참고문헌 자동파싱 및 참조링킹을 위한 Citation Matcher 연구 및 개발

Research and Development of Citation Matcher for Reference Parsing and Cross-Reference Linking

이상기, 김선태, 이용식, 이태석
한국과학기술정보연구원

Lee Sang-gi, Kim Sun-tae, Lee Yong-sik, Yi Tae-seok
Korea Institute of Science and Technology Information

요약

CrossRef에서는 DOI 식별자를 기반으로 출판사 간 참고문헌을 링크하기 위한 참조링킹 인프라를 제공하고 있으며, CrossRef과 연계한 참고문헌의 전자원문 링크 서비스인 참조링킹 체제를 구축하는 기관이 점차 늘고 있다. 본 연구에서는 참조링킹 체제를 효율적으로 구축하기 위해 Citation Matcher를 개발하였다. Citation Matcher는 과거 수작업에 의존하던 참고문헌 DB구축 및 식별자 매칭 프로세스를 자동화하고 패턴화한 것으로, 참고문헌을 원형 그대로 Copy & Paste하면, 참고문헌의 패턴을 분석하여 참고문헌으로부터 저널명, 저자, 권/호 등 Citation 정보를 일목요연하게 파싱하고, 파싱한 정보를 표준화된 방식으로 CrossRef, Pubmed, yesKISTI 등의 메타데이터와 매칭하여 식별자를 획득하여 링크하는 솔루션이다. Citation Matcher는 과거 사람이 수행하던 국내 학술논문의 참고문헌 구축 및 매칭 프로세스를 완전 자동화함으로써 업무 프로세스를 혁신하고, 정보자원 간 연계성 제고 및 논문 간 Seamless한 접근을 통해 이용 편리성을 제고하기 위한 것이다.

Abstract

CrossRef operates a cross-publisher citation linking system based on the DOI® global identifier. The number of organization building a reference citations linking structure through CrossRef is increasing. This paper concentrates on developing a Citation Matcher Solution to effectively build the reference linking structure. Citation Matcher automatically builds and processes the reference citation and identifier mapping which used to be handled manually. After the copy & paste of the reference citation, analyzation is processed to parse the journal title, author name, volume, issue, and start pages from the free style text. CrossRef, PubMed, and YesKISTI's identifiers are collected by through a standardized method. Renovation of the building process for domestic scholastic resources' reference linking and matching will be made possible by using a Citation Matcher. The connection between resources and seamless access for the electronic full-text will enhance the usability.

1. 서 론

최근 들어 링킹혁명(Linking Revolution)이라는 용어가 등장할 정도로 전자원문에 대한 링킹의 중요성이 부각되고 있다. 특히, 이용자들은 도서관 내부자료는 물론이고 외부자료도 단 한번의 방문으로 제공받고 싶어 한다[1].

국내 주요 학술단체와 출판사들을 중심으로 전자간행물에 대한 이용을 극대화시킬 수 있는 방안으로 여러 기관 간에 동종 또는 이종 간행물간 연계를 지칭하는 참조링킹(cross-reference linking)에 대한 관심이 높아지고 있다.

기존의 간행물간 링킹이라는 개념은 주로 개별도서와 표, 목차간의 상관관계를 표시하는 것을 의미했다. 하지만 하이퍼링크 기술의 발전과 함께 참조링킹이 등장하면서 논문초록, 색인과 해당 원문간의 연계는 물론이고 논문을 읽을 때 발생할 수 있는 인용문헌 또는 참고문헌간의 연계, 본문의 특정 구·단어·문장과 타 원문과의 연계 등 다양한 종류의 연계를 가능하게 하는 하이퍼텍스트 저널의 개념으로 확장하여 표현하고 있다[2].

본 연구에서는 참고문헌의 패턴을 분석하여 참고문헌으로부터 저널명, 저자, 권/호 등 Citation 정보를 일목요연하게 파싱하고, 파싱한 정보를 표준화된 방식으로 CrossRef, Pubmed, yesKISTI 등의 메타데이터와 매칭하여 식별자를 획득하는 Citation Matcher 솔루션을 개발하였다. 이를 통해 과거 수작업에 의존하던 참고문헌 DB구축 및 식별자 매칭 프로세스를 기계화하고 자동화하였으며, 업무 효율성을 대폭 개선하고 정보자원 간 연계성을 제고하였다.

2. 연구동향

참조링킹의 경우 참고한 논문들이 다른 출판사인 경우 그 논문들에 접근하기 위해서는 상대방의 데이터베이스를 이용해야 하며, 이는 출판사들이 동의해야 가능한 것으로 수많은

참고문헌들이 각기 다른 출판사에 소속되어 있으므로 참고문헌을 링크시키기 위해서는 모든 출판사들의 동의를 얻어야 하는 불편함이 있었다.

CrossRef에서는 이를 해결하기 위해 출판사들 사이에 각 출판사들의 데이터베이스(DB)를 상호 운영할 수 있는 중간시스템을 두고 그 중간시스템에 각 출판사들이 출판하는 저널과 논문들의 식별자와 정보를 저장하고, 각 출판사들이 자기 DB 사용에 대한 동의를 중간시스템과 체결하면 참고논문의 전자원문(full text)을 제공하는 서비스를 개발하였다. 아울러 학술논문이 인용한 다른 출판사의 논문들을 DOI를 이용하여 연결시킴으로써 참고문헌의 전자원문(Full-Text)으로 링크를 제공하고 있다.

▶▶ 그림 1. CrossRef Simple Text Query[3]

그림 1은 CrossRef에서 서비스하는 Simple Text Query 솔루션으로 출판사나 정보서비스 기관에서는 DOI를 이용하여 손쉽게 참조링킹 체제를 구축할 수 있다. 즉, Simple Text Query 입력폼에 참고문헌 원형을 Copy & Paste한 후 submit하면 참고문헌을 자동 파싱한 다음 DOI를 매칭하여 리턴해 준다.

Pubmed에서도 대량으로 참고문헌을 자동 매칭할 수 있는 솔루션을 개발하여 서비스하고 있다.

그림 2와 같이 입력폼에 Pubmed가 지침화한 파이프 방식의 포맷으로 저널명, 발행연도, 권/호, 시작페이지 등을 입력하여 submit하면 Pubmed 식별자(pmid)를 자동으로 매칭하여 리턴해 준다. Pubmed Batch Citation Matcher의 경우 온라인으로는 100건까지 처리가능하며, 100건을 넘는 경우 파일로 형태로 작성하여 e-mail로 전송하면, Pubmed에서 매칭한 후 결과를 e-mail로 보내준다.

▶▶ 그림 2. Pubmed Batch Citation Matcher[4]

3. KISTI Citation Matcher

KISTI Citation Matcher(KCM)는 KISTI에서 수작업으로 구축 중인 참고문헌 및 식별자 매칭 업무 프로세스를 자동화하기 위해 개발한 솔루션으로, 웹 기반으로 한건씩 처리하는 Single Citation Matcher, 여러 건을 동시에 처리하는 Multi Citation Matcher, 출판된 원형 그대로 입력하면 자동으로 파싱하여 처리하는 Free Citation Matcher 그리고 대량의 정보를 e-mail를 통해 처리하는 Batch Citation Matcher로 구성되어 있다.

이중 특히, Free Citation Matcher의 경우 그림3과 같이 사람이 일일이 참고문헌의 Citation 항목을 분리하여 처리하지 않고, 논문에서 입력한 원형 그대로 Copy & Paste하면 자동 파싱하여 매칭해 주는 솔루션으로 효율성과 편리성이 매우 높은 지능화된 프로그램이다.

참고문헌 자동파싱 및 매칭 솔루션인 Free Citation Matcher 프로세스는 그림 3과 같다.

▶▶ 그림 3. 자동파싱 및 매칭 프로세스

참고문헌이 입력되면 가장 먼저 참고문헌의 기술형식을 확인한다. 학술논문의 참고문헌은 각 학회별로 기술형식이 표준화되어 있기 때문에 기술형식이 어떤 형식을 기준으로 작성되었느냐에 따라 파싱 알고리즘이 달라진다. 미국의 경우 주요 참고문헌 표기형식은 표 1과 같이 5개 형식으로 이중 APA, MLA 방식을 가장 많이 사용하고 있다[5,6].

[표 1] 주요 참고문헌 기술형식

구분	종류	참고문헌 기술형식
APA형식 (사회과학, 자연과학분야)	단행본	저자명. (출판년도). 서명: 부서명. (판차). 출판지: 출판사.
	학술지	저자명. (출판년도). 논문제목: 부제목. 학술지명, 권 (호), 페이지.
MLA형식 (인문과학분야)	단행본	저자명. 서명: 부서명. 판차. 출판지: 출판사, 출판년도.
	학술지	저자명. "논문제목." 학술지명 권.호 (출판년도): 페이지.
Chicago (대학에서 많이사용)	단행본	저자명. *서명: 부서명*. 판차. 출판지: 출판사, 출판년도.
	학술지	저자명. "논문제목." *학술지명* 권.호(출판년도): 페이지.
Turabian (대학에서 많이사용)	단행본	저자명. *서명: 부서명*. 판차. 출판지: 출판사, 출판년도.
	학술지	저자명. "논문제목: 부서명." *학술지명* 권. 호(출판월 년): 페이지.
ISO 690 (국제표준)	단행본	저자명. *서명*. 판차. 출판지 : 출판사, 출판년도.
	학술지	저자명. 논문제목. *학술지명*. 출판년도, 권, 호, p. 페이지.

참고문헌 기술형식이 파악되면 APA, MLA 등 참고문헌 기술형식을 토대로 저널명, 출판년도, 권/호, 저자, 논문 시작페이지 등 Citation 항목을 파싱한다. 파싱된 권/호 및 페이지정보는 매칭 성공률을 제고하기 위해 필터링 과정을 거치는데, 필터링은 다양한 표기에 대응하기 위해 미리 패턴정보를 수집해 놓은 테이블과 비교하여 통상적으로 가장 많이 사용하는 패턴으로 변경하는 작업이다. 필터링된 Citation 정보는 참조링킹을 위해 DOI, PMID, KOI 등 식별자를 순차적으로 매칭하여 DB화한다.

식별자인 DOI, PMID, KOI를 매칭하기 위해서는 각 식별자가 요구하는 표준화된 포맷으로 Citation 정보를 제공해야 한다.

▶▶ 그림 4. KISTI Free Citation Matcher 입력화면

그림 4는 식별자를 매칭하기 위해 참고논문을 출판된 원형 그대로 Copy & Paste한 입력화면이며, 그림 5는 자동파싱 및 매칭 프로세스를 통해 식별자를 매칭한 결과화면으로 CrossRe의 DOI와 Pubmed의 PMID가 성공적으로 매칭된 것을 확인할 수 있다.

▶▶ 그림 5. KISTI Free Citation Matcher 매칭결과 화면

그림 6은 KISTI Single Citation Matcher로 Free Citation Matcher를 이용하여 처리할 수 없거나, 한건씩 정밀 파싱할 필요가 있는 경우에 사용하는 솔루션이다. Free Citation Matcher와 다른 점은 사람이 Citation 정보를 일일이 분리하여 야하며, 한 번에 한건씩밖에 처리할 수 없는 점이다. 하지만 표준화된 방식으로 작성하지 않은 참고문헌이나 Free Citation Matcher로 처리하기 어려운 참고문헌을 정밀 매칭할 때 유용하게 사용할 수 있다. 또한, Single Citation Matcher의 경우 일부분만 알고 있는 Citation 정보를 활하여 관련정보를 매칭할 수 있으며, 저자가 잘못 입력한 참고문헌을 작업자가 수정하면서 매칭할 수 있는 장점이 있다.

▶▶ 그림 6. KISTI Single Citation Matcher

그림 6은 참고문헌 "Acta physiologica Scandinavica" 저널, vol(180), issue(4), 405 page를 매칭한 결과로 CrossRef

DOI와 Pubmed PMID 및 yesKISTI 식별자가 성공적으로 매칭된 것을 확인할 수 있다.

표 2는 KISTI에 구축되어 있는 인용색인DB(KSCI) 참고문헌 103,443건을 대상으로 KISTI Citation Matcher를 이용하여 매칭한 결과이다. 매칭결과 DOI의 경우 34,792건(33.6%), ISSN은 38,236건(49%)이 매칭되어 사람이 수작업으로 매칭한 결과와 별 차이가 없는 것을 확인할 수 있었으며, 이를 통해 개발된 솔루션이 매우 우수한 것을 확인할 수 있었다.

[표 2] 참고문헌 매칭율

구 분	ISSN	KOI	DOI	PMID	yesKISTI
매칭건수	38,236	2,254	34,792	1,692	29,315
매칭율	40%	2.2%	33.6%	1.6%	28.3%

※ 참고문헌 103,443건 대비 매칭율

4. 결 론

본 연구에서는 효율적인 참고문헌DB 구축 및 식별자 매칭을 위해 KISTI Citation Matcher를 개발하였다. 이것은 기존 수작업을 통해 참고문헌 DB를 구축하고 식별자를 매칭하던 방식을 완전 자동화한 것이다.

KISTI Citation Matcher는 참고문헌을 APA, MLA 등 규칙화된 패턴을 이용하여 자동 파싱하는 프로세스와 권/호, 페이지정보를 표준화된 포맷으로 변환하는 프로세스 그리고 CrossRef DOI, Pubmed PMIID, KISTI KOI 식별자를 자동 매칭하는 프로세스로 구성되었다.

KISTI Citation Matcher를 이용하여 KISTI가 구축한 인용색인DB(KSCI)의 참고문헌 103,443건을 매칭한 결과 DOI의 경우 34,792건(33.6%), ISSN은 38,236(40%)로 사람이 수작업으로 일일이 매칭한 결과와 별 차이가 없는 것을 확인하였다.

KISTI에서는 Citation Matcher 솔루션을 개발하여 참고문헌 구축 및 매칭 프로세스를 완전 자동화함으로써 업무 프로세스를 혁신하고 경제비용을 절감할 수 있었으며, 이를 기반으로 참조링킹 체제를 구축하여 정보자원 간 연계성 및 접근성을 제고하였다.

▌참 고 문 헌▐

[1] 김성희 “OpenURL을 이용한 전자자원 링킹시스템 비교・분석”, 정보관리학회, 제22권, 제4, 58호, pp.221-234, 2002.

[2] 디지털타임스, “디지털콘텐츠 연계 프로그램(UCL)”, <http://www.dt.co.kr/contents.html?article_no=2006120802012060650001>, 2006.

[3] CrossRef, <http://crossref.org/>, 2007.

[4] Pubmed, <http://www.ncbi.nlm.nih.gov/sites/entrez/>, 2007.

[5] 박온자 “온라인 자료의 인용 및 참고문헌 수록 양식과 국내대학 및 학술잡지에서 사용하고 있는 인용 및 참고문헌 수록 양식 조사연구”, 정보관리학회, 제16권, 제2호, pp81-104, 1999.

[6] 남영준 “참고문헌의 서지기술 표준에 관한 연구”, 한국문헌정보학회지, 제39권, 제4호, pp261-279, 2005.