http://dx.doi.org/10.14386/SIME.2024.32.2.23

디지털 헬스 분야 지식생산기관 식별 정확도 제고 방안 연구*

최성윤** · 문성욱***

<목 차>

국문초록 : 현대에는 개인 연구자 대부분이 지식생산기관에 소속되어 지식생산기관의 유형과 지식생산기관 간의 협력이 과학 지식생산에 미치는 영향이 높음에도 불구하고, 지식생산기관이 정확히 식별되지 않아 과학 지식생산 과정을 실증적으로 파악하는 데는 한계가 있다. 본 연구는 지식생산기관의 식별 정확도를 높이는 방법을 제안하였다. 구체적으로 디지털 헬스 분야의 PubMed 서지정보를 수집한 후 알고리즘을 적용하기 전 데이터 처리 단계에서 '맥락적 연결'을 활용하여 기관정보의 불완전성을 해소하고, 알고리즘 적용단계에서는 기관명 모호성(IND)을 개선하는 방법을 제시하였다. 본 연구가 산출한 '지식생산기관 데이터셋'과 동일한 서지정보를 대상으로 하는 기존 공개 데이터셋인 'PKG datasets'을 비교했을 때, 본 연구가 제시한 방법은 지식생산기관 데이터셋에 포함된 대상 데이터 수를 2배로 증가시

* 이 논문은 2023년도 서강대학교 교내연구비(과제번호: 202312032.01)와 산업통상자원부 산업전문인력역량강화사업(과제번호: P0012783) 지원을 받아 수행된 연구 내용입니다.
** 서강대학교 일반대학원 협동과정 기술경영전공 박사수료 (sychoi@sogang.ac.kr)
*** 서강대학교 기술경영전문대학원 교수, 교신저자 (seongwuk@sogang.ac.kr)

켰으며, 국가별 순위도 보다 정확하게 반영하였다. 또한 한국 지식생산기관의 디지털 헬스 분야 기여도가 과소 또는 과대 평가되고 있다는 사실도 발견하였다. 본 연구에서 제시한 방법은 향후 과학지식을 생산하고 과학 혁신을 달성하는 데 있어 지식생신기관의 역할을 실증적으로 연구하는 데 기여할 것으로 판단된다.

주제어 : 서지정보 기반 지식생산 기관 식별, 서지정보의 부정확성, 저자정보 기반 기관 고유 식별자, 규칙 기반 알고리즘, 정확한 기관 식별의 임팩트

Research on Improving the Identification Accuracy of Knowledge Production Institutions in the Digital Health Field

Choi, Seongyun · Moon, Seongwuk

Abstract : Despite the important roles of institutions and their collaboration in producing knowledge for innovation, the lack of accurate methods for identifying such knowledge-producing institutions has restricted empirical research on the role of institutions in innovation. This study explores methods to enhance the accuracy of identifying institutions involved in innovation process. To this end, we propose ways to improve accuracy in both aspects of information - data and algorithms - using bibliographic information in the digital health field. Specifically, in the data processing stage before applying algorithms, we address contextual inaccuracies of bibliographic information; in the algorithm application stage, we propose methods to improve the ambiguity of institution names (IND). When compared with the PKG dataset, which is publicly available datasets based on the same bibliographic information, our methods doubled the number of cases available for subsequent analysis. We also discovered that the contribution of Korean institutions in the digital health field is either underestimated or overestimated. The method presented in this study is expected to contribute to empirically researching the role of knowledge-producing institutions in innovation process and ecosystem.

Key Words : Author affiliations information; Inaccurate bibliographic information; New methods for affiliations disambiguation algorithm; Impact of precise identification of affiliations

Ⅰ. 서 론

현대 과학기술 분야의 연구자는 과학기술 연구의 폭발적 증가로 인한 지식적 부담(burden of knowledge)으로 특정 기관에 소속되어 연구를 진행하는 추세를 보인다(Jones, 2009; Jones et al., 2008). 대부분의 연구자가 기관에 소속됨에 따라 다양한 유형의 지식생산기관 간의 협력관계는 과학 연구 성과와 진보에 지대한 영향을 미치고 있다. 일례로 최근 코로나라는 갑작스러운 인류의 난제를 극복하면서 생명과학 분야에서의 지식생산기관 간의 협력관계는 더욱 공고해지고 체계화되었다. 그 결과로 통상적으로 9년 이상 걸리던 백신 개발 기간이 무려 9년 이상 앞당겨진 1년으로 단축되었다. 미국의 제약회사 화이자(Pfizer)와 독일의 생명공학기업 바이오엔테크(BioNTech)는 협력하여 화이자의 '백신 개발 노하우와 인프라'를 바이오엔테크의 'mRNA(messenger RNA) 원천기술'과 결합함으로써 'mRNA 백신 플랫폼 기술'을 개발하였고, 이를 활용하여 빠르고 안전하게 백신을 개발하는 과학적 진보를 이루었다. 이러한 현상은 과학기술 연구의 일부 분야에만 국한되지 않으며, 다양한 지식생산기관의 협력에 의한 과학기술의 진보가 이루어지고 있다(Jones, 2009; Jones et al., 2008).

과학 지식을 생산하는 연구자들이 기관에 소속됨에 따라 과학 지식의 유형과 확산 방식 및 경로는 대학이나 기업 등 연구자들이 소속된 기업의 유형에 따라 차이가 생긴다(문성욱 2012; Moon, 2011; Williams, 2013). 특히 지식생산기관의 유형은 과학 지식의 주요 생산 방식인 협력에 큰 영향을 미치고 있다(Bikard et al., 2019). 그러나 지금까지는 지식생산기관의 유형을 정확히 식별하여 지식생산기관의 유형이 과학 지식에 미치는 영향을 밝혀내는 데는 큰 관심이 없었다(Williams, 2013). 또한 지식생산기관 단위에서 실증연구를 진행하기 위해 필요한, 정확도 높은 대규모 지식생산기관 식별 데이터셋이나 기관 식별 알고리즘이 거의 없다는 점도 연구의 한계로 작용한다. 이는 비정형 텍스트 데이터인 논문 서지정보가 완전하지 못하며, 동시에 기관의 명칭 변화나 기관 간 통폐합 등의 동태적 변화를 정확히 반영하지 못한 데서 기인한다. 이와 같은 한계는 지식생산기관의 특성이 과학 지식생산에 미치는 영향을 실증적으로 분석하는 데 있어 여전히 해결해야 할 과제로 남아 있다.

본 연구는 PubMed에서 '디지털 헬스(digital health)' 관련 키워드를 사용하여 추출한 논문들의 서지정보 데이터베이스를 활용하여 지식생산기관을 정확하게 식별하는 방안을 두 가지 측면에서 제시하고자 한다. 첫째, 데이터 측면에서 서지정보 상 기관 관련 정보

가 내포하고 있는 부정확성 및 오류를 최소화하는 방법을 제시한다. 둘째, 고유 식별자(unique key) 기반의 알고리즘을 사용하여 기관명 모호성(Institution Name Disambiguation, IND)을 개선함으로써 지식생산기관의 식별 정확도를 높이는 방안을 제시한다. 이렇게 구축된 데이터셋을 PubMed 서지정보를 이용하여 지식생산기관을 식별한 공개 데이터셋인 'PKG(PubMed Knowledge Graph) datasets'(Xu et al., 2020)과 비교 및 분석하는 과정을 거칠 것이다.

본 논문은 다음과 같이 구성된다. 제2장에서는 지식생산기관 식별의 중요성에 대한 선행 연구를 고찰하고, 제3장에서는 지금까지 서지정보에서 지식생산기관 식별이 어려웠던 이유를 도출하여 개선 방향을 제시한다. 제4장은 PubMed를 활용하여 디지털 헬스 분야의 대규모 서지정보를 수집하는 과정을 설명한다. 제5장에서는 데이터 전처리와 기관명 모호성을 활용한 개선 방안을 제시하였다. 이를 바탕으로 제6장에서는 공개 데이터셋인 PKG datasets과 비교하여 분석하며, 마지막으로 제7장에서는 결론 및 시사점을 도출하였다.

Ⅱ. 과학 발전과 지식생산기관 식별의 중요성

과학기술의 발전 과정을 계량적으로 분석한 최근 연구들에 의하면 현대 과학 분야에서의 연구는 개인적인 활동에서 집단적이고 조직적인 활동으로 성격이 바뀌고 있다(Wuchty et al., 2007; Jones et al., 2008; Stephan, 2012). 과학 연구가 더욱 전문적이고 복잡해지면서 과학 지식생산에서 분업과 협업이 빠른 속도로 증가하였고, 집단적이고 조직적인 연구가 지배적인 방식이 된 것이다(Jones et al., 2008). 이러한 이유로 현대 과학 분야의 연구자는 대체로 연구기관에 소속되어 소속기관의 조직 구조나 관리 방식에 영향을 받으며 분업과 협업을 통해 과학 지식을 생산하고 있다. 연구자의 특성은 연구 환경과 방식의 변화, 그리고 집단 수준 규범의 효과에 의해 강력하게 조절된다는 점에서(Louis et al., 1989) 조직의 특성이 과학 지식생산에 미치는 영향에 관한 연구는 매우 중요하다(문성욱 2012; Moon, 2011; Williams, 2013; Perkmann et al., 2013).

선행 연구에 따르면 지식생산기관의 특성은 과학 지식생산에 다음과 같은 영향을 미친다. 첫째, 기관의 특성에 따라서 지식생산 패턴이 달라진다(문성욱 2012; Moon, 2011). 큰 범주에서 보면 산업계와 학계는 일반적으로 연구에 접근하는 방식이 다르며(Bikard et al., 2019), 이 외에도 공공 분야와 병원 등 구분된 특징을 가진 기관들이 있다. 선행

연구에서는 산업계와 학계의 연구 접근 방식의 차이를 연구 결과 공개에 대한 의사결정 권한(authority)의 분배(Moon, 2011; Murray et al., 2016), 특허 등 지식재산권의 문제(Hall, 2012; Williams, 2013; Sampat and Williams, 2019), 과학자 동기 부여 방식(Azoulay and Manso, 2011) 등 경제적 관점에서 원인을 찾는 경향이 강해지고 있다.

예를 들면 대학의 주요 임무는 지식을 생산하고 전파하는 것이기 때문에(Bush, 1945; Dasgupa and David, 1994), 학계 연구자는 자율성과 기초연구 중심의 개방 과학(open science)을 추구한다는 특징이 있다. 즉 연구자는 자신이 수행하는 프로젝트를 자유롭게 선택하고 자신이 추구하는 연구 방법을 적용하며, 연구 방법과 연구 결과를 창의적으로 공개하고 통제할 수 있다(Aghion et al,. 2008; Moon, 2011).

반면에 기업은 '독점 기술(proprietary technology)'을 갖춤으로써 이윤을 추구한다. 따라서 기업 연구자들은 확실히 상업화될 수 있는 응용 연구를 추구하며, 그 결과를 전유하려는 경향이 강하다. 재량권을 가진 경영진의 지시에 따라 연구를 수행하며, 상업화 가능성이 큰 과제를 강요받는다(Moon, 2011).

공공연구소는 과학기술 분야의 모호성 및 불확실성을 극복하고 전문성과 운영의 신축성을 확보하기 위해 설립되었다(Pollitt et al., 2004). 공공연구소가 정부 기관으로 설립될 경우 계층제로 인한 통제가 강해 연구 수행과 관련된 자율성, 신축성이 현저하게 제약된다. 이에 각국에서는 연구소를 연구공동체에 위임하는 방식이나(예: 독일의 막스 플랑크 협회(Max Planck Gesellschaft), 한국의 국가과학기술연구회 등), 대학 등과의 계약을 맺어 연구를 진행하거나 연구소를 출연하는 방식으로 설립하고 있다. 공공연구소는 설립 목적에 따라 연구 분야 및 과제 선정에 제약이 있으나, 대학 수준의 연구 자율성을 갖는다는 점과 기업처럼 상업화를 목적으로 한다는 점에서 학계와 산업계의 특징을 일부 가지고 있다.

병원은 대학 및 연구소의 기초연구 결과를 기업과 임상 진료 현장으로 연결하는 등 중개연구(translation research)를 실행하여 협력체계의 핵심적인 요소를 담당한다. 병원의 임상 연구는 건강 영역에서의 과학지식 생산의 변화를 현장에서 실증한다는 점에서 더욱 중요해지고 있다(Hicks and Katz, 1996; Thune and Mina, 2016). 이러한 점에서 앞서 언급한 유형의 지식생산기관과 구분되는 특징을 가지고 있지만 지금까지 공공 부문에 속한다고 인식되어 연구가 충분히 이루어지지 않고 있다(문성욱, 2012).

둘째, 지식생산이 협력적 방식으로 급격히 전환됨에 따라 협력하는 기관들의 유형과 특성이 협력적 지식생산의 성과에 영향을 미친다. Jones et al.(2008)은 대학 간 협력 연구의 유형을 단독 연구, 동일 대학 내 협력 연구, 타 대학 간 협력 연구 세 가지로 분류하였으며, 분석 결과 타 대학 간 협력 연구가 크게 증가하였다고 주장하였다. 특히, 대학

간의 계층화(Stratification by university rank)가 심화하면서 협력을 통해 과학 지식을 생산하는 데 있어서 엘리트 대학의 역할이 매우 중요해졌다고 주장하였다(Jone et al., 2008). 또한 Agrawal and Goldfarb(2008)는 BITNET의 도입을 '기술 충격(technology shock)'으로 간주하고 협력 주체들의 특성이 공동연구에 미치는 영향을 분석하였는데, 여러 협력 연구 유형 중 엘리트 대학과 중간층 대학 간의 협력이 가장 높은 비중을 차지하는 것을 확인하였다.

연구 능력이 상이한 대학 간의 공동연구 외에 학계와 산업계의 공동연구의 증가 추이에 대한 분석을 보면, Adams et al.(2005)은 논문 저자로 참여한 200개의 중점기업을 활용하여 대학-기업 간 공동연구의 변화 추이를 측정했다. 그 결과 1981년부터 1990년까지 민간기업의 참여도는 약 2배 증가하였으며, 특히 1990년대 초반에 증가 속도가 급격히 빨라졌다. 따라서 민간기업 또한 엘리트 대학과 마찬가지로 협력을 통한 과학 지식생산에 적극적으로 기여하고 있는 것으로 판단된다. 산학협력의 효과와 관련해서 Bikard et al.(2019)은 업계 협력자가 연구 프로젝트에 참여할 경우 여러 과학자가 동시에 거의 동일한 발견을 하는 경우인 '동시 발견'에서 성공 가능성을 높여 그렇지 않은 연구 프로젝트보다 질적으로 높은 효과를 볼 수 있다고 주장하였다.

이처럼 지식생산기관의 유형과 협력이 과학 지식생산에 미치는 중요도를 분석한 선행연구들은 다수 있지만, 지식생산기관을 체계적으로 식별하여 지식생산기관에 따른 지식생산 방식과 협력 유형, 그리고 성과 분석에 활용한 연구는 수는 많지 않다(문성욱 2012; Moon, 2011; Williams, 2013). 또한 기존의 지식생산기관 식별 데이터는 서지정보 상 불완전한 기관정보와 같은 기관을 표현하는 다양한 명칭을 다른 기관으로 식별하는 알고리즘 등으로 인해 지식생산기관의 역할을 분석하는 데 어려움이 있다. 이에 본 연구는 서지정보 상 기관정보를 정확하게 식별해 내는 방안을 제시하고자 한다.

Ⅲ. 지식생산기관 식별의 한계

본 장에서는 지식생산기관 식별에 활용되는 논문 서지정보의 부정확성과 지식생산기관을 식별하는 알고리즘, 두 가지 측면에서 지식생산기관 식별의 어려움을 살펴본다. 이를 위해 지식생산기관 식별정보를 포함하고 있는 대규모 공개 데이터셋인 PKG datasets을 활용하여 서지정보의 부정확성을 유발하는 요인의 유형을 분류한다.

1. 부정확한 서지정보

논문의 서지정보는 출판연도, DOI(Digital Object Identifier) 등의 정형 데이터뿐만 아니라 논문 요약(abstract)과 저자 소속(affiliation) 등의 비정형 데이터로 구성되어 있다. 특히 저자 소속 정보는 개인과 기관 등 다양한 분석 단위(unit of analysis)에 대한 정보를 포함하고 있어 서지정보 분석에서 활용도가 높은 항목이지만 저자 소속 문자열을 나열하는 표준 형식이 없어 정확한 정보를 추출하기 매우 어렵다. 저자 소속 정보는 기관, 도시, 주 및 국가의 순서로 나열하는 약간의 문화적 선호도가 있는 자유형식 텍스트(free form text) 필드이지만, 다양한 스타일과 형식 변형을 보여(Jonnalagadda and Topham, 2010; Torvik, 2015) 원 데이터를 직접 활용하기 매우 어렵다. 서지정보 상 저자 소속 식별을 부정확하게 만드는 요인은 정보 부족, 정보 중첩, 정보 오류, 부가 정보, 기타 다섯 가지 유형으로 분류할 수 있다(<표 1> 참고).

먼저 '정보 부족' 유형은 전체 주소가 나오지 않거나 특정 단어만 존재하는 경우이다. 정보가 부족할 경우 국가명이나 저자의 소속기관 등을 정확히 알 수 없어 저자가 소속된 기관 식별을 어렵게 하는 가장 큰 원인이다(Torvik, 2015). 정보가 부족하게 되는 이유로 관행(practice)에 따라 저자 소속에서 중복되는 단어를 생략하는 경우를 들 수 있다. 대부분은 소속기관에 포함되어 있는 지리정보(도시, 국가 등)가 생략되지만, 저자의 소속기관이 생략되는 경우도 있으며, 반대로 소속기관만 제공되거나 부서 정보만 존재하기도 한다.

'정보 중첩' 유형은 저자와 저자 소속이 제대로 연결되지 않는 오류이다. 아무리 저자 소속 정보가 정확하더라도 소속 정보가 저자와 제대로 연결되어 있지 않으면 분석 결과의 신뢰성을 보장하기 어렵다. 구체적으로 공저자가 있는 경우 저자 소속 정보에 나열된 두 개 이상의 공저자의 소속 정보를 하나의 소속 정보로 중첩하여 특정 저자[1]에게만 연결하거나, 공저자의 소속 정보를 하나로 합쳐서 모든 저자에게 일괄적으로 연결하는 경우가 있다. 예를 들어 <표 1>의 정보 중첩의 첫 번째 사례(PMID:25989387)를 보면 PKG datasets에서는 '저자-저자 소속'을 제대로 연결하지 않았다. 모든 공저자의 소속기관을 'University of Washington'으로 분류함에 따라 협력 연구에서 가장 큰 비중을 차지하는

1) 주로 제1 저자에 해당하지만, 교신저자에 나타나는 경우도 있다(PMID: 28157405). 따라서 제1 저자만을 한정해서 분석하는 경우 잠재적인 오류를 내포하고 있다는 것을 고려해야 한다.

산학협력을 식별할 수 없다. <표 1>의 두 번째 사례(PMID:21564255)는 제1 저자에게만 중첩된 정보를 연결함으로써 저자의 국가 및 소속기관이 잘못 식별되었다. 정보 중첩은 한 명의 저자가 겸직하고 있는 다수의 기관을 표시하는 경우와는 구별된다.[2)]

'정보 오류' 유형은 논문의 서지정보와 실제 논문의 정보가 다른 경우이다. 이를 자동으로 찾아내는 것은 거의 불가능하다. 이러한 경우는 매우 적기는 하지만 논문 서지정보에 어느 정도의 오류가 존재한다는 점을 인정해야 한다.

'부가 정보' 유형은 저자 소속에 저자의 이력이나 연구 분야 등이 서술된 경우나 일반적으로는 저자 소속에 포함되지 않는 부수적인 정보(예: "The first two authors should be regarded as joint First Authors" 혹은 "current affiliation")를 포함하는 경우이다. 이러한 부가 정보는 일반적인 형식(주소 형식)에서 벗어나기 때문에 저자 소속을 식별하는 데 장애 요소가 된다.

마지막으로 '기타' 유형에는 잘못된 구분자와 같은 특수한 형태가 포함된다. 예를 들어 <표 1>의 첫 번째 사례(PMID: 323413796)는 두 개의 저자 소속 간 구분자가 존재하지 않아(예: LondonBermuda) PKG datasets에서 저자의 국가를 영국(UK)이 아닌 버뮤다(Bermuda)로 식별한 것을 보여준다. 이러한 오류는 국가 간 비교 시 왜곡된 결과를 산출하도록 한다. 또한 두 번째 사례(PMID: 31986076)와 같이 고유 명칭이나 특정 단어에 대한 오탈자(예: ollege) 등이 있다. 이외에도 텍스트 기반의 비정형 데이터에서 비롯되는 다양한 오류들은 정확한 지식생산기관의 식별 난도를 높이는 요인으로 작용한다.

이러한 요인들로 인해 국가, 기관 등의 주요 정보가 분석 대상에서 누락되어 의도하지 않은 선택 편의(selection bias)를 유발하거나, 잘못된 변수 추출로 인한 왜곡된 분석 결과를 산출할 위험이 존재한다. 따라서 논문의 저자 소속 정보에서 지식생산기관을 식별하기에 앞서, 서지정보 중 비정형 데이터로 구성된 저자 소속 정보의 특징을 고려한 데이터 처리가 필요하다.

2) 겸직의 경우 하나의 문자열에서 세미콜론(;)과 같은 구분자로 분리되거나, 개별 문자열이 별도의 행으로 구분되어 있는 경우도 있다.

<표 1> 저자 소속 오류 유형

유형	설명
정보 부족	**전체 주소가 나오지 않거나 특정 단어만 존재**(PMID: 31923894)
	1 Department of Pediatrics. 2 Institute for Computational and Mathematical Engineering, and.
정보 중첩	**전체 저자 소속 정보가 합쳐져서 표기되는 경우**
	(PMID: 25989387) From the Department of Radiology, University of Washington, 1715 Columbia Road N, Portage Bay Building, Suite 222, Seattle, WA 98195-7987 (L.A.P., B.F.E., P.E.K.); PixelMed Publishing, Bangor, Pa (D.A.C.); and MiM Software, Cleveland, Ohio (D.N.)
	(PMID: 21564255) Computing and Information Science, Masdar Institute of Science & Technology, UAE Technology & Development Program, Massachusetts Institute of Technology, USA School of Informatics, University of Edinburgh, UK Faculty of Informatics, British University in Dubai, UAE CLLE, CNRS and University of Toulouse, France Faculty of Education, British University in Dubai, UAE
정보 오류	**서지정보와 논문의 정보가 서로 다른 경우**(PMID: 26121100)
	(서지정보: 7명) Philip Toltzis, Gerardo Soto-Campos, Christian R Shelton, Evelyn M Kuhn, Ryan Hahn, Robert K Kanter, Randall C Wetzel (논문저자: 6명) Toltzis, Philip; Soto-Campos, Gerardo; Kuhn, Evelyn M.; Hahn, Ryan; Kanter, Robert K.; Wetzel, Randall C.
부가 정보	**저자의 이력에 대한 설명을 기술하거나 특정 저자의 기여도, 현 소속기관 등의 부가적인 정보가 포함된 경우**(PMID: 27477463)
	Department of Bioengineering, University of Washington, Seattle WA, 2419 8th Ave N Apt 402, Seattle, WA 98109. This author's research studies the underlying cortical organization of human sensorimotor function, specializing in multi-day micro-electrocorticographic array recordings in human subjects. This type of research is applicable in many areas of cognitive neuroscience, from brain-computer interfacing to neural engineering and robotics.
기타	**특정 단어나 명칭의 오탈자 등**
	(PMID: 23413796) Library and Information Manager, Royal College of Psychiatrists, LondonBermuda Hospital Board, King Edward VII Memorial Hospital, Health Sciences Library, Paget, Bermuda
	(PMID: 31986076) Biostatistics(Mr Long) and Proficiency Testing (Ms Vasalos), *ollege* of American Pathologists, Northfield, Illinois

2. 지식생산기관 식별 알고리즘의 한계

과학기술 분야 논문의 서지정보에서 저자의 소속 기관 식별을 위한 기관 정보 데이터를 구축하는 것은 매우 어렵다(Caron and Daniels, 2016). 그 이유로 첫째, 지식생산기관의 명칭은 특정 국가나 기관에서 코드화하여 통합적으로 관리되는 데이터가 아니다. 공식 행정 기관이 통일된 형식으로 기관 목록을 작성하고 갱신하는 것이 최선의 방식이지만 현실적으로 쉽지 않으며, 저자 또는 출판사마다 사용하는 형식의 차이로 인해 더욱 어렵다.

둘째, 시간의 흐름에 따른 기관의 설립, 폐지, 변경, 합병, 분할 등의 동태적인 변화를 포착하지 못한다. 예를 들면 '프랑스 국립 농업 연구원(INRA)'과 '프랑스 물 환경 및 농업 기술 연구소(IRSTEA)'는 2020년 1월 1일 '프랑스 국립 농업 연구원(INRAE)'으로 합병하였다. 그러나 명칭이 변경된 후에도 연구자들이 여전히 이전 명칭을 사용하는 경우가 있다.

셋째, 기관이 약어 등 다양한 명칭을 가지고 있거나 번역(비영어권 국가) 등으로 인한 표기상의 오류가 발생할 수 있다. 이러한 오류를 기술적으로 해결하는 것은 쉽지 않다. 예를 들면 벨기에의 다른 도시에 소재한 두 대학인 'Katholieke Universiteit Leuven'과 'Université catholique de Louvain'은 둘 다 영어로 번역할 경우 'Catholic University of Louvain'으로 번역된다.

넷째, 오탈자가 자주 발생한다. 특히 기관 명칭의 주요 부분에서 철자 오류가 자주 발생하는데 이는 저자에 의해 혹은 데이터베이스를 생성하는 중에 발생한다.

다섯째, 연구의 주체가 되는 기관을 어느 수준에서 입력하는지에서 차이가 발생한다. 일부 저자는 논문의 주요 기관명으로 대학명을 적는 반면, 일부 저자는 학과명, 부서명 또는 캠퍼스명만 나열하는 경우도 있다. 부서명만 보고 기관과 해당 부서 간의 정확한 소속 관계를 인지하는 것은 어렵지만, 신뢰성 있는 결과를 얻기 위해서는 부서명에서 유추하여 기관명을 식별할 필요가 있다.

<표 2>는 규칙 기반 알고리즘이 적용된 PKG datasets의 한계 사례이다. 'Harvard Medical School'은 여러 명칭으로 입력되어 있어서 같은 소속이더라도 다른 기관명이나 잘못된 기관명으로 분류되었다. 이러한 오류는 기관을 잘못 식별함으로써 기관의 성과를 과대 또는 과소 측정하거나, 기관 간 협력관계를 측정하지 못하거나 잘못된 상관관계를 형성하여 왜곡된 연구 결과로 이어질 수 있다(Huang et al., 2014).

<표 2> PKG datasets의 기관명 분류 오류

저자 소속 정보	분류된 지식생산기관명
Harvard Medical School, Boston, MA, USA	Harvard Medical School
1 Harvard Medical School, Boston, MA, USA	1 Harvard Medical School
7Harvard Medical School, Boston, MA 02115 USA	7Harvard Medical School
a Harvard Medical School, Boston, MA, USA	a Harvard Medical School
Harvard Medical School, Boston, MA, USA	Massachusetts General Hospital, Harvard Medical School, Shriners Burns Hospital
Harvard Medical SchoolBostonMA02115USA	Harvard Medical SchoolBostonMAUSA.

이러한 문제를 해결하기 위해서 다양한 시도가 있었다. 그중 지식생산기관 식별에 가장 많이 적용된 방법은 규칙 기반(rule-based) 접근법이다(Jonnalagadda and Topham, 2010; Huang et al., 2014; Torvik, 2015; Caron and Daniels, 2016; Xu et al., 2020). Yu et al.(2007)은 저자 소속 문자열을 기반으로 연구자의 프로필을 생성하기 위해 기관명을 추출하는 방법을 제시하였고, Jonnalagadda and Topham(2010)은 저자 소속 문자열에서 기관명을 추출하여 동의어 사전을 구축한 후 이를 활용하여 정식 기관명으로 정규화(normalization)하는 방안을 제시하였다. 유사한 연구로 Torvik(2015)는 저자 소속 문자열을 세분화하여 지리정보를 정확하게 연결시켰으며, Xu et al.(2020)은 이를 확장하여 기관의 명칭을 다른 정보들과 통합하여 학술적 영향, 지식 사용 및 지식 이전을 측정하는 연구를 수행하였다. 규칙 기반 접근법은 특정 작업 수행에 적합한 규칙이나 조건을 만들어 적용하기 때문에 단순하면서도 제한된 영역에서는 높은 정확도를 보이지만, 다양한 형태의 지식생산기관의 명칭과 문자열의 변형(예: 약어, 구두점, 띄어쓰기)을 고려해 단일 기관으로 정확히 식별하는 데는 한계가 있다(Huang et al., 2014; Caron and Daniels, 2016).

최근에는 GPT(Generative pre-trained transformer) 기반의 최신 기술인 개체 인식(Named Entity Recognition, NER)을 사용하려는 시도가 있지만, 주석이 달린 데이터가 많이 존재해야 한다는 한계가 있어 적용하기 어렵다(Jonnalagadda and Topham, 2010; Caron and Daniels, 2016).

3. 개선 전략

논문 서지정보에서 지식생산기관을 정확히 식별하는 것은 지식생산기관 특징에 따른 지식생산 활동 유형과 협력관계를 밝히는 데 도움이 되지만(문성욱, 2012; Moon, 2011), 그 가능성에 비해서 활발하게 연구가 진행되고 있지 않다. 특히 논문 서지정보의 부정확성을 지적한 연구는 거의 없으며, 현재까지 진행된 관련 연구는 대부분 연구 방법으로 규칙 기반 접근법을 활용하고 있다. 하지만 규칙 기반 접근법은 다양한 지식생산기관의 명칭과 문자열의 변형을 모두 고려하는 데는 한계가 있다. 본 연구는 규칙 기반 접근법에 기반하여 지식생산기관 식별의 정확도를 높이기 위해서 앞서 살펴본 서지정보 상 저자 소속 정보의 부정확성과 기존 지식생산기관 식별 알고리즘의 한계를 극복하는 방안을 제시하고자 한다.

첫째, 서지정보에서 저자와 저자 정보의 맥락적인 연결을 고려하여 입력 데이터의 정확도를 향상한다. 기존 연구는 서지정보의 저자 소속 텍스트를 그대로 사용하여 지식생산기관을 식별하였다. 그러나 정보 중첩의 오류와 같이 저자와 저자 소속이 제대로 연결되지 않은 경우가 다수 존재한다. 대부분의 서지정보 제공업체는 저자나 출판사에서 제공하는 정보를 그대로 사용하여 데이터가 부정확할 가능성이 높다. 하지만 입력 요소인 원천데이터의 정확성 여부가 연구 결과에 결정적인 영향을 미친다. ‘Garbage in - garbage out’로 표현되는 바와 같이 품질이 낮은 데이터 입력은 신뢰할 수 없는 데이터 출력으로 이어지기 때문이다. 수집된 정보는 매우 정확해야 하고, 그렇지 않으면 데이터 분석, 애플리케이션 또는 비즈니스 프로세스를 신뢰할 수 없게 된다(Kilkenny and Robinson, 2018). 이는 지식생산기관의 식별에도 마찬가지로 적용된다. 따라서 서지정보의 부정확성을 제거하고 정확한 데이터 입력을 위해 불필요한 단어(stopwords)와 구두점(punctuation) 등을 제거하고 저자와 저자 소속을 정확하게 연결하는 전처리 과정을 거친다. 최종적으로 순수 문자만으로 구성된 값을 기준으로 전처리된 데이터 간 중복을 제거함으로써 식별 과정의 일관성을 높이고 식별 대상도 줄여 식별 절차의 효율성을 높일 수 있다.

둘째, 규칙 기반 접근법의 한계로 지적되는 기관명 모호성을 개선하기 위해 데이터 정규화와 고유 식별자를 활용한 알고리즘을 적용한다. 이를 통해 ‘대표 저자 소속’을 식별함으로써 다양한 형태로 존재하는 저자 소속을 하나의 단일한 기관명으로 연결한다. 다양한 지식생산기관의 명칭을 단일 개체에 연결하여 신뢰할 수 있는 테이블을 생성하는

것은 중요하다(De Bruin and Moed, 1990). 구체적으로는 앞서 전처리 과정을 거친 저자 소속의 약어와 관용어 등을 데이터 정규화를 통해 표준 용어로 대체한다. '지리정보 사전(dictionary)'을 활용하여 지리정보(예: 도시, 국가)를 식별하고, 지식생산기관의 속성(예: 기관 유형, 부서명)을 별도의 변수로 생성하여 저자 소속에서 분리한다. 최종적으로 지식생산기관 식별에 불필요한 단어를 제거하고 고유 식별자를 부여하는 알고리즘을 사용하여 저자가 소속된 기관명의 변이를 하나의 기관명으로 전환한다. 이를 통해 기관명을 정확하게 식별하는 확률을 높일 수 있다.

Ⅳ. 디지털 헬스 분야 서지정보 구축

지금까지 지식생산기관의 중요성과 식별의 어려움 및 개선 전략에 대해서 논의하였다. 본 장에서는 PubMed 서지정보의 특징을 살펴보고, PubMed에서 대규모 서지정보를 수집하는 방법을 살펴본다.

1. PubMed 서지정보의 특징

본 논문에서는 PubMed 데이터베이스를 활용한다. PubMed는 의사, 의사 과학자, 바이오 분야 등에서 연구하는 연구자 등에게 필수적인 데이터베이스이다. MeSH(Medical Subject Headings)와 같이 상용 서지정보 데이터베이스에서는 제공하지 않는, 미국국립보건원(National Institutes of Health, NIH)에서 자체적으로 분류하고 관리하는 가치 있는 정보도 제공하기 때문에 생명과학 및 의학 분야에서 대체 불가능한 데이터베이스라고 할 수 있다. 특히 서지정보 검색 시 MeSH를 활용하면 민감도(sensitivity)와 특이도(specificity)가 향상되어 연구자가 검색하고자 하는 주제를 더 정교하게 검색할 수 있다는 장점이 있다(Torvik, 2015).

PubMed 서지정보는 다음과 같은 고유의 특징을 가지고 있다. 첫째, 저자의 형식은 성(last name)과 이름(first name) 그리고 이니셜(initial)로 구성된다. 2002년 이전에는 성과 이름을 포함한 저자명(full name)이 PubMed 인용문에 포함되지 않았기 때문에, 2002년 이후부터 논문에 성과 이름이 함께 게재되는 경우에만 제공되고 있다(PubMed Help,

2020). 또한 수년 간 저자 수를 표기하는 방법에 대한 규정을 만들고 수정하며 유지해 왔다(<표 3> 참조). 1983년까지는 저자 수를 제한하지 않았으나 이후 1995년까지는 10명으로 제한하였고 1999년까지는 25명까지로 제한하였으며, 2000년 이후로는 다시 저자 수에 대한 제한을 두지 않는 등 일련의 변화를 거쳤다.

<표 3> 저자 수에 따른 표기 규정 변화

기간	규정 변경사항
1966-1983	이 기간 동안 생성된 레코드의 경우 저자 수를 제한하지 않음
1984-1995	저자 수를 10명으로 제한하고 추가 저자의 존재를 나타내는 “et al.” 표시 (1983년 10월 29일에 작성된 인용부터 시작)
1996-1999	저자 수를 10명에서 25명으로 증가 (저자가 25명 이상이면 처음 24명을 나열하고, 마지막 저자를 25번째로, 26번째 이상을 “et al.”로 대체)
2000-현재	2000년에 발행된 저널 발행 호부터 목록 저자 수를 제한하지 않음

출처: https://www.nlm.nih.gov/bsd/mms/medlineelements.html#au 재구성

둘째, PubMed의 저자 소속은 처음에는 동명이인 문제를 해결하기 위해서 기록하기 시작했다. 출판사에서 서지정보를 제출한 경우 저자, 기업 저자 및 연구자의 소속이 포함되며, 저자 소속 필드에 “동등하게 기여(Contributed equally)”라는 노트를 포함한다(PubMed Help, 2020). 그러나 모든 출판사가 미국국립의학도서관(National Library of Medicine, NLM)에 제출하는 자료에 저자 소속을 제공하는 것은 아니며, 시간이 지남에 따라 색인 규정도 변경되었다(<표 4 참조>). 예를 들어 1988년부터는 첫 번째 목록에 있는 저자에 대해서만 저자 소속을 체계적으로 색인화하기 시작했으며,[3] 1995년부터는 필요한 경우 “USA”가 저자 소속 정보 끝에 추가되었다. 1996년부터 이메일 주소가 추가되었고 1999년에는 거리 정보 또는 중복 데이터 삭제를 위해 저자 소속 편집을 중단하였다(NLM Technical Bullin, 1999). 2013년에는 저자 소속에 대한 편집 및 품질 관리를 중단하였고(NLM Technical Bullin, 2013), 2014년에는 저자 소속 정보를 논문에 연결하던 규정을 논문의 저자에 연결하는 것으로 변경하고(NLM Technical Bullin, 2014), 인용문에 복수의 소속기관을 추가하였다.

3) 교신저자에게만 입력된 경우도 존재한다(PMID: 25436904).

<표 4> 저자 소속 표기 규정 변화

변경 시점	변경 사항
1988-	미국의 경우 제1저자의 소속과 함께 학회지가 제공하는 경우 우편번호, 기관, 도시 및 주를 포함하였고 미국 외 국가는 학회지가 제공하는 경우 소속 및 주소를 포함
1995-2013	“USA” 명칭은 첫 번째 저자의 소속이 미국 50개 주나 콜롬비아 특별구에 있는 경우 주소 끝에 추가
1996-	저널에 있는 경우 기본 저자의 전자메일(e-mail) 주소가 소속 필드 끝에 포함
2003-	완전한 제1저자 주소는 생략된 단어 없이 논문에 나타난 대로 입력
2013.10-	모든 저자 및 기여자의 소속을 수용하기 위해 이 필드의 품질 관리(Quality Control) 중단
2014.12.-	각 저자 또는 기여자에 대한 여러 소속을 포함[4]

출처: https://www.nlm.nih.gov/bsd/mms/medlineelements.html#ad 재구성

저자 소속 데이터는 출판사가 제공한 대로 입력되며, 미국국립의학도서관은 가능한 경우 “기관 구분, 기관 이름, 도시, 주, 우편번호, 국가(미국의 경우 USA), 마침표, 공백 뒤에 이메일 주소” 형태의 데이터를 쉼표로 구분하여 포함하도록 요청하고 있다.[5]

2. 디지털 헬스 분야 PubMed 데이터의 수집

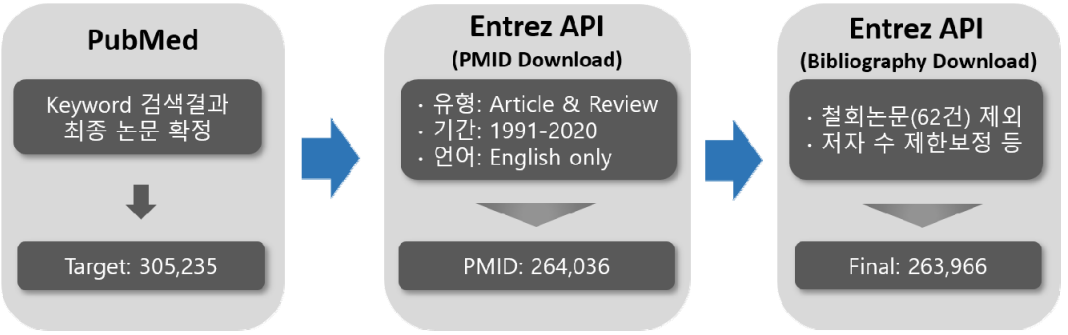
최근 정보기술 발달에 따라 서지정보에 API(Application Programing Interface) 방식의 사용자 인터페이스를 활용하는 것이 보편화되었다. 많은 연구자가 기하급수적으로 증가한 대규모 데이터의 효율적인 관리와 활용을 위해 데이터베이스 관리 시스템(Database Management System, DBMS)을 활용하는 추세에 있다(Varian, 2014; Einav and Levin, 2013)

PubMed에서도 프로그래밍을 활용하여 논문 서지정보를 자동으로 쉽게 다운로드할

4) 2014년 10월 이전에도 제1 저자의 소속 정보에 모든 공저자의 소속 정보를 제공하는 경우가 있으며, 특정 학회지의 경우(예: IOP Publishing) 2014년 이후에도 전체 저자에 대한 저자 소속 정보를 제공하지 않아 직접 논문을 확인하여 보완하였다.

5) https://www.nlm.nih.gov/bsd/mms/medlineelements.html#ad 참조.

수 있는 Entrez API를 제공하고 있다. 본 연구는 이를 활용하여 대용량 데이터를 수집하고[6] 수집된 데이터를 오픈소스 데이터베이스를 활용하여 저장하였다.



<그림 1> 디지털 헬스 데이터베이스 구축을 위한 데이터 수집 절차

본 연구에서 사용한 논문들의 다운로드 절차는 <그림 1>과 같다. 1) PubMed 웹사이트에서 키워드 검색을 통해서 최종 대상 논문을 선정하고(305,235건) 2) 해당 키워드에서 논문 유형(Article 및 Review), 검색 기간(1991-2020), 출판 언어(English) 등의 필터를 적용하여[7] Entrez API로 대상 Article의 PMID를 먼저 다운로드하였다(264,036건). 3) 다운로드 한 PMID를 기반으로 다시 Entrez API를 사용하여 개별 PMID에 대한 기초 데이터(Raw data)를 다운로드하였다(263,966건)[8].

검색에 사용된 키워드는 디지털 헬스 분야의 주요 키워드인 'Artificial Intelligence', 'Machine Learning', 'Digitalization', 'Digital Health'를 포함하는 검색 키워드 조합을[9] 사용하였다.

---

6) 자세한 내용은 https://www.ncbi.nlm.nih.gov/home/develop/api/ 참조.

7) 논문의 출판 유형을 학술논문(article)과 리뷰논문(review)으로, 언어를 영어로 제한하면 26,401건이 감소한다. 그리고 1991년 이전에 출간된 논문 13,799건, 2020년에 게재 확정되었으나 실제로는 2021년에 게재된 논문 999건을 제외하여 총 264,036건을 대상으로 하였다.

8) 게재되었다가 철회된 논문 62건을 제외하였다. 그리고 MEDLINE의 저자 수 제한 정책의 영향을 받는 논문 87건 중에서 실제 논문을 찾지 못한 7건을 제외하고, 나머지 80건의 저자 데이터를 보정하여 총 263,966건을 최종 확정하였다.

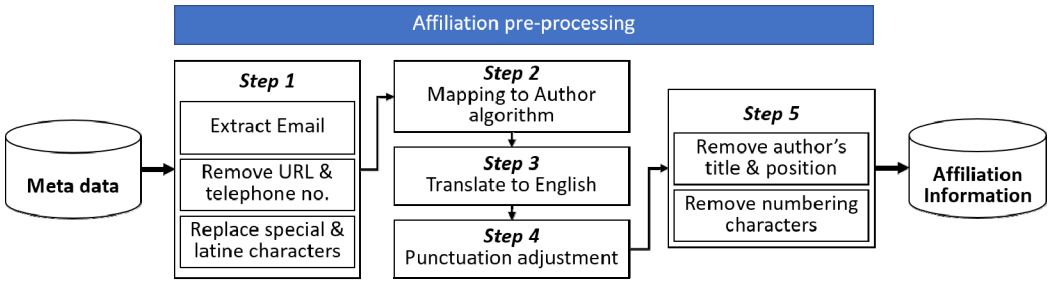
9) (artificial intelligence OR machine learning OR digitalization OR Digital Health) AND (("1862"[Date - Publication] : "2020"[Date - Publication]))

Ⅴ. 지식생산기관 식별 개선: 데이터와 알고리즘

본 장에서는 앞서 수집한 디지털 헬스 분야 서지정보 중 저자와 저자 소속 정보의 맥락적 연결 등을 통해 데이터의 부정확성을 개선하고, 고유 식별자를 활용한 규칙 기반의 알고리즘을 적용하여 지식생산기관 식별 정확도를 높이는 방안을 제시한다.

1. 저자 소속 데이터 전처리: 서지정보의 부정확성 개선

논문 서지정보의 저자 소속에는 국가명, 저자명, 기관명과 같이 명확하게 확인되어야 하는 다양한 개체가 포함되어 있으나, 다양한 입력 형식과 코딩 규칙(Richardson, 2010)으로 인해 일부 모호함은 불가피하다. 본 절에서는 서지정보에 내재한 오류를 개선하기 위해 저자와 저자 소속의 맥락적 연결을 위한 데이터 전처리 방안을 제시한다. 데이터 전처리는 지식 발견 및 비즈니스 분석 이전에 필요한 단계로(Caron and Daniels, 2016) 서지정보 분석에서도 마찬가지로 적용된다. 저자 소속의 원본 데이터는 5단계의 전처리 과정을 거친다(<그림 2> 참조). 1단계는 '사전 준비 단계'이다. 정규 표현식(regular expression) 등을 사용하여 보조적으로 활용이 가능한 이메일과 URL, 전화번호 등의 정보를 추출하여 별도의 변수로 생성하고 특수문자 및 라틴문자는 본래의 의미에 맞는 단어로 대체하였다.



<그림 2> 서지정보 전처리 절차

2단계는 '저자-저자 소속' 연결 단계이다. 데이터 측면에서 서지정보의 부정확성을 높이는 정보 중첩 오류를 해결하는 중요한 과정이다. 대부분의 경우에는 중첩된 저자 소속

에 저자의 이니셜을 제공하고 있어 이를 기준으로 중첩된 정보를 분리하고 저자명과 매칭하는 별도의 알고리즘을 구현하였다. 구체적으로는 두 가지 유형으로 나누어 먼저 저자 소속이 완전히 존재하는 경우는 이니셜을 기준으로 저자 소속을 분리하여 바로 연결하고, 그 외 반복되는 정보(예: 도시 및 국가 정보)가 생략된 경우에는 앞뒤 문자열에서 생략된 부분을 유추하여 완전한 문자열을 구성한 뒤 '저자-저자 소속'을 연결하도록 구현하였다. 다만 정보 중첩의 약 16%가 저자의 이니셜을 포함하지 않고 있어 실제 논문을 확인하여 직접 연결하는 과정을 거쳤다.

3단계는 '영어 번역' 단계이다. 저자 소속 정보를 영어 이외의 언어로 작성한 경우 번역하여 영어로 전환하였다[10]. 이는 논문은 영어로 작성하였어도 저자 소속 정보는 저자의 모국어로 작성한 경우가 다수 존재하기 때문이다.[11] 이 경우 동일한 저자 소속 정보가 서로 다른 두 개의 기관으로 분류될 수 있으므로 영어로 전환하는 작업이 필수적이다(Jonnalagadda and Topham, 2010). 그러나 최종 기관명은 해당 기관의 관례(practice)를 반영하였다. 예를 들어 드레스덴 공과 대학의 경우 'Technische Universität Dresden'을 고유명사이자 상표로 등록하여 항상 독일어 기관명을 사용하도록 공식적으로 권고하고 있다.

4단계는 '구두점 제거' 단계로 불필요한 구두점 및 용어(예: ¶, §, ‖) 등을 제거하였다. 또한 PubMed는 의미 있는 데이터를 쉼표(comma)로 구분하는 것을 원칙으로 하고 있어(PubMed Help, 2020) 쉼표가 아닌 마침표(period)로 구분된 경우 마침표를 쉼표로 대체하였다.

5단계는 '불용어(stop-words) 제거' 단계이다. 저자의 직급이나 직위와 관련된 항목(예: Professor, Student, MD, MS, phD 등)과 개별 저자 소속 정보에 부여된 순번 혹은 순서 문자 등 데이터 분석 과정에서 오류가 발생할 가능성이 있는 항목들을 제거하였다.

마지막으로 전처리된 저자 소속 데이터에서 구두점 및 대소문자 등으로 인한 오류를 처리하기 위해 중복된 저자 소속 데이터를 제거하였다. 먼저 전체 문자열을 소문자로 전환한 뒤 숫자 및 특수문자 등을 제거하여 문자로만 구성하였다. 그리고 전환된 문자를 그룹화하여 가장 문자열의 길이가 긴(가장 많은 정보를 가진) 데이터를 대표 저자 소속

10) 먼저 딥러닝(deep learning) 기반의 FastText(Joulin et al., 2016) 텍스트 분류기를 이용해 각 저자 소속이 작성된 언어를 분류하였고, 영문이 아닌 경우 구글(Google)의 googletrans API (https://pypi.org/project/googletrans/)를 사용하여 영문으로 전환하였다.

11) 예를 들면 PubMed ID 8358495는 논문은 영어로 작성했지만 저자 소속 정보는 저자의 모국어인 독일어로 작성하였다. 이와 같은 경우는 영어로 작성한 논문의 약 6.11%인 65,600여 건으로 전부 영어로 전환히였다.

으로 지정하여 중복을 제거하였다. 예를 들면 <표 5>와 같이 동일한 저자 소속이라도 국가명을 다양하게 표시하기도 하고 우편번호가 들어가는 경우와 아닌 경우도 있으며, 관례상 소속기관 뒤에 세미콜론(;)을 붙이거나 쉼표가 표시되지 않는 등 다양한 형태가 존재한다. 따라서 동일한 중복제거 문자 'arizonastateuniversitytempeazusa'에 대해서는 대표 저자 소속으로 'Arizona State University, Tempe, AZ 85281, USA.'가 설정된다.

<표 5> 저자 소속 데이터의 중복제거 그룹 문자 생성 규칙

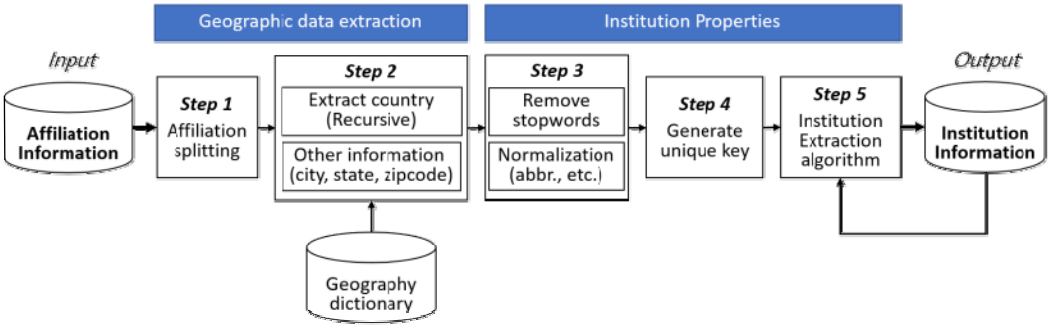
전처리 데이터	중복제거 그룹 문자
Arizona State University, Tempe, AZ 85281, USA.	arizonastateuniversitytempeazusa
Arizona State University; Tempe, AZ USA.	
Arizona State University, Tempe, AZ, U. S. A.	
Arizona State University, Tempe, AZ, USA.	
Arizona State University Tempe AZ USA.	

특히, 대부분의 서지정보는 쉼표를 구분자로 특정 단위(segment)로 세분화할 수 있는데(Torvik, 2005; Huang et al., 2014), 이러한 과정을 통해 원본 서지정보에서 중요한 식별 기준이 되는 쉼표 누락으로 발생할 수 있는 오류를 개선하면서도 가장 많은 정보가 담긴 문자열을 선택하여 지식생산기관 식별 과정에서 더 많은 정보 추출을 할 수 있도록 한다.

2. 고유 식별자를 활용한 알고리즘을 통한 지식생산기관의 식별

본 절에서는 기관명 모호성을 개선하기 위해서 고유 식별자를 활용한 규칙 기반 알고리즘을 제안한다. 고유 식별자는 데이터를 신뢰성 있게 관리하고 검색할 수 있도록 특정한 목적에 맞게 할당된 중복 또는 변경되지 않는 값을 의미한다. 이는 다양한 분야에서 활용되고 있으며, 대표적인 예로 연구자 고유 식별 번호인 ORCID(Open Research and Contributor ID)가 있다. ORCID는 저자명 모호성(author name disambiguation)을 해결하기 위해서 2012년부터 활성화되면서 다수의 연구에 활용되고 있다(Kim and Owen-Smith, 2021). 이와 유사하게 기관명 모호성을 해결하기 위해서 기관 고유 식별자를 적용하였다.

그러나 지식생산기관명은 저자명에 비해 상대적으로 많은 동태적 변화를 거친다. 이에 정규화된 저자 소속에 고유 식별자를 부여하였다. 이 방법을 사용할 경우 동일 기관에 복수의 식별자가 부여될 가능성은 여전히 존재하지만, 기존의 규칙 기반 접근법이 가지고 있는 한계인 기관명 모호성을 상당 부분 개선할 수 있을 것으로 판단된다. 동시에 분석에 필요한 다양한 정보가 포함되어 있는 저자 소속을 '지리정보 추출(geographic data extraction)'과 '지식생산기관 속성(institution properties)' 두 가지 영역으로 나누어 변수화하였다. 전체 과정은 다섯 단계로 진행되며 <그림 3>과 같다.



<그림 3> 저자 소속에서 주요 정보 추출 알고리즘

지리정보 추출 영역에는 1단계와 2단계가 포함된다. 먼저 1단계에서는 2개 이상의 저자 소속이 존재하는 경우 저자 소속을 분리하였다. 이는 모든 공저자의 정보가 합쳐져서 발생하는 정보 중첩이 아닌 대부분이 저자의 겸직을 나타낸다. 겸직도 정보 중첩과 동일한 모호함을 발생시키기 때문에, 분리된 저자 소속이 존재하는 경우 가장 첫 번째 항목을 저자의 저자 소속으로 지정하였다.

2단계는 지식생산기관의 지리정보를 추출하는 과정이다. 이 과정은 다소 복잡한 과정을 거친다. 서로 다른 국가에 이름이 같은 도시가 존재하기 때문에(예: London은 영국과 미국 양국에 존재함) 국가명을 먼저 추출하고 외부 지리정보 데이터를[12] 활용하여 도시, 주(존재하는 경우), 우편번호를 추출한다.

다음으로 3단계부터 5단계까지는 지식생산기관 속성 영역에 해당한다. 3단계는 저자 소속 데이터를 정규화하는 과정이다. 주요 지리정보 이외의 주소 정보(예: Ave., P.O.

12) simplemaps(https://simplemaps.com/data/world-cities) 데이터셋을 기반으로 PubMed에서 사용하는 명칭과 일치하지 않는 경우 등을 보정하였다.

Box. 등)를 제거한 뒤, 기관의 약어를 대표기관명으로 대체하고 표준 국가명을 적용하여 정규화한다. 또한 부서 정보를 분리하여 변수화한다(department, division 등의 키워드 기반으로 추출). 이는 지식생산기관 정보를 추출하는 과정에서 오류를 감소시키기 위한 필수적인 과정이다.

4단계는 고유 식별자를 부여하는 과정이다. 지금까지의 과정으로 정형화된 저자 소속에 고유 식별자를 부여하여 그룹화하였다. 예를 들면 <표 6>에서 정규화된 데이터는 'Arizona State University, Tempe, AZ, USA'가 되고 여기에 고유 식별자를 적용하면 'arizonastateuniversitytempeazusa'가 된다. 이러한 고유 식별자는 동일한 저자 소속을 식별하거나 향후 추가적인 작업에서 반복된 작업을 줄임으로써 일관성과 정확도를 향상시킬 것으로 판단된다.

<표 6> 저자 소속 정보의 고유 식별자 생성 규칙

저자 소속 전처리 데이터	정규화 데이터
	고유 식별자
Arizona State University, Tempe, AZ, U.S.A	Arizona State University, Tempe, AZ, USA
Arizona State University, 660 S Mill Ave, Tempe, AZ 85281, USA	
Arizona State University, Tempe, AZ, United States of America	arizonastateuniversitytempeazusa

5단계에서는 사전에 정의된 '지식생산기관 전환 규칙'에 따라 지식생산기관의 주요 정보를 식별한다. 지식생산기관 전환 규칙에는 규칙 기반 알고리즘을 구현하기 위한 규칙이 정의되어 있다. 여기에는 지식생산기관을 식별하기 위해 국가별로 상이한 기관 식별 정보(예: 미국의 대표 기업형태는 'Corp.'이고 독일은 'GmbH') 등과 기관 유형을 식별하기 위한 정보(예: university, hospital 등) 등을 반영하여 정의된 규칙이 포함된다. 또한 기관의 유형은 해당 기관의 운영 목적에 따라 대학(u), 병원(h), 연구소(r), 민간기업(p) 등 총 네 개로 분류하도록 규칙이 적용되어 있다. 그 외에 의대와 대학병원 여부('medical sch' 등의 키워드 기반으로 추출) 등 추가로 산출할 수 있는 정보도 변수로 생성하였다.

Ⅵ. 지식생산기관 식별 개선 결과

본 장에서는 제안된 데이터 전처리 절차와 알고리즘을 적용하여 생성한 지식생산기관 데이터셋의 기관 식별 정확도 개선 정도를 추정한다. 이를 위해 저자 소속 문자열을 기준으로 한 통계분석 결과 및 기존 PKG datasets과 비교분석 결과를 제시하고 시사점을 도출한다.

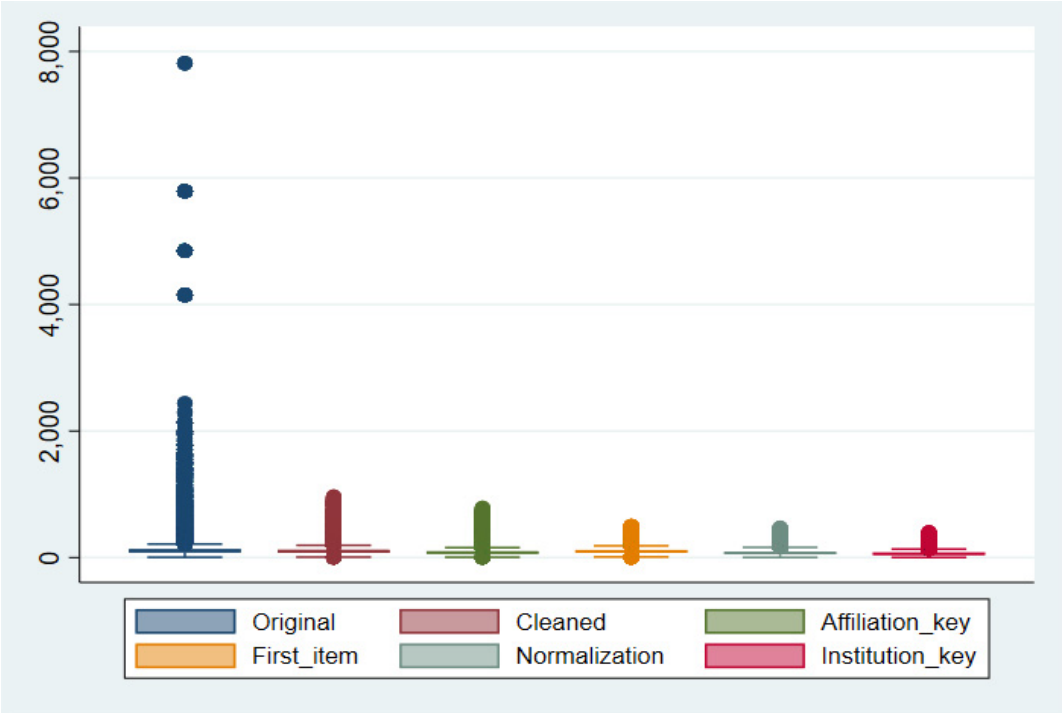
1. 지식생산기관 식별 정확도 개선

서지정보에서 식별된 지식생산기관의 정확도는 식별 기준에 따라 달라지기 때문에 (예: 대학의 연구실을 연구실로 식별할 것인지 상위조직인 대학으로 인식할 것인지 여부 등) 정확도를 평가하기는 쉽지 않지만, 저자 소속의 문자열 길이와 중복도를 통해서 정확도를 추정해 볼 수 있다.

첫째, 데이터 전처리 과정을 통해서 서지정보의 부정확성을 개선할 수 있다. 저자 소속 문자열은 표준 형식은 없으나 어느 정도는 규칙성 있는 형식을 가지고 있어 데이터 정제를 통해서 부정확성을 제거하고 일정한 길이로 수렴시킬 수 있다. 예를 들어 다수의 기관에 겸직하고 있는 저자의 소속을 구분자로 나눠 하나의 문자열로 표시한다면 겸직이 아닌 경우보다 문자열이 길 수는 있으나, 아무리 길어도 문자열의 길이가 1,000자 이상인 경우는 드물다. 이 경우 겸직보다는 정보 중첩 오류일 가능성이 높다.

<그림 4>는 데이터 측면에서 저자 소속 원본 데이터(original)와 원본 데이터에서 지식생산기관을 식별하기 위한 서지정보 전처리(Cleaned), 저자 소속 중복제거(affiliation key), 그리고 알고리즘을 적용하는 3단계(겸직 분리(First item), 정규화(Normalization), 기관 식별자(Institution key)) 등 총 6개 항목에서 저자 소속 문자열의 길이 변화에 대한 통계적 수치를 나타내고 있다.

원본 데이터(1)는 저자 소속 문자열 길이의 최댓값이 7,813자로 바로 다음의 서지정보 전처리 단계(2) 최댓값 996자와 비교했을 때 약 1/8 수준으로 감소하였다. 또한 중간값은 큰 차이가 없지만 표준편차는 거의 2배의 차이가 나 데이터 전처리 과정을 통해서 지식생산기관 식별에 불필요한 정보가 제거되었고 일정한 형태의 저자 소속으로 전환되었다고 판단할 수 있다.



	(1)	(2)	(3)	(4)	(5)	(6)
최대값	7,813	966	787	500	472	402
최소값	4	3	3	2	2	2
평균값	117.56	105.95	86.11	100.96	77.71	64.32
표준편차	99.72	49.12	40.01	38.41	36.39	30.67
중간값	104	97	96	79	71	59

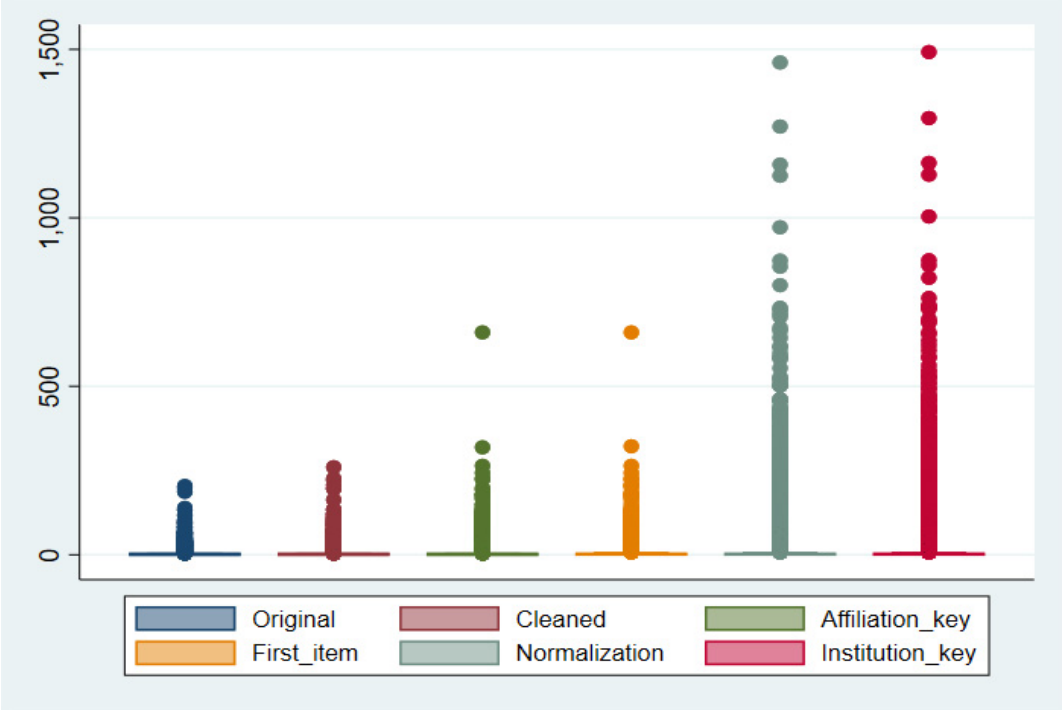
<그림 4> 소속기관 문자열 길이 분석 결과

이후의 단계를 거치면서도 문자열의 최댓값을 포함한 수치들이 큰 격차 없이 전체적으로 통계 수치가 감소하는 형태가 나타난다. 이는 저자 소속 데이터가 지식생산기관 식별 단계를 거치면서 PubMed 상 부정확성이 일정 부분 보정된다는 점을 시사한다고 판단된다.

둘째, 고유 식별자 기반의 알고리즘을 통해서 동일한 저자 소속을 그룹화함으로써 지식생산기관 식별 정확도를 향상시킬 수 있다. 우선 알고리즘을 적용하여 실질적으로 동일하지만 작은 차이로 다르게 식별되었던 문자열들이 하나의 대표 문자열으로 그룹화한다. 이 경우 하나의 고유 식별자에 연결된 저자 소속 데이터가 증가하고, 이에 비례해서 동일한 저자 소속 문자열을 다르게 표시할 가능성은 작아지면서 지식생산기관 식별

정확도가 높아질 수 있다.

<그림 5>에서 보면 전체 문자열은 모든 단계에서 동일하지만 원본 데이터(1) 대비 마지막 단계인 기관 식별자(6)의 최대중복 문자열 수는 7.3배이다. 이는 지식생산기관을 식별하는 단계를 거칠수록 유일한 문자열과 반복된 문자열은 감소하는 반면, 평균 중복과 표준편차는 증가하는 것을 알 수 있다. 특히 정규화(5) 단계에서 급격히 상승하는 것은 동일한 저자 소속 문자열이 다양하게 표현되어 있다는 것을 반영한다. 이는 동일한 저자 소속들이 그룹화된 것으로써 이 단계 데이터를 활용하여 지식생산기관을 식별하면 정확도가 증가할 가능성이 높다.



	(1)	(2)	(3)	(4)	(5)	(6)
전체 문자열	1,017,236	1,017,236	1,017,236	1,017,236	1,017,236	1,017,236
유일한 문자열	558,167	474,863	436,073	418,401	311,186	297,141
반복된 문자열	162,906	181,254	173,836	175,184	139,801	135,021
최대중복	204	260	660	660	1,461	1,492
최소중복	1	1	1	1	1	1
평균중복	1.82	2.14	2.33	2.43	3.27	3.42
표준편차	2.32	2.84	3.72	3.87	11.99	13.07

<그림 5> 저자 소속 문자열 중복도 분석 결과

결론적으로 데이터 전처리 과정과 고유 식별자 기반의 알고리즘을 적용하는 절차를 통해 저자 소속 문자열을 정형화하였고, 그 결과 유사한 저자 소속 정보를 그룹화하면서 일관성과 정확도가 개선되었다고 판단된다.

2. PKG datasets과의 비교

본 절에서는 앞서 구축한 지식생산기관 데이터셋을 PKG datasets과 비교 분석한다. PKG datasets는 PubMed 서지정보에서 학술적 영향, 지식 사용 및 지식 이전을 측정하고 바이오(bio) 개체와의 연결을 기반으로 저자 및 조직을 프로파일링하기 위해 만들어졌다는 점에서 차이가 있다. 그러나 두 데이터셋은 동일한 PubMed 서지정보를 대상으로 규칙 기반 접근법을 사용한다는 점에서 비교가 가능할 것으로 판단된다. 다만 PKG datasets은 MapAffil 2016 데이터셋[13]을 기반으로 하고 있어 MapAffil 2016에서 제공하지 않는 2015년부터 2019년까지는 오픈소스 라이브러리를[14] 활용하여 만들어졌다(Xu et al., 2020). 따라서 두 데이터셋 간의 비교는 2019년까지만 수행하였다.

구체적으로는 PubMed에서 제공하는 PMID를 기준으로 동일한 서지정보 분류 결과와 국가별 비교, 그리고 한국의 주요 현황 비교분석을 통해 시사점을 도출한다.

2.1. 기관 식별에 활용 가능한 데이터 규모

동일한 서지정보를 사용하더라도 접근법에 따라 실제로 식별에 활용되는 데이터 범위와 식별된 기관의 속성이 달라진다(<표 7> 참고). 전체 데이터 수에서는 두 데이터셋 간에 큰 차이가 없으나, 분류된 데이터 수에서 상당한 차이가 발생하며 분류된 국가 수도 15개국의 차이가 존재한다.

이러한 차이가 발생하는 이유는 앞서 정의한 다섯 가지 부정확성 발생 요인에서 찾을 수 있다. 먼저 분류/미분류 데이터 수의 차이는 주로 정보 중첩 및 정보 부족 오류를 처리하는 방식에서 발생하는 것으로 판단된다. PKG datasets은 저자와 저자 소속의 맥락

13) MapAffil 2016 데이터셋은 2016년 10월 첫째 주에 획득한 PubMed의 스냅샷을 기반으로 구성되었으며, PubMed 저자 소속 문자열을 전 세계 도시 및 관련 geo-code로 해석하고 있다.
14) Affiliation Parser: https://github.com/titipata/affiliation_parser

적인 연결을 고려하지 않고 저자 소속 텍스트 자체에 중점을 두어 가장 마지막에 존재하는 정보에서 도시 및 국가정보를 분류하고 있다(Torvik, 2015). 다양한 국가 출신의 저자들의 저자 소속이 하나로 묶여 제1 저자에게만 연결되거나 모든 저자에게 동일하게 연결되는 정보 중첩이 발생하는 경우, 가장 마지막에 있는 정보를 기준으로 식별하면 나머지 저자 소속의 국가는 누락 될 수밖에 없다. 정보 부족 또한 국가를 제대로 식별하지 못하는 주요 원인에 해당한다. 그러나 지식생산기관 데이터셋은 중복을 제거하고 대표 저자 소속을 지정하는 과정에서 일부 오류들이 수정된다. 이러한 점에서 주로 차이가 발생하는 것으로 판단된다.

<표 7> PKG datasets과 비교 현황

구분	PKG datasets	지식생산기관 데이터셋	차이
전체 데이터 수	886,741	884,105	2,623
분류된 데이터	587,950	881,531	293,581
미분류 데이터	298,791	2,574	296,217
분류된 국가 수	181	196	15

같은 이유로 지식생산에 참여하는 지식생산기관의 국가별 순위가 변동되었다(<표 8> 참고). 구체적으로 국가 간 비교에서는 단순히 수적인 차이를 넘어 국가별 순위도 변동된다. 예를 들어 중국(China)의 경우 지식생산기관 데이터셋이 PKG datasets 보다 약 1.7배 더 많이 식별되었고, 이스라엘(Israel)의 경우에는 두 데이터셋 간 2단계의 순위 차이가 발생하였다. PKG datasets의 기반이 되는 MapAffil 2016 데이터셋에서 저자 소속 데이터의 지리정보 연결 정확도는 97.7%가 넘는데, 이러한 차이는 여러 가지 측면에서 시사하는 바가 크다.

결론적으로 아무리 지식생산기관 식별 알고리즘이 정확하더라도 저자 소속 데이터의 부정확성을 고려하지 않고 텍스트 그 자체로만 식별하면 연구 결과가 왜곡될 가능성이 존재한다는 것을 알 수 있다.

<표 8> 빈도수 기준 국가별 순위 비교

순위	PKG datasets		지식생산기관 데이터셋		순위변동	차이
1	USA	175,212	USA	246,478		71,266
2	China	74,795	China	126,636		51,841
3	UK	39,134	UK	56,488		17,354
4	Germany	31,314	Germany	48,045		16,731
5	Japan	25,748	Japan	35,807		10,059
6	Italy	21,435	Italy	33,498		12,063
7	Canada	19,064	Canada	27,218		8,154
8	Korea	16,283	Korea	25,786		9,503
9	France	14,851	France	23,811		8,960
10	Australia	14,721	Australia	22,827		8,106
11	Netherlands	13,305	Spain	21,352	-1	8,436
12	Spain	12,916	Netherlands	19,962	+1	6,657
13	India	11,091	India	16,870		5,779
14	Brazil	10,691	Brazil	15,598		4,907
15	Switzerland	8,787	Switzerland	13,254		4,467
16	Taiwan	7,038	Taiwan	9,891		2,853
17	Sweden	6,337	Sweden	9,196		2,859
18	Belgium	5,827	Belgium	8,162		2,335
19	Iran	5,573	Turkey	7,733	-1	2,168
20	Turkey	5,565	Iran	7,602	+1	2,029
21	Austria	4,614	Austria	6,503		1,889
22	Denmark	4,184	Denmark	5,943		1,759
23	Singapore	4,029	Singapore	5,900		1,871
24	Israel	3,657	Finland	5,552	-1	2,020
25	Finland	3,532	Poland	5,383	-1	1,902
26	Poland	3,481	Israel	5,195	+2	1,538

2.2. 기관 식별 데이터 질 비교: 한국 디지털 헬스 분야 지식생산기관 식별을 중심으로

한국으로 분류된 데이터를 중심으로 지식생산기관 식별 내역을 비교한 결과 지식생산기관의 기여도가 과소 또는 과대 평가되는 것을 발견하였다. 무엇보다 PKG datasets은 한국 기관이 생산한 논문 수를 약 9,500건 정도 적게 측정하고 있다(<표 8>).

또한 개별지식생산기관의 식별 결과도 많이 달라져는 것을 발견하였다. 첫째, 지식생산기관의 다양한 명칭을 하나로 집계하지 못하여 특정 지식생산기관의 기여도가 과소평가되고 있다(<표 9> 참고). 예를 들어 한국과학기술원(KAIST)의 경우 본 연구에서 구축된 '지식생산기관 데이터셋'에서는 5위 'Korea Advanced Institute of Science and Technology'로만 존재하는 반면에, 'PKG datasets'에서는 15위 'Korea Advanced Institute of Science and Technology', 17위 'KAIST', 18위 'Korea Advanced Institute of Science and Technology (KAIST)' 등 총 세 개로 구분되어 기여도가 하나로 집계되지 못하고 분산되는 결과를 초래하고 있다.

<표 9> 한국과학기술원(KAIST) 사례

순위	PKG datasets	건수	순위	지식생산기관 데이터셋	건수
15	Korea Advanced Institute of Science and Technology	259	5	Korea Advanced Institute of Science and Technology	813
17	KAIST	256			
18	Korea Advanced Institute of Science and Technology (KAIST)	212			
합 계		727	합 계		813

둘째, 지식생산기관의 기관 유형을 고려한 결과에서도 기관의 기여도가 과소평가 되고 있다(<표 10> 참고). 예를 들어 서울대(Seoul National University)와 고려대(Korea University)의 경우 같은 기관명에 대해서 기관 유형을 '대학'과 '미지정(UNK)' 두 개로 분류하고 있으며, 연세대의 경우 'Yonsei University'와 'Yonsei University College of Medicine'는 사실상 동일한 기관이지만 다르게 표현함으로써 기여도가 과소평가 되고 있다.

<표 10> 한국의 디지털 헬스 분야 상위 20위 지식생산 유형 비교

PKG datasets	유형	건수	지식생산기관 데이터셋	유형	건수
Unknown	UNK	1,634	Seoul National University	대학	2,259
Samsung Medical Center, Sungkyunkwan University School of Medicine	대학 병원	644	Yonsei University	대학	1,901
Seoul National University	UNK	437	Samsung Medical Center	병원	1,380
Seoul National University	대학	411	Korea University	대학	1,034
Sungkyunkwan University	UNK	315	Korea Advanced Institute of Science and Technology	대학	813
Yonsei University College of Medicine	대학	301	Seoul National University Hospital	병원	718
Seoul National University College of Medicine	대학	290	Asan Medical Center	병원	709
Korea University	대학	285	Sungkyunkwan University	대학	620
Korea University	UNK	282	Kyung Hee University	대학	569
Seoul National University Hospital	대학 병원	261	University of Ulsan	대학	567
Samsung Medical Center, Sungkyunkwan University School of Medicine	UNK	229	Seoul National University Bundang Hospital	병원	503
Asan Medical Center, University of Ulsan College of Medicine	대학 병원	223	Ajou University	대학	452
Yonsei University	UNK	210	Hanyang University	대학	435
Seoul National University Bundang Hospital	UNK	210	Kyungpook National University	대학	416
Yonsei University	대학	208	Chonnam National University	대학	294
Seoul National University Bundang Hospital	대학 병원	175	Pohang University of Science and Technology	대학	289
Kyung Hee University	UNK	171	National Cancer Center	병원	289
Kyungpook National University	UNK	154	The Catholic University of Korea	대학	279
Hanyang University	UNK	150	Chung-Ang University	대학	273
Ajou University School of Medicine	대학	148	Pusan National University	대학	261

셋째, 지식생산기관의 유형별로 살펴보면 민간기업의 기여도가 가장 과소평가 되고 있다(<표 11> 참고). 미분류된 항목을 제외하면 모든 기관 유형에서 약 2배 정도의 차이가 나지만, 민간기업의 경우 약 6배의 차이가 발생하였다. 유형별로 차지하는 비중에서

도 1%에도 미달하는 0.7% 비중을 가진다. 민간기업은 산학협력 등 협력 연구에서 중요한 역할을 하지만 다른 유형 대비 높은 기관 식별 난이도로 인해서 미분류 항목(65.5%) 속에 묻혀서 제대로 표현이 안 된 경우가 많은 것으로 판단된다.

<표 11> 한국의 디지털 헬스 분야 기관 유형 현황

PKG datasets			지식생산기관 데이터셋		
대학	9,998	27.4%	대학	22,178	60.7%
병원	5,870	16.1%	병원	9,798	26.8%
연구소	1,233	3.4%	연구소	3,014	8.3%
민간기업	267	0.7%	민간기업	1,493	4.1%
미분류	23,911	65.5%	미분류	33	0.1%

넷째, 한국과 협력하고 있는 국가별 기여도가 과소 및 과대 평가되는 국가가 존재한다(<표 12> 참고). 먼저 협력 건수는 두 데이터셋 간 대부분 약 2배의 차이가 존재한다. 국가 순위의 경우 미분류된 건을 제외하고 1위에서 4위까지의 국가 순위는 동일하지만 5위부터의 순위는 상당한 차이를 보인다. 예를 들면, 'PKG datasets'에서 5위를 차지하고 있는 일본(Japan)의 경우에는 순위에서 1단계가 과대평가 되었고 '지식생산기관 데이터셋'에서 5위를 차지하고 있는 독일(Germany)은 4단계가 과소평가 되었다.

<표 12> 한국과 협력하고 있는 상위 10개국

순위	PKG datasets		지식생산기관 데이터셋	
1	Korea	12,512	Korea	25,785
2	USA	1,924	USA	3,653
3	UK	379	UK	636
4	China	259	China	618
5	Japan	232	Germany	496
6	Australia	195	Japan	388
7	Iran	187	Canada	337
8	Canada	165	Australia	325
9	Germany	154	Iran	314
10	Spain	125	Italy	299
~	Unknown	23,675	Unknown	33

Ⅵ. 결 론

본 연구는 실증연구에서 과학분야 지식생산기관이 정확히 식별되지 않는 이유와 식별의 정확도를 높이는 방안을 디지털 헬스 분야를 중심으로 분석하였다. 이를 위해 전처리를 통해 서지정보의 부정확성을 줄이고 고유 식별자를 활용한 알고리즘을 개발하여 지식생산기관을 식별하였다.

먼저 PubMed 서지정보의 부정확성을 유발하는 오류 유형과 지식생산기관 식별의 한계점을 각각 다섯 가지로 정의하였고, 이를 해결하기 위해 데이터 측면과 알고리즘 측면에서 접근 방법을 제시하였다. 이를 바탕으로 PubMed 데이터베이스에서 디지털 헬스 분야의 서지정보를 구축하고 PKG datasets과 비교함으로써 저자 소속 데이터의 일관성과 정확도가 개선되는 것을 확인하였다. 다만 비교분석에서 나타난 수치상의 차이 외에 추출된 소속기관의 정확도는 어떠한 기준을 적용했는지에(예: 상위 기관을 고려했는지의 여부 등) 따라서 달라지기 때문에 우열을 가리기에는 한계가 있다. 그러나 동일한 데이터와 동일한 규칙 기반 접근법을 연구 방법으로 사용했다는 점에서 본 연구의 분석 결과는 기관의 역할이 과소 또는 과대 평가될 위험을 줄이는 방법을 제시하였다.

본 연구가 제시한 방법-논문에서 저자 소속 데이터를 중심으로 고안하여 제시한 데이터 전처리 방법과 고유 식별자 기반의 알고리즘은 디지털 헬스 분야뿐 아니라 다른 분야에도 동일하게 활용할 수 있다. 또한 본 연구 결과로 생성된 데이터베이스는 범용성을 가져 향후 유사한 연구에도 활용할 수 있을 것으로 판단된다. 구체적으로는 동일한 데이터 전처리 과정을 통해 분석하고자 하는 저자 소속의 고유 식별자를 생성하고, 이를 데이터베이스에서 검색하여 동일한 값이 존재하면 별도의 처리 과정 없이 지식생산기관을 바로 연결할 수 있다. 또한 동일한 값이 존재하지 않는 경우에도 제안된 알고리즘을 적용하여 지식생산기관을 식별하고 데이터베이스에 추가하는 방식으로 지속적인 확장이 가능하다. 이를 통해 과학기술 분야 등 여러 분야에서 그동안 논문 서지정보의 부정확성으로 인해 수행하기 어려웠던 많은 연구가 가능해져 연구 저변을 넓히는 데 기여할 수 있을 것으로 판단된다.

그러나 비정형 데이터의 특성과 정보 부족 등으로 인해 모든 데이터를 알고리즘으로 정제하고 식별하는 데는 한계가 있어 여전히 적지 않은 수작업이 요구된다. 따라서 본 연구 결과로 구축된 정확도 높은 지식생산기관 데이터베이스를 기반으로 하여, 저자를

구분하는 정보가 존재하지 않더라도 '저자-저자 소속' 쌍(pair)을 식별하는 방법 등의 보다 자동화된 해결 방법이 필요하다. 또한 대형 언어 모델(Large Language Model, LLM) 등 고도화된 기법을 적용하여 지식생산기관을 식별하는 알고리즘을 구현하는 등의 연구를 후속 연구로 진행할 계획이다.

참고문헌

(1) 국내문헌

문성욱, 2012, “공공-민간 협력구조와 과학기술연구의 생산성: 인간 배아줄기세포 연구를 중심으로”, 한국개발연구원

(2) 국외문헌

Aghion, Philippe, Mathias Dewatripont, and Jeremy C. Stein. “Academic Freedom, Private sector Focus, and the Process of Innovation.” The RAND Journal of Economics 39, no. 3 (September 2008): 617 – 35.

Bikard, Michaël, Keyvan Vakili, and Florenta Teodoridis. “When Collaboration Bridges Institutions: The Impact of University – Industry Collaboration on Academic Productivity.” Organization Science 30, no. 2 (March 2019): 426 – 45.

BRUIN, R E de, and H F Moed. “The Unification of Address in Scientific Publications,” n.d.

Bush, Vanevar. “Science: The Endless Frontier.” National Science Foundation – EUA. Washington, 1945.

Einav, Liran, and Jonathan Levin. “The Data Revolution and Economic Analysis.” Innovation Policy and the Economy 14, no. 1 (2014): 1 – 24.

Hall, Bronwyn H., and Dietmar Harhoff. “Recent Research on the Economics of Patents.” Annual Review of Economics 4, no. 1 (September 1, 2012): 541 – 65.

Hicks, Diana, and Sylvan Katz. “Hospitals: The Hidden Research System,” 1996.

Huang, Shuiqing, Bo Yang, Sulan Yan, and Ronald Rousseau. “Institution Name Disambiguation for Research Assessment.” Scientometrics 99, no. 3 (June 2014): 823 – 38.

Jones, Benjamin F. “The Burden of Knowledge and the ‘Death of the Renaissance Man’: Is Innovation Getting Harder?” The Review of Economic Studies 76, no. 1 (2009): 283 – 317.

Jones, Benjamin F., Stefan Wuchty, and Brian Uzzi. “Multi-University Research Teams: Shifting Impact, Geography, and Stratification in Science.” Science 322, no. 5905 (November 21, 2008): 1259 – 62

Jonnalagadda, Siddhartha, and Philip Topham. “NEMO: Extraction and Normalization of Organization Names from PubMed Affiliation Strings.” Journal of Biomedical Discovery and Collaboration 5 (2010): 50.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. “Fasttext. Zip: Compressing Text Classification Models.” arXiv Preprint arXiv:1612.03651, 2016.

Kilkenny, Monique F., and Kerin M. Robinson. “Data Quality:‘Garbage in – Garbage Out.’” Health Information Management Journal. SAGE Publications Sage UK: London, England, 2018.

Kim, Jinseok, and Jason Owen-Smith. “ORCID-Linked Labeled Data for Evaluating Author Name Disambiguation at Scale.” Scientometrics 126, no. 3 (March 2021): 2057 – 83.

Louis, Karen Seashore, David Blumenthal, Michael E. Gluck, and Michael A. Stoto. “Entrepreneurs in Academe: An Exploration of Behaviors among Life Scientists.” Administrative Science Quarterly, 1989, 110 – 31.

Moon, Seongwuk. “How Does the Management of Research Impact the Disclosure of Knowledge? Evidence from Scientific Publications and Patenting Behavior.” Economics of Innovation and New Technology 20, no. 1 (2011): 1 – 32.

“NLM Technical Bulletin.” U.S. National Library of Medicine, December 1999.

“NLM Technical Bulletin.” U.S. National Library of Medicine, December 2013.

“NLM Technical Bulletin.” U.S. National Library of Medicine, December 2014.

Partha, Dasgupta, and Paul A. David. “Toward a New Economics of Science.” Research Policy 23, no. 5 (1994): 487 – 521.

Perkmann, Markus, Valentina Tartari, Maureen McKelvey, Erkko Autio, Anders Broström, Pablo D’este, Riccardo Fini, Aldo Geuna, Rosa Grimaldi, and Alan Hughes. “Academic Engagement and Commercialisation: A Review of the Literature on University – Industry Relations.” Research Policy 42, no. 2 (2013): 423 – 42.

Pollitt, Christopher, Colin Talbot, Janice Caulfield, and Amanda Smullen. Agencies: How Governments Do Things through Semi-Autonomous Organizations. Springer, 2004.

“PubMed Help.” Bethesda (MD): National Center for Biotechnology Information (US), 2020.

Sampat, Bhaven, and Heidi L. Williams. “How Do Patents Affect Follow-On Innovation? Evidence from the Human Genome.” American Economic Review 109, no. 1 (January 1, 2019): 203 – 36.

Stephan, Paula. How Economics Shapes Science: Harvard University Press, 2012

Thune, Taran, and Andrea Mina. “Hospitals as Innovators in the Health-Care System: A Literature Review and Research Agenda.” Research Policy 45, no. 8 (2016): 1545 – 57.

Torvik, Vetle I. “MapAffil: A Bibliographic Tool for Mapping Author Affiliation Strings to Cities and Their Geocodes Worldwide.” D-Lib Magazine 21, no. 11/12 (November 2015).

Varian, Hal R. “Big Data: New Tricks for Econometrics.” The Journal of Economic Perspectives 28, no. 2 (2014): 3–27.

Williams, Heidi L. “Intellectual Property Rights and Innovation: Evidence from the Human Genome.” Journal of Political Economy 121, no. 1 (February 2013): 1–27.

Wuchty, Stefan, Benjamin F. Jones, and Brian Uzzi. “The Increasing Dominance of Teams in Production of Knowledge.” Science 316, no. 5827 (May 18, 2007): 1036–39.

Xu, Jian, Sunkyu Kim, Min Song, Minbyul Jeong, Donghyeon Kim, Jaewoo Kang, Justin F. Rousseau, et al. “Building a PubMed Knowledge Graph.” Scientific Data 7, no. 1 (June 26, 2020): 205.

Yu, Wei, Ajay Yesupriya, Anja Wulf, Junfeng Qu, Marta Gwinn, and Muin J Khoury. “An Automatic Method to Generate Domain-Specific Investigator Networks Using PubMed Abstracts.” BMC Medical Informatics and Decision Making 7, no. 1 (December 2007): 17.

□ 투고일: 2024.03.07. / 수정일: 2024.04.01. / 게재확정일: 2024.04.26.