연속간행물 종합목록의 중복레코드 최소화 방안 연구

A Study on the Duplicate Records Detection in the Serials Union Catalog

이혜진*, 김순영**, 김완종*, 최호남***
한국과학기술정보연구원 해외정보팀 연구원*,
한국과학기술정보연구원 해외정보팀 선임연구원**,
한국과학기술정보연구원 해외정보팀장***

Lee, Hye-jin*, Choi, Ho-nam**, Kim, Wan-jong*,
Kim Soon-young***
KISTI Researcher.*, KISTI Senior Researcher**,
KISTI Overseas Information Team Leader***

요약

연속간행물 종합목록은 국내 여러 기관에 산재한 연속간행물의 정보를 통합하여 공유하고, 정보자원화하기 위한 필수 도구로서 최적화된 목록 및 소장 정보를 생성하여 이용자에게 학술지에 대한 신뢰성 있는 정보를 제공하는 것이 목적이다. 이를 위해서는 데이터의 일관성이 무엇보다 중요하며 레코드의 중복성은 종합목록 품질평가에 있어 중요한 척도 중에 하나가 된다. 본 연구는 연속간행물 기반의 종합목록 데이터의 품질을 개선하기 위하여 오류 데이터로 인한 중복레코드를 최소화하기위한 방안을 마련하는데 있다. 이를 위하여 연속간행물의 중복레코드 검증 요소를 분석하고 검증 프로세스를 제안하였다.

Abstract

A Serials Union Catalog is an essential Bibliographic Control tool for integrated and shared the serials information which is scattered to the domestic libraries. It provides reliable informations about serials to user through creating optimized catalogs and holding informations. It is important of the consistency of the bibliographic record and the record's duplication ratio is an important criterion about Database Quality Assessment. This paper checks bibliographic data elements and proposes the duplicate detection process to improve union catalog quality for minimizing duplicate detection.

I. 서 론

연속간행물 종합목록은 국내 여러 기관에 산재한 연속간행물의 정보를 통합하여 공유하고, 정보자원화하기 위한 필수도구로서 최적화된 목록 및 소장 정보를 생성하여 이용자에게 학술지에 대한 신뢰성 있는 정보를 제공하는 것이 목적이다. 또한 다양한 학술지에 대한 정보를 공동 활용함으로써 도서관의 업무효율을 향상시키는데 기여하고 있다. 그러나 우리나라의 종합목록은 해외의 OCLC Worldcat이나 NII의 Webcat 등과 같이 분담목록을 통해 구축된 종합목록이 아닌 여러 기관에서 기 구축된 서지정보와 소장정보를 한 데이터베이스에 통합한 종합목록이기 때문에 표준화된 규칙에 따른 중복 및 오류데이터를 선별하는 작업이 무엇보다 중요하다. 종전의 중복레코드 선별작업은 간단한 중복 알고리즘을 통해 소량의 데이터의 중복 및 오류데이터를 검증하고, 대다수의 데이터는 육안으로 분석하는 방법을 병행하는 것이 사실이다. 그러나 종합목록에 통합되어지는 데이터의 양이 점점 늘어나면서 그 규모는 방대해지기 때문에 좀 더 정밀한 방법의 알고리즘을 통해 수작업을 통한 중복데이터 처리방법을 최소화할 필요가 있다. 특히, 연속간행물의 경우는 단행본과 달리 계속적인 서명변경 및 표준번호의 변경, 연속간행물의 이전, 후속 저록의 관계 등의 연속간행물의 서지적 특성을 반영하여 중복레코드를 구별해야 한다. 따라서 본 연구에서는 연속간행물 종합목록의 품질을 개선하고, 각 기관에서 생성한 연속간행물의 서지정보 및 소장정보를 통합함으로써 발생되는 비표준화된 데이터에 대한 중복레코드를 식별하기 위하여 중복데이터 유형을 살펴보고, 이를 통해 중복 레코드 검증 요소와 검증 프로세스를 도출하고자 한다.

II. 선행연구

우리나라는 종합목록의 역사가 길지 않기 때문에 중복레코드 검증에 관한 연구는 많지는 않지만 국내 최초로 조순영[3]은 KERIS의 단행본 종합목록을 기반으로 비 표준화된 데이터와 오류 데이터로 인한 중복데이터를 색출하기 위해 기존의 중복 알고리즘의 한계를 분석하고 비교요소간 유사성과 가중치를 이용하여 새로운 알고리즘을 개발하였으며, 이에 대한 효용성을 실험 마스터 파일을 대상으로 입증하였다. 이 밖에 국내에서 관련연구로 이제환[2]은 KERIS 종합목록의 이용자 설문을 통해 종합목록 데이터의 중복성과 데이터 품질과는 밀접한 관계가 있음을 도출하고 제공정보의 중복성을 종합목록 DB 품

질 평가의 지표로 선정하였으며, 김선애, 이수상[1]은 중복성은 DB의 절대적 유용성을 평가하는데 있어 매우 중요한 기준으로 보았으며, 레코드 필드 구조의 완전성과 일관성을 위해서는 DB 품질 평가 시, 레코드의 고유성을 적극 반영할 필요가 있다고 하였고, 표본 레코드의 중복율을 평가지표 측정방법의 하나로 포함시켰다. 또한 최인숙[4]은 디지털자료실지원센터의 종합목록 데이터를 분석한 결과, 동일 저작에 대해 상이한 레코드가 다수 존재하는 중복성의 문제가 존재했으며, 종합목록 데이터 기술의 일관성 결여로 인한 데이터의 중복성은 데이터 품질을 저해하는 상당히 심각한 문제의 원인이라고 보고 있다.

국외의 경우는 종합목록의 역사가 오래되어서 중복레코드와 관련된 품질개선에 대한 많은 연구가 이루어져왔고, 특히 O'Neil[5]은 OCLC Worldcat 중복레코드 발생의 요인으로 입력오류, 가변장과 고정장 필드 간의 비일치 등이라고 보았으며, 이 때문에 중복레코드를 검증하기가 용이하지 않다고 하였다. 또한 O'Neil은 기존의 중복알고리즘의 개선하여 검증되지 않았던 레코드를 대상으로 새로운 알고리즘을 적용하여 전체 중복의 1/3을 줄이는 성과를 이루었다.

III. 연속간행물 종합목록 중복 레코드 검증 요소 및 유형 추출

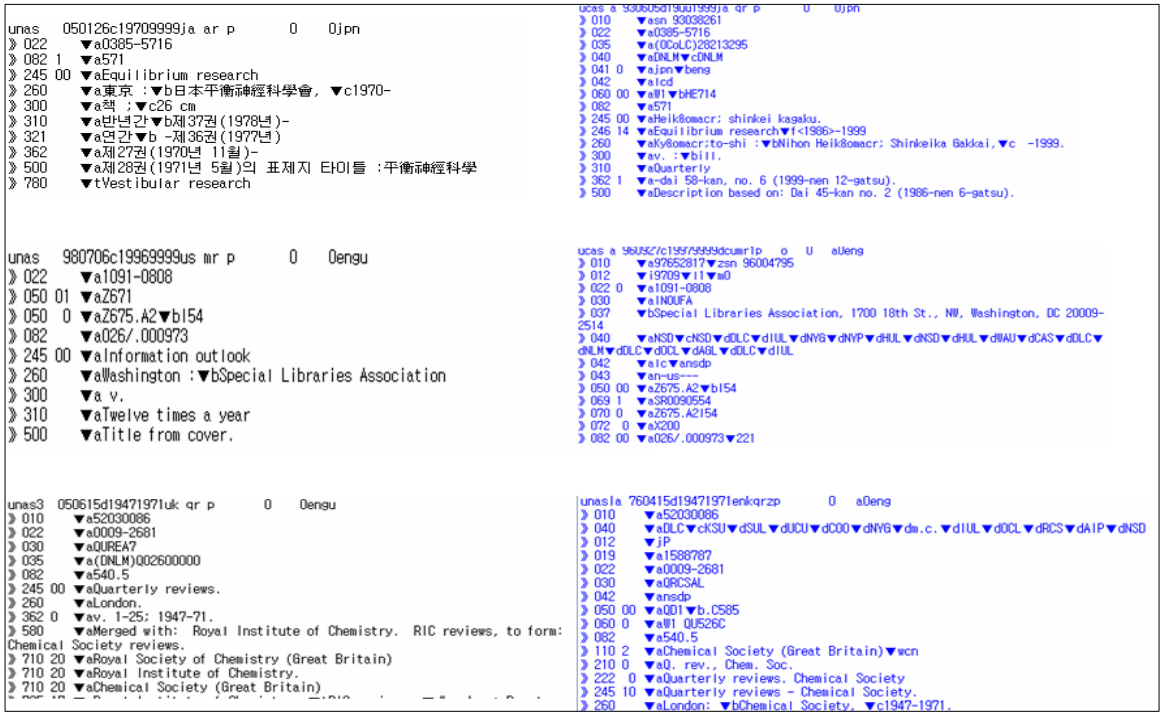
1. 레코드 중복 오류 유형

연속간행물 종합목록 레코드의 중복 데이터를 검증하기 위해서는 여러 가지의 검증 요소와 유형을 선별하는 작업이 필요하다. 인식번호, 서명 등과 같은 기본적인 서지 데이터 요소뿐만 아니라, 연속간행물이 가지고 있는 고유의 특성을 반영하는 데이터 요소를 중복레코드 검증요소로서 채택할 수 있다. 특히, 연속간행물의 전후관계에서 발생되는 출판년도와 폐간년도, 표제변동사항 등은 연속간행물의 중복레코드를 선별하는데 있어서 중요 요소 중에 하나이다.

연속간행물 중복레코드 사례로는 그림 1과 같이

① 서명은 다르나(로마나이즈화) ISSN과 창폐간년이 동일한 경우,
② 서명과 ISSN은 동일한데 창간년이 다른 경우,
③ 서명과 ISSN, 창폐간년은 모두 동일한데 출판사가 누락된 경우

등이 있으며 대개 중복레코드는 MARC 필드 오기, 고정장과 가변장 필드의 불일치, 전후관계에서 오는 창폐간년 오류, 동양서 본표제의 로마나이즈화, 출판사 오기 등의 문제가 존재한다. 특히, 동양서의 경우, copy cataloging을 하면서 로마나이즈화된 서명들이 많으며, 서명에 대한 전거가 없으면 육안으로써 식별하는 방법뿐이 없다. 그리고 KORMARC와 MARC 21 등의 기술형식에 따라 데이터를 표준화시켜서 중복레코드를 검증하도록 해야 한다. 예를 들어, 245 필드의 경우, KORMARC에서는 $x를 대등서명으로 사용하지만, MARC 21에서는 =$b를 사용하며, 발행국의 경우에는 두 형식 모두



▶▶ 그림 1. 중복데이터 오류 사례

ISO 3166 두 자리 국가코드를 따르지만 MARC21은 미국과 영국 등을 도시까지 적용하여 세 자리로 기술하고 있으며, KORMAR은 우리나라를 시도를 적용하여 세 자리로 기술하고 있으므로 주의해야 할 사항이다.

2. 중복레코드 검증 요소 추출

앞서 중복레코드의 중복유형 및 오류 사례를 바탕으로 중복 가능성이 높은 요소들 7개를 추출하였다.

① LCCN(010), OCLCCN(035 (OCoLC))과 같은 인식번호는 한 레코드당 유일한 값으로 존재한다. 양서의 경우, LC나 OCLC 등의 copy cataloging으로 목록이 구축되는 경우 위와 같은 인식번호는 중복레코드 검증에 유용한 요소가 된다.

② ISSN은 연속간행물 표준번호로써 중복레코드 검증에 기초적인 역할을 한다. 하지만 ISSN은 변함이 없고, 서명이 변경되는 경우도 종종 존재하므로 이를 식별할 수 있는 체계가 필요하다.

③ 서명은 관사를 제외한 본서명, 부서명, 대등서명(245 $a, :$b, =$b)을 검증요소로 하되, 편제(245 $p)가 있는 경우 편제도 중복레코드를 검증하기 위한 요소로 사용되어야 한다. 연속간행물이 여러 편으로 나오는 경우, 편제는 중복여부의 중요한 요소가 된다.

④ 출판사는 동일한 서명의 연속간행물이라고 하여도 국가에 따라 출판되는 경우 중요한 요소로 이용될 수 있다. 특히 출판사명은 전거 통제가 이루어지지 않으면 중복검증의 방해요인이 되기도 한다.

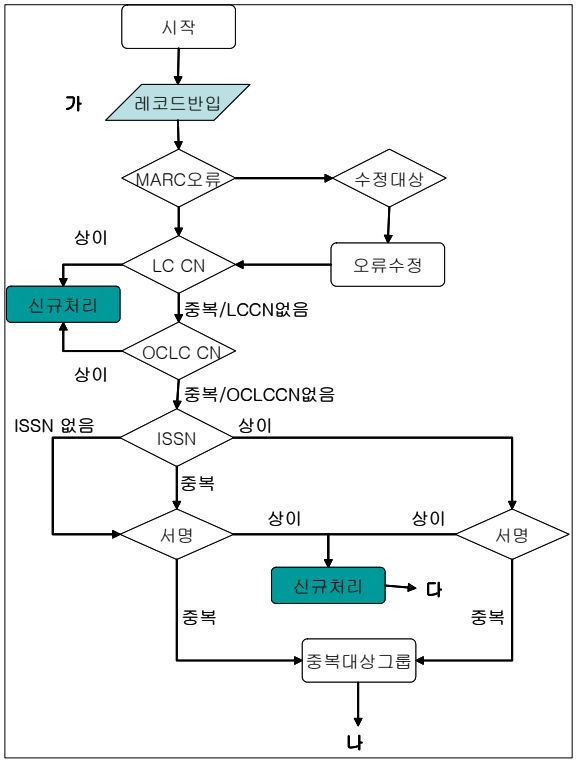
⑤ 연속간행물의 출판년, 폐간년은 매우 중요한 요소로 작용된다. 특히, 전후관계의 연속간행물의 경우, 780필드와 785필드의 지시기호를 잘 구분하여 명확한 관계를 지어야 중복레코드를 줄일 수 있을 것이다.

IV. 중복레코드 검증 프로세스 도출

연속간행물의 중복레코드 검증 요소들을 바탕으로 수작업을 최소화시킬 수 있는 검증프로세스를 그림 2., 그림 3과 같이 도출하였다.

먼저, 1단계에서는 서명과 식별기호를 중심으로 중복레코드를 색출하고, 2단계에서는 1단계에서 추출된 데이터를 대상으로 출판사항을 중심으로 중복레코드 검증을 한다.

1단계에서 검증요소는 크게 LCCN, OCLCCN, ISSN, 서명이다. 먼저, 대상 레코드의 010필드에서 기호와 스페이스를 모두 없애고 문자와 숫자만을 추출하여 비교한다.

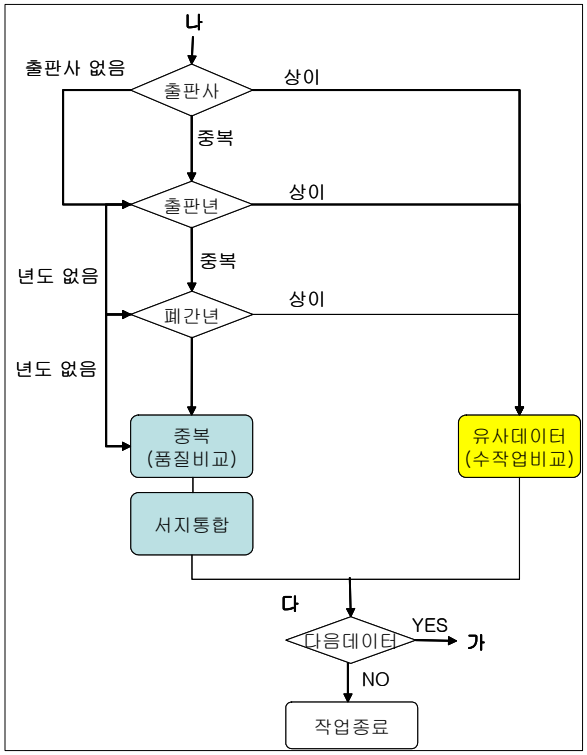


▶▶ 그림 2. 1단계 중복 데이터 검증 프로세스

중복되거나 LCCN 값이 없는 데이터를 대상으로 OCLCCN을 비교한다. OCLCCN은 035필드의 '(OCoLC)'값이 앞에 붙어 있는 경우만 추출해서 괄호값을 없애고 데이터 비교를 한다. 035필드는 기타 시스템 제어번호들이 기입되므로 OCLC 제어번호를 선별하는 과정이 중요하다. 역시 중복되거나 값이 없는 데이터를 대상으로 ISSN 비교를 한다. ISSN은 '-' 삭제하고 숫자만 비교하며, 중복이 된 데이터를 대상으로 서명을 비교한다. 서명은 245 필드의 $a, $b, $x에 기입된 데이터를 관사제거, 스페이스 제거, 접속사를 일치시켜서 추출한다. 위의 서브필드는 교차비교하며, 중복된 데이터 중 $p가 있는 경우는 $p끼리 서로 비교하도록 한다. 중복된 데이터는 1단계 중복대상 그룹이 되며, 상이한 데이터는 모두 신규 데이터로 구축한다. 1단계에서 중복된 데이터는 2단계 기준요소를 비교하여 일치하지 않더라도 기본적으로 유사데이터가 된다.

2단계의 검증요소는 출판사, 출판년, 폐간년이며, 1단계에서 중복된 데이터를 대상으로 출판사의 260 $b를 추출하여 비교한다. 이때 출판사는 전거 통제가 되어 있으면 유사데이터로 처리되는 부분이 감소되기 때문에 수작업이 훨씬 줄어들게 된다. 출판사 비교를 통해 중복된 데이터는 출판년도를 비교하게 된다. 출판년도는 008필드의 07-10를 추출하며 중복되지 않은 데이터는 모두 유사데이터로 처리하여 수작업으로 확인한다. 출판년도 비교작업에서 중복된 데이터는 마지막으로 폐간년도를 비교해야 한다. 폐간년도는 008필드의 11-14자리의 데이

터를 추출하며 역시 중복되지 않은 데이터는 유사데이터로 처리한다.



▶▶ 그림 3. 2단계 중복 데이터 검증 프로세스

유사데이터로 처리된 데이터는 비교요소들을 고려하여 불가피하게 수작업으로 비교한다. 중복 처리된 데이터는 품질비교를 통해 품질우위의 데이터를 기준으로 서지통합을 진행한다. 서지통합 후 다음데이터가 있으면 앞선 프로세스를 다시 반복하고 없으면 작업을 종료하게 된다.

V. 결론 및 제언

우리나라의 종합목록은 기 구축된 목록을 통합하는 형식을 취하고 있기 때문에 중복레코드를 선별하고 처리하는 과정이 매우 중요하다. 특히 데이터의 중복성은 데이터베이스 품질을 좌우하는 하나의 지표가 되므로 중복되지 않는 데이터를 제공함으로써 이용자에게 신뢰를 주어야 한다.

본 연구에서는 중요 학술정보 커뮤니케이션 도구인 연속간행물 정보를 통합, 공유하여 서비스하는 연속간행물 종합목록의 중복레코드를 최소화하기 위하여 7개의 중복데이터 검증요소를 추출하여 프로세스를 도출하였다. 7개의 요소는 검증순서대로 LCCN, OCLCCN 인식번호, ISSN 표준번호, 서명사항, 출판사, 출판년, 폐간년이며, 각 요소에 해당하는 데이터 값을 추출하기 위해서는 스페이스와 관사 삭제, 접속사가 일치되도록 데이터를 정제하는 작업이 필요하다. 또한 중복데이터를 검증하기 이전에 MARC 규칙의 여부와 오기가 있는지 검토, 분석하여 데이터 수정을 해야 한다.

중복데이터의 검증은 사전 작업이 많이 요구된다. 사전작업을 철저하고 정밀하게 해야만 사후 수작업 처리가 줄어들게 된다. 특히, 전거의 필요성은 이 과정에서도 매우 중요한 이슈가 되며 특히, 서명과 출판사처럼 다양하게 표현될 수 있는 중복 검증 요소는 전거 데이터를 확보하는 것이 중요한 작업이라고 할 수 있다.

앞으로 관련연구로 본 연구에서 제시한 중복레코드 검증 프로세스를 바탕으로 실제 DB에서 검색의 재현율과 정확율을 측정하고, 데이터 품질 개선에 얼마나 영향을 미치는지에 대한 확대 연구가 필요하며, 더 나아가 요소별 가중치 산출 및 가중치를 적용한 측정을 통한 연구도 필요할 것이다.

▌참 고 문 헌▐

[1] 김선애, 이수상. “KORIS-NET 종합목록 DB의 품질평가”, 한국문헌정보학회지, 제40권 1호, pp.95-117, 2006.
[2] 이제환, “공동목록 DB의 품질평가와 품질관리: KERIS의 종합목록 DB를 중심으로” 한국문헌정보학회지. 제38권 1호, pp.61-90, 2002.
[3] 조순영, “종합목록의 중복레코드 검증을 위한 알고리즘 연구”, 한국문헌정보학회지, 제37권 4호, pp.69-88, 2003.
[4] 최인숙, “디지털자료실지원센터 종합목록 데이터 품질평가 및 관리방안”, 한국문헌정보학회지, 제38권 3호, pp.119-139, 2004.
[5] E.T.O'Neil. "Duplication Detection", Annual Review of OCLC Research. 15-16, 1988-89.
[6] S.A.Cousins, "Duplicate Detection and Record Consolidation in Large Bibliographic Databases".Journal of Information Science, Vol.24 no.4, pp.231-40, 1998.