog	diseas	trp
inflamm	genet	neg	compar	analysi	clinic	protein	antiphospholipid	secret	prolifer
Topic21	**Topic22**	**Topic23**	**Topic24**	**Topic25**	**Topic26**	**Topic27**	**Topic28**	**Topic29**	**Topic30**
cancer	breast	portug	protein	analysi	human	membran	diabet	protein	protein
proteom	cancer	databas	meta	podocyt	protein	antigen	type 2	diseas	gene
cell	human	azor	analysi	biologi	genom	cancer	mellitu	locu	interact
follicl	metastat	diseas	polymorph	target	cancer	protein	glucos	missens	express
itraq	cost	promot	membran	young	miss	signal	express	mutat	insert
prolifer	her2	glucagon	prognosi	phenotyp	proteom	proteom	gene	analysi	bind
plaqu	immunodefici	allel	popul	cancer	languag	glycom	diseas	famili	polymorph

〈그림 11〉 기술명 membrane proteome structure의 토픽 해석 예시(토픽별 상위 7개 용어만 표기)

술을 선정할 수 있었고, 그 중 성격이 유사한 기술들은 같은 그룹으로 묶어 총 16개의 상위 기술 그룹을 최종 선정하였다(〈표 3〉 참조). 〈표 2〉의 BT 분야 유망 기술 목록은 각 용어에 T로 시작하는 용어코드를 부여하고 있으며 확장된 상위 용어인 〈표 3〉에서는 G로 시작하는 용어코드를 각기 부여하고 있다. 또한 하위 기술용어와의 연계 항목에 표기를 하여 상호 참조가 가능하도록 하였다.

3.2.2 BT 기반 ICT 분야의 메타데이터 수집과 토픽 모델링

본 연구에서는 두 가지 방식으로 데이터를 수집한다. 첫 번째는 정보 시스템을 이용해 검색 결과를 일괄 다운로드 받는 방식이며, 두 번째는 프로그래밍을 통해 실시간 검색 서비스를 활용하는 방식이다. 앞서 생명공학 분야의 지적 구조를 파악하고 주요 토픽을 해석하는 과정에서는 기술용어 목록을 바탕으로 Pubmed

〈표 5〉 ICT 융합을 위해 상위 개념으로 확장된 BT 유망 기술 목록

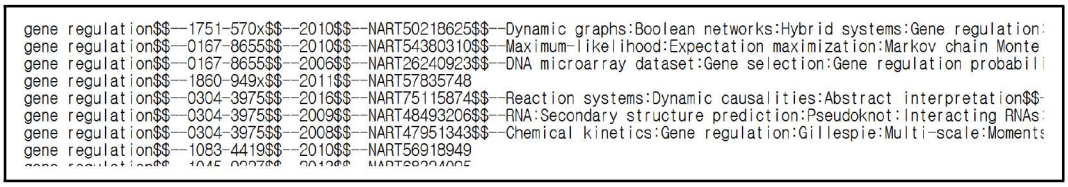
용어 코드	BT 유망기술의 상위 기술명 *(굵은 글씨체가 대표 용어임)*	대표 용어의 한글 기술명	하위 기술용어 연계코드
G01	**Gene expression analysis** / gene expression profiling	유전자 발현 분석	T01, T03, T27, T30
G02	**Gene ontology**	유전자 온톨로지	T23
G03	**Gene regulation**	유전자 조절	T30
G04	**Genetic engineering**	유전 공학	T01, T07, T13, T18, T20, T26, T29
G05	**Genome sequencing** / high-throughput sequencing / next-generation sequencing	유전체 서열분석	T21, T22
G06	**Metabolic engineering**	대사 공학	T04, T06, T17, T19
G07	**Mhealth** / telehealth / telemedicine / e-medicine	모바일 헬스	T14, T24, T25, T28
G08	**Microfluidics** / microfluidic device	미세유체기술	T03, T22
G09	**Morphogenesis**	형태 발생	T02, T07, T29
G10	**Prosthesis** / robotic arm / robotic surgery	보철	T09
G11	**Protein design** / protein folding	단백질 설계	T05, T18
G12	**Genomics**	유전체학	T21, T22
G13	**Cancer immunotherapy**	암 면역치료	T08
G14	**Neuro technology**	뇌신경 기술	T03, T10
G15	**Pharmaceutical discovery**	신약 개발	T08, T11, T12, T15
G16	**Tissue engineering**	조직 공학	T09, T16

사이트를 이용해 직접 검색한 메타데이터를 일괄(batch) 방식으로 다운로드 받아 처리하였다(〈그림 6〉 참조). 이 장에서는 두 번째 수집 방식인 서비스 프로그래밍 방식으로 데이터를 수집하는 OpenAPI 기반의 방식을 사용한다. 자동화를 기반으로 효과적인 대용량 데이터의 검색 및 처리가 가능한 방식이지만, 대량의 학술정보를 제공하는 시스템이 필요하며 실시간 서비스를 제공하는 OpenAPI 서비스가 구축되어 있어야 제한 없이 활용이 가능하다. 본 연구에서는 한국과학기술정보연구원(KISTI)가 서비스하는 NOS(NDSL Open Service, http://nos.ndsl.kr)를 이용하였다. 단, OpenAPI 서비스를 이용하면 서버를 통한 실시간 검색 기능만을 수행할 수 있기 때문에 검색 결과 데이터를 저장 및 처리하거나 데이터간 매핑 작업을 위한 일련의 소프트웨어는 별도의 추가 개발이 필요하였다.

OpenAPI를 활용해 ICT 분야의 학술 메타데이터를 자동 수집하기 위해 우선 관련 저널 정보를 수집하였다. KISTI NDSL 서비스의 저널브라우징 기능을 통해 얻은 생물정보학(bioinformatics) 저널 52종을 포함한 ICT 분야의 저널을 수집하였다. 2000년 이후 최근까지 발행되고 있는 저널을 중심으로 선정하여, 최종 수집된 융합 대상 저널은 1,058종 이었다. 〈표 4〉는 저널 브라우징을 통해 수집한 해당 저널 정보의 주요 데이터 항목이다. 〈그림 12〉는 저널 정보를 이용해 NOS 서비스를 활용한 용어의 실시간 매칭 결과 예시이다. 추출된 데이터는 특정 기술 용어에 대한 매칭된 논문의 저널정보, 문헌번호, 키워드 등으로 구성되어 있

〈표 4〉 수집한 정보통신기술 및 생물정보학 저널 목록 중 일부 예시

저널명	발행기관	창간년	언어	주제분야	ISSN
Current bioinformatics	Bentham Science Pub	2006	eng	570:793	1574-8936
Applied bioinformatics	Open Mind Journals	2002	eng	570:793	1175-5636
Evolutionary bioinformatics	Libertas Academica	2005	eng	570.285	1176-9343
Trends in bioinformatics	Asian Network for Scientific Info.	2008	eng	570.285	1994-7941
Open access bioinformatics	Dove Medical Press	2009	eng	570.285	1179-2701
Advances in bioinformatics	Hindawi Pub. Corp	2008	eng	570.283	1687-8035 1687-8027
Briefings in bioinformatics	H. Stewart Publications	2000	eng	570.285	1477-4054 1467-5463



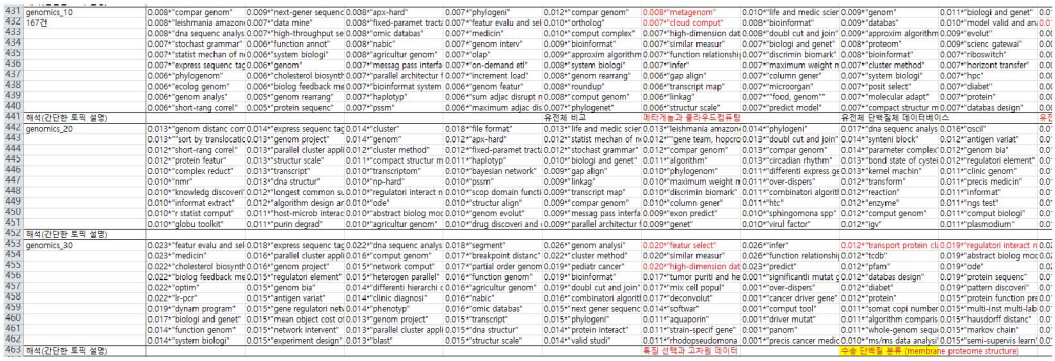
〈그림 12〉 빅데이터 기반 OpenAPI를 활용한 분야 간 연계 데이터 수집 결과(일부)

다. 이 데이터를 활용해 내용 분석을 위한 2차 토픽 모델링을 실시한다.

앞서 언급한 바와 같이 최초의 BT와 ICT 두 분야 간의 유망 기술 용어 매칭은 유효한 수준의 데이터가 수집되지 않았기 때문에 BT 분야의 30개 대표 유망 기술을 다시 일반적인 개념의 16개 상위어로 기술용어 목록을 확장하여 ICT 대상 저널에 재 수집을 실시하였다. 수집 결과, 16개 중 10개의 용어가 매칭에 성공하였으며 총 600건의 문헌이 수집되었다(〈표 5〉 참조). 문헌 수가 너무 적은 경우 중복 토픽이 지나치게 발생하여 토픽 모델링이 용이하지 않으므로 10건 이하로 수집된 용어를 제외한 결과이다. 수집된 용어와 문헌의 통계는 〈표 5〉와 같다. 광범위하게 수집된 수만 건의 BT 문헌 데이터에 비하면 양이 적지만 적정한 양의 문헌이 검색된 용어에 대해서는 토픽을 전수 해석한 결과, 충분히 활용이 가능한 수준이라 판단되었다(〈그림 13〉 참조). BT-ICT 연계 토픽에 대한 상세한 해석은 다음 장인 융합 기술 아이템 도출 과정에서 상세히 다루고자 한다.

〈표 5〉 확장된 기술용어 목록(BT)을 이용한 ICT 저널 매칭 결과

용어 코드	BT 상위 용어 (대표어만 기재)	ICT 저널 매칭 문헌수
G01	gene expression analysis	27
G02	gene ontology	59
G03	gene regulation	35
G05	genome sequencing	121
G07	mhealth	73
G08	microfluidics	14
G09	morphogenesis	30
G10	prosthesis	42
G11	protein design	32
G12	genomics	167
합 계		600



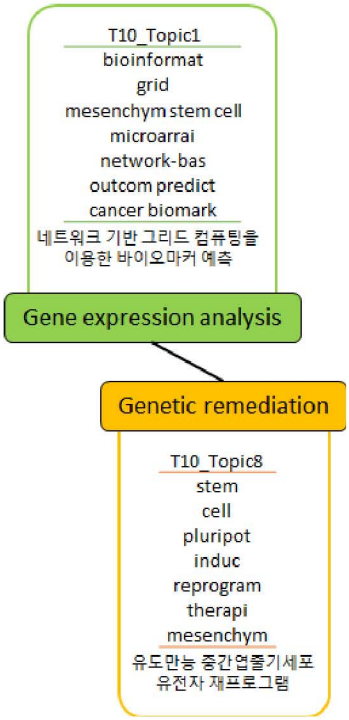
〈그림 13〉 BT-ICT 매칭을 통해 수집된 문헌의 토픽 해석 및 평가 수행 예시

4. BT-ICT 학제 간 융합 기술 아이템 도출

본 장에서는 학문 분야의 도메인 지식(domain expertise)을 기반으로, BT와 ICT 분야의 주요 기술 토픽들을 분석함으로써 융합 가능 기술들을 도출하는 과정을 소개한다. 토픽 해석에 있어, 앞서 〈표 5〉에서 언급한 바와 같이 일부 BT 기술에 대해서는 BT-ICT 간 연관 논문의 수가 적어서 T20, T30 토픽 모델에서 일부 중복 토픽이 발견되었지만 ICT의 세부 기술 단위에서 BT와의 연관성을 분석함에 있어 문제는 없었다. 본 연구에서 제안한 융합기술 발굴 방법은 통해 도출할 수 있었던 대표 사례를 소개하고자 한다.

4.1 고성능 그리드 컴퓨팅 기술 기반 줄기세포 암 유발 인자 예측과 줄기세포 치료제 개발

첫 번째는 중간엽 줄기세포 치료제의 안전성 확보를 위한 암 관련 바이오마커 발굴에 그리드 기반의 고성능 컴퓨팅 기술을 융합한 사례이다(〈그림 14〉 참조). 최근 급성이식편대숙주병 소아 환자에 대한 임상3상 성공으로 최초의 중간

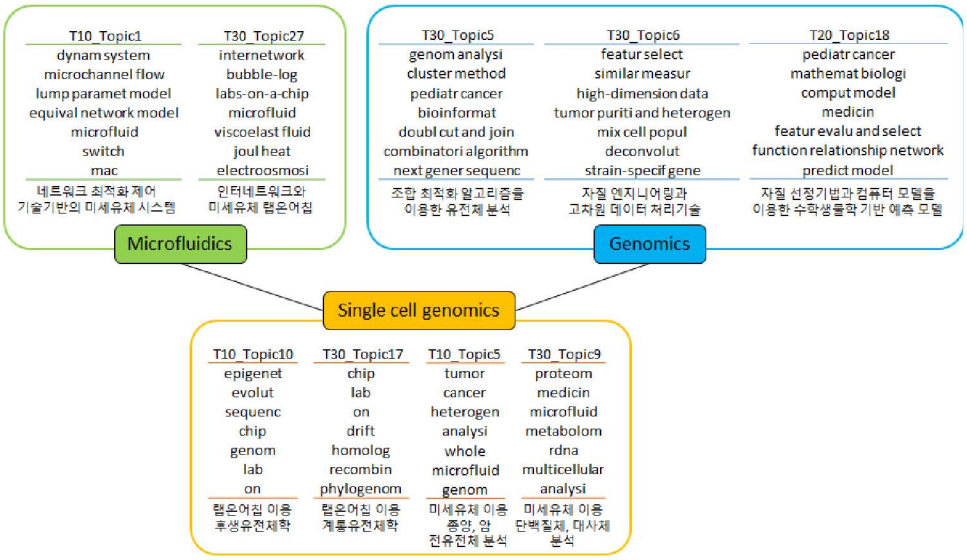


〈그림 14〉 유전자 발현 분석과 유전자 교정 간의 토픽 매칭

엽 줄기세포 치료제의 FDA 승인 전망을 밝게 하고 있음에도 불구하고(Galipeau & Sensebe, 2018), 중간엽 줄기세포와 암과의 관련성에 대한 의문은 지속적으로 제기되고 있는 상황이다(Lee & Hong, 2017; Ridge, Sullivan, & Glynn, 2017). 이에 네트워크 기반의 그리드 컴퓨팅(grid computing) 기술을 채용한 고성능 컴퓨팅 환경을 암 바이오마커 예측에 활용하여 중간엽 줄기세포의 유전자 교정을 위한 치료제로서의 안전성 향상을 도모할 수 있을 것으로 보인다.

4.2 미세유체 조작기술과 고차원 데이터 조합 최적화 기술을 이용한 단일세포 유전체학 및 암 연구

두 번째 기술 융합 사례는, 서로 다른 두 기술인 미세유체학(microfluidics)과 유전체학(genomics)이 단일세포 유전체학(single cell genomics)을 중심으로 융합되는 경우이다(〈그림 15〉 참조). 미세유체 기술은 마이크로미터 단위의 미세한 유로의 정밀 제작이 필수적인 분야로 반도체 공정에서 주로 사용되는 광식각(photolithography) 기술이 사용되는 등(Whitesides, Ostuni, Takayama, Jiang, & Ingber, 2001) ICT 기술과 연관이 깊으며, 마이크로미터 단위의 크기(microarray)를 갖는 단일세포를 조작하고 분석하는데 필요한 플랫폼을 제공할 수 있다는 장점도 가지고 있다. 동시에 단일 세포 유전체학에서 요구되는 대량의 다양한 유전자 변이와 같은 고차원 데이터를 효과적으로 분석하기 위해 조합 최적화 알고리즘(combinatorial algorithm)과 같은 IT 기술을 도입할 수 있을 것으로 보인다. 이와 관련하여 Prakadan, Shalek, Weitz(2017)와 Caen, Lu, Nizard, Taly(2017)가 언급한 것처럼 미세

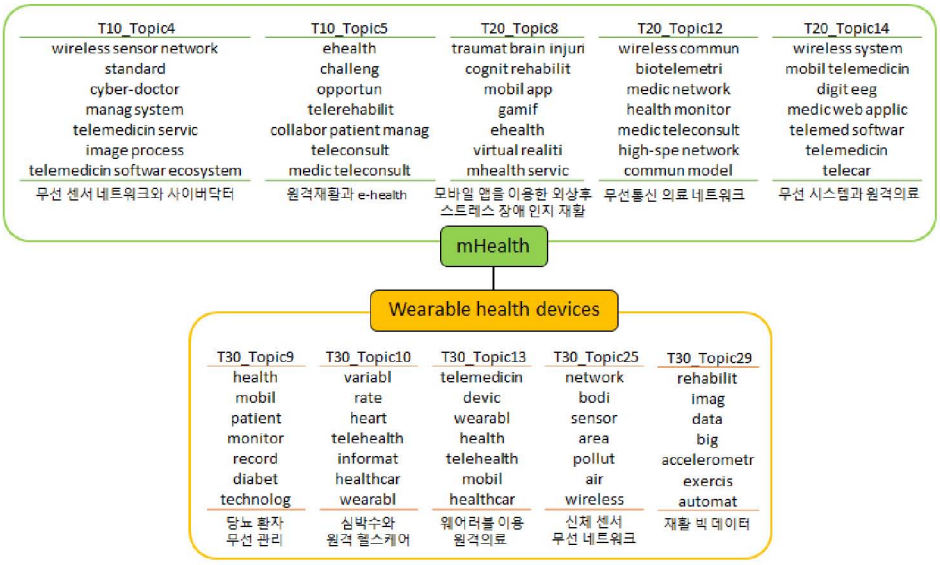


〈그림 15〉 미세유체학, 유전체학, 단일세포 유전체학의 토픽 매칭

유체학 융합 기술이 분석 속도와 비용 면에서도 큰 장점을 가지고 있어 단일 세포 오믹스에 광범위하게 응용되기 시작하고 있음을 확인할 수 있었다.

4.3 무선 네트워크를 통한 원격 의료 모바일헬스 기술과 웨어러블 건강 보조 장치를 통한 환자 관리 및 재활 지원, 의료기술 보완

웨어러블 건강 보조장치(wearable health devices)와 모바일헬스(mhealth) 기술을 중심으로 하는 융합 기술은 ICT와의 융합 성격이 가장 큰 사례라 할 수 있다(〈그림 16〉 참조). 모바일헬스 기술에서는 주로 무선 네트워크를 통한 원격 의료에 관한 토픽들로 구성되어 있으며, 웨어러블 건강 보조장치 미래 유망 기술에서는 웨어러블과 무선 기술을 이용한 원격 의료, 재활 등과 관련된 토픽들이 분석되었다. 많은 토픽들이 무선 기술과 웨어러블을 주제로 하여 분석되어 이 융합 기술 분야가 현재에도 가장 활발하게 연구되고 있는 분야 중 하나임을 확인할 수 있었다. 세계적인 비즈니스 컨설팅 기업인 Gartner는 2015년에는 웨어러블 기술이 아직 임상 시험을 포함한 모바일헬스 생태계에 적용하기 위한 준비가 되지 않았다고 보고하였으나(Gartner, 2015), 불과 3년만에 인공지능 기술 발달과 데이터의 질적 양적 증가에 힘입어 웨어러블 기술이 헬스케어 생태계의 열쇠를 쥐게 되었다고 재평가하기도 하였다(Gartner, 2018). 전세계 웨어러블 기기의 2017년 300억 달러 시장 규모와 17%의 성장률(Gartner, 2017), 모바일헬스 분야의 2017년 230억 달러 시장 규모와 35%의 성장률(Research and Markets, 2017) 등의 경제 지표 또한 이를 뒷받침하고 있다.



〈그림 16〉 모바일헬스와 웨어러블 건강 보조장치의 토픽 매칭

5. 결 론

본 연구는 학제 간 연구를 통해 새로운 융합 기술을 발굴하기 위한 내용 분석 방법을 제안하고 도출된 기술 사례를 제시하는 것을 목표로 하였다. 융합을 위해 선정한 두 개의 학문 분야는 연구 및 산업 시장에서 큰 자본 시장을 가지며 사회적 파급력이 큰 분야인 바이오공학 기술 분야와 정보통신 기술 분야로 선정하였으며, 학문의 특성상 특정성이 강한 BT 분야를 중심으로, 범용적이며 응용성이 강한 기술 특성을 갖는 IT 분야를 접목하고자 하였다.

연구 수행 과정을 크게 세 단계로 요약하면, 첫 번째 단계로 BT분야의 유망 기술을 선정하여 기술 중심의 지적 구조를 파악한 후 상세 내용 분석을 실시하였다. 미래 유망 기술의 용어 목록을 작성하여 대량의 학술논문 메타데이터를 수집한 후, 패스파인더 네트워크 알고리즘을 이용해 지식 구조를 시각화하였다. 이를 통해 전체적인 유망 기술의 위상 구조와 연결성을 파악하였다. 대량의 데이터의 차원을 줄임으로써 효과적으로 내용을 분석하기 위해 LDA 토픽 모델링 기법을 사용하였다. 두 번째 단계에서는 OpenAPI 서비스를 이용하여 BT와 ICT 분야가 동시에 등장하는 학술논문 메타데이터를 자동으로 수집하였다. 1단계에서 작성한 BT 유망 기술을 이용한 검색 시, 상세 분야의 전문 용어를 검색 키워드로 삼은 이유로 인해 검색 결과가 거의 도출되지 않는 문제가 발생하였다. 이를 해결하는 방법으로, 유망 기술들의 상위어를 재선정하여 기술용어 목록을 확장한 후 ICT 분야와 연결함으로써 학제 간 해석이 가능하도록 하였다. 마지막 단계에서, BT와 ICT 분야가 접목되는 융합 기술의 후보 아이템을 찾기 위해 토픽 모델의 내용 분석을 실시하였다. 연구를 통해 발굴한 기술들에 대한 현황 조사를 실시한 결과, 본 연구에서 제안한 여러 단계의 데이터 처리 및 분석 기법 적용 과정이 학제 간 융합 연구가 가능한 최종 아이템들을 발굴하는 데 효과적이었음을 확인할 수 있었다.

본 연구의 몇 가지 기여점을 정리하면 다음과 같다. 첫째, 기술융합을 위한 지식베이스로 활용할 BT 분야의 유망 기술의 목록을 작성하였다. 특정 용어를 대표어로 통제하였고, 검색 재현율을 보장하기 위한 유사어 확장 및 상위 개념어를 추가 작성하였다. 작성된 기술용어 목록은 향후 관련 연구 및 응용 연구에 활용될 수 있다. 둘째, 기술 융합의 사례를 발굴하는 과정에서 필연적으로 나타나는 데이터 희소(data sparseness) 문제를 해결하는 방안을 제시하였다. 융합기술 후보를 도출함에 있어 가장 큰 이슈는 한 분야의 세부 기술이 다른 분야에서는 거의 나타나지 않아 초기 실험을 위한 테스트 문헌을 확보할 수 없는 데이터 부족 문제였다. 본 연구에서는 이를 해결하기 위한 방법으로, 상위 개념으로 기술 용어를 보다 일반화하여 타 분야와 연관된 데이터를 추출한 후 토픽 모델링을 통해 세부 내용을 해석하는 방식을 선택하였다. 셋째, 문헌 분석 시 키워드 기반의 단순 분석이 아닌 대량의 다차원 데이터를 함축하여 표현할 수 있는 텍스트 마이닝 기법을 적용하였다는 점이다. LDA 기법을 통해 방대한 데이터를 도메인 전문가가 전수 해석할 수 있는 수의 토픽으로 축소할 수 있었다. 따라서 도메인 지식을 가진 분석가가 심층적인 내용 분석을 하는 데 매우 효과적이었다.

학제 간 융합기술 발굴 방법을 제안함에 있어 본 연구의 한계점과 향후 연구 계획은 다음과 같다. 우선 본 연구에서는 도출된 기술의 활용 사례를 다루지 않고 있으므로 향후 추가 연구를 통해 제안된 방법론이 실용적으로 사용될 수 있는 활용 시나리오를 검토할 필요가 있다. 또한 서로 다른 영역 간의 융합 기술을 발굴하는 과정에서의 주요 이슈와 도출된 결과의 해석을 주로 다루었기 때문에, 아직 자동화된 프로세스 및 모델을 제시하지 못하고 있다. 향후 연구 수행을 통해 다양한 학문 기술 분야로부터 나타나는 유형의 패턴화와 성능 평가가 필요할 것이며, 대량의 데이터로부터 주요 연구 이슈들을 효과적으로 발굴하고 매칭하기 위해 다양한 텍스트 마이닝 기법 연구가 필요할 것으로 보인다.

참 고 문 헌

강태규, 박성희, 장일순, 김인수, 한동원 (2009). 녹색성장 LED 융합 기술 동향 분석. 전자통신동향분석, 24(5), 30-37.

과학기술정책연구원 (2011). 사회문제 해결을 위한 과학기술-인문사회 융합방안. STEPI 정책연구 2011-14.

박지호, 권순, 이충희, 정우영 (2011). 위성항법시스템과 비전시스템 융합 기술 기반의 신뢰성 있는 위치 측위에 관한 연구. 전자공학회논문지, TC48(10), 20-28.

백현미, 김명숙 (2013). 특허 네트워크 분석을 통한 융합 기술 트렌드 분석: 한국・미국・유럽・일본의 특허데이터를 중심으로. 벤처창업연구, 8(2), 11-19.

산업연구원 (2014). 한국의 기술융합 발전 트렌드 및 융합기술개발 결정요인 분석. 산업연구원, 연구보고서 2014-709.

생명공학정책연구센터 (2015). 2015 바이오 미래유망기술 - ICT융합 바이오헬스 10대 미래유망기술 -. BioINsay No.2(총서 제223권)

생명공학정책연구센터 (2017). 2017 바이오 미래유망기술 - 바이오헬스 이슈를 선도하는 10대 미래유망기술 -. BioINsay No.14(총서 제242권)

생명공학정책연구센터 (2018). 2018 바이오 미래유망기술 - Core, Red, Green, White Bio로 살펴본 10대 미래유망기술 -. BioInsay No.27(총서 제261권)

육지희, 송민 (2018). 토픽모델링과 딥 러닝을 활용한 생의학 문헌 자동 분류 기법 연구. 정보관리학회지, 35(2), 63-88. http://dx.doi.org/10.3743/KOSIM.2018.35.2.063

정도헌 (2017). 자동 분류 기법과 지적 구조 분석 기법을 융합한 처방적 분석 시스템 구현 방안 연구. 정보관리학회지, 34(4), 33-57. http://dx.doi.org/10.3743/KOSIM.2017.34.4.033

조아, 이경희, 조완섭 (2015). LDA 기법을 이용한 버스 승객의 잠재적 이동패턴 분석. 한국데이터정보과학회지, 26(5), 1061-1069. http://dx.doi.org/10.7465/jkdi.2015.26.5.1061

진설아, 송민 (2016). 토픽 모델링 기반 정보학 분야 학술지의 학제성 측정 연구. 정보관리학회지, 33(1), 7-32. http://dx.doi.org/10.3743/KOSIM.2016.33.1.007

최호창, 곽기영, 김남규 (2018). 기술 성숙도 및 의존도의 네트워크 분석을 통한 유망 융합 기술 발굴 방법론. 지능정보연구, 24(1), 101-124. http://dx.doi.org/10.13088/jiis.2018.24.1.101

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

Caen, O., Lu, H., Nizard, P., & Taly, V. (2017). Microfluidics as a strategic player to decipher single-cell omics?, Trends in Biotechnology, 35(8), 713-727. http://doi.org/10.1016/j.tibtech.2017.05.004

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society of Information Science, 41(6), 391-407.

Farrahi, K., Gatica-Perez, D., & Gatica-Perez, D. (2012). Extracting mobile behavioral patterns with the distant N-gram topic model. In Proceedings of the 16th International Symposium on Wearable Computers (ISWC), 1-8. http://doi.org/10.1109/ISWC.2012.20

Galipeau, J., & Sensebe, L. (2018). Mesenchymal stromal cells: Clinical challenges and therapeutic opportunities. Cell Stem Cell, 22(6), 824-833. http://doi.org/10.1016/j.stem.2018.05.004

Gartner (2015). Are wearables fit for clincal trials?. smarter with gartner(2015.10.15.). Retrieved from http://www.gartner.com/smarterwithgartner/are-wearables-fit-for-clinical-trials/

Gartner (2017). Gartner says worldwide wearable device sales to grow 17 percent in 2017. newsroom press release(2017.08.24.). Retrieved from http://www.gartner.com/en/newsroom/press-releases/2017-08-24-gartner-says-worldwide-wearable-device-sales-to-grow-17-percent-in-2017

Gartner (2018). Wearables hold the key to connected health monitoring. smarter with gartner (2018.03.08.) Retrieved from http://www.gartner.com/smarterwithgartner/wearables-hold-the-key-to-connected-health-monitoring/

Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval, 50-57.

Jeong, D.H., & Song, M. (2014). Time gap analysis by the topic model-based temporal technique. Journal of Informetrics, 8(3), 776-790. http://dx.doi.org/10.1016/j.joi.2014.07.005

Jung, S.Y., Ahn, S., Nam, K.H., Lee, J.P., & Lee, S.J. (2012). In vivo measurements of blood flow in a rat using X-ray imaging technique. The International Journal of Cardiovascular Imaging, 28(2), 1853-1858. http://doi.org/10.1007/s10554-012-0029-1

Lee, H.Y., & Hong, I.S. (2017). Double-edged sword of mesenchymal stem cells: Cancer-promoting versus therapeutic potential. Cancer Science, 108(10), 1939-1946. http://doi.org/10.1111/cas.13334

McKinsey Global Institute (2011). Big data: The next frontier for innovation, competition, and productivity. Retrieved from http://www.mckinsey.com/business-functions/digital-mckinsey/our-insights/big-data-the-next-frontier-for-innovation

MIT (2015). MIT technology review, 10 breakthrough technologies, 2015. Retrieved from http://www.technologyreview.com/lists/technologies/2015/

Prakadan, S.M., Shalek, A.K., & Weitz, D.A. (2017). Scaling by shrinking: empowering single-cell 'omics' with microfluidic devices. Nature Reviews Genetics, 18(6), 345-361. http://doi.org/10.1038/nrg.2017.15

Quirin, A., Cordon, O., Guerrero-Bote, V.P., Vargas-Quesada, B., & Moya-Anegon, F. (2008). A quick MST-Based algorithm to obtain pathfinder networks(∞, n-1). Journal of the American Society for Information Science and Technology, 59(12), 1912-1924. http://doi.org/10.1002/asi.20904

Research and Markets (2017). mHealth (Mobile Healthcare) Ecosystem Market: 2017-2030- $23 Billion Opportunities, Challenges, Strategies & Forecasts. Globe Newswire(2017.03.02.). Retrieved from http://globenewswire.com/news-release/2017/03/02/930109/0/en/mHealth-Mobile-Healthcare-Ecosystem-Market-2017-2030-23-Billion-Opportunities-Challenges-Strategies-Forecasts.html

Ridge, S.M., Sullivan, F.J., & Glynn, S.A. (2017). Mesenchymal stem cells: key players in cancer progression. Molecular Cancer, 16(31). http://doi.org/10.1186/s12943-017-0597-8

Salton, G., & McGill, M.J. (1983). Introduction to Modern Information Retrieval. McGraw-Hill (NY).

Schvaneveldt, R.W., Durso, F.T., & Dearholt, D.W. (1989). Network structures in proximity data. In G. Bower(Ed.), The psychology of learning and motivation: Advances in research and theory, 24, 249-284. New York: Academic Press.

Song, M., & Kim, S.Y. (2013). Detecting the knowledge structure of bioinformatics by mining

full-text collections. Scientometrics, 96(1), 183-201.
http://doi.org/10.1007/s11192-012-0900-9

The Science Times (2016). '와해성 기술'이 내년 R&D 이끈다(2016.12.14.). Retrieved from http://www.sciencetimes.co.kr/?news=%EC%99%80%ED%95%B4%EC%84%B1-%EA%B8%B0%EC%88%A0%EC%9D%B4-%EB%82%B4%EB%85%84-rd-%EC%9D%B4%EB%81%88%EB%8B%A4

Vretos, N., Nikolaidis, N., & Pitas, I. (2012). Video fingerprinting using latent dirichlet allocation and facial images. Pattern Recognition, 45(7), 2489-2498.
http://doi.org/10.1016/j.patcog.2011.12.022

Whitesides, G.M., Ostuni, E., Takayama, S., Jiang, X., & Ingber, D.E. (2001). Soft lithography in biology and biochemistry. Annual Review of Biomedical Engineering, 3, 335-373.
http://doi.org/10.1146/annurev.bioeng.3.1.335

Wikipedia (2018). Disruptive Innovation. Retrieved from http://en.wikipedia.org/wiki/Disruptive_innovation

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Baek, Hyun Mi, & Kim, Myung Seuk (2013). Technological convergence trend through patent network analysis: focusing on patent data in Korea, U.S., europe, and Japan. Asia-Pacific Journal of Business Venturing and Entrepreneurship, 8(2), 11-19.

Biotech Policy Research Center (2015). 2015 Discovering future emerging biotechnologies - ICT-converged biohealth top 10 future emerging biotechnologies -. BioINsay No.2(Series No.223)

Biotech Policy Research Center (2017). 2017 Future emerging biotechnologies - top 10 future emerging technologies leading biohealth issues -. BioINsay No.14(Series No.242)

Biotech Policy Research Center (2018). 2018 Future emerging biotechnologies - top 10 future emerging technologies from the aspects of core, red, green, white bio -. BioInsay No.27 (Series No.261)

Cho, Ah, Lee, Kyung Hee, & Cho, Wan Sup (2015). Latent mobility pattern analysis of bus passengers with LDA. Journal of the Korean Data & Information Science Society, 26(5), 1061-1069. http://dx.doi.org/10.7465/jkdi.2015.26.5.1061

Choi, Hochang, Kwahk, Kee-Young, & Kim, Namgyu (2018). Discovering promising convergence technologies using network analysis of maturity and dependency of technology. Journal

of Intelligence and Information Systems, 24(1), 101-124. http://dx.doi.org/10.13088/jiis.2018.24.1.101

Jeong, Do-Heon (2017). Prescriptive analytics system design fusing automatic classification method and intellectual structure analysis method. Journal of the Korean Society for Information Management, 34(4), 33-57. http://dx.doi.org/10.3743/KOSIM.2017.34.4.033

Jin, Seol A, & Song, Min (2016). Topic modeling based interdisciplinarity measurement in the informatics related journals. Journal of the Korean Society for Information Management, 33(1), 7-32. http://dx.doi.org/10.3743/KOSIM.2016.33.1.007

Kang, T.G., Park, S.H., Jang, I.S., Kim, I.S., & Han, D.W. (2009). The convergence technology analysis of green growth led illumination. Electronics and telecommunications trends, 24(5), 30-37.

Korea Institute for Industrial Economics & Trade (2014). An analysis on the trends and determinants of technology convergence of Korea. R&D Report 2014-709.

Park, Chi-Ho, Kwon, Soon, Lee, Chung-Hee, & Jung, Woo-Young (2011). A study of a reliable positioning based on technology convergence of a satellite navigation system and a vision system. Journal of the Institute of Electronics and Information Engineers, TC48(10), 20-28.

STEPI (2011) Science, technology and society studies for societal challenges. STEPI Policy Research 2011-14.

Yuk, JeeHee, & Song, Min (2018). A study of research on methods of automated biomedical document classification using topic modeling and deep learning. Journal of the Korean Society for Information Management, 35(2), 63-88. http://dx.doi.org/10.3743/KOSIM.2018.35.2.063