Journal of Korean Library and Information Science Society
Vol. 56, No. 1

공공도서관 MARC 데이터 중복검증 알고리즘 개선 방안 연구

- 부산 지역 M도서관 사례를 중심으로 -

A Study on Improving Duplicate Verification Algorithm for Public Library MARC Data: Focusing on the Case of M Library in Busan

송 민 건(Min-geon Song)*
이 수 상(Soo-Sang Lee)**

〈 목 차 〉

Ⅰ. 서 론
Ⅱ. 중복검증 알고리즘
Ⅲ. MARC 데이터 중복검증 알고리즘 적용
Ⅳ. 알고리즘 개선 및 재적용
Ⅴ. 결 론

요 약 : 본 논문은 본 연구자가 기존에 수행한 중복검증 알고리즘의 적용 연구의 한계점을 보완하고자 수행한 후속 연구에 대한 논문이다. 부산 지역의 M도서관으로부터 직접 MARC 데이터를 제공받아 KERIS의 중복검증 알고리즘을 Python으로 구현하여 적용하였다. 도서기호가 일치하는 레코드 쌍을 추출하고 이를 별치기호와 권·연차기호를 기준으로 동일 집단과 불일치 집단으로 나누어 알고리즘 적용 결과를 비교하였다. 동일 집단은 98.10%가, 불일치 집단은 0.43%만이 동일 자료로 판정되었다. 알고리즘 적용 결과 불일치로 판정된 중복레코드 쌍을 분석하여 알고리즘의 개선 방안을 다음과 같이 3가지로 제안하였다. 첫째, 세트(SET) ISBN을 제거하고 판정. 둘째, 발행처 항목 판정에서 전방 또는 후방일치는 일치로 간주. 셋째, 저자 항목 판정에서 전방 또는 후방일치는 일치로 간주. 알고리즘 개선 결과 동일 집단에서는 동일 판정이 98.29%로 상승하였고, 불일치 집단에서는 동일 판정의 변화 없이 불일치 판정이 93.40%에서 93.63%로 상승하였다. 이에 따라 개선 방안이 다른 자료를 중복 자료로 판정하는 오류를 억제하면서 알고리즘 성능을 높일 수 있음을 확인하였다.

주제어 : 공공도서관, 목록데이터, MARC, 중복검증, 통합도서관

ABSTRACT : This paper is a follow-up study to compensate for the limitations of the previous research on the application of the duplicate verification algorithm. MARC data was provided directly from M Library in Busan, and the duplicate verification algorithm of KERIS was implemented and applied in Python. We extracted pairs of records with matching book numbers and divided them into 'same group' and 'mismatch group' based on matching location symbols and volumes, and compared the results of the algorithm. As a result of applying the algorithm, 98.10% of the 'same group' and only 0.43% of the 'mismatch group' were determined to be the same material. By analyzing the duplicate record pairs that were determined to be mismatched as a result of the algorithm, we proposed three ways to improve the algorithm as follows. First, remove ISBNs that contain the phrase SET. Second, consider forward or backward matches as matches in the publisher category. Third, forward or backward matches for author entries were considered matches. As a result of the algorithmic improvements, the identical judgment increased to 98.29% in the same group, and the mismatch judgment increased from 93.40% to 93.63% with no change in the identical judgment in the mismatch group. This shows that the improvements can increase algorithm performance while suppressing the error of labeling different materials as duplicates.

KEYWORDS : Public Library, Catalog Data, MARC, Duplicate Verification, Integrated Library

* 부산대학교 문헌정보학과 박사과정(mgs207@pusan.ac.kr / ISNI 0000 0005 1420 3658) (제1저자)
** 부산대학교 문헌정보학과 교수(sslee@pusan.ac.kr / ISNI 0000 0000 6434 9851) (교신저자)

• 논문접수: 2025년 2월 24일 • 최초심사: 2025년 3월 6일 • 게재확정: 2025년 3월 11일
• 한국도서관·정보학회지, 56(1), 289-305, 2025. http://dx.doi.org/10.16981/kliss.56.1.202503.289

Ⅰ. 서 론

1. 연구의 배경과 목적

국내 공공도서관의 목록레코드는 대체로 개별 도서를 단위로 하여 작성되고 있다. 즉, 권차나 복본에 상관없이 개별 도서 1책이 모두 하나의 레코드를 이루고 있다. 이러한 방식은 검색 결과 레코드가 매우 과다하게 나타나 이용자가 도서를 찾는 데 큰 장애 요소가 된다는 문제점이 있다.

이는 어느 한 도서관만의, 어느 한 지역만의 문제가 아니다. 노지현과 이은주(2023)의 연구에 따르면 부산 지역의 모든 공공도서관은 "1책당 1개의 레코드"를 원칙으로 서지데이터가 구축되고 있다. 부산 지역 이외에도 16개 광역대표도서관을 조사한 결과 동일 자료(복본)를 책 단위로 분리 구축하는 방식이 절대적으로 많았으며, 서울도서관, 대전한밭도서관 등 2개 관만이 동일 자료의 복본에 대해 하나의 서지레코드에 복수의 소장 정보를 연결하는 목록 작성 방식을 적용하고 있었다.

특히, 광역대표도서관은 도서관법 제26조에 따라 지역도서관의 협력네트워크를 구축 및 운영하고 지역 단위의 종합적인 도서관 자료를 수집 정리 보존 및 제공하기 위해서는 지역 내 도서관들의 통합목록을 구축해야 한다. 이러한 통합목록에서는 많은 도서관에서 소장한 같은 복본이 모두 개별 레코드를 구성하기 때문에 과도한 서지레코드로 인한 문제점이 극대화된다.

예를 들어, M도서관의 홈페이지에서 '불편한 편의점'을 검색(검색일: 2024년 2월 17일)하면 『불편한 편의점』의 1권이 6건, 2권이 1건, 1권의 큰글자도서가 5건, 2권의 큰글자도서가 2건 총 14건의 레코드가 검색된다. 이는 개별 저작 단위의 목록레코드의 경우 4건의 서지레코드로 충분하며, 다권본의 통합기술방식(Multipart Monograph Analyzed and Classed Separately)을 적용할 경우, 1건의 서지레코드에 14건의 소장 정보를 입력하는 것으로 충분하다. 부산 지역의 대표도서관인 부산도서관에서 구축한 부산 지역 공공도서관의 통합목록인 부산도서관포털에서는 '불편한 편의점'을 검색(검색일: 2025년 2월 17일)하면 무려 918건의 레코드가 검색된다.

이러한 문제점을 해소하고 RDA, BIBFRAME 등의 차세대 서지기술 개념이 적용된 목록 규칙의 발전 방향에 대응하기 위해서는 개별 도서 단위의 목록레코드를 통합해야 한다. 이러한 목록레코드 통합 과정의 첫 단계는 우선 동일한 도서의 복본을 식별하는 것, 즉 중복검증이다. 즉, 중복검증은 개별 도서 단위로 작성된 목록레코드를 통합하고 과도한 레코드 생성으로 인한 문제점을 해결하고 차세대 목록규칙에 적응하기 위한 첫 단계라는 점에서 중요하다.

본 연구는 연구자가 기존에 공공도서관의 OPAC 데이터에 대해 중복검증 알고리즘을 적용한 연구(송민건, 이수상, 2024)의 후속 연구로서, 기존 연구의 한계점과 시사점을 토대로 연구를 수행하였다. 기존 연구에서는 공공도서관 홈페이지에서 확인할 수 있는 OPAC 데이터를 활용하였다. 이러한 OPAC 데이터는 MARC 데이터에 기반하고 있기는 하지만, 이용자가 주로 활용하는 데이터 위주로

가공되어 제공된다. OPAC 데이터에서는 판사항과 총서사항 요소, 반복기술된 ISBN 등을 비롯한 MARC 데이터 필드에서 다양한 필드에서 나타나는 데이터 요소를 모두 파악할 수 없다. 이에 온전한 중복검증 알고리즘을 적용하지 못하여, 중복검증 알고리즘의 문제점인지 실제 입력된 데이터의 문제점인지 판정하기 어려워 중복검증 알고리즘 자체에 대한 개선 방안을 제시하지 못한 한계점이 있었다. 이러한 한계로 인해 본 연구자는 MARC 데이터를 직접 추출하고 이를 KERIS 중복검증 알고리즘에 적용하여 더욱 정밀하게 알고리즘을 적용하는 후속 연구가 필요하다고 논의하였다.

이에 본 연구에서는 실제 공공도서관에서 구축한 MARC 데이터에 완전한 중복검증 알고리즘을 적용하여 중복검증 알고리즘의 개선 방안을 도출하고자 하였다. 이에 부산 지역의 M도서관에서 실제 사용하는 MARC 데이터를 제공받아 중복검증 알고리즘을 적용하였다. 본 연구에서 설정한 연구 문제는 다음과 같다.

연구 문제 1. 현행 중복검증 알고리즘을 공공도서관의 MARC 데이터에 적용할 때, 중복 판정의 정확도는 어떠한가?

연구 문제 2. 현행 중복검증 알고리즘을 공공도서관의 MARC 데이터에 적용할 때, 정확한 판정이 내려지지 않는 레코드는 어떤 유형을 하고 있는가?

연구 문제 3. 현행 중복검증 알고리즘을 어떻게 개선할 수 있는가? 구체적으로 정확한 판정이 내려지지 않는 레코드를 정확히 판정하기 위해 어떤 수정을 가해야 하는가?

2. 연구방법

본 연구는 다음과 같은 과정을 통해 수행되었다. 첫째, 연구에 활용할 부산 지역 M도서관의 MARC 데이터를 제공받았다. 둘째, 본 연구에 적용할 중복검증 알고리즘에 대해 파악하고 M도서관의 MARC 데이터 형태에 적용할 중복검증 알고리즘을 Python으로 구현하였다. 셋째, 수집한 M도서관의 MARC 데이터에 대해 Python으로 구현한 중복검증 알고리즘을 적용하였다. 구체적으로, 090 필드가 일치하는 레코드 쌍을 추출하고 049 필드의 $v(권·연차기호)와 $f(별치기호) 서브필드가 일치하는 집단과 불일치하는 집단을 비교하였다. 넷째, 알고리즘 적용 결과 나타난 중복검증률과 불일치 레코드에 대해 분석하였다. 다섯째, 현행 공공도서관의 MARC 데이터에 적용할 수 있는 중복검증 알고리즘 개선 방안을 제안하였다. 여섯째, 제안된 개선 중복검증 알고리즘을 적용하고 기존 결과와 비교하였다.

3. 선행연구

국내 공공도서관이 구축하고 있는 목록레코드에 대해 분석한 연구로는 노지현과 이은주(2023)가

부산 지역 49개 공공도서관의 서지데이터의 품질에 대해 분석한 연구가 대표적이다. 분석 결과 부산 지역 서지데이터의 가장 근본적인 문제는 동일 자원(복본)에 대해 아이템 단위로 서지레코드를 중복 생성하는 구조임을 지적하며 서지데이터를 통합해야함을 강조하였다. 이러한 서지데이터의 통합을 위해서는 중복검증 알고리즘의 활용이 필수적이다.

또한, 서지데이터의 내용적 품질에서 동일한 자료의 발행년도가 서로 다르게 입력되고 일부 데이터가 누락되는 등 다양한 문제점을 내포하고 있음을 확인하였다. 이러한 문제점들은 중복검증 알고리즘을 적용 과정에서 오류를 일으키는 원인이 될 수 있어, 중복검증 알고리즘을 최대한 정교하게 구축해야 할 필요가 있다.

마지막으로, 이러한 연구 결과가 부산 지역의 특수한 결과가 아니라 국내 대부분의 공공도서관에 해당한다고 강조하였다.

해외에서는 세계 최대의 통합목록인 WorldCat을 구축한 OCLC에서 중복레코드를 판정하고 제거하기 위하여 AI 기술을 도입하는 연구를 진행하고 있다. 전 세계에서 다양한 언어로 작성된 목록레코드에 대해 중복 여부를 판정하기 위해 AI 기술을 적용하고 있다. Proffitt(2023)은 OCLC에서 336명의 이용자가 34,000개의 중복레코드 쌍을 지정한 것을 기계 학습 모델에 학습시켜 첫 기계학습 모델을 구현한 것을 발표하였다.

본 연구자인 송민건과 이수상(2024)은 부산 지역의 공공도서관의 OPAC을 웹 크롤링으로 수집하여 KERIS 중복검증 알고리즘을 적용하였다. 그 결과로 공공도서관의 목록레코드에 대해 중복레코드를 판정하는 도구로서 KERIS 중복검증 알고리즘의 활용 가능성을 확인하였다. 이에 더해 알고리즘 적용 결과 중복으로 판정되지 않은 중복레코드 쌍을 분석하였으며, 대부분이 OPAC 데이터의 오류가 원인으로, 간단한 데이터 교정만으로 중복 검증의 품질을 높일 수 있음이 파악되었다. 연구의 한계점으로 OPAC에서는 KERIS 중복검증 알고리즘에 활용하는 일부 요소를 파악할 수 없어, 실제 알고리즘의 개선 방안을 제안하기 위해서는 MARC 데이터를 직접 활용한 연구가 필요하다고 논의하였다.

Ⅱ. 중복검증 알고리즘

국내의 주요 도서관 통합목록으로는 국립중앙도서관의 KOLIS-NET과 KERIS 종합목록이 있다. 두 종합목록 모두 다양한 기관에서 생산하는 중복레코드를 식별하기 위해 중복검증 알고리즘을 개발하여 활용하고 있다.

KOLIS-NET 담당자로부터 메일(2024년 3월 11일)로 확보한 중복검증 알고리즘은 MARC의 035 필드(기관제어번호)와 020 필드, 표제/발행자/발행년을 순서대로 비교하여 일치하는 경우 중복으로 판정한다. 이는 고유 식별자에 대한 완전일치를 기준으로 중복레코드를 판정하는 것으로,

고유 식별자 입력의 오류만 없으면 중복검증에 활용하기에 문제는 없다. 하지만 035 필드의 경우 개별 공공도서관 레코드에서는 작성하지 않고, 나머지 식별자의 경우에도 다양한 오류가 발생할 수 있어 공공도서관의 목록레코드의 중복검증에 적용하기에는 어려움이 있다. 이에 본 연구에서는 기존 연구에서 활용한 KERIS의 중복검증 알고리즘을 사용하였다.

다만 기존 연구에서는 MARC 데이터의 원본을 직접 사용하지 못하고 MARC 데이터를 통해 가공된 OPAC 데이터를 활용하여 중복검증 알고리즘을 간소화하여 적용하였다. 그 과정에서 1xx, 7xx에 기술된 저자사항과 판사항, 총서사항, 그리고 그 외에 반복기술된 필드 등을 확인하지 못하는 문제점이 있었다. 이에 본 연구에서는 MARC 데이터 원본을 직접 입수하여 KERIS의 현행 중복검증 알고리즘을 완전하게 적용할 수 있도록 하였다.

KERIS 중복검증 알고리즘은 KERIS의 담당자로부터 메일(2023년 11월 16일)을 통해 확보한 것을 그대로 활용하였다. 우선 중복을 확인하려는 MARC 레코드 쌍에 대해 각 비교요소의 항목별로 데이터를 추출하고 비교하여 점수를 부여한다. 그리고 각 항목별 점수를 중복 판정 점수표에 적용하여 두 레코드의 중복 여부를 최종 판정하는 과정을 거친다. MARC 데이터에서 추출하는 비교요소는 〈표 1〉과 같고, 비교요소 항목별 점수 부여 기준은 아래 〈표 2〉와 같다.

〈표 1〉 KERIS 알고리즘의 9가지 비교요소와 MARC 데이터 필드

비교요소	MARC 데이터 필드
서명	245 $a(관제관칭 제외), 245 $a(관제관칭 포함), 245 $ab, 245 $abp, 245 $ap, 245 $x, 245 $b, 245 $p, 246 $a, 740 $a, 940 $a
저자명	245 $d(또는 $c)의 정보를 역할어까지 추출 100 $a, 110 $a, 110 $ab, 111 $a, 700 $a, 710 $a, 710 $ab, 711 $a, 900 $a, 910 $a, 910 $ab, 911 $a 를 순서대로 추출 저자정보가 존재하지 않는 경우, 260 $b 정보를 저자정보로 추출
발행처	ISBN 번호를 출판사보다 먼저 비교 008TAG의 26~27 번째 한국대학출판부호와 38~39 한국정부기관부호 추출 260 $b의 출판사정보 추출 502 $b의 학위수여기관정보 추출
발행년	008TAG의 07~10 ('||||'는 발행년 없는 것으로 판정), 260 $c에서 최초 발견되는 연속된 4자리 숫자 260 $c에서 최초 발견되는 연속된 숫자정보 (4자리 X) 008TAG의 07~10 정보가 숫자정보가 아니어도 해당 정보를 그대로 추출
페이지	300 $a
판	250 $a('[…]' 사이의 정보 제거 안 함)
총서	490 $a, 490 $v, 830 $a, 830 $v, 440 $a, 440 $v, 400 $a, 400 $v, 410 $a, 410 $v, 411 $a, 411 $v, 245 $a
인식번호	020 $a에서 10자리(13자리)의 ISBN 추출 020에서 추출 할 때, ISBN 계산식에 의해, 10자리 혹은 13자리 ISBN 판별, 10자리 ISBN은 13자리 추가 추출, 13자리 ISBN의 경우 10자리 ISBN 추가 추출, 이후 MARC대 MARC 비교에서 ISBN 비교에 모두 사용 022 $a, $z, $y에서 9자리의 ISSN 추출 010 $a, $z에서 13자리의 LCCN 추출
권차	245 $n (245 $n이 존재하지 않는 경우, 090 $c 등의 정보를 참조하지 않는다.)

〈표 2〉 KERIS 알고리즘의 각 비교요소별 점수 산정 기준

비교요소	설명	점수 산정
서명	정보비교는 동자이음어, 사전변환, 표준형 변환을 한 정보로 비교한다. 서명에 한자가 있을 경우, 한자 독음이 2개 이상 나올 수 있을 경우, 각각을 따로 추출하여 비교	245 $ab, 245 $ap, 245 $abp 정보간 완전일치시 5점
		완전일치하고 두 정보 중 하나는 245 $ab, 245 $abp, 245 $ap 중 하나이고 다른 하나는 245 $x이면 4점
		완전일치하나 두 정보 중 하나라도 245 $a($b/$p는 다르고 $a만 일치), 245 $b, 245 $p, 740 $a, 940 $a, 246 $a이면 3점
		서명 부분일치시 (최소 서명 길이 6자 이상, 80% 일치시) 2점
		기타 0점
저자명	정보비교는 표준형 변환을 한 정보로 비교한다. 245 $d/$c의 경우, 역할어까지 추출 1xx, 7xx, 9xx 의 경우, ()내의 정보는 제거하고 추출 저자명 정보 추출시, ‘,’ ‘;’ 이후의 정보는 무시 저자명에 한자가 있을 경우, 한자 독음이 2개 이상 나올 수 있을 경우, 각각을 따로 추출하여 비교	245 $d/$c가 일치하면 3점
		245 $d/$c를 제외하고 첫번째 저자 추출정보가 일치하면 3점
		저자정보 중 하나라도 일치하는 게 있으면 1점
		해당사항이 없으면 0점
발행처	ISBN 비교 점수가 4점 이상이면, 발행처 점수를 4점 ISBN 비교 점수가 4점 미만이면, 이래의 순서로 출판사 점수 비교함. 발행처 비교 점수와 ISBN 비교 점수 중 더 높은 것을 발행처 비교 점수로 평가	추출한 260 $b 정보가 동일하면 4점, ( )안의 데이터는 무조건 제외
		추출한 260 $b가 반복될 경우, 각각을 추출해서 점수 비교 → 하나만 일치해도 4점
		추출한 502 $b 정보가 동일하면 4점
		260 $b 정보가 Head/Tail match일 경우 2점
		해당사항이 없으면 0점
발행년	4자리 숫자정보를 가지는 발행년에 대해서는 +/- 계산까지 처리 발행년이 ‘19uu’ 등의 문자정보를 포함하거나, 4자리 숫자정보가 아닌 경우, 완전일치만 처리	추출한 출판년 정보가 완전일치하면 4점
		4자리 숫자인 출판년 정보를 +/- 1 했을 때 출판년이 동일해지면 2점
		해당사항이 없으면 0점
페이지	300a의 숫자그룹을 모두 추출하여, 페이지 정보를 각각 비교	두 서지에 페이지 정보의 개수와 내용이 모두 일치하면 5점
		두 서지에 페이지 개수는 일치하지 않으나, 동일한 내용이 있으면 3점
		두 서지에 모두 페이지 정보가 존재하지 않거나 한쪽 서지만 페이지 정보가 존재할 경우 2점
		해당사항이 없으면 0점(페이지 정보가 완전히 다른 경우)
판	정보비교는 사전변환, 표준형 변환을 한 정보로 비교한다.	추출한 정보가 완전일치하면 3점
		두 서지에 판 정보가 모두 없으면 3점
		해당사항이 없으면 0점
총서	정보비교는 사전변환, 표준형 변환을 한 정보로 비교한다.	추출한 총서사항 정보의 $a와 $v가 모두 같으면 3점
		두 서지에 모두 총서사항이 없으면(245 $a는 제외) 3점
		추출한 총서사항 정보의 $a만 같거나, ‘총서태그 $a’와 ‘245 $a’가 같을 경우 2점 (245$a 끼리는 서로 비교하지 않음)
		해당사항이 없으면 0점
인식번호	정보비교는 표준형 변환을 한 정보로 비교한다. ISBN의 경우, 중간에(‘-’문자가 있는 경우는 ‘-’문자를 제거한 후 비교 인식번호의 비교는 동일유형(ISBN, ISSN, LCCN)간에만 비교	$a 의 추출갯수와 내용이 모두 동일하면 5점
		$a 의 추출 갯수는 다르나, 일치하는 내용이 있는 경우 4점
		$a, $z, $y 에 상관없이 일치하는 내용이 하나라도 있으면 3점
		둘 다 없으면 2점
		해당사항이 없으면 0점
권차	정보비교는 사전변환, 표준형 변환을 한 정보로 비교한다. 권차 정보에 로마자 숫자가 있는 경우, 일반 숫자로 변환한다. 권차 정보의 맨 처음이 “제3권” 형식으로 되어 있으면, “제”문자를 제거하고 비교한다.	두 서지에 모두 권차가 있으면서 일치하면 3점
		두 서지에 모두 권차가 없으면 2점
		한 서지에만 권차가 있으면 1점
		두 서지에 모두 권차가 있으나 다르면 0점

위의 기준에 따라 매겨진 점수를 중복 판정 점수표에 적용하여 최종 판정을 내린다. 우선순위(역순)에 따라 각 항목의 점수가 모두 기준표에 제시된 점수 이상일 경우 해당 판정을 내리고, 만족하지 못하면 다음 우선순위의 기준 점수를 적용하는 과정을 차례로 반복한다. 해당 과정을 통해 동일 또는 유사 판정을 내리며 아무 기준을 만족하지 못하면 불일치로 판정한다. 단행본, 연속간행물, 학위논문, 다권본, 고서, 비도서의 자료유형별로 다른 점수표를 적용하며, 이 중에서 단행본에 적용되는 판정 점수표는 아래 〈표 3〉과 같고, 다권본에 적용되는 점수표는 아래 〈표 4〉와 같다.

〈표 3〉 KERIS 알고리즘의 단행본 중복 판정 점수표

판정	우선순위(역순)	서명	저자	발행처	발행년	페이지	판	총서	인식번호	권차
동일	6	5	3	4	4	5	0	0	0	0
	5	4	1	4	0	5	3	0	0	0
	4	3	3	4	0	5	3	0	0	0
	3	5	3	2	0	5	3	0	0	0
	2	4	3	2	2	3	0	2	2	0
	1	0	1	2	4	0	0	2	5	2
유사	7	5	3	0	0	0	0	0	0	0
	6	4	0	2	0	0	0	0	0	0
	5	3	3	2	4	0	3	3	2	2
	4	3	0	0	0	5	0	0	0	0
	3	0	0	0	0	3	0	2	3	0
	2	2	0	2	0	5	0	0	0	0
	1	0	0	0	2	0	0	0	5	0

〈표 4〉 KERIS 알고리즘의 다권본 중복 판정 점수표

판정	우선순위(역순)	서명	저자	발행처	발행년	페이지	판	총서	인식번호	권차
동일	5	5	3	4	2	5	0	0	0	2
	4	5	3	4	4	5	3	3	5	0
	3	5	1	2	0	5	3	0	0	2
	2	4	3	4	0	5	3	0	0	2
	1	3	1	2	0	0	3	0	5	2
유사	6	2	3	2	4	5	0	2	0	2
	5	3	0	4	4	3	3	0	2	2
	4	0	1	2	0	5	0	2	0	0
	3	2	0	0	0	2	0	3	5	0
	2	5	0	2	0	0	0	0	2	0
	1	2	1	0	0	0	0	0	2	2

이러한 KERIS의 중복검증 알고리즘의 경우, 대체로 특정 요소가 완전 일치하는 자료들을 동일한 자료로 판정하는 다른 알고리즘들과는 다르게 한 요소가 완전히 일치하거나 완전히 불일치하더라

도 다른 요소들을 종합적으로 판정하여 중복 여부를 판단한다는 점이 특징이다. 인식번호 점수가 0점이라도 동일 자료로 판정될 수 있으며, 반대로 인식번호 점수가 5점이라도 동일 자료가 아니라고 판정될 수 있다. 즉 ISBN이 동일하더라도 다른 자료로 판정될 수 있고, ISBN이 다르더라도 동일 자료로 판정될 수 있다. 이러한 특징은 공공도서관의 목록 업무 과정에서 발생할 수 있는 MARC 데이터의 오탈자와 같은 사소한 오류를 보완할 수 있는 특징이 있다. 이에 KERIS의 중복검증 알고리즘이 공공도서관에 적용하기에 적절하다고 판단하여 활용하였다.

Ⅲ. MARC 데이터 중복검증 알고리즘 적용

1. 분석 대상 MARC 데이터

저자의 기존 연구에서는 공공도서관 내부 데이터를 확보하지 못하여 MARC 데이터를 기반으로 표시되는 OPAC 데이터를 활용하여 중복검증 알고리즘 적용에는 큰 무리는 없었으나, 1xx, 7xx에 기술된 저자사항과 판사항, 총서사항, 그리고 그 외에 반복기술된 필드 등 일부 MARC 데이터가 누락되어 온전한 알고리즘을 적용하지 못하는 등의 문제가 있었다. 이에 따라 알고리즘의 문제보다는 수집된 데이터의 형태에서 문제가 많이 발견되어 알고리즘의 개선 방안을 논의하지 못하고 데이터 처리 위주의 해결 방안을 제시하였다는 한계가 있었다.

이에 따라, 본 연구에서는 알고리즘의 개선 방안을 논의하기 위해 부산의 M도서관으로부터 온전한 MARC 데이터를 직접 제공받아 완전한 중복검증 알고리즘을 적용하여 문제점을 파악하고 개선 방안을 도출하고자 하였다.

M도서관의 MARC 데이터 수집은 2024년 10월 26일과 11월 8일, 2차례에 걸쳐 진행되었다. 10월 26일 기준으로 M도서관의 종합자료실과 어린이자료실에 소장된 101,864건의 도서레코드 중에서 추출 과정에서 오류가 발생한 1,907건을 제외한 총 99,957건의 MARC 데이터를 xml 형태로 수집하였다.

2. 후보 레코드 쌍 추출

99,957건의 레코드를 모두 1:1로 중복검증을 수행하려면 약 50억 번의 중복검증을 수행해야 한다. 모든 레코드에 대해 중복검증을 수행하기에는 어려움이 따른다. 그래서 일부 유사한 데이터가 존재하는 레코드 쌍을 중복 후보 레코드 쌍으로 지정하여 세부 중복검증을 수행하였다.

또한 중복검증 알고리즘은 동일한 중복 자료를 판별하는 것과 동시에 일치하지 않는 자료를 구별할 수 있어야 한다. 동일하지 않은 자료를 동일하다고 판정하는 것 역시 중복검증 알고리즘의

오류로 일종의 제2종 오류[1]이다. 이를 확인하기 위해 동일 집단 레코드 쌍과 불일치 집단 레코드 쌍을 각각 추출하여 알고리즘을 적용 결과를 비교하도록 하였다. 우선 Python을 활용하여 090 필드 전체를 JSON 문자열로 변환하여 동일한 090 필드 값이 있는 레코드 쌍을 추출하였다. 이 중에서 049 필드의 $f(별치기호)와 $v(권·연차기호) 서브필드가 모두 일치하는 집단을 동일 집단으로, 일치하지 않는 집단을 불일치 집단으로 구분하여 중복검증 알고리즘 적용 결과를 비교하였다.

3. 중복검증 알고리즘의 적용

최종적으로, 090 필드가 일치하는 동일 집단 2,521쌍과 불일치 집단 3,047쌍에 대해 Python을 통해 구현한 중복검증 알고리즘을 적용하였다. 비교 대상 중복 후보 레코드 쌍에 대해 위의 〈표 2〉에 해당하는 알고리즘 비교요소에 해당하는 MARC 레코드 요소를 추출하고 〈표 3〉에 해당하는 기준에 따라 항목별 점수를 부여하였다.

다권본/단행본 여부에 따라 서로 다른 점수 기준표를 적용하여 중복 판정을 수행하기 때문에 다음과 같이 다권본/단행본 여부를 판정하는 기준을 설정하였다.

첫째, 049 필드의 $v(권·연차기호), 245 필드의 $n(권차/편차기호), 440 필드 또는 490 필드의 $v(총서번호)가 존재하는 경우

둘째, 020 필드 내에서 '세트' 또는 'SET'라는 문구가 기술된 경우

위 2가지 기준 중 하나를 만족하는 경우를 다권본으로, 나머지를 단행본으로 설정하여 각각에 맞는 점수 기준표를 적용하였다.

M도서관의 MARC 데이터를 통해 추출한 중복 후보 레코드 쌍에 중복검증 알고리즘을 적용한 결과는 〈표 5〉와 같다.

〈표 5〉 M도서관 데이터에 관한 KERIS 변경 알고리즘 적용 결과

(단위: 쌍)

구분	중복 후보 레코드 쌍	판정 결과		
		동일	유사	불일치
동일 집단	2,521	2,473 (98.10%)	20 (0.79%)	28 (1.11%)
불일치 집단	3,047	13 (0.43%)	188 (6.17%)	2,846 (93.40%)

090 필드와 049 $f, $v가 모두 일치하는 동일 집단의 경우 98.10%가 동일 자료로 판정되었으며, 불일치 집단의 경우 0.43%만이 동일 자료로 판정되었다.

1) 귀무가설이 거짓임에도 불구하고 귀무가설을 채택하는 오류

Ⅳ. 알고리즘 개선 및 재적용

1. 불일치 판정 도서 사례 파악

동일 집단에서 유사 또는 불일치 판정이 나온 48건의 사례를 분석하여 실제 동일한 자료인지 아닌지, 동일 자료라면 왜 동일 판정이 나오지 않았는지를 분석하였다. 분석 결과 다음과 같이 개선이 필요한 레코드의 사례를 3가지 파악하였다.

〈표 6〉 알고리즘 적용 결과 유사 판정 동일 자료 레코드 예시 1

필드	서브 필드	레코드 1	레코드 2
	001	KMO201606782	KMO201701369
	005	20161213135942	20170404152809
	007		ta
	008	161118s2016 ulka 000a kor	170328s2016 ggka 000a kor
020	a	9791195444847	9791195444854(set)
	g	13730	13730
	c	₩13500	-
020	a	-	9791195444847
	g	-	13730
	c	-	₩13500
041	a	jpn	jpn
040	a	121015	121015
	c	121015	121015
049	l	MJ0000227878	MJ0000227878
	c	-	2
056	a	730	730
	2	6	6
090	a	730	730
	b	118	118
100	a	-	임지호
245	a	시즈의 일본어 노트	시즈의 일본어 노트
	b	시즈와 함께하는 감성 일본어	시즈와 함께 하는 감성 일본어
	d	김연진 지음	김연진 지음
260	a	서울	파주
	b	Orbita(오르비타)	Orbita
	c	2016	2016
300	d	187 p.	183p.
	e	천연색삽화	천연색삽화
	f	21 cm	22cm
546	a	-	본문은 한국어, 일본어가 혼합수록됨
653	a	일본어학습	외국어학습
700	a	김연진	김연진
740	a	시즈와 함께하는 감성 일본어	-
740	a	시즈와 함께하는 감성 일본어	-
950	b	₩13500	₩13500

동일 집단 내 유사 판정을 받은 사례 중에서 위 〈표 6〉의 사례는 020 $a의 추출 개수가 달라 인식번호 점수가 5점이 아니라 4점으로 판정된 사례이다. 다권본 자료는 개별자료의 ISBN과 전체 세트 ISBN을 모두 발급받고 이를 모두 MARC 데이터에 입력한다. 문제는 세트 ISBN의 경우, 동일 시리즈의 도서인지를 판정하는 데는 매우 유용하지만 개별자료의 중복 여부를 판정하는 과정에서 혼란을 준다. 세트 ISBN의 조정을 통해 인식번호 점수를 4점에서 5점으로 조정하는 것은 인식번호 점수 5점을 요구하는 동일 판정 기준이 존재하기에(단행본의 우선순위 1, 다권본의 우선순위 4, 1) 의미가 있다.

〈표 7〉 알고리즘 적용 결과 유사 판정 동일 자료 레코드 예시 2

필드	서브 필드	레코드 1	레코드 2
001		KMO200800173	KMO200802541
005		20080129155008	20080704144754
008		080129s2007 ulk 000a kor	070212s2007 ulka 000a kor
020	a	8946415850	9788946415850
	g	03810	03810
	c	₩15000	₩15000
040	a	121015	121015
	c	121015	121015
049	l	EM0000153712	MJ0000156431
	c	-	2
056	a	594.504	594.504
	2	4	4
090	a	594.504	594.504
	b	3	3
100	a	-	임지호
245	a	마음이 그릇이다 천지가 밥이다	마음이 그릇이다 천지가 밥이다
	b	당신을 위해 차리는 29가지 밥상	-
	d	지은이: 임지호	임지호 지음
246	i	-	관제
	a	-	당신을 위해 차리는 29가지 밥상
260	a	서울	서울
	b	샘터사	샘터
	c	2007	2007
300	d	247p	247p
	e	색채삽도	컬러삽도
	f	23cm	23cm
500	a	등록 2008.1.24	등록 2008.7.9
653	a	밥상	밥
700	a	임지호	-
	a	지은이: 임지호	-
950	b	₩15000	₩15000

동일 집단 내 유사 판정을 받은 사례 중에서 위 〈표 7〉의 사례는 발행처 정보의 끝에 '사'라는 문구가 들어가 발행처가 불일치한 사례이다. ISBN 점수가 4점 이상인 경우 발행처 점수도 자동으로 4점이 부여되기 때문에 대부분 큰 문제가 되지는 않지만, 위 사례와 같이 ISBN 작성에서도 오류가 발생하면 문제가 된다. 특히 중복 판정 점수표에 따르면 ISBN 점수가 0점인 경우에는 동일 판정을 받기 위해 발행처 점수를 4점을 받을 것을 요구하는 기준이 많아, 해당 사례에 대해서도 개선이 필요하다.

〈표 8〉 알고리즘 적용 결과 유사 판정 동일 자료 레코드 예시 3

필드	서브 필드	레코드 1	레코드 2
001		KMO201909304	KMO201905354
005		20191113135110	20190626145949
007			ta
008		191112s2019 ulk 000a kor	190619s2019 bnka 000a kor
020	a	9788990969002	9788990969002
	g	03660	03660
	c	₩13000	₩13000
040	a	121015	121015
	c	121015	121015
	d	121015	121015
049	l	BRN000251225	BRN000247318
	c	2	-
056	a	668	668
	2	6	6
090	a	668	668
	b	163	163
100	a	글: 이인미	-
245	a	기억하는 도시, 부산	기억하는 도시, 부산
	b	이인미가 기억하고 사진을 찍다	이인미가 기억하고 사진을 찍다
	d	글: 이인미	이인미 글·사진
	e	사진: 이인미	-
260	a	부산	부산
	b	비온후	비온후
	c	2019	2019
300	d	1책	1책
	e	천연색삽화	천연색삽화
	f	18 cm	18cm
653	a	부산자료	부산출판
700	a	글: 이인미	이인미
950	b	₩13000	₩13000

동일 집단 내 유사 판정을 받은 사례 중에서 위 〈표 8〉의 사례는 저자 정보의 역할어를 작성한 양식이 일치하지 않아 저자 점수가 0점이 부여되어 동일 판정을 받지 못하였다. 점수표에 따르면, 동일 판정에 해당하는 기준은 모두 저자 점수가 3점 또는 1점 이상을 요구하기 때문에 저자 점수가 0점인 경우에는 어떠한 기준에서도 동일 판정을 내릴 수 없다. 이에 따라 역할어의 구분 없이 주요 저자에 대해 비교하여 판정을 내릴 수 있는 기준이 꼭 필요하다.

2. 중복검증 알고리즘 개선 제안

동일한 자료에 대해 유사/불일치 판정이 내려진 사례의 데이터를 분석하여 중복검증 알고리즘의 개선안을 마련하였다.

첫째, 인식번호 점수 비교에 적용하는 020 필드(ISBN)를 추출하는 과정에서 '세트' 또는 'SET'라는 문구가 포함된 필드는 추출하지 않도록 한다. 이는 세트 ISBN은 개별자료의 중복 여부를 판정하기에는 적절하지 않기 때문이다.

둘째, 발행처 항목 비교에서 전방일치[2] 또는 후방일치하는 경우 일치하는 것으로 판정하여 4점을 부여하도록 조정하였다. 발행처 항목의 경우 접두사나 접미사의 형태로 '㈜'나 '사'와 같이 사업체임을 뜻하는 글자가 첨가되는 경우가 있으며, 출판사의 하위 브랜드인 임프린트에서 발행한 도서의 경우에는 모기업의 이름이나 임프린트의 이름을 혼용하여 발행처를 기술하는 사례도 흔히 발견된다. 이러한 사례들에 대해서도 발행처가 일치하는 것으로 간주하여 발행처 기술 과정에서 발생하는 오류를 보완하고자 하였다.

셋째, 저자 항목 비교에서도 전방일치 또는 후방일치하는 경우 일치하는 것으로 판정을 조정하여 3점 또는 1점을 부여하였다. 245 $d/e에서 일치하거나 1xx, 7xx, 9xx에서 추출한 저자정보 중에서 제일 먼저 등장한 저자정보가 일치하면 3점을 부여하고, 그 외에 하나라도 일치하면 1점을 부여하는 데 여기서 전방일치 또는 후방일치의 경우도 일치로 간주하여 판정하는 것이다. 이러한 조정을 통해 역할어 없이 이름만 입력된 경우 외에도 저자의 호나 법명 등과 같은 별칭이 함께 입력된 경우에 대해서도 동일한 판정이 가능하다.

3. 개선 중복검증 알고리즘 재적용

중복검증 알고리즘의 개선 방안을 M도서관의 MARC 데이터에 재적용한 결과는 아래 〈표 9〉와 같다.

---

2) 단순히 서로의 첫 부분이 같은 경우가 아닌 한쪽이 다른 한쪽으로 시작하는 형태를 전방일치로 한다. 예를 들어, '중복'과 '중복검증'은 전방일치이나 '중복검증'과 '중복 후보'는 전방일치가 아니다. 이는 후방일치에서도 마찬가지로 적용한다.

〈표 9〉 M도서관 데이터에 관한 KERIS 알고리즘 개선 방안 적용 결과 비교

(단위: 쌍)

구분	개선 여부	중복 후보 레코드 쌍	판정 결과		
			동일	유사	불일치
동일 집단	개선 이전	2,521	2,473 (98.10%)	20 (0.79%)	28 (1.11%)
	개선 이후		2,478 (98.29%)	12 (0.48%)	31 (1.23%)
불일치 집단	개선 이전	3,047	13 (0.43%)	188 (6.17%)	2,846 (93.40%)
	개선 이후		13 (0.43%)	181 (5.94%)	2,853 (93.63%)

개선 방안 적용 결과, 동일 집단 내에서 현행 알고리즘에서 유사 판정을 내렸던 20쌍 중 5쌍은 동일로, 3쌍은 불일치로 판정이 재조정되었다. 그리고 불일치 집단에서 유사 판정을 내렸던 188쌍 중 7쌍이 불일치로 판정이 재조정되었다.

동일 집단에서 유사 판정에서 동일 판정으로 조정된 5쌍 중 3쌍은 저자 점수의 조정으로, 나머지 2쌍은 각각 인식번호 점수와 발행처 점수가 조정되어 판정이 바뀐 것으로 확인되었다.

동일 집단에서 유사 판정에서 불일치 판정으로 바뀐 3쌍과 불일치 집단에서 유사 판정에서 불일치 판정으로 바뀐 7쌍의 경우 모두 다권본의 다른 권차의 도서들로, 공유하는 세트 ISBN이 판정에서 제외되면서 인식번호 점수가 낮아져 불일치 판정으로 조정된 사례이다. 즉, 현행 알고리즘에서는 세트 ISBN이 판정에 활용되기 때문에 020 $a 중 세트 ISBN이 일치하여 4점이 부여되었으나, 개선 알고리즘에서는 세트 ISBN이 제거되어 개별 ISBN만을 활용하므로 일치하는 것이 없어 0점으로 조정되었다.

이에 따라, 개선 알고리즘이 현행 알고리즘에 비해 다른 자료를 중복 자료로 판정하는 오류는 억제하면서 중복 자료를 식별하는 중복검증률을 높일 수 있음을 확인하였다.

Ⅴ. 결 론

본 연구에서는 연구자가 기존 연구를 수행하는 과정에서 발생한 한계점을 보완하여 후속 연구를 수행하였다. 공공도서관의 실제 MARC 데이터를 제공받아 KERIS 중복검증 알고리즘을 온전히 적용하여 중복검증 알고리즘의 활용 가능성을 확인하였다. 중복검증을 수행하기 위한 중복 후보 레코드 쌍을 동일 집단과 불일치 집단으로 나누어 추출하여 비교한 결과 동일 집단의 98.10%가 동일 자료로, 불일치 집단에서는 0.43%만이 동일 자료로 판정되었다. 이에 공공도서관에서 개별 도서 단위로 작성된 수많은 중복레코드를 통합하기 위한 도구로서 KERIS 중복검증 알고리즘의 적용 가능성을 검증하였다.

이에 더해 중복검증 알고리즘을 통해 판정이 잘못된 사례들을 분석하여 중복검증 알고리즘의 개선 방안을 다음과 같이 제안하였다. 첫째, 세트(SET) ISBN을 제거하고 판정할 것. 둘째, 발행처 항목의 판정 과정에서 전방일치 또는 후방일치는 일치로 간주할 것. 셋째, 저자 항목의 판정 과정에서 전방일치 또는 후방일치는 일치로 간주할 것.

이러한 개선 방안을 적용한 결과, 현행 알고리즘에 비해 동일 집단에서 5쌍이 추가로 동일 판정을 받은 것으로 개선되었다. 추가로 동일 집단에서 3쌍, 불일치 집단에서 7쌍이 유사 판정에서 불일치 판정으로 수정되었으며 이들은 모두 같은 다권본의 다른 권차의 도서였다. 결과적으로, 개선 알고리즘이 현행 알고리즘에 비해 오류는 억제하고 중복검증률은 높일 수 있다는 점을 확인하였다.

본 연구는 부산 지역의 M도서관이라는 개별 도서관 도서의 목록레코드에 대해서만 중복검증을 수행하였다. 이로 인해 중복검증 알고리즘의 개선 성능이 통계적으로 유의미한 것인지에 대해 검증하지 못했고, 서로 다른 도서관에서 작성한 다양한 형태의 목록레코드에 대한 중복검증 과정에서 발생할 수 있는 문제점과 개선 방안에 대해 파악하지 못했다는 한계점이 있다. 하지만 이러한 한계점에도 불구하고 KERIS 중복검증 알고리즘이 공공도서관의 목록레코드에 적용할 수 있으며, 그 과정에서 필요한 개선 방안을 제안하고 검증한 점에 그 의의가 있다.

본 연구에서 수행한 개별 도서관의 중복검증뿐 아니라 알고리즘을 통해 파악한 중복레코드를 앞으로 어떻게 통합할 것인지에 대해서도 논의와 후속 연구가 필요하다. 모든 중복레코드의 서지사항이 같다면 큰 문제 없이 통합할 수 있겠지만 동일한 도서에 대해서도 서지기술 방식이 천차만별이다. 개별 도서관의 중복레코드를 파악하고 통합한 다음에는 더 나아가 지자체를 비롯한 다수의 도서관의 목록을 통합한 통합목록의 중복검증 및 통합 과정에 대한 후속연구가 필요하다.

이러한 대규모 통합목록의 중복레코드 검증 및 통합 과정에서는 머신러닝과 AI 기술의 활용 여부를 검토할 수 있을 것이다. 도서관 목록레코드의 중복검증은 대규모의 데이터에서 유사한 패턴을 파악해낸다는 특성상 머신러닝과 AI를 적용하려는 시도가 활발히 이루어질 것으로 보인다. 도서관 목록레코드의 중복검증과 데이터 통합 과정에서 머신러닝과 AI의 활용 가능성에 대해서도 후속연구가 필요하다.

참 고 문 헌

국립중앙도서관 (2023). 한국문헌자동화목록형식(KORMARC) - 통합서지용.

국립중앙도서관 (발행년불명). KOLIS-NET 종합목록 중복검증 알고리즘. [KOLIS-NET 종합목록 담당자로부터 메일(2024. 3. 11.)로 전달받음]

노지현, 이은주 (2023). 공공도서관 서지데이터의 품질 제고 방안 - 부산시 공공도서관을 중심으로 -.

한국도서관·정보학회지, 54(3), 105-128. https://doi.org/10.16981/kliss.54.3.202309.105
도서관법. 법률 제19592호.
송민건, 이수상 (2024). 공공도서관 목록데이터의 중복검증에 관한 연구 - 부산 지역 G도서관 사례를 중심으로 -. 한국도서관·정보학회지, 55(1), 1-26. https://doi.org/10.16981/kliss.55.1.202403.1
조순영 (2003). 종합목록의 중복레코드 검증을 위한 알고리즘 연구. 한국문헌정보학회지, 37(4), 69-88. http://uci.or.kr/G704-000226.2003.37.4.001
한국교육학술정보원 (2006). 중복검사 및 품질평가(종합목록). [KERIS 종합목록 담당자로부터 메일(2023. 11. 16.)로 전달받음]
LC Program for Cooperative Cataloging (2015). Series Statements and Series Authority Records Session 10: Multipart Monographs. Available: https://www.loc.gov/catworkshop/courses/naco-full%20series-RDA/Session10MultipartMonographs.pdf
Proffitt, M. (2023, August 14). Machine Learning and WorldCat: improving records for cataloging and discovery. Hanging Together the OCLC Research blog. Available: https://hangingtogether.org/machine-learning-and-worldcat-improving-records-for-cataloging-and-discovery/

• 국한문 참고문헌의 영문 표기
(English translation / Romanization of references originally written in Korean)

Cho, Sun-Yeong (2003). A study on duplicate detection algorithm in union catalog. Journal of the Korean Society for Library and Information Science, 37(4), 69-88. http://uci.or.kr/G704-000226.2003.37.4.001
Korea Education and Research Information Service (2006). Duplicate Verification and Quality Assessment(Union Catalog). [Received via email from KERIS officer (2023, November 16)]. National Library of Korea (n.d.).
Libraries Act. Act No. 19592.
National Library of Korea (2023). Korean Machine Readable Cataloging Format - Integrated Format for Bibliographic Data.
National Library of Korea (n.d.). KOLIS-NET Duplication Verification Algorithm [Received via email from National Library of Korea officer (2024, March 11)].

Rho, Jee-Hyun & Lee, Eun-Ju (2023). Improving the quality of bibliographic data in public libraries: focusing on public libraries in Busan Metropolitan City. Journal of Korean Library and Information Science Society, 54(3), 105-128. https://doi.org/10.16981/kliss.54.3.202309.105

Song, Min-geon & Lee, Soo-Sang (2024). A study on duplication verification of public library catalog data: focusing on the case of G library in Busan. Journal of Korean Library and Information Science Society, 55(1), 1-26. https://doi.org/10.16981/kliss.55.1.202403.1