종합목록의 중복레코드 검증을 위한 알고리즘 연구
A Study on Duplicate Detection Algorithm in Union Catalog

조 순 영(Sun-Yeong Cho)

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 두개의 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

참고문헌
국립중앙도서관. 2002. 『한국문헌자동화목록형식 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 データベースの 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 CO-PAC Database Experience in Great Britain.” Journal of Information Sci-ence, 24(4): 231-40.
Cousins, S. A. 1999 “Virtual OPACs ver- sus Union Database: Two Models of Union Catalogue Provision.” The Electronic Library, 17(2): 97-103.
Intner, Shelia S. 1989 “Quality in Biblio-graphic Databases : An Analysis of Member-Contributed Cataloging in OCLC and RLIN.” Advances in Library Administration and Or-ganization : 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 Biblio-graphic Data. Washington, DC : Library of Congress.
Library Technologies, Inc. 2002. Data- base 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 Stan-dards. 4th ed. Dublin, Ohio : OCLC,
___. 2002. OCLC Batchloading Guide. 3.ed.
<http://www.oclc.org/oclc/man/7123bach/>.
Rittberger, M., W. Rittbeger. 1997. “Mea-suring Quality in the Production of Databases.” Journal of Information Science, 23(1): 25-37.
Stankowski, Rebecca House. 1991. “Biblio-graphic Record Maintenance in a Consortium Database.” Cataloging & Classification Quarterly, 12(2): 47-62.