저자역할용어사전 구축 및 저작군집화에 관한 연구

Designing a FRBR Work Grouping Algorithm of Bibliographic Records using a Role Term Dictionary of Authors

윤재혁 (Jaehyuk Yun)*
도슬기 (Seulki Do)**
오삼균 (Sam G. Oh)***

초 록

본 연구는 통합서지용 한국문헌자동화목록(KORMARC)으로 작성된 서지레코드를 FRBR의 저작(Work) 단위로 군집화하는 과정에서 나타난 이슈사항들을 분석하고, 이에 대한 해결방안을 고안하였다. 특히 기존의 연구에서는 대표저작자를 식별하고 처리하는 기준이 명확하게 드러나지 않거나 파생저작 레코드의 대표저작자를 선정하는 방법에 대한 논의가 충분히 이루어지지 않았다. 따라서 본 연구는 저작을 창작하는 데 기여한 사람이 다수일 때 대표저작자를 명확하게 식별하기 위한 방법을 고안하는 데 초점을 맞추었다. 이를 위해 책임표시사항(245) 필드의 책임표시 태그(▼d, ▼e)에서 추출한 역할용어를 토대로 표준화된 저자역할용어사전을 개발하여 대표저작자 판별에 활용하는 방안을 마련하였다. 또한 저자명의 유사도와 표제의 유사도를 각각 계산하여 유사도가 일정 수준 이상인 경우 동일한 저작으로 군집화 하는 방법을 채택하였다. 각각의 유사도를 계산하여 동일 저작을 판단하므로 공백, 관제처리, 괄호제거와 같은 데이터 정제 조건을 조정하여 6가지 패턴에 따른 군집화의 정확도를 비교하였고, 저자명과 표제의 유사도가 모두 80퍼센트 이상일 때의 정확도가 가장 높게 나타났다. 본 연구는 대표저작자 선정을 위한 역할용어사전 개발, 대표저작자와 표제의 유사도를 별도로 측정하여 저작군집화를 시도한 실험연구이며 후속 연구에서는 표제 간 유사도 측정의 정확도를 향상시키는 방안과 FRBR 1그룹의 다른 개체(표현형, 구현형, 개별자료) 수준으로 확대하여 활용하는 방안, 국내에서 사용하고 있는 다른 형태의 MARC 데이터에 적용하는 방안을 고안할 예정이다.

ABSTRACT

The purpose of this study is to analyze the issues resulted from the process of grouping KORMARC records using FRBR WORK concept and to suggest a new method. The previous studies did not sufficiently address the criteria or processes for identifying representative authors of records and their derivatives. Therefore, our study focused on devising a method of identifying the representative author when there are multiple contributors in a work. The study developed a method of identifying representative authors using an author role dictionary constructed by extracting role-terms from the statement of responsibility field (245). We also designed another way to group records as a work by calculating similarity measures of authors and titles. The accuracy rate of WORK grouping was the highest when blank spaces, parentheses, and controling processes were removed from titles and the measured similarity rates of authors and titles were higher than 80 percent. This was an experiment study where we developed an author-role dictionary that can be utilized in selecting a representative author and measured the similarity rate of authors and titles in order to achieve effective WORK grouping of KORMARC records. The future study will attempt to devise a way to improve the similarity measure of titles, incorporate FRBR Group 1 entities such as expression, manifestation and item data into the algorithm, and a method of improving the algorithm by utilizing other forms of MARC data that are widely used in Korea.

키워드: 저자역할용어사전, 저작군집화, 대표저작자, 저자명표제유사도
FRBRizing, Author roleterm dictionary, Work grouping, Representative author, Similarity of representative name and title, KORMARC to FRBRizing

* 성균관대학교 일반대학원 문헌정보학과 석박사 통합과정 수료(seisiel@g.skku.edu) (제1저자)
** 성균관대학교 일반대학원 문헌정보학과 박사수료(sinhwask@gmail.com) (공동저자)
*** 성균관대학교 문과대학 문헌정보학과 교수(samoh@g.skku.edu) (교신저자)

■ 논문접수일자 : 2020년 5월 26일 ■ 최초심사일자 : 2020년 6월 11일 ■ 게재확정일자 : 2020년 6월 22일
■ 정보관리학회지, 37(2), 197-223, 2020. http://dx.doi.org/10.3743/KOSIM.2020.37.2.197

1. 서 론

1.1 연구의 목적과 필요성

1998년 IFLA에서 서지레코드의 기능적 요구사항(FRBR; Functional Requirements for the Bibliographic Records, 이하 FRBR)을 발표한 이래로 각 국가별, 기관별로 기존의 서지레코드 구조를 FRBR에 맞게 재구조화하는 작업이 수행되었다. 수행된 연구들의 주된 목적은 이용자가 서지세계를 탐색할 때의 출발점이 되는 저작(Work) 개념을 정립하는 데 있다. 더 정확하게는 MARC 인코딩 스킴에 따라 작성된 기존의 서지 목록에서 저작 개념을 찾아내는 과정을 규격화하기 위한 '저작세트 알고리즘 개발'로 볼 수 있다.

대표적인 FRBR 구조화 알고리즘 중 하나인 미국 OCLC(Online Computer Library Center)의 Work-set 알고리즘은 전거데이터를 활용하여 '저자명 또는 저자명/표제'의 형태를 지닌 식별자(key)의 대표형과 변형을 자동으로 생성하고, 서지레코드에서 추출한 저자명과 표제를 앞서 생성한 식별자 목록과 매핑하는 방식을 취하고 있다(Hickey & Toves, 2009). 한편 미의회도서관(Library of Congress, 이하 LC)이 개발한 FRBR Display 도구는 OCLC의 알고리즘과 달리 전거레코드를 사용하지 않고, 시스템에서 표제 또는 저자명을 검색한 결과로 나타난 서지레코드만을 이용하였다. 각 서지레코드로부터 '저자명/표제' 혹은 '표제'를 추출하여 레코드를 매칭 하는 방식이며 부출표목 필드(7XX)의 유형이 분출일 때, 즉 합집(collection)에 수록된 개별 저작을 입력하기 위해 부출표목을 사용했을 때에만 매칭 과정에 포함시킨다는 특징이 있다(Library of Congress, 2004).

오랜 기간 동안 하나의 인코딩 스킴에 따라 입력되어 온 서지레코드를 새롭게 재구조화하기 위한 알고리즘을 개발하는 작업은 지난한 과업이다. OCLC의 알고리즘을 적용하기 위해서는 인명전거나 통일표제와 같은 전거형 접근점의 역할이 중요하며, LC의 알고리즘을 적용하기 위해서는 기본표목(1XX)이 필수적이다. 그러나 서지레코드의 품질검증이 제대로 되어 있지 않고, 기본표목과 통일표제를 적용하고 있지 않은 국내의 목록 환경에서 OCLC나 LC에서 제시한 알고리즘을 그대로 적용하기란 어려운 문제이다(노지현, 2008).

국내에서 수행된 MARC를 FRBR 구조로 변환하는 연구는 김현희, 유영준, 박서은(2007), 노지현(2008), 이미화와 정연경(2008), 김정현, 이성숙, 이유정(2015)의 연구가 대표적이다. 이 연구들은 MARC 레코드에서 저자명 또는 표제에 관련된 데이터들을 모두 추출하여 이들의 조합을 통해 '저자명+표제'와 같은 접근점을 제공하는 방법을 채택하였다. 저자 정보와 표제 정보를 추출하는 데에는 표제 및 책임표시사항(245), 기본표목(1XX), 부출표목(7XX) 필드가 가장 많이 사용되었다. 김정현, 이성숙, 이유정(2015)은 전거레코드와 통일표제가 입력된 서지레코드에 대해서는 전거형 접근점을 추출하여 '저자명+표제' 조합에 활용하였다.

이 연구들은 MARC에 직접 입력된 데이터를 추출하여 활용함으로써 국내의 서지환경에서 기존의 서지레코드를 저작 단위로 군집화할 수 있는 토대 데이터를 제공하였다는 데에 큰 의미가 있다. 또한 앞서 지적한 국내의 목록

환경에서 나타날 수 있는 이슈사항들을 확인하고, 이를 위해 해결해야 할 사항들을 제언하여 앞으로도 계속해서 이와 같은 방향성을 가지고 연구해야 할 과제들을 제시해주었다. 서지세계에서 자료의 형태는 다양하고, 관련 자료들이 계속해서 생산되고 있기에 추후에도 실제 서지데이터를 대상으로 기존과는 다른 접근방법으로 테스트를 수행하여 문제점을 발견하고 이에 대한 해결방안을 고안하는 작업이 의미 있을 것으로 보인다.

본 연구는 이러한 문제인식 하에 서지레코드를 저작 중심으로 군집화 하는 기존의 연구들이 저작자를 식별하고 처리하는 데 있어 기준이 불분명하다는 점을 발견하였다. 앞선 연구들은 대체적으로 기본표목(1XX)과 부출표목(7XX)으로부터 저자명을 추출하여 저작 군집화를 시도하였는데 이 과정에서 대표저작자를 선정하는 기준을 명확하게 제시하고 있지 않다. 단일저자의 작품을 대상으로는 대표저작자를 선정하는 데 있어 문제가 없을 것이지만, 단일작품 레코드에 원작자와 파생저작자가 복수저자로 함께 존재하는 경우에는 누구를 대표저작자로 선정할 것인지에 대한 기준을 마련할 필요가 있다.

대표저작자를 선정하기 위한 기준을 만들기 위해 우선적으로 표제와 책임표시사항(245) 필드로부터 책임표시 태그(▼d, ▼e)의 저자명과 역할용어를 추출하여 '표준화된 역할용어' 사전을 개발하였다. 다음으로 개발된 역할용어 사전을 토대로 복수저자로 구성된 레코드의 역할용어 데이터를 표준화시켰으며 마지막에는 Tillett(2004)의 'Family of work'를 활용하여 대표저작자를 선정하였다. 역할용어 데이터를 표준화하는 아이디어는 Lee와 Park(2012)의 연구에서 언급된 것으로, 다양한 형태로 입력된 역할용어들을 표준화된 용어로 변환하면 복수저자로 이루어진 레코드에서 대표저작자를 선정하는 데 활용할 수 있는 강점이 있다. 역할용어의 표준화는 역할용어 데이터를 지니고 있으나 표준화되지 않은 245 필드의 한계점과, 데이터는 표준화되어 있지만 역할용어가 부재한 기본표목(1XX)과 부출표목(7XX) 필드의 한계를 극복하기 위한 방안이다.

역할용어 사전의 개발과 대표저작자의 선정 이후에는 대표저자명과 표제 데이터를 조합하고 유사도를 비교하여 일정 수준 이상 유사한 것으로 나타나는 레코드들을 동일한 저작으로 군집화 하였다. 기존 연구자들이 활용한 방법과 마찬가지로 본 연구에서도 저자명과 표제를 접근점으로 사용하였지만, '저자명+표제' 형태의 저작식별자를 비교하는 방식이 아닌 저자명 유사도와 표제 유사도를 별도로 계산하여 각각의 유사도가 일정 수준 이상인 경우 동일한 저작으로 그룹화 하는 방법을 채택하였다는 점에서 차이가 있다. 또한 역할용어사전을 구축하고, 이를 토대로 한 대표저작자를 선정하는 기준을 마련하여 추후 전거데이터 구축 시 활용 가능한 아이디어를 제공하는 기초연구의 의미를 갖는다.

1.2 연구대상 선정 및 데이터 수집

본 연구에서 사용한 샘플 데이터는 국립중앙도서관 홈페이지의 KORMARC 데이터를 크롤링하여 수집하였다. 데이터 크롤링은 프로그래밍 언어인 R을 사용하였으며, LC의 FRBR Display 알고리즘 개발에서 사용한 방법과 유

사하게 시스템에서 표제를 검색하여 제시된 검색 결과에 대한 서지데이터를 수집하였다.[1] 연구대상 작품은 다양한 경우의 수와 서지유형, 파생작품을 보유하고 있는 단행자료로 한정하였다.[2] 수집한 데이터에서 연구에 필요한 필드를 추출한 후 시트 형태로 생성하여 정제 작업을 수행하였는데, 이 과정에서 총서자료를 제외[3]하였다. 최종적으로 연구에 사용된 서지데이터의 수를 정리하면 〈표 1〉과 같다.

2. 선행연구

2.1 FRBR 저작세트(Work-Set) 알고리즘 개발에 관한 선행연구

본 연구는 현재 KORMARC 인코딩 스킴에 따라 작성된 국립중앙도서관의 레코드들을 추출하여 FRBR의 저작세트를 만드는 실험적 연구로 기존에 진행된 MARC에서 FRBR 저작세트 알고리즘을 구현에 관한 연구들을 살펴보는 작업이 필요하다.

널리 알려져 있는 연구로는 OCLC의 Hickey와 Toves(2009)가 개발한 Work-Set 알고리즘, Library of Congress(2004)의 FRBR Display 도구가 있다. FRBR Work-Set 알고리즘은 전거레코드를 활용하여 '저자명/표제' 식별자의 대표형과 변형 목록을 생성하고, 서지레코드에서 추출한 저자명과 표제를 식별자 목록과 매핑한다. 다음으로 알고리즘에서 정의한 규칙에 따라 동일한 저작세트끼리 군집화를 실시하고, '저자명/표제'의 대표형을 저작의 대표키로 할당한다. 미국의회도서관의 FRBR Display 도구는 전거레코드를 활용하지 않고 서지레코드에서 '저자명+표제' 또는 '표제'를 추출하여 레코드끼리 매칭하고 정렬하는 작업을 통해 저작

〈표 1〉 연구에 사용된 샘플 서지데이터의 수

작품명(검색 키워드)	최초 '표제' 검색 결과	총서자료가 제외된 최종 샘플 데이터
노인과 바다	223	149
로미오와 줄리엣	160	100
오만과 편견	211	192
해리포터	115	114
햄릿	328	216
합계	1,037	771

1) '표제'를 검색한 결과를 토대로 데이터를 크롤링하는 방법을 사용하였기에 저자명 전거데이터는 수집 데이터에 포함되지 않았다.
2) 국립중앙도서관의 단행자료는 도서, 학위논문, 전자책의 세 범주로 세분화되는데 전자책의 경우는 MODS 인코딩 스킴으로 레코드가 작성되어 있어 데이터 추출 작업 시 제외되었다. 다양한 경우의 수와 서지유형, 파생작품이 다양한 대상 선정은 노지현(2008), 김정현, 이성숙, 이유정(2015)의 연구를 참고하여 연구자가 직접 선정하였다.
3) 일반적으로 하나의 레코드가 하나의 저작 정보를 담고 있는 단행자료와 달리 총서자료는 하나의 레코드가 다수의 저작을 포함하고 있다. 총서자료는 ① 내용주기(505)의 존재 여부, ② 표제 및 책임표시사항(245) ▼a의 반복 사용 여부, ③ 제어번호(001)의 중복 여부를 확인하는 과정을 거쳐 제외되었다.

수준을 유지하기 위한 기준을 만들었다. '저자명+표제'의 데이터는 100, 110, 111 + 240, 243, 245 필드를 순서대로 매칭하여 저작세트를 만들었으며, '표제'를 기준으로 저작을 판별하는 기준으로는 130, 240, 243, 245 필드의 데이터를 사용하였다.

국내에서도 MARC 레코드를 FRBR 구조로 변환하여 저작 중심의 군집화를 시도한 연구가 다수 수행되었다. 연구자들은 모두 표제와 저자명과 관련된 필드를 추출하여 표제+저자명 혹은 표제+무저자와 같은 조합으로 결합하여 알고리즘을 개발하였고, 연구 과정에서 나타난 이슈사항들을 토대로 MARC의 FRBR 구조화에서의 고려사항에 대해 논하였다. 김정현, 이성숙, 이유정(2015)은 국립중앙도서관의 서지레코드를 추출하여 FRBR 구조화 알고리즘을 개발하였고, 검색시스템을 구현하여 검색 시 저작, 표현형, 구현형이 관계가 어떻게 나타나는지를 검증하였다. 저작의 군집화는 RDA 6.27(주석 달기)의 내용처럼 저작의 전거형 접근점을 만들어 관련 저작을 모두 모을 수 있도록 '저작+표제'(저자가 없는 경우 '표제')세트를 만드는 데 주력하였다. 저자와 관련된 데이터는 전거레코드와 서지레코드의 기본표목(100)과 부출표목(700)에서 추출하였다. 표제 데이터는 표제와 표제관련필드(240, 245, 246)와 부출표목-비통제관련/분출표제(740), 그리고 통일표제(130, 730)로부터 추출하여 사용하였다. 이를 통해 100+245, 100+246, 700+245▼a, 700+740, 700+245▼b, 700+740, 245+무저자, 130(무저자-통일표제), 730(통일표제)로 저작을 군집화하였다. 노지현(2008)은 국립중앙도서관의 '햄릿'에 대한 서지레코드 161건을 대상으로 군집화를 시도하였는데, 표제는 240, 245, 740에서 추출하고 저자명은 100, 700, 900에서 추출하여 100+240, 100+245, 100+740, 700+240, 700+245, 700+740, 900+245, 900+740의 저작세트를 생성하였다. 기본표목 필드(100)을 기준으로 하되, 기본표목을 적용하고 있지 않은 경우에는 부출표목(700) 중 첫 번째 데이터를 기준으로 산정하였다. 김현희, 유영준, 박서은(2007)은 음악자료 387건을 대상으로 FRBR 구조화하는 알고리즘을 설계 후 KERIS의 레코드 107건을 대상으로 적용하였다. 저자명과 표제를 활용할 때에는 기본표목(100, 110, 111)/부출표목(700, 710, 711)을 순서대로 체크한 후 240, 245를 체크하였고, 표제만을 활용할 때에는 130을 순서대로 체크하고 240, 245를 체크하였다. 이미화와 정연경(2008)은 기본표목(1XX)이 있는 경우 표제+저자명으로, 기본표목이 없는 경우 부출표목(7XX)에서 저자명을 2인 색인하여 저자가 하나만 일치할 경우에도 동일저작으로 처리하였다. 통일표제(130)가 있는 경우 표제로만 색인을 작성하고, 통일표제가 부재한 경우 240 ▼a를 확인하여 표제의 대표형으로 처리하였다. 저자명은 100, 110, 111을 추출하여 표제+저자명 색인을 추출하였다.

2.2 서지적 관계 유형에 관한 선행연구

기존의 목록레코드는 구현형 수준에서 구조화되어 있고, 저작 수준에서의 서지구조가 논의된 것은 FRBR가 등장한 이후라고 볼 수 있다. FRBR는 서지세계의 탐색에 있어 저작을 중심으로 다른 서지데이터와의 관계성을 중요시한다. 이와 관련한 이슈사항은 하나의 저작

과 새로운 저작의 경계기준을 어디에 두느냐의 문제와, 동일저작으로 묶을 수 있는 서지적 관계들에 대한 명확한 기준이 필요하다는 것이다.

Tillett는 목록규칙을 분석하여 서지적 관계를 대등관계, 파생관계, 기술관계, 전체-부분관계, 딸림자료관계, 전후관계, 특성공유관계의 7가지를 제안하였다(Tillett, 1987; 김순희, 이성숙, 2005에서 재인용). FRBR에서는 1집단인 저작, 표현형, 구현형, 개별자료 간의 상호관계와 1집단과 2-3집단 간의 서지적 관계가 핵심이 된다. 박지영(2009)은 FRBR의 서지적 관계를 기반으로 RDA의 서지적 관계를 1차관계(하나의 저작과 표현형, 구현형, 개별자료 간의 관계), 파생관계, 기술관계, 대등관계, 전체부분관계, 딸림관계, 전후관계로 정리하였다. 김정현(2007)은 MARC 연관저록필드, Tillett, Bertha, Smiragila의 서지적 관계 유형, FRBR와 RDA의 서지적 관계 유형을 분석하여 한국어저작의 서지적 관계 유형을 분석하기 위한 준거로 삼고자 하였는데, Tillett의 7가지 관계유형 중 특성공유관계를 제외한 대등, 파생, 기술, 전체-부분, 딸림, 전후관계로 구분하였고, 후속 연구(김정현, 2015)에서는 사서오경 자료의 저작 유형을 대등(복제, 사본, 영인본 등), 파생(번역, 개정, 요약 등), 기술(주석, 역주, 비평 등), 전체-부분, 딸림(부록, 색인집 등), 전후관계로 구분하였다.

자료에 따른 서지적 관계 유형을 분석하고, KORMARC의 대응 필드와 매핑하는 연구도 다수 수행되었다. 김순희와 이성숙(2005)은 FRBR의 서지적 관계 유형과 KORMARC-통합서지용 필드를 매핑하였다. 저작과 새로운 저작과의 관계는 후속, 부록, 보유, 요약, 개작, 변형, 모방, 종속적인 구성요소, 독립된 구성요소, 필수적인 지적인 면, 저작에 책임을 진 개인과 단체, 저작의 주제로 나타나는데, 그 중에서 보유, 요약, 개작, 변형, 모방, 종속적인 구성요소, 독립된 구성요소, 필수적인 지적인 면은 매핑되는 필드가 부재하다는 점을 발견하였다. 이성숙과 이현주(2013)는 기존연구에서 조사된 국악자료 관계유형을 토대로 FRBR에서 제시한 관계유형 및 범주를 보완한 후에 KORMARC의 대응필드와 매칭하는 작업을 수행하였다. 국악자료의 관계유형은 대등(복제, 대체, 재구성), 파생(번역, 개정, 축약, 변조, 편곡, 채보/악보, 연주/공연), 기술(해제, 비평, 안내), 전체-부분, 딸림자료(부록), 전후(후속)관계로 분류하였다. 이에 더하여 RDA에서의 서지적 관계 유형을 내용측면과 용기측면으로 나누어 서지적 관계 유형을 분류하였고, KORMARC 필드와의 매핑 작업을 수행하였다. 이때 서지적 관계 유형의 대범주는 Tillett가 제시한 서지적 관계 유형을 토대로 하였다. 김정현(2015)은 서지적 관계 유형에 관한 기존 연구들을 분석하여 서지적 관계를 분류한 후 사서오경의 서지적 관계 유형 분석의 준거로 삼았다. 사서오경에 나타난 서지적 관계 유형은 대등(복제, 사본, 영인본 등), 파생(번역, 개정, 요약 등), 기술(주석, 역주, 비평 등), 전체-부분, 딸림(부록, 색인집 등), 전후관계인 것으로 확인되었고, 이를 KORMARC 형식의 대응 필드와 매핑하는 작업을 수행하였다. 그 중 기술관계에 해당하는 해설이나 평론은 고유의 필드번호가 없이 일반주기 필드에 기술하고 있음이 확인되었다.

FRBR의 서지적 관계에서는 저작과 다른 저작을 구분하는 서지적 관계의 정의 또한 매우 중

요하다. Tillett는 저작의 계통(Family of Work)을 정리하면서 하나의 저작과 새로운 저작을 구분하는 기준점을 제시하였다(Tillett, 2004). 이에 따르면 원저작에서 내용의 변경이 이루어졌을 때 등가, 파생, 기술관계의 세 가지 서지적 관계가 형성되며 이에 따라 동일저작인지 새로운 저작인지를 구분할 수 있다. 김정현(2015)은 사서오경에 나타난 저작과 새로운 저작 간의 관계를 후속, 부록, 요약, 개작, 변형, 각색, 모방 등으로 정리하였다.

2.3 선행연구 분석

FRBR 저작세트 알고리즘을 개발하는 기존의 연구들을 살펴보면 기본표목과 부출표목에 기입된 저자명 및 표제와 관련된 필드들을 최대한 추출하여 '저자명+표제' 혹은 '표제+무저자'의 결합을 통해 저작세트를 생성하였다. 또한 연구자들은 일관되게 전거형 접근점의 중요성에 대해 강조하고 있다. 저자명(개인명/단체명 등)은 어느 정도 제어가 되고 있지만, 표제의 경우는 통일표제와 관련된 필드의 사용이 미비함을 지적하였다. 특히 고전작품의 경우 통일표제의 사용이 매우 효과적인데 실제로는 '춘향전'에서 통일표제 사용의 사례를 일부 확인할 수 있었고 그 이외에는 전무하였다. 또한 역할용어의 활용과 연관저록의 입력요소 강화, 서지의 관계유형을 일반주기가 아닌 연관저록에 기록해야 한다는 점을 제언하고 있다.

서지관계의 유형에 대한 기존 연구들은 KORMARC의 실제 데이터요소와 서지적 관계를 매핑하여 활용할 수 있는 필드와 부재한 필드를 상세하게 분석해줌으로써 후속연구의 방향성을 제시해주었다. 특히 하나의 저작과 새로운 저작의 경계기준을 삼는 작업이나 동일한 저작으로 군집화 할 때의 서지적 관계의 기준 정립과 같은 연구에 있어 기준점이 될 수 있다. 무엇보다 선행연구에서 서지적 관계의 유형을 대범주-중범주로 분류하는 작업에 더하여 KORMARC의 필드와 매핑하는 연구가 다수 수행되었는데, 본 연구에서 KORMARC 데이터 크롤링 및 데이터 정제 과정에서 필요한 필드를 결정함에 있어 중요한 정보원이 되어주었다. 추후에 KORMARC의 데이터를 크롤링하여 레코드의 각 필드의 데이터들을 활용 및 분석하는 연구를 수행하고자 할 때에도 매우 유용한 자료가 될 것이다.

본 연구는 선행연구자들이 수행한 연구와 마찬가지로 저자명과 표제를 결합하여 저작을 식별하고 군집화 할 수 있는 방안을 개발하는 것이며, 그 중에서도 다수의 저자가 창작한 저작물에서 대표저작자를 선정하기 위해 245 필드에 입력된 역할용어를 활용하는 방법을 개발하는 데 초점을 맞추었다. 이는 Lee와 Park(2012)[4]이 ISTC(International Standard Text Code, 이하 ISTC)의 식별자와 전거레코드의 식별자, 저자역할용어코드를 결합한 정보를 통해 FRBR의 저작을 식별하는 데 도움이 될 것이라고 언

4) Lee와 Park(2012)에 따르면 ISTC의 Work 개념은 FRBR의 저작(Work) 개념과 완전히 일치하는 개념이 아니며, 오히려 표현형(Expression)과 더 관련될 수 있다고 언급하면서, 다만 국제표준인 ISTC의 역할용어 개념과 KORMARC에서의 역할용어, 그리고 전거레코드의 식별자 등을 결합하면 저작의 식별에 훨씬 더 도움이 될 것이라고 제언하였다. 이들은 나아가 ISTC나 ISNI Agency처럼 저작 단에서의 식별자(국제표준이면 더욱 더 효과적)를 관리하는 기관의 역할이 필요하다고 하면서 국가도서관이 그 역할을 수행할 필요가 있음을 강조하였다.

급하고, 이미화와 정연경(2008)이 저작의 정확한 식별을 위해 2인 이상의 공저자를 갖는 저작물의 경우 각 저자별로 역할용어를 기술하도록 제언하고 있어 역할용어의 활용이 FRBR의 저작을 식별하는 데 도움이 될 것이라는 판단에서 비롯되었다. 역할용어의 활용을 위해 본 연구는 KORMARC을 기반으로 작성된 레코드로부터 저자의 역할을 나타내는 용어를 추출하고, 이를 표준화 및 세분화하여 '역할용어 사전'을 만들었으며, 서지적 관계 유형에서 Tillett의 'Family of Work'를 대표저작자를 선정하는 기준을 세우기 위한 참고자료로 활용하였다.

3. 표준화된 역할용어 사전을 활용한 대표저자명 식별자 부여

연구를 수행하기 위해 사용된 도구는 Microsoft Excel, VBA(Visual Basic for Application), 그리고 프로그래밍 언어인 R이 사용되었다. 데이터 정제는 주로 Excel을 이용하였는데, 그 이유는 시트 단위로 데이터가 구성되고 정제 단계에 따라 변화되는 데이터의 모습을 육안으로 확인 가능하기 때문이다. 함수의 반복 사용이나 데이터 간 유사도 비교와 같이 Excel로 구현할 수 없는 과정들은 VBA와 R을 보조로 사용하여 수행하였다.

3.1 샘플 데이터 상세 분석

수집된 샘플 데이터의 레코드에서 저자명 및 통일표제와 관련된 필드의 입력 건수를 살펴보면 다음의 〈표 2〉와 같다.

저자명과 관련된 기본표목(1XX) 및 부출표목(7XX)을 살펴본 결과에 따르면 총 771개의 작품 중에서 1) 단독저자의 작품이 317건(41.12%)이었으며, 그 중 2) 기본표목을 사용한 작품이 71

〈표 2〉 샘플 데이터의 저자명, 통일표제 관련 필드 상세 분석

전체 레코드 수		771건
저자명	100 필드의 수	69
	110 필드의 수	2
	700 필드의 수	1,219
	710 필드의 수	49
	900 필드의 수	662
	910 필드의 수	3
통일표제	130 필드의 수	0
	730 필드의 수	0
	930 필드의 수	0

1) 단독저자 작품의 수: 317 / 771 (41.12%)
2) 기본표목(100, 110)이 입력된 작품은 모두 단독저자의 작품: 71 / 771 (9.21%)
3) 단독저자의 작품 중 부출표목을 사용하여 표현한 경우: 246 / 771 (31.91%)
4) 복수저자의 작품 중 기본표목 없이 부출표목으로만 표현한 경우: 454 / 771 (58.88%)
5) 부출표목(700, 710) 필드 중 역할용어를 가진 필드의 수: 117 / 1,268 (9.23%)

건(9.21%), 3) 부출표목을 사용한 작품은 246건(31.91%)으로 기본표목과 부출표목을 혼용한 레코드는 발견되지 않았다. 한편, 4) 나머지 복수저자의 작품 454건(58.88%) 중 기본표목을 사용한 경우는 없었으며, 모두 부출표목만을 사용하였다. 저자의 수에 관계없이 샘플 데이터에 존재하는 부출표목의 수는 전체 1,268개였으며 이들 중 5) 역할용어를 가진 경우는 117개(9.23%)인 것으로 확인되었다. 수집된 샘플 데이터에서 통일표제와 관련된 필드(130, 730, 930)에는 데이터가 입력되어 있지 않은 것으로 확인되었다. 단독저자의 작품 317건은 저자가 한 명이므로 기본표목, 부출표목 사용 여부와 상관없이 저작자가 명확하게 드러난다. 반면, 복수저자를 부출표목으로만 표현한 경우에는 대표저작자를 파악하는 작업이 부수적으로 필요하다. 대표저작자 식별은 각 작품에 기여한 저자(기여자)들의 역할을 비교함으로써 달성될 수 있는데, 문제는 샘플 데이터가 역할용어를 거의 보유하고 있지 않다는 점이다. 전체 샘플 데이터에서 부출표목이 총 1,268건이나 사용되었음에도 불구하고 10퍼센트도 되지 않는 부출표목만 역할용어를 지니고 있다. 역할용어의 부재는 대표저작자를 파악하기 어렵게 만드는 요인이며, 저작식별자로 '저자명+표제'의 조합을 사용하게 될 경우 군집화의 정확률을 낮추는 결과를 초래할 수 있다. 따라서 본 연구자는 저작식별자를 부여하는 작업에 앞서 2인 이상의 저자(기여자)가 저작물에 관여했을 경우 각자 어떤 역할을 했는지를 파악하는 방법을 고안하였다.

3.2 대표저작자 선정을 위한 표준화된 역할용어 사전 구축

앞서 기술한 바와 같이 단독저자가 창작한 작품은 대표저작자 식별에 큰 어려움이 없는 반면 복수저자가 저작물에 관여하였을 경우에는 대표저작자를 선정하는 기준이 필요하다. 기존의 연구에서는 부출표목의 첫 번째 저자를 대표저작자로 선정하거나(노지현, 2008), 부출표목에서 2번째까지 입력된 저자를 추출하여 비교하는 방법(이미화, 정연경, 2008)을 사용하였다. 이처럼 부출표목을 활용한 대표저작자 선정은 KORMARC이 '서명주기입방식'으로 기본표목을 사용하지 않는 것을 원칙으로 하고 있는 상황에서는 최선의 방법인 것으로 판단된다. 본 연구 또한 기존의 연구들과 마찬가지로 다수의 부출표목들 중에서 대표저작자를 선정하는 데 초점을 두었으나, 보다 더 높은 정확도를 얻기 위해 245 필드에 사용된 역할용어를 활용하여 대표저작자를 식별하는 방법을 고안하였다.

Lee와 Park(2012)은 KORMARC 레코드를 FRBR 구조로 변환함에 있어 역할용어의 활용과 국제표준 식별자의 활용을 제언한 바 있다. 그들은 KORMARC의 로컬표목(9XX)에 역할용어를 입력하는 필드가 존재하지만 활용이 미비하다는 점을 언급하였다. 이는 KORMARC이 KCR4에 따라 기본표목보다 부출표목을 주로 사용하는 전통에서 기인한 것으로 보인다. 샘플 데이터를 분석해 본 결과 부출표목(700, 710)에 사용된 역할용어는 약 9퍼센트밖에 되지 않았다. 주로 사용하는 부출표목의 활용률이 이처럼 저조한 것을 고려하면 로컬표목의 활용률은 더더욱 낮게 나타날 수밖에 없다.

245 필드에서 역할용어를 포함하고 있는 경우는 전체 771개의 작품 중 511개(66.28%)에 해당한다. 특히 복수저자(기여자)의 작품 454개는 모두 245 필드에 역할용어를 포함하고 있었는데, 이는 부출표목이 역할용어를 약 9퍼센트밖에 지니지 못한 것과는 차이가 극명하게 갈리는 결과이다.[5] 따라서 다수의 저자(기여자)들이 작품의 창작에 기여했을 때 이들의 역할을 구별함에 있어 기본표목, 부출표목 필드의 데이터보다 245 필드를 활용하는 것이 더 적합할 것으로 판단하였다. 그러나 245 필드의 역할용어들은 한글, 한자, 영어 등의 다양한 언어로 작성되거나 '저', '지음', '글'과 같이 동일한 역할이더라도 다양한 표현이 존재할 수 있다. 그러므로 역할용어들 간의 정확한 비교를 위해서는 245 필드에 나타난 역할용어들의 종류를 파악하고 표준화시킬 수 있는 과정이 필요하며, 이를 위해 역할용어 사전을 구축하여 활용하였다.

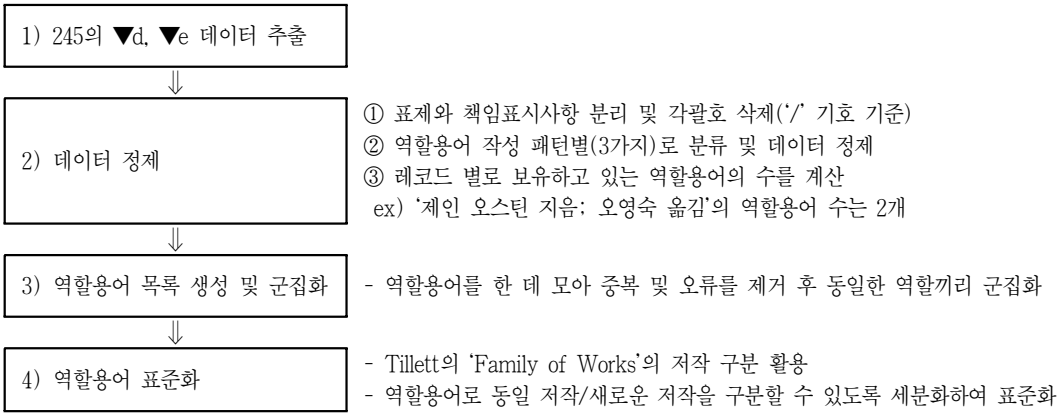
역할용어 사전 구축 과정을 개괄하면 〈그림 1〉과 같다.

3.2.1 245 필드의 ▼d, ▼e 데이터 추출

245 필드는 첫 번째 저자를 ▼d에 기술하고, 두 번째 저자부터는 ▼e를 반복적으로 사용하여 기술한다. 따라서 해당 데이터를 모두 수집하였다.

3.2.2 데이터 정제

표제와 책임표시사항은 빗금(/) 부호를 중심으로 구분되므로 Excel의 텍스트 나누기 기능을 사용하여 표제와 책임표시사항을 분리한다. 이때 표제에서 동일한 기호가 사용될 수도 있으므로 '/▼d'와 같이 식별기호를 포함한 문자열을 역사선(\) 기호로 변경하여 텍스트를 분리하였다. 각괄호가 사용된 경우 각괄호를 모두 제거하고, 저자의 역할이 다를 때 사용하는 기호인 쌍반점(;)이 공백과 결합된 경우



〈그림 1〉 역할용어 사전 구축 과정 개괄

5) 245 필드가 역할용어를 다수 보유하고 있는 것은 목록 작성의 특성에 의한 것으로, 목록을 작성하는 단계에서 자원에 역할용어가 명확하게 드러나지 않은 경우에는 목록작성자가 각괄호([ ])를 사용하여 직접 기술하고 있다.

('[공백];'), 역할용어를 처리하는 과정에서 역할용어와 공백이 결합되는 문제('지음[공백]')가 발생할 수 있으므로 공백을 제거하고 쌍반점만 남긴다.

다수의 저자(기여자)가 작품 창작에 기여했을 때 동일한 역할을 수행한 저자는 쉼표(,)를 사용하여 나열하며, 저자 간 수행한 역할이 다르면 저자 사이에 쌍반점(;)을 입력하여 구분한다. 이러한 작성 원칙을 염두에 두고 실험 데이터를 살펴본 결과, 역할용어를 작성하는 패턴을 〈표 3〉과 같은 세 가지로 구분할 수 있다.

패턴 Ⓑ는 역할용어와 이름 사이에 쌍점(:)을 입력함으로써 역할과 이름을 명확히 구분할 수 있을 뿐만 아니라 Ⓐ와 Ⓒ 패턴과도 구별되는 특징을 지닌다. Ⓐ와 Ⓒ는 이름과 역할을 나열한다는 점은 공통적이나 나열의 순서가 반대임을 확인할 수 있다. 이때 Ⓒ는 저자가 수행한 역할을 '피동형 동사 + by'를 결합하여 사용하므로 'by'를 기준으로 삼으면 역할과 이름의 구분이 가능하다. 즉, 쌍점(:)과 'by'의 포함 여부에 따라 레코드의 저자 사항을 분류할 수 있다. 하지만 단순히 'by' 문자열의 포함 여부를 기준으로 사용하면 문자열에 'by'가 포함된 모든 사례를 Ⓒ 패턴으로 분류할 수 있으므로 '[공백]by[공백]'과 같이 'by'의 앞과 뒤에 공백이 들어가 있는 문자열의 포함 여부를 기준으로 삼았다.

본 연구에서 각 패턴이 가지고 있는 특징에 따라 고안한 역할용어 추출 방식은 다음과 같다. Ⓐ는 쌍반점(;)을 기준으로 문자열을 분리한 후, 각 문자열의 가장 마지막에 위치한 단어를 역할용어로 추출하였다. Ⓑ는 식별기호(▼d 또는 ▼e)와 쌍점(:) 사이의 단어를 역할용어로 추출한다. Ⓒ는 'by'[6]까지의 문자열을 역할용어로 추출하였다.

책임표시사항으로부터 역할용어를 추출하는 과정은 레코드가 보유하고 있는 역할용어의 수만큼 반복하여 이루어졌다. 또한 추후에 역할용어 사전을 참조하여 기존의 역할용어를 표준화된 용어로 변환하는 작업에서 Excel의 VLOOKUP 함수를 활용하게 되는데, 이때에도 역할용어의 수만큼 함수가 반복하여 적용되어야 한다. 따라서 각각의 레코드가 보유하고 있는 역할용어의 수를 계산할 필요가 있다.

3.2.3 역할용어 목록 생성 및 군집화

단일 저자의 작품은 책임표시사항에 역할용어가 입력된 사례가 57건(7.40%)으로 복수 저자의 작품에 비해 매우 적으며, 역할용어의 입

〈표 3〉 역할용어의 입력 패턴 및 예시

역할용어 입력 패턴	예시
Ⓐ 이름a 역할1; 이름b 역할2	▼d어네스트 헤밍웨이 지음;▼e황종호 옮김
Ⓑ 역할1: 이름a; 역할2: 이름b	▼d지은이: 어니스트 헤밍웨이 ;▼e옮긴이: 김옥수
Ⓒ 영문역할동사1 by 이름a; 영문역할동사2 by 이름b;	▼d[written by] Ernest Hemingway ;▼eannotated with critical introduction by Byung-chul Kim

6) 실험 데이터 중에서는 'written by'로 쓰지 않고 '▼dby'와 같이 'by'만 사용한 사례가 있었는데, 정확도를 높이기 위해 '▼dby'를 모두 '▼dwritten by'로 변경한 이후 역할용어를 추출하였다.

력 여부를 정확하게 자동으로 판단할 수 있는 요소가 부족하다. 이에 따라 단일 저자의 작품을 대상으로는 역할용어를 추출하지 않기로 결정하였고, 역할용어 추출 과정을 위의 세 가지 패턴에 단일 저자인 경우를 추가한 네 개의 분기로 구성하였다.

각 분기별 역할용어 추출 방법을 정리하면 아래와 같다.

① Ⓑ 패턴: 식별기호(▼d 또는 ▼e)와 쌍점(:) 사이의 단어를 역할용어로 추출
② Ⓐ 패턴: 쌍반점(;)을 기준으로 문자열을 분리한 후, 각 문자열의 가장 마지막에 위치한 단어를 역할용어로 추출
③ 단일 저자의 작품은 역할용어를 추출하지 않음
④ Ⓒ 패턴: '[공백]by[공백]' 까지의 문자열을 역할용어로 추출

레코드로부터 추출한 역할용어들은 245 필드에서 저자의 역할이 달라질 때 쌍반점(;)을 입력하는 규칙을 동일하게 적용하여 '지음;옮김'과 같은 형태로 한 데 모으며, 이처럼 각각의 레코드가 지닌 역할용어의 모음을 '역할용어세트'라 명명하였다.

역할용어세트의 생성이 완료되면 전체 세트를 한 셀에 모으고 쌍반점을 기준으로 삼아 역할용어를 모두 쪼갠다. 이후로는 중복된 역할용어를 제거하고, 공백 셀이나 사람의 이름이 역할용어로 입력된 오류 데이터를 수정하는 작업을 수행하였다. 이 단계에서의 수정작업은 KORMARC 서지레코드 원본을 수정하는 것이 아니라, 역할용어 추출 과정에서 발생한 오류로 인해 비어있는 셀과 역할용어로 사람의 이름이 추출된 것을 삭제하는 것을 의미한다. 오류가 발생한 레코드를 살펴본 결과, 서지레코드가 잘못 입력되었을 때 오류가 발생하며, 사람의 이름이 역할용어로 추출된 경우에는 해당 오류가 발생한 레코드를 찾아 오류 레코드임을 표시하였다. 마지막으로 수집된 역할용어를 동일한 역할끼리 분류하여 정리하였으며, 그 결과는 〈표 4〉와 같다.

3.2.4 역할용어 표준화

수집된 역할용어를 비교하는 과정에서 발생

〈표 4〉 패턴별 처리 과정을 통해 추출된 역할용어 목록

written by	저	譯	만화	譯註	번역, 개작
공저	著	譯	일러스트	註譯	영화
공지음	著者	譯者	엮음	해설	원작
글	지은이	옮긴이	편	해설·주석	原作
만듦	지음	옮김	編	解題	원저
작	共譯	옮김이	편역	retold by	原著
作	공옮김	Comics	annotated with critical introduction by	각본	편저
작곡	번역	illustrated by	번역·해설	다시쓴이	編著
작시	역	그림	역주	번안	단일

할 수 있는 혼란을 최소화하기 위해서는 역할용어를 표준화시킬 필요가 있다. 본 연구에서는 역할용어의 표준화 과정에 이미 존재하는 표준을 활용하는 방법을 먼저 고려하였다. ISTC는 ISO(국제표준화기구)의 기준에 따라 문자 기반 작품을 대상으로 하는 국제표준식별체계로 Contributor role[7]이라는 메타데이터 요소를 명시하고 있으며, 이를 활용하여 일차적으로 표준화시킨 이름은 〈표 5〉와 같다.

ISTC의 Contributor role을 사용하여 역할용어를 어느 정도 표준화하였지만 다음과 같은 문제를 가지고 있는 것으로 확인되었다. 하나는 ISTC의 용어로 표현할 수 있는 역할용어의 수가 8개(author, author of supplementary text, creator of other non-text content, editor or reviser, traslator, compiler, excepter, unspecified)로 매우 적다는 점이다. 특히 ISTC에 따라 표준화시킨 역할용어세트를 모아놓은 〈표 6〉을 참조하면 Unspecified로 표준화된 역할용어가 한 레코드 내에 다수가 존재하는 것을 확인할 수 있는데, 이는 대표저작자를 선정하는 데 어려움을 준다. 또 다른 문제는 일부 역할용어(ex: creator of other non-text content 등)의 경우 문자열의 길이가 길고 공백이 존재하여 데이터 정제 시 오류 발생 가능성이 존재한다는 점이다.

본 연구에서는 이와 같은 ISTC의 Contributor role을 활용하는 방안의 한계점을 극복하기 위해 8개의 역할용어로 표준화된 데이터를 2차적으로 세분화하는 단계를 추가하였다. 역할용어를 세분화하는 기준은 Tillett(2004)의 'Family of Works'를 참조하였다. 이에 따르면 원저작에서 콘텐츠의 변경이 이루어졌을 때, 원저작과 등가관계일 때에는 동일한 저작/동일한 표현형이, 파생관계에서는 동일한 저작이 될 수도 있고 새로운 저작이 될 수도 있다. 기술관계일 때

〈표 5〉 ISTC - Contributor role을 사용하여 추출된 역할용어 표준화

추출된 역할용어	ISTC 기반 역할용어
written by / 공저 / 공지음 / 글 / 만듦 / 작 / 作 / 작곡 / 작시 / 저 / 著 / 著者 / 지은이 / 지음	Author
共譯 / 공옮김 / 번역 / 역 / 譯 / 譯 / 譯者 / 옮긴이 / 옮김 / 옮김이	Translator
Comics / illustrated by / 그림 / 만화 / 일러스트	Creator of other non-text content
엮음	Compiler
편 / 編 / 편역	Editor or reviser
annotated with critical introduction by / 번역·해설 / 역주 / 譯註 / 註譯 / 해설 / 해설·주석 / 解題	Author of supplementary text or Excerpter
retold by / 각본 / 다시쓴이 / 번안 / 번역, 개작 / 영화 / 원작 / 原作 / 원저 / 原著 / 편저 / 編著 / 단일[8]	Unspecified

7) ISTC User Manual의 메타데이터 요소 중 Contributor role은 "The role of each contributor identified, from one of the following types: Author, Author of supplementary text, Creator of other non-text content, Editor or reviser, Translator, Compiler, Excerpter, Unspecified."와 같이 8개의 역할용어를 사용하고 있다.

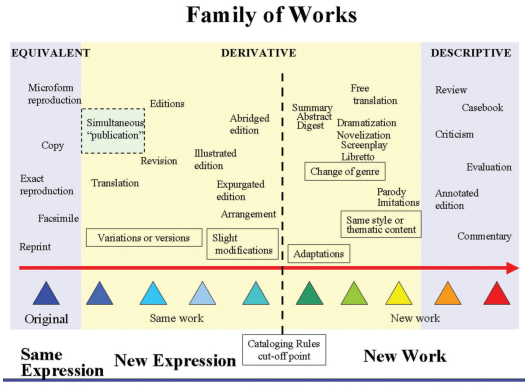
8) 역할용어를 추출하지 않은 단일 저자의 작품임을 표시하기 위해 임시로 부여하였다.

〈표 6〉 초기 역할용어세트(ISTC 기반 표준화)를 각 레코드에 적용한 결과 목록

Author
Author of supplementary text
Author of supplementary text;Creator of other non-text content;Translator;Unspecified
Author of supplementary text;Unspecified
Author;Author
Author;Author of supplementary text
Author;Author of supplementary text;Author of supplementary text
Author;Author of supplementary text;Compiler;Translator;Unspecified
Author;Author of supplementary text;Translator
Author;Author;Author;Translator
Author;Author;Translator
Author;Creator of other non-text content
Author;Creator of other non-text content;Translator
Author;Creator of other non-text content;Unspecified
Author;Editor or reviser
Author;Translator
Author;Translator;Unspecified
Author;Unspecified
Compiler;Creator of other non-text content
Compiler;Translator;Translator
Compiler;Unspecified
Creator of other non-text content;Unspecified;Unspecified
Editor or reviser
Editor or reviser;Unspecified
Translator;Unspecified
Unspecified
Unspecified;Unspecified;Unspecified

는 원저작에서 완전히 새로운 저작이 된다. 예를 들어 〈그림 2〉에서 볼 수 있듯이 원래의 저작에서 '각색'을 했을 경우에는 ② 새로운 저작 - 파생(Dramamatization)에 해당되어 새로운 저작으로 간주되며, 〈표 5〉의 샘플 데이터의 패턴별 처리 과정을 통해 추출된 역할용어 목록에서는 '각본', '다시쓴이' 등이 이에 해당한다.

〈표 7〉은 최종적으로 구축한 역할용어사전으로, 추출된 역할용어를 ISTC에 따라 1차적으로 표준화한 역할용어와 Tillett의 기준을 참조하여 2차적으로 세분화한 역할용어를 정리한 것이다. 예를 들어 각본이나 번안은 Tillett의 기준에 따르면 새로운 저작을 만들어낸 자의 역할을 의미하나 표준화된 역할용어에서는 'Unspecified'로 지정되어 역할이 명확하게 드러나지 않는다. 이에 세분화된 역할용어인 'Renewer'를 부여하였다. 역할용어의 명명은 임의로 이루어졌지만 역할용어에 따라 동일한 저작과 새로운 저작을 구분하는 기준이 될 수 있다. '단일'은 단독 저자가 창작한 작품이므로 'Author'를 부여하였다.



저작구분	Edition(English)	Edition(Korean)
① 동일 저작 (파생관계)	Translation	번역본
	Editions	제n판
	Revision	개정본
	Abridged	축약본
	Illustrated	삽화본
	Expurgated	삭제본
	Arrangement	편집본
② 새로운 저작 (파생관계)	Summary	요약본
	Abstract	초록본
	Digest	요약본
	Free translation	무상 번역본
	Dramatization	각색본
	Novelization	소설화본
	Screenplay	시나리오본
	Libretto	오페라(악극) 대본
	Parody	패러디본
	Imitations	모방본
③ 새로운 저작 (기술관계)	Review	리뷰
	Casebook	사례 모음집
	Criticism	비평
	Evaluation	평가
	Annotated	주석
	Commentary	논평

〈그림 2〉 Tillett(2004)를 토대로 재정리한 서지적 관계에서 동일한 저작/새로운 저작을 구분하는 기준

〈표 7〉 Tillett의 Family of Works를 활용하여 역할용어를 세분화 한 결과

추출된 역할용어	ISTC 기반 초기 역할용어	표준화된 역할용어(우선어)
written by / 공저 / 공지음 / 글 / 만듦 / 작 / 作 / 작곡 / 작시 / 저 / 著 / 著者 / 지은이 / 지음	Author	Author
共譯 / 공옮김 / 번역 / 역 / 譯 / 譯 / 譯者 / 옮긴이 / 옮김 / 옮김이	Translator	Translator
Comics / illustrated by / 그림 / 만화 / 일러스트	Creator of other non-text content	Nontext
엮음 / 편 / 編 / 편역	Editor or reviser	Editor
annotated with critical introduction by / 번역·해설 / 역주 / 譯註 / 註譯 / 해설 / 해설·주석 / 解題	Author of supplementary text	Annotator
원작 / 原作 / 원저 / 原著	Unspecified	Original
retold by / 각본 / 다시쓴이 / 번안 / 번역, 개작 / 편저 / 編著	Unspecified	Renewer
단일	Unspecified	Author
영화	Unspecified	Unspecified[9]

9) '영화'라는 역할용어가 정제 과정 중 확인되어 서지제어번호를 통해 원데이터와 실제 표지정보를 확인하였다. 원래는 존재하지 않는 역할용어로 잘못 입력된 것으로 보인다. 따라서 표준화된 역할용어를 부여하는 것이 적절하지 않다고 판단하여 임의로 'Unspecified'로 분류하였다.

3.3 대표저작자 선정 및 저자명식별자 부여

새롭게 구축한 역할용어사전을 활용하여 역할용어세트를 표준화시키고, Tillett의 기준에 따라 세트별로 역할용어를 비교하여 대표저작자를 선정한 결과는 〈표 8〉과 같다. 역할용어세트는 중복을 제외하고 총 25가지의 형태를 갖고 있는 것으로 나타났으며, 이 중에서 19가지는 대표저작자가 수행한 역할이 하나이므로 역할에 맞는 이름 데이터를 그대로 출력할 수 있다. 그러나 ①, ⑧, ⑨, ⑩, ㉑, ㉔의 6가지 세트에서는 대표저작자의 역할이 두 개 이상 존재하여 사람이 직접 대표저작자를 선정하는 처리를 거칠 수 있게끔 '수동검사(Manual)'를 표기하였다.

수동검사 과정을 포함시킨 이유는 목록 데이터의 입력 오류로 인해 공동의 대표저작자가 아님에도 불구하고 대표저작자가 두 명 이상인 것으로 나타났을 가능성이 있기 때문이다. 대표저작자의 역할을 둘 이상 포함하고 있는 6가지의 역할용어세트의 오류 원인 및 유형을 파악하기 위해 해당 데이터를 직접 살펴본 결과는 이를 뒷받침한다.

대표저작자가 2인 이상인 레코드는 동일한 역할을 수행한 경우(역할: 이름1, 이름2)와 서로 다른 역할을 수행한 경우(역할1: 이름1; 역할2: 이름2)의 두 가지로 분류할 수 있다. 전자는 나열되어 있는 복수의 저자명을 그대로 저작을 식별하는 데 활용한다. 만일 나열되어 있는 저자명을 '이름1+표제'와 '이름2+표제'와 같이 분리하면 저작을 매칭할 때 저자명 및 표제는 같지만 완전히 다른 저작이 동일 저작으로 묶일 수 있기 때문이다.

한편, 서로 다른 역할을 수행했지만 공동 대표저작자로 선정된 후자는 앞서 언급한 바와 같이 대표저작자 추출 과정에서 오류가 발생했을 가능성이 있다. 수동검사가 필요한 역할용어세

〈표 8〉 역할용어세트의 패턴별 대표저작자 선정

역할용어세트	대표저작자	역할용어세트	대표저작자
① Annotator;Annotator;Author	Manual*	⑭ Author;Nontext;Translator	Author
② Annotator;Author	Annotator	⑮ Author;Original	Author
③ Annotator;Author;Editor;Translator;Unspecified	Annotator	⑯ Author;Original;Translator	Author
④ Annotator;Author;Translator	Annotator	⑰ Author;Renewer	Renewer
⑤ Annotator;Nontext;Original;Translator	Annotator	⑱ Author;Translator	Author
⑥ Annotator;Original	Annotator	⑲ Editor;Nontext	Editor
⑦ Author	Author	⑳ Editor;Original	Original
⑧ Author;Author	Manual*	㉑ Editor;Translator;Translator	Manual*
⑨ Author;Author;Author;Translator	Manual*	㉒ Nontext;Original;Renewer	Renewer
⑩ Author;Author;Translator	Manual*	㉓ Original;Renewer	Renewer
⑪ Author;Editor	Author	㉔ Original;Renewer;Renewer	Manual*
⑫ Author;Nontext	Author	㉕ Original;Translator	Original
⑬ Author;Nontext;Renewer	Renewer		

* Manual은 수동으로 검사할 필요가 있는 역할용어세트임을 나타냄.

트 6가지를 살펴보면 ①은 Annotator, ⑧, ⑨, ⑩은 Author, ㉑은 Translator, ㉔는 Renewer 역할용어가 반복하여 나타나며, 이에 해당하는 레코드의 수는 총 18개임을 확인하였다. 대표저작자로서의 역할을 실제로 공동 수행한 레코드는 4개였고 나머지 16개는 공동 대표저작자로 잘못 추출된 레코드였다. 대표저작자 추출 과정에서 오류가 발생한 원인은 공동의 역할을 수행한 사람들을 쉼표(,)를 통해 나열해야 하는데 쌍반점(;)을 사용하거나 '이름 역할' 패턴과 '역할:이름' 패턴을 혼용했기 때문이었으며 오류 레코드의 수는 각각 11건, 5건임을 확인하였다.

다음 단계는 선정된 대표저작자의 역할을 기준으로 삼아 245 필드로부터 대표저작자의 이름을 추출하고 저자명 식별자를 부여하는 과업이다. 본 연구에서 245 필드를 활용한 근본적인 이유는 245 필드가 역할용어를 가장 많이 보유하고 있다는 사실 때문이다. 그러나 역할용어는 동일한 역할임에도 불구하고 각기 다른 형태로 기술되어 있어 표준화 과정을 필요로 하는데, 이는 저자명 또한 마찬가지이다. 단적인 예로 '헤밍웨이', '어네스트 헤밍웨이' 및 'E.M. 헤밍웨이'는 모두 헤밍웨이(Hemingway)를 가리키나 모두 다르게 표현되어 있다. 서로 다른 레코드가 같은 저작인지 파악하기 위한 방법으로 '저자명+표제'의 유사도를 비교하는 본 연구에서는 비교의 정확도를 높이기 위해 통일된 형태의 대표저작자 이름이 필요하다. 따라서 245 필드에서 추출한 대표저작자의 이름과 부출표목(7XX) 필드의 이름을 매칭하여 최종적으로는 표준화된 형태를 갖춘 부출표목의 이름을 저자명 식별자로 활용하였다.

3.3.1 저자명 식별자 부여 - 단일 저자의 작품

단일 저자가 창작한 작품은 기본표목 또는 부출표목을 하나만 사용하며 이들은 표준화된 형태로 입력되므로 그대로 저자명 식별자로 사용하였다. 한편, 단일 저자의 작품이 317건이지만 322개의 표목 데이터가 있는 것으로 확인되어 이를 살펴본 결과, 245 필드에 나타나 있지 않은 원저작자 또는 기관명이 부출표목으로 함께 입력된 오류 레코드가 5건이 있었다. 이와 같은 오류 레코드는 연구자가 직접 원저작자의 표목을 선택하였다.

3.3.2 저자명 식별자 부여 - 복수 저자의 작품

복수 저자가 창작한 작품에서 추출한 대표저작자의 이름을 살펴보면, 다음과 같은 세 가지 패턴이 나타나는 것을 확인할 수 있다. 첫째는 띄어쓰기가 없는 이름, 둘째는 외국인명이나 기관명과 같이 띄어쓰기가 있는 이름이며, 셋째는 공동의 역할을 수행하여 다수의 이름이 나열된 경우이다. 이러한 세 개의 패턴에 수동검사가 필요한 경우를 더하여 총 네 개의 패턴으로 레코드를 분류하였고, 각각의 패턴에 따라 다음과 같이 정제 작업과 언어변환 처리를 수행하였다.[10]

10) 본 연구에서는 대표저작자의 저자명을 찾기 위한 것이므로 로컬필드(9XX)는 제거한 후 작업을 진행하였다. 추후 저자명을 군집화하여 전거데이터를 과업이 진행된다면 로컬필드의 데이터도 데이터 매칭 작업에 포함시켜 정확도를 높이는 데 기여할 수 있을 것으로 기대한다. 언어처리시 한자에서 한글로 변환하는 기능은 엑셀의 기본 기능을 활용, 영어에서 한글로 변환하는 기능은 구글 스프레드 시트에 데이터를 복사하여 붙여넣기 한 다음 구글 번역기 함수를 자동으로 활용할 수 있어 이를 활용하였다.

① 띄어쓰기가 없는 이름
- '김유조', '헤밍웨이'와 같이 띄어쓰기가 없는 형태의 저자명
- 이러한 유형의 패턴은 별도의 처리를 하지 않고 부출표목과 매칭

② 띄어쓰기가 있는 이름
- 대다수가 외국인명, 한글로 표기되지만 띄어쓰기를 사용한 형태인 '어니스트 헤밍웨이'나 한글과 영어가 혼합된 'E.M. 헤밍웨이'와 같은 경우도 포함
- 기관명 데이터도 종종 확인됨
- 이러한 형태는 부출표목 필드의 외국인명 기입 순서인 '성, 이름'과 같은 형태로 변형시키고 구두점을 모두 공백으로 변환함('E.M. 헤밍웨이'에서 '헤밍웨이, E M'으로 변환됨)

③ 다수의 저자명이 나열된 경우
- 저자명을 하나씩 분리한 후 각각 부출표목과 매칭하고 재조합
- 이와 같은 처리를 하는 이유는 만일 나열되어 있는 저자명들을 '이름1+표제'와 '이름2+표제'와 같이 분리하면 '저자명+표제'의 유사도를 비교하는 과정에서 저자명과 표제의 표기는 같지만 실제로는 완전히 다른 저작을 동일한 저작으로 판단할 가능성이 있음

④ 오류로 분류되어 수동검사가 필요한 레코드
- 이러한 유형은 전체 771건의 작품 중 20건으로 별도의 수동검사를 실시
- 역할용어사전 구축 시 사람의 이름이 역할용어로 잘못 추출되어 오류로 표시했던 레코드와 역할용어세트에서 대표저작자의 역할이 두 개 이상이었던 레코드가 이에 해당
- 수동검사로 분류된 레코드는 연구자가 직접 부출표목을 살펴본 후 저자명 식별자를 생성

또한 245 필드 및 표목에서 추출한 대표저작자의 이름은 한글과 영어, 한자가 혼용되어 있어 문자들 간의 유사도를 직접 비교할 수 없으므로 언어를 한글로 통일하는 전처리 과정을 추가하였다. 한자를 한글로 변환하는 데에는 Excel의 기본 기능을 활용하였고, 영어를 한글로 변환하는 데에는 Google 스프레드 시트에서 제공하는 번역 함수를 활용하였다.

데이터 정제 및 언어변환 처리를 진행한 후에는 5가지 저자명 데이터(245, 100, 110, 700, 710)를 대상으로 코사인 유사도(cosine similarity)[11]를 비교하여 최종적으로는 부출표목과 동일한 형태의 저자명 식별자를 모든 레코드에 부여하였다. 245 필드의 저자명과 부출표목의 저자명이 제대로 매칭 되었는지 확인한 결과, 복수 저자의 작품 434건 중에서 426건(98.16%)이 제대로 할당되었음을 알 수 있었다. 오류가 발생한 원인은 목록 데이터가 잘못된 경우와 언어변환 과정이 제대로 이루어지지 않은 경우의 두 가지였다. 언어변환 과정에서 발생한 오류는 5건으로, 245 필드의 'Mr. Sun'이 '해씨'로

11) 용어(term)는 하나의 차원을 구성하고 문서(document)는 각 용어가 등장하는 빈도에 따라 벡터 값을 가진다. 코사인 유사도(cosine similarity)는 두 벡터 간 각도의 코사인 값을 이용해 벡터 간 유사도를 비교하는 방법으로 정보 검색 및 텍스트 마이닝 분야에서 두 문서의 유사도를 측정하는데 널리 사용되고 있다(Singhal, 2001).

번역이 되어 부출표목의 '선진호'와 매칭이 이루어지지 않은 사례, 'Whisp, Kennilworthy'와 같이 번역기로는 변환이 되지 않는 사례를 예로 들 수 있다. 나머지 3건의 오류는 표목의 ▼a에 이름과 역할용어가 같이 기입되었거나 대표저작자의 이름이 부출표목으로 아예 입력되어 있지 않은, 목록 데이터 자체의 문제로 인해 나타난 것으로 확인하였다.

4. 저자명 + 표제 유사도 비교를 통한 저작군집화

저자명과 표제를 활용하여 저작을 군집화 하는 데에는 몇 가지 방법이 있겠지만 본 연구에서는 다음과 같은 두 가지 방법에 주목하였다. 첫 번째는 저자명과 표제를 결합한 '저자명+표제' 세트의 유사도를 비교하여 레코드 간 유사도가 일정 수준 이상일 때 동일한 저작으로 판단하는 방법이다. 두 번째는 저자명의 유사도와 표제의 유사도를 개별적으로 계산하고 각각의 유사도가 일정 수준 이상인 레코드들을 동일한 저작으로 판단하는 방법이다. 문자열 간의 유사도를 측정하는 기준으로는 정보검색 분야에서 두 문서의 유사성을 파악하는 데 널리 사용하고 있는 코사인 유사도를 채택하였다.

먼저 '저자명+표제' 세트의 유사도를 계산하는 방법을 살펴본 결과, 저자명과 표제가 서로에게 미치는 영향이 커서 전체적인 유사도가 오히려 떨어지는 것으로 나타났다. 예를 들어 동일한 저자의 다른 작품들인 'Shakespeare, William+로미오와 줄리엣'과 'Shakespeare, William+햄릿'의 유사도가 88퍼센트를 넘는 것으로 나타났고, 시리즈물로 부제만 다른 'Rowling, J. K.+해리포터:혼혈 왕자'와 'Rowling, J. K.+해리포터:아즈카반의 죄수' 또한 유사도가 85퍼센트에 달하는 것으로 나타났다. 저자명과 표제를 조합한 세트의 유사도를 비교하는 방식은 이처럼 저자명이 완전히 일치하는 경우 제목의 차이가 상당함에도 불구하고 유사도가 매우 높게 나타나는 한계점이 있다. 또한 반대로 표제 간 유사도가 매우 높지만 저자가 다른 상황에서도 동일한 문제가 발생할 수 있다.

따라서 저자명과 표제의 유사도가 서로에게 미치는 영향을 최소화하기 위해 저자명끼리의 유사도와 표제끼리의 유사도를 각각 계산하고, 두 개의 유사도 값이 일정 수준 이상일 때 동일한 저작으로 판단하는 방법을 채택하게 되었다. 그러나 이 방법 또한 유사도 수준을 설정하는 데 대한 논의가 필요하므로 데이터 정제 조건을 달리하여 최종 결과물의 정확도를 비교하였다.

4.1 저자명 + 표제 유사도 비교를 위한 사전 작업

저자명과 표제, 각각의 유사도를 계산하기에 앞서 데이터를 정제하였다. 시리즈물은 표제 사항의 ▼b에 부제를 입력하는 경우가 있으므로 본표제(▼a)와 부제(▼b)를 결합하였고, 권차 및 권제(▼n, ▼p)는 제외하였다. 또한 245 필드의 저자명과 표목의 저자명을 매칭할 때와 마찬가지로 표제의 한자를 모두 한글로 변환하였으나 영문 표제는 별도의 번역 과정을 거치지 않았다. 그러나 본표제가 영어이면서 대등표제가 있는 경우에는 본표제를 대등표제로 변환하였다. 이와 같은 정제 과정은 표제를 최대한 한

글로 표기하여 유사도 비교의 정확성을 높이기 위함이다.

목록을 작성하면서 본표제를 입력할 때 첫 문자열이 관제나 관사면 원괄호를 사용한다. 관제는 표제를 비교하는데 영향을 줄 수 있으며 괄호만 제거할 것인지 괄호 안의 내용까지 모두 제거할 것인지 결정할 필요가 있다. 245 필드는 관제를 표제와 함께 부출할 것인지 여부를 지시기호로 설정하는데, 관제를 부출하는 경우는 관제에 의미를 둔 것으로 판단하여 괄호만 제거하고 괄호 안의 내용은 그대로 유지하며, 부출하지 않을 시에는 괄호 안의 데이터까지 모두 제거하였다. 괄호는 관제 이외에도 표제 중간에서 '패러디(Parody)'와 같이 단어의 의미를 명확히 하는 데 사용되기도 하는데, 이때에도 괄호 및 괄호 안의 문자열을 모두 제거하였다. 하지만 관제나 괄호의 처리에 따라 정확도가 달라질 수 있음을 고려하여 표제를 다음과 같은 세 가지 형태로 나누어 저작 단위 군집화의 정확도를 각각 비교하였다: 1) 본래의 문자열 그대로 사용가능한 본표제와 부제의 조합, 2) 관제를 부출하는 레코드의 경우 괄호는 제거하고 괄호 내의 데이터는 유지하며, 표제 중간에 사용된 괄호 및 괄호 안의 문자열은 모두 제거, 3) 관제 부출 여부에 관계없이 모든 괄호와 그 안의 내용을 제거.

문자열을 비교하는 과정에서 문자열 간 띄어쓰기 또한 유사도에 영향을 줄 수 있으며, 이를 고려하여 공백을 제거한 표제와 제거하지 않은 표제의 두 형태로 한 번 더 분리하였다. 이러한 과정을 거쳐 최종적으로 관제 및 괄호 처리, 공백 처리의 차이에 따른 총 6가지 형태의 표제를 대상으로 표제 간 유사도를 계산하였다. 저자명은 앞선 단계에서 이미 표목을 기준으로 삼아 표준화된 형태를 갖춘 상태였기에 별도의 정제 과정을 거치지 않고 유사도를 계산하였다.

4.2 저자명 + 표제 유사도 비교 결과

저자명 유사도와 표제 유사도를 개별적으로 계산한 결과를 유사도 순으로 정렬하여 직접 살펴보았으며, 각 유사도가 80퍼센트 이상일 때 비교적 높은 수준의 정확도로 군집화가 잘 되는 것을 확인하였다. 이에 따라 표제 유사도 80퍼센트, 저자명 유사도 80퍼센트 이상을 조건으로 분리 집합(disjoint sets)[12] 연산 알고리즘을 적용하여 레코드를 군집화 하였고 그 결과는 〈표 9〉와 같다.

군집화 결과는 P, PE, F, FT의 총 네 가지 항목으로 분류하였다. P(Pass)는 저작 단위 군집화가 잘 이루어졌음을 의미하며, PE(Pass but having Error)는 군집화가 잘 이루어졌으나 데이터 자체에 문제가 있음을 의미한다. 예를 들어 앞서 245 필드와 표목 필드의 저자명 매칭 과정에서 나타난 '김유조 해설'과 같이 데이터가 잘못 입력된 사례, 총서를 제외하는 과정에서 제외되지 않은 사례, 본표제가 외국어로 표기되었으나 대등표제가 없어서 한글 표제와 매칭이 이루

12) 분리 집합(disjoint sets)은 서로소 집합이라 불리기도 하며 서로소라는 명칭에서 알 수 있듯이 공통 원소가 없는, 즉 교집합이 없는 두 개 이상의 집합을 의미한다(Cormen, Leiserson, Rivest, & Stein, 2009). 분리 집합을 연산하기 위한 방법으로는 두 개의 집합을 하나의 집합으로 합치는 합집합(union)과 특정 원소가 속한 집합을 찾는 찾기(find)가 대표적이다. 본 연구에서는 각각의 레코드를 하나의 집합으로 설정하고, 저자명과 표제의 유사도 기준에 따라 동일성을 판단하여 하나로 합치는 합집합 연산을 적용하는 방식으로 군집화를 수행하였다.

〈표 9〉 저자명 유사도 80퍼센트, 표제 유사도 80퍼센트 이상의 조건에서 분리집합 알고리즘을 적용하여 레코드를 군집화 한 결과

조건	항목	단일(N=317)	복수(N=454)	전체(N=771)
1-1 표제에 별도의 처리를 하지 않음	P*	305 (96.21)	376 (82.82)	681 (88.33)
	PE**	7 (2.21)	12 (2.64)	19 (2.46)
	F***	5 (1.58)	61 (13.44)	66 (8.56)
	FT****	0 (0)	5 (1.10)	5 (0.65)
1-2 1-1의 조건에서 문자열 간 공백 제거	P	306 (96.53)	395 (87.00)	701 (90.92)
	PE	7 (2.21)	13 (2.86)	20 (2.59)
	F	4 (1.26)	41 (9.03)	45 (5.84)
	FT	0 (0)	5 (1.10)	5 (0.65)
2-1 부출한 관제의 괄호 제거, 표제 중간에 사용된 괄호와 내용 모두 제거	P	305 (96.21)	379 (83.48)	684 (88.72)
	PE	7 (2.21)	12 (2.64)	19 (2.46)
	F	5 (1.58)	58 (12.78)	63 (8.17)
	FT	0 (0)	5 (1.10)	5 (0.65)
2-2 2-1 조건에서 문자열 간 공백 제거	P	306 (96.53)	396 (87.22)	702 (91.05)
	PE	7 (2.21)	13 (2.86)	20 (2.59)
	F	4 (1.26)	40 (8.81)	44 (5.71)
	FT	0 (0)	5 (1.10)	5 (0.65)
3-1 관제 부출 여부와 상관없이 모든 괄호와 내용 제거	P	305 (96.21)	382 (84.14)	687 (89.11)
	PE	7 (2.21)	12 (2.64)	19 (2.46)
	F	5 (1.58)	55 (12.11)	60 (7.78)
	FT	0 (0)	5 (1.10)	5 (0.65)
3-2 3-1 조건에서 문자열 간 공백 제거	P	306 (96.53)	400 (88.11)	706 (91.57)
	PE	7 (2.21)	13 (2.86)	20 (2.59)
	F	4 (1.26)	36 (7.93)	40 (5.19)
	FT	0 (0)	5 (1.10)	5 (0.65)

* P(Pass): 군집화가 잘 이루어진 레코드
** PE(Pass but having Error): 군집화가 잘 이루어졌으나 오류가 있는 레코드
*** F(Fail): 군집화가 이루어지지 않은 레코드
**** FT(Failure of Translation): 저자명 표준화 과정에서 번역문제로 군집화 오류가 발생한 레코드

어지지 않은 사례 등이 해당된다. F(Fail)는 군집화가 정상적으로 수행되지 않은 레코드를 나타내며, 원인을 살펴본 결과 부제의 문자열이 너무 길어 원작과 유사도 비교가 잘 이루어지지 않은 사례가 대부분을 차지하였다. 이 외에도 80퍼센트의 유사도 수준으로 매칭되지 못한 레코드들이 F로 분류되었다. 마지막으로 FT(Failure of Translation)로 분류된 레코드는 저자명 표준화 과정에서 번역기를 이용할 때 제대로 번역되지 않아 군집화가 이루어지지 않은 사례이다.

정상적으로 군집이 형성되지 않는 문제의 원인을 알아보기 위해 F로 분류된 레코드들을 살펴본 결과 부제(▼b)로 인해 표제 간 유사도를 제대로 비교하지 못하는 데에서 대부분의 문제가 발생하고 있음을 확인하였다. 〈표 9〉에서 91.57퍼센트로 가장 높은 정확도를 나타내고 있는 3-2 조건

의 군집화 결과를 살펴보면, F로 분류된 레코드는 40건(5.19%)이며 이들 중 37건의 레코드가 부제(▼b)의 문자열이 길어서 원작과 매칭이 이루어지지 않은 경우이다. 예를 들어 "노인과바다"와 "노인과바다:큰글씨책"의 경우, 쌍점(:) 이후에 나타나는 "큰글씨책"이라는 문자열로 인해 표제 간 유사도가 떨어져 군집이 형성되지 않는 것이다. 나머지 3건의 레코드는 표제 간 유사도가 80퍼센트보다 높아져야 하는 레코드가 1건, 80퍼센트보다 낮아져야 하는 레코드가 1건, 저자명 유사도가 80퍼센트보다 높아져야 하는 레코드가 1건이었다. 이러한 문제점에 대한 해결책으로는 부제(▼b)를 제거한 뒤 표제 간 유사도를 비교하거나 유사도 수준을 변경하는 방안이 있어 두 가지 방법을 각각 적용해보았다.

먼저 부제(▼b)를 제거하고 본표제(▼a) 간의 유사도를 비교하는 방법은 레코드의 특성에 따라 군집화 결과가 매우 달라져 채택하지 못하였다. 실험 데이터에서 "노인과 바다"를 비롯한 "로미오와 줄리엣", "오만과 편견", "햄릿"은 한 편으로 완결되는 저작인 반면에 "해리포터"는 본편이 7가지의 다른 책들로 구성된 시리즈 작품이다. 앞의 네 개 저작은 판(edition)과 관련된 사항이나 부연 설명을 부제에 입력하였기 때문에 본표제만을 비교하는 방법이 군집화 정확도를 상당히 높일 수 있다. 그러나 시리즈 작품인 "해리포터"는 본표제에 "해리포터"를 입력하고 부제에는 시리즈 각 편의 이름을 입력하는 방식으로 목록을 작성한다. 따라서 시리즈 한 편만으로도 하나의 저작을 형성할 수 있기에 이 경우 본표제만을 비교하면 시리즈 전체가 해리포터라는 하나의 저작으로 묶일 가능성이 높다.

다음으로 군집화 정확도를 최대치로 끌어올리기 위해 표제 간 유사도 수준을 75퍼센트, 70퍼센트로 더 세분화하여 측정해보았다. 결과는 다음의 〈표 10〉에서 확인할 수 있듯이, 표제 유사도가 80퍼센트 미만으로 내려갈 경우 군집화가 정상적으로 이루어진 사례가 706건(91.57%)

〈표 10〉 3-2 조건에서 표제 간 유사도를 80, 75, 70퍼센트로 세분화하여 적용한 결과

조건 및 유사도		항목	단일(N=317)	복수(N=454)	전체(N=771)
3-2	80(%)	P	306 (96.53)	400 (88.11)	706 (91.57)
		PE	7 (2.21)	13 (2.86)	20 (2.59)
		F	4 (1.26)	36 (7.93)	40 (5.19)
		FT	0 (0)	5 (1.10)	5 (0.65)
	75(%)	P	305 (96.21)	396 (87.22)	701 (90.92)
		PE	7 (2.21)	12 (2.64)	19 (2.46)
		F	5 (1.58)	41 (9.03)	46 (5.97)
		FT	0 (0)	5 (1.10)	5 (0.65)
	70(%)	P	305 (96.21)	396 (87.22)	701 (90.92)
		PE	7 (2.21)	11 (2.42)	18 (2.33)
		F	5 (1.58)	42 (9.25)	47 (6.10)
		FT	0 (0)	5 (1.10)	5 (0.65)

에서 701건(90.92%)으로 줄어든 것을 확인할 수 있다. 또한 유사도가 75퍼센트일 때와 70퍼센트일 때는 군집 결과의 차이가 아주 미미하다. 표제 유사도를 조정함으로써 나타난 군집화 결과의 차이는 레코드의 수 측면에서는 큰 변화가 없는 것처럼 보일 수 있지만 실제 데이터를 확인해 본 결과 유사도를 조정하는 것에 대한 큰 의미를 발견하였다.

표제 간 유사도가 80퍼센트일 때와 70퍼센트로 낮추었을 경우 총 25건의 레코드에서 군집 결과가 다르게 나타났다. 25건 중 9건은 표제 간 유사도 수준이 80퍼센트일 때는 부제가 길어서 매칭이 되지 않았다가 유사도가 70퍼센트로 낮아지면서 군집이 제대로 형성되는 긍정적인 사례이다. 하지만 나머지 16건은 기존에 잘 매칭(P, PE) 되었으나 유사도 수준을 낮추면서 오히려 매칭이 잘못 이루어지는 부정적인 사례로 확인되었다. 부제(▼b)로 인해 군집이 제대로 형성되지 않는 문제를 해결하기 위해서 잘 매칭되었던 레코드를 오류로 만드는 것은 어불성설이다. 이러한 결과로 볼 때 표제의 유사도 수준은 80퍼센트 이상으로 설정하는 것이 가장 적합한 것으로 판단된다. 또한 부제가 길어서 군집이 이루어지지 않는 문제를 해결하기 위해서는 본 연구에서 시도한 방법이 아닌 다른 방안을 모색할 필요가 있다.

5. 결 론

방대한 양의 서지레코드를 FRBR의 저작 단위로 재구조화 하는 작업은 소요되는 시간과 비용이 매우 높아 기계화된 처리가 불가피하며 군집화를 자동화할 수 있는 방안이 필요하다. 본 연구는 이러한 필요에 따라 서지레코드를 저작 단위로 군집화 하는 하나의 시도로 시작되었다. KORMARC 서지레코드를 FRBR의 저작 단위로 군집화하기 위해서는 레코드마다 저작 식별자를 부여해야 하며, 저작의 식별 작업은 레코드의 저자명과 표제를 비교함으로써 수행된다. 이때 다수의 사람에 의해 창작된 저작의 레코드는 2개 이상의 저자명 필드를 보유하므로 대표저작자를 선정하는 과정을 필요로 한다. 본 연구는 다수의 저자명 필드가 있을 때 대표저작자를 선별하기 위한 방법으로 각 저자의 역할을 비교하고자 저자역할용어사전을 구축하고, 이를 활용하여 저작군집화를 실험하였다. 그 결과 저작을 창작하는데 기여한 사람이 다수인 레코드를 대상으로 선행연구들에 비해 보다 높은 정확도의 군집화 결과물을 얻을 수 있었다. 저작 단위의 군집화 방안과 연구결과를 요약하면 다음과 같다.

첫째, 245 필드에 입력된 역할용어를 대상으로 ISTC 및 Tillett의 Family of Works를 참조하여 역할용어를 표준화시켜 역할용어사전을 구축하고, 각 레코드 별로 대표저작자를 선정하는 근거로 활용하였다. 본 연구에서는 다양한 형태로 기입된 역할용어를 표준화된 용어로 변경할 수 있는 수준의 사전을 구축하였으나 국립중앙도서관의 전체 레코드를 대상으로 역할용어를 추출하여 시소러스(Thesaurus) 형태로 구축하여 관리한다면 저작군집화 이외에 목록을 작성하는 작업에도 활용할 수 있을 것이다. 또한 본 연구에서는 레코드의 역할용어세트에서 대표저작자를 선정하는 기준에 한계가 존재하므로 대표저작자를 선정하기 위한 합의된 기준을 마

련하기 위한 연구가 이어져야 할 것이다.

둘째, 저자명과 표제를 조합하지 않고 저자명 유사도와 표제 유사도를 각각 비교하는 방식으로 저작을 식별하였다. 유사도를 비교하기에 앞서 저자명은 245 필드의 데이터가 다양한 형태로 입력되어 있음을 고려하여 기본표목 또는 부출표목의 ▼a로 표준화시켰으며, 표제는 정제 방식을 달리하여 군집화 정확도를 최대치로 끌어올리고자 하였다. 그 결과 표제 정보에 포함된 모든 괄호 및 괄호 안의 내용, 그리고 공백을 제거하고 80퍼센트의 유사도 수준을 설정했을 때 군집화 정확도가 가장 높게 나타나는 것을 확인하였다.

셋째, 저작 단위의 군집이 제대로 형성되지 않는 주요 원인은 표제에서 비롯됨을 확인하였다. 특히 부제(▼b)의 문자열이 길어 유사도가 크게 떨어지는 사례가 대부분이었으며, 부제를 제외한 본표제(▼a)의 유사도를 비교하거나 유사도 수준을 조정하는 방법만으로는 문제를 해결하기 어려웠다. 부제(▼b)로 인해 군집이 제대로 형성되지 않는 문제를 해결하는 방안으로는 본 연구의 방법을 통해 1차적으로 군집화를 실시한 후, 저작의 대표 제목에서 부제에 해당하는 문자열을 제거하고 유사도를 한 번 더 비교하는 2차 군집화 방법을 고려해 볼 수 있을 것이다.

이외에도 본 연구의 과정에서 나타난 문제점들을 해결하거나, 오류를 최소화하는 방안을 모색할 필요가 있다. 예를 들어, 저자명의 표준화 과정은 서로 다른 언어로 쓰인 저자명을 비교하기 위한 언어 변경을 필요로 한다. 본 연구는 언어 변경의 수단으로 번역 기능을 사용하였으나 '선진호, Mr. Sun'의 사례와 같이 음차(音借)가 필요하지만 단어의 사전상 정의(Sun, 해)로 변경하는 오류, '케닐워디, Kennilworthy'와 같이 사전에 등재된 단어가 아니어서 음차가 전혀 이루어지지 않는 오류가 나타났다. 따라서 언어를 변경할 때 번역이 아닌 로마자 표기법(Romanization)을 활용하여 데이터의 언어를 음차하는 방안을 고려해 볼 수 있다.

현재의 서지레코드 구조를 다른 형태로 변경하는 과업에서 직면하는 문제들은 대부분 데이터의 품질 문제와 관련되어 있다. 본 연구도 비교적 품질관리가 잘 이루어지고 있는 국립중앙도서관의 데이터를 사용하였지만 몇 가지 오류와 문제를 발견하였고, 데이터 오류가 있는 상황에서 최대한의 정확도를 끌어내기 위한 방법을 모색하는 시도에서 비롯되었다. 따라서 가장 우선적으로 고려할 사항은 서지데이터의 품질을 최대한 일관되게 고수준으로 유지하는 것이다.

본 연구는 KORMARC으로 작성된 서지레코드를 FRBR로 구조화하기 위한 하나의 실험적 연구로 기존의 연구와는 차별화된 저작 단위의 군집화 방안을 마련하였다. 하지만 실제 디지털 도서관에서 이용자들에게 제공하는 구현형 단위의 디스플레이에 대한 내용을 포함하고 있지 않다. 엄청난 비용과 시간을 들여 저작 단위의 군집화를 시도하는 이유는 검색 결과를 저작 단위로 디스플레이함으로써 이용자에게 FRBR의 유용성을 제공하기 위함이다. 따라서 표현형, 구현형까지 확장하기 위한 후속 연구가 이루어질 필요가 있다. KORMARC 외에도 국내에 적용하고 있는 다른 형태의 MARC 데이터를 대상으로 한 연구도 필요하다. 궁극적으로는 저작 군집화 대상을 국립중앙도서관의 서지데이터뿐만 아니라 다른 기관이 가진 서지

데이터를 모두 포함하여 전체적인 서지세계를 아우를 수 있어야 한다. 중요한 점은 저작 군집화 시도에 앞서 데이터의 품질을 충분히 검증하여 데이터와 품질 자체를 최대로 높인 상태에서 과업이 이루어져야 한다는 것이다. 마지막으로 기존의 레코드를 정비하여 전거데이터, 특히 통일표제(uniform title)를 적용시킬 수 있다면 한층 더 저작 군집화의 정확도를 높이고 자동화된 알고리즘을 구현하기 좀 더 수월해질 것으로 기대한다.

참 고 문 헌

국립중앙도서관 (2014). 한국문헌자동화목록 - 통합서지용. Retrieved from http://www.nl.go.kr/common/jsp/kormarc_2014/index.html

김순희, 이성숙 (2005). FRBR 모형의 서지적 관계에 관한 연구. 사회과학연구, 16, 25-47.

김정현 (2007). 서지적 관계를 기반으로 한 한국어 도서의 저작유형 분석. 한국도서관·정보학회지, 38(3), 183-200. https://doi.org/10.16981/kliss.38.3.200709.183

김정현 (2015). 사서오경의 서지적 관계 특성에 따른 FRBR 적용에 관한 연구. 한국도서관·정보학회지, 46(2), 317-336. https://doi.org/10.16981/kliss.46.2.201506.317

김정현, 이성숙, 이유정 (2015). KORMARC 서지레코드의 FRBR 알고리즘 개발에 관한 연구. 한국도서관·정보학회지, 46(1), 1-23. https://doi.org/10.16981/kliss.46.1.201503.1

김현희, 유영준, 박서은 (2007). FRBR 모형의 KORMARC 데이터베이스로의 적용 가능성에 대한 실험적 연구: 음악자료를 중심으로. 한국도서관·정보학회지, 38(2), 185-202. https://doi.org/10.16981/kliss.38.2.200706.185

노지현 (2008). KORMARC 레코드에 대한 FRBR 모델의 적용 실험. 한국도서관·정보학회지, 39(2), 291-312. https://doi.org/10.16981/kliss.39.2.200806.291

박지영 (2009). 자원의 기술과 접근(RDA). 국립중앙도서관연구소 웹진, 40, 1-23. Retrieved from https://wl.nl.go.kr/webzine/publish/krili/200907_02/pdf/policy01_0731.pdf

이미화, 정연경 (2008). 저작 클러스터링 분석을 통한 FRBR의 목록 적용에 관한 연구. 정보관리학회지, 25(3), 65-82. https://doi.org/10.3743/KOSIM.2008.25.3.065

이성숙, 이현주 (2013). 한국전통음악의 서지적 관계 특성에 따른 FRBR 모형 적용방안. 사회과학연구, 24(2), 399-421. https://doi.org/10.16881/jss.2013.04.24.2.399

Cormen, T. H., Leiserson, C. E., Rivest, R. L., & Stein, C. (2009). Introduction to algorithms (3rd ed.). Cambridge, MA: The MIT Press.

Hickey, T. B., & Toves, J. (2009, August). FRBR work-set algorithm: version 2.0. dublin, ohio:

OCLC online computer library center, Inc. Retrieved from http://www.oclc.org/research/activities/past/orprojects/frbralgorithm/2009-08.pdf

Lee, H., & Park, Z. (2012). FRBRizing bibliographic records focusing on identifiers and role indicators in the Korean cataloging environment. Cataloging & Classification Quarterly, 50(5-7), 688-704. http://dx.doi.org/10.1080/01639374.2012.681599

Library of Congress (2004). FRBR display tool version 2.0. Retrieved from http://www.loc.gov/marc/marc-functional-analysis/tool.html

Singhal, A. (2001). Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 24(4), 35-43.

Tillett, B. B. (1987). Bibliographic relationships: Toward a conceptual structure of bibliographic information used in catalog. Los Angeles: University of California. 재인용: 김순희, 이성숙 (2005). FRBR 모형의 서지적 관계에 관한 연구. 사회과학연구, 16, 25-47.

Tillett, B. B. (2004). What is FRBR? a conceptual model for the bibliographic universe. Retrieved from https://www.loc.gov/cds/downloads/FRBR.PDF

• 국문 참고문헌에 대한 영문 표기
(English translation of references written in Korean)

Kim, H., Yoo, Y., & Park, S. (2007). An experimental study on the FRBR model adaptation to KORMARC database: Focusing on music materials. Journal of Korean Library and Information Science Society, 38(2), 185-202. https://doi.org/10.16981/kliss.38.2.200706.185

Kim, J. (2007). An analysis on the work types of Korean books based bibliographical relationship. Journal of Korean Library and Information Science Society, 38(3), 183-200. https://doi.org/10.16981/kliss.38.3.200709.183

Kim, J. (2015). A study on the adoption of the FRBR according to the bibliographic relationships of Five Classics and Four Books. Journal of Korean Library and Information Science Society, 46(2), 317-336. https://doi.org/10.16981/kliss.46.2.201506.317

Kim, J., Lee, S., & Lee, Y. (2015). A study on the development of FRBR algorithm for KORMARC bibliographic record. Journal of Korean Library and Information Science Society, 46(1), 1-23. https://doi.org/10.16981/kliss.46.1.201503.1

Kim, S., & Lee, S. (2005). A study on bibliographic relationships of the FRBR model. Journal of Social Science, 16, 25-47.

Lee, M., & Chung, Y. (2008). A study of FRBR implementation to catalog by using work

clustering. Journal of the Korean Society for Information Management, 25(3), 65-82. https://doi.org/10.3743/KOSIM.2008.25.3.065

Lee, S., & Lee, H. (2013). A study on the adoption of the FRBR model according to the bibliographic relationships of Korean classical music. Journal of Social Science, 24(2), 399-421. https://doi.org/10.16881/jss.2013.04.24.2.399

National Library of Korea (2014). Korean machine readable cataloging format - integrated format for bibliographic data. Retrieved from http://www.nl.go.kr/common/jsp/kormarc_2014/index.html

Park, J. (2008). Resource Description and Access (RDA). Korea Research Institute for Library and Information, 40, 1-23. Retrieved from https://wl.nl.go.kr/webzine/publish/krili/200907_02/pdf/policy01_0731.pdf

Roh, J. (2008). An application of FRBR model to KORMARC records. Journal of Korean Library and Information Science Society, 39(2), 291-312. https://doi.org/10.16981/kliss.39.2.200806.291

Tillett, B. B. (1987). Bibliographic relationships: Toward a conceptual structure of bibliographic information used in catalog. Los Angeles: University of California. Quoted in Kim, S., & Lee, S. (2005). A study on bibliographic relationships of the FRBR model. Journal of Social Science, 16, 25-47.