종합목록의 중복레코드 검증을 위한 알고리즘 연구*

A Study on Duplicate Detection Algorithm in Union Catalog

조 순 영(Sun-Yeong Cho)**

목 차

초 록

본 연구는 KERIS 종합목록의 품질 개선을 위하여 새로운 유형의 중복 데이터 색출 알고리즘을 개발한 것이다. 새로운 알고리즘에서는 현재 적용하고 있는 것과 같은 MARC 데이터 일치여부 비교 방식에서 탈피하여 언어별 서지 유형별 다른 비교방식을 적용하였다. 아울러 비교 요소간의 유사성을 측정하고, 각 요소의 중요도에 따라 가중치를 차등 부여하는 방식을 병행하였다. 새로 개발한 알고리즘의 효용성을 입증하기 위하여 최근 종합목록에 업로드된 데이터 210,000건을 추출하여 실험용 마스터 파일을 구축하고 7,649건을 두 개의 알고리즘으로 처리한 결과 새로운 알고리즘에서 중복레코드의 색출 비율이 36.2% 더 높게 나타났다.

ABSTRACTS

This study intends to develop a new duplicate detection algorithm to improve database quality. The new algorithm is developed to analyze by variables of language and bibliographic type, and it checks elements in bibliographic data, not just MARC fields. The algorithm computes the degree of similarity and the weight values to avoid possible elimination of records by simple input error. The study was performed on the 7,649 newly uploaded records during the last one year against the 210,000 sample master database. The findings show that the new algorithm has improved the duplicates recall rate by 36.2%.

키워드: 종합목록, 오류데이터, 중복데이터, 데이터 품질관리
Union Catalog, Duplicate Detection Algorithm, MARC

---

* 본 논문은 박사학위논문을 축약한것임.

** 한국교육학술정보원(KERIS) 학술연구정보화실장(chosy@keris.or.kr)

논문접수일자 2003년 11월 11일
게재확정일자 2003년 11월 26일

1. 서 론

미국 OCLC의 WorldCat이나 일본 NII의 NACSI-CAT과 같이 역사가 오랜 종합목록은 각 회원 도서관이 DB를 구축하기 전에 분담목록터미널을 이용하여 센터의 종합목록부터 검색하고 데이터가 없는 경우에는 신규 데이터를 입력하는 방식을 취하였다. 그러나 우리나라 KERIS의 종합목록은 대학에서 이미 구축를 일괄 통합하여 초기DB를 구축하고 신규데이터 입력도 회원 도서관 센터로 업로드하는 방식도 수용하고 있기 때문에 양적인 성과는 거두었지만 비 표준화된 데이터 및 오류데이터 등으로 인한 중복데이터가 다량 발생하였다. 따라서 표준화된 데이터를 대상으로 한 기존의 중복 알고리즘으로 모든 중복 데이터를 색출하는 것은 한계가 있기 때문에, 현재 중복 알고리즘에서는 모든 데이터를 기계적으로 처리하지 않고 중복 가능성이 있는 데이터를 별도로 구분하여 품질 검증 요원이 육안으로 식별하는 작업을 병행하고 있다. 그러나 업로드되는 데이터 양이 급속히 증가하면서 수작업 처리는 한계에 이르렀고 이 같은 상황을 반영한 새로운 개념의 알고리즘이 절실하게 필요하게 되었다. 이 같은 문제는 KERIS와같이 종합목록을 운영하는 센터 뿐 아니라 여러 개의 분관 데이터를 통합하는 대학도서관에서도 마찬가지이다. 따라서 본고에서는 중복데이터 발생의 원인이 되는 데이터의 오류 유형을 분석하고 그 결과를 반영한 한국형 알고리즘을 개발하여 실제 적용함으로서 중복데이터 검색의 효율성이 향상되는 것을 입증하고자 한다.

2. 선행연구

국내 종합목록은 그 역사가 짧기 때문에 구체적으로 데이터베이스의 중복 알고리즘에 관한 연구가 행해진 바는 없으나, 관련 연구에서 이제환[1]은 종합목록에서의 중복레코드 다량 발생으로 인한 데이터 품질 관리상의 문제점을 지적하고 있고, 이지은[2]은 중복 데이터 발생의 가장 큰 원인으로 사서들의 목록시스템 사용 미숙과 데이터의 입력오류, 또한 총서명, 주제명, 부출서명 등 필수 데이터에 대한 입력기준이 미비한 점을 지적하고 있다. 반면 김지훈[3]은 서지 레코드의 오류를 원문헌, 목록기술, 주제 관련, 편집상의 원인으로 지적하면서 그 중 목록기술상의 오류는 전문적인 지식을 요구하는 기술목록규칙과 MARC에 대한 이해 부족이 주 원인이라고 밝히고 있다.

한편 해외에서는 종합목록의 역사가 오랜만큼 이미 종합목록의 중복레코드에 관하여 많은 연구가 진행되었고 대부분의 연구는 복수의 데이터베이스를 대상으로 데이터를 무작위

1) 이제환, KERIS 서지 데이터베이스의 품질관리를 위한 평가 모델 개발 및 개선방안 수립, 서울: 한국교육학술정보원, 2001.
2) 이지은, 첨단학술정보센터 종합목록데이터베이스 품질관리에 관한 연구, 석사학위논문, 숙명여자대학교 : 문헌정보학과, 1999.
3) 김지훈, 서지데이터베이스의 품질관리-k관의 MARC 레코드 분석을 중심으로, 도서관학논집 제21집 (1994) : 401-429.

추출하여 데이터의 품질 수준을 비교하거나, 중복 또는 오류 알고리즘에 의해 추출된 데이터를 육안으로 판별함으로서 그 원인을 분석하는 방법을 택하고 있다. 많은 연구에서 데이터 오류율을 철자 오기입의 문제와 MARC의 입력 표준 사용의 적절성을 기준으로 측정하고, 중복알고리즘에 대해서는 필수 데이터 필드를 중심으로 비교하고 있다. 그리고 데이터의 질적 향상을 위해 정확성을 최대화하고, 중복성을 최소화하는 방안으로서 입력자의 철저한 교육, 제도적인 업무흐름의 개선 등 수동적인 방법을 제안하고 있다. 대표적인 연구자로는 O'Neill[4]을 중심으로 한 OCLC의 시스템 분석팀을 들 수 있다. 그들은 중복레코드 발생의 가장 큰 원인으로 입력 오류, 잘못된 MARC 필드사용, 생략된 정보, 가변장과 고정장 필드간의 비 일치 등이라고 지적하고 오류 레코드 때문에 중복레코드가 검색되지 않는 점을 강조하고 있다. 그들은 WorldCat 데이터에서 실제 중복임에도 불구하고 28% 이상이 두개 이상의 요소를 다르게 기입하고, 한개의 요소가 다르게 기입된 경우도 26%라고 지적하고 있다. 또한 O'Neill[5]은 기존의 중복알고리즘으로 찾아내지 못한 레코드를 대상으로 클러스터링과 평가알고리즘을 적용한 새로운 자동중복처리 프로그램을 개발하여 전체 중복의 1/3을 줄이는 결과를 얻었다고 발표한 바 있다. Ridley[6]는 서지 데이터베이스의 질적 관리와 중복제거를 위한 전문가시스템에 관한 논문에서 비교대상 레코드 간에 발생할 수 있는 다양한 경우의 조건을 모두 제시하고 중복레코드 색출 작업을 단계적으로 처리하였다. 그 외 Cousins[7]는 470만 서지레코드를 보유하고 있는 영국 CURL의 종합목록 COPAC를 다른 데이터 임에도 불구하고 중복레코드로 잘못 검증된 경우와 중복레코드 임에도 불구하고 신규레코드로 처리되는 경우에 대해 연구한바 있다.

3. 연구방법

새로운 알고리즘의 효율성 입증을 위한 테스트 베드로 KERIS 종합목록에 최근 1년간 업로드된 21만건(KORMARC 137,000, MARC21 74,000)의 단행본 레코드를 마스터 데이터베이스로 구축하고 2일간 회원 도서관에서 작성한 7,649건의 레코드를 별도의 업로드용 파일로 생성하였다. 동일한 실험용 화일을 기존의 알고리즘과 신규 알고리즘으로 각각 업로드 한후 처리 결과를 분석함으로서 다음의 네가지 가설에 대해 비교 평가하였다.

평가 1) 중복레코드 검증알고리즘에서 데이터의 언어속성이 반영된 별도의 프로그램이 적용될 경우 중복레코

4) E. T. O'Neill, "Duplicate Detection," Annual Review of OCLC Research (1988/89) : 15-16.
5) Ibid.., 12-13.
6) M. J. Ridley, "An Expert System for Quality Control and Duplicate Detection in Bibliographic Databases," Program 26 : 1 (1992) : 1-18.
7) S. A. Cousins, "Duplicate Detection and Record Consolidation in Large Bibliographic Databases: the COPAC Database Experience in Great Britain," Journal of Information Science 24: 4 (1998) : 231-40.

드 검색의 재현율은 높아지는가?

평가 2) 학위논문이나 번역서와 같이 원자료의 서지적 특성이 다른 경우 중복알고리즘을 다양화 한다면 중복레코드 검색의 재현율은 높아지는가?

평가 3) 오류데이터가 많은 데이터베이스의 경우 완전일치 방식으로 데이터를 비교하지 않고, 비교요소의 유사도를 값으로 측정하고, 각 요소별로 가중치를 달리 주어 중복값을 부여하면 중복레코드 검색의 재현율은 높아지는가?

평가 4) 중복레코드 검증 알고리즘에서 MARC 데이터의 필드별 비교 방식을 요소별 비교방식으로 바꿀 경우 중복레코드 검색의 재현율은 높아지는가?

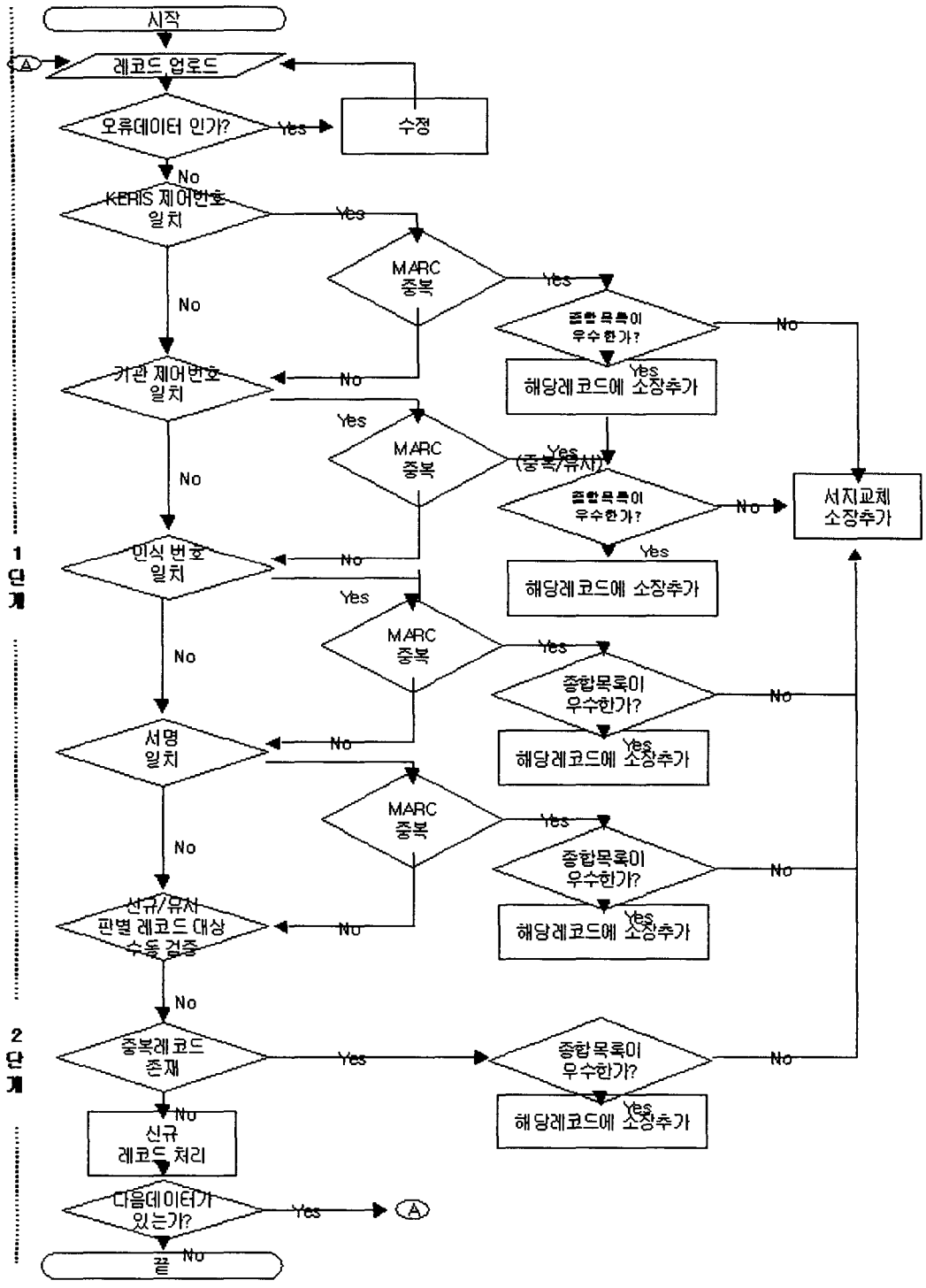
평가를 위하여 신규레코드를 종합목록 데이터베이스에 업로드할 때 전체 중복레코드 중 자동 알고리즘에 의해 걸러지는 중복레코드의 비율 즉 재현율을 평가요소로 택하고, 2년 이상 검색 경험자가 수작업한 중복 테스트 결과를 기준값으로 설정하여 비교한다. 그러나 중복레코드의 기준과 알고리즘의 적용 단계에 따라 알고리즘의 효율성이 달라질 수 있고, 〈표 1〉에서와 같이 데이터의 처리방식, 구분방식, 중복값 산출방식, 판정방식, 데이터 추출방식 등에 따라 재현율이 영향을 받을 수 있기 때문에 신규 프로그램에서는 중복 여부를 가부로 결정하지 않고 중복의 정도에 대한 값을 수치화하고 중복레코드의 정의에 따라 운영자가 한계값을 조정함으로서 중복레코드에 대한 재현율의 수준을 선택할 수 있도록 하였다.

4. 현행 중복 알고리즘

KERIS의 중복레코드 처리과정은 〈그림 1〉과 같고 전체 과정은 크게 두 단계로 구분된다. 1 단계에서는 고유번호인 센터의 제어번호, 도서관의 제어번호, ISBN, ISSN, LC 제어번호와 같은 인식번호를 우선적으로 비교하고, 빠른 비교를 위한 방법으로 각 레코드의 245 필드에서 서명의 앞부분 20자리 만을 추출하여 마스터 데이터베이스에서 일치하는 중복대상 레코드 파일을 생성한다. 그 다음 2 단계에서는 서명, 저자, 발행자, 발행년, 면수, 판사항, 총서사항, 인식번호 등 8개 항목에 대해 문자열 비교를 실시하여 중복 여부를 가린다.

〈표 1〉 독립요인과 종속요인

독립요인	통제요인	종속요인
- 데이터 처리 방식: 언어구분 유무 - 데이터 구분 방식: 서지 유형별 구분 - 중복값 산출 방식: 유사값 및 가중치 부여 - 중복 판정 방식: 조건비교 -〉 중복값 - 데이터 추출 방식: 필드별 -〉요소별	→ 한계값 수준 →	검색의 재현율



〈그림 1〉 현행 중복 알고리즘

중복 가능성이 있는 8개 항목에 대하여 비교한 결과 다음 조건 중 하나를 만족시키는 경우 중복 가능성이 있는 유사 데이터그룹으로 분류한다.

① ISBN/ISSN/LCCN과 같은 인식번호의 개수가 같고 모두 일치하며, 발행년이 일치하는 경우.
② 245 필드 ǂa 서명의 전방 또는 후방이 일치하고, 면장수가 모두 일치하고, 인식번호가 한쪽만 존재하는 경우.
③ 245 필드 ǂa 서명이 완전하게 일치하지 않아도 80% 이상 일치하고, 발행자가 전방 또는 후방이 일치하고, 면장수가 각각 존재하며 일치하는 경우.
④ 245 필드 ǂb 또는 740 필드 ǂa 서명이 완전하게 일치하고, 발행년은 하나만 존재하고, 면장수가 모두 일치하고, 인식번호가 한쪽만 존재하는 경우.
⑤ 245 필드 ǂa 서명이 완전하게 일치하고, 발행자의 전방 또는 후방이 일치하는 경우.
⑥ 245 필드 ǂa 서명이 완전하게 일치하고, 245 필드 이외의 첫 번째 저자사항이 일치하는 경우.

이상의 조건을 만족 시키는 데이터를 대상으로 다시 다음의 조건을 추가 비교하여 다섯 가지 조건 중 하나라도 만족시키면 중복레코드로 처리한다. 여기서의 조건은 최소한의 수준을 나타내는 것으로 8개 요소를 비교할 때 각 요소가 모두 해당 조건 이상이 되어야 한다. 유사 그룹이지만 이들 조건에 부합되지 않는 데이터는 중복데이터 검증요원이 각각 육안으로 판별하는 과정을 거친다.

① 필드 245 ǂa 본서명의 전방 또는 후방이 일치하고, 245 필드의 저자는 일치하지 않으나 기타 필드의 저자 중 하나가 완전하게 일치하며, 발행자가 완전하게 일치하지는 않으나 발행자의 전방 또는 후방이 일치하고, 발행년이 완전하게 일치하고, ISBN/ISSN/LCCN과 같은 인식번호가 모두 일치하는 경우
② 245 필드 ǂb 또는 740 필드 ǂa 서명이 완전히 일치하고, 245 필드의 저자가 일치하고, 발행자, 면장수가 일치하고, 판사항이 두 레코드 모두 존재하지 않고, 인식번호는 한쪽만 존재하는 경우이다.
③ 245 필드 ǂa 서명이 일치하고, 245 필드의 저자는 일치하지 않으나 기타 필드의 저자 중 하나가 일치하고, 발행자의 전방 또는 후방이 일치하고, 면장수가 일치하고, 판사항이 두 레코드 모두 존재하지 않고, 인식번호와 권차는 한쪽만 존재하는 경우이다.
④ 245 필드 ǂa 서명 및 저자가 일치하고, 발행자의 전방 또는 후방이 일치하고, 발행년, 면장수가 모두 일치하고, 인식번호는 한쪽만 존재하는 경우이다.
⑤ 245 필드 ǂa 서명이 완전하게 일치하고, 245 필드의 저자가 일치하고, 발행자가 완전하게 일치하고, 면장수가 모두 일치하고, 판사항이 두 레코드 모두 존재하지 않고, 인식번호는 한쪽만 존재하는 경우이다.

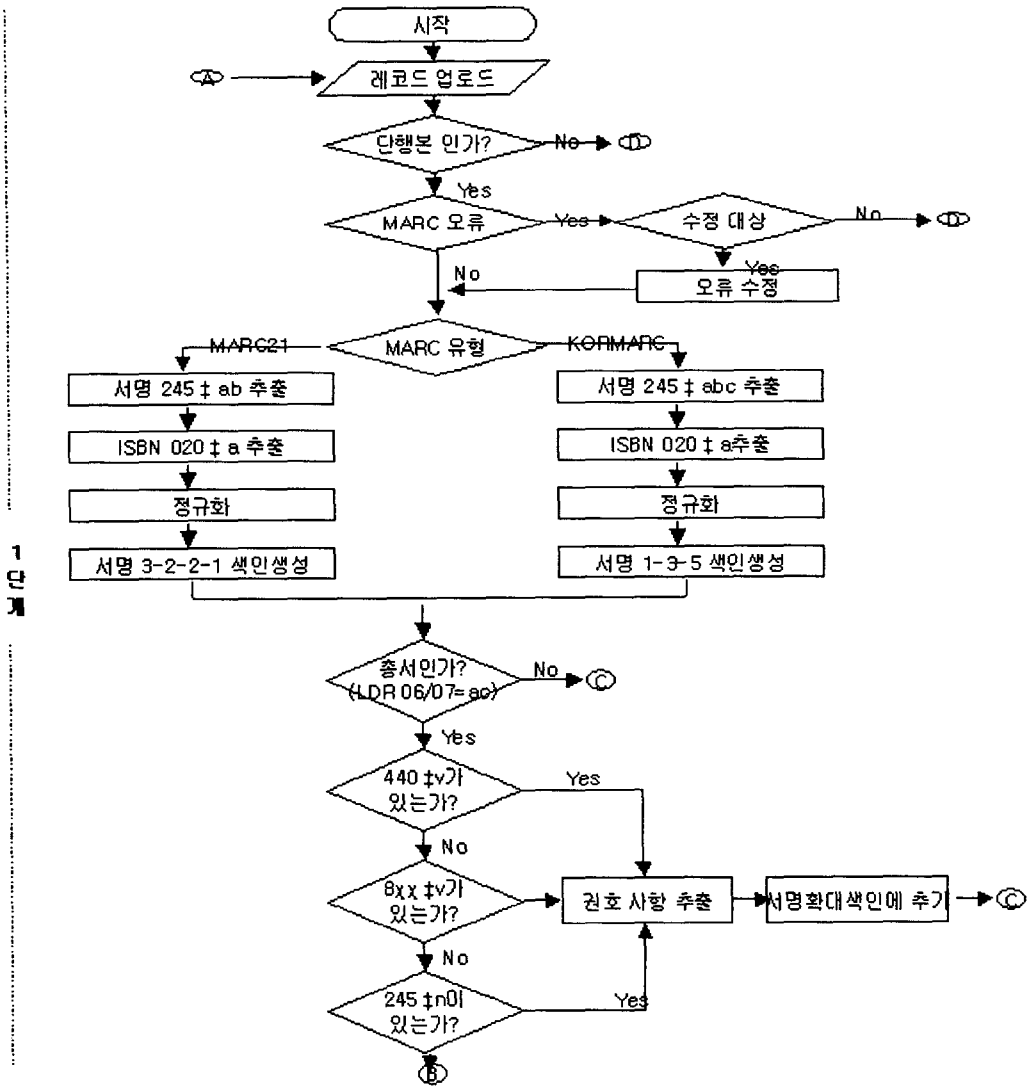
5. 신규중복 알고리즘

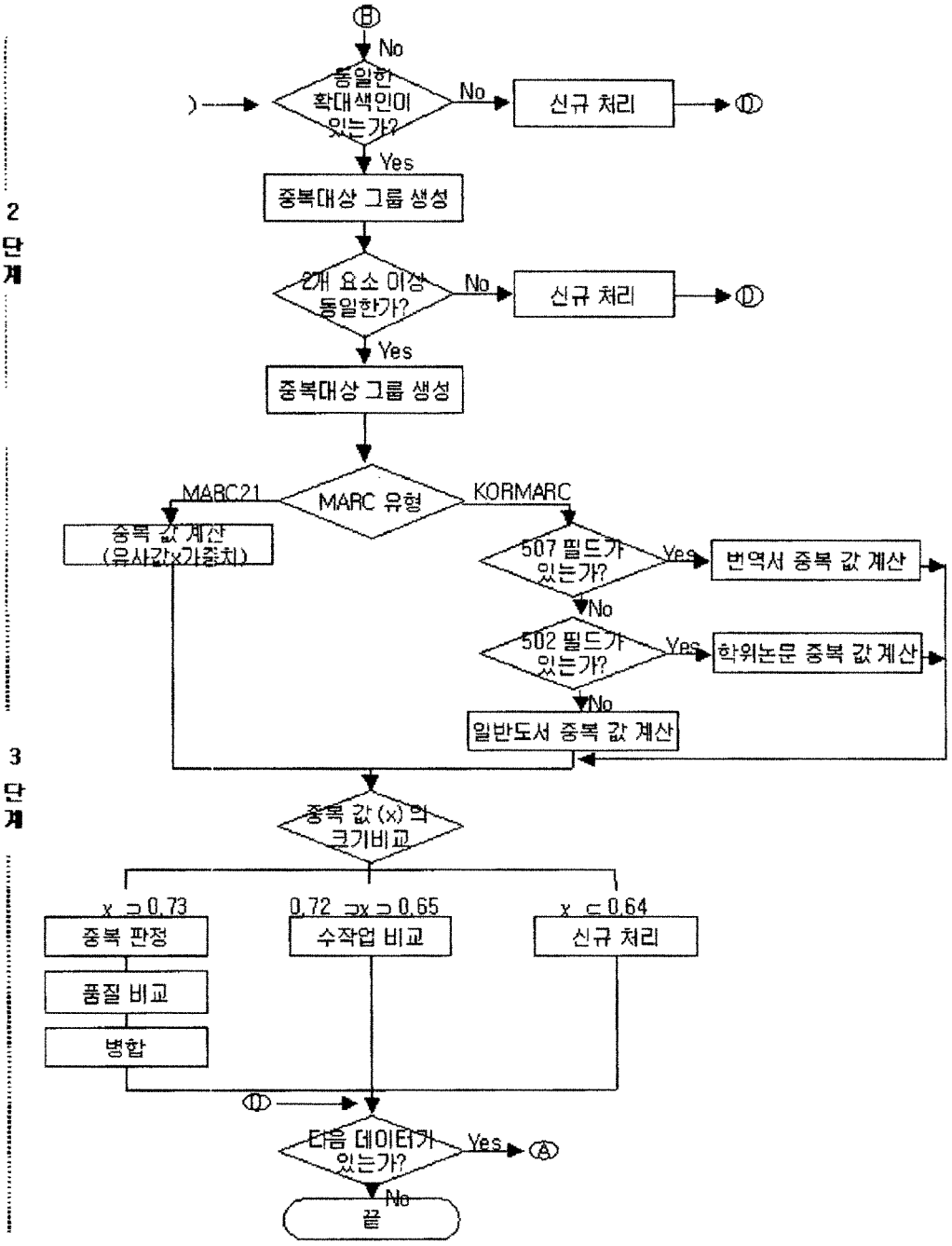
5. 1 프로그램 개발 환경

신규 알고리즘을 적용한 프로그램은 WINDOWS 2000 Professional 운용체제하에서 개발하고 언어는 Visual C++과 Informix-esqlc를 사용하였다. Unix에서도 호환가능하며 테스트는 Sun Enterprise 5000 서버에서 실시하고 관계형 데이터베이스 관리시스템으로는 Informix 7.24를 사용하였다.

5. 2 중복대상 그룹생성

신규 알고리즘의 전체 과정은 〈그림 2〉과 같이 3 단계로 구분할 수 있다. 1 단계에서는





〈그림 2〉 신규 중복 알고리즘

서명을 중심으로 중복 가능성이 있는 데이터를 모두 색출한다. 2 단계에서는 1 단계에서 색출된 레코드를 대상으로 8개의 요 정수값으로 확장 색인을 생성한다

서는 최종적으로 중복인지의 여부를 판정하기 위해 12개 요소에 대해 상세한 하위 필드까지 레코드 쌍간의 비교 과정을 거친다.

5. 2. 1 1단계 중복대상 색인 생성

1 단계에서는 중복 가능성이 있는 데이터를 최대한 많이 추출하기 위하여 서명 필드를 대상으로 간략한 색인을 생성한다. MARC21 데이터는 245 필드의 ǂa 본서명, ǂb 부서명을 대상으로 색인을 생성하고, KORMARC 데이터는 ǂc 잡제까지도 포함한다. KORMARC 데이터의 경우 모든 단어를 붙인 후 1-3-5자를 추출한다. 예를 들어 '學生의 敎師에 대한 期待와 自己期待' 라는 서명에서는 '학의사'를 추출한다. MARC21 데이터는 불용어를 포함한 서명을 대상으로 첫 단어에서부터 차례대로 3-2-2-1자를 추출한다. 예를 들어 'Introduction to computer based library system' 이라는 서명에서는 'inttocob' 라는 색인을 생성한다. 정규화 규칙에 의해 서명을 기본 키로 추출하되 서명의 치명적 오류나 대등서명의 문제로 1 단계 대상에서부터 누락되는 것을 막기 위하여 ISBN까지 추출하여 1 단계 색인에 포함시킨다.

5. 2. 2 2 단계 중복대상 확장 색인 생성

2 단계에서는 1 단계에서 중복 대상으로 선정된 레코드를 대상으로 〈표 2〉의 확장 색인 대상 필드의 여덟개 요소를 숫자 값으로 치환하여 비교한다.

- 제어번호 : LC제어번호의 경우 빈칸을 모두 생략하여 붙이고, 숫자 데이터 앞의 문자는 삭제하여 비교한다.
- 발행년도 : 발행년은 008 고정장 필드의 07/10에서 추출한다. 코드 값에 따라 판권년, 복수년, 추정년, 복간년 등 다양한 경우가 있으나 자료상에서 명확하게 구분될 수 있는 발행년 1 만으로 비교한다.
- 면수 : 형태사항 300의 ǂa 면장수 표시는 단순한 사항이지만 기입의 차이가 많은 요소이기 때문에 ǂa에서 가장 큰 숫자 네자리까지 만 추출한다. 즉 ǂa ix, 40 p., ǂa a[40]면인 경우에는 두 레

〈표 2〉 2단계 확장색인 추출요소

요 소	KORMARC	MARC21	형태
KERIS, 제어번호	035 ǂa	035 ǂa	b
LC 제어번호	·	010 ǂa	b
I S B N	020 ǂa	020 ǂa	b
발행년도	008 07/10	008 07/10	b
면 수	300 ǂa	300 ǂa	b
서 명	245 ǂa,b,c	245 ǂa,b	b(ui)
저 자	(100 ; 111 ; 700 ; 711) ǂa,	(100 ; 111 ; 700 ;711) ǂa,	b(ui)
	(110 ; 710) ǂa,b	(110 ; 710) ǂa,b	
발 행 자	260 ǂb, 502 ǂb	260 ǂb	b(ui)

b: binary, ui: unsigned integer

코드에서 모두 '40' 만 추출한다.

- 서명 : 서명을 구성하는 문자의 ASCII 값을 모두 더하여 정수 값을 부여한다. 예를 들어 서명 '현대사회복지실천이론' 이란 서명의 ASCII 값은 현(199+246)+대(180+235)+사(187+231)+회(200+184)+복(186+185)+지(193+246)+실(189+199)+천(195+181)+이(192+204)+론(183+208)=4023이다. 그러나 ISBN과 LC 제어번호는 정수값 최대 크기인 65497을 넘을 수도 있기 때문에 아래의 공식으로 나눈 몫에서 그 나머지 값에 +1을 더한 값을 부여한다.

ASCII값 합계 / 65497(양의 정수값 :2 byte 기준) = x ... a(나머지 값)+1

- 저자 : 저자명은 전거통제를 받지 않는 경우 동일인명의 표기가 도서관마다 다양하기 때문에 중복 비교가 어려운 요소이다. 따라서 이형 표기의 확률이 높은 번역서의 경우 507 ǂa 원저자명으로 비교하여 일치율을 높인다.
- 발행자 : 발행자명은 260 필드 ǂb에서 추출하여 정규화 규칙을 적용한다. 발행자가 둘 이상인 경우 구두점 ' : ' 앞부분의 첫 번째 발행자 명만 추출한다.

KORMARC 데이터 중 학위논문은 발행자를 [s.n] 불명으로 표기한 사례가 많기 때문에 학위논문주기 502 필드의 ǂb에서 대학명을 앞에서 5문자 즉 10 byte만 대신 추출하는 것이 바람직하다. 서명은 1 단계 색인생성에서의 정규화 규칙을 그대로 적용하고 서명 필드 전체를 대상으로 하여 정수 값을 부여한다.

이상의 방법으로 하나의 서지레코드 당 하나의 확장 색인 레코드를 생성하여 2 단계 비교를 한다.

5. 2. 3 3 단계 중복대상 레코드 비교

2 단계에서 중복 대상으로 선정된 레코드를 대상으로 각기 쌍을 이루어 다시 〈표 3〉의 12개 요소를 대상으로 MARC 데이터를 추출한다. 요소라 함은 발행년도를 비교할 때 008 07/10과 260 ǂc, 학위논문 수여기관 필드 502 ǂb와 008의 26/27 한국대학부호와 같이 관련된 필드를 동시에 비교하는 것을 의미한다.

1) 비교 요소 :

2 단계 확장 색인에서 비교대상이 되었던 필드 이외에 언어, 책임표시사항, 판사항, 크기, 총서명 등 다섯개 요소를 추가 하고 정규화규칙을 확장하여 텍스트 데이터를 비교한다.

① LCCN은 번역서와 MARC21데이터에 대해서만 비교하고, ISBN은 학위논문의 경우 비교 요소에서 제외시킨다.

② 언어 : 언어는 번역서를 비교할 수 있는 중요한 값으로서 고정장의 008/35-37과 041의 ǂa,h를 완전일치 방식으로 비교한다.

③ 저자 : 단체명, 회의명을 포함한 저자는 목록규칙 및 전거통제 여부에 따라 다양하게 표기될 수 있기 때문에 동일인 임에도 불구하고 일치하지 않는 사례가 발생 할 수 있다. 따라서 KORMARC은 필드 245 ǂd, MARC21은

〈표 3〉 중복 대상 레코드의 비교 요소

요소명		KORMARC			MARC21
		일반도서	학위논문	번역서	
1	LCCN	x	x	010 ǂa	010 ǂa
2	ISBN	020 ǂa	x	020 ǂa	020 ǂa
3	언 어	x	x	008 /35-37 041 ǂa,h	x
4	저 자	100,111,700,711 ǂa 110, 710 ǂa,b	100, 700 ǂa	100,111,507,700, 711 ǂa 110, 710 ǂa,b	100,111,700,711 ǂa 110, 710 ǂa,b
5	서 명	130,240,730,740 ǂa,n,p 245 ǂa,b,n,p 505 ǂt	245 ǂa,b 740 ǂa	130,240,730,740 ǂa,n,p 245 ǂa,b,n,p 507 ǂt	130,240,730,740 ǂa,n,p 245 ǂa,b,n,p 505 ǂt
6	책임표시사항	245 ǂd	x	x	245 ǂc,e
7	판사항	250 ǂa	x	250 ǂa	250 ǂa,b
8	발행자	260 ǂb	502 ǂb, 008 26/27	260 ǂb	260 ǂb
9	발행년도	008 07/10, 260 ǂc,g	008 07/10 260 ǂc, 502 ǂd	008 07/10 260 ǂc,g	008 07/10 260 ǂc,g
10	면 수	300 ǂa	300 ǂa	300 ǂa	300 ǂa
11	크 기	300 ǂc	x	300 ǂc	300 ǂc
12	총서명	440 ǂ, 830 ǂall	x	440 ǂ, 830 ǂall	440, 830 ǂall 800-811 ǂa,b,t,v

필드 245 ǂc 전체를 문자열방식으로 비교하여 일치하는 부분에 대해 유사값을 부여한다.

④ 서명 : 2 단계 정규화 규칙을 적용하고, 추가로 한자 서명의 경우 동자이음어에 대해서는 복수로 색인을 생성한다. 또한 본서명과 n:n으로 비교 대상이 되는 총서명의 경우 MARC21에서 AACR2 이전의 800-811도 비교 대상에 포함시킨다.

⑤ 책임표시사항 : 책임표시사항은 자료의 표제면에 있는 그대로 기입하는 것을 원칙으로 하기 때문에 기입의 오차는 없지만 기술순서가 다를 수 있다. 따라서 두 개의 레코드 전체를 문자열 비교방식으로 하여 유사값을 측정한다.

⑥ 판사항 : 판사항은 전체 데이터의 4%에 기입되어 있는 중요한 정보이나 인쇄 년도와 명확히 구별되지 않고 기입형식 또한 다양하기 때문에 완전일치 방식으로 비교한다.

⑦ 발행자 : 발행자의 경우 MARC21 데이터는 260 ǂb를 하나의 문자열로 연결해서 앞에서부터 여섯자리까지만 비교한다. KORMARC 데이터에서는 ǂa의 앞에서 4문자까지 비교한다. 단, 학위논문의 경우 발행자를 대학명과 [s.n.]으로 사용한 기관이 많기 때문에 발행기관을 ǂb 수여기관명으로 대체 비교한다. 아울러 학위논문은 대학간행물로서

고정장 008/26-27에 기입하는 대학명 코드 값과 비교한다.

⑧ 발행년도 : 260 ǂc의 발행년 사항을 추가로 포함시키되, 발행년의 경우 복간년, 배포년 등 확실치 않은 년도로 인한 중복의 누락이 많기 때문에 -1 ≦ x ≦ 1의 범위에 드는 경우 동일한 것으로 인정한다. 특별히 학위논문의 경우 학위 수여년도와 논문의 발행년도에 대한 혼동으로 1년의 차이가 많이 나기 때문에 범위 안에 드는 것은 모두 동일 본으로 간주한다.

⑨ 면수 : 300 필드 자체가 KORMARC데이터에서는 반복불가이나 MARC21에서는 반복사용 가능이므로 둘 중에 하나만 일치하면 일치 하는 것으로 한다.

⑩ 크기 : 자료의 크기는 필드 300 ǂc 세로의 길이만 비교하는 것을 원칙으로 하고 문자 앞의 숫자 값만 추출한다. 크기가 다양한 다권본의 비교대상에서 제외시킨다. 형태사항 기술 규칙에서는 소수점이 있는 경우 무조건 올림으로 기술하도록 되어 있으나 실제 크기가 일정한 학위논문의 사례에서는 3%의 오차가 있는 것을 고려하여 -1 ≦ x ≦ 1 범위에서의 허용오차를 허락한다.

⑪ 총서명 : MARC21 데이터 중 필드 8XX는 1XX와 동일한 형식이지만 AACR 2 이전의 데이터에서는 총서명에서 비교한다.

2) 비교방식

① 복수비교

정규화 된 요소를 대상으로 〈표 4〉에서와 같은 방식으로 비교한다. 여기서 n:n 비교라 함은 대상 레코드의 요소에 포함되는 필드를 모두 복수로 비교한 후, 가장 큰 유사값을 택하여 다른 요소의 유사값과 합한다. 이 방식은 비교 횟수와 대상의 종류를 최대한 다양하게 하기 때문에 결과의 정확성을 높일 수 있고 목록 기술의 차이로 중복에서 누락되는 레코드 수를 줄일 수 있다.

〈표 4〉 요소별 비교 방식

요소명	비교	요소비교	비교방식
LCCN	1:1		완전일치
ISBN	n:n		완전일치 (check digit 계산후)
언 어	1:1	○(번역서)	완전일치(ǂa,h.의 각 3 byte만 비교)
저자, 서명	n:n	○	유사값 산출
책임표시사항	1:1	○	유사값 산출
판사항	1:1		완전일치(KORMARC: 문자 MARC21: 숫자)
발행자	n:n	○(학위논문)	유사값(KORMARC: 8 , MARC21: 6 byte)
			학위논문 10 byte
발행년도	n:n	○	완전일치 · 허용오차 -1≦x≦1적용
면 수	1:1		완전일치 · 허용오차 -3≦x≦3적용
크 기	1:1		완전일치 · 허용오차 -1≦x≦1적용
총서명	n:n		유사값(4xx, 830: all, 800-811: ǂabtv)

② 유사값 비교

중복 대상으로 구분된 레코드 쌍은 각 요소별로 비교하되 완전일치 방식이 아니고 유사값을 측정하는 방식을 취한다. 유사값의 측정은 아래의 공식으로 계산한다.

유사값(S) = b/a + c/a * (a-b) / a

(a : 정규화 한 문자열 길이에서 가장 긴 값 , b : 앞에서부터 완전일치 하는 문자의 개수, c : b 이후 a 끝까지의 길이에서 일치하는 문자 수)

③ 가중치 비교

요소별 가중치는 〈표 5〉와 같이 고유성이 높은 요소일수록 높은 값을 부여한다. 서명은 가장 중요한 요소임에도 불구하고 텍스트의 길이가 길고 기입 내용의 표기 형식 또한 다양하기 때문에 상대적으로 낮은 값을 부여하였다. 반면에 ISBN과 같은 고유 인식번호에 높은 값을 부여하였다. ISBN은 고유한 번호일 뿐만 아니라 공식에 따라 체크 디지트를 확인하기 때문에 오류 입력의 확률이 거의 없다고 할 수 있다. 또한 가중치 부여로 중복 검색의 정확성을 높이기 위해 동일한 요소 내에서도 일치하는 경우에 따라 다른 값을 부여하였다.

④ 중복값 비교

위와 같이 부여된 가중치와 유사값 즉 데이터의 일치 정도를 기준으로 중복의 정도를 나타내는 중복값을 다음의 공식에 의하여 산출한다.

중복값(D) = SUM [요소의 가중치(W) x 유사값(S)] / SUM (각 요소의 가중치)

중복레코드를 비교하는 과정에서는 1:n으로 유사 레코드가 존재할 수있다. 즉, 2 단계 확장 색인에서 업로드 대상인 레코드와 마스터 파일의 여러개 레코드가 정수 값이 일치하면 각기 쌍을 이루어 연속적으로 중복값을 산출한다. 중복 값은 0에서 1사이에 분포하게 되고 1에 가까울수록 중복의 확률이 높은 것이다. 중복대상이 되는 모든 레코드의 중복값이 산출되면 그 값을 근거로 중복 여부를 판정하는 한계값을 정하게 된다. 여기서 한계값은 수작업상에서 중복으로 판정된 레코드의 값을 근거로 정하게 되나, 이값은 자료의 유형 또는 중복레코드의 기준, 유사레코드 구분 여부에 따라 달라질 수 있다. OCLC나 NII처럼 유사 레코드를 구분하지 않는다면 한계값의 설정이 절대적인 판단 기준이 되기 때문에 일부 중복레코드가 포함되더라도 재현율이 100%가 넘는 수준에서 설정을 하게 되고, 현재 KERIS처럼 데이터 검증 단계가 있다면 정확율이 100%가 되는 값을 기준으로 설정하게 될 것이다.

6. 데이터 처리 결과

신규 알고리즘 처리 결과를 평가하기 위해서는 중복 레코드로 판정할 수 있는 한계값을 설정해야 하기 때문에 전체 업로드 레코드를 대상으로 중복값을 계산한 것이 〈표6-1〉이다. 이 표에서 수작업에서 중복 레코드로 판정된

〈표 5〉 요소별 가중치

요소명	가 중 치	
LCCN	- 두 레코드의 값이 일치하는 경우 : 3.0	
ISBN	- 두 레코드의 값이 일치하는 경우 : 5.0	
언 어 (008/35-37, 041 ǂa)	- 두 레코드 모두 두 요소의 값이 일치하는 경우 : 1.0 - 한 레코드에서만 두 요소의 값이 일치하는 경우 : 0.75 - 두 레코드 모두 두 요소의 값이 일치하지 않는 경우 : 0.5	
저 자	- 두 레코드의 값이 일치하는 경우 : 1.5	
서 명	- 두 레코드의 값이 일치하는 경우 : 2.0	
책임표시	- 두 레코드의 값이 일치하는 경우 : 1.0	
판사항	- 숫자 데이터가 일치하는 경우 : 1.0	
발행자 (008/26-27, 502 ǂb)	학위 논문	- 두 레코드 모두 두 요소의 값이 일치하는 경우 : 3.0 - 한 레코드에서만 두 요소의 값이 일치하는 경우: 2.5 - 두 레코드 모두 두 요소의 값이 일치하지 않는 경우 : 2.0
	기타	- 두 레코드의 값이 일치하는 경우 : 2.0
발행년도 (008/07-10, 260 ǂc)	- 두 레코드 모두 두 요소의 값이 일치하는 경우 : 2.0 - 한 레코드에서만 두 요소의 값이 일치하는 경우 : 1.5 - 두 레코드 모두 두 요소의 값이 일치하지 않는 경우 : 1.0	
면 수	- 두 레코드의 값이 일치하는 경우 : 2.0	
크 기	- 두 레코드의 값이 일치하는 경우 : 1.0	
총서명	- 두 레코드 모두 ǂv가 있는 경우 : 1.5 - 하나의 레코드 만 ǂv가 있는 경우 : 1.0 - 두 레코드 모두 ǂv가 없는 경우 : 0.75	

3,275레코드에 대한 중복값의 분포도를 근거로 재현율이 100%이고 정확율이 97.28%인 중복값 0.73을 중복레코드 판정의 한계값으로 설정하였다. 〈표 6〉에 의하면 3,275번째 레코드가 포함되는 중복값 0.72 이상인 경우가 중복대상이 되어야 하지만 실제 수작업에서 중복레코드로 판정된 모든 레코드가 이 범위의 중복값을 갖게 되는지에 대해서는 모든 레코드를 하나씩 비교해야만 알 수 있고, 그 결과를 근거로 정확율 및 저현율을 산출해서 최적정 수준의 한계값을 정하게 된다. 따라서 그 한계값 이내에 수작업 결과에서의 중복레코드가 몇 퍼센트 포함되는가 즉 재현율이 얼마인가 또한 그 한계값 이내의 레코드 중에서 중복이 아닌 레코드가 얼마나 포함되어 있는가가 신규 프로그램의 효율성을 측정할 수 있는 척도가 될 것이다.

6. 1 현행 알고리즘 처리 결과

현재 운용중인 KERIS 알고리즘으로 7,649건을 업로드한 결과 〈표 7〉와 같이 신규로 판정된 37.90%의 레코드와 중복으로 판정된 26.11%의 레코드를 합한 64.01%의 레코드만이 수작업으로 처리한 결과와 일치한다. 그 외 중복임에도 신규로 처리된 레코드는 KORMARC

〈표 6〉 실험데이터의 중복값 분포도

중 복 값	레코드건수	누 계	누진율	정확율	재현율
0.77~1	2,74263	2,742	37.50%	100%	83.72%
0.76	99	2,841	38.85%	100%	86.74%
0.75	135	2,976	40.70%	100%	90.87%
0.74	53	3,029	41.43%	100%	92.48%
0.73	157	3,186	43.57%	100%	97.28%
0.72	95	3,281	44.87%	99.81%	97.52%
0.71	24	3,305	45.20%	99.09%	97.74%
0.70	5	3,310	45.27%	98.92%	97.98%
0.69	20	3,330	45.54%	98.34%	98.44%
0.68	31	3,361	45.97%	97.43%	98.77%
0.67	21	3,382	46.25%	96.82%	99.05%
0.66	21	3,403	46.54%	96.23%	99.57%
0.65	80	3,483	47.64%	94.01%	≒100%
0.64-0	3828	7,311	100%	52.00%	≒100%

〈표 7〉 KERIS 알고리즘과 수작업 결과 비교

KERIS 알고리즘	수작업	건 수	
신 규	중 복	94*(1.23%)	2,993(39.13%)
신 규	신 규	2,899(37.90%)	
유 사	중 복	1,184(15.48%)	2,657(34.73%)
유 사	신 규	1,473(19.25%)	
중 복	중 복	1,997(26.11%)	1,999(26.14%)
중 복	신 규	2**(0.03%)	
계		7,649(100%)	

70건과 MARC21 24건을 합친 94건으로 전체의 1.23%에 해당한다.

이들 레코드를 대상으로 누락의 원인을 분석한 결과 대부분 저자와 서명 필드에서의 기술방식 차이와 오기입에서 오는 문제로 나타났다. 저자명은 미 전거 통제, 서명은 하위필드 사용의 미숙, 관제관칭, 대등서명의 오기입이 주원인이었다. 또한 중복에서의 누락은 아니나 전체 레코드의 34.73%에 달하는 2,657건이 유사그룹으로 분류되어 다시 육안으로 구분해야 한다. 유사그룹으로 구분된 2,657건을 수작업으로 검색 한 결과 그 중 중복레코드가 1,184건 (44.56%), 신규레코드가 1,473 (55.44%)로 나타났다. 따라서 기존 알고리즘은 유사그룹을 크게 형성함으로서 정확율을 높이는 대신 재현율을 낮추어 수작업의 분량이 많은 경우이다. 중복레코드로 검색된 레코드 1,999건 중 2건의 신규레코드가 포함됨에 따라 중복레코드 검색

에 대한 정확율은 99.9%에 이르고, 신규레코드 검색에 대한 정확율은 96.8%만에 이른다. 또한 전체 중복레코드 3,275 (94+1,184+1,997)건 중 1,997건만 중복으로 판정함에 따라 재현율은 60.97%에 이른다고 할 수 있다.

6. 2 신규 알고리즘 처리 결과

전체 업로드 대상 레코드 7,649건에 대해 1단계에서 80.16%인 총 6,132건이 중복대상으로 구분되었다. 〈표 7〉에서의 수작업 결과에서 중복레코드가 3,275(94+1,184+1,997)건인 것과 비교하면 187.2%가 중복 대상으로 구분된 것이다. 2 단계에서 12개 요소를 추출하여 정수 값으로 비교한 결과 두개 이상 일치하는 데이터는 3,804건으로 수작업에서의 중복데이터 건수의 116.15%이다. 3,804건을 대상으로 유사값을 산출하고 n:n방식으로 비교한 결과 전체 마스터에서 일치하는 레코드는 총 7,311건으로 1개당 평균 1.92개의 중복대상 레코드가 존재하는 것으로 나타났다.

KERIS 알고리즘과 신규 알고리즘에서의 처리 결과를 비교한 것이 〈표 8〉이다. 중복임에도 입력 오류로 누락된 데이터에 대해서는 유사값을 비교하고, 오류 확률의 고저를 기준으로 가중치를 차등부여 함으로서 검색의 재현율을 높였다. 신규 알고리즘에서도 누락된 3건은 ISBN과 같이 가중치가 높은 요소 들이 비교 대상 레코드에 없고 서명에서의 입력 오류와 부서명의 미기입으로 중복값이 0.64이하로 떨어진 경우이다. 따라서 기존 알고리즘에서 검색 오류가 된 데이터만을 대상으로 신규 알고리즘으로 처리한 결과 중복 검색의 재현율은 전체 중복 1,278건(94+1,184건) 중 1,049건(61+988건)이 검색되어 82.28%로 나타났고 전체 대상 레코드에서는 한계값을 0.73으로 했을 때 재현율이 97.28%로 나타났다.

6. 2. 1 언어별 비교 평가

KORMARC 데이터를 MARC21 데이터

〈표 8〉 신규 알고리즘과 수작업 결과와의 비교

KERIS 알고리즘	수작업	레코드수	신규 알고리즘		
신규	중복	94건	중복값	0.73≦x≦1	61 건
				0.65≦x≦0.72	32 건
				0≦x≦0.64	1 건
유사	중복	1,184건	중복값	0.73≦x≦1	988 건
				0.65≦x≦0.72	194 건
				0≦x≦0.64	2 건
유사	신규	1,473건	중복값	0.73≦x≦1	0 건
				0.65≦x≦0.72	257 건
				0≦x≦0.64	1216 건
중복	신규	2건	중복값	0.73≦x≦1	0 건
				0.65≦x≦0.72	0 건
				0≦x≦0.64	2 건

와 동일한 방식으로 비교했을 때와 데이터의 언어별 특성을 고려해 다른 방식으로 비교했을 때의 중복 검색율을 비교하였다. 비교 결과 〈표 9〉와 같이 문자열 1-3-5방식으로 추출했을 때 중복 대상 레코드로 추출된 건수가 단어의 3-2-2-1방식으로 추출한 것보다 월등히 많다. 또한 한글의 띄어쓰기로 인한 결과의 차이를 조사하기 위해 3-2-2-1방식을 사용하되 처음엔 형태소 분석을 하지 않은 상태에서 비교하고 그 다음엔 형태소 분석을 한 후 비교하였다. 그 결과 국내서의 서명은 띄어쓰기를 하지 않은 상태로 출판되는 것이 대부분이기 때문에 목록자에 따라 임의로 띄어쓰기를 하는 사례가 많아 추출된 문자와의 일치율이 매우 낮게된다. 따라서 국내서 데이터의 서명 추출은 서양서의 추출 방식과는 다른 문자열 1-3-5방식을 적용하는 것이 가장 적합한 것으로 나타났다.

6. 2. 2 서지유형별 비교 평가

원자료의 서지적 특성을 반영하여 중복알고리즘을 다양화 한다면 중복레코드 검색의 재현율은 높아질 것이라는 가정 하에 신규 알고리즘의 효율성을 평가하기 위해 학위논문과 번역서를 일반도서와는 다른 처리 과정으로 처리한 후 〈표 10〉와 같이 중복레코드 검색의 재현율을 비교하였다.

1) 학위논문 비교

비교 결과 학위논문은 일반도서의 처리과정보다 중복값이 평균 0.1153 높아졌다. 303개 레코드 중 1 단계에서 중복대상이 된 142개 레코드에서 중복값 0.73이상인 중복레코드의 수가 학위논문의 처리 과정을 별도로 한 경우 94개의 레코드가 검색되었고, 일반도서와 같은 과정으로 처리한 경우 88개의 레코드가 검색되어 재현율이 높아졌다. 또한 중복값이 1로서 완전 일치로 판명된 레코드 수는 48개에서 60개로 늘어났다. 즉 학위논문의 특성을 살려 발행기관이 대학 코드와 일치하는 경우 가중치를 높임에 따라 중복값이 상승하는 효과를 가져왔다.

〈표 9〉 언어별 처리결과 비교

구 분	문자열 1-3-5 방식	단어별 3-2-2-1 방식	
		형태소 분석 전	형태소 분석 후
중복대상 레코드수	4,631	4,135	4,467
마스터 DB중복대상 레코드수	69,069	210,483	363,806

〈표 10〉 서지 유형별 처리 결과비교

구 분		일반도서 처리과정	서지유형별 처리과정
평 균	학위논문	0.727895	0.843234
중복값	번역서	0.842442	0.892554
중 복	학위논문	88	94
레코드수	번역서	1,666	1,680

2) 번역서 비교

번역서에서도 일반도서의 처리 과정보다 중복값이 평균 0.050112 우수하게 나타났고, 중복값이 0.73이상으로 중복인 레코드 수는 일반과정으로 처리한 경우 1,666개의 레코드에서 1,680개의 레코드로 증가하였다. 또한 중복대상 1,922개 레코드 중 중복값이 1인 레코드 수가 170개에서 446개로 높아졌다. 입력자의 판단에 따르지 않고 책에 쓰여 있는 그대로 전사한 원서명 필드를 비교함으로서 중복 일치율이 높아졌고, 필드 041의 언어코드와 고정장의 언어코드가 일치하는 경우 가중치를 높여 줌으로서 중복값이 상승되었다. 따라서 학위논문과 번역서를 별도로 처리함으로서 중복 검색의 재현율이 월등히 높게 나타났다.

6. 2. 3 유사값 및 가중치 비교 평가

오류 데이터가 많은 데이터베이스를 대상으로 개발한 신규 알고리즘에서는 데이터의 요소를 완전일치 방식으로 비교하지 않고, 각 요소의 유사도를 값으로 측정하면서 요소별로 가중치를 달리 주어 중복 가능성에 대한 값을 부여하였기 때문에 중복레코드 검색의 재현율은 높아질 것이다. 이 같은 원리를 적용한 신규 알고리즘의 효율성을 평가하기 위하여 기존 방식에서 잘못 처리된 레코드를 대상으로 검색하였다. 그 결과 수작업에서 중복임에도 불구하고 누락된 94건과 유사 그룹으로 구분되어 중복 여부를 다시 육안으로 판별해야 하는 총 2,657건의 레코드를 유사값 및 중복값을 측정하여 한계값을 0.73으로 정한 경우 중복레코드는 82%, 신규레코드는 82.5%가 기계적으로 처리되었다. 따라서 유사한 정도 및 중요도를 수치로 부여한 방법이 기존의 비교 방식에 비해 중복 검색의 재현율이 월등히 높은 것으로 나타났다.

6. 2. 4 요소별 비교 평가

중복레코드 검증 알고리즘에서 MARC 데이터의 필드별 비교 방식을 요소별 비교방식으로 바꾼다면 중복레코드 검색의 재현율은 높아질 것으로 추정하여 학위논문과 번역서를 대상으로 요소별 비교를 시도함으로서 신규 알고리즘의 효율성을 평가하였다. 기존 알고리즘에서는 입력 오류나 데이터 필드의 누락으로 단일 필드만을 비교했을 때 필드가 미기입인 경우 중복임에도 불구하고 중복 대상에서 누락되기 때문에 중복 검색율이 낮아지는 사례가 많았으나, 번역서에서 본서명 외에 원서명도 비교하고, 학위논문에서 발행기관 외에 대학코드까지 모두 비교하면서 그 중 하나만 일치해도 중복대상으로 처리하였다. 그 결과 번역서에서는 중복레코드 1,680건 중 232건이 원서명으로 일치하여 추출됨으로서 13.8%의 재현율을 높였다. 또한 학위논문에서도 94건의 검색 결과중 14건은 학위논문 주기사항에서의 대학명과 고정장 필드에서의 대학코드가 다른 것이 추출됨으로서 14.89%의 재현율을 향상시켰다. 이와 같이 실험 데이터를 운용한 결과 언어별, 서지유형별, 유사값 및 가중치, 요소별의 네가지 비교 평가에서 현행 알고리즘에 비해 그 성능이모두 향상된 것으로 나타났다.

7. 결 론

기존 프로그램에서는 기계적인 중복 레코드 검색의 재현율이 60.97%이고 수작업에 의한 검색의 재현율이 36.1%로 총96%였다. 그러나 새로 개발한 알고리즘으로 전체 실험 데이터를 업로드한 결과 재현율이 97.28%로 나타나 기계적인 처리 방식만으로 비교하면 기존의 알고리즘 보다 36.31%가 높아졌다. 그러나 검색 효율성 측정에서 또 하나의 중요한 요소인 정확율은 현행 프로그램에서는 유사그룹이 존재하고신규 프로그램에서는 운영자의 한계값 설정에 따라 조정이 가능하므로 비교의 의미가 없다. 또한 신규 알고리즘에서도 데이터 자체의 오류로 인한 누락 데이터는 100% 완전하게 색출되지 않았기 때문에 기계적으로 모두 해결하는 것은 불가능하므로 유사 그룹이 존재할 수 밖에 없다. 따라서 신규 알고리즘은 데이터의 비교 방식을 다양화함으로서 검색의 효율성을 높이고 수작업의 분량을 최소화하는데 의미가 있다고 할 수 있다. 즉 비교 대상이 되는 모든 값을 수치화하여 운영자가 한계값을 조정함으로서 가장 적정 수준의 재현율과 정확율을 선택할 수 있도록 한 것이 특징이다. 즉 비 표준화된 데이터의 낮은 입력 수준을 최대한 반영한 알고리즘이라고 할 수 있을 것이다. 본 연구에서는 중복 레코드를 추출해내는 검색에 초점을 맞추었으나 중복으로 처리된 레코드를 기계적으로 어떻게 병합할것인가 즉 중복으로 판정된 레코드의 질적 우위를 기계적으로 비교하여 병합 처리하는 부분에 대한 확대 연구가 필요할 것이다.

참 고 문 헌

국립중앙도서관. 2002. 『한국문헌자동화목록 형식 및 기술규칙』. 〈http://www.nl.go.kr/main.php3?top=10&main=kormarc/kormarc.html〉.

최석두 외. 1997. 『대학도서관 분담편목용 입력 기본 표준에 관한 연구』. 서울: 첨단학술정보센터.

酒井清彦. 1994. "オンライン總合目錄 デ-タベ-スの 重複排除." 『情報の科學 と技術』, 44(4): 183-189.

Cousins, S. A. 1998. "Duplicate Detection and Record Consolidation in Large Bibliographic Databases: the COPAC Database Experience in Great Britain." *Journal of Information Science,* 24(4): 231-40.

Cousins, S. A. 1999 "Virtual OPACs versus Union Database: Two Models of Union Catalogue Provision." *The Electronic Library,* 17(2): 97-103.

Intner, Shelia S. 1989 "Quality in Bibliographic Databases : An Analysis of Member-Contributed Cataloging in OCLC and RLIN." *Advances in Library Administration and Organization : A Research Annual.*

Greenwich : JAI Press. 1-24.

Library of Congress. 1999. *Library of Congress Rule Interpretations: Contents.* 〈http://www.tlcdelivers.com/tlc/crs/lcri0000.htm〉.

_____. 2002. *MARC21 Format for Bibliographic Data.* Washington, DC : Library of Congress.

Library Technologies, Inc. 2002. *Database Preparation Services : How Many Duplicates Are There?.* 〈http://www.librarytech.com/D-DEDU-C.HTM〉.

OCLC. 2002. *Bibliographic Formats and Standards,* 〈http://www.oclc.org/oclc/bib.htm〉.

_____. 1999. *Bibliographic Input Standards.* 4th ed. Dublin, Ohio : OCLC,

_____. 2002. *OCLC Batchloading Guide. 3.ed.* 〈http://www.oclc.org/oclc/man/7123bach/〉.

Rittberger, M., W. Rittbeger. 1997. "Measuring Quality in the Production of Databases." *Journal of Information Science,* 23(1): 25-37.

Stankowski, Rebecca House. 1991. "Bibliographic Record Maintenance in a Consortium Database." *Cataloging & Classification Quarterly,* 12(2): 47-62.