

공공도서관 동시 대출 도서의 주제 연관성 분석 연구*

A Study on the Topical Associations of Simultaneously Borrowed Books
in Public Libraries

강 우 진 (Woojin Kang)**
정 인 영 (In Yeong Jeong)***
이 종 육 (Jongwook Lee)****

〈 목 차 〉

- | | |
|------------|------------|
| I. 서 론 | IV. 연구 결과 |
| II. 이론적 배경 | V. 요약 및 결론 |
| III. 연구방법 | |

요 약: 공공도서관 대출데이터를 활용하여 이용자의 정보 이용행위를 이해하려는 노력이 꾸준히 이어지고 있다. 본 연구에서는 공공도서관 이용자가 동시에 대출한 도서들의 주제 분야를 살펴보고, 이를 간의 연관성을 파악하고자 하였다. 이를 위해 2019년에 이루어진 대출 도서 984,790권의 한국십진분류기호를 활용하여 같은 날 동일한 이용자에 의해 대출된 도서 목록의 집합인 동시에 대출 건수 22,443,699에 대해 ITEM2VEC 기법을 적용하여 분류기호를 벡터로 변환하였다. 다음으로 연관성이 높은 분류기호를 10개씩 추출하였으며, 총 522개의 분류기호를 활용하여 네트워크를 생성하였다. 네트워크에서는 15개 커뮤니티를 식별하였으며, 커뮤니티별 주요 특성을 파악하였다. 15개의 커뮤니티 가운데 두 개 이상의 주류로 구성된 커뮤니티에서는 요목 수준에서 의미 있는 주제적 연관성을 파악할 수 있었다. 본 연구는 이용자의 도서 이용행위에 기초하여 함께 대출될 가능성이 있는 자료의 주제를 파악한 것으로 도서관 장서 구성 및 배치, 관련 주제 분야 자료 추천, 분류표 개정 등에 유익한 시사점을 제공할 수 있을 것이다.

주제어: 공공도서관, 동시 대출 도서, 주제 연관성, 대출데이터

ABSTRACT: There has been research to understand users' information behaviors using book circulation data of public libraries. In this study, we examined the subject areas of books simultaneously borrowed by users of public libraries and aimed to identify the relationships among the subject areas. To accomplish this, we utilized the Korean Decimal Classification codes of 984,790 loaned books in 2019 to transform the lists of concurrently borrowed books, totaling 22,443,699 records, by the same users on the same day, into vectors using the ITEM2VEC technique. Next, we extracted ten highly related classification codes for each classification code, utilizing a total of 522 classification codes to create a network. We identified 15 communities within this network and examined the characteristics of each community. Among the 15 communities, those consisting of two or more main classes allowed us to identify meaningful thematic associations. This study, grounded in users' book usage behaviors, has suggested the topics of books that could be borrowed together. The findings offer valuable insights for library collection development and placement, recommending related subject materials, and revising classification systems.

KEYWORDS: Public Library, Simultaneously Borrowed Books, Topical Association, Circulation Data

* 본 논문은 2022년도 한국도서관·정보학회 춘계 학술대회에서 발표된 논문을 수정·보완하여 작성하였음.
** 경북대학교 문헌정보학과 박사과정(rkddnwls1234@knu.ac.kr / ISNI 0000 0005 0659 7247) (제1저자)
*** 경북대학교 일반대학원 기록학전공 석사과정(jiy1138@knu.ac.kr / ISNI 0000 0005 1269 2320) (공동저자)
**** 경북대학교 문헌정보학과 부교수(jongwook@knu.ac.kr / ISNI 0000 0004 6830 6145) (교신저자)

• 논문접수: 2023년 8월 23일 • 최초심사: 2023년 8월 24일 • 게재확정: 2023년 9월 11일
• 한국도서관·정보학회지, 54(3), 33-55, 2023. <http://dx.doi.org/10.16981/kliss.54.3.202309.33>

※ Copyright © 2023 Korean Library and Information Science Society
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives 4.0
(<https://creativecommons.org/licenses/by-nc-nd/4.0/>) which permits use, distribution and reproduction in any medium, provided that
the article is properly cited, the use is non-commercial and no modifications or adaptations are made.

I. 서 론

공공도서관은 누구나 각종 지식정보자원에 자유롭게 접근할 수 있도록 국가, 지방자치단체, 교육청 등에 의해 설립된 기관이며 공중의 생활권역 내에서 정보 이용, 독서 문화 활동 및 평생교육 등을 위한 정보센터로 역할을 담당한다(도서관법 2조; IFLA, 1994). 공공도서관 이용자는 연령, 직업, 경제적 상황 등이 매우 다양하며, 이들의 요구 또한 매우 다양한 편이다(김성희, 이승민, 2022; 심지영, 2021). 이에 따라 도서관은 지역사회 내에서 다양한 이용자의 요구를 반영하여 적절한 서비스를 제공하기 위해 노력해야 한다. 이에 공공도서관에서는 이용자의 정보요구 및 이용행위에 대한 이해를 기반으로 구체적이고 전략적인 접근을 통해 서비스 제공 방안을 마련할 필요가 있다.

여러 선행연구에서 공공도서관 이용자의 인식을 조사(김성희, 이승민, 2022; 조재인, 2019)하거나 장서 데이터를 분석(심지영, 2021; 조재인, 2019)하여 도서관의 현황과 이용자의 정보 이용 행위를 조사한 바 있다. 특히 공공도서관 대출데이터를 활용한 연구에서는 도서관 이용자의 독서 경향(이순영, 이수상, 2021), 도서 이용 패턴(심지영, 2021) 등을 조사하여 이에 근거한 장서 및 프로그램 개발의 필요성을 제안하였다.

공공도서관 대출데이터를 활용한 연구는 도서관 빅데이터 분석활용 체계가 구축되면서 활발하게 이루어졌는데, 이러한 체계가 구축된 과정을 살펴보면 다음과 같다. 우선 2014년에는 국립중앙도서관과 한국과학기술정보연구원이 빅데이터 기반 도서관 서비스 환경 기반을 조성하기 위하여 도서관 빅데이터 플랫폼 구축을 위한 계획을 수립하였다. 이에 기초하여 2016년에는 ‘사서 의사결정지원 서비스 솔로몬’, ‘도서관 정보나루’(<https://www.data4library.kr/>) 등이 개발되었다(온정미, 박성희, 2020; 한국과학기술정보연구원, 2015). 2019년에는 한국문화정보원이 ‘문화 빅데이터 플랫폼’(<https://www.culture.go.kr/bigdata/>)을 구축하고, 국립중앙도서관이 이 플랫폼에 참여하면서 전국 공공도서관 대출데이터를 제공하고 있다. 이에 공공도서관 대출데이터를 활용한 국내 대다수 연구는 ‘도서관 정보나루’ 또는 ‘문화 빅데이터 플랫폼’을 통해 데이터를 활용하였다.

본 연구에서는 공공도서관 대출데이터를 활용하여 이용자의 도서 이용행위를 부분적으로 확인하고자 한다. 특히, 서로 다른 주제가 함께 대출된 도서를 통해 주제 간 연관성을 살펴보거나 주제 특성을 살펴본 선행연구(김완종, 2014; 조재인, 2019)가 존재하였다. 이를 바탕으로 이용자가 동시에 대출한 복수 도서들의 주제 분야를 살펴보고, 이를 주제 간의 연관성을 분석하고자 한다. 구체적으로 WORD2VEC 기법 알고리즘을 변경한 ITEM2VEC을 적용하였다. WORD2VEC 기법은 자연어 처리 분야에서 유사한 의미를 가진 단어들을 인접한 벡터 공간에 위치시키는 기법이다 (Mikolov, Chen, et al., 2013). ITEM2VEC에서는 분석 단위를 단어로 처리하는 WORD2VEC과 달리 아이템 단위로 변환하여 유사한 이용 패턴이 있는 아이템을 탐색하여 추천한다(Barkan &

Koenigstein, 2016). 이 기법을 활용하여 주제를 분석 단위로 설정한 다음 주제 간 연관성을 정리한 후, 관련 있는 주제를 연결하여 네트워크를 생성하였다. 다음으로 커뮤니티 탐지(community detection) 기법을 적용하여 관련 있는 주제 분야 간 커뮤니티를 형성하여, 특성을 확인하였다. 이러한 시도는 이용자의 복수자료 이용 행위를 부분적으로 이해할 수 있으며, 이를 통해 도서관 장서 구성이나 배치, 자료 추천, 분류표 개정 작업 등에 유익한 시사점을 제공할 수 있다.

II. 이론적 배경

1. 공공도서관 대출데이터 분석 연구

본 연구의 목적은 공공도서관 이용자별 대출데이터를 분석하여 도서의 주제 간 연관성을 탐색하기 위한 것이다. 대출데이터를 활용한 국내 선행연구는 크게 이용자의 정보 이용 행위를 이해하기 위한 연구(김완종, 2014; 심지영, 2021; 이순영, 이수상, 2021; 조재인, 2019)와 도서 이용에 영향을 미치는 요인을 분석한 연구로 나눌 수 있다.

가. 정보 이용 행위 분석 연구

정보 이용 행위를 이해하기 위한 연구는 특정 지역이나 성별, 연령대와 같은 이용자의 특성에 따른 정보 이용 행위를 파악하기 위해 도서의 대출데이터를 분석하였다. 김완종(2014)은 서울특별시 한 자치구의 9개 공공도서관에서 수집된 데이터를 한국십진분류체계에 따라 동시 대출 빈도 분석과 함께 상관계수를 산출하여 주제 네트워크를 생성하고 군집 분석을 수행하였다. 연구 결과, 대출 빈도가 높은 소수의 주제 분야가 전체 대출의 80% 이상을 차지하고 있다는 것을 확인하였다. 이를 통하여 주제별 유사성과 대출 빈도를 파악하였으나, 한국십진분류법의 주류와 강목 수준에서 분석함에 따라, 세부 주제에 대한 분석이 필요하다는 보완 사항이 존재하였다.

심지영(2021)은 도서관의 한국십진분류체계에 따른 주제와 도서의 이용에 있어 다각적인 맥락을 제공하는 인터넷 서점 분류체계를 함께 활용하였다. 이를 위해 인터넷 서점 정보 데이터에서 속성 정보, 저자 정보를 수집하여 도서의 속성 정보와 도서관 대출 빈도와 대응 일치 분석을 수행하였다. 이를 통해 도서관별 정보요구의 특성과 도서 이용 양상을 확인하였다.

이순영과 이수상(2021)은 추출한 대출데이터를 통하여 기술통계 분석, 연도별 소장 장서 증가량 추이, 장서 구성비 등을 장서 비율로 나눈 값인 Bonn의 이용계수에 대한 분석을 수행하였다. 연구 결과, 대부분 도서관이 연도별 장서 증가량 추이, 미대출 장서 및 소장 장서의 주제별 구성비가 비슷하며 주제별 장서 구성이 이용자의 요구를 반영하지 못하고 있음을 확인하였다.

조재인(2019)은 남녀 노령자의 독서 주제를 파악하고자 하였다. 이를 위해 연구대상 집단의 공공도서관 인기 대출 도서와 각 인기 도서와 함께 대출된 도서를 추출하여 동시 대출 도서 행렬을 생성하였다. 집단별 차이를 확인하기 위하여 행렬의 상위 50건을 대상으로 가중네트워크를 생성, 군집 분석을 수행하였다. 분석 결과, 연령대별, 성별에 따라 독서 경향이 다른 것을 확인하였다.

앞선 선행연구들을 살펴본 결과, 대부분 하나의 지역이나 하나의 이용자 집단을 대상으로 연구를 수행하였다. 위 연구들은 해당 지역과 집단의 정보 이용행위를 확인할 수 있다는 점에서 의미가 있으나, 국내 도서관을 포괄적으로 살펴보지 못했다는 한계가 있다. 따라서, 본 연구는 선행연구와 달리 국내 도서관 860여 곳의 데이터를 활용하여 전반적인 대출 도서의 주제 간 연관성을 확인하고자 하였다.

나. 도서 이용 영향요인 분석 연구

도서 이용에 영향을 미치는 요인을 분석한 연구는 문화적, 사회경제적 요인과 도서 이용의 상관관계를 확인하고자 하였다. 이종육, 강우진, 이명(2021)은 지역사회의 사회경제적 박탈 수준이 공공도서관 도서 대출에 미치는 영향을 확인하고자 하였다. 이를 위해 수도권 지역사회의 박탈 지수를 산출하고 박탈 지수와 대출 책수 간의 관계를 파악하기 위해 다층모형 로버스트 선형회귀 분석을 실시하였다. 연구 결과, 지역사회의 사회경제적 박탈 수준이 높아질수록 1인당 공공도서관 대출 책수는 감소하는 것을 확인하였다. 그리고 이종육, 강우진, 박중규(2021)는 베스트셀러 순위가 공공도서관에서의 평균 대출 건수에 미치는 영향을 패널분석을 통해 살펴보고자 하였다. 이를 위해 국립중앙도서관이 제공하는 공공도서관 대출데이터 세트와 YES24 베스트셀러 목록 데이터 세트를 고정효과모형으로 분석하였다. 연구 결과, 도서의 베스트셀러 순위와 해당 도서 평균 대출 건수가 유의미한 상관관계가 있으며, 이는 도서의 내용분류에 영향을 받는 것을 알 수 있었다.

2. WORD2VEC과 ITEM2VEC

자연어 처리(Natural Language Processing) 영역에서는 기존의 복잡한 학습 방식보다 빠르고 효율적으로 학습하여 단어를 벡터 공간 내 벡터값으로 위치하게 하는 임베딩 기법인 WORD2VEC 기법이 제시되었다(Mikolov, Chen, et al., 2013). WORD2VEC은 문장 내 함께 등장하는 단어 간에는 유사성이 있으며, 유사한 단어는 유사한 벡터값을 가질 것이라고 가정한다. 이러한 가정 아래 대량의 단어를 자연어에서 벡터로 변환하는 임베딩(embedding)을 거쳐 다차원의 벡터로 표현한다(Mikolov, Chen, et al., 2013; Mikolov, Yih, & Zweig, 2013).

WORD2VEC은 의미의 유사도에 따라 벡터 공간 내에서 단어를 위치시키며, 성별이나 단·복수 등의 단어 간 관계까지 고려한다. 또한, 대수적 연산을 수행하는 단어 오프셋 기술(word offset technique)을 적용한다(Mikolov, Chen, et al., 2013). 이러한 WORD2VEC 기법은 여러 방식으

로 활용되어 단어, 문장 벡터화, 분류와 같은 자연어 처리뿐만 아니라 협업 필터링과 같은 추천 시스템에서도 적용되고 있다.

협업 필터링(collaborative filtering)은 이용자의 아이템(item) 이용 경험을 기반으로 유사한 아이템을 탐색하여 추천하는 기법이다. 즉, 과거 이용 데이터에 기반하여 이용자에게 적절한 아이템을 추천할 수 있다. 각종 추천 시스템에서는 이용자와 아이템 간의 관계를 중심으로, 이용자 집단과 아이템을 2가지 측면에서 유사성을 측정하여 구현하고 있다. 최근 딥러닝 기술이 발전하면서, 추천 시스템의 예측 성능을 높이기 위하여 CNN(Convolutional Neural Network), LSTM(Long Short-Term Memory), RNN(Recurrent Neural Network) 등과 같은 딥러닝 기반의 추천 기법이 제시되었다(김철진 외, 2019). 하지만 이는 이용자와 아이템의 수가 많을 경우, 많은 계산량이 필요 하다는 한계점이 존재한다(Barkan & Koenigstein, 2016). 또한 도서관 대출 서비스의 경우, 도서관의 아이템인 도서에 대하여 협업 필터링에 있어 중요한 사항 중 하나인 이용자가 만족도나 선호도 등을 파악하지 못하여, 이를 반영하지 못한다는 한계가 있다(임정훈, 조창제, 김종현, 2022).

ITEM2VEC은 WORD2VEC의 알고리즘을 변형하여 기존 이용자-아이템의 관계를 파악하던 기존 추천 시스템과 다르게 아이템 간의 유사성을 파악하는 기법이다(Barkan & Koenigstein, 2016). ITEM2VEC은 WORD2VEC과 마찬가지로 특정 범위 내 단어를 예측하는 학습 기법인 Skip-gram 모델(Mikolov, Chen, et al., 2013)과 학습하는 과정에서 잡음(noise)의 영향을 줄여 성능을 높이는 방법인 Negative-sampling(Mikolov, Sutskever, et al., 2013)을 활용한 SGNS(Skip-Gram with Negative Sampling)을 적용한다(Barkan & Koenigstein, 2016). 다만 ITEM2VEC은 WORD2VEC과 다르게 학습 과정에 아이템의 순서는 고려하지 않는다(Barkan & Koenigstein, 2016).

ITEM2VEC은 LSTM, CNN과 같은 다른 신경망(Neural network) 기법에 비해 학습 시간과 데이터 처리에 효율적인 것으로 알려져 있다. 또한, 일반적으로 추천에 활용되는 SVD 협업 필터링 방식에 비해 높은 성능을 가지고 있으며, 트랜잭션에서 희소한 아이템 추천에도 우수한 성능을 보인다(김철진 외, 2019: Barkan & Koenigstein, 2016). 따라서, 동시 대출 도서의 주제별 연관관계를 분석하기 위하여 ITEM2VEC을 적용하여 주제 간 유사성과 관련성을 파악하고자 한다.

3. 네트워크 분석

네트워크 분석(network analysis)은 구조화를 통해 개체 간 상호 관계를 체계적으로 표현하고 패턴을 파악하기 위하여 사용하는 방법이다. 이는 다양한 분야에서 연구 분야의 지식구조와 동향을 파악하거나, 맥락분석 등을 위하여 사용된다. 대규모 네트워크의 경우, 개별 노드의 특성을 파악하는 것이 어려우므로 노드 간 그룹화를 수행하고 특성을 파악하는 것이 효율적이다. 이에 따라 노드의 그룹을 파악하는 커뮤니티 탐지에 대한 연구가 지속적으로 이루어져 왔다.

기준의 연구는 계층적 클러스터링이나 Kernighan-Lin 알고리즘을 통해 네트워크 내 그룹을 추출하여 연구를 수행하였다. 그렇지만 이러한 알고리즘은 네트워크 내 그룹을 형성하기 위해 만들어진 것이 아니므로 커뮤니티 탐지에 적합하지 않았다. 따라서, 네트워크 그룹을 탐지하기 위해 매개 중심성이 높은 링크를 순서대로 제거하는 방법인 Girvan-Newnam 알고리즘이 제안되었다 (Newman, 2004). 하지만 이 알고리즘은 전체 링크의 매개 중심성을 계산하고 이러한 중심성이 낮은 링크를 순서대로 제거하기 때문에 시간이 오래 걸려 대규모 커뮤니티 탐지 방법으로는 적절하지 않다. 이러한 문제를 해결하기 위해, 네트워크 내 커뮤니티 내부 및 외부의 연결 정도를 측정하는 지표인 모듈성 지표(modularity)를 이용한 Rouvain 알고리즘이 제안되었다. 이 방식은 모듈성이 최적이 되는 커뮤니티를 탐지하고 통합된 네트워크를 생성할 수 있다(Blondel et al., 2008).

본 연구에서는 ITEM2VEC을 적용하여 파악한 주제 간 유사성 및 관련성을 기반으로 주제 네트워크를 생성하였으며, 생성된 네트워크에 Rouvain 알고리즘을 적용하여 주제 간 그룹화를 수행하고, 각 커뮤니티의 특성을 파악하였다.

III. 연구방법

본 연구는 이용자가 동시에 대출한 도서들의 주제 분야 간 연관관계를 분석한다. 다시 말해, 이용자가 여러 권의 도서를 대출하는 경우, 비슷한 주제나 관심 분야를 대출할 것이라는 가정에 근거하여 동시 대출 도서를 분석하였다. 데이터 수집·전처리, 분석 단계에 대한 세부 내용은 다음과 같다.

1. 데이터 수집 및 전처리

공공도서관 대출데이터와 소장 도서 데이터는 ‘문화 빅데이터 플랫폼’(<https://www.culture.go.kr/bigdata/>)을 통해 수집하였다. 다만 문화 빅데이터 플랫폼을 통해 수집 가능한 전체 대출데이터는 15억 건이 넘을 정도로 규모가 방대하다. 또한 대출데이터와 소장 도서 데이터는 문화 빅데이터 플랫폼을 통하여 매달 지속적인 데이터가 업로드되고 있다. 하지만, 대출데이터 중에서 대출이 최신에 이루어진 건의 경우, 도서관마다 매달 제공하는 데이터의 시간적 범위가 상이하며, 도서의 반납, 예약 등 데이터 변화의 폭이 커 데이터의 완전성에 대한 한계가 존재한다. 이러한 한계점을 보완하기 위하여 연도를 기준으로 필터링하고 데이터를 확인한 후, 상대적으로 데이터의 변화가 안정되어 완전성이 보장된 2019년 데이터를 분석 범위로 설정하였다.

전처리 단계에서는 소장 도서 데이터의 국제표준도서번호(ISBN)와 한국십진분류체계(KDC)

를 정제하는 작업을 수행하였다. 먼저 수집한 대출데이터 내 ISBN 정보의 경우, ISBN13 체계를 통하여 소장 도서 정보를 제공함에 따라 ISBN이 13자리가 아니면서 접두부가 도서를 나타내는 ‘978’이나 ‘979’로 시작하는 경우가 아닌 경우 소장 도서 데이터에서 제외하였다(양수완, 2020). 소장 도서 데이터의 ISBN의 정보에 대하여 세트 ISBN 여부를 함께 제공하고 있다. 이를 모두 개별도서 ISBN으로 변환하였다. 또한, KDC의 빈 데이터는 전처리한 ISBN 데이터를 기준으로 공공도서관에서 가장 많이 채택한 KDC 즉, 최빈값을 사용하여 보간하였다.

위의 작업에 따라 본 연구에 활용된 2019년 공공도서관 대출데이터는 866개 도서관에서 중복을 제거하고 고유한 대출자 4,430,811명에 의해 이루어진 대출 106,720,843건이다. 이 가운데 같은 날 동일한 이용자에 의해 대출된 도서 목록의 집합을 ‘동시 대출 건수’로 규정하였으며, 그 결과 동시 대출 건수는 22,443,699건(중복을 제거한 대출 도서의 수는 984,790권)이 식별되었다.

2. 데이터 분석

앞선 단계를 통해 파악된 도서 984,790권의 KDC 분류기호를 활용하여 동시 대출 건수 22,443,699건에 대해 ITEM2VEC 기법을 적용하여 분류기호를 벡터로 변환하였다. ITEM2VEC은 출현 빈도가 적게 나타나는 항목도 적절히 반영할 수 있어 KDC의 강목 수준에서 수행한 선행 연구(김완종, 2014)와 달리 요목 수준에서 분석하였다. ITEM2VEC 적용 과정에서 구축한 데이터 세트 내에서 출현 빈도가 낮은 분류기호를 제외한 요목 수준의 분류기호별 연관성이 높은 분류기호 상위 10개를 추출하였다. 상위 10개 분류기호를 활용한 이유는 동시 출현 빈도가 낮은 분류기호의 경우, 분석과정과 의미 해석상의 왜곡을 발생시킬 수 있기 때문이다.

이러한 과정을 통해 정리된 522개의 분류기호를 대상으로 네트워크를 생성하였다. 이때, 각 분류기호를 노드로 설정하고, 링크는 유사한 분류기호 간 연결, 그리고 링크의 가중치는 각 분류기호 간 유사도로 설정하였다. 네트워크 생성 후, 커뮤니티 탐지 기법인 Rouvain 알고리즘을 적용하고, 각 커뮤니티에 구성된 분류기호들을 살펴보며 특성을 파악하였다. 커뮤니티 탐지 기법을 통하여 산출된 커뮤니티에 대해 커뮤니티별로 연결 중심성, 매개 중심성, 근접 중심성을 산출하였다. 또한 커뮤니티 내 속하는 분류기호에 대하여 해당 커뮤니티 내에서의 연결 중심성, 매개 중심성, 근접 중심성을 확인하여, 해당 커뮤니티에서의 중요도를 파악하였다.

마지막으로 커뮤니티 간 중요성뿐만 아니라 유사성을 파악하기 위해 분류기호들을 이루고 있는 커뮤니티를 벡터화하였다. 이때, 문장을 이루고 있는 단어들의 벡터값에 대하여 가중 합 평균을 내어 문장의 벡터를 도출한 방법인 DOC2VEC을 응용하여 커뮤니티 내의 분류기호들의 가중 합의 평균을 활용하였다(Le & Mikolov, 2014). 이후, 코사인 유사도를 통하여 커뮤니티 간 유사도를 산출하여 이들 간의 관계를 파악하였다.

IV. 연구 결과

1. 주제 분야별 도서대출 현황

주제 분야에 따른 대출 현황을 살펴보기 위해 분류기호별 대출 건수(106,720,843건)를 KDC 주류와 강목 수준에서 살펴보았다. 주류별 비중을 확인한 결과, 문학(800)의 비율(46.18%)이 가장 높았으며, 다음으로 사회과학(300)이 12.71%, 자연과학(400)이 6.12%, 역사(900)가 8.42%로 뒤를 이었다. 비중이 가장 낮은 분야는 1.52%를 차지하는 종교(200)였으며, 예술(600)이 2.77%, 총류(000)가 3.30%, 철학이 3.91%로 상대적으로 대출 비중이 낮은 것으로 나타났다(〈표 1〉 참고).

〈표 1〉 KDC 주류별 대출 분포

000 총류	100 철학	200 종교	300 사회과학	400 자연과학
3,521,804 (3.3%)	4,171,590 (3.91%)	1,625,370 (1.52%)	13,566,030 (12.71%)	9,943,601 (9.32%)
500 기술과학	600 예술	700 언어	800 문학	900 역사
6,527,906 (6.12%)	2,953,997 (2.77%)	6,145,692 (5.76%)	49,280,143 (46.18%)	8,984,710 (8.42%)

대출이 활발히 이루어진 상위 20개의 강목을 살펴본 결과, 문학(800)이 5개(800, 810, 830, 840, 860)로 가장 많았으며, 사회과학(300)이 4개(320, 330, 370, 380), 역사(900)가 4개(900, 910, 980, 990)였다. 또한 자연과학(400)과 기술과학(500)은 2개(400, 410, 510, 590)였고 언어(700), 철학(100), 총류(000)가 1개(000, 180, 740)씩 목록에 있었으나, 예술(700)과 종교(200)는 존재하지 않았다. 강목 중 한국문학(810)은 대출 비중에서 18.43%를 차지하여 문학(800)을 제외한 다른 주류 항목보다 높은 대출 비중을 차지하고 있었다. 주류별로 강목의 비중을 정리하면 〈표 2〉와 같다.

〈표 2〉 KDC 강목별 대출 분포

강목	0 총류	1 철학	2 종교	3 사회과학	4 자연과학
00	1,334,898 (1.25%)	430,494 (0.4%)	34,569 (0.03%)	696,837 (0.65%)	5,287,892 (4.95%)
10	50,723 (0.05%)	25,189 (0.02%)	565,514 (0.53%)	17,398 (0.02%)	2,294,166 (2.15%)
20	503,146 (0.47%)	48,110 (0.05%)	292,419 (0.27%)	5,310,547 (4.98%)	256,119 (0.24%)
30	576,399 (0.54%)	6,886 (0.01%)	715,711 (0.67%)	1,446,383 (1.36%)	145,093 (0.14%)

공공도서관 동시 대출 도서의 주제 연관성 분석 연구

강목	0 종류	1 철학	2 종교	3 사회과학	4 자연과학
40	5,962 (0.01%)	72,304 (0.07%)	1,651 (0%)	470,882 (0.44%)	265,291 (0.25%)
50	141 (0%)	168,713 (0.16%)	2,256 (0%)	44,257 (0.04%)	441,965 (0.41%)
60	16,597 (0.02%)	211,533 (0.2%)	1,727 (0%)	333,039 (0.31%)	1,631 (0%)
70	41,016 (0.04%)	31,596 (0.03%)	4,742 (0%)	3,760,048 (3.52%)	441,513 (0.41%)
80	992,686 (0.93%)	1,993,637 (1.87%)	3,982 (0%)	1,441,429 (1.35%)	155,560 (0.15%)
90	236 (0%)	1,183,128 (1.11%)	2,799 (0%)	45,210 (0.04%)	654,371 (0.61%)
강목	5 기술과학	6 예술	7 언어	8 문학	9 역사
00	329,398 (0.31%)	347,472 (0.33%)	56,318 (0.05%)	9,715,778 (9.1%)	1,319,339 (1.24%)
10	1,569,499 (1.47%)	97,580 (0.09%)	1,159,377 (1.09%)	19,665,929 (18.43%)	3,114,094 (2.92%)
20	305,874 (0.29%)	5,146 (0%)	141,465 (0.13%)	955,012 (0.89%)	206,128 (0.19%)
30	347,243 (0.33%)	368,808 (0.35%)	174,015 (0.16%)	4,558,771 (4.27%)	17,850 (0.02%)
40	151,761 (0.14%)	17,333 (0.02%)	4,505,807 (4.22%)	10,572,916 (9.91%)	34,773 (0.03%)
50	321,977 (0.3%)	1,071,570 (1%)	13,114 (0.01%)	1,259,789 (1.18%)	5,293 (0%)
60	86,242 (0.08%)	89,438 (0.08%)	25,125 (0.02%)	1,556,548 (1.46%)	2,439 (0%)
70	184,905 (0.17%)	304,186 (0.29%)	33,940 (0.03%)	322,835 (0.3%)	1,794 (0%)
80	21,549 (0.02%)	137,026 (0.13%)	5,801 (0.01%)	373,490 (0.35%)	2,412,769 (2.26%)
90	3,209,458 (3.01%)	515,438 (0.48%)	30,730 (0.03%)	299,075 (0.28%)	1,870,231 (1.75%)

총류(000)는 총류(000), 철학(100)은 심리학(180), 종교(200)에서는 기독교(230), 사회과학(300)에서는 경제학(320)이 큰 비중을 차지하였다. 자연과학(400)에서는 자연과학(400), 기술과학(500)에서는 생활과학(590)의 비중이 커으며 예술(600)에서는 회화, 도화, 디자인(650), 언어(700)에서는 영어(740)의 비율이 높았다. 특히, 언어에서는 영어 관련 도서의 대출 비중이 약 73.32%를 차지하였는데, 이는 언어 관련 도서 이용자들은 다양한 언어 중 특히 영어에 관심이 높은 것을 확인할 수 있다. 문학(800)에서는 앞서 살펴본 바와 같이 한국문학(810)이 가장 많이

대출되었으며, 역사(900)에서는 아시아(910) 항목이 높은 것으로 나타났다. 위 강목들은 모두 해당 주류 내에서 대출 건수의 30% 이상을 차지하고 있었다.

2. 동시 대출 도서 주제 분석

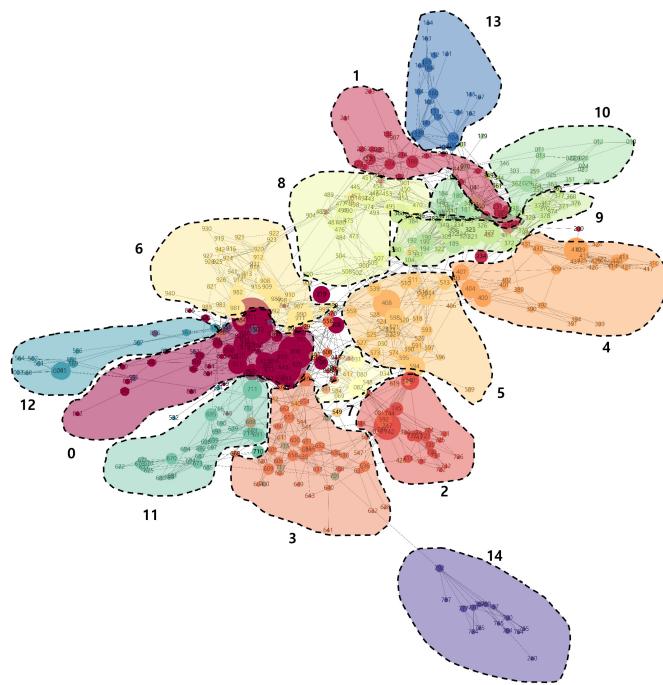
개별 이용자가 하루에 빌린 도서의 주제를 그룹화하고 분석하여 특성을 파악하였다. 즉, 2019년 도서관 이용자의 22,443,699건의 동시 대출 건수에 대한 목록을 생성하였다. 생성한 동시 대출 건수 목록에서 동시에 대출된 도서의 KDC 분류기호를 추출하였으며, ITEM2VEC 기법을 적용하여, 분류기호들을 벡터로 변환하였다. 벡터값을 통하여 분류기호마다 높은 연관성을 가진 분류 기호 10개를 산출·연결하여 네트워크를 생성하였다.

생성된 네트워크에서 유사한 이용 패턴이 나타나는 분류기호로 이루어진 커뮤니티를 형성하기 위해 Louvain 알고리즘을 통한 커뮤니티 탐지를 수행하였으며 결과로, 15개의 커뮤니티를 생성하였으며, 커뮤니티를 이루고 있는 분류기호를 살펴보았다(〈그림 1〉, 〈표 3〉 참조). 또한 네트워크 내에서 중요한 역할을 하는 커뮤니티를 살펴보기 위하여 커뮤니티 수준의 연결 중심성(Degree Centrality), 매개 중심성(Betweenness Centrality), 근접 중심성(Closeness Centrality)을 산출하였다(〈표 4〉 참조). 그 후, 커뮤니티마다 서브 그래프를 생성하여, 커뮤니티 내에서 연결 중심성, 매개 중심성, 근접 중심성을 산출하여 중요한 분류기호를 살펴보았다.

〈표 3〉 커뮤니티별 주요 분류 항목

커뮤니티	주요 분류항목
C0	문학 총류, 한국문학, 중국문학, 일본문학 및 기타 아시아 제문학, 영미문학, 독일문학, 프랑스문학, 스페인 및 포르투갈문학, 이탈리아문학, 기타 제문학
C1	인도철학, 심령, 초심리학 및 운명판단, 종교 총류, 비교종교, 불교, 기독교, 이슬람교
C2	중국어, 일본어, 영어
C3	자동차공학, 의복, 미술 총류, 조각, 공예, 서예, 디자인 및 회화, 사진예술
C4	국방 및 군사학, 자연과학 총류, 통계자료 및 통계학, 수학, 물리학, 화학
C5	의학, 농학, 환경공학(위생), 건축학, 식품공학, 제조, 생활과학,
C6	역사 총류, 아시아사, 유럽사, 아프리카사, 북아메리카사, 지리, 전기
C7	사전, 박물관학, 총서류
C8	천문학, 지학, 생명과학, 식물학, 동물학, 기술과학 총류
C9	사회과학 총류, 경제학, 사회학, 정치학, 행정학, 교육학, 민속학
C10	도서 및 서지학, 문헌정보학, 신문 및 저널리즘, 논리학, 심리학, 윤리학
C11	음악, 공연예술, 오락 및 스포츠, 한국어
C12	시스템, 컴퓨터과학, 프로그래밍, 토목공학, 기계공학, 전기·통신·전자공학
C13	철학 총류, 형이상학, 인식론, 철학의 세계, 경학, 동양철학, 서양철학
C14	독일어, 프랑스어, 스페인어 및 포르투갈어, 기타제어

공공도서관 동시 대출 도서의 주제 연관성 분석 연구



〈그림 1〉 분류기호 커뮤니티 형성 결과

〈표 4〉 커뮤니티별 연결, 매개, 근접 중심성 결과

커뮤니티	연결 중심성	매개 중심성	근접 중심성
C0	0.074	0.172	0.031
C1	0.026	0.030	0.024
C2	0.033	0.034	0.026
C3	0.07	0.209	0.031
C4	0.039	0.015	0.025
C5	0.057	0.169	0.033
C6	0.051	0.045	0.033
C7	0.018	0.004	0.025
C8	0.04	0.064	0.025
C9	0.065	0.125	0.032
C10	0.055	0.073	0.026
C11	0.039	0.074	0.024
C12	0.010	0.005	0.025
C13	0.020	0.012	0.021
C14	0.002	0.000	0.013

커뮤니티는 하나의 주류에 속한 분류기호들로 구성된 커뮤니티와 두 가지 이상의 주류에 속한

분류기호들로 구성된 경우로 나눌 수 있다. 이들 커뮤니티의 특징에 대한 설명은 다음과 같다.

가. 하나의 주류로 구성된 커뮤니티

하나의 주류로 구성된 커뮤니티는 C0, C2, C5, C6, C7, C9, C13과 C14이다(〈그림 1〉, 〈표 3〉 참고). 먼저 'C0'은 전반적인 문학 주제로 구성되어 있다. 주제 네트워크에서 가장 많은 주제 분류 기호(61개)가 모여 커뮤니티를 이루고 있다. 문학 주제의 경우 가장 많은 대출이 이루어지는 주제 이기도 하며, 이에 따라 커뮤니티별 중심성을 산출하였을 때 다른 커뮤니티에 존재하는 주제 분류 기호와 가장 많이 연결되어 있으며(0.74), 이를 중심으로 다른 주제 분류기호들을 매개(0.172)하고 있음을 확인할 수 있다(〈표 4〉 참조). 커뮤니티 내에서는 프랑스 소설(863)이 가장 높은 연결 중심성(0.283)을 가지고 있다. 프랑스 소설의 경우, 「개미」, 「고양이」, 「나무」 등을 집필한 베르나르 베르베르와 「아가씨와 밤」, 「센트럴 파크」, 「7년 후」 등을 집필한 가욤 뮤소와 같이 국내 독자들이 흥미를 갖는 작가의 도서가 많이 대출이 이루어진 것이 이에 영향을 미쳤을 것으로 추측할 수 있다.

'C2'는 중국어(720), 일본어(730), 영어(740)와 같은 국내의 주류 외국어로 이루어져 있으며 이들은 모두 언어(700)의 하위 주제이다. 이들은 언어학적 접점이 없으나, 영어(740)는 국내의 제1외국어이며, 중국어(720)와 일본어(730)는 국내에서 빈번하게 채택되는 제2외국어이다. 이는 도서 주제 이용에 있어, 언어학적인 체계의 유사성보다는 학습 수요에 기반하여 이용되고 있는 것으로 보인다. 해당 분류기호를 가지고 있는 도서의 경우, 관련 수험서나 참고서와 같이 개인학습을 위주로 이용되는 모습을 보인다(이종우, 강우진, 박중규, 2021). 이러한 모습은 도서 이용에도 반영되어 다른 커뮤니티 내 분류기호들과 연결이 낮은 모습, 즉 낮은 연결 중심성(0.026)을 보여 준다. 또한, 커뮤니티 내에서는 일본어 문법이 연결 중심성이 0.382, 매개 중심성이 0.444 그리고 근접 중심성이 0.523으로 높은 것으로 나타났다. 또한 일본어 독본, 회화, 해석은 매개 중심성이 0.406, 근접 중심성이 0.523으로 높게 나타났으며, 중국어 독본, 회화, 해석은 근접 중심성이 0.515로 높게 나타났다. 또한 독본, 회화, 해석뿐만 아니라 어휘도 문법과 독본, 회화, 해석을 뒤이어 높은 매개 중심성(0.378)을 가진 것으로 나타났다. 문법과 독본, 회화, 해석 그리고 어휘의 주제가 높은 커뮤니티 내에서 높은 중심성을 보여줌에 따라, 외국어 개인학습과 이용에 있어 문법과 회화, 어휘를 중요히 여기고 있음을 확인할 수 있다.

'C5'는 기술과학(500)의 하위 주제인 의학(510), 농업·농학(520), 건축학(540), 화학공학(570), 제조업(580), 생활과학(590)과 KDC 5판의 건축술(610)로 구성된다. 위 주제는 일상생활과 관련 있는 항목으로 생명과 관련된 의학(510) 전반, 주거와 관련된 건축 시공 및 적산(542), 식생활과 관련된 음료기술(573) 등이 포함된다. 한편, KDC 5판에서 예술(600)에 속했던 건축술(5판 610)이 건축학(540)과 같은 커뮤니티에 속하였다. 건축술의 경우, 이용자가 기술적 성격의 건축공학과 예술적 성격의 건축술을 함께 이용할 가능성이 높다. 5판은 이 두 주제가 분리되어 이용이 불편하

다는 지적이 있었다. 이를 개선하기 위하여 건축학(540)으로의 통합이라는 개선안 제시되었으며, 6판 이후 통합되었다(여지숙, 공성훈, 오동근, 2013). 특히, 'C5'는 'C3'과 'C0' 다음으로 커뮤니티 수준에서 높은 매개 중심성(0.169)을 가지고 있는데, 이는 상대적으로 사회과학, 자연과학, 역사와 같이 일상생활에 도움을 주는 실용적인 특성이 있기 때문으로 판단된다. 이를 반영하듯 화장, 식품과 음료, 농업, 주택관리 및 가정설비, 건강과 같이 실생활 관련된 주제가 높은 연결 중심성, 매개 중심성, 근접 중심성을 가지고 있는 것으로 나타났다.

'C6'에 속한 강목은 역사 총류(900), 아시아사(910), 유럽사(920), 아프리카사(930), 북아메리카사(940), 지리(980), 전기(990)로 모두 역사(900)의 하위 주제이다. 커뮤니티 내에서는 주로 유럽과 아시아의 역사, 지리가 커뮤니티 내에서 높은 연결 중심성, 매개 중심성, 근접 중심성을 가지고 있는 것으로 나타났다. 'C7'은 백과사전(030), 박물관학(069), 총서(082)로, 주로 주류 총류(000)에 해당하는 주제로 이루어져 있다. 다음으로 'C9'의 경우, 사회과학 총류(300), 경제학(320), 사회학(330), 정치학(340), 행정학(350), 법률·법학(360), 교육학(370), 민속학(380)으로 구성된다. 위 주제는 모두 통계자료(310)와 국방·군사학(390)을 제외한 사회과학(300)의 하위 주제이다. 사회과학 주제의 경우, 문학 다음으로 높은 비중의 대출 빈도를 가졌음에도 불구하고, 커뮤니티별 중심성에 있어서는 상대적으로 낮은 수치를 가지고 있다. 이는 'C5'와 달리, 다른 커뮤니티와 비교하여 커뮤니티 내 주제 이용이 활발히 이용되어 상대적으로 낮은 중심성을 가지고 있을 것으로 추측된다. 커뮤니티 내 중심성의 경우, 사회학 주제가 'C5' 내 높은 연결 중심성(0.333)을 가져 다른 주제들과 유사성을 가지고 있는 것으로 파악하였다. 또한, 행정학(0.182)과 교육학(0.221)이 높은 매개 중심성을 가져, 커뮤니티 내 다른 분류기호를 매개하는 것으로 나타났다.

'C13'은 철학 총류(100), 형이상학(110), 인식론(120), 철학의 세계(130), 경학(140), 동양철학(150), 서양철학(160)과 같이 철학(100)의 하위 주제로 구성된다. 낮은 주제 대출 빈도를 반영하듯, 커뮤니티 수준에서 낮은 중심성을 가지고 있다. 커뮤니티 내 중심성의 경우, 서양철학이 연결 중심성이 0.524, 매개 중심성이 0.543, 그리고 근접 중심성이 0.618로 커뮤니티 내 가장 높은 중심성을 가지고 있다. 뒤이어 맹자, 노자, 한비자 등을 다룬 중국철학이 0.429로 높은 연결 중심성을 가지고 있는 것으로 나타났다.

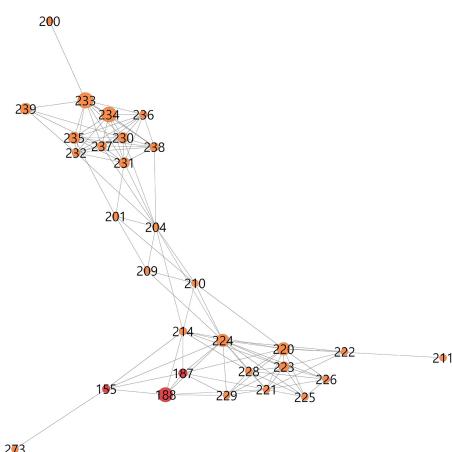
'C14'는 국내 주류 외국어가 속한 'C2'와 대조적으로 독일어(750), 프랑스어(760), 스페인어 및 포르투갈어(770)와 같은 상대적으로 학습이 덜 이루어지고 있는 외국어 주제로 이루어져 있다. 특히 'C2'에 속하는 영어(740)의 경우, 중국어(720), 일본어(730)보다는 'C14'의 언어와 언어학적 체계가 비슷함에도 불구하고 이용 패턴에 있어서는 다른 행태를 띠고 있음을 재확인할 수 있다. 특히, 해당 커뮤니티에 속하는 주제 분류의 경우, 같은 외국어 분류기호를 가지고 있는 'C2'와 비교하여도 낮은 비중을 가지고 있으며, 이에 따라 더 낮은 중심성을 가지고 있는 것으로 나타났다.

종합적으로, 커뮤니티가 하나의 주류로 이루어진 경우, 보통 하나의 커뮤니티로 구성되나, 일부

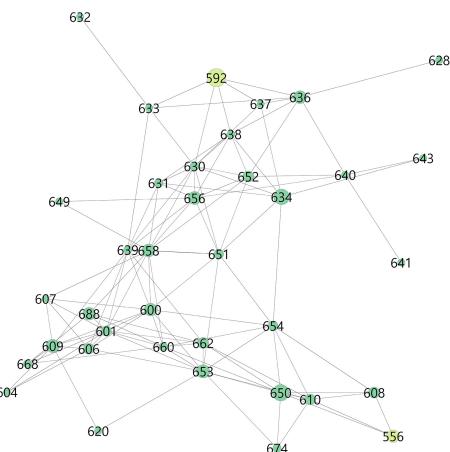
복수의 커뮤니티로 분화되어 형성되는 경우도 존재한다. 이는 유사한 주제를 가졌으나, 주제 이용 측면에서 다소 상이하여 나타났음을 대략적으로 추론할 수 있었다. 이에 대한 예시로는 같은 언어 (700)에 속하지만 다른 커뮤니티로 나뉜 'C2'와 'C14'를 들 수 있다.

나. 둘 이상의 주류로 구성된 커뮤니티

복수의 주류로 구성된 커뮤니티는 C1, C3, C4, C8, C10, C11, C12이다. 이 커뮤니티는 모두 하나의 주류가 대부분을 차지하지만 다른 주제 분야를 일부 포함하고 있다.



〈그림 2〉 C1 서브 그래프



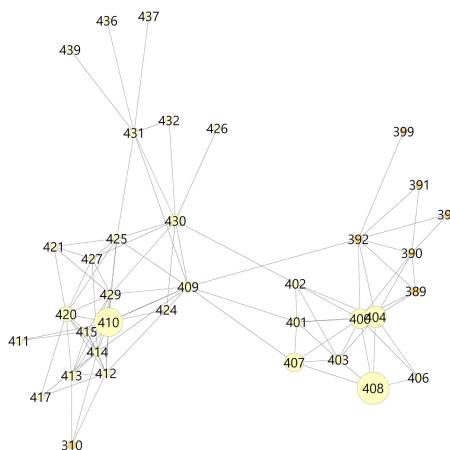
〈그림 3〉 C3 서브 그래프

‘C1’은 종교(200)의 하위 주제가 대부분을 차지하고 철학(100)에 속하는 주제가 일부 포함되어 있다. 종교(200)에 속하는 주제는 비교종교(210), 불교(220), 기독교(230)와 이슬람교(270)이다. 한편, 철학(100)에 속하는 주제는 인도철학 및 사상(155)과 심령연구 및 초심리학(187), 운명판단(188)과 같은 일부 심리학 분야가 혼재되어 있다. 철학과 종교, 심리학은 오랜 기간 유사한 주제로 인식된 것으로 보이는데 가령, Brown이나 Cutter가 제시했던 과거 분류체계에도 함께 분류된 것이 확인된다. ‘C1’의 경우, 비슷한 대출 비중을 가지고 있는 ‘C2’와 비교하였을 때 커뮤니티 수준에서의 낮은 연결 중심성, 매개 중심성 그리고 근접 중심성을 가지고 있다. 이는 다른 커뮤니티들과의 연결이 현저히 낮은 걸 확인할 수 있으며, 커뮤니티 내 주제들과의 이용이 매우 활발하게 이루어지고 있음을 추측할 수 있다. 커뮤니티 내 중심성에서는 주로 불교, 기독교와 관련된 주제가 주로 높은 연결, 매개, 근접 중심성을 가진 것으로 나타났다.

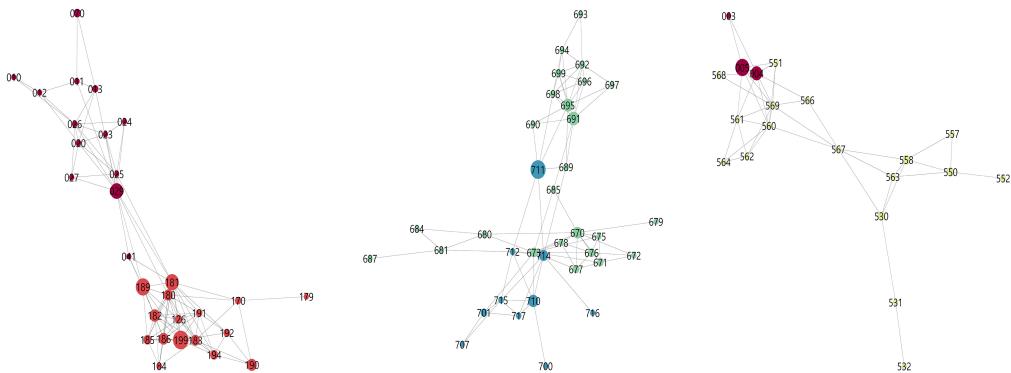
'C3'은 예술(600), 조형미술(620), 공예(630), 서예(640), 디자인 및 회화(650), 사진예술(670)

과 같이 예술(600)의 하위 주제가 중심을 이루고 있다. 이에 더하여, 기술과학의 하위 주제인 자동차공학(556)과 의복(592)이 함께 발견되었는데, 자동차 공학과 의복은 디자인과 밀접한 관련이 있기 때문이다. 특히 자동차 영역에서 디자인과 관련된 연구가 이루어진 바 있다는 점과 대출 데이터의 「자동차 디자인 교과서」라는 책을 통해 위 주제 간 연관성을 추론할 수 있다. 또한, 커뮤니티 수준에서 가장 높은 매개 중심성(0.209)을 가지고 있는데, 'C5'와 같이 다른 주제를 가진 커뮤니티의 주제 도서와 활발히 이용되어 이들을 매개하는 특성을 가졌음을 추측할 수 있다.

'C4'는 사회과학(300)의 하위 주제인 통계자료(310)와 국방·군사학(390), 자연과학(400)의 하위인 총류(400), 수학(410), 물리학(420), 화학(430)으로 이루어져 있다. 물리학과 화학의 경우, 각 분야 내 물리화학(Physical chemistry), 화학물리학(Chemical physics)이 존재할 정도로 밀접한 관계가 있다. 또한, 사회과학 내 통계자료의 경우 자연과학의 통계학(413)뿐만 아니라 다른 기초과학과도 밀접한 연관이 있어, 같은 커뮤니티에 속한 것으로 보인다.



뮤니티 수준에서 상대적으로 낮은 연결중심성(0.055)과 대출 비중에 비하여 높은 매개 중심성(0.073)을 가지고 있다. 이는 다른 커뮤니티 내 주제와는 이용 특성에 있어서 직접적인 유사성을 적을 수 있으나, 인문, 사회, 과학 등 복합학적인 특징을 가진 덕분에 다른 커뮤니티 사이를 연결하는 커뮤니티로 해석할 수 있다. 이는 이용자들이 다른 주제로 관심을 환기할 수 있는 주제들로 이루어진 커뮤니티로 추측할 수 있다. 커뮤니티 내 연결, 매개, 근접 중심성은 대부분 다양한 분야가 복합된 심리학 분야와 독서 및 정보매체 활용 분야가 높은 것으로 나타났다.



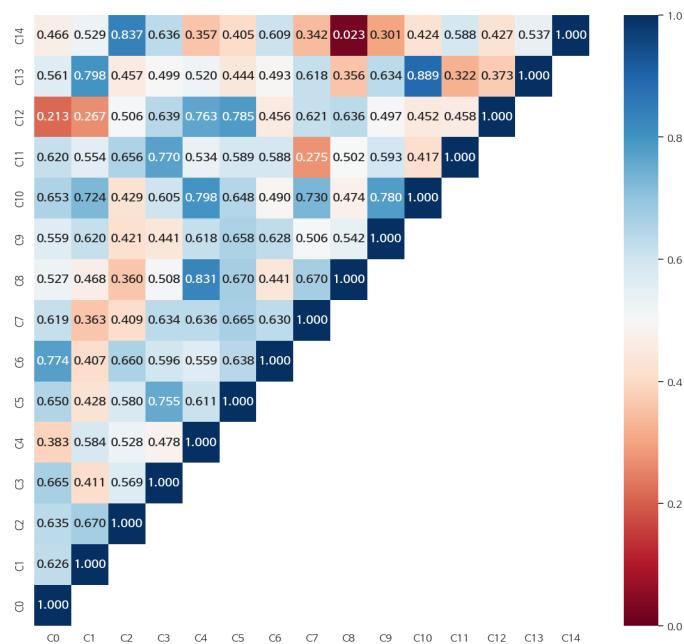
〈그림 6〉 C10 서브 그레프 〈그림 7〉 C11 서브 그레프 〈그림 8〉 C12 서브 그레프

'C11'은 예술(600) 주제의 음악(670), 공연예술(680), 오락·스포츠(690)와 언어(700)의 언어총류(700), 한국어(710) 주제가 함께 나타났다. 'C12'는 공업, 토목공학(530), 전기공학, 통신공학, 전자공학(560)과 같은 기술과학(500) 주제와 총류의 이론 체계 및 시스템(003), 컴퓨터과학(004), 프로그래밍, 프로그램, 데이터(005)가 함께 나타났다. 이 커뮤니티에 속한 주제 분야는 분류체계에서는 총류와 기술과학에 나누어져 있지만, 학술 연구 분야에서는 공학에 속하는 것이 일반적이다. 더하여, 현대 기술과학 분야는 프로그래밍과 소프트웨어를 적극적으로 활용하고 있다. 이 커뮤니티는 이러한 상황을 명확하게 보여주는 예시이다. 특히 'C5'와 같이 주로 기술과학 분야의 분류기호로 커뮤니티를 이루고 있으나, 대비적으로 커뮤니티 수준에서 연결, 매개, 근접 중심성은 상당히 낮은 편으로 나타났다. 이는 일상생활에서 활용되어 실용적인 성격을 가진 'C5'와 다르게 'C12'의 경우, 전자공학, 전기공학과 같이 전문적인 성격이 강하게 나타나 이용에 있어 이러한 지식이 필요한 이용자들을 대상으로 이용이 되어, 낮은 중심성을 가지고 있는 것으로 추측된다.

다. 커뮤니티 간 유사도

분류기호를 벡터화한 것과 유사하게, 커뮤니티를 이루고 있는 주제를 벡터화하였다. 커뮤니티 내의

분류기호들의 가중 합의 평균을 내어 커뮤니티를 벡터화하였으며, 코사인 유사도를 통하여 커뮤니티별 유사도를 산출하였다. 또한, 유사한 커뮤니티를 3개씩 연결하여 기존 분류기호 네트워크를 간략화하였다. 커뮤니티 간 유사도(〈그림 9〉)와 커뮤니티 네트워크 시각화 결과(〈그림 10〉)는 다음과 같다.

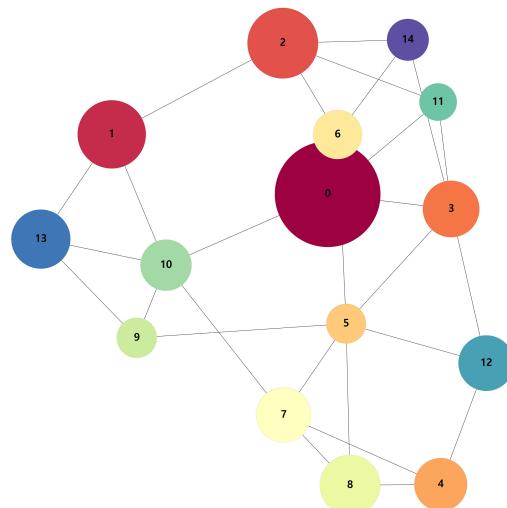


〈그림 9〉 커뮤니티 간 유사도

〈그림 9〉에 제시된 코사인 유사도는 0과 1 사이로 나타나며 1에 가까울수록 유사도가 높은 것을 의미한다. 본 연구의 14개 커뮤니티 간 유사도는 가장 높은 수준인 0.889에서 가장 낮은 수준 0.023까지 분포하고 있다. 유사도가 높은 커뮤니티와 낮은 커뮤니티의 주요 사례를 살펴보면 다음과 같다. 가장 유사도가 높은 커뮤니티는 ‘C10’과 ‘C13’으로 유사도는 0.889이다. ‘C10’의 논리학, 심리학, 윤리학과 ‘C13’에 속한 대부분이 철학(100)의 하위 주제이기 때문에 두 커뮤니티 간 유사도가 높게 나타난 것으로 추측된다. ‘C10’과 두 번째로 유사한 커뮤니티는 ‘C9’로 유사도는 0.78이다. ‘C10’의 문현정보학, 신문, 심리학이 분류체계에서는 종류나 철학으로 분류되나 학술 연구 분야에서는 사회과학 분야로 분류된다. 따라서 사회과학 분야가 다수 속한 ‘C9’와 유사도가 높게 나타났을 것이다.

한편, ‘C13’과 두 번째로 유사한 커뮤니티는 ‘C1’로 유사도는 0.798이다. 이는 두 커뮤니티가 모두 종교의 하위 주제를 포함하고 있기 때문으로 보인다. 더하여, ‘C1’의 하위 주제 중 심리학과 철학은 Cutter의 전개 분류법과 미 의회 도서관 분류법 등 다른 분류체계에서 종교가 함께 하나의

대주제로 구성되기도 한다(곽철완, 2016). 이러한 근거를 통해 철학과 종교 간 높은 유사성을 설명할 수 있을 것이다. 그리고 'C4'는 'C12'와도 0.763이라는 높은 유사도를 보이는데, 이는 자연과학과 기술과학의 연관성에서 기인한 것으로 보인다. 'C4'는 다수의 자연과학 하위 주제가 속하고, 'C12'는 기술과학의 하위 주제가 많다. 두 주제가 불가분의 관계이며 서로를 기반으로 발전하는 것은 널리 알려진 사실이다.



〈그림 10〉 커뮤니티 네트워크 시각화 결과

문학(800)의 하위 주제로 이루어진 'C0'과 역사·지리·전기(900)의 하위 주제로 이루어진 'C6'과 또한 높은 유사성을 가지고 있다. 데이터를 확인한 결과, 역사적 배경이나 인물을 대상으로 한 문학 작품이 역사책과 함께 대출되거나, 지리나 기행과 연관된 책을 함께 대출한 사례를 확인할 수 있었다. 이 밖에도 같은 주류를 공유한 커뮤니티의 유사성이 높았다. 가령, 'C2'와 'C14'는 유사도가 0.837로 매우 높은 편이다. 이 두 커뮤니티는 모두 언어(700)의 하위 주제로 이루어졌으므로 유사도가 높다. 'C4'와 'C8' 또한 주로 자연과학(400)으로 구성되어 유사도가 높은 것으로 나타났다.

반면 'C8'과 'C14'는 가장 낮은 유사성을 가지고 있다. 'C14'는 제2외국어 분류기호로 이루어져 있으며 주로 개인의 학습을 위한 학습/수험서로 이용되는 반면 'C8'은 자연과학의 하위 주제로 상대적으로 학술적인 지식이 요구되는 전문 서적이 많다(심지영, 2021; 이종욱, 박중규, 강우진, 2021). 이와 같은 독서 이용 동기의 차이는 이용 기반 주제 유사성에도 반영되는 것을 볼 수 있다. 이와 유사한 사례로 전문지식을 요구하는 주제 분야인 사회과학 주제로 구성된 'C9'도 낮은 유사성을 보인다.

V. 요약 및 결론

이용자의 정보요구 및 이용행위에 대한 이해가 공공도서관 전략 수립에서 중요하게 고려되고 있는 가운데 본 연구는 도서관의 대출데이터를 활용하여 도서관 이용자의 대출 행위를 살펴본 뒤, 도서 동시 대출에 따른 주제 분야 간 연관성을 살펴보았다. 특히, 기존 연구가 단일 지역이나 이용자 집단을 대상으로 하였고, 주제 분야 연관성 분석도 분류표의 강목 수준에서 이루어진 한계가 있었다. 이에 따라 본 연구에서는 2019년 전국 866개 공공도서관의 약 1억 건에 달하는 대출데이터에서 요목 수준의 KDC를 추출한 뒤 ITEM2VEC을 적용하여 주제 분야 간 연관성을 살펴보았다. 이는 강목 수준에서 분석한 선행연구와 비교하여 분석 수준을 요목 수준까지 확장하였으며, 분석 대상이 된 도서관의 범위도 전국으로 확대하여 분석 결과의 일반화 가능성을 높이고자 하였다. 특히 본 연구에서는 커뮤니티를 구성하는 보다 세분된 분류기호를 제시하였고, 기호 간 연관성을 살펴보았다는 점에 차별성이 있다.

연구 결과를 요약하면 다음과 같다. 먼저, 주제별 대출 비율을 분석하면, 가장 큰 비중의 주류는 약 46%를 차지한 문학(800)이고, 가장 적은 비중의 주류는 1.5%를 차지한 종교(200)였다. 강목 수준에서는 한국문학이 약 18%로, 문학을 제외한 다른 주류보다 높은 대출 비중을 가지고 있는 것으로 나타났다. 또한 대출 비중 상위 20개의 강목 중 문학(800)과 사회과학(300) 그리고 역사(900)는 각각 5개와 4개의 강목이 있던 것에 비해 예술과 종교는 목록에 존재하지 않았다. 이를 통해 주제에 따라 이용에 있어 편차가 있는 것을 확인할 수 있었다.

다음으로 ITEM2VEC 기법을 통하여, 한 이용자에 의해 동시에 대출된 도서의 분류기호 간 연관성을 살펴본 후, 커뮤니티 탐지 기법을 통해 분류기호 간 커뮤니티를 형성하였다. 형성된 15개의 커뮤니티는 하나 혹은 두 개의 주제(주류)로 구성되었는데, 하나의 주제로 이루어진 커뮤니티는 총류(사전, 박물관학, 총서류), 철학, 사회과학, 기술과학, 언어, 문학, 역사 등으로 구성되어 주로 사회과학 및 인문학 분야의 주제인 것으로 파악되었다. 두 개의 주제로 구성된 커뮤니티를 살펴보았을 때, 총류에 속한 컴퓨터 공학, 프로그래밍 등이 기계공학, 전자 공학과 같은 기술 과학과 함께 나타나는가 하면, 총류에 속한 서지학, 문헌정보학, 신문 및 저널리즘과 철학의 논리학, 심리학이 함께 이루고 있었다. 이외에도 철학과 종교, 사회과학 내 통계자료와 자연과학 내 수학, 통계학, 물리학 등이, 자연과학과 기술과학, 기술과학과 예술이 같이 커뮤니티를 구성하고 있었다. 이는 두 가지 주류로 이루어진 커뮤니티를 분석한 결과 모두 요목 수준에서 의미 있는 주제적 연관성을 파악할 수 있었다. 그렇지만 국방과학과 자연과학의 사례와 같이 학문 간의 연관성을 확인하기 힘든 경우도 존재하였는데, 이에 대한 추가적인 분석이 요구된다.

본 연구는 자연어 처리 분야에서 단어 임베딩 기법 중 하나인 WORD2VEC 기법을 응용하여 아이템 간 이용 유사성을 파악하는 ITEM2VEC 기법을 활용하였다. 이 기법은 이용자의 도서

이용행위에 기반하여 함께 대출될 가능성이 높은 자료의 주제를 파악하는 데 도움이 되어, 궁극적으로 도서관 장서 구성 및 배치, 관련 주제 분야 자료 추천, 분류표 개정 등에 활용할 수 있을 것이다. 다시 말해, 도서관은 이용자의 도서 대출 특성을 반영하여 장서를 배치할 수 있다. 이를 통해 이용자의 편의성을 도모하고, 관련 자료에 대한 우연한 발견 가능성을 높여 자료 이용률을 높이는 데 기여할 수 있다. 나아가 ITEM2VEC 기법은 도서의 분류 기호뿐만 아니라 주제명이나 키워드 등에도 적용할 수 있어, 이용자의 자료 이용행위를 이해하는 데 도움이 될 수 있을 것이다.

한편 이 연구에서는 도서 분류기호의 요목 수준을 활용하였으나 일부 주제 분야(예: 문학)에서는 이보다 더 하위 수준의 분석이 필요함을 알 수 있었다. 그리고 특정 연도(2019년)의 대출데이터만 활용한 한계가 존재한다. 이에 후속 연구에서는 대출데이터의 시간적 범위를 확대하는 것과 더불어 국립중앙도서관 주제명표목표 등을 활용한 더욱 심층적인 주제 연관성 분석을 실시하고자 한다.

참 고 문 헌

- 곽철완 (2016). Cutter의 전개분류법에 대한 연구. *한국문현정보학회지*, 50(3), 249-265.
<https://doi.org/10.4275/KSLIS.2016.50.3.249>
- 김성희, 이승민 (2022). 미래 공공도서관 발전 방향에 대한 사서 및 이용자 인식 조사. *한국비블리아학회지*, 33(1), 499-519. <https://doi.org/10.14699/kbiblia.2022.33.1.499>
- 김완종 (2014). 공공도서관의 주제별 자료 이용 현황 분석: 서울특별시 A구 산하 공공도서관을 중심으로. *정보관리학회지*, 31(3), 353-369. <https://doi.org/10.3743/KOSIM.2014.31.3.353>
- 김철진, 정지현, 조천우, 유제광 (2019). 상품 추천 기법의 성능평가 분석. *한국지식정보기술학회논문지*, 14(5), 515-525. <https://doi.org/10.34163/jkits.2019.14.5.008>
- 도서관법. 법률 제18763호.
- 문화체육관광부 [발행년불명]. 도서관 정보나루. 출처: <https://www.data4library.kr/>
- 심지영 (2021). 도서대출 데이터를 이용한 공공도서관 이용자 정보요구 분석: 서울시 공공도서관을 중심으로. *정보관리학회지*, 38(2), 173-199. <https://doi.org/10.3743/KOSIM.2021.38.2.173>
- 양수완 (2020). 도서관 공공데이터의 품질에 관한 연구: 도서관 정보나루의 도서 상세 조회 API를 중심으로. *정보관리학회지*, 37(4), 181-206. <https://doi.org/10.3743/KOSIM.2020.37.4.181>
- 온정미, 박성희 (2020). 도서관 빅데이터 플랫폼을 활용한 공공도서관 빅데이터 분석 연구: 대전한밭도서관을 중심으로. *정보관리학회지*, 37(3), 25-50.
<https://doi.org/10.3743/KOSIM.2020.37.3.025>
- 여지숙, 공성훈, 오동근 (2013). KDC 제5판 건축학 분야 전개의 개선방안. *한국문현정보학회지*,

- 47(2), 359-376. <https://doi.org/10.4275/KSLIS.2013.47.2.359>
- 이순영, 이수상 (2021). 부산지역 공공도서관의 빅데이터 분석 연구: 도서관 정보나루 장서/대출데이터를 중심으로. *한국문헌정보학회지*, 55(4), 89-114.
- 이종우, 강우진, 박중규 (2021). 베스트셀러 순위가 공공도서관 대출에 미치는 영향 분석: 패널자료 분석을 중심으로. *정보관리학회지*, 38(4), 1-23. <https://doi.org/10.3743/KOSIM.2021.38.4.001>
- 이종우, 강우진, 이명 (2021). 지역사회 사회경제적 박탈이 공공도서관 대출 책수에 미치는 영향. *한국문헌정보학회지*, 55(4), 219-243. <https://doi.org/10.4275/KSLIS.2021.55.4.219>
- 임정훈, 조창제, 김종현 (2022). 연관규칙을 활용한 학교도서관 도서추천시스템 개발에 관한 연구. *정보관리학회지*, 39(3), 1-22. <https://doi.org/10.3743/KOSIM.2022.39.3.001>
- 조재인 (2019). 도서 대출데이터를 활용한 남녀 노령자의 독서 주제 분석. *한국도서관·정보학회지*, 50(1), 23-41. <https://doi.org/10.16981/kliss.50.1.201903.23>
- 한국과학기술정보연구원 (2015). 2014 도서관 빅데이터 분석활용 체계 구축. 문화체육관광부.
- 한국문화정보원 [발행년불명]. 문화빅데이터 플랫폼. 출처: <https://www.bigdata-culture.kr/>
- Barkan, O. & Koenigstein, N. (2016). ITEM2VEC: neural item embedding for collaborative filtering. 2016 IEEE 26th International Workshop on Machine Learning for Signal Processing (MLSP), 1-6.
- Blondel, V. D., Guillaume, J., Lambiotte, R., & Lefebvre, E. (2008). Fast unfolding of communities in large networks, *Journal of Statistical Mechanics: Theory and Experiment*, P10008. <https://doi.org/10.1088/1742-5468/2008/10/p10008>
- International Federation of Library Associations and Institution (1994). Public Library Manifesto.
- Le, Q. V. & Mikolov, T. (2014). Distributed representations of sentences and documents. *Proceedings of Machine Learning Research*, 1188-1196.
- Newman, M. E. J. (2004). Detecting community structure in networks. *The European Physical Journal B*, 38, 321-330. <https://doi.org/10.1140/epjb/e2004-00124-y>
- Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv*. <https://doi.org/10.48550/arXiv.1301.3781>
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. *arXiv*. <https://doi.org/10.48550/arXiv.1310.4546>
- Mikolov, T., Yih, W. T., & Zweig, G. (2013) Linguistic regularities in continuous space word representations. *NAACL HLT*. 746-751.

• 국한문 참고문헌의 영문 표기

(English translation / Romanization of references originally written in Korean)

- Cho, Jane (2019). Analysis of reading domian of men and women elderly using book lending data. *Journal of Korean Library and Information Science Society*, 50(1), 23-41. <https://doi.org/10.16981/kliss.50.1.201903.23>
- Kim, Chul Jin, Jeong, JiHyun, Jo, Cheon-Woo, & Yoo, Je Kwang (2019). A performance evaluation analysis of product recommendation techniques. *Journal of Knowledge Information Technology and Systems*, 14(5), 515-525. <https://doi.org/10.34163/jkits.2019.14.5.008>
- Kim, Seonghee & Lee, Seungmin (2022). A study on the perception of librarians and library users on the future direction of public library development. *Journal of the Korean Biblia Society for Library and Information Science*, 33(1), 499-519. <https://doi.org/10.14699/kbiblia.2022.33.1.499>
- Kim, Wan Jong (2014). A study on collection use of an public libraries focused of the clustering analysis of circulation statistics of the seoul borough a library users. *Journal of the Korean Society for Information Management*, 31(3), 353-369. <https://doi.org/10.3743/KOSIM.2014.31.3.353>
- Korea Culture Information Service Agency [n.d.]. BigData MarketC. Available: <https://www.bigdata-culture.kr/>
- Korea Institute of Science and Technology Information (2015). 2014 establishment of a system to utilize big data analysis in libraries. Ministry of Culture, Sports and Tourism.
- Kwak, Chul wan (2016). A study of cutter's expansive classification. *Journal of the Korean Society for Library and Information Science*, 50(3), 249-265. <https://doi.org/10.4275/KSLIS.2016.50.3.249>
- Lee, Jongwook, Kang, Woojin, & Lee, Myeong (2021). The effects of socioeconomic deprivation on public library book circulation: a community-level study. *Journal of the Korean Society for Library and Information Science*, 55(4), 219-243. <https://doi.org/10.4275/KSLIS.2021.55.4.219>
- Lee, Jongwook, Kang, Wooin, & Park, Jungkyu (2021). The effects of the bestseller ranks on public library circulation: based on panel data analysis. *Journal of the Korean Society for Information Management*, 38(4), 1-23.

<https://doi.org/10.3743/KOSIM.2021.38.4.001>

Lee, Soon-young & Lee, Soosang (2021). A study on big data analysis of public library in Busan: based on the library collection/circulation data. *Journal of the Korean Society for Library and Information Science*, 55(4), 89-114.

Libraries Act. Act No. 18763.

Lim, Jeonghoon, Cho, Changje, & Kim, Jongheon (2022). A study on the development of the school library book recommendation system using the association rule. *Journal of the Korean Society for Information Management*, 39(3), 1-22.

<https://doi.org/10.3743/KOSIM.2022.39.3.001>

Ministry of Culture, Sports and Tourism [n.d.]. Data For Library. Available:

<https://www.data4library.kr/>

On, Jeongmee & Park, SungHee (2020). Big data analysis for public libraries utilizing big data platform: a case study of Daejeon Hanbat Library. *Journal of the Korean Society for Information Management*, 37(3), 25-50.

<https://doi.org/10.3743/KOSIM.2020.37.3.025>

Shim, Jiyoung (2021). Identifying information needs of public library users based on circulation data: focusing on public libraries in seoul. *Journal of the Korean Society for Information Management*, 38(2), 173-199.

<https://doi.org/10.3743/KOSIM.2021.38.2.173>

Yang, Suwan (2020). Quality diagnosis of library-related open government data: focused on book details api of data for library. *Journal of the Korean Society for Information Management*, 37(4), 181-206. <https://doi.org/10.3743/KOSIM.2020.37.4.181>

Yeo, Ji-suk, Kong Sunghoon, & Oh, Dong-Geun (2013). Improvements and modifications of the subject, architecture engineering, in the 5th edition of the Korean decimal classification. *Journal of the Korean Society for Library and Information Science*, 47(2), 359-376. <https://doi.org/10.4275/KSLIS.2013.47.2.359>

