바이오화학분야 연구 지원을 위한 논문 정보 수집 및 저장 시스템 개발

엄정호[1], 김병정[2]
[1]한국과학기술정보연구원 국가과학기술데이터본부 책임연구원
[2]한국과학기술정보연구원 데이터분석본부 책임기술원
jhum@kisti.re.kr, bjkim@kisti.re.kr

Development of Biochemistry Research Publication Collecting and Archiving System

Jung-Ho Um[1], Byeong-jeong Kim[2],
[1]Div. of National S&T Data, KISTI
[2]Div. of Data Analysis, KISTI

요 약

최근 ESG 경영 등 환경에 대한 관심이 고조됨에 따라, 기존 화학산업을 대체할 수 있는 바이오화학산업이 성장하고 있다. 바이오화학산업규모는 연평균 성장률 10%로 2050년에는 화학산업 시장의 약 50% 정도를 차지할 것으로 예상될 정도로 유망 분야로 성장하고 있다. 본 논문에서는 신산업으로 성장하고 있는 바이오화학분야의 연구자들이 해당 분야의 유망 소재에 대하여 최신 연구정보를 빠르게 파악하고, 미래 유망 바이오화학물질의 발굴등 바이오화학 분야에 다양하게 활용할 수 있도록 관련 논문 정보를 수집, 저장, 검색할 수 있는 시스템을 개발하였다. 해당 수집 논문정보는 바이오화학 산업분류와 연관된 바이오화학물질에 대한 정보와 연계되어 있어, 향후 인공지능 데이터 분석 등에 활용할 수 있는 데이터를 제공할 수 있을 것이라 기대한다.

1. 서론

바이오화학산업규모는 연평균성장률(최근 5년) 10.47%이고, 2023년에는 972억 달러, 2050년에는 전체 화학산업 시장의 약 50%를 점유할 것으로 예측[1]될 만큼, 향후 유망한 분야로서 자리매김하고 있다. 바이오화학산업과 관련하여, 유망한 소재를 발굴하는 것은 매우 중요한 주제로서, 연구자들은 후보물질과 관련된 논문 정보를 찾기 위해 노력할 것이다. 이를 한 사이트에서 관련 정보들을 모아 연구자에게 관련 분야와 유사성 높은 논문정보를 제공하기 위해, 바이오화학분야 오픈데이터 서비스를 구축하였다. 논문을 관련 있는 바이오화학물질 중심으로 분류할 수 있다면 연구자의 검색 편의성을 높일 수 있을 수 있다. 이를 위해, CAS STNext 데이터베이스를 활용하여 바이오화학물질별로 관련 논문을 탐색하였다. CAS STNext 데이터베이스로부터 Web of Science 등의 수집 및 정보 활용이 가능한 논문정보들을 수집하였으며, 이를 이용자들이 쉽게 활용할 수 있도록 논문 메타 정보를 구축하는 서비스에 적재하였다. 이를 통해, 바이오화학분야에 관심 있는 연구자들이 관련 정보를 빠르게 탐색하거나 LDA와 같은 토픽모델링[2]또는 GPT와 같은 언어모델[3]에서 분석할 수 있는 데이터를 제공함으로써 해당 분야 기술 연구 지원에 기여 할 것으로 기대한다.

2. 바이오화학분야 연구 지원을 위한 논문 정보 수집 및 저장 시스템 개발

본 시스템을 개발하기 위하여, 다음과 같이 논문 수집 및 저장 절차를 수립하고, 단계별 필요한 데이터와 시스템을 정의한다. 첫째, 논문 정보를 수집할 바이오화학물질을 선정한다. CAS STNext 데이터베이스에서 선정된 바이오화학물질의 CAS 번호로 검색하여 해당 바이오화학물질과 관련된 논문 제목을 추출한다. STNext 데이터베이스에는 이미 CAS 번호별로 관련 논문 및 특허 등 문헌 정보에 대한 정보가 큐레이팅 되어 제공하고 있다. 둘째, 추출된 논문 제목을 활용하여, 관련 논문에 대한 정보를 ScienceON과 오픈액세스 저널을 검색하여 수집한다. 이를 통해 수집한 논문 정보는 총 18,727건이다. 한편, 논문을 이용하는 사용자가 관심 분야에 따라 논문을 탐색할 수 있도록, 바이오화학산업의 활용분야별로 논문을 분류한다. 먼저, 바이오화학산업을 크게 바이오 플라스틱, 바이오정밀화학, 화장품용 기능성소재, 의료용 화학소재의 4대 대분야로 분류한

다. 바이오 플라스틱은 고무, 플라스틱, 기타의 소분류로 나눈다. 바이오정밀화학은 용매, 화학제품, 연료, 기타의 소분류로 나눈다. 화장품용기능성소재는 향수, 보습재, 기능성, 계면활성제/증점제, 기타로 분류한다. 의료용 화학소재는 치료제, 건강보조식품, 식품첨가제로 분류한다. 해당 분류와 관련된 바이오화학물질들을 찾기 위해서 바이오화학 전문가가 수작업을 통해 연관 바이오화학산업분야별로 분류하였다. 즉, 수집한 논문은 바이오화학물질별로 연관성을 가지고 있기에 바이오화학산업분야와도 연결될 수 있다. 이같이 산업분류로 해당 논문들을 분류한 이유는 분야별 연구자들의 관심도가 서로 다를 것이라고 예상했기 때문이다. 수집한 논문을 각 산업분류별로 나누면 표 1과 같다.

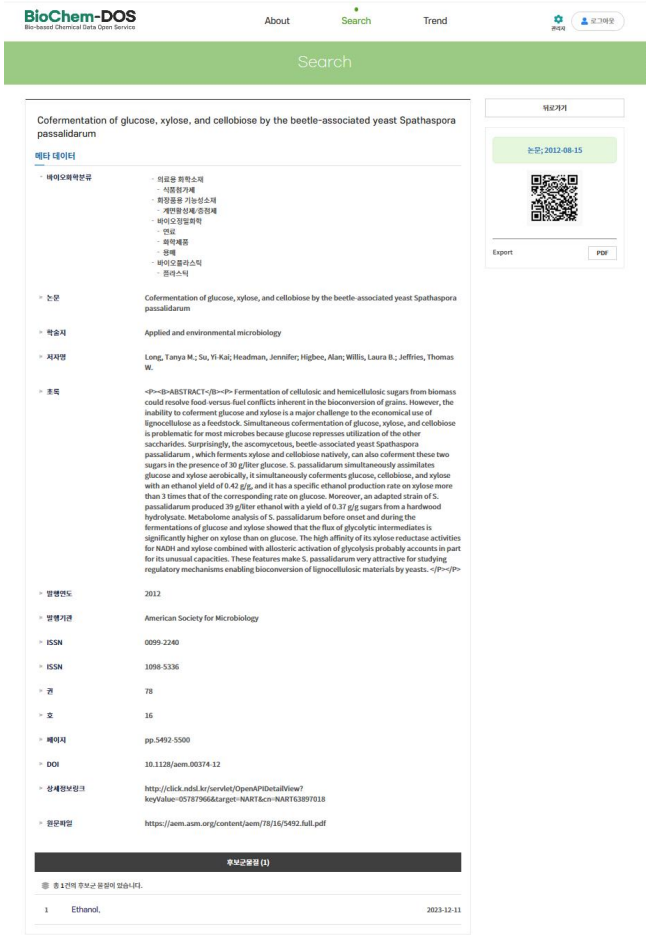
<표 1> 바이오화학산업분류별 수집 논문의 건수

바이오화학 산업분류(대분류)	바이오화학 산업분류(소분류)	논문 건수
바이오플라스틱	고무	6,585
	플라스틱	19,903
	기타	12,829
바이오정밀화학	용매	19,095
	화학제품	20,394
	연료	18,927
	기타	20,609
화장품용기능성소재	향수	2,036
	보습제	1,045
	기능성	4,962
	계면활성제/증점제	20,257
	기타	10,561
의료용 화학소재	치료제	11,375
	건강보조식품	4,822
	식품첨가제	25,473

셋째, 수집한 논문을 저장하고, 검색할 수 있도록 연구데이터 저장 관리 시스템인 NaRDA[4]를 활용한다. NaRDA는 오픈사이언스를 위해 연구데이터를 제출하고, 출판하는 시스템이지만, 기존 논문 리포지터리의 역할도 수행할 수 있을 뿐만 아니라, 향후 관련 분야의 다양한 리소스(연구데이터, 보고서, 특허, 소프트웨어)등을 수집하기 위해 연구데이터 저장 시스템을 활용하여 저장한다. 해당 수집된 논문은 [5]에서 서비스되고 있다. 그림 1은 구현된 서비스에서 수집한 논문 정보 조회 화면을 보인다. 논문의 저자, 초록, 키워드, ISBN/ISSN,권, 호, 페이지, 논문 DOI, 상세 링크 정보를 조회할 수 있다. 아울러, QR 코드도 제공하고 있어, 해당 논문 정보를 바로 활용할 경우, QR 코드를 통한 접속 또한 가능하다.

3. 결론

본 논문에서는 바이오화학분야 연구지원을 위해 논문을 수집 및 저장하는 시스템에 대해 기술하였다. 본 연구에서 수집한 논문은 각 CAS 화합물 및 바이오화학산업분야별로 연결 정보를 제공하고 있다. 따라서, 해당 논문 정보들은 연구자들이 바이오화학분야의 논문 정보를 한 장소에서 바로 탐색할 수 있도록 도와줄 뿐만 아니라 인공지능을 활용한 논문 추천 시스템 등에 학습 데이터셋으로도 활용할 수 있을 것으로 기대한다. 향후 연구로 관련 논문 정보를 지속적으로 수집하고, 이를 자동으로 분류할 수 있도록 인공지능 분류 알고리즘을 본 구축한 서비스에서 제공하는 데이터를 활용하여 적용할 예정이다.



(그림 1) 논문 정보 게시 화면

Acknowledgments

본 연구는 산업통상자원부의 바이오산업 기술개발 사업에서 2020-2024년도에 지원받아 수행된 연구임 (과제번호 : 20008945).

참고문헌

[1] Emerging Trends in Bio-chemicals, frost & sullivanv. 2013.

[2] Jelodar, Hamed and Wang, Yongli and Yuan, Chi and Feng, Xia and Jiang, Xiahui and Li, Yanchao and Zhao, Liang,"Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey", Multimedia tools and applications, vol. 78, pp. 15169-15211, 2019.

[3] Kalyan, Katikapalli S., "A survey of GPT-3 family large language models including ChatGPT and GPT-4", Natural Language Processing Journal, 2023.

[4]Youngho Shin, JungHo Um, Dongmin Seo, Sungho Shin, "Development of a National Research Data Platform for Sharing and Utilizing Research Data", Journal of Information Science Theory and Practice, pp. 25-38, 2022.

[5] 바이오화학오픈서비스, https://idr.kisti.re.kr/biochem-dos/