토픽 성장 분석을 통한 오픈액세스 분야 연구 동향 분석*

Understanding Research Trends of Open Access via Topic Growth Analysis

정재민 (Jaemin Chung)**
김완종 (Wan Jong Kim)***

초 록

전통적인 학술 커뮤니케이션 체제의 문제점을 해결하기 위한 대안으로 오픈액세스 패러다임에 대한 국제적 관심과 확산이 지속되고 있다. 하지만 데이터 기반의 정량적인 방법을 통해 오픈액세스 분야의 글로벌한 동향이나 성장 추세를 파악하려는 노력은 아직까지 부족한 실정이다. 본 연구는 오픈액세스 분야의 학술논문 데이터에 토픽 모델링을 적용하여 세부 연구토픽을 식별하고, 성장곡선을 적합하여 각 연구토픽의 성숙도와 예상 잔여수명을 계산한다. 본 연구는 오픈 사이언스의 세 가지 핵심요소인 오픈액세스, 오픈데이터, 오픈협업과 관련된 14개 토픽들을 식별하였으며, 오픈액세스 분야가 앞으로 약 65년간 꾸준히 성장할 것으로 예상하였다. 본 연구의 분석 결과는 연구자들과 정책 의사결정자들이 오픈액세스 분야의 동향과 성장 추세를 이해하는 데 도움을 줄 수 있을 것으로 기대된다.

ABSTRACT

To solve the problems of the traditional scholarly communication system, global interest in the open access paradigm continues. Nevertheless, there is still a lack of research to understand global research and growth trends in the field of open access through data-based quantitative methods. This study aims to identify which sub-fields exist in open access and analyze how long each research field will grow in the future. To this end, topic modeling and growth curve analysis were applied to global academic papers in the field of open access. This study identified 14 research topics related to open access, open data, and open collaboration, which are three key elements of open science, and foresaw that the field of open access will grow over the next 65 years. The results of this study are expected to support researchers and policymakers in understanding global research trends of open access.

키워드: 오픈액세스, 오픈 사이언스, 오픈데이터, 오픈협업, 토픽 모델링, 성장곡선 분석
open access, open science, open data, open collaboration, topic modeling, growth curve analysis

* 본 연구는 2022년 한국과학기술정보연구원(KISTI)의 기본사업 과제로 수행되었음.
** 한국과학기술정보연구원 오픈액세스센터 AccessON개발팀 연구원(jmchung@kisti.re.kr) (제1저자)
*** 한국과학기술정보연구원 오픈액세스센터 AccessON개발팀 팀장(wjkim@kisti.re.kr) (교신저자)
■ 논문접수일자 : 2022년 11월 14일 ■ 최초심사일자 : 2022년 12월 5일 ■ 게재확정일자 : 2022년 12월 9일
■ 정보관리학회지, 39(4), 75-97, 2022. http://dx.doi.org/10.3743/KOSIM.2022.39.4.075

1. 서 론

정보기술의 발전과 전자 저널의 확산에 따라 기존 학술 생태계의 문제를 해결하기 위해 새로운 학술 커뮤니케이션 모델인 오픈액세스 패러다임이 등장했다(Craig et al., 2007). 전통적으로 학술 생태계는 논문에 대한 독점권을 가진 상업출판사들이 수익성을 목적으로 이끌어 오고 있었다. 하지만 출판사들의 상업화가 심해짐에 따라 연구를 수행하는 국가, 기관, 대학, 도서관, 개인이 학술 자료에 접근하기 위한 벽이 점차 높아지고 있다. 최근에는 이러한 문제를 해결하기 위한 대안으로 오픈액세스가 관심을 받고 있다. 오픈액세스는 금전적, 제도적, 기술적 장벽 없이 합법적인 목적으로 저작물을 자유롭게 이용할 수 있도록 허용하는 것을 의미하며(윤희윤, 김신영, 2007), 오픈액세스 학술지에 출판을 장려하는 골드 오픈액세스(Gold Open Access) 모델과 출판물을 리포지터리에 기탁하는 그린 오픈액세스(Green Open Access) 모델을 통해 실천된다(김선겸 외, 2019; Costa & Leite, 2016). 저자가 이와 같은 오픈액세스 모델을 기반으로 연구결과를 공개하면 언제 어디서든 누구나 해당 학술 자료를 무료로 열람하고 활용할 수 있다. 따라서 오픈액세스는 정보의 접근성 향상 및 격차 해소, 연구자의 영향력 확대 등 학술 생태계에 다양한 이점을 가져다주는 학술 커뮤니케이션 모델이다(최재황, 조현양, 2005). 이러한 장점 때문에 최근에는 미국 보건복지부의 Public Access Policy, 유럽연합 집행위원회의 Horizon 2020 및 Plan S 등의 정책 시행이나 오픈액세스 전환계약 체결과 같이 정부 혹은 기관 수준에서 오픈액세스 문화를 정착시키고자 하는 노력도 이루어지고 있다.

오픈액세스에 대한 관심과 확산에 따라서 오픈액세스의 학술적 영향력에 관한 연구, 이해관계자들의 인식과 행동에 대한 조사분석 연구, 도메인별 리포지터리 및 데이터베이스에 관한 연구 등 오픈액세스에 대한 연구가 활발하게 수행되고 있다(Craig et al., 2007; Palmer, Dill, & Christie, 2008; Van Santen et al., 2019). 이와 더불어 대량의 논문 데이터에 텍스트 마이닝을 적용하여 오픈액세스 분야에 대해 분석하려는 시도도 등장하였다. 서선경과 정은경(2013), 김선겸 외(2019)는 오픈액세스 분야의 지적구조를 분석하는 연구를 수행하였다. 두 연구는 각각 1998년 1월 1일부터 2012년 7월 31일까지, 2013년 1월 1일부터 2018년 11월 31일까지의 Web of Science 문헌정보학 카테고리 내 오픈액세스 관련 문헌들에 동시출현단어 분석 기법을 적용하여 오픈액세스 분야의 지적구조를 제시하고 주제영역이 어떻게 구성되었는지를 확인하였다. 또한 Cho(2020)는 Web of Science에서 수집한 2005년부터 2019년까지의 오픈액세스 관련 문헌들을 3년 주기로 나누어 오픈액세스 연구의 지적구조 변화를 분석하였다. 신주은과 김성희(2021)는 KCI(Korea Citation Index)와 RISS(Research Information Sharing Service)를 통해 수집한 국내 오픈액세스 관련 연구물을 통해 국내 연구 동향 및 지적구조를 분석하였으며, 김판준(2021)은 LISTA(Library, Information Science and Technology Abstract) 데이터베이스에서 추출한 키워드를 통해 국외 오픈액세스 분야의 키워드 유형별 특성에 대해 분석하였다.

이러한 선행연구에도 불구하고 데이터 분석

기반의 오픈액세스 분야에 관한 연구에는 여전히 한계점이 존재한다. 첫째, 선행연구들은 분석 대상이 되는 데이터를 문헌정보학 분야 혹은 국내 연구로 한정하여 수집하였다. 이에 따라 선행연구의 분석 결과는 전반적인 오픈액세스 동향이 아니라 특정 분야 혹은 특정 국가에 한정된 인사이트만을 제공한다. 둘째, 데이터는 과거의 이력을 담고 있는 동시에 미래를 내다볼 수 있는 기반이 된다. 하지만 일부 선행연구들은 정적인 관점에서 오픈액세스 분야의 과거 연구 동향과 지적구조를 분석하는 데 그쳤으며 오픈액세스 분야의 성장이나 미래를 관찰하고자 하는 연구는 수행된 바 없다. 따라서 본 연구는 선행연구가 가지는 다음 두 가지 한계점을 연구문제로 제기한다.

- 연구문제 1: 오픈액세스 내에 어떤 세부 연구분야가 존재하는가?
- 연구문제 2: 오픈액세스와 그 세부 연구분야는 앞으로 얼마나 더 성장할 것인가?

위 연구문제에 답하기 위해 본 연구는 데이터 수집 범위를 넓혀 전 세계의 오픈액세스 관련 학술논문 데이터를 수집하고, 토픽 모델링 기법을 적용하여 오픈액세스 분야의 세부 연구토픽을 식별한 후, 각 토픽의 연도별 기여도 누적 합을 성장곡선에 적합하여 연구토픽들의 성장 정도와 잔여수명 등을 예상해본다.

본 연구는 오픈액세스와 그 세부 분야의 진화 및 성장 추세를 동적으로 분석하고자 한 초기 시도라는 기여점을 가진다. 본 연구는 오픈액세스 내에 어떤 세부 분야가 연구되고 있는지를 식별하고 각 분야의 성숙도와 예상 잔여수명 등에 대한 분석을 수행한다. 또한, 본 연구는 분석 대상이 되는 데이터의 수집 범위를 넓힘으로써 다양한 학문 분야에서 오픈액세스를 바라보는 관점에 대한 인사이트를 제공할 수 있다. 마지막으로 본 연구는 데이터 기반의 체계적인 방법론을 적용하기 때문에 오픈액세스 분야에 대한 객관적이고 복합적인 이해를 제공한다. 따라서 본 연구는 오픈액세스 관련 연구자들이나 정책 결정자들의 의사결정을 지원할 수 있는 결과를 제공할 것으로 기대된다.

본 논문의 구성은 다음과 같다. 제2장에서는 본 연구에서 활용하는 분석 기법에 관한 선행연구를 짚어본다. 제3장에서는 본 연구의 분석 데이터와 연구에서 활용하는 방법론을 소개한다. 제4장에서는 제3장의 데이터 및 방법론에 따른 분석 내용 및 결과를 제시한다. 마지막으로 제5장에서는 본 연구의 결론 및 추후연구에 관해 기술한다.

2. 선행연구

2.1 토픽 모델링

토픽 모델링은 키워드를 의미론적으로 군집화하여 문서 집합 내에 존재하는 잠재적인 주제를 식별하는 통계적 텍스트 마이닝 기법이다(Ko et al., 2017). 문서는 키워드의 집합으로 이루어져 있으며 같은 문서에서 자주 함께 등장하는 키워드들은 하나의 주제를 이루고 하나의 문서는 하나 이상의 주제를 포함한다(Jeong, Yoon, & Lee, 2019). 이러한 관점에서 토픽 모델링은 대량의 문서 집합에 등장한 키워드의 분

포를 분석하여 각 문서가 어떤 주제에 해당하는지와 각 주제가 어떤 키워드들로 구성되는지를 식별할 수 있다(Chung et al., 2021). 주로 잠재 의미 분석(Latent Semantic Analysis, LSA), 확률적 잠재 의미 분석(Probablistic Latent Semantic Analysis, pLSA), 잠재 디리클레 할당(Latent Dirichlet Allocation, LDA) 등의 기법들이 토픽 모델링에 사용되며(Blei, Ng, & Jordan, 2003; Deerwester et al., 1990; Hofmann, 1999), 최근에는 언어 모델을 기반으로 개발된 토픽 모델링 기법들도 다양하게 제시되고 있다(Bianchi et al., 2020; Dieng, Ruiz, & Blei, 2020).

대량의 문서 집합에 숨겨진 주제를 식별할 수 있다는 장점 때문에 분석 대상이 되는 도메인의 세부 주제를 파악하기 위해 토픽 모델링을 활용하는 연구들이 다수 수행되었다. Ji와 Cha(2021)는 학술 커뮤니케이션 분야의 논문 데이터에 토픽 모델링을 적용하여 19개의 토픽을 식별한 뒤 각 토픽의 증감 추세를 기반으로 상향세인 토픽을 조사하고 토픽 네트워크 분석을 통해 중심성이 높은 주요 토픽을 분석하였다. Ma et al.(2018)은 논문 데이터에 토픽 모델링을 적용한 결과와 기관별 보유 논문의 행렬 연산을 통해 연구기관의 경쟁력을 계산하는 방법을 제시하고, 컴퓨터 공학 분야에 제시한 방법을 적용하여 각 기관의 연구 강점과 약점을 분석하는 사례연구를 진행하였다. 송성전과 심지영(2022)은 도서 리뷰 데이터를 수집하여 토픽 모델링을 수행하고 그 결과에 층화 샘플링을 적용하여 도서 추천에 영향을 미치는 7개 범주 내 90개 선호요인 관련 개념과 선호요인의 양상을 식별 및 파악하였다. 이러한 선행연구들은 토픽 모델링을 단순히 주제 식별에만 활용하고 그친 것이 아니라 토픽 모델링의 결과물들을 변형하여 분석거나 다른 기법 및 이론에 접목하여 더욱 심도있는 인사이트를 제공하였다.

본 연구는 토픽 모델링을 적용하여 오픈액세스 분야의 연구 동향을 분석한다. 본 연구는 토픽 모델링의 결과물인 토픽-키워드 분포 행렬을 통해 오픈액세스 분야 내에 어떤 세부적인 연구토픽이 있는지를 식별하고, 문서-토픽 분포 행렬에서 연도별로 각 연구토픽의 기여도를 누적 합하여 성장곡선 분석에 활용한다.

2.2 성장곡선

성장곡선은 생물의 개체군 크기가 시간이 지남에 따라 S자 형태의 곡선을 따른다는 사실에서부터 발전되었다(Elvers et al., 2016). 개체군의 크기는 생성 초반에 느린 속도로 증가하다가 급격하게 성장한 뒤에 다시 성장 추세가 완만해지는 과정을 거쳐 성장을 멈춘다(Yoon et al., 2018). 이러한 S자 곡선의 특성을 반영하기 위해 로지스틱 모델(Logistic model), 곰페르츠 모델(Gompertz model), 배스 모델(Bass model) 등 다양한 성장곡선 모델이 제시되었으며(Young, 1993), 미생물이나 질병의 확산과 성장을 모델링하는 연구에서 널리 활용되었다(Berger, 1981; Zwietering et al., 1990).

최근에는 특허 혹은 논문 데이터를 기반으로 특정 도메인의 진화와 성장 추세를 분석하기 위해 성장곡선을 활용한 연구들이 등장하였다. Yoon et al.(2018)은 특허 데이터를 기반으로 전자 피부 기술의 진화 추세를 파악하였다. 위

연구는 전자 피부 기술 내 세부 기술 분야의 특허 수를 로지스틱 모델에 적용하여 각 분야의 성숙도, 기대 잔존수명 등을 계산 및 분석하였다. Braun, Schubert, Kostoff(2000)와 Du et al.(2019)은 각각 풀러린 분야, 도시 기반시설 분야의 논문 데이터를 수집한 뒤 연도별 논문 수 혹은 피인용 수를 성장곡선에 적용하여 각 분야의 성장 단계를 파악하였다. Adamuthe와 Thampi(2019)와 Cho와 Daim(2016)은 각각 계산 기술과 OLED 기술 분야의 논문과 특허 데이터를 수집하여 동향을 파악하였다. 이 두 연구는 연도별 논문 수와 특허 수를 기반으로 연구개발 수준을 파악하거나 두 성장곡선의 시간 차이를 통해 연구와 기술개발의 시간 차이를 분석하는 데 성장곡선을 활용하였다.

이처럼 성장곡선을 활용하면 S자 곡선의 기울기와 변곡점을 기반으로 특정 분야의 성장 단계를 정성적으로 판단할 수 있으며 수리적 계산을 통해 해당 분야의 성숙도나 잔여수명을 예상해볼 수 있다(Yoon et al., 2014). 본 연구는 오픈액세스와 그 세부 연구 분야의 성장 추세를 분석하기 위해 성장곡선을 활용한다. 따라서 본 연구는 각 세부 분야가 어느 정도 성장했는지, 어느 정도 더 많은 연구가 진행될 것인지와 같은 결과를 전망할 수 있을 것이다.

3. 데이터 및 방법론

3.1 데이터

본 연구는 오픈액세스 관련 논문 데이터를 수집한다. 수집 프로세스는 분석 도메인과 관련된 검색 조건을 설정한 뒤 Web of Science와 같은 논문 데이터베이스를 통해 검색 및 수집하는 과정을 거친다. 수집된 데이터는 분석에 필요한 정보인 논문의 제목, 초록, 주제어, 게재연도 등을 포함한다. 다만 검색 조건에 따라 수집된 데이터는 분석 대상 도메인과 관련되지 않은 논문을 포함하고 있을 수 있다. 따라서 정성적인 판단을 기반으로 분석에 사용할 논문을 선별하는 작업이 필요하다.

논문 데이터는 비정형 데이터인 텍스트 정보를 담고 있으므로 데이터 선별이 완료된 뒤에 전처리 과정을 거쳐 데이터를 문서-키워드 행렬 형태로 정형화한다. 우선 텍스트 마이닝 툴을 활용하여 각 논문에서 키워드를 추출한다. 추출된 키워드에는 분석에 노이즈로 작용할 수 있는 키워드가 포함될 수 있으므로 글자 수가 적어 명확한 용어 구분이 어려운 키워드, 등장빈도가 너무 높거나 너무 적은 키워드, “method”, “result” 등 너무 일반적인 키워드 등 분석자의 기준에 따라 불용어를 제거한다. 키워드 선별이 완료되면 최종적으로 분석 대상이 되는 유효 키워드 셋이 구축된다. 그다음 문서별 유효 키워드를 기반으로 각 문서를 벡터 형태로 나타낸다. 본 연구에서는 TF-IDF(Term Frequency-Inverse Document Frequency) 기반의 벡터 공간 모델(Vector Space Model, VSM)을 활용한다. VSM은 텍스트 문서를 단어 벡터로 정형화하는 방법으로 벡터의 요소들은 각 단어의 가중치를 의미하며, TF-IDF는 문서 내 중요한 단어에 높은 중요도를 부여하는 가중치이다(Sidorova et al., 2008). TF-IDF는 〈수식 1〉과 같이 한 문서 내 단어의 등장빈도에는 비례하지만, 해당 단어를 포함하고 있는 문서의 수에는 비례

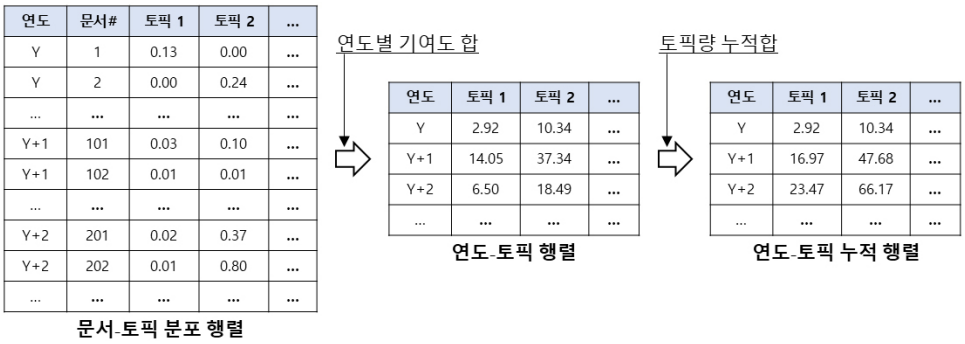
하지 않기 때문에 특정 문서 안에서의 중요도를 나타낼 수 있다는 특징을 가진다. 수식에서 w_{ij}는 문서 j에서 단어 i가 가지는 TF-IDF 가중치이며, tf_{ij}는 단어 i가 문서 j에서 등장한 빈도를, df_i는 단어 i를 포함하고 있는 문서의 수를, N은 전체 문서의 수를 의미한다. 최종적으로 모든 문서를 TF-IDF 기반의 단어 벡터로 표현하여 문서-단어 행렬을 구축한다.

$$w_{ij} = tf_{ij} \times \ln\left(\frac{N}{1+df_i}\right) \qquad \langle \text{수식 } 1 \rangle$$

3.2 LDA

본 연구는 오픈액세스 분야의 연구토픽을 식별하기 위해 가장 널리 활용되는 토픽 모델링 기법인 LDA를 활용한다. LDA는 대량의 문서 처리와 잠재 토픽의 해석에 있어 타 토픽 모델링 기법보다 성능이 우수하다는 특징이 있다(Jeong, Yoon, & Lee, 2019). LDA 기반 토픽 모델링은 문서-키워드 행렬과 토픽의 수를 입력으로 하여 문서-토픽 분포 행렬과 토픽-키워드 분포 행렬을 출력한다. 문서-단어 행렬은 앞서 구축한 TF-IDF 기반의 행렬을 사용한다. 토픽의 수는 각 토픽이 명확하게 구분되어 토픽 간 중복되는 내용을 최소화할 수 있는 최적의 개수를 찾아야 한다. 일반적으로 토픽 간 평균 코사인 유사도가 가장 낮은 지점을 선정하는 등의 체계적인 방법을 통해 최적의 토픽 수를 구할 수 있다(Chung et al., 2021).

본 연구는 LDA의 출력물인 토픽-키워드 분포 행렬을 통해 각 연구토픽이 어떤 키워드를 중심으로 구성되어 있는지와 문서-토픽 분포 행렬을 통해 각 문서가 어떤 연구토픽에 관한 내용을 다루는지를 파악하여 도출된 모든 연구토픽의 이름을 레이블링한다. 그리고 각 연구토픽의 동향을 시간의 흐름에 따라 분석하기 위해 문서-토픽 분포 행렬을 성장곡선 분석에 적합한 형태로 변환한다. 문서-토픽 분포는 한 논문이 각 연구토픽에 대한 내용을 얼마나 포함하고 있는지를 확률값으로 나타낸다. 따라서 연구토픽별로 특정 연도에 게재된 논문들의 기여도를 더하면 해당 연도에 각 연구토픽에 관련된 연구가 어느 정도로 수행되었는지를 파악할 수 있다(Chen et al., 2017). 본 연구는 〈그림 1〉과 같이 연도별 문서-토픽 분포 행렬의 기



〈그림 1〉 연도-토픽 누적 행렬 구축 예시

여도 합인 토픽량을 요소로 가지는 연도-토픽 행렬을 구축한다. 그 다음 연도-토픽 행렬의 토픽량을 연도별로 누적하여 연도-토픽 누적 행렬을 구축하고 이를 성장곡선에 활용한다.

3.3 성장곡선 분석

본 연구는 분석에 적절한 성장곡선 모델을 선정한 뒤 해당 모델에 각 토픽의 연도별 누적 토픽량을 적용하여 오픈액세스 분야 연구토픽의 성장을 모델링한다. 기술 및 연구 분야의 성장 분석에 자주 사용되는 로지스틱 모델과 곰페르츠 모델을 본 연구의 분석에 활용할 후보 모델로 정하였다. 로지스틱 모델과 곰페르츠 모델의 수식은 각각 〈수식 2〉, 〈수식 3〉과 같다.

$$Y_t = \frac{L}{1 + ae^{-b(t-t_0)}} \qquad \langle \text{수식 } 2 \rangle$$

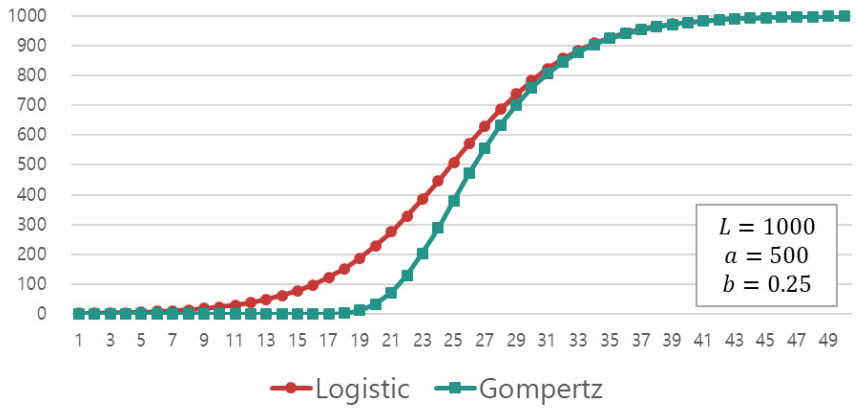
$$Y_t = Le^{-ae^{-b(t-t_0)}} \qquad \langle \text{수식 } 3 \rangle$$

수식에서 L은 성장곡선의 상한을, a와 b는 곡선의 모양을 결정하는 모수를 의미한다. 로지스틱 모델은 좌우 대칭인 S자 곡선인 반면, 곰페르츠 모델은 변곡점이 전체 수명주기의 약 37% 지점에 위치한 비대칭 형태의 S자 곡선이다(Franses, 1994)(〈그림 2〉 참조). 따라서 분석하는 데이터나 대상에 따라 적절한 성장곡선이 상이할 수 있다. 본 연구는 선별한 논문 데이터의 연도별 누적 논문 수를 로지스틱 모델과 곰페르츠 모델에 각각 적용하여 우수한 성능을 보이는 모델을 선정한다. 모델 선정에는 회귀 모델의 성능 평가에 널리 사용되는 지표인 MSE(Mean Squared Error), RMSE(Root Mean Squared Error), MAE(Mean Absolute Error)를 활용하며, 각 지표의 수식은 〈수식 4〉, 〈수식 5〉, 〈수식 6〉과 같다.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \qquad \langle \text{수식 } 4 \rangle$$

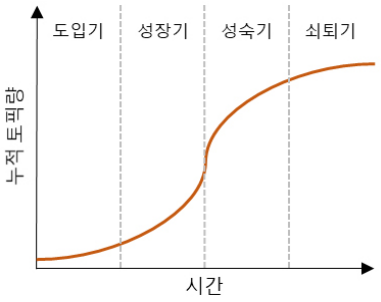
$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2} \qquad \langle \text{수식 } 5 \rangle$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \qquad \langle \text{수식 } 6 \rangle$$



〈그림 2〉 로지스틱 곡선과 곰페르츠 곡선의 형태

적절한 모델이 선정되면 각 연구토픽의 연도-토픽 누적 벡터를 모델에 적용하고, 연구토픽들의 성장 단계, 성숙도(Maturity Ratio, MR), 예상 잔여수명(Expected Remaining Life, ERL)을 살펴본다. 성장 단계는 〈그림 3〉과 같이 성장곡선을 도입기(Emerging), 성장기(Growth), 성숙기(Maturity), 쇠퇴기(Senility)의 네 구간으로 구분한 단계로, 현재 특정 연구토픽이 어느 단계에 해당하는지를 나타낸다. 성장 단계를 구분하는 명확한 기준은 존재하지 않기 때문에 분석자가 성장곡선의 변곡점을 참고하여 직관적인 판단으로 구분한다(김정욱, 정병기, 윤장혁, 2016).



〈그림 3〉 성장곡선의 네 단계

성숙도와 예상 잔여수명은 각각 특정 연구토픽이 성장 상한 대비 얼마나 성장하였는지, 특정 연구토픽이 소멸할 때까지 얼마나 남았는지를 의미한다(Yoon et al., 2014). 이때 성장곡선이 상한에 닿기까지는 무한한 시간이 필요하므로 분석 대상의 소멸 시기를 별도로 지정해야 한다. 본 연구에서는 곡선이 상한의 90%에 도달하는 시점을 소멸 시기로 지정한다. 성숙도와 예상 잔여수명의 수식은 〈수식 7〉, 〈수식 8〉과 같다.

$$MR = \frac{L_{now}}{L} \times 100 \qquad \text{〈수식 7〉}$$

$$ERL = T_{\rho} - T_{now} \qquad \text{〈수식 8〉}$$

수식에서 L_{now}는 현재 시점의 누적 토픽량, T_{ρ}는 누적 토픽량의 예상치가 성장 상한의 일정 비율 ρ에 도달하는 연도($\rho = 0.90$), T_{now}는 현재 연도를 의미한다.

4. 결 과

4.1 데이터 및 분석

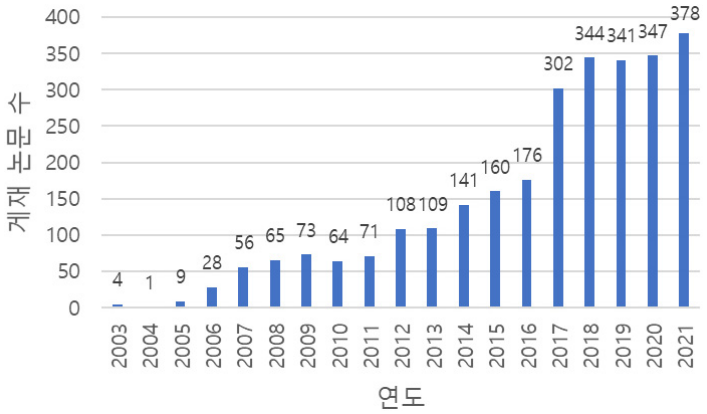
4.1.1 데이터

본 연구는 오픈액세스 분야와 관련된 논문을 수집하였다. Web of Science 등재 학술지에 게재된 문서들을 대상으로 제목(Title), 저자 키워드(Author Keywords), 추가 키워드(Keywords Plus) 필드에서 “open access” 키워드를 검색하였다. 검색 기간은 2001년 1월 1일부터 2021년 12월 31일까지, 문서 유형은 순수 연구 논문(Article)으로 설정하여 총 4,087편의 논문 데이터를 수집하였다. 분석을 위해 논문의 게재연도 정보가 필수적인데 수집한 데이터에는 게재연도가 기재되지 않은 논문이 일부 존재한다. 또한, 수집한 데이터는 본 연구에서 분석하고자 하는 문헌정보학 분야의 오픈액세스와 관련이 없는 논문을 포함하고 있을 수 있다. 예를 들어 내시경, 어업, 송전 등의 분야에서 “open access gastroscopy,” “open access colonoscopy,” “open access fishery,” “open access transmission” 등의 전문 용어들이 사용되고 있다. 본 연구는

게재연도 정보가 기재되지 않았거나 오픈액세스와 직접적인 관련이 없는 논문 1,252편을 제외한 논문 2,835편을 일차적으로 선별하였다.

이후 선별한 데이터에서 다음 절차를 통해 논문의 키워드 리스트를 구축하였다. 첫째, Python 자연어 처리 패키지인 spaCy[1]를 통해 각 논문의 제목과 초록에서 명사(구)를 추출하여 단수형태로 변환하고 각 키워드의 출현빈도를 구하였다. 둘째, 각 논문의 저자 키워드와 추가 키워드에 기재된 키워드들을 세미콜론 단위로 구분하고 중복을 제거하였다. 셋째, 앞선 두 단계에서 수집한 키워드와 출현빈도를 결합한 뒤 중복된 키워드가 있는 경우 출현빈도를 합하여 최종적으로 키워드 리스트를 구축하였다. 일반적으로 분석 대상 분야인 오픈액세스는 "open access" 혹은 그 약어인 "OA"로 표현된다. 본 연구에서는 분석에 있어 두 용어를 동일한 것으로 간주하기 위해 키워드 내 "OA"라는 용어를 모두 "open access"로 변경하여 저장하였다. 예를 들어 키워드 "gold oa"는 "gold open access"로, 키워드 "oa journal"은 "open access journal"로 저장되지만, 키워드 "bibliographic approach"는 별도로 변경 프로세스를 거치지 않고 그대로 저장된다. 모든 논문에 대하여 위 세 단계를 반복 수행한 뒤 마지막으로 다음과 같은 기준으로 불용어를 제거하여 유효 키워드 집합을 구축하였다: 1) 3글자 이하의 키워드 제거; 2) 문서빈도(Document Frequency)가 1인 키워드 제거; 3) 논문의 연구 내용과 관련이 없는 키워드 ("abstract," "purpose," "method" 등) 제거; 4) 절반 이상의 문서에서 등장하여 토픽 구분에 노이즈로 작용할 수 있는 키워드 "open access" 제거. 이와 같은 작업을 통해 본 연구는 4,338개의 유효 키워드를 선별하였으며 2,835편의 논문 가운데 이 유효 키워드를 포함하는 2,777편의 논문이 분석에 활용된다. 마지막으로 유효 키워드와 분석 대상 논문을 TF-IDF 기반의 문서-단어 행렬로 표현하였다.

본 연구의 분석에 활용하는 데이터의 연도별 논문 수는 〈그림 4〉와 같다. 2002년 부다페스트



〈그림 4〉 연도별 게재 논문 수

1) https://spacy.io/

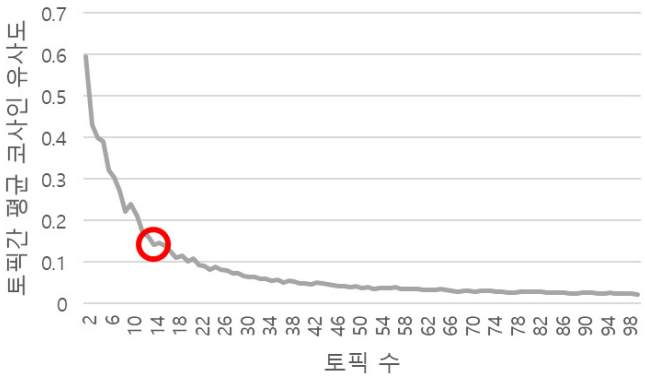
오픈액세스 이니셔티브(Budapest Open Access Initiative) 이후로 2003년부터 “open access”라는 키워드를 포함하는 관련 연구들이 수행되기 시작하였다. 2004년도와 2010년도에 출판된 논문의 수가 소량 감소한 것을 제외하면 최초 등장 이후 현재까지 꾸준히 연구 규모가 증가하고 있음을 확인할 수 있다. 특히 2017년에는 논문 수가 전년도 대비 약 1.7배 눈에 띄게 증가하는 현상이 관찰되었다.

4.1.2 데이터 분석

본 연구는 Python 패키지인 scikit-learn[2]을 활용하여 LDA 기반의 토픽 모델링을 수행하였다. LDA의 입력값 중 하나인 문서-단어 행렬은 4.1.1장에서 구축한 TF-IDF 기반의 문서-단어 행렬을 사용한다. LDA의 또 다른 입력값 중 하나인 토픽의 개수는 〈그림 5〉와 같이 토픽의 단어 분포 행렬 간 평균 코사인 유사도의 감소 추세가 안정화되는 지점인 14개로 선정하였다.

토픽 모델링으로 도출된 토픽-단어 분포 행렬과 문서-토픽 분포 행렬을 통해 토픽별로 비중이 높은 단어와 문서를 파악하여 오픈액세스 분야 내 14개 연구토픽의 이름을 레이블링하였다(〈표 1〉 참조). 본 연구는 오픈액세스 분야를 이해하기 위하여 관련 문헌을 수집했음에도, 오픈 사이언스(Open Science)의 세 가지 핵심 요소인 오픈액세스(Open Access), 오픈데이터(Open Data), 오픈협업(Open Collaboration)에 대한 연구토픽이 모두 도출되었다(최희윤, 서태설, 2020).

연도별 누적 논문 수를 기반으로 로지스틱 모델과 곰페르츠 모델을 적용한 결과는 〈표 2〉와 같다. 두 모델의 모수들에 대한 t-검정 결과 모든 p-value가 0.01 이하로 유의한 것으로 확인되었다. 로지스틱 모델의 상한은 5,848, 곰페르츠 모델의 상한은 57,976으로 두 모델의 상한이 큰 차이를 보인다. 다시 말해 로지스틱 모델은 오픈액세스 분야의 논문이 5,848편가량 게재될 때까지 성장할 것으로 예상한 반면, 곰페르츠 모델은 관련 논문이 57,976편 게재될 때까지 성장할 것으로 예상하였다. 성능평가 측면



〈그림 5〉 토픽 수에 따른 토픽간 평균 코사인 유사도 추세

2) https://scikit-learn.org

〈표 1〉 오픈액세스 분야 연구토픽 및 주요 키워드

토픽 #	연구토픽	주요 키워드	분류
토픽 1	데이터 활용 및 평가	open access data, business model, open access resource, wikipedia, quality metric	오픈데이터
토픽 2	학술지 평가	article processing charge, scholarly publishing, predatory publisher, scholarly communication, mega journal	오픈액세스
토픽 3	기타	data, latin america, self citation, society publisher, medical informatics	기타
토픽 4	조직적 협력	open access publishing, international collaboration, access journal, data management, research institute	오픈액세스
토픽 5	구독 학술지	subscription journal, sci hub, plan s, transformative agreement, hybrid open access	오픈액세스
토픽 6	약탈적 출판	predatory journal, open science, open access publication, predatory publishing, open access policy	오픈액세스
토픽 7	과학적 파급력	open access article, google scholar, citation, journal article, impact	오픈액세스
토픽 8	학술 소셜 네트워킹 서비스	citation rate, digital repository, open access practice, academic social network, student	오픈협업
토픽 9	결정학 데이터베이스	crystallography open database, research data, open access collection, crystal structure, research output	오픈데이터
토픽 10	유전학 데이터베이스	academic journal, deep learning, gene expression, binding site, gene	오픈데이터
토픽 11	예측 분석	data sharing, acceptance rate, machine learning, scientific publication, borrowing request	오픈데이터
토픽 12	기관 리포지터리	institutional repository, academic librarian, academic staff, university library, academic library	오픈액세스
토픽 13	오픈액세스 학술지	open access journal, directory of open access journal, social science, open scholarship, directory of open access journals doaj	오픈액세스
토픽 14	동료심사 및 출판	peer review, document supply, academic publishing, e book, electronic publishing	오픈협업

〈표 2〉 로지스틱 모델과 곰페르츠 모델 적용 결과

모델	MSE	RMSE	MAE
로지스틱 모델 ($L=5,848$, $a=177.9631$, $b=0.2677$)	1,096.2000	33.1089	28.2113
곰페르츠 모델 ($L=57,976$, $a=8.1816$, $b=0.0523$)	**958.6116**	**30.9615**	**27.2673**

노트: 두 모델의 모수 L, a, b에 대한 t-검정 결과 p-value가 모두 0.01 이하임.

에서 곰페르츠 모델이 세 가지 평가지표 모두 낮은 값을 기록했기 때문에 본 연구에서 분석하는 데이터에는 곰페르츠 모델이 더욱 적합하다고 판단하였다. 따라서 본 연구에서는 14개 연구토픽의 누적 토픽량을 곰페르츠 모델에 적용하여 성장곡선 분석을 수행하였으며 연구토픽별 모델 적용 결과는 〈표 3〉과 같다. 본 연구에서는 곡선의 기울기가 가팔라지기 시작하여 특정 분야가 본격적인 성장에 진입했다고 판단되는 지점인 성숙도 10%를 기준으로 도입기와 성장기를 구분하였다.

〈표 3〉 토픽별 곰페르츠 모델 적용 결과

토픽 #	성장상한 L	모수 a	모수 b	예상 잔여수명 ERL	성숙도 MR	성장단계
토픽 1	5,012	8.7005	0.0491	71	3.32	도입기
토픽 2	7,784	8.1933	0.0478	74	2.72	도입기
토픽 3	12,533	8.8783	0.0379	99	1.31	도입기
토픽 4	1,631	7.6846	0.0669	46	11.53	성장기
토픽 5	15,653	9.0162	0.0374	100	1.18	도입기
토픽 6	11,149	8.6006	0.0451	79	2.58	도입기
토픽 7	2,734	8.0051	0.0627	51	8.71	도입기
토픽 8	5,237	8.0946	0.0463	75	3.38	도입기
토픽 9	6,890	8.3018	0.0442	80	2.80	도입기
토픽 10	3,745	8.5422	0.0560	60	5.09	도입기
토픽 11	2,355	7.2398	0.0545	52	7.64	도입기
토픽 12	1,404	7.2853	0.0711	41	15.11	성장기
토픽 13	5,284	8.4640	0.0501	69	3.73	도입기
토픽 14	1,737	7.2055	0.0619	50	10.80	성장기

4.2 분석 결과

4.2.1 연구문제 1: 오픈액세스 내에 어떤 세부 연구분야가 존재하는가?

연구문제 1에 답하기 위하여 본 장에서는 〈표 1〉의 토픽 모델링 결과에 대해 상세하게 분석한다. 본 연구는 레이블링이 모호한 토픽 3(기타)을 제외하고 전체 연구토픽을 오픈 사이언스의 세 가지 핵심요소에 따라 구분하였다(최희윤, 서태설, 2020). 오픈액세스와 관련된 연구토픽은 토픽 2(학술지 평가), 4(조직적 협력), 5(구독 학술지), 6(약탈적 출판), 7(과학적 파급력), 12(기관 리포지터리), 13(오픈액세스 학술지), 오픈데이터와 관련된 연구토픽은 토픽 1(데이터 활용 및 평가), 9(결정학 데이터베이스), 10(유전학 데이터베이스), 11(예측 분석), 오픈협업과 관련된 연구토픽은 토픽 8(학술 소셜 네트워킹 서비스), 14(동료심사 및 출판)이다.

본 연구의 주요 분석 대상 분야가 오픈액세스인 만큼 오픈액세스와 관련된 연구토픽은 전체 14개 토픽 중 7개로 가장 많이 식별되었다. 각 연구토픽은 오픈액세스 내 이슈 및 주제어들을 포함하고 있다. 토픽 2는 다양한 관점에서 학술지를 평가하는 연구들이 주를 이룬다. 예를 들어, 오픈액세스 학술지 및 메가 저널에 대한 정량적 평가와 저자의 학술지 선정에 미치는 요인에 대한 조사 등의 연구가 진행되었다. 토픽 4는 오픈액세스를 확산하기 위한 조직적 협력에 관한 토픽이다. 연구자 혹은 연구기관의 개별적인 노력만으로는 오픈액세스를 확산하기가 어려운 것이 현실이다. 따라서 정부 부처 및 연구기금 지원 기관, 대학/연구소 등 연구수행 기관, 도서관, 출판사 및 학회, 개인 연구자 등 각 주체가 적극적으로 나서야 한다. 이러한 관점에서, 오픈액세스 학술지 출판을 위한 연구기관과 출판사의 협력에 관한 연구, 오픈액세스 학술지 출판을 위한 비즈니스 모델을 제시하고

이해관계자의 역할을 정리하는 연구 등이 수행되었다. 토픽 5는 구독 기반 학술지의 오픈액세스화에 관한 다양한 키워드를 포함하고 있다. 예를 들어, 구독 기반 학술지의 구독료를 출판료로 전환하는 전환계약(Transformative Agreement), 저자의 셀프 아카이빙을 허용하는 그린 오픈액세스와 Sci-Hub 등의 시스템을 통해 불법적으로 논문을 공개하는 블랙 오픈액세스(Black Open Access)의 비교, 구독 기반의 학술지에 저자가 논문 출판 비용(Article Processing Charge, APC)을 지불하여 오픈액세스 논문을 출판하는 하이브리드 골드 오픈액세스(Hybrid Gold Open Access) 학술지의 수익 구조 등의 연구들이 진행되었다. 토픽 6은 약탈적 출판에 관한 연구토픽이다. 약탈적 출판은 학술출판의 디지털화 및 오픈액세스의 부작용 중 하나로 투명하고 엄격하지 않은 동료심사를 통해 가능한 많은 논문을 출판하면서 저자에게 논문 출판 비용을 청구하여 이익을 얻는 비즈니스 모델을 의미한다(Beall, 2012; Eriksson & Helgesson, 2017). 이러한 현상을 이해하고 나아가 올바른 학술출판 문화를 조성하기 위해 저자의 약탈적 학술지에 대한 인식 및 이용 동기를 조사하는 연구, 약탈적 학술지의 특성을 분석하는 연구, 약탈적 학술지에 게재된 논문의 영향력을 조사하는 연구 등 다양한 연구들이 수행되어왔다. 토픽 7은 오픈액세스의 과학적 파급력에 관한 연구토픽이다. 이 토픽에 해당하는 연구들은 피인용이나 온라인 접근성 및 가시성 등을 과학적 파급력의 지표로 설정하고 오픈액세스 학술지와 구독 기반 학술지를 학술지 관점 혹은 논문 관점에서 비교하였다. 이러한 연구들은 각 관점에서 오픈액세스 학술지와 오픈액세스 논문의 파급력을 통계적으로 증명하였다. 토픽 12는 학술연구 성과를 수집하고 저장하는 리포지터리에 관한 연구토픽이다. 리포지터리는 수집 주체 및 수집 대상에 따라 기관 리포지터리와 주제 리포지터리 등으로 구분되는데 이 토픽에서는 대학의 기관 리포지터리에 관한 연구들이 주를 이룬다. 예를 들어 특정 국가의 대학 도서관이 운영하는 리포지터리에 대한 비즈니스 모델, 대학 도서관 사서들의 셀프 아카이빙에 대한 인식과 그린 오픈액세스 확산을 위해 기관 리포지터리를 운영하는 사서에게 필요한 역량 등의 연구들이 수행되어왔다. 마지막으로 토픽 13은 오픈액세스 학술지에 관한 연구토픽이다. 특히 글로벌 오픈액세스 학술지 색인 서비스인 Directory of Open Access Journals(DOAJ)와 관련하여 DOAJ 등재의 효과 및 DOAJ의 한계점을 논의하는 등의 연구가 진행되었다. 이 외에는 각기 다른 분야에서 오픈액세스 논문을 기반으로 수행한 계량 분석 연구가 다양하게 진행되었음을 확인하였다.

오픈데이터와 관련된 연구토픽은 총 4개 도출되었다. 오픈데이터는 모든 사람이 공개적으로 접근, 활용, 편집, 공유할 수 있는 데이터를 의미한다. 본 연구에서는 토픽 모델링 결과를 통하여 "자유롭게 접근할 수 있는" 데이터라는 의미를 나타내기 위해 오픈액세스라는 표현이 사용되고 있음을 확인하였다. 많은 연구자가 오픈데이터를 오픈액세스 데이터(Open Access Data) 혹은 오픈액세스 데이터베이스(Open Access Database)로 표현하고 있으며, 특정 분야의 오픈데이터와 그 집합을 의미하는 오픈액세스 컬렉션(Open Access Collections)이나 오픈액세스 시리즈(Open Access Series)와 같은 용어도 확립되

어 있다. 이러한 관점에서 토픽 1은 공개된 데이터를 평가하는 연구나 공개된 데이터를 각기 다른 목적으로 활용한 연구들이 주를 이룬다. 예를 들어 오픈데이터를 평가할 수 있는 데이터 품질 지표를 제시하는 연구와 특정 도메인의 데이터를 공개하기 위한 목적의 연구 등이 토픽 1에 포함되어 있다. 토픽 9와 토픽 10은 각각 결정학 분야와 유전학 분야에서 오픈데이터를 활용한 연구에 대한 토픽이다. 토픽 9에 속한 대다수의 연구는 "crystallography open database"라는 키워드를 포함한다. Crystallography Open Database[3]는 바이오 폴리머를 제외한 유기, 무기, 금속-유기 화합물 및 광물의 결정구조에 대한 오픈액세스 컬렉션이다. 이 토픽에 해당하는 문서들은 공개적으로 사용 가능한 결정 데이터를 통하여 결정구조에 관해 연구한 경우가 대다수를 이루기 때문에 토픽 9는 결정학 분야에서 오픈데이터의 활용에 관한 연구 흐름을 나타낸다고 해석할 수 있다. 토픽 10에 속한 문서들은 무료로 공개된 유전자 발현 데이터베이스를 활용한 연구들이 대다수를 이룬다. 유전자 발현의 요인을 분석하는 연구, 염기서열 패턴을 학습한 딥러닝 모델을 구축하는 연구, 특정 생물군의 유전자 발현 데이터 및 리포지터리에 대한 연구 등이 수행되었다. 이외에 다양한 오픈데이터를 활용하여 달리 분류하기 어려운 연구들은 토픽 11로 식별되었다. 이 토픽에 속한 연구들은 대체로 특정 도메인에서 공개된 데이터를 활용하여 분야 내에 존재하는 다양한 문제를 해결하고자 머신러닝 및 딥러닝 모델을 개발하는 목적을 가진다.

오픈협업과 관련된 연구토픽은 두 개 도출되었다. 유럽 위원회에서는 오픈 사이언스의 관행 중 하나로 오픈협업을 "학계 내, 혹은 시민, 시민사회, 시민 과학과 같은 타 지식 행위자와의 열린 협력"이라고 정의한다(European Commission, 2021). 이러한 관점에서 본 연구에서는 오픈액세스와 관련도가 높은 두 개의 오픈협업 연구토픽이 식별되었다. 토픽 8은 학술 소셜 네트워킹 서비스와 관련된 연구토픽이다. 학술 소셜 네트워킹 서비스는 연구자들이 논문, 특허, 프로젝트와 같은 연구 결과물을 공개하고 질의응답을 통해 학술적 지식을 공유하는 등 전세계 연구자들과 소통할 수 있는 기능을 한다(Ovadia, 2014). 실제로 관심 있는 연구 혹은 연구자를 팔로우하여 협업을 요청하는 사례도 다수 존재하는 만큼 학술 소셜 네트워킹 서비스는 다양한 지식 행위자들이 자유롭게 상호작용하는 오픈협업의 장으로서 역할을 하고 있다. 본 연구에서는 오픈액세스와 관련하여 디지털 리포지터리와 학술 소셜 네트워킹 서비스를 비교하는 연구뿐만 아니라 오픈협업과 관련하여 학술 소셜 네트워킹 서비스를 분석하거나 새로운 서비스를 제안하는 연구들이 토픽 8로 식별되었다. 예를 들어 유명 학술 소셜 네트워킹 서비스(Academia.edu, Mendeley.com, ResearchGate.net, Zeroto.org, Google Scholar 등)의 기능에 대한 비교, ResearchGate.net의 알트메트릭(Altmetric)과 SCOPUS 서지 지표의 관계에 대한 분석, 연구자들이 웹 기반 소셜 네트워크에서 타 연구자들의 연구결과를 평가하는 소셜 피어리뷰(Social Peer Review) 방식을 제안하는 연구들이 토픽 8에 속해 있다.

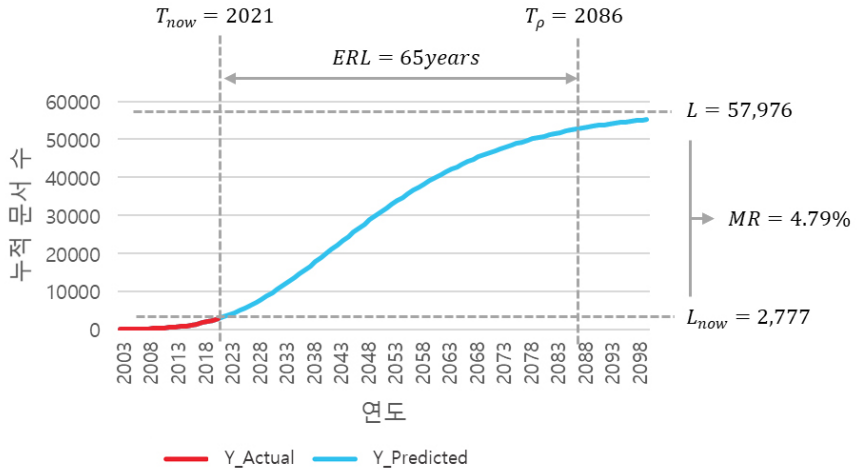
3) http://www.crystallography.net/cod/

한편 토픽 14는 동료심사 및 출판에 관한 연구 토픽이다. 동료심사는 출판되는 학술논문의 질을 개선함과 동시에 과학의 발전에 크게 기여하는 학술 활동이다. 하지만 전통적인 암맹 동료심사는 심사의 신뢰성이나 일관성, 이해충돌 등의 문제를 가지고 있다(Ross-Hellauer, 2017). 따라서 암맹 동료심사의 문제점을 해결하기 위한 방식으로 개방형 동료심사(Open Peer Review)가 제시되었다. 개방형 동료심사는 동료심사자의 심사 보고서를 공개하거나 연구자들이 논문에 대한 의견을 공개적으로 게시함으로써 과학의 발전에 기여한다는 점에서 오픈 사이언스와 오픈협업의 주요 도구로도 관심을 받고 있다. 본 연구에서는 오픈액세스 학술출판이나 전자출판을 통한 접근성 확대나 알트메트릭 등에 관한 연구들뿐만 아니라 암맹 동료심사의 문제점을 제시하고 오픈액세스와 개방형 동료심사에 관해 논의하는 연구들이 토픽 14로 식별되었다.

4.2.2 연구문제 2: 오픈액세스와 그 세부 연구 분야는 앞으로 얼마나 더 성장할 것인가?

연구문제 2에 대해 답하기 위해 본 장에서는 성장곡선 분석 결과에 대해 상세하게 해석한다. 정량적으로 계산된 성숙도, 예상 잔여수명과 정성적으로 판단한 성장 단계를 통해 오픈액세스와 각 연구토픽의 성장에 대해 살펴본다.

오픈액세스 분야의 성장은 모델 선정을 위해 2,777개 논문에 대하여 연도별 누적 논문 수를 곰페르츠 모델에 적용한 결과를 통해 해석할 수 있다. 〈표 2〉에 기재된 곰페르츠 모델의 모수를 시각화한 그래프는 〈그림 6〉과 같다. 그래프에서 빨간색 선은 실제 연도별 누적 논문 수를 의미하며 파란색 선은 곰페르츠 모델이 예상한 연도별 누적 논문 수이다. 곰페르츠 곡선의 상한(L)은 57,976이기 때문에 오픈액세스 분야는 관련 논문이 약 58,000건 출판될 때까지 성장할 것으로 예상된다. 성장곡선이 상한의 90%에 도달하는 지점인 소멸 시기(T_ρ)는 2086년으로, 예상 잔여수명(ERL)은 약 65년 남은 것

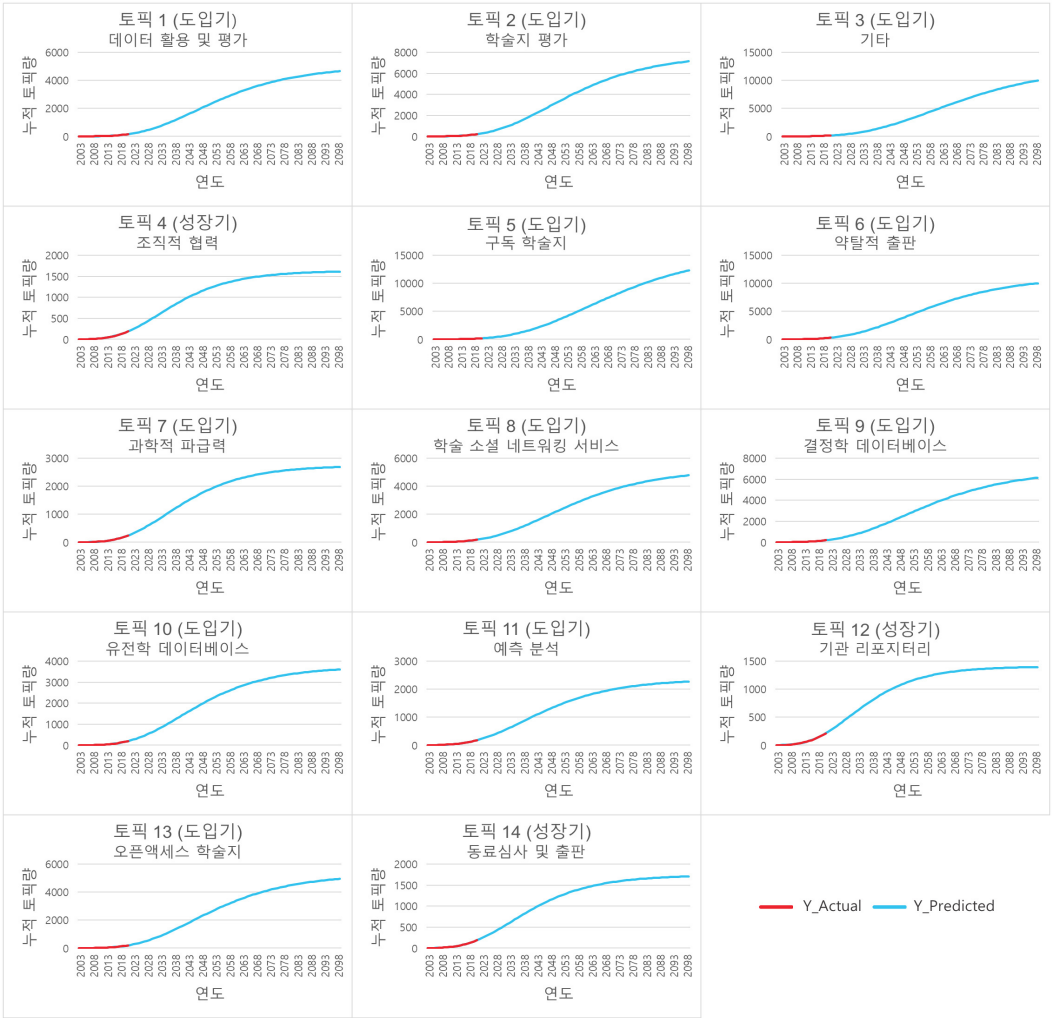


〈그림 6〉 오픈액세스 분야의 성장곡선

으로 확인된다. 마지막으로, 오픈액세스 분야의 2021년 기준 성숙도(MR)는 4.79%로 전체 성장 단계 중 도입기에 해당하는 것으로 보인다.

세부 연구 분야의 성장을 분석하기에 앞서 〈표 2〉와 〈표 3〉을 비교해본 결과 연도별 누적 논문 수를 기반으로 적용한 곡선과 모든 연구토픽의 곡선이 비슷한 수치의 모수(a, b)로 도출되었음을 확인하였다. 또한, 연구토픽들의 평균 예상 잔여수명은 67.64년, 평균 성숙도는 5.7%로 〈그림 6〉에서 분석한 전체 오픈액세스 분야의 결과와 근접한 수치를 보였다. 따라서 본 연구에서 적용한 총 15개의 성장곡선 모델 모두가 오픈액세스와 그 세부 연구 분야의 성장을 일관되게 나타내고 있음을 확인할 수 있다.

14개 연구토픽에 대한 성장곡선은 〈그림 7〉과 같다. 대다수 연구토픽이 현재 도입기 단계



〈그림 7〉 연구토픽별 성장곡선

인 것으로 보인다. 특히 성숙도가 5% 미만인 토픽 1(데이터 활용 및 평가), 2(학술지 평가), 3(기타), 5(구독 학술지), 6(약탈적 출판), 8(학술 소셜 네트워킹 서비스), 9(결정학 데이터베이스), 13(오픈액세스 학술지)은 전체 오픈액세스 분야와 비교했을 때 현재 성장이 더딘 분야이다. 따라서 해당 분야에서는 앞으로 더욱 활발한 연구가 진행될 것이라 예상된다. 반면 성장 상한이 낮아 현재 성숙도가 10% 이상으로 높게 도출된 토픽 4(조직적 협력), 12(기관 리포지터리), 14(동료심사 및 출판)는 성장기에 진입한 것으로 보인다. 해당 연구토픽들의 예상 잔여수명은 짧게는 41년, 길게는 50년으로 계산되었다. 변곡점이 전체 곡선의 35%에 위치하는 곰페르츠 곡선의 특성으로 미루어 보아 현재 성장기인 세 연구토픽은 10년 이내로 그 성장 추세가 완만해지는 성숙기에 진입할 것으로 예상된다.

5. 결 론

본 연구는 오픈액세스의 세부 연구 분야를 식별하고 각 분야의 진화와 성장 추세를 파악하고자 Web of Science에서 수집 및 전처리한 오픈액세스 관련 학술논문 2,777건에 토픽 모델링과 성장곡선 분석을 수행하였다. 본 연구에서 수행한 분석의 세부 결과는 다음과 같다. 첫째, 오픈 사이언스의 세 가지 핵심요소인 오픈액세스, 오픈데이터, 오픈협업과 관련된 14개 연구토픽을 식별하였다. 이는 오픈액세스와 관련된 연구자들이 단순히 학술정보의 개방뿐만 아니라 학술연구 활동의 성과물인 데이터의 공개와 연구주제에 대한 비판적 의견을 주고받으며 커뮤니케이션하는 것에도 관심이 있음을 미루어 짐작해볼 수 있다. 또한, 본 연구에서는 “자유롭게 접근할 수 있는”이라는 뜻의 형용사로 오픈액세스라는 용어가 사용되고 있음을 확인하였다. 이는 전 세계 연구자들이 오픈액세스라는 패러다임에 대하여 아직은 완전히 이해하고 있지는 않음을 의미한다고 볼 수 있다. 둘째, 오픈액세스 분야는 약 65년의 잔여 수명을 가지고 있으며 성장 상한까지 5% 남짓 성장한 도입기 단계인 것으로 보인다. 그리고 오픈액세스의 세부 연구 분야들 역시 현재까지 적게는 1%, 많게는 15%가량 성장하여 앞으로 널리 성장할 여지가 충분하다. 이는 첫 번째 결과와 관련하여 오픈액세스가 아직은 성숙한 분야가 아니므로 오픈액세스 대한 인식이 완전히 적립되지 않았을 수 있다는 점을 뒷받침할 수 있다.

본 연구는 데이터를 기반으로 오픈액세스의 글로벌한 동향을 분석하였다는 점에서 의의를 갖는다. 특히, 분석 대상 데이터를 특정 학술지 카테고리 혹은 특정 국가로 한정하지 않았기 때문에 전 세계적으로 그리고 전 분야에서 오픈액세스와 관련하여 어떤 연구가 수행되고 있는지에 대한 이해를 제공하였다. 또한, 오픈액세스 분야의 미래 성장 추세를 전망한 초기 시도라는 점 역시 본 연구의 기여점이다. 과거의 추세를 통해 미래를 예상해볼 수 있는 데이터 분석의 강점을 활용하여 본 연구는 오픈액세스 분야가 현재 얼마나 성장했는지, 앞으로 얼마나 더 성장할 수 있을지에 대한 인사이트를 제공하였다. 따라서 본 연구의 결과는 학술출판 관계자, 문헌정보학 분야 연구자, 나아가 연구기금 지원 기관과 정부의 정책 결정자들에게

오픈액세스에 대한 포괄적인 이해를 제공할 수 있을 것으로 사료된다.

그럼에도 불구하고 본 연구는 몇 가지 한계점과 후속 연구에 대한 가능성을 가진다. 첫째, 본 연구는 Web of Science에서 간단한 검색식을 통해 대량의 논문 데이터를 수집한 뒤 분석자의 정성적인 판단을 통하여 분석 대상이 되는 데이터를 선별하였다. 이 과정에서 오픈액세스와 관련되지 않은 데이터가 일부 포함되었을 수 있다. 엄격한 기준으로 분석 대상 데이터를 선별했다면 실제 오픈액세스와 더욱 밀접한 관련이 있는 연구토픽들에 대해 깊이 있는 분석이 가능했을 것이다. 다만 본 연구에서는 오픈 사이언스라는 큰 틀 안에서 오픈액세스 데이터, 오픈액세스 데이터베이스와 같은 표현 역시 대중들이 오픈액세스를 인식하는 방식이라 가정했고, 이를 통해 오픈액세스에 대한 정확한 개념과 이해가 아직은 널리 확산되지 않았다는 점을 발견하였다. 추후에는 오픈액세스에 관한 논문을 식별할 수 있는 엄격한 검색식을 개발하고 검증하는 연구를 수행할 필요가 있으며, 개발된 검색식을 통해 수집한 데이터에 본 연구에서 사용한 분석 기법을 적용해 볼 수 있을 것으로 생각된다. 둘째, 본 연구는 토픽 모델링 결과물을 기반으로 연도-토픽 누적 행렬을 분석에 사용했기 때문에 연구토픽의 성장은 연도별 논문 수를 기반으로 분석된 전체 오픈액세스 분야의 성장에 의존적이다. 따라서, 추후 각 토픽의 연도별 가중치로 연도-토픽 누적 행렬을 보완하거나 연도에 따라 토픽을 식별하는 동적 토픽 모델링 기법을 사용한다면 논문 수에 의존적이지 않은 새로운 인사이트를 식별할 수 있을 것으로 생각된다. 셋째, 성장곡선 분석은 개체가 S자 형태를 따르며 성장한다는 가정을 기반으로 한다. 하지만 개체의 성장은 쇠퇴기까지 다다르지 않을 수 있으며 도입기, 성장기, 성숙기 중에도 언제든지 그 성장이 멈출 수 있다. 본 연구 역시 오픈액세스 분야가 S자 형태로 성장한다는 가정을 기반으로 수행되었기 때문에 시간이 지나 본 연구의 성장곡선 분석결과가 사실인지를 확인해보는 연구를 수행해볼 수 있을 것이다. 넷째, 본 연구의 토픽 레이블링은 토픽 모델링 결과를 기반으로 저자 간 합의를 통해 수행되었다. 추후 오픈액세스 및 학술 커뮤니케이션 분야 전문가를 통해 14개 연구토픽을 해석하고 레이블링 결과를 검증하는 연구를 수행한다면 본 연구의 타당성을 더욱 강화할 수 있을 것이다. 마지막으로, 본 연구는 2003년부터 2021년까지의 데이터를 기반으로 오픈액세스 내 세부 연구 분야를 식별하고 식별된 연구 분야들의 성장을 파악하였다. 따라서 현재 하나의 주된 분야로 성장하지 못한 연구 분야 혹은 미래에 발생할 수 있는 연구 분야에 대한 고려가 부족하다. 추후에는 이상치 탐지 기법 등을 활용하여 미래에 떠오를 오픈액세스 관련 연구 분야들을 예상해보는 연구를 수행해볼 수 있을 것이다.

참 고 문 헌

김선겸, 김완종, 서태설, 최현진 (2019). 동시출현단어 분석을 활용한 오픈액세스 분야의 지적구조 분석: 2013 년부터 2018 년까지 출판된 문헌정보학 저널을 기반으로. 한국도서관·정보학회지, 50(1), 333-356. https://doi.org/10.16981/kliss.50.1.201903.333

김정욱, 정병기, 윤장혁 (2016). 네트워크분석과 기술성장모형을 이용한 기술기획: 증강현실 기술의 특허를 활용하여. 대한산업공학회지, 42(5), 337-351. https://doi.org/10.7232/JKIIE.2016.42.5.337

김판준 (2021). 동시출현단어분석에 기초한 지적구조 분석에서 키워드 유형별 특성에 관한 연구: 국외 오픈액세스 분야를 중심으로. 한국문헌정보학회지, 55(3), 103-129. https://doi.org/10.4275/KSLIS.2021.55.3.103

서선경, 정은경 (2013). 동시출현단어 분석 기반 오픈 액세스 분야 지적구조에 관한 연구. 한국비블리아학회지, 24(1), 207-228. https://doi.org/10.14699/kbiblia.2013.24.1.207

송성전, 심지영 (2022). 리뷰 정보를 활용한 이용자의 선호요인 식별에 관한 연구. 정보관리학회지, 39(3), 311-336. https://doi.org/10.3743/KOSIM.2022.39.3.311

신주은, 김성희 (2021). 국내 오픈액세스 분야의 지적구조 분석에 관한 연구. 한국문헌정보학회지, 55(2), 147-178. https://doi.org/10.4275/KSLIS.2021.55.2.147

윤희윤, 김신영 (2007). 국내외 문헌정보학 학술지의 오픈 액세스 동향 분석. 한국도서관·정보학회지, 38(1), 257-276. https://doi.org/10.16981/kliss.38.1.200703.257

최재황, 조현양 (2005). 오픈 액세스 운동의 동향과 학술적 이해관계자의 대응전략. 정보관리학회지, 22(3), 307-326. https://doi.org/10.3743/KOSIM.2005.22.3.307

최희윤, 서태설 (2020). 글로벌 연대와 상생의 길 오픈사이언스. 한국과학기술정보연구원. 출처: https://repository.kisti.re.kr/handle/10580/15590

Adamuthe, A. C. & Thampi, G. T. (2019). Technology forecasting: a case study of computational technologies. Technological Forecasting and Social Change, 143, 181-189. https://doi.org/10.1016/j.techfore.2019.03.002

Beall, J. (2012). Predatory publishers are corrupting open access, Nature, 489(7415), 179-179. https://doi.org/10.1038/489179a

Berger, R. D. (1981). Comparison of the gompertz and logistic equations to describe plant disease progress. Phytopathology, 71(7), 716-719. https://doi.org/10.1094/phyto-71-716

Bianchi, F., Terragni, S., Hovy, D., Nozza, D., & Fersini, E. (2020). Cross-lingual contextualized topic models with zero-shot learning. https://doi.org/10.48550/arXiv.2004.07737

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

Braun, T., Schubert, A. P., & Kostoff, R. N. (2000). Growth and trends of fullerene research as reflected in its journal literature. Chemical Reviews, 100(1), 23-38. https://doi.org/10.1021/cr990096j

Chen, B., Tsutsui, S., Ding, Y., & Ma, F. (2017). Understanding the topic evolution in a scientific domain: an exploratory study for the field of information retrieval. Journal of Informetrics, 11(4), 1175-1189. https://doi.org/10.1016/j.joi.2017.10.003

Cho, J. (2020). Intellectual structure evolution of open access research observed through correlation index of keyword centrality. Scientometrics, 125(3), 2617-2635. https://doi.org/10.1007/s11192-020-03682-4

Cho, Y. & Daim, T. (2016). OLED TV technology forecasting using technology mining and the Fisher-Pry diffusion model. Foresight, 18(2), 117-137. https://doi.org/10.1108/fs-08-2015-0043

Chung, J., Ko, N., Kim, H., & Yoon, J. (2021). Inventor profile mining approach for prospective human resource scouting. Journal of Informetrics, 15(1), 101103. https://doi.org/10.1016/j.joi.2020.101103

Costa, M. P. D. & Leite, F. C. L. (2016). Open access in the world and Latin America: a review since the Budapest Open Access Initiative. TransInformação, 28, 33-46. https://doi.org/10.1590/2318-08892016002800003

Craig, I. D., Plume, A. M., McVeigh, M. E., Pringle, J., & Amin, M. (2007). Do open access articles have greater citation impact?: a critical review of the literature. Journal of Informetrics, 1(3), 239-248. https://doi.org/10.1016/j.joi.2007.04.001

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6), 391-407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6〈391::AID-ASI1〉3.0.CO;2-9

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. Transactions of the Association for Computational Linguistics, 8, 439-453. https://doi.org/10.1162/tacl_a_00325

Du, H., Liu, D., Lu, Z., Crittenden, J., Mao, G., Wang, S., & Zou, H. (2019). Research development on sustainable urban infrastructure from 1991 to 2017: a bibliometric analysis to inform future innovations. Earth's Future, 7(7), 718-733. https://doi.org/10.1029/2018ef001117

Elvers, D., Song, C. H., Steinbüchel, A., & Leker, J. (2016). Technology trends in biodegradable

polymers: evidence from patent analysis. Polymer Reviews, 56(4), 584-606. https://doi.org/10.1080/15583724.2015.1125918

Eriksson, S. & Helgesson, G. (2017). The false academy: predatory publishing in science and bioethics. Medicine, Health Care and Philosophy, 20(2), 163-170. https://doi.org/10.1007/s11019-016-9740-3

European Commission (2021). Horizon Europe, open science: early knowledge and data sharing, and open collaboration. Available: https://data.europa.eu/doi/10.2777/18252

Franses, P. H. (1994). A method to select between Gompertz and logistic trend curves. Technological Forecasting and Social Change, 46(1), 45-49. https://doi.org/10.1016/0040-1625(94)90016-7

Hofmann, T. (1999). Probabilistic latent semantic indexing. In Proceedings of the 22nd annual international Annual Conference of the Association for computing Machinery Special Interest Group in Information Retrieval Conference on Research and Development in Information Retrieval, 50-57.

Jeong, B., Yoon, J., & Lee, J. M. (2019). Social media mining for product planning: a product opportunity mining approach based on topic modeling and sentiment analysis. International Journal of Information Management, 48, 280-290. https://doi.org/10.1016/j.ijinfomgt.2017.09.009

Ji, H. & Cha, M. (2021). Topic analysis of scholarly communication research. Journal of Information Science Theory and Practice, 9(2), 47-65. https://doi.org/10.1633/JISTaP.2021.9.2.4

Ko, N., Jeong, B., Choi, S., & Yoon, J. (2017). Identifying product opportunities using social media mining: application of topic modeling and chance discovery theory. Institute of Electrical and Ectronics Engineers Access, 6, 1680-1693. https://doi.org/10.1109/access.2017.2780046

Ma, T., Li, R., Ou, G., & Yue, M. (2018). Topic based research competitiveness evaluation. Scientometrics, 117(2), 789-803. https://doi.org/10.1007/s11192-018-2891-7

Ovadia, S. (2014). ResearchGate and Academia.edu: academic social networks. Behavioral & Social Sciences Librarian, 33(3), 165-169. https://doi.org/10.1080/01639269.2014.934093

Palmer, K. L., Dill, E., & Christie, C. (2008). Where there's a will, there's a way?: survey of academic librarian attitudes about open access. College & Research Libraries, 70(4), 315-335. https://doi.org/10.5860/0700315

Ross-Hellauer, T. (2017). What is open peer review? a systematic review. F1000Research, 6, 588. https://doi.org/10.12688/f1000research.11369.2

Sidorova, A., Evangelopoulos, N., Valacich, J. S., & Ramakrishnan, T. (2008). Uncovering the intellectual core of the information systems discipline. Management Information Systems

Quarterly, 32(3), 467-482. https://doi.org/10.2307/25148852

Van Santen, J. A., Jacob, G., Singh, A. L., Aniebok, V., Balunas, M. J., Bunsko, D., Neto, F. C, Castaño-Espriu, L., Chang, C., Clark, T. N., Cleary Little, J. L., Delgadillo, D. A., Dorrestein, P. C., Duncan, K. R., Egan, J. M., Galey, M. M., Haeckl, F. P. J., Hua, A., Hughes, A. H., Iskakova, D., Khadilkar, A., Lee, J., Lee, S., LeGrow, N., Liu, D. Y., Macho, J. M., McCaughey, C. S., Medema, M. H., Neupane, R. P., O'Donnell, T. J., Paula, J. S., Sanchez, L. M., Shaikh, A. F., Soldatou, S., Terlouw, B. R., Tran, T. A., Valentine, M., Van der Hooft, J. J. J., Vo, D. A., Wang, M., Wilson, D., Zink, K. E., & Linington, R. G. (2019). The natural products atlas: an open access knowledge base for microbial natural products discovery. American Chemical Society Central Science, 5(11), 1824-1833. https://doi.org/10.1021/acscentsci.9b00806

Yoon, J., Jeong, B., Lee, W. H., & Kim, J. (2018). Tracing the evolving trends in electronic skin (e-skin) technology using growth curve and technology position-based patent bibliometrics. Institute of Electrical and Ectronics Engineers Access, 6, 26530-26542. https://doi.org/10.1109/access.2018.2834160

Yoon, J., Park, Y., Kim, M., Lee, J., & Lee, D. (2014). Tracing evolving trends in printed electronics using patent information. Journal of Nanoparticle Research, 16(7), 1-15. https://doi.org/10.1007/s11051-014-2471-6

Young, P. (1993). Technological growth curves: a competition of forecasting models. Technological Forecasting and Social Change, 44(4), 375-389. https://doi.org/10.1016/0040-1625(93)90042-6

Zwietering, M. H., Jongenburger, I., Rombouts, F. M., & Van't Riet, K. J. A. E. M. (1990). Modeling of the bacterial growth curve. Applied and Environmental Microbiology, 56(6), 1875-1881. https://doi.org/10.1128/aem.56.6.1875-1881.1990

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

Choi, Hee-yoon & Seo, Tae-sul (2020). Open Science, the Path to Global Solidarity and Coexistence. Korea Institute of Science and Technology Information. Available: https://repository.kisti.re.kr/handle/10580/15590

Choi, Jae-Hwang & Cho, Hyun-Yang (2005). The recent trends of open access movements and the ways to help the cause by academic stakeholders. Journal of the Korean Society for

Information Management, 22(3), 307-326. https://doi.org/10.3743/KOSIM.2005.22.3.307

Kim, Jungwook, Jeong, Byeongki, & Yoon, Janghyeok (2016). A technology planning approach based on network and growth curve analyses: the case of augmented reality patents. Journal of Korean Institute of Industrial Engineers, 42(5), 337-351. https://doi.org/10.7232/JKIIE.2016.42.5.337

Kim, Pan Jun (2021). A study on the characteristics by keyword types in the intellectual structure analysis based on co-word analysis: focusing on overseas open access field. Journal of the Korean Society for Library and Information Science, 55(3), 103-129. https://doi.org/10.4275/KSLIS.2021.55.3.103

Kim, Sun-Kyum, Kim, Wan-Jong, Seo, Tae-Sul, & Choi, Hyun-Jin (2019). Domain analysis on the field of open access by co-word analysis: based on published journals of library and information science during 2013 to 2018. Journal of Korean Library and Information Science Society, 50(1), 333-356. https://doi.org/10.16981/kliss.50.1.201903.333

Seo, SunKyung & Chung, EunKyung (2013). Domain analysis on the field of open access by co-word analysis. Journal of the Korean Biblia Society for Library and Information Science, 24(1), 207-228. https://doi.org/10.14699/kbiblia.2013.24.1.207

Shin, Jueun & Kim, Seonghee (2021). A study on the intellectual structure of domestic open access area. Journal of the Korean Society for Library and Information Science, 55(2), 147-178. https://doi.org/10.4275/KSLIS.2021.55.2.147

Song, Sungjeon & Shim, Jiyoung (2022). Identification of user preference factor using review information. Journal of the Korean Society for Information Management, 39(3), 311-336. https://doi.org/10.3743/KOSIM.2022.39.3.311

Yoon, Hee-Yoon & Kim, Sin-Young (2007). Trends analysis of open access for foreign and domestic scholarly journals in the field of library and information science. Journal of Korean Library and Information Science Society, 38(1), 257-276. https://doi.org/10.16981/kliss.38.1.200703.257