https://doi.org/10.5392/JKCA.2019.19.11.262

학술논문과 참고문헌의 자동매핑 사례 분석

Case study of Journal Article and Reference Mapping

김재훈, 김순영, 임석종, 황혜경
한국과학기술정보연구원

**Jayhoon Kim(jay.kim@kisti.re.kr), Soon Young Kim(maya@kisti.re.kr),
Seok Jong Lim(seoklim@kisti.re.kr), Hyekyong Hwang(hkhwang@kisti.re.kr)**

요약

학술논문의 말미에 기재하는 참고문헌은 저자가 연구윤리를 준수하고 독자들이 관련 선행연구를 참고할 수 있도록 돕는 정보이자 논문간의 인용과 피인용 관계를 연결시키는 데 유용한 정보이다. 계량서지학이 발전하면서 참고문헌 데이터는 국가, 기관, 개인의 학술 영향력을 평가하는 중요한 데이터로 활용되고 있다. 하지만 참고문헌 형식의 다양성, 학술지명과 저자명 축약 기재로 인한 정보 손실, 저자들의 오타 등으로 인해 참고문헌을 식별하여 연결하는 것은 쉽지 않다. 본 연구에서는 학술논문 참고문헌 데이터를 구축하고 매핑하는 과정에서 발생한 오류 사례를 분석함으로써 참고문헌 데이터 매핑율 제고 방안을 고찰하였다. 연구결과 참고문헌 식별 실패의 주요 원인은 유사 학술지명 식별 문제로 밝혀졌으며 식별과 매핑율 향상을 위한 방안으로 학술지명 전거파일 활용, 논문 DOI 등록율 제고를 제시하였다. 본 연구는 연구 대상 데이터에서 차별성이 있다. 국내에서 주로 구독, 이용, 출판, 인용되는 국내 및 해외 학술지 통합 데이터베이스를 대상으로 참고문헌 매핑을 시도하였다. 참고문헌 구축량 및 매핑율 향상을 통해 해외 인용색인 데이터베이스와는 차별화된 국내 상황을 반영한 인용 분석 및 서비스 기반 데이터베이스로 활용이 가능하다.

■ 중심어 : | 참고문헌 매핑 | 인용 분석 | 인용 지수 | 인용색인 데이터베이스 | 학술논문 |

Abstract

References at the end of an academic paper are information that helps authors keep their research ethics, readers refer to related prior studies. Also references are useful information for linking citations and citations between articles. As bibliography metrics develops, bibliographic data is used as an important data for assessing the academic influence of countries, institutions and individual researchers. However, it is not easy to identify and link the reference data due to the diversity of the bibliographic citation formats, the loss of information due to the abbreviation of journal names and author names, and typos by authors. This study investigated the method of improving the bibliographic data mapping rate by analyzing the unmapped cases. As a result, it was found that the main cause of the article-reference mapping failure was the similarity of abbreviated journal names. Research team suggested that continuous management of journal title authority data and improving the DOI registration rate as ways to improve the identification and mapping rate. This study is differentiated from other studies in used database. Bibliography mapping was attempted for domestic and foreign integrated journal database that is mainly subscribed, used, published and cited in Korea. Through reference construction volume and mapping rate improvement, it can be used as citation analysis and service database reflecting domestic situation that is different from overseas citation index database.

■ keyword : | Reference Mapping | Citation Analysis | Citation Index | Citation Database | Research Paper |

* 본 연구는 2019년도 한국과학기술정보연구원(KISTI)의 주요사업 과제로 수행한 것입니다.

접수일자 : 2019년 11월 04일
심사완료일 : 2019년 11월 15일
수정일자 : 2019년 11월 15일
교신저자 : 김순영, e-mail : maya@kisti.re.kr

I. 서 론

참고문헌은 저자가 연구윤리를 준수하고 독자들이 관련 선행연구를 참고할 수 있도록 돕는 정보이자 논문 간의 인용과 피인용 관계를 연결시키는 데 유용한 정보이다. 계량서지학이 발전하면서 참고문헌 데이터는 국가, 기관, 개인의 학술 영향력을 평가하는 중요한 데이터로 활용되고 있다. 캠브리지 사전[1]에서는 인용(Citation)을 저작물로부터 취해진 단어 또는 문자로 정의하고 있다. Wikipedia[2]에서는 인용(Citation)의 중요한 목적을 저자의 주장에 대한 근거 제시, 지적 도덕성 준수(표절 방지), 아이디어의 출처 명시 , 독자 스스로 판단할 수 있도록 근거 제시 등으로 제시하고 있다. Laborie와 Halperin[3]은 인용 분석을 색인, 초록 또는 학술지 등의 원 출판물을 인용하여 이용되거나 생산된 특정 주제 분야 문헌의 특성을 밝히는 것으로 정의하였다. 서울대학교 도서관 LibGuide[4]에는 인용(citation)을 하나의 저작물을 원저자를 밝히고 널리 알려진 형식을 사용하여 다른 저작물에 이용하는 행위로 정의되어 있다.

인용의 개념이나 인용과 피인용의 관계는 명확하나 인용율, 인용 분석 등 실제 활용에 있어 다소 까다롭게 여겨진다. Egghe와 Rousseau[5]의 인용 분석 초기 연구에서는 인용과 피인용의 개념을 명확히 하기 위해 다이어그램을 제시하고 설명하였다. “Reference”를 인용하는 것인 역방향 개념으로 “Citation”을 인용되는 것인 순방향 개념으로 설명하고 있다. Ashikuzzaman[6]은 문헌정보학 학술 블로그 LIS BD Network의 인용분석 관련 기사에서 인용(citation)은 각주(footnote)나 참고문헌(reference)과 같이 두 문헌간의 연결 관계를 표현하는 서지정보라고 정의하며 인용(citation)이 참고문헌(Reference)을 포함하는 개념으로 설명하고 있다.

초기의 인용 분석 사례로는 1927년 Gross와 Gross의 도서관 장서 선정을 위한 화학 분야 학술지 참고문헌 분석 사례가 있다[7]. 다양한 인용 분석 방법 중 Eugene Garfield가 1955년 고안한 피인용 지수(impact factor)는 학술지를 평가하는 대표적인 기준이 되었다.

참고문헌은 연구주제 발굴, 연구 근거 제시와 방향을 잡아가는데 필요하다. 더불어 연구 성과를 계량 분석하는데 참고문헌을 활용한 학술영향력 지수를 산출하는데 활용되고 평가에 중요해지고 있다. 국가과학기술정보센터 NDSL(http://www.ndsl.kr)에서 "인용분석"이라는 키워드로 논문을 검색한 결과를 보면 참고문헌의 용도는 학술지 평가도구, 연구자 업적 평가, 국가별, 주제 분야별 인용도 비교, 연구성과 분석, 지적 구조 분석, 장서개발 및 폐기, 학술지 이용행태 분석 등으로 나타난다.

학술 영향력을 산출하는데에는 학술지 영향력을 산출하는 Impact factor, Cited half-life(피인용 반감기), Median Impact facto(중간인용지수), Eigen factor, SNIP(Source Normalized Impact per Paper) 등과 연구자 영향력을 산출하는 h-index, g-index 등이 있다. 학술 영향력은 인용건수를 기준으로 측정되기 때문에 측정하는 서지 데이터베이스의 품질이 중요하다. 참고문헌정보가 누락 없이 구축되고 참고문헌의 학술지, 논문이 정확히 식별되어야 한다. 연구자 영향력까지 산출하기 위해서는 서지 데이터베이스에 구축된 각 논문별 저자까지 식별되어야 한다. 그러나 전 세계 학술지들의 참고문헌 형식의 다양성[8], 학술지명과 저자명 축약 기재로 인한 정보 손실, 저자들의 오타 등으로 인해 참고문헌을 식별하여 인용 문헌과 피인용 문헌을 연결하는 것은 쉽지 않다.

Web of Science, SCOPUS 등 대표적인 인용색인 DB가 구축되어 학술논문 검색, 인용분석, 연구실적 평가 등에 활용되고 있으나 영문 학술지 중심의 선정 논란이 꾸준히 제기되고 있으며 여러 산학연 기관들이 고가의 사용료를 지불하고 있다. 국내외를 아우르는 학술지 논문 데이터베이스를 구축하는 국내의 공공기관은 한국과학기술정보연구원이 거의 유일하다. 근래 구축 개시한 해외 참고문헌 데이터를 활용하면 전 세계에 출판되는 국내 연구자들의 연구성과를 포괄적으로 분석하는 등 한국 상황에 최적화된 인용 분석과 서비스가 가능해 질 것으로 기대된다.

본 연구에서는 한국과학기술정보연구원의 국가가용 학술논문DB (이하 e-Gate DB)를 대상으로 참고문헌 데이터 구축 과정을 살펴보고 매핑 오류 사례를 분석함

으로써 참고문헌 데이터 구축시 과제와 품질 향상 방안을 모색해 보았다.

II. 이론적 배경

Eugene Garfield[9]는 Science Citation Index(SCI) 인용 분석 연구에서 학술지명 데이터 표기 다양성의 문제로 인한 인용 분석의 어려움을 밝혔다. 첫째 학술지 약어명 기재에 다양성 문제를 지적하였다. 실험 데이터로 사용한 참고문헌 분석결과 수록 학술지는 12,000여종이었는데 이들 학술지의 약어명이 10여만 개나 되었음을 지적하였다. 둘째 학술지 합병 분리로 변경된 학술지명으로 인용 분석을 수행하기 위한 데이터 처리의 어려움을 지적하였다. 셋째 저자들의 학술지명 기재 오타 문제를 지적하였다. 이러한 문제는 선행연구가 수행된 1971년 이후 현재까지도 완벽한 해결책이 도출되지 못하고 있다.

한정민 등[10]은 논문과 참고문헌 매핑 효율을 향상시키기 위하여 한의학 분야 학술논문 자동매핑 방법을 연구하였다. 연구결과로 참고문헌 인용지수 검색 시스템을 구축하기 위해서는 논문의 선후관계를 파악할 수 있도록 정확한 매핑이 필요함을 주장하였다. 이를 위해 한의학 분야 학술지를 대상으로 참고문헌 데이터 형식을 통일하고 한의학 사전을 활용하여 데이터 구축시 오타를 자동 수정할 수 있는 시스템을 개발함으로써 매핑 성능 수준을 측정한 결과 성능을 약 3.6% 개선하였다. 여기서 사용된 “자동 매핑”의 개념은 작업자가 참고문헌 데이터를 입력하여 구축하는 과정에서 작업자에게 이미 구축된 전통의학정보포털(OASIS) 논문 정보를 추천해주고 선택하게 함으로써 기 구축된 논문과 매핑 하는 방식으로 보인다. 본 연구에서는 사용한 “자동 매핑” 방식은 이미 구축된 논문과 참고문헌 데이터를 매핑 대상 필드별 문자열 일치 여부로 매핑 하는 방식으로 수작업을 수반하지 않는 자동 매핑 방식이라는 점에서 차이가 있다.

최근에는 국제 표준 디지털 객체 식별자(Digital Object Identifier, DOI)를 활용하여 논문과 참고문헌과의 매핑이 용이해졌다. 하지만 학술지에 따라 참고문헌에 DOI 기재를 의무화하지 않는 경우가 있고 DOI 오류도 있어 완벽한 매핑을 보장할 수는 없다. Xu 등[11]은 Web of Science의 참고문헌 데이터 중 DOI 오류 사례를 분석하였으며 대부분 (약 92%)의 오류는 DOI 접두사(prefix) 오류이나 전체적으로는 매우 다양한 유형의 오류가 있어 DOI 오류 데이터 정제(cleansing)가 용이하지 않음을 지적하였다. 따라서 DOI이외의 매핑 기준을 유지하는 것이 필요하다.

구희관 등[12]은 이러한 논문과 참고문헌간의 매핑율을 높이기 위하여 학술지명, 저자, 권호, 발행년도 등 7개 인용 필드의 데이터를 정규화하고 이를 검색엔진을 활용하여 TF/IDF 기반으로 인용 매칭 방법을 연구하였다. 게재페이지, 게재년도, 권/호, 저자명 중 하나와 논문제목을 결합한 경우와 저자명과 게재년도를 결합한 경우의 총 다섯 가지의 필드결합이 0.8 이상의 인용 매칭 성능을 보인 것으로 나타났다.

또한 주요 인용색인 데이터베이스의 학술지 선정 기준 편향성에 대한 지속적인 논란이 있다. 예로 Mongeon과 Paul-Hus[13]는 주제 분야 및 영어권 학술지 중심의 수록 편향성 문제 지적하였고 Rafols[14] 등은 남미 국가들의 학술지의 주요 인용색인 데이터베이스 과소 등재 문제를 지적하였다. 반면 Wagner와 Wong[15]은 BRICs 국가에서 출판되는 학술지의 Web of Science, SCOPUS 등재율은 합당하고 언어 장벽, 학술지 형식, 디지털화 부족 등의 결과라고 주장하였다. 이러한 등재 편향성의 문제로 인하여 국가 특성이 반영된 인용색인 데이터베이스가 반드시 필요하다고 본다.

김홍렬과 정경희[16]는 국내 참고문헌 데이터베이스 비교 연구에서 구축 대상 학술지 망라성 부족, 국내 참고문헌만을 대상으로 DB를 구축함으로 인한 인용빈도 및 지수 신뢰성 저하, 학술지 선정 기준의 모호성, 학술지명 전거 통제 부재 문제 등을 지적하였다. 본 연구의 분석 대상 데이터베이스는 국내외 학술지를 아우르는 망라성, 국내에서 주로 구독, 이용, 출판, 인용되는 국내외 학술지 중심의 선정 기준, 국내외 학술지명 전거 관리로 선행연구에서 지적한 문제점을 상당 부분 해소한 점에서 국내의 타 인용색인 데이터베이스와의 차별성이 있다.

III. 연구 방법

1. 연구 절차 및 데이터 선정

본 연구에서는 통상적인 참고문헌 데이터 구축 절차에 따라 [그림 1]과 같이 참고문헌 데이터를 구축하고 논문(피인용 문헌)과 참고문헌(인용문헌)을 매핑하였고 미매핑 원인을 분석하였다.

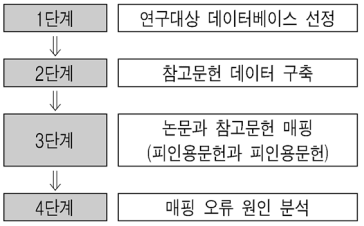


그림 1. 본 연구의 수행 절차

e-Gate DB에는 주요 색인DB(Web of Science, SCOPUS, 한국연구재단 KCI)에 등재된 전 세계 피어리뷰(peer review) 학술지 논문 정보가 8천만 건 이상 수록되어 있다. 또한 전 세계 40여 학술 출판사로부터 직접 생산되는 논문 메타데이터를 입수하고 있어 데이터 품질이 높다. 2017년부터는 참고문헌 데이터를 구축하고 있으며 2019년 7월 기준으로 7억4천만건의 참고문헌 데이터가 구축된 상태이다. 국내외 주요 색인DB에 등재된 양질 학술지를 수록하고 있으며 논문 정보의 양도 방대하고 참고문헌 구축시 학술지명 전거, 인용 논문과의 매핑 등 가공 수준이 높아 본 연구의 대상으로 선정하였다.

2. 참고문헌 데이터 구축 절차

e-Gate DB 구축에 사용되는 참고문헌 원형 데이터는 Crossref 및 해외 학술출판사로부터 입수하고 있다. 대부분 구조화된 데이터로 되어 있어 품질이 높다.

```
<bib xml:id="bib10">
<label>[10]</label>
  <citation type="journal" xml:id="cit10">
    <author><givenNames>D.A.</givenNames><familyName>Medvetz</familyName>
    </author>
    <author><givenNames>K.M.</givenNames><familyName>Hindi</familyName>
    </author>,
    <author><givenNames>M. J.</givenNames><familyName>Panzner</familyName>
    </author>,
    <author><givenNames>A. J.</givenNames><familyName>Ditto</familyName>
    </author>,
    <author><givenNames>Y. H.</givenNames><familyName>Yun</familyName>
    </author>,
    <author><givenNames>W. J.</givenNames><familyName>Youngs</familyName>
    </author>,
    <journalTitle>Metal&#8208;Based Drugs</journalTitle>
    <pubYear year="2008">2008</pubYear>, article ID 384010,
    <pageFirst>7</pageFirst> pages,
    <accessionId ref="info:doi/10.1155/2008/384010">http://dx.doi.org/10.1155/2008/384010
    </accessionId>.
  </citation>
</bib>
```

그림 2. 구조화된 참고문헌 원형 데이터 예시

입수된 참고문헌 소스 데이터는 로더에 의해 파싱된 후 데이터베이스에 적재된다. 기존 참고문헌 소스 데이터와 구조가 다른 경우 파싱 오류가 발생하는데 이 경우 데이터 구조를 분석하고 로더를 수정한다. 파싱된 데이터는 항목별로 유효성 검사를 수행한다. 유효성에 오류가 있는 경우 소스 데이터를 재입수하여 구축한다. 마지막 단계로 참고문헌 데이터는 e-Gate DB에 구축되어 있는 논문과 매핑된다.

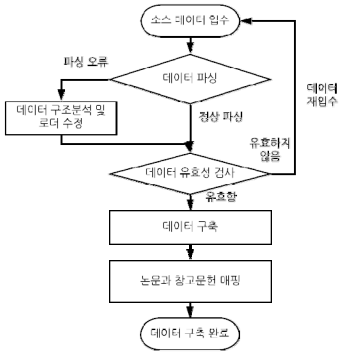


그림 3. 본 연구에서의 참고문헌 데이터 구축 절차

3. 논문과 참고문헌 매핑 방법

참고문헌 데이터 구축시마다 참고문헌 레코드와 e-Gate DB 논문 데이터를 비교하여 일치하는 레코드가 있는지 조사한다. 비교 조건으로 참고문헌 레코드에 DOI가 있으면 충분하고, DOI가 없는 경우 참고문헌 레코드의 6개 항목과 기 구축되어 있는 논문 레코드의 해당 항목을 비교한다. 매핑 기준은 [표 1]과 같다.

논문과 참고문헌 레코드를 비교하여 일치하면 매핑에 성공한 것으로 판단하고 참고문헌 레코드의 매핑 필드에 e-Gate DB의 논문 레코드 식별번호를 기록하여 참고문헌과 논문을 연결한다. 본 연구에서는 2018년 출판된 해외 학술논문 1,000건을 임의로 선정하여 참고문헌 자동 매핑을 수행하였다.

표 1. 참고문헌과 논문의 매핑 기준

구분	매핑 기준
DOI가 있는 경우	DOI 일치하면 매핑
DOI가 없는 경우	① 학술지명(또는 ISSN) 일치 ② 논문명 일치 ③ 논문 시작페이지 일치 ④ 출판년도 일치 ⑤ 권 일치 ⑥ 1저자의 성 일치 ①~⑥ 조건을 모두 만족해야 매핑

IV. 연구 결과

1. 참고문헌 데이터 구조화 수준

KISTI에 입수된 참고문헌 총 데이터는 741,020,621건이고 이 중 689,495,179건이 구조화된 데이터이다. 참고문헌 중 학술지 논문의 경우 구조화 비율 99.92%에 달한다(2019년 7월 기준). 소스 데이터를 구조화된 형태로 확보하는 것은 데이터 품질 우수성과 구축비용 절감 측면에서 큰 이득이 된다. 소스 데이터가 비구조화된 상태로 입수된다면 수작업 또는 CERMINE과 같은 논문 메타데이터 파싱(parsing) 도구를 사용하여 구조화해야 하는데 기계학습 기술까지 적용한 CERMINE의 경우에도 비구조화 참고문헌의 논문명 파싱 정확율 90%, 저자명 파싱 정확율 80% 수준이다[17].

표 2. 참고문헌 전체 소스 데이터 구조화 수준

참고문헌 유형	구조화 데이터		비구조화 데이터		총 건수
	건수	비율(%)	건수	비율(%)	
학술지 논문	529,474,029	99.92	425,068	0.08	529,899,097
단행본	71,763,343	99.77	163,440	0.23	71,926,783
보고서	224,585	99.80	444	0.20	225,029
학술대회 논문	687	99.42	4	0.58	691
기타	88,032,535	63.35	50,936,486	36.65	138,969,021
합계	689,495,179	93.04	51,525,442	6.95	741,020,621

2. 참고문헌 매핑 결과 분석

연구방법에서 기술한 바와 같이 2018년 출판 논문 1,000건에 대한 참고문헌 구축 결과, 수록된 참고문헌 147,854건 중 유형이 학술논문인 참고문헌은 124,407건이었다. 이 가운데 e-Gate DB에 기 구축된 논문 정보와 자동 매핑을 시도한 결과 성공 건수는 104,902건(성공율 84.3%), 실패 건수는 19,505건이었다. 참고문헌 매핑 결과는 [표 3]과 같다.

표 3. 참고문헌 자동 매핑 실험 결과

구분	건수	비고
샘플 논문 건수	1,000	
참고문헌 건수	147,854	논문 당 평균 참고문헌 148건
참고문헌 중 학술지 건수	124,407	참고문헌 중 학술지 논문 비율 84.1%
지동매핑 성공 건수	104,902	참고문헌 자동식별율: 84.3%

본 연구에서는 SCOPUS DB와의 참고문헌 매핑 수준을 비교하고자 내부 분석용으로 도입한 SCOPUS 데이터로 비교 분석하였다. SCOPUS에서도 논문과 참고문헌이 매핑된 경우 참고문헌 테이블에 논문 식별자(eID)가 기록되는데 이 건수를 측정한 결과 802,303,502건으로 전체 참고문헌 1,469,246,077의 54.6%인 것으로 확인되었다.

e-Gate DB의 자동 매핑 실패 19,505건 중 일부 1,394건을 샘플링 하여 실패 원인을 분석하였다. 샘플링 기준은 참고문헌 게재지명이 "A"로 시작하는 전량으로 하였다. 매핑 실패 건과 일치하는 e-Gate DB내의 논문 데이터를 수작업으로 찾아 건별로 데이터 내용을 비교하였다. 비교 결과와 매핑 실패 원인을 기술하였으며 분석 완료 후 원인별로 그룹핑하였다.

표 4. 참고문헌 자동 매핑 실패 원인

구분	건수	비율(%)
학술지명 유사	1,200	86.1
학술지 아님	77	5.5
게재지 확인 불가(소스 데이터 오류)	48	3.4
주요 색인 미등재(e-Gate 구축 비대상)	41	2.9
부실 학술지(Predatory Journal) 구축 비대상	8	0.6
참고문헌정보 오타(소스 데이터 오류)	7	0.5
기타 사유	13	1.0
합계	**1,394**	**100**

[표 4]와 같이 자동 매핑 실패 원인을 다섯 가지로 구분하였다. 첫째는 학술지명이 유사한 경우로 대부분의 경우가 이에 해당한다(전체의 86.1%에 해당). 예를 들어 참고문헌에 학술지 약어명 "Ann Geophys"로 기재된 경우 "Annales geophysicae"과 "Annals of geophysics" 두 개의 학술지에 해당될 수 있다. 둘째로는 참고문헌 레코드가 학술지가 아님에도 소스 데이터에는 학술지로 잘못 표기된 경우이다. 이는 출판사에서 참고문헌 소스 레코드 제작시 발생하는 오류로 본 실험 데이터의 5.5%를 차지하였다. 셋째로는 참고문헌 매핑 대상 논문이 e-Gate DB 구축 비대상이기 때문으로 확인되었다. e-Gate DB는 Web of Science, SCOPUS, KCI 등재 학술지 중심으로 논문 레코드를 구축하고 있다. 구축되지 않은 논문이 참고문헌으로 입수된 경우 매핑이 불가하다. 넷째로는 부실 학술지로 인한 매핑 불가로 나타났다. 빈도는 실험 데이터의 0.6%로 참고문헌 1,394건 중 8건이 발견되었다. e-Gate DB는 부실 학술지 정보를 관리하고 있으며 해당 학술지의 논문 정보 구축을 배제하고 있다. 다섯째는 참고문헌 정보 오타로 인한 것으로 나타났다(0.5%).

3. 참고문헌 매핑 향상 방안

가. 학술지명 전거데이터 활용

분석결과 유사 학술지명을 구분하지 못하는 문제로 인한 매핑 실패 건수가 1,200건(86.1%)로 대부분의 경우에 해당하였다. 유사 학술지명 문제를 해결하는데 학술지명 전거 데이터를 활용하는 방안이 있다. e-Gate DB에는 전 세계 학술지 120,894종에 대한 이형(異形)명 327,302건이 구축되어 있다. 참고문헌 매핑율을 높이기 위해서는 매핑 기준을 다소 완화하여 논문명, 논문 시작페이지, 저자명 등의 다양한 요소를 복합적으로 비교함으로써 매핑 성능을 상당히 향상시킬 수 있을 것이다. 구희관 등(2008)[12]의 연구결과에 따르면 미매핑건의 80% 정도는 매핑이 가능할 것으로 추정된다.

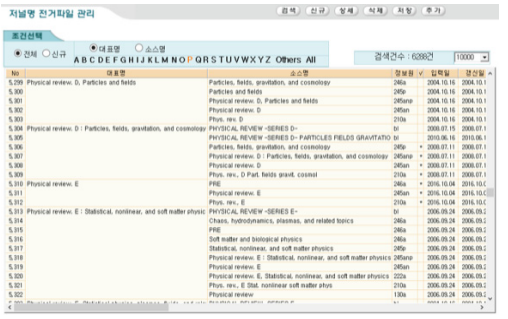


그림 4. 학술지명 전거 데이터 예

나. 논문 DOI 정보 구축

논문과 참고문헌에 각각 DOI가 부여되어 있는 경우 매핑 충분조건이 되므로 DOI 등록율이 높아지면 매핑율이 증가될 수 있다. 현재 e-Gate DB에 구축된 참고문헌 논문의 DOI 등록율은 45% 수준이다. 하지만 DOI는 콘텐츠 권리권자가 등록하므로 DB구축 기관에서 DOI 등록율을 높일 수는 없다. 다만 출판사들이 과거 출판 논문에도 소급하여 DOI를 등록하고 있으므로 DB구축기관으로서는 갱신된 정보를 지속적으로 모니터링하고 수집하여 구축하고 주기적으로 재 매핑하는 것이 참고문헌 매핑율을 높이는 최선의 방안으로 사료된다.

V. 결론

본 연구에서는 국내의 학술논문 참고문헌 데이터 구축 사례를 알아보고 인용 분석 활용이 가능한 수준으로 데이터 품질 제고 방안을 알아보고자 하였다.

본 연구에서 대상 데이터로 사용한 e-Gate DB에는 참고문헌 데이터를 포함한 소스 데이터가 대부분 구조화된 상태로 입수되어 데이터베이스의 품질이 매우 높은 수준이다. 하지만 여러 정보원으로부터 입수한 소스 데이터를 통합하므로 여전히 품질관리의 어려움이 있다. 이러한 품질은 다양한 정보원으로부터 입수된 소스 데이터 구조, 데이터 내용의 다양한 오류 검출 및 보정, 메타데이터 통합에 대한 깊은 지식과 풍부함 경험에서

비롯된다고 볼 수 있다.

현재 한국과학기술정보연구원(KISTI)의 e-Gate DB 전체 논문과 참고문헌간의 매핑율은 2019년 7월 기준으로 69.4% 수준이다. KISTI가 연구용으로 구매한 SCOPUS 데이터 비교한 결과 SCOPUS DB의 논문과 참고문헌간의 매핑율은 54.6% 수준으로 나타났다. 본 연구에서 적용한 매핑 기준은 DOI가 있으면 충분조건이고 DOI가 없는 경우 학술지명, 논문명, 시작페이지, 출판년도, 권, 1저자의 성 등 6개 항목 완전 일치되는 경우에만 매핑되는 매우 엄격한 수준이다. 따라서 참고문헌 원형 데이터의 오류나 일부 항목이 누락된 경우 매핑이 되지 않았다. 매핑율 향상을 위해서는 보다 다양한 방법을 활용하는 것이 필요하다. 본 연구에서는 참고문헌 매핑율 향상 방안으로 학술지명 전거 데이터를 활용한 매핑 프로세스 개선과 글로벌 식별자인 DOI를 활용한 정기적 매핑 방안을 제시하였다. 특히 참고문헌의 매핑율을 높이기 위해서는 학술지명 전거 데이터를 지속적으로 구축하고 활용하는 것이 필수적이다.

KISTI의 e-Gate DB의 참고문헌 데이터 구축 사례는 e-Gate DB를 활용한 학술 영향력 평가를 위한 마스터 DB로 활용가능성을 보여준다. e-Gate DB는 국내외 학술지를 포함하고 있으며(망라성), 국내에서 주로 구독, 이용, 출판, 인용되는 국내외 학술지를 수록하고 있고(선정 기준의 명확성), 국내외 학술지명 전거 관리를 하는 등 선행 연구에서 제기한 국내 인용색인 데이터베이스들의 문제점을 상당 부분 해소하였다. 이러한 점에서 국내의 타 인용색인 데이터베이스와 명확히 차별화 된다. 또한 Web of Science나 SCOPUS 같은 해외 색인DB들의 비영어권 학술지 선정 형평성 문제를 고려하면 국내 상황에 맞는 참고문헌DB를 보유할 필요성은 충분하다. e-Gate DB는 논문 정보의 양이 많고 품질이 우수하며 KISTI의 NTIS DB, 국내학술지DB와 연계함으로써 한국 R&D 연구 성과에 특화된 분석에 매우 적합한 강점을 가지고 있다. 더불어 참고문헌 매핑율을 높임으로써 한국 상황에 맞는 다양한 학술적 연구 성과를 분석하고 서비스하는데 최적의 데이터베이스로 활용될 수 있을 것이다.

본 연구에서 사용한 데이터베이스 이외의 참고문헌 데이터베이스들의 매핑 기법과 성능 비교를 수행하지 못한 점은 한계로 남는다. 후속으로 논문과 참고문헌 매핑 효율을 더욱 제고하기 위해 다양한 매핑 방법과 기술을 적용하는 실증적 연구를 수행해 보고자 한다.

참 고 문 헌

[1] https://dictionary.cambridge.org/ko/사전/영어/citation, 2019.8.16.

[2] https://en.wikipedia.org/wiki/Citation, 2019.9.4.

[3] T. Laborie and M. Halperin, "Citation Patterns in Library Science Dissertations," Journal of Education for Librarianship, Vol.16, No.4, pp.271-283, 1976.

[4] https://libguide.snu.ac.kr/citation, 2019.8.16.

[5] L. Egghe and R. Rousseau, Introduction to Informetrics : quantitative methods in library, documentation and information science, Elsevier Science Publishers, 1990.

[6] http://www.lisbdnet.com/citation-analysis, 2019.8.16.

[7] 박성미, "한국 통계학 문헌의 계량서지학적 분석," 정보관리학회지, 제5권, 제1호, pp.104-130, 1988.

[8] https://www.endnote.com/downloads/styles, 2019.9.9.

[9] E. Garfield, "Citation Analysis as a Tool in Journal Evaluation : Journals can be ranked by frequency and impact of citations for science policy studies," Science, Vol.178, No.4060, pp.471-479, 1972.

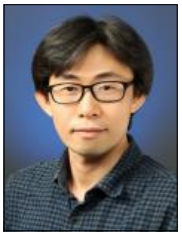
[10] 한정민, 장현철, 김진현, 예상준, 김상균, 김철, 송미영, "학술논문의 참고문헌 자동매핑 방법에 관한 연구," 정보관리연구, 제41권, 제3호, pp.155-173, 2010.

[11] S. Xu, L. Hao, X. An, D. Zhai, and H. Pang, "Types of DOI errors of cited references in Web of Science with a cleaning method,"

Scientometrics, Vol.120, No.3, pp.1427-1437, 2019.
[12] 구희관, 정한민, 성원경, “인용 필드 정규화와 타입이 인용매칭에 미치는 영향,” 한국콘텐츠학회논문지, 제8권, 제11호, pp.395-403, 2008.
[13] P. M. Mongeon and A. Paul-Hus, “The journal coverage of Web of Science and Scopus: a comparative analysis,” Scientometrics, Vol.106, No.1, pp.213-228, 2016.
[14] https://digital.csic.es/handle/10261/162452, 2019.11.14.
[15] C. S. Wagner and S. K. Wong, “Unseen science? Representation of BRICs in global science,” Scientometrics, Vol.90, No.3, pp.1001-1013, 2012.
[16] 김홍렬, 정경희, “국내 참고문헌 데이터베이스 운영현황 및 실태에 관한 분석,” 정보관리학회지, 제22권, 제2호, 2005.
[17] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. Dendek, and Ł. Bolikowski, “CERMINE: automatic extraction of structured metadata from scientific literature,” International Journal on Document Analysis and Recognition, Vol.18, No.4, pp.317-335, 2015.

저 자 소 개

김 재 훈(Jayhoon Kim) 정회원



- 1999년 2월 : 연세대학교 문헌정보학과(학사)
- 2008년 2월 : 성균관대학교 경영학과(석사)
- 2006년 ~ 현재 : 한국과학기술정보연구원 선임연구원

〈관심분야〉 : 디지털 콘텐츠, 콘텐츠 큐레이션

김 순 영(Soon Young Kim) 정회원

- 1992년 2월 : 충남대학교 문헌정보학과(학사)
- 1992년 ~ 2005년 : 한국과학기술원 도서관
- 2006년 ~ 현재 : 한국과학기술정보연구원 선임연구원

〈관심분야〉 : 해외학술논문 데이터베이스, DB설계

임 석 종(Seok Jong Lim) 정회원

- 1998년 2월 : 중앙대학교 문헌정보학과(석사)
- 2009년 8월 : 중앙대학교 문헌정보학과(박사)
- 2005년 ~ 현재 : 한국과학기술정보연구원 선임연구원

〈관심분야〉 : 오픈액세스, 출판 플랫폼, 리포지터리

황 혜 경(Hyekyong Hwang) 정회원

- 1992년 2월 : 서울여자대학교 문헌정보학과(학사)
- 1999년 2월 : 연세대학교 문헌정보학과(석사)
- 2014년 2월 : 연세대학교 문헌정보학과(박사)
- 1999년 ~ 현재 : 한국과학기술정보연구원 책임연구원

〈관심분야〉 : 콘텐츠 큐레이션, 오픈 액세스