목차 정보와 kNN 분류기를 이용한 사회과학 분야 도서 자동 분류에 관한 연구

A Study on Book Categorization in Social Sciences Using kNN Classifiers and Table of Contents Text

이용구 (Yong-Gu Lee)*

초 록

이 연구에서는 한 대학도서관의 신착 도서 리스트 중 사회 과학 분야 6,253권에 대해 목차 정보를 이용하여 자동 분류를 적용하였다. 분류기는 kNN 알고리즘을 사용하였으며 자동 분류의 범주로 도서관에서 도서에 부여한 DDC 300대 강목을 사용하였다. 분류 자질은 도서의 서명과 목차를 사용하였으며, 목차는 인터넷 서점으로부터 Open API를 통해 획득하였다. 자동 분류 실험 결과, 목차 자질은 분류 재현율과 분류 정확률 모두를 향상시키는 좋은 자질임을 알 수 있었다. 또한 목차는 풍부한 자질로 불균형인 데이터의 과적합 문제를 완화시키는 것으로 나타났다. 법학과 교육학은 사회 과학 분야에서 특정성이 높아 서명 자질만으로도 좋은 분류 성능을 가져오는 점도 파악할 수 있었다.

ABSTRACT

This study applied automatic classification using table of contents (TOC) text for 6,253 social science books from a newly arrived list collected by a university library. The k-nearest neighbors (kNN) algorithm was used as a classifier, and the ten divisions on the second level of the DDC's main class 300 given to books by the library were used as classes (labels). The features used in this study were keywords extracted from titles and TOCs of the books. The TOCs were obtained through the OpenAPI from an Internet bookstore. As a result, it was found that the TOC features were good for improving both classification recall and precision. The TOC was shown to reduce the overfitting problem of imbalanced data with its rich features. Law and education have high topic specificity in the field of social sciences, so the only title features can bring good classification performance in these fields.

키워드: 목차, kNN 분류기, 도서 범주화, DDC (Dewey Decimal Classification)
Table of contents, kNN classifier, book categorization, DDC (Dewey Decimal Classification)

* 계명대학교 문헌정보학과 부교수(yonggulee@kmu.ac.kr)

■ 논문접수일자 : 2020년 2월 27일 ■ 최초심사일자 : 2020년 3월 8일 ■ 게재확정일자 : 2020년 3월 24일
■ 정보관리학회지, 37(1), 1-21, 2020. http://dx.doi.org/10.3743/KOSIM.2020.37.1.001

1. 서 론

기계 학습에 의한 자동 분류(automatic classification) 기법은 다양한 분야에서 각광받고 있다. 구체적으로 빅데이터나 데이터 과학(data science)을 비롯하여 다수의 영역에서 자동 분류 기법을 활발하게 활용하고 있다. 자동 분류의 대상도 텍스트뿐만 아니라 이미지 등 범위를 넓혀가고 있다. 이러한 모습은 자동 분류가 좋은 성능을 가져오고 있음을 방증하기도 한다. 특히 최근 인공지능 기술이 적용된 자동 분류 기법은 기존 다른 기법에 비해 더 좋은 성능을 보여주는 경향이 있어 더욱 주목받고 있다.

어떤 대상물을 기계에 의해 자동으로 분류하거나 범주화(categorization) 하기 위해서는 일반적으로 그에 따른 적절한 자질(feature)이 필요하다. 이는 좋은 자질이 분류 성능에 직접적으로 영향을 미치기 때문이다. 예를 들어 전문(full-text)으로 구성된 텍스트를 미리 지정된 범주(주제)로 자동 분류하고자 할 때, 일반적으로 이들 텍스트에 출현한 많은 수의 단어들을 자질로 사용한다. 구체적으로 학술 논문의 경우 기본적으로 표제뿐만 아니라 초록과 본문을 대상으로 자질을 추출한다. 즉 짧은 표제에서 추출한 소수의 자질을 포함할 뿐만 아니라 초록과 본문의 다수의 자질을 포함한다. 짧게 압축되어 소수의 단어로 표현된 표제뿐만 아니라 초록이나 본문에 나타난 풍부한 자질이 분류 성능에 많은 영향을 미치기 때문이다. 이렇듯 분류 성능을 향상시키기 위해서는 보다 좋은 자질의 수를 일정 정도 증가시키는 것이 필요하다.

도서관은 단행본 도서의 이용을 촉진하고 높이기 위해 다양한 서비스를 개발하고 제공해야 한다. 일반적으로 단행본 도서가 도서관 장서에서 가장 많은 부분을 차지하기 때문에, 이용자가 이들을 더 많이 이용할수록 도서관 경영 측면에서는 더 효율적이라 할 수 있다. 현재 이용자는 자신의 정보요구를 해결하기 위해 주로 도서관의 온라인 열람 목록(OPAC)에서 검색 활동을 수행한다. 이를 위해 도서관은 자관의 목록 규칙에 따라 필요한 메타데이터를 생성하고 도서관 시스템 또는 온라인 열람 목록에서 해당 자원이 검색되도록 지원한다. 즉 도서관 이용자는 자신이 정보요구를 충족하기 위해 이들 메타데이터를 사용하여 해당 도서 자원을 식별하고 탐색한다.

자동 분류 방법이 다양한 분야에서 활용되듯이 도서관의 단행본 도서를 대상으로 적용하여 기존의 도서관 서비스를 개선시키거나 새로운 서비스를 시도할 필요가 있다. 도서를 자동 분류할 때 앞서 말했듯이 좋은 분류 성능을 가져오려면 기계 학습에 필요한 적절한 자질이 필요한데, 도서관에서 생성한 도서에 대한 메타데이터를 살펴보면, 전문 형태의 추출 가능한 자질은 대부분 표제(서명)로 한정된다. 자동 분류의 성능 측면에서 이러한 상황을 살펴보면 대체로 짧은 문장인 표제만으로 좋은 성능을 가져오기 어렵다는 뜻이다.

도서 관련하여 보다 넓게 추출 가능한 메타데이터를 고려해 보면, 도서의 내용을 대표하면서 풍부한 자질을 제공하는 목차가 있다. 다행히 최근에는 도서 목차 정보에 쉽게 접근할 수 있다. 국립중앙도서관은 담당자에 의하면 예산 허용 범위 안에서 매년 3-40만 건의 단행본 도서

에 대해 목차 정보를 구축 제공하고 있으며, 한국교육학술정보원(KERIS)도 대학도서관의 종합목록에 대해 부분적으로 목차 정보를 제공하고 있다. 다수의 국내 대학도서관도 그들의 홈페이지에서 직접 또는 간접적으로 목차 정보를 제공하고 있다. 사실 인터넷 서점들도 적극적으로 목차정보를 구축하고 제공하는데, 이는 구매자가 도서를 구입할 때 목차 정보를 통해 구입 여부에 대한 판단에서 도움 주기 때문인 것으로 보인다.

도서관 장서의 대부분을 차지하는 단행본 도서에 대해 새로운 서비스 개발 측면에서 목차 정보의 다양한 활용 가능성을 알아보는 것이 필요하다. 이러한 배경에서 이 연구는 도서관의 신착 자료 목록에 포함된 도서의 서명과 목차를 이용하여 자동 분류를 수행하였다. 이를 통해 목차 텍스트가 자동 분류에서 가지는 특성을 파악하고자 하였다. 이때 도서관 환경에서 활용 가능한지 알아보기 위해, 도서에 부여된 DDC(Dewey Decimal Classification)의 분류기호를 자동 분류의 범주로 삼았다. 또한 분류기로는 텍스트를 대상으로 자동 분류를 수행하는데 많이 사용되며, 비교적 이해하기 쉽고 구현이 간단한 kNN(k-Nearest Neighbor) 알고리즘을 적용하였다.

2. 이론적 배경

단행본 도서를 자동 분류할 때 자질로 도서의 목차를 활용하려면, 도서의 목차가 가지는 통계적 특성이나 언어학적 특성을 살펴볼 필요가 있다. 먼저 Winke(1999)는 미국 의회도서관 목록으로부터 648권의 도서를 추출하여 목차가 가지는 주요 통계적 특성을 분석하였는데, 이들 도서 중 92.75%가 목차를 가지며 평균적으로 67.75개의 단어를 포함하는 것으로 파악하였다. 또한 대부분의 목차가 한 단계 또는 두 단계의 계층 구조를 가진다고 하였다. 비록 조사 대상의 전체 도서 수가 648권으로 비교적 적지만, 대부분의 도서가 목차를 가진다는 것에 주목할 필요가 있다.

국내 사회과학 분야 도서를 대상으로 목차 텍스트의 형태소 분석과 통계적 특성을 파악한 연구(이용구, 2019)에서는 도서의 서명과 목차 텍스트에서 추출한 형태소 중, 명사가 대략 절반 정도를 차지하며 동사는 상대적으로 그 비율이 낮다는 것을 제시하였다. 특히 서명과 비교하여 목차는 어절이나 명사 빈도가 대략 50배 정도 더 많은 어휘를 갖는 것으로 파악하였으며, 서명에는 없고 목차에만 유일하게 출현한 명사의 비율은 95.2%에 달하는 것으로 파악하였다. 이렇듯 목차에서 다수의 명사가 출현하는 특징은 그 목차를 자동 분류의 자질로 충분히 사용가능하다는 것을 보여준다고 할 수 있다. 다만 목차가 주제적 측면에서 도서의 내용을 나타내기 위해 어떠한 특성을 가지는지 좀 더 연구할 필요가 있다.

목차는 도서에 나타난 장이나 절의 내용을 대표하는 문장으로 구성되어 있는데, 이용자의 정보 과업 측면에서 다음과 같은 특징을 가진다. 우선 특정 도서를 검색하는 이용자는 그 도서가 자신의 정보요구에 부합하는지 판단하기 위해서 목차 정보를 활용할 수 있다. 또한 검색 시스템은 목차에 출현한 단어를 색인어로 추가하여, 이용자가 특정 도서를 탐색할 때 검색 성능

을 높일 수 있다(Pappas & Herendeen, 2000; Chercourt & Marshall, 2013). 이러한 측면은 저자가 도서를 생산하고 그 도서의 주요한 내용을 목차로 압축하여 나타내기에 충분히 가능하다고 볼 수 있다.

실제 검색시스템에 목차 정보를 반영하여 검색 성능 평가를 수행한 연구들을 살펴보면, 다음과 같다. Dillon과 Wenzel(1990)은 검색 효과 측면에서 목차 텍스트를 추가하여 개선시킨 서지 레코드가 그렇지 않은 일반 서지 레코드보다 10%의 재현율을 끌어올렸으나 정확률은 낮아져 최종적으로 소폭의 검색 성능만 향상되는 결과를 가져왔다. 유사하게 Van Orden(1990)도 디지털 시스템에서 목차와 초록 같이 내용을 풍부하게 지니는 구성요소가 짧은 표제보다 많은 단어를 포함하여 검색될 경향이 높으나, 낮은 정확률을 동반한다고 주장하였다. 이러한 결과로 볼 때, 이용자에게 목차는 단순히 표제나 저자 정보를 넘어 그들의 정보탐색 과정에서 자원 발견의 기회를 높이는 도구로서 역할을 한다고 볼 수 있다. 마찬가지로 목차에 출현한 풍부한 단어는 자동 분류와 같은 특정 영역에서 도서에 대한 다양한 활용을 가능하게 할 수 있다.

도서의 분류, 주로 주제 분류를 자동으로 수행하고자 하는 초기 연구는 규칙에 기반한 방법을 적용하거나, 통계에 기반한 방법을 적용하였다. 특히 다수의 연구들이 사용한 방법은 정보검색 기법이다(Larson, 1992; Godby & Stuler, 2003). 이 연구들은 현재의 자동 분류 기법과는 차이를 보이는데, 주로 기존의 장서에서 분류대상 도서와 가장 유사한 도서나 문헌을 검색하여 그 도서의 분류기호를 이용하였다. 구체적으로 가장 유사한 문헌을 검색하고 그 분류기호를 새로 분류해야 할 도서나 문헌의 분류기호로 부여하였다. 즉 기계 학습에 의해 분류기호를 자동으로 부여하기 보다는 검색의 결과로 가장 유사도가 높은 도서의 분류기호를 사용하는 방식이다.

선행 연구 측면에서 보면, 도서를 기술하는 서지 데이터를 이용한 자동 분류 연구도 비교적 소수이지만, 도서 자동 분류에서 기계 학습 또는 범주화(categorization)에 의한 지도 학습(supervised learning)을 이용한 연구는 더욱 적다. 서지 데이터를 이용하여 도서관에서 사용하는 DDC나 LCC(Library of Congress Classification)와 같은 분류체계로 자동 분류하는, 즉 도서의 분류기호를 자동으로 할당하려는 연구들은 다음과 같다.

우선 Frank와 Paynter(2004)는 대학도서관의 목록 레코드를 이용하여 자동 분류 실험을 수행하였는데, 이를 통해 특정 도서에 대해 LCC 분류기호를 자동으로 부여하였다. 이때 목록 레코드에서 사서가 수작업으로 부여한 두 종류의 데이터를 추출하였다. 하나는 LCC 분류기호(MARC 필드 050 또는 090의 하위 필드 $a)이며, 다른 하나는 LCSH(Library of Congress Subject Headings)의 주제명(제2지시기호가 0인 650과 651 필드)이다. 즉 도서관 목록의 서지 레코드에 부여된 LCSH 주제명을 분류 자질로 사용하고, 마찬가지로 같은 서지 레코드에 부여된 LCC 분류기호를 분류할 범주(class)로 간주하여 자동 분류를 수행하였다. 분류기로 지지 벡터 기계(support vector machine: SVM) 분류기를 이용하였으며, 실험 데이터는 80만 건의 서지 레코드를 학습 집단으로 설정하고 5만

건을 검증집단으로 설정하여 자동 분류를 수행하였으며, 그 결과 분류 정확도(classification accuracy) 55%를 얻었다. 실험의 최종 목적으로 LCSH 주제명이 수작업으로 부여된 웹 자원을 자동 분류하여 LCC 분류기호를 할당하는 데 활용하였다. 이 연구의 경우 도서의 서명이나 목차를 분류 자질로 사용하지 않았다.

Wang(2009)도 서지 데이터를 대상으로 지도 학습 방법의 기계학습을 이용하여 DDC 분류기호(범주)를 자동으로 부여하고자 하였다. 학습과 검증에 필요한 실험 데이터는 미의회도서관(Library of Congress)이 1994년에서 2004년 동안 생성한 과학과 기술 분야(DDC의 주류 500과 600 분야)에 해당하는 88,440건의 레코드를 OCLC의 WorldCat 시스템으로부터 내려 받아 사용하였다. Wang은 데이터의 희소성 문제를 해결하기 위해 소수의 문헌을 갖는 분류 범주를 통합하고 수평적으로 배열하는 평탄화(flattening) 방법을 통해 DDC 분류체계를 재구조화하였다. 그 결과 SVM 분류기가 0.493의 마이크로평균 F_1 값을 얻었다. 또한, 비록 그 횟수를 제한하긴 하였지만, 분류 효과를 높이고자 앞서 재구조화된 계층적 구조에서 분류기가 올바른 경로를 따라가도록 인간이 개입하여 실험한 결과, SVM 분류기가 0.849의 마이크로평균 F_1 값을 얻었다. 다만, Wang은 계층 분류에서 인간의 개입을 허용하는 부분에 대해서 실용적 측면에서 접근이라고 주장하였으나, 이러한 개입으로 최상의 계층부터 최하위 범주 상의 경로에 대한 기계에 의한 분류가 없어져 결과적으로 최하위 계층에 대한 분류 문제로 회귀하게 된 한계를 가진다.

3. 자동 분류 실험 설계

3.1 실험집단 구축

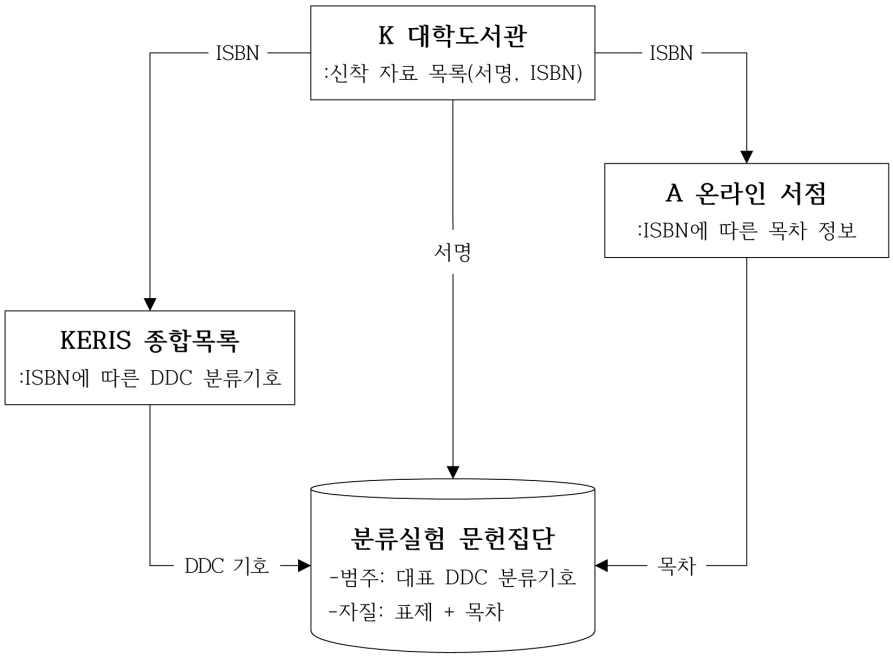
기계 학습에 의해 대상물을 자동 분류하기 위해서는 학습 데이터가 필요하다. 기계는 학습 데이터로부터 필요한 정보를 추출(학습)하여 분류기를 만들고 이를 자동 분류에 사용한다. 이때 학습 데이터는 기계 학습에 필요한 구성 요소로 개개의 분류 대상물과 그 대상물을 분류한 범주 체계를 포함한다. 범주 체계는 자동 분류를 적용하고자 하는 분야나 업무에 따라 달라진다. 예를 들어 정상 이메일과 스팸 이메일을 가려내야 한다면 스팸 여부가 범주가 된다.

이 연구의 목적은 도서를 자동 분류할 때 도서가 지니는 목차 정보의 활용 가능성을 알아보는 것이다. 이를 위해 분류 대상물을 목차를 가지는 도서로 한정했으며 범주 체계는 그 도서에 부여된 주제를 나타내는 분류기호를 사용하였다. 도서관은 수서되는 도서에 대해 분류·목록 작업을 통해 분류기호와 서지사항을 추출한다. 도서관에서 많이 채택하고 있는 십진분류법은 동일 또는 유사 주제를 다루는 도서에 대해 동일한 또는 유사한 분류기호를 부여하도록 그 규칙을 제시한다. 따라서 이러한 주제적 측면을 나타내는 분류기호는 자동 분류를 위한 학습 데이터의 범주 체계로 적합하다. 도서관이나 외주 기관에서 사서가 수작업으로 도서의 분류기호를 생성하며, 편목 과정을 통해 한 도서의 서지 정보를 추출한다. 일반적으로 이러한 데이터 중에서 서명이 자동 분류의 자질로 대표적으로 사용 가능하다.

이 연구에서 사용한 학습 데이터를 구축하기 위한 전반적인 과정은 〈그림 1〉과 같다. 우선 한 대학도서관의 신착 자료 목록을 도서관 홈페이지에서 수집하여 단행본 도서 리스트를 확보하였다. 이 리스트에서 도서의 서명과 국제표준도서번호(ISBN)를 추출하였다. 이렇게 추출된 ISBN를 이용하여 온라인 서점 API(application programming interface)를 통해 해당 도서의 목차 텍스트를 수집하였다. 또한 해당 도서의 대표 DDC 분류기호를 확보하기 위해 KERIS의 종합목록을 이용하였다. 이러한 과정을 통해 한 도서에 대한 주제 범주에 해당하는 대표 DDC 분류기호와 분류 자질에 해당하는 서명과 목차 정보를 조합하여 분류 실험에 필요한 문헌집단(학습 문헌과 검증 문헌)을 구축하였다. 각 과정의 구체적인 내용은 다음과 같다.

자동 분류에서 학습 문헌과 검증 문헌으로 사용할 단행본 도서 리스트를 얻기 위해 연구중심 대학교의 도서관에서 2014년과 2015년의 신착 자료 목록을 수집하였다. 실제 이러한 도서목록을 분류실험에 사용할 때 장점은 향후 도서 관련 자동 분류 또는 텍스트 범주화 기술을 도서관 시스템에 활용할 때 적용 가능성을 더 높일 수 있다는 것이다. 도서관 환경 속에서 실제 구축된 학습 데이터와 자동 분류 자체를 위해 구축된 데이터는 상황에 따라 매우 상이하고 최종 분류 성능도 달라질 수 있어, 이러한 소장 도서목록을 활용하는 것이 적용 가능성을 보다 높일 수 있다.

수집 대상 신착 자료 목록에 포함된 단행본 도서의 총 수는 45,705종인데, 이 목록은 다른 연구(이용구, 2019)에 사용된 것과 동일하다.



〈그림 1〉 분류 실험 문헌집단 수집 및 구축 과정

다만 해당 대학도서관의 웹사이트에서 메타데이터를 직접 크롤링하여 목록에 있는 도서에 대해 보다 상세한 서지 사항을 추출하였다. 이때 크롤링된 웹 페이지에서 기본적인 정보(자료유형, 서명/저자, 청구기호, 개인저자, 발행사항, 형태사항, ISBN, 내용주기, 주제 등)를 추출하였다. 그 결과를 주요 자료 유형을 중심으로 살펴보면 국내서 단행본이 32,394종이며, 외국서 단행본이 13,127종이었다. 특히 ISBN의 경우 한 도서의 다양한 이판(하드바운드와 페이퍼백, 또는 종이책과 전자책, 큰글씨책 등)에 각각 다른 번호가 부여되거나, 다권본 또는 총서로 구성된 도서의 경우 각 권에 부여된 낱권 ISBN과 전체에 부여된 세트 ISBN이 구별되어 제공되었다. 이러한 각각의 ISBN을 모두 합쳐 최종적으로 55,435개의 ISBN이 수집되었다. 이 중에서 세트 ISBN은 다수 반복되므로 중복을 제거하였다. 예를 들어 ISBN이 979-11-304-4100-9인 『커뮤니케이션이해총서』 세트는 낱권이 100종으로 이루어져 있어, 세트 ISBN이 100번 나타나므로 중복 제거하여 하나만 남겨두었다. 이러한 중복을 제거하여 최종적으로 53,694개의 고유한 ISBN를 추출하였다.

대학도서관의 신착 자료 목록에서 추출된 ISBN에 대해 목차 정보를 추출하기 위해 한 인터넷 서점의 OpenAPI를 이용하였다. 구체적으로 상품 조회 API를 통해 도서의 목차 정보를 수집하였다. 앞서 고유하게 추출된 ISBN에 대해 인터넷 서점에서 42,230개의 레코드가 조회되었다. 이 중에서 목차 정보를 가진 레코드, 더 정확히는 목차 콘텐츠의 길이가 1 이상인 레코드는 30,302건에 해당하였다. 이들 목차 콘텐츠는 인터넷 서점에서 웹으로 서비스되기 때문에 HTML 태그를 포함하고 있어 이를 제거하는 사전처리 작업을 수행하였다.

자동 분류에서 범주 체계의 일관성은 매우 중요하다. 대상물에 범주명(class label)을 부여하는 기준의 혼란은 분류기의 성능에 많은 영향을 미치기 때문이다. 일반적으로 도서관에서는 자관에 소장될 자료를 조직하는데 사용되는 분류 규칙이나 정책이 존재한다. 각각의 도서관이 처한 상황이 다르므로 자료 분류 규칙도 도서관마다 다를 수 있다. 또한 아무리 자관에서 정한 하나의 분류 규칙이나 정책을 적용하더라도 다수의 분류자가 존재하면 일관된 자료 분류는 어려울 수 있다. 따라서 이러한 한계를 다소 완화하기 위해 이 연구에서는 국내의 주요 도서관을 선정하고 이들이 특정 도서에 부여한 분류기호를 추출하여 가장 높은 빈도를 갖는 분류기호를 대표 분류기호로 적용하였다. 이는 KERIS 자체의 대표 분류기호를 적용한 이전 연구(이용구, 2019)와 다른 부분인데, 현장의 적용가능성은 높이고 분류기호의 오류 가능성은 낮출 것으로 의도하였다. 즉 이러한 방식으로 대표 DDC 분류기호를 추출하는 이유는, 한 도서의 범주명인 분류기호를 부여할 때 한 대학도서관의 분류기호 대신 다수의 도서관에서 가장 많이 부여한 분류기호를 사용하는 것이 개별도서관 또는 사서 개개인의 오류로부터 자유로워 해당 도서의 분류기호로 보다 신뢰성이 높을 것으로 판단하였기 때문이다.

도서에 대표 분류기호를 할당하는 과정을 설명하면 다음과 같다. 우선 앞서 추출된 ISBN에 대해 KERIS의 종합목록을 검색하고 해당 ISBN의 도서를 소장한 기관들을 추출하였다. 이때 소장 기관의 기준은 국가 대표도서관인 국립중앙

도서관과 9개의 대학도서관을 대상으로 하였으며 이들 기관의 DDC 분류기호를 추출하였다. 도서자료의 종수 상위 도서관을 기준으로 선정하되, DDC 분류를 자체적으로 수행하는 충분한 인력이 있는 대학도서관을 기준으로 하여 9개의 대학도서관을 선정하였다. 이를 위해 대학도서관의 통계 데이터를 제공하는 학술정보통계시스템(http://www.rinfo.kr/)에서 필요한 데이터를 내려 받아 활용하였다.

앞서 기술한 방법에 따라 대학도서관의 목록에서 고유하게 추출된 ISBN 53,694개에 대해 KERIS의 종합목록으로부터 10개 도서관의 분류기호를 추출하여 53,625개를 얻었다. 이들 레코드에 대해 강목 수준으로 대표 DDC 분류기호를 처리하였다.

실험 데이터의 구축 과정을 요약해보면, 대학도서관의 수서된 도서에 대해 ISBN를 이용하여 인터넷 서점으로부터 목차 정보를 추출하고, KERIS로부터는 대표 DDC 분류기호를 추출하였다. 최종적으로 특정 도서의 서명과 목차, 그리고 분류기호를 하나로 통합하여 전체 실험 집단을 구축하였다. 통합 과정에서 자동 분류 실험 데이터에 부합하게 다음의 몇 가지 원칙을 전체적으로 적용하다.

첫째, 다권본이나 총서의 세트 ISBN은 이들 ISBN이 목차를 포함하고 있더라도 최종 실험 데이터에서 제외하였다. 이유는 이들 ISBN에 분류기호가 부여되기보다 이들의 구성원인 낱권 ISBN에 분류기호가 부여되기에 적절한 주제범주로 볼 수 없기 때문이다. 즉 분류의 대상이 낱권에 해당하므로 세트 ISBN에 대한 분류기호는 적절하지 않을 수 있다.

둘째, 다권본의 경우 첫 번째 ISBN만 처리 대상으로 삼았다. 예를 들어 3권으로 이루어진 다권본이라도 해당 시리즈의 첫 번째 도서만 실험 대상에 포함하였다. 그 이유는 시리즈의 두 번째 이후의 도서의 목차가 첫 번째 도서와 유사할 확률이 높아 최종 분류 성능에 영향을 미칠 수 있다고 보았기 때문이다. 학습 문헌으로 다권본 중의 한 도서가 배정되면 다권본의 나머지 도서를 분류할 때 배정된 도서와 유사할 확률이 높아지므로 최종 분류 성능이 높아질 수 있다. 목차 유무에 관계없이 세트 ISBN과 두 번째 이후에 ISBN에 해당하는 레코드 수가 대략 15,000건에 해당하며 이를 제외하였다.

셋째, 서명과 목차가 영문으로만 이루어진 도서는 제외하였다. 이들 도서는 대개 외국서 단행본에 해당하는데, 대부분의 외국서는 목차를 수집하는 과정에서 제외되었다. 다만 목차가 수집된 일부 레코드가 있으며, 그 수가 2,271건에 해당하였다. 이 연구에서는 이들 도서를 제외하였다.

마지막으로, DDC 분류기호가 없거나 목차 텍스트를 형태소 분석하여 추출된 자질이 없는 경우 해당 레코드도 실험 데이터에서 제외하였다. 이들 레코드는 자동 분류에 반드시 필요한 범주 정보와 분류 자질을 갖지 않기에 자동 분류용 실험 데이터로 사용할 수 없다.

이러한 원칙에 의해 최종적으로 수집된 분류실험 문헌집단은 23,629건이 되었다. 이 중에서 가장 많은 수의 레코드가 수집된 분야가 6,253건에 해당하는 사회과학분야(DDC 300대)이었다. 기계 학습을 이용하는 자동 분류의 경우 일정 규모의 학습 문헌을 갖는 것이 중요하므로 이 연구에서는 DDC 주류 300대의 사회과학분

야를 분석 대상으로 선정하였다.

최종적으로 수집된 문헌집단의 서명과 목차에 대해 형태소 분석을 수행하였으며, 이를 통해 명사(고유명사와 일반명사)를 분류 자질로 추출하였다. 이때 사용한 형태소 분석기는 울산대학교 한국어처리연구실(http://nlplab.ulsan.ac.kr/)에서 공개한 UTagger 파이썬 공개 2018(우분투) 버전을 사용하였다.

3.2 분류기와 성능 평가 척도

텍스트 범주화 또는 자동 분류를 위해서는 기계가 학습 과정을 통해 분류기를 생성해야 한다. 기계 학습 분야에는 다양한 분류기가 사용되고 있다. 그중에서 kNN 분류기는 사례 기반 학습(instance-based learning) 또는 게으른 학습(lazy learning)의 유형에 속한다. 이 유형의 알고리즘은 입력 문헌을 실제 분류하기 전까지 학습 과정이 미루어진다.

kNN 분류기는 분류하고자 하는 입력 문헌에 대해 학습 문헌집합으로부터 유사도가 가장 높은 k개의 최근접 이웃문헌을 찾아낸 다음, 이 이웃문헌들에 할당된 범주 정보를 이용하여 입력문헌에 부여할 하나 이상의 범주를 선정한다(정영미, 2012, pp. 214-215). 이때 유사도 산출을 위한 유사계수나 거리계수를 선택해야 하며 가장 유사한 이웃의 개수를 나타내는 k값도 미리 선정해야 한다.

kNN 분류기에 적용된 알고리즘은 다양한 문제를 해결하기 위한 방법으로 사용되는데, 가장 흔하게는 분류(Yang & Lin, 1999)나 회귀(Altman, 1992), 최근에는 이상치 탐지(Campos et al., 2016) 등에도 사용된다. kNN 분류기의 성능을 향상시키기 위한 자질 선정에 대한 연구(이재윤, 2005; 이용구, 2013; Azam & Yao, 2012) 등도 있다.

자동 분류에서 분류기의 성능을 평가하기 위해 척도는 정보 검색의 성능 평가 척도와 유사하다. 하지만 그 계산 방식은 주로 2×2 분할표의 일종인 오차 행렬(confusion matrix)을 이용한다. 오차 행렬은 〈그림 2〉와 같이 분할표로 구성되는데, 이 그림에서 a 값은 분류기가 옳은 범주에 제대로 분류한 문헌의 수를 나타내며, b 값은 분류기가 틀린 범주에 잘못 분류한 문헌의 수를 나타낸다. c 값은 옳은 범주에 분류하지 못한 문헌 수이며, d 값은 틀린 범주에 분류하지 않은 문헌 수를 나타낸다(정영미, 2012, pp. 225-226).

		(문헌의) 실제 범주	
		옳은 범주	**틀린 범주**
(분류기의) 예측 범주	**범주에 분류**	*a* (True Positive)	*b* (False Positive)
	범주에 비분류	*c* (False Negative)	*d* (True Negative)

〈그림 2〉 분류 성능 평가를 위한 오차 행렬

분류기의 성능 평가 척도인 분류 재현율은 민감도(sensitivity)라고도 하는데 a/(a+c) 식으로 계산하며, 분류 정확률은 a/(a+b) 식으로 계산하다. 자동 분류에서는 분류를 정확하게 하는지를 파악하기 위해 민감도뿐만 아니라 특이도(specificity)를 동시에 평가한다. 이 척도는 정보 검색에 배제율에 해당하는데 틀린 범주에 분류하지 않아야 하는 정도를 나타내며 d/(b+d) 식으로 계산한다. 정확도(accuracy)는 (a+d)/(a+b+c+d)로 산출하여 전체 분류 중 옳은 분류의 비율을 나타낸다. 정확도의 경우 데이터 집단이 불균형을 가질 때 다수가 포함된 범주가 유리해지는 단점이 있어 신뢰성의 문제를 야기할 수 있다. 또 다른 단일가 척도로 F_1 척도나 BEP(break-even point) 등이 있다.

4. 자동 분류 실험

이 연구는 앞서 설명된 실험집단을 처리하고 자동 분류를 위한 자질을 처리하기 위한 프로그램으로 Python을 전반적으로 사용하였고, 도서의 서명과 목차로부터 추출한 자질을 이용하여 DDC 300대 강목의 주제 코드를 자동으로 부여하기 위한 분류기 프로그램은 Python용 기계학습 실험 패키지인 Scikit-learn 버전 0.22(Pedregosa et al., 2011)를 사용하였다.

4.1 실험집단 및 자질 분석

한 대학도서관에서 입수한 신착 자료 목록 중에서 목차 정보를 인터넷 서점으로부터 추출하고, KERIS 종합목록의 주요 대학도서관에서 제시한 DDC 분류기호가 300대 사회과학분야인 6,253권을 분석하여 〈표 1〉을 얻었다.

실험 대상 전체 도서 중에서 가장 높은 비율을 보이는 분야로는 DDC 33X대 강목인 '경제학'이 1,428권으로 22.8%를 차지했으며, 다음으로 34X대 '법학'이 1,419권으로 근소한 차이로 2위였으며, 30X대 사회학 및 인류학을 포함

〈표 1〉 실험집단의 분류기호 분석(DDC 300대 강목)

요목	30X	32X	33X	34X	35X	36X	37X	38X	39X	합계
0	60	286	303	148	6		200		12	1,015
1	92	26	101	35	115	122	224	26	21	762
2	155	5	403	248	86	180	116	49	15	1,257
3	161	37	96	182	27	106	10	2	6	627
4	26	39	14	114	4	78	7	47	38	367
5	167	35	31	182	92	2	17	2	5	533
6	283		57	410	1					751
7	49	135	41	82	1			34		342
8		9	341	5	3	35	61	12	58	524
9			41	13	5		11	5		75
합계 (비율 %)	993 (15.9)	572 (9.1)	1,428 (22.8)	1,419 (22.7)	340 (5.4)	523 (8.4)	646 (10.3)	177 (2.8)	155 (2.5)	6,253 (100)

하는 '사회과학 일반'(15.9%), 37X대 '교육학'(10.3%) 순으로 나타났다. 즉 '경제학'과 '법학' 분야 도서가 약 절반 정도 차지한 반면, 38X대 '상업, 통신, 교통'과 39X대 '풍속, 민속학'은 약 5%대에 머물러 학문별 주제 분야 사이의 매우 큰 차이를 보였다. 31X대 '통계학'의 경우 단행본 도서가 외국서만 수서되어 실험 집단에서 제외되었다. 이러한 편차는 대상 대학도서관의 수서 정책이나 활동 보다 출판 동향에 따른 원인에서 기인하는 것으로 보인다. 실제 이 연구의 실험 데이터를 수집한 대학도서관은 연구중심형 대학에 해당되고, 이 도서관의 국내서의 비중이 대학도서관 중에서 상위권에 해당하여 국내 단행본 도서를 거의 대부분 수집한다고 보아도 무방하기 때문이다.

DDC 주류 300대 도서에 대해 요목별로 살펴보면 346대 '상법'이 410권으로 가장 많고, 332대 '금융' 403권, 338대 '생산, 산업경제' 341권 순으로 나타났다. 다만 총류 성격의 30X대 사회과학 일반 분야의 요목별 비율은 다른 주제 분야에 비해 상대적으로 고른 모습을 보였다.

자동 분류에서 자질의 종류나 특성은 분류 성능에 많은 영향을 미치므로 이들의 통계적 특징을 살펴볼 필요가 있다. 도서의 서명이나 목차는 텍스트 또는 전문 형식을 띠므로 그 안에 출현하는 단어(자질)의 기본적인 기술 통계를 구하여 〈표 2〉를 얻었다. 다른 연구(이용구, 2019)에서도 제시하였듯이 전통적으로 서명과 목차는 명사 위주로 표현된다. 이 표에서 짧은 문장의 서명은 명사 중심의 자질을 평균 5.54개 가지며, 그에 비해 목차는 평균 237.4개를 가지는 것으로 나타났다. 도서에서 추출한 두 텍스트가 많은 차이를 보임을 알 수 있다. 한 단어가 여러 번 중복하여 출현하는 것을 제거한 고유빈도의 경우는 상대적으로 그 차이가 크지 않았다. 다만 표준편차 측면에서 출현빈도를 살펴보면 서명은 3.1개로 비교적 작으나 목차는 403.01개로 매우 커서 두 텍스트가 전혀 다른 모습을 보였다. 이를 통해 목차의 길이가 매우 다양함을 알 수 있다. 특히 목차의 출현빈도 최댓값이 6,757개로 매우 긴 사례도 존재함을 알 수 있다.

실험 집단의 통계적 특성을 강목별로 좀 더 자세히 살펴보기 위해 강목별 자질의 평균과 표준편차를 구하여 〈표 3〉을 얻었다. 출현빈도 측면에서 서명은 36X대가 6.29개로 평균이 가장

〈표 2〉 서명과 목차 자질의 기술 통계

	서명		목차	
	출현빈도	고유빈도	출현빈도	고유빈도
mean	5.54	5.27	237.40	102.41
std	3.10	2.83	403.01	108.35
min	0	0	1	1
25%	3	3	64	35
50%	5	5	134	72
75%	7	7	262	131
max	36	27	6,757	1,363

〈표 3〉 300대 강목별 자질의 평균과 표준편차

강목	서명				목차			
	출현빈도		고유빈도		출현빈도		고유빈도	
	mean	std	mean	std	mean	std	mean	std
30X	5.13	2.81	4.84	2.61	159.74	137.59	94.35	74.18
32X	5.26	2.74	4.98	2.53	170.83	172.59	97.95	85.40
33X	6.00	3.12	5.70	2.89	339.26	630.55	135.15	147.26
34X	5.05	3.14	4.86	2.82	261.82	456.86	93.62	117.60
35X	5.86	2.99	5.64	2.78	215.18	230.76	91.04	77.18
36X	6.29	3.78	5.99	3.40	214.09	213.04	89.44	72.01
37X	5.72	2.71	5.39	2.44	201.40	216.93	91.43	79.23
38X	5.93	3.71	5.72	3.33	194.71	279.36	85.76	89.22
39X	5.04	2.67	4.79	2.45	145.03	147.27	82.96	81.17

높았으며 다음으로 33X대(6.0), 38X대(5.93) 순으로 나타났다. 목차는 33X대가 339.26개로 평균이 가장 높았으며, 34X대(261.92), 35X대(215.28) 순으로 나타났다. 이들은 표준편차도 커서 다른 강목 보다 목차의 길이가 다양함을 알 수 있다. 즉 이들 강목은 대체로 길이가 긴 목차를 가진 도서가 다수 존재한다는 것을 의미한다. 고유빈도 측면에서도 목차는 33X대가 135.15개로 평균이 가장 높았으며, 다음으로 32X대(97.95), 30X대(94.35) 순으로 나타났다. 다만 목차의 고유빈도에서 32X대와 30X대가 2위, 3위를 차지하여 독특한 특징을 보였다. 이들 강목은 출현빈도에서 하위권(각각 7, 8위)에 위치하는데 고유한 자질은 2, 3위에 위치하므로 이들 강목은 중복되는 단어가 비교적 적으며 고유한 단어가 상대적으로 많이 출현함을 알 수 있다. 이와 비교하여 34X대 강목의 목차는 평균에서 출현빈도를 기준으로는 2위이나 고유빈도는 4위에 위치한 것을 보면, 다른 강목에 비해 고유 단어의 중복 출현이 상대적으로 많음을 알 수 있다.

기계 학습을 통한 자동 분류를 하려면, 학습 집단(training set)과 검증 집단(test set)이 필요하다. 학습 집단은 분류기를 구축하는데 필요하며, 구축된 분류기를 검증 집단을 통해 성능을 평가한다. 이 연구는 〈표 4〉와 같이 6,253개의 실험 집단에 대해 DDC 300대 강목을 기준으로 학습 집단과 검증 집단을 8:2로 나누어 층화 무작위 추출(stratified randomization)로 처리하였다. 그 결과 학습 집단과 검증 집단은 각각 5,002개와 1,251개의 문헌으로 구성되었다. 학습 집단과 검증 집단의 문헌에서 추출한 자질에 대해 로그TF×IDF 가중치를 적용하였으며 kNN 분류기에서 최근접 이웃문헌을 결정하기 위한 유사도는 유클리디언(Euclidean) 거리계수를 적용하였다.

이 연구는 도서를 자동 분류하기 위해 분류 자질로서의 서명과 목차가 가지는 특성을 파악할 필요가 있다. 서명과 목차의 차이를 분석하기 위해 서명 텍스트, 목차 텍스트, 그리고 서명과 목차를 결합한 텍스트(서명+목차) 3가지 추출 방식으로 구성하였다. 각각의 텍스트에 대해 분류기를 구축하기 위해 문헌-자질 행렬을 작성하였는데, 학습 집단에서 서명만 이용

〈표 4〉 학습 집단과 검증 집단 비율

강목	학습 집단		검증 집단	
	문헌 건수	비율	문헌 건수	비율
30X	794	15.9%	199	15.9%
32X	458	9.2%	114	9.1%
33X	1,142	22.8%	286	22.9%
34X	1,135	22.7%	284	22.7%
35X	272	5.4%	68	5.4%
36X	418	8.4%	105	8.4%
37X	517	10.3%	129	10.3%
38X	142	2.8%	35	2.8%
39X	124	2.5%	31	2.5%
합계	5,002	100%	1,251	100%

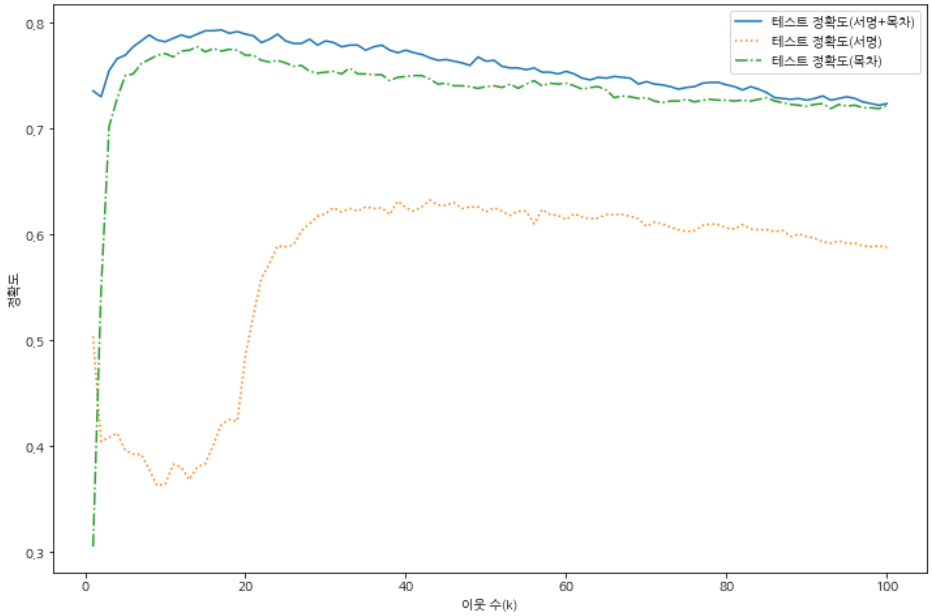
한 경우 자질의 크기는 5,648개였으며, 목차를 이용한 자질은 50,346개였으며, 서명과 목차를 결합한 경우(서명+목차)는 54,579개의 자질이 추출되었다.

4.2 분류 실험 결과

입력 문헌에 어떤 범주를 부여할지 결정하기 위해서 kNN 분류기는 주요하게 두 개의 매개변수의 설정이 필요하다. 하나는 입력 문헌과 유사도가 가장 높은 몇 개의 최근접 이웃문헌을 사용할 것인가, 즉 k값을 결정하는 부분이다. 다른 하나는 찾아낸 k개의 최근접 이웃문헌을 이용하여 어떤 방식으로 입력 문헌에 범주를 부여할 것인가이다. 후자의 경우는 찾아낸 최근접 이웃문헌에 부여된 범주 중 다수의 범주를 선택하는 투표 방식과 이웃문헌 간 유사도 가중치를 반영하여 결정하는 가중치 방식이 있다. 이 연구에서는 투표 방식을 적용하였다.

다른 주요 매개변수인 k값을 결정하기 위해서는 일반적으로 경험규칙(heuristics)을 적용한다. 즉 학습 집단을 이용하여 k값을 1부터 시작하여 일정한 값까지 적용하여 그 중에서 가장 좋은 성능을 가져오는 값을 결정한다. 학습 집단에서 서명, 목차, 서명+목차 3개 텍스트의 자질에 대한 최적의 k값을 결정하기 위해 사전 실험을 수행하였다. k값을 1부터 100까지 증가시키면서 각각 자질에 대해 정확도를 구하여 가장 좋은 정확도를 보이는 k값을 구하였다. 그 결과 〈그림 3〉과 같이 서명 자질은 43, 목차 자질은 14, 서명+목차 자질은 17일 때 가장 좋은 정확도를 보였다. 따라서 이후에 각각의 자질에 대해 kNN 분류기에 이들 값을 적용하여 분류 실험을 수행하였다. 서명 자질의 경우 작은 k값에서 낮은 정확도를 보이다가 30부터 상승하여 40 전후에서 가장 높은 성능을 보였다. 이와 달리 목차 자질의 경우 k값이 5부터 급격히 상승하여 10 전후에서 가장 좋은 성능을 보였다. 목차의 경우 소수의 최근접 이웃만으로도 좋은 성능을 보이나 서명은 소수 최근접 문헌만으로는 낮은 성능을 가져옴을 알 수 있다.

서명, 목차, 서명+목차 3가지 텍스트의 자질을 이용하여 DDC 300대 주류에 대한 자동 분류 성능을 다양한 척도로 나타내면 〈표 5〉와 같



〈그림 3〉 분류 자질에 따른 k값의 정확도

〈표 5〉 300대 주류 실험집단의 자질별 분류 성능

평가 척도	서명	목차	서명+목차
정확도(accuracy)	0.6323	0.7770	0.7930
balanced accuracy	0.4860	0.7204	0.7406
매크로 정확률	0.7527	0.7913	0.8018
매크로 재현율	0.4860	0.7204	0.7406
매크로 F_1	0.5254	0.7483	0.7630
마이크로 정확률	0.6323	0.7770	0.7930
마이크로 재현율	0.6323	0.7770	0.7930
마이크로 F_1	0.6323	0.7770	0.7930

다. 전체적으로 살펴보면 서명만 사용하여 자동 분류한 것보다 자질의 수가 많아지는 목차, 서명과 목차를 결합한 자질이 모든 평가 척도에서 좋은 성능을 보였다. 다만 세부적으로 들여다보면 척도별로 향상 정도가 차이가 있다. 먼저 성능 향상이 작게 이루어진 경우를 살펴보면 매크로 평균 정확률이 서명만 이용한 자질에 비해 목차만 이용한 자질에서 약 3.9% 정도 향상되었으며, 서명+목차의 자질에서는 약 4.9% 정도 향상되었다. 이와 비교하여 매크로 평균 재현율의 경우 서명 자질에 비하여 목차와 서명+목차 자질은 각각 약 23.4%, 약 25.5% 정도 향상되어 매우 큰 성능 향상이 이루어졌음을 알 수 있다. 또한 매크로 정확률과 재현율을 동일하게 가중치를 적용한 매크로 평균 F_1 값은 서명 자질에 비해 목차 자질이 약 22.3%, 서

명+목차 자질은 약 23.8% 정도 향상되었다.

전체적으로 목차가 분류 재현율이나 분류 정확률을 모두 향상시키는 것을 알 수 있다. 다만, 정확률은 소폭으로 향상되고 재현율이 큰 폭으로 향상된 것은 풍부한 자질을 갖는 목차의 특성이 자동 분류에서 반영된 것으로 보인다. 이는 목차가 재현율을 높이지만, 정확률을 떨어뜨려 결국 최종 성능에서 소폭의 상승을 가져오는 정보 검색 결과(Dillon & Wenzel, 1990; Van Orden, 1990)와는 차이가 있음을 알 수 있다. 이러한 결과는 자동 분류 분야에서 목차의 활용 가능성이 한층 더 높아진다는 것을 의미한다. 향후 목차 정보를 사용할 수 있는 환경에서 도서 분류를 고려한다면, 목차를 분류 자질로 적극적으로 활용해 볼 필요가 있음을 뜻한다.

〈표 5〉에서 하나의 자질 집합에 대해 동일한 문헌과 범주를 가지므로 마이크로 정확률, 재현율, F_1 값이 모두 같은 값을 가질 수밖에 없지만, 자질 종류 간의 차이를 보면 서명 자질에 비해 목차 자질은 마이크로 평균이 약 14.5% 정도, 서명+목차 자질은 약 16.1% 정도 분류 성능이 향상된 것을 알 수 있다. 또한 정확도와 balanced accuracy 모두 서명 자질보다 목차 자질과 서명+목차 자질이 더 좋은 분류 성능을 가져왔다. 범주 불균형에 보다 적합한 척도인 balanced accuracy가 더 많이 향상된 것을 알 수 있다.

앞서 자동 분류에서 목차 텍스트가 가지는 분류 성능을 사회과학 분야 주류(300대) 전체 수준에서 파악하였으나 이 분야의 세부 분야인 강목 단위로 목차 자질이 보이는 특성을 파악할 필요가 있다. 이를 위해 사회과학 분야의 강목별로 서명, 목차, 서명+목차 자질의 분류 성능을 분류 재현율, 분류 정확률, 그리고 F_1 척도를 중심으로 측정하면, 〈표 6〉과 같았다. 9개 강목에 대해 서명 자질의 F_1 분류 성능 척도는 큰 편차를 보인다. 아주 낮은 값인 39X대의 0.0625부터 다수의 강목이 0.4에서 0.5 사이에 값을 가진다. 반면 목차 자질과 서명+목차 자질은 대부분의 강목에서 고르게 높은 값을 갖는다. 최솟값이 0.6207이지만, 대부분의 강목은 주로 0.7과 0.8의 값을 가지는 것을 알 수 있다.

〈표 6〉 300대 강목에 따른 자질별 분류 성능

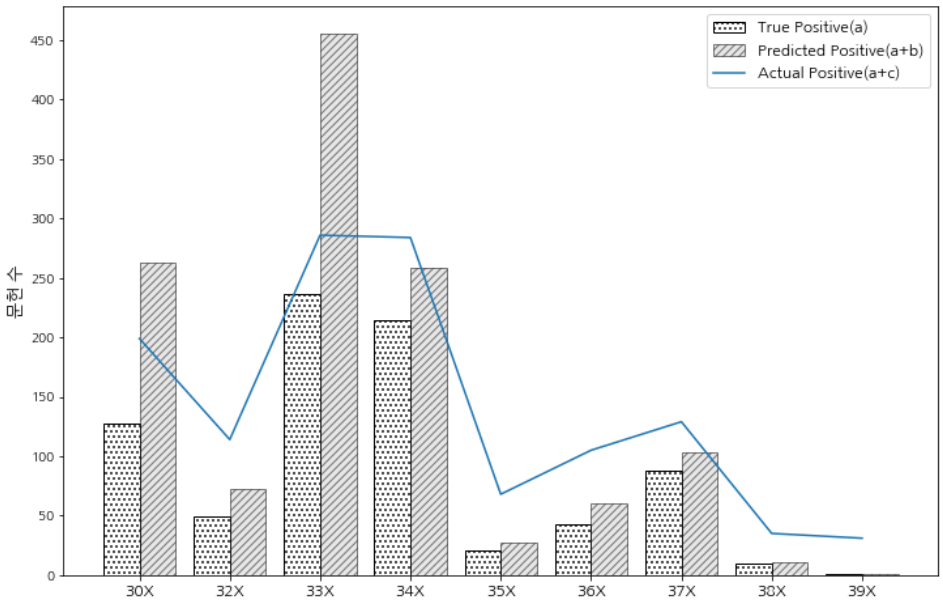
자질	척도	30X	32X	33X	34X	35X	36X	37X	38X	39X
서명	Precision	0.4867	0.6806	0.5197	0.8295	0.7778	0.7167	0.8544	0.9091	1.0000
	Recall	0.6432	0.4298	0.8287	0.7535	0.3088	0.4095	0.6822	0.2857	0.0323
	F_1	0.5541	0.5269	0.6388	**0.7897**	0.4421	0.5212	0.7586	0.4348	0.0625
목차	Precision	0.6959	0.7143	0.7900	0.8070	0.8475	0.6739	0.8582	0.7826	0.9524
	Recall	0.6784	0.7018	0.8287	0.8979	0.7353	0.5905	0.8915	0.5143	0.6452
	F_1	0.6870	0.7080	0.8089	0.8500	0.7874	0.6294	**0.8745**	0.6207	0.7692
서명+목차	Precision	0.7293	0.7167	0.7915	0.8644	0.8125	0.6939	0.8298	0.7778	1.0000
	Recall	0.6633	0.7544	0.8497	0.8979	0.7647	0.6476	0.9070	0.6000	0.5807
	F_1	0.6947	0.7350	0.8196	**0.8808**	0.7879	0.6700	0.8667	0.6774	0.7347

먼저 단일 척도 F_1을 중심으로 살펴보면 서명 자질에서 가장 좋은 성능을 보인 분야는 34X대 '법학'으로 0.7897 값을 보였으며, 다음으로 '교육학' 0.7586, '경제학' 0.6488 순이었다. 목차 자질은 서명 자질과 유사하게 37X대 '교육학'이 0.8745로 가장 좋은 성능을 보였으며, 다음으로 '법학' 0.8500, '경제학' 0.8089 순이었다. 서명+목차 자질은 0.8808로 '법학'이 1위이며, 다음으로 '교육학' 0.8667, '경제학' 0.8196 순이었다. 3가지 자질에서 '법학'과 '교육학'은 번갈아 1,2위를 자치하였으며, '경제학'이 모두 3위를 차지했다.

'법학'과 '교육학'이 3가지 자질에서 모두 좋은 분류 성능을 보인 이유를 좀 더 자세히 분석하기 위해 이들 강목의 분류 재현율과 분류 정확률을 자세히 살펴볼 필요가 있다. 이들 강목의 정확률은 자질이 적은 서명 텍스트부터 매우 높은 것을 알 수 있다. 즉 짧은 문장으로 소수의 자질로 구성된 서명에서도 0.8 이상의 정확률을 보이는데, 이는 해당 강목에 부여된 다수의 문헌이 서명만으로도 정확하게 분류되는 것을 의미한다. 그 이유를 실례로 들어보면 '법학'의 경우 서명 안에서 '민법', '형법' 등 각종 법의 명칭이나 '판례', '로스쿨' 등과 같이 법학을 나타내는 용어가 출현하였다. 이들 단어는 9개 강목 내에서 특정성이 높은 단어에 해당한다. 비슷하게 37X대(교육학)도 '공부', '교사', '학교' 등과 같은 교육학을 나타내는 단어가 많이 출현하였다. 즉 이 두 분야는 300대 사회과학 분야에서 다른 세부 분야보다 주제적으로 특정성이 높아 소수의 자질 내지 용어만으로도 높은 분류 성능을 가져오는 것을 알 수 있다. 이러한 특성은 〈표 3〉의 출현빈도와 고유빈도에 대한 분석에서도 찾아볼 수 있다. 요약하면 '법학'이나 '교육학' 처럼 특정성이 높은 분야는 소수의 자질만으로도 충분히 좋은 분류 성능을 가져올 수 있음을 의미한다.

이와 달리 '경제학'의 경우 정확률은 상대적으로 낮으나 재현율이 높아 F_1 척도가 3위의 높은 분류 성능을 보였다. 물론 목차 자질과 서명+목차 자질에서는 정확률과 재현율 모두 높았으나 여전히 둘 중에서 재현율이 더 높은 것을 알 수 있다. 이러한 주된 원인은 학습 문헌의 범주별 불균형과 그로 인한 과대적합(overfitting)으로 해석할 수 있다. 실제 실험 결과의 오차 행렬을 분석하면, 소수의 자질을 갖는 서명에서도 '경제학' 분야는 실험 집단의 문헌 수의 대략 1/3에 해당하는 456건이 33X대로 분류되었다. 이를 성능 척도로 설명하면 '경제학' 분야가 정확률은 다른 강목에 비해 상대적으로 낮지만 재현율은 매우 높은 값을 가짐을 알 수 있다. 전체 9개의 강목에 대한 좀 더 구체적인 서명 자질의 오차 행렬의 수치를 그래프로 나타내면 〈그림 4〉와 같았다. 경제학(33X)의 경우 검증 집단의 문헌 수(Actual Positive)가 286개이며 kNN 분류기가 이 분야로 분류한 문헌 수(Predicted Positive)가 456개인데, 이 중에서 정확하게 분류한 문헌 수(실선인 True Positive)가 237개였다. 각각의 수치를 〈그림 2〉의 오차 행렬의 표기 방식으로 표현하면 a+c, a+b, a에 해당한다.

〈그림 4〉에는 33X대의 '경제학' 뿐만 아니라 다른 강목에 대한 수치들도 제시되었는데, 과적합을 보이는 33X대 '경제학'과 비슷한 패턴을 보이는 강목이 30X대인 '사회과학 일반'이다. 이 강목은 학습 문헌이 794개(15.9%)로 3위를 차지하는데, 이러한 많은 학습 문헌에 비



〈그림 4〉 서명 자질을 이용한 강목별 분류 결과

해 서명과 목차 자질 모두에서 분류 성능이 향상되지 않았다. 이러한 주된 이유로 30X대 강목의 오분류(False Positive)에서 찾을 수 있는데, 주로 39X대, 36X대, 32X대 강목에 해당하는 도서가 30X대 강목으로 분류되었다. 실제로 39X대 강목의 지역의 풍속이나 민속 관련 주제와 30X대 강목의 요목 306(Culture & institutions)과 307(Communities)의 지역명과 관련 주제에서 지명 같은 공통된 용어를 가지며, 30X대 강목이 가지는 다수의 학습 문헌으로 인해 분류기가 오류를 일으키는 것으로 추정할 수 있다.

〈표 6〉을 보면, 실험 집단에서 소수의 문헌을 가진 35X대, 38X대, 39X대 강목의 분류 성능 F_1 척도를 보면 서명 자질에 비해 목차 자질에서 각각 34.5%, 18.6%, 70.7%가 향상된 것을 알 수 있다. 이것은 목차 자질에서 재현율이 큰 폭으로 향상되었기 때문이다. 실제 서명 자질에서 이들 소수 문헌을 가진 강목들의 재현율은 모두 30% 이하였다. 적은 학습 문헌과 동시에 짧은 서명에서 기인하는 부족한 자질로 낮은 분류 성능을 가져오는 것을 알 수 있다. 따라서 소수의 문헌을 가진 강목들이라도 목차 자질을 사용하면 분류 성능이 큰 폭으로 향상되는 것을 알 수 있다.

5. 결 론

이 연구는 점점 활용 가능성이 높아지고 있는 목차 정보를 도서의 자동 분류에 활용하고자 하였다. 연구를 위한 데이터로 도서관에서 개별 도서에 대해 사서가 수작업으로 부여한 주제 코드인 DDC의 분류기호와 서지사항 중에 키워드 추출이 가능한 표제(서명), 그리고 인

터넷 서점의 Open API를 통해 획득한 목차를 사용하였다. 분류기호는 분류기가 최종적으로 분류할 범주명(class label)으로 간주하고, 분류기가 기계 학습에 필요한 자질로 도서의 서명과 목차를 활용하였다. 대학도서관 신착 자료 목록에서 수집한 사회 과학 분야(DDC 300대 주류) 도서 6,253권을 자동 분류 실험을 위한 문헌 집단으로 선정하였으며, 이 중에서 무작위 층화 추출을 통해 20%에 해당하는 1,251권의 도서를 검증 문헌 집단으로 사용하였다. 자동 분류에 사용될 분류기로는 텍스트 자질에 적합하며 이해하기 쉬운 모형인 kNN 분류기를 적용하였다.

연구 결과를 정리하면 다음과 같다. 첫째, 분류 성능을 평가하는 다양한 척도에서 서명만 사용한 것보다 목차 또는 서명과 목차를 결합한 자질을 사용한 자동 분류가 더 우수한 성능을 보였다. 구체적으로 정확도, balanced accuracy, 매크로 평균의 정확률과 재현율 및 F_1 척도, 그리고 마이크로 평균 척도 등에서 최저 3.9%에서 최고 25.5%까지 모두 좋은 성능을 보였다. 일반적으로 재현율과 정확률을 동시에 향상시키기 어려운 데, 목차는 자동 분류에서 재현율과 정확률 모두를 높여 좋은 자질임을 알 수 있다. 또한 서명과 목차를 결합한 자질도 목차만 사용한 경우보다 모두 더 좋은 성능을 보였다.

둘째, '경제학'이나 '법학'과 같은 DDC 300대 강목 분야별로 분류 재현율과 분류 정확률 그리고 F_1 척도에서 많은 성능의 차이를 보였다. 단일 척도인 F_1 척도를 살펴보면 서명 자질에서는 '법학', 목차 자질에서는 '교육학'이 가장 좋은 성능을 보였다. 이는 이 분야에 출현한 용어들이 특정성이 높아 좋은 분류 성능을 가져오는 것으로 파악하였다. 주제적으로 특정성이 높은 분야 또는 강목은 좋은 분류성능을 가져옴을 유추할 수 있어, 향후 사회 과학이 아닌 다른 주류에서도 이러한 현상이 보이는지 파악할 필요가 있다.

셋째, DDC 300대(사회 과학 분야) 강목인 학문들 사이의 출판물 성장 비율 차이로 인해 범주별 실험 문헌 집단의 불균형이 존재하여 자동 분류에서 과적합의 문제를 초래하였으나 목차 자질의 사용은 이러한 문제를 해결하는 것으로 보인다. 이 연구에서는 가장 많은 문헌(도서) 수를 가지는 경제학(33X대 강목)이 분류 자질이 적은 서명에서 높은 재현율을 보였으나 반대로 정확률은 과대 적합으로 다소 낮은 값을 보였다. 하지만 목차에서는 높은 재현율은 유지하면서 정확률이 향상되어 이러한 사실을 뒷받침하는 것으로 보인다.

넷째, 반대로 소수의 문헌을 가진 강목들이라도 목차 자질을 사용한 경우에 재현율과 그에 따른 F_1 척도의 분류 성능이 큰 폭으로 향상되었다. 이러한 재현율의 큰 폭의 향상은 목차의 풍부한 자질에서 기인하는 것으로 보인다.

향후 보다 더 큰 실험 집단을 구축할 필요가 있다. 이 연구에서는 최종적으로 6,253권의 실험 데이터를 구축하였는데, 다른 선행연구처럼 더 큰 규모의 실험 데이터를 구축하고 자동 분류 실험을 수행한다면, 이러한 연구결과를 일반화 하고 목차의 보다 다양한 특성을 분석할 수 있을 것으로 생각된다.

또한 주제 범위를 DDC의 사회 과학 분야뿐만 아니라 전 주제를 대상으로 자동 분류 실험을 수행하고 각 주류별 특성을 파악하거나 분류 성능을 향상시킬 수 있는 다양한 분류기의

적용을 시도해 볼 가치가 있다. 더 나아가 DDC 분류 체계의 특성을 반영하는 자동 분류 연구가 필요하다. 예를 들어 주제가 총류의 성격을 띠거나 2가지 이상을 다룰 때 십진 분류 체계에서는 하나의 분야로 분류를 해야 하는데, 이는 분류 전문가도 쉽지 않다. 이러한 분류 문제의 어려움은 여전히 자동 분류에서도 나타난다고 볼 수 있다. 이러한 문제는 향후 보다 많은 연구를 통해 해결 되어야 할 것으로 보인다.

참 고 문 헌

이용구 (2013). 문헌빈도와 장서빈도를 이용한 kNN 분류기의 자질선정에 관한 연구. 한국도서관·정보학회지, 44(1), 27-47. http://dx.doi.org/10.16981/kliss.44.1.201303.27

이용구 (2019). 사회과학 분야 도서의 목차 텍스트에 대한 통계적 특성에 관한 연구. 정보관리학회지, 36(2), 255-273. http://dx.doi.org/10.3743/KOSIM.2019.36.2.255

이재윤 (2005). 자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 대한 연구. 한국문헌정보학회지, 39(2), 123-146. http://dx.doi.org/10.4275/kslis.2005.39.2.123

정영미 (2012). 정보검색연구(증보판). 서울: 연세대학교 출판문화원.

Altman, N. S. (1992). An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. The American Statistician, 46(3), 175-185. http://dx.doi.org/10.1080/00031305.1992.10475879

Azam, N., & Yao, J. (2012). Comparison of term frequency and document frequency based feature selection metrics in text categorization. Expert Systems with Applications, 39(5), 4760-4768.

Campos, G. O., Zimek, A., Sander, J., Campello, R. J. G. B., Micenková, B., Schubert, E., … & Houle, M. E. (2016). On the evaluation of unsupervised outlier detection: Measures, datasets, and an empirical study. Data Mining and Knowledge Discovery, 30(4), 891-927. https://doi.org/10.1007/s10618-015-0444-8

Chercourt, M., & Marshall, L. (2013). Making keywords work: Connecting patrons to resources through enhanced bibliographic records. Technical Services Quarterly, 30(3), 285-295. http://dx.doi.org/10.1080/07317131.2013.785786

Dillon, M., & Wenzel, P. (1990). Retrieval effectiveness of enhanced bibliographic records. Library Hi Tech, 8(3), 43-46. https://doi.org/10.1108/eb047797

Frank, E., & Paynter, G. W. (2004). Predicting library of congress classifications from library

of congress subject headings. Journal of the American Society for Information Science and Technology, 55(3), 214-227. https://doi.org/10.1002/asi.10360

Godby, C. J., & Stuler, J. (2003). The library of congress classification as a knowledge base for automatic subject categorization. In Subject Retrieval in a Networked Environment: Proceedings of the IFLA Satellite Meeting, Dublin, OH, 14-16.

Larson, R. R. (1992). Experiments in automatic library of congress classification. Journal of the American Society for Information Science, 43(2), 130-148. https://doi.org/10.1002/(SICI)1097-4571(199203)43:2<130::AID-ASI3>3.0.CO;2-S

Pappas, E., & Herendeen, A. (2000). Enhancing bibliographic records with tables of contents derived from OCR technologies at the american museum of natural history library. Cataloging & Classification Quarterly, 29(4), 61-72. http://dx.doi.org/10.1300/J104v29n04_05

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12(Oct), 2825-2830.

Van Orden, R. (1990). Content-enriched access to electronic information: Summaries of selected research. Library Hi Tech, 8(3), 27-32. https://doi.org/10.1108/eb047795

Wang, J. (2009). An extensive study on automated dewey decimal classification. Journal of the American Society for Information Science and Technology, 66(11), 2269-2286. https://doi.org/10.1002/asi.21147

Winke, R. C. (1999). An analysis of tables of contents in recent english-language books. Library Resources & Technical Services, 43(1), 14-27. http://dx.doi.org/10.5860/lrts.43n1.14

Yang, Y., & Lin, X. (1999). A re-examination of text categorization methods, In: Proceedings of the 22nd annual international ACM SIGIR conference on research and development in the information retrieval(1999), 42-49.

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

Chung, Young-Mee (2012). Research in information retrieval. Seoul: Yonsei University Press.

Lee, Jae Yun (2005). An empirical study on improving the performance of text categorization considering the relationships between feature selection criteria and weighting methods.

Journal of the Korean Society for Library and Information Science, 39(2), 123-146. http://dx.doi.org/10.4275/kslis.2005.39.2.123

Lee, Yong-Gu (2013). A study on feature selection for kNN classifier using document frequency and collection frequency. Journal of Korean Library and Information Science Society, 44(1), 27-47. http://dx.doi.org/10.16981/kliss.44.1.201303.27

Lee, Yong-Gu (2019). A study on the statistical characteristics for table of contents text of the books in social sciences field. Journal of the Korean Society for Information Management, 36(2), 255-273. http://dx.doi.org/10.3743/KOSIM.2019.36.2.255