메타데이터를 활용한 기록물 자동분류 성능 요소 비교*

Comparison of Performance Factors for Automatic Classification of Records Utilizing Metadata

김영범 (Young Bum Gim)**
장우권 (Woo Kwon Chang)***

초 록

이 연구의 목적은 기록물의 맥락정보를 담고 있는 메타데이터를 활용하여 기록물 자동분류 과정에서의 성능요소를 파악하는데 있다. 연구를 위해 2022년 중앙행정기관 원문정보 약 97,064건을 수집하였다.수집한 데이터를 대상으로 다양한 분류 알고리즘과 데이터선정방법, 문헌표현기법을 적용하고 그 결과를 비교하여 기록물 자동 분류를 위한 최적의 성능요소를 파악하고자 하였다. 연구 결과 분류 알고리즘으로는 Random Forest가, 문헌표현기법으로는 TF 기법이 가장 높은 성능을 보였으며, 단위과제의 최소데이터 수량은 성능에 미치는 영향이 미미하였고 자질은 성능변화에 명확한 영향을 미친다는 것이 확인되었다.

ABSTRACT

The objective of this study is to identify performance factors in the automatic classification of records by utilizing metadata that contains the contextual information of records. For this study, we collected 97,064 records of original textual information from Korean central administrative agencies in 2022. Various classification algorithms, data selection methods, and feature extraction techniques are applied and compared with the intent to discern the optimal performance-inducing technique. The study results demonstrated that among classification algorithms, Random Forest displayed higher performance, and among feature extraction techniques, the TF method proved to be the most effective. The minimum data quantity of unit tasks had a minimal influence on performance, and the addition of features positively affected performance, while their removal had a discernible negative impact.

키워드: 기록물 분류, 기록물 자동분류, 자동분류, 문헌분류, 메타데이터
records classification, records automatic classification, automatic classification, document classification, metadata

---

* 이 연구는 기록관리학 석사학위논문을 수정・요약한 것임.

** 전남대학교 대학원 기록관리학 석사(smartb112@naver.com) (제1저자)
*** 전남대학교 문헌정보학과 교수(wk1961@jnu.ac.kr) (교신저자)

■ 논문접수일자 : 2023년 8월 16일 ■ 최초심사일자 : 2023년 9월 12일 ■ 게재확정일자 : 2023년 9월 12일
■ 정보관리학회지, 40(3), 99-118, 2023. http://dx.doi.org/10.3743/KOSIM.2023.40.3.099

1. 서 론

AI(Artificaial Intelligence)의 시대가 시작되었다. Gates(2023)는 인공지능 기술이 컴퓨터, 인터넷, 휴대전화만큼이나 근본적으로 사람들의 일상을 바꿀 것이며, 모든 산업과 기업이 인공지능 중심으로 재편될 것이라고 하였다. 실제로 인공지능 기술은 음성인식, 자동번역, 추천알고리즘 등 다양한 형태로 삶에 녹아들어 세계의 패러다임을 바꾸는 핵심 기술이 되고 있다. 이러한 변화는 문헌정보・기록관리 영역에서도 예외는 아니어서 다양한 인공지능 기술을 접목한 많은 연구가 수행되고 있다.

그중에서도 특히 업무 효율화를 위한 자동분류에 관한 연구는 1960년대 이후 꾸준히 이루어졌다. 그러나 자동분류는 비용을 절감하는 데에는 효과적이었지만 수작업에 의한 분류만큼의 정확성을 담보하지는 못한다는 한계를 지니고 있었다. 공공기록관리분야에서 이러한 문제는 더 치명적이었는데, 정부기능분류체계(Business Reference Model, 이하 BRM)에 의한 기록물분류가 이루어지는 기록관리 현장에서 오분류(誤分類)는 오분류에 그치지 않기 때문이다. 분류된 공공기록물은 범주화된 단위과제에 따라 기록물평가와 연동되며 이는 곧 보존기간 책정에 직결된다(설문원, 2013). 책정된 보존기간을 기준으로 기록물의 보존/폐기 여부가 결정되기 때문에 공공기록관리분야에서 분류의 정확성은 더욱 강조될 수밖에 없었다.

그러나 정확성을 위한 수작업 분류도 막대한 기록물이 그 대상이 되어 실질적인 분류가 이루어지지 못하는 환경에서는 정확성을 담보할 수 없었다. 실제로 국가기록원(2021)에서 소장 경제기록물 10,049건을 분석한 결과 78%에 달하는 기록물에서 오분류 가능성이 제기되었다. 국가기록원은 매년 대량으로 이관되는 공공기록물을 수작업으로 분류해야 하는 환경에서 결국 이관기록물의 양을 감당하지 못하고 일부 기록물만 우선순위에 따라 선별적으로 검수하고 있었다(정지혜 외, 2022). 소수의 기록관리 전문요원이 대량의 기록물을 수작업으로 분류해야 하는 환경에서 이러한 문제는 필연적인 결과였다.

대량의 기록물을 수작업으로 분류하는 데에서 비롯되는 업무부담을 경감하면서도 정확도를 담보하는 방법으로 방재현(2018)이 제안한 것은 지능형 아카이브 시스템의 주요 개념 중 하나로 인공지능 기술을 접목한 반자동화 기록물분류 서비스였다. 이 서비스는 인공지능을 활용하여 자동분류를 수행하면서도 분류 결과를 의사결정 지원 도구로만 사용하여 업무담당자가 최종 판단을 내릴 수 있는 구조로 이러한 개념적 진전을 통해 기록관리 분야에서의 자동분류 적용방안이 구체화되었다. 그럼에도 불구하고 기록물 자동분류에 관한 연구는 미비했는데, 특히 기록물의 특성을 반영하여 자동분류 과정에 적용되는 다양한 기법들을 비교하고 최적의 성능을 보이는 요소를 탐색하는 연구는 수행되지 않았다.

따라서 이 연구에서는 기록물의 특징인 맥락정보를 반영할 수 있도록 분류자질로 메타데이터를 활용하고, 다양한 기법들을 적용하여 기록물 자동분류에서 최적의 성능을 보이는 요소는 어떤 것인지 파악하는 것을 목적으로 한다. 자동분류에서 사용되는 주요 단계인 문헌표현단계, 데이터선정단계, 분류기생성단계에서 적

용 가능한 다양한 기법에 따른 성능변화를 비교하여 기록물 자동분류를 위한 최적의 방법이 무엇인지 분석한다.

이 연구는 기록물 자동분류를 위한 최적 성능 요소 도출을 통해 기록관리 현장에서 적용 가능한 최선의 자동분류 방법을 밝힌다는 점에서 의의가 있다. 방대한 양으로 인해 실질적인 분류가 이루어지지 못하는 기록물을 대상으로 최적성능요소를 기반으로 한 자동분류를 수행할 수 있다면 최소한의 분류 작업이 이행되어 체계적인 기록관리가 이루어지는 데에 기여할 수 있을 것이다.

2. 이론적 배경

2.1 기록물 분류제도

기록물은 문헌과 달리 내용보다 맥락이 중요하므로 분류 시에도 유관 주제를 한데 모은 주제분류보다 생산기관의 구조와 생산부서의 기능을 중요시하는 기능분류가 사용된다(이원영, 2000). 이러한 기록물의 특성은 공공기관의 기록물 분류제도에 그대로 반영이 되었다. 1999년 제정된 「공공기관의 기록물관리에 관한 법률」(이하 「공공기록물법」)은 공공기록관리에 조직과 업무에 따른 분류가 이루어질 것을 규정하였으며 2007년 전부개정된 「공공기록물관리에 관한 법률 시행령」(이하 「공공기록물법 시행령」)에서는 BRM을 도입하며 기능분류의 원칙을 더욱 강조하였다. Business Reference Model을 직역하면 업무참조모형인데, 효율적인 정부기능 관리를 위해 조직구조보다는 업무 흐름 중심으로 체계를 재정의하는 것을 의미하며 그 결과 도출된 것이 정부기능분류체계이다. BRM을 정부기능분류시스템이라고 지칭하기도 하는데, 「정부기능의 분류 및 관리에 관한 규정」에 따르면 "행정기관이 기능분류업무를 전자적으로 처리할 수 있도록 지원하고, 연계정보 등의 공동 활용을 확대하기 위한 정보화시스템"을 말한다. 또한, 기능분류란 기능별 분류체계 및 목적별 분류체계를 작성·관리하여 행정기관의 업무 및 관련 정보를 체계적으로 활용할 수 있도록 기능 중심으로 분류하는 것을 의미한다.

기록물분류에서 2007년 「공공기록물 법 시행령」 개정 전과 후의 가장 큰 차이는 기록물분류기준표에 따른 기록물분류에서 기록관리기준표에 따른 기록물분류로 변경된 점이다. 국가기록원이 분류방법을 제시하고 대기능, 중기능, 소기능, 단위업무, 단위사안에 따라 5단계로 분류단계를 구분하며 단위업무를 기준으로 운영하던 기록물 분류제도를 BRM에 따라 각급기관이 분류방법을 제시하고 정책분야, 정책영역, 대기능, 소기능, 단위과제 등 6단계로 분류단계를 구분하며 단위과제를 기준으로 분류하도록 변경하였다. 이를 정리하면 〈표 1〉과 같다(국가기록원, 2012).

2.2 선행연구

공공기록물의 분류를 위해서는 앞서 살펴본 바와 같이 기록관리기준표에 의거하여 분류단계의 최하위 범주인 단위과제를 기준으로 하는 기능분류가 수행되어야 한다. 그러나 1960년대 시작된 자동분류에 관한 연구는 문헌을 주된

〈표 1〉 기록물 분류제도 변경 구분

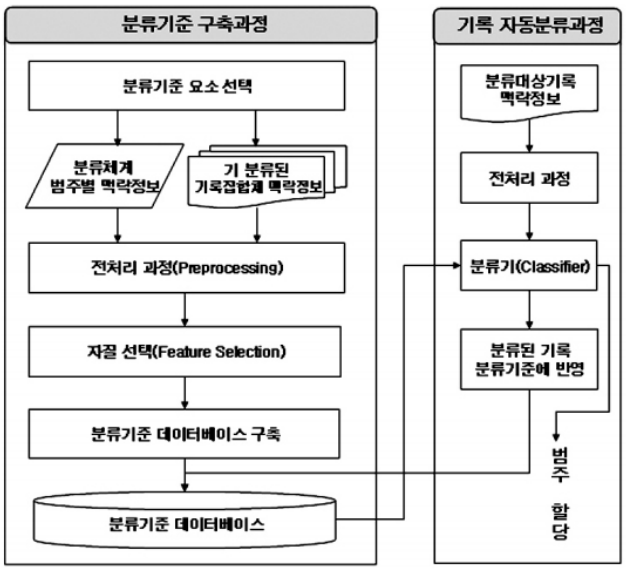
	기록물분류기준표	기록관리기준표
사용년도	2000 ~ 2006	2007 ~
근거	공공기관의 기록물관리에 관한 법률 시행령 [대통령령 제16609호]	공공기록물 관리에 관한 법률 시행령 [대통령령 제19985호]
분류방법	국가기록원 분류기준	정부기능분류체계(BRM)
고시주체	국가기록원	각급기관
분류기준	단위업무	단위과제
분류단계	(5단계) 대기능, 중기능, 소기능, 단위업무, 단위사안	(6단계) 정책분야, 정책영역, 대기능, 중기능, 소기능, 단위과제

대상으로 삼아왔기 때문에 문헌의 내용을 중심으로 하는 주제분류가 기본적인 개념이었다(최윤수, 최성필, 2019). 국내에서 본격적인 기록학에 관한 연구가 이루어진 것이 2000년 이후이고 2010년대 초반까지도 기록물분류라는 개념에 보편적인 합의가 이루어지지 않았던 점(설문원, 2012)을 감안하면 2000년대 후반 기록물 자동분류에 관한 연구가 수행된 것은 주목할만하다.

장지숙, 이해영(2009)은 문헌 자동분류에 적용된 다양한 이론과 기법을 기록 자동분류에 접목하고자 하며 기록 자동분류 분야에 새로운 지평을 열었다. 문헌과 기록물이 지닌 고유의 속성과 그에 따른 분류 원칙의 차이를 밝히고 기록물 자동분류에서 중요한 분류기준 요소를 정리하고자 하였다. 기록물 자동분류에서 중요한 것은 문헌 분류에서 중요했던 내용정보가 아니라 BRM, 메타데이터, 기능시소러스 등의 맥락정보임을 강조하며 이를 활용한 자동분류 시스템을 〈그림 1〉과 같이 제안하였다.

남은경, 안혜림, 송민(2013)은 실질적인 기록관리가 이루어지지 않는 행정기관 웹사이트 게시물의 문제를 해소하기 위해 자동분류 시스템을 구현하고자 하였다. 단어 출현빈도와 BRM 단계별 가중치를 활용하여 자동분류를 수행하였는데, 분류 결과에 대해 성능평가를 수행한 결과 52%의 정확도(accuracy)를 확인할 수 있었다. 이는 별도의 분류 알고리즘을 적용하지 않고 자체 계산식을 사용하여 도출된 결과로 SVM 등의 분류 알고리즘 적용 필요성을 암시했다.

본격적인 기록물 자동분류 연구는 차세대 기록관리 모델 재설계 연구(국가기록원, 2017; 김해찬솔 외, 2017)가 수행되며 이루어졌다고 볼 수 있다. 해당 연구는 세종 테그셋 기반 시퀀스 레이블링(sequence labeling) 형태소분석과 Structural SVMs 분류 알고리즘을 사용하여 대기능을 분류한 결과 최대 97%의 정확도를 보이는 놀라운 결과의 자동분류를 수행할 수 있었다. 또한, 오분류 분석을 통해 더 높은 정확도와 이를 위해 선행되어야 할 고려사항들을 제시하였다. 그러나 이 결과는 높은 정확도의 자동분류를 수행하기 위해 수동으로 변별력 있는 대기능 선별, 조정 및 통합을 통해 얻은 결과로 실제 기록관리 실무에서 적용되기에는 한계가 있었다.



〈그림 1〉 기록 자동분류시스템 구성도

상기 연구들은 저마다의 특정 방법론을 사용하여 기록물 자동분류 연구를 수행하였지만 정교한 성능 측정을 진행하지 않았거나 진행하더라도 단일 사례의 성능을 확인하는 데 그쳤다. 최적의 기록물 자동분류를 위해서는 정교한 성능 측정뿐만 아니라 통제된 환경에서의 다양한 방법론 적용과 그에 따른 결과의 비교가 필요하다. 따라서 이 연구에서는 여러 가지 분류 알고리즘 및 데이터선정기준과 문헌표현기법의 변경에 따른 성능을 비교하며 기록물 자동분류를 위한 최적의 성능 요소를 파악하고자 한다.

3. 연구설계

3.1 연구문제

이 연구는 기록물을 대상으로 한 자동분류를 수행할 때 주요 단계별 최적 성능 요소가 어떤 것인지 탐색하고자 하였다. 이를 위해 자동분류의 주요 단계를 문헌표현단계, 데이터선정단계, 분류기학습단계로 구분하고 각 단계에 따라 여러 가지 기법들을 적용하고 그 결과를 비교하였다. 이에 따른 연구문제는 다음과 같다.

- [연구문제 1] 분류기학습단계에서 가장 우수한 성능을 보이는 분류 알고리즘은 무엇인가?
- [연구문제 2] 데이터선정단계에서 데이터 선정 방법에 따른 성능 변화는 어떠한가?
- [연구문제 3] 문헌표현단계에서 가장 우수한 성능을 보이는 문헌표현기법은 무엇인가?

연구문제 1에서는 일반적으로 문헌분류에 사용되는 Support Vector Machines(이하 SVM),

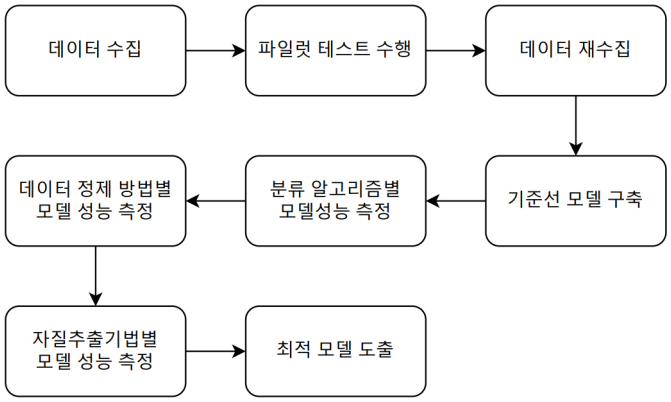
k-Nearest Neighbor(이하 kNN), Naive Bayes(이하 NB), Dicision Tree(이하 DT)와 DT기반의 앙상블 모델로 배깅(bagging)과 부스팅(boosting)을 활용하여 좋은 성능을 보여주고 있는 대표적인 두 가지 분류 알고리즘인 Random Forest(이하 RF)와 XGBoost(이하 XGB) 등 총 6개 기계학습기법을 scikit-learn 라이브러리를 통해 사용하였다. 또한, 도출된 최적 분류 알고리즘을 활용하여 다음 연구문제를 탐구한다. 연구문제 2에서는 종속변수인 단위과제의 범주 선정 기준 변경에 따른 성능 변화와 독립변수인 자질의 변경에 따른 성능 변화 및 자질별 영향력을 파악한다. 데이터선정단계는 단어 자질 집합인 텍스트뿐 아니라 기관명 혹은 생산일자와 같은 메타데이터 자질의 변경과 범주선정기준에 따라 변화하는 전체 데이터셋의 크기 및 성능을 포괄하는 단계이다. 범주선정기준은 자동분류에 사용되는 범주를 선정하는데 필요한 최소 데이터 수량을 의미한다. 연구문제 3에서는 문헌표현기법으로 TF 기법과 TF-IDF 기법, TF-ICF 기법 등 총 3가지 기법을 사용하고 그에 따른 성능을 비교하여 최적의 문헌표현기법이 무엇인지 확인한다.

연구에 사용된 기법들의 성능을 파악하기 위한 주요 지표로는 정확도(accuracy)를 사용하였다. 그러나 분류 알고리즘의 경우, 알고리즘별 특성에 따라 정확도 외에 정밀도(precision) 혹은 재현율(recall)에서 우수한 성능을 보이는 알고리즘이 있을 수 있으므로 추가적인 성능지표로 정밀도와 재현율을 종합적으로 고려한 F1-Score를 사용하였다. 특히, 각 범주에 대한 F1-Score를 계산한 뒤 그 값들의 산술평균으로 도출되는 macro 방식과 범주별 데이터 비율에 따른 가중평균으로 도출되는 weighted 방식을 이용하여 추가로 성능 평가를 진행하였다. 또한, 원활한 모델 구성을 위해 분류 알고리즘별 훈련시간 및 예측시간을 측정하였다. 따라서 최적 분류 알고리즘 탐색을 위해서는 정확도와 macro F1, weighted F1과 훈련시간 및 예측시간을 복합적으로 고려하여 최적 분류 알고리즘을 도출하였다.

3.2 연구 절차

최적 구성요소 파악을 위한 가장 이상적인 방법은 자동분류 과정에서의 모든 구성요소를 조합하며 성능을 비교하는 것이다. 그러나 이러한 방식은 각 단계의 구성요소 가지 수를 모두 곱하며 기하급수적으로 늘어나는 모든 경우의 수를 분석해야 하므로 현실적으로는 무리가 있다. 실제로 자동분류에 관한 선행연구들을 보면 대부분 기존 연구에 비해 개선된 성능을 보이는 새로운 기법을 제안하거나(육지희, 송민, 2018), 성능요소를 비교하더라도 분류 알고리즘별 비교 혹은 문헌표현기법별 비교와 같이 동일한 단계 내에서 비교하는 차원의 연구(김판준, 2016)들이 수행되고 있었다. 따라서 이 연구에서는 선행연구들을 기반으로 기본이 되는 모델인 기준선 모델(baseline model)을 먼저 구축하고 자동분류 각 과정에서 적용 가능한 다양한 방법들에 따른 결과를 과정별로 비교함으로써 각 과정에서 가장 높은 성능을 보이는 방법이 어떤 것인지 파악하고자 하였다.

이에 따른 구체적인 연구흐름도는 〈그림 2〉와 같다. 먼저 데이터를 수집한 뒤, 파일럿 테스



〈그림 2〉 연구흐름도

트를 수행하여 연구설계의 수정사항을 파악하고자 하였다. 파일럿 테스트로 파악된 개선사항을 반영하기 위해 데이터를 재수집하고 재수집 데이터를 사용하여 본 연구를 수행하기 위한 기준선 모델을 구축하였다. 이후 기준선 모델을 기반으로 연구문제에 따른 분류 알고리즘별 모델 성능, 데이터선정방법별 모델성능, 문헌표현 기법별 모델 성능을 측정하고 기록물 자동분류를 위한 최적성능요소를 도출하고자 하였다.

3.3 연구대상

이 연구는 정보공개포털(open.go.kr)에 공개된 51개 중앙행정기관의 2022년 공개원문정보를 대상으로 하여 기록물 자동분류 성능요소를 비교하였다. 수집 대상을 공개정보로 설정하여 개인정보 침해의 여지를 방지하는 한편, 수집데이터에 대한 접근성과 연구에 대한 재현가능성을 보장하고자 하였다. 데이터의 수집범위는 2022년 1월 1일부터 2022년 12월 31일까지 1년으로 설정하여 최신기록물 분류현황을 반영하고자 하였다. 또한, 원문정보에서 제공하고 있는 메타데이터를 주요 연구대상으로 설정하였다. 자동분류를 위한 다양한 연구가 이루어지고 있지만 대부분의 연구는 원문 전체를 연구범위로 설정하고 있다(정지혜 외, 2022). 그러나 기록물의 효과적인 관리를 위해서는 메타데이터를 대상으로 하는 연구가 필수적이며(이현실, 한성국, 2006) 앞서 살펴본 바와 같이 기록물 분류에 있어 중요한 점은 맥락정보에 의한 기능분류가 이루어져야 한다는 점이다.

원문정보에서 제공되는 메타데이터 목록은 제목, 기관명, 담당부서명, 담당자명, 생산일자, 문서번호, 보존기간, 단위업무, 공개여부, 분류체계, 본문파일 등과 같다. 이 중 연구에 적절하지 않은 메타데이터를 제외하여 제목, 기관명, 담당부서명, 담당자명, 생산일자, 문서번호, 분류체계 등의 메타데이터만 사용하였다. 보존기관과 공개여부는 기록관리기준표 작성 및 관리 절차에 따라 단위과제에 귀속되어 책정 및 설정되는 속성이 있으므로 연구에서 제외하였다. 또한, 본문파일은 파일명으로 내용이 결정되는데 이는 제목과 중복되는 값이고, 첨부파

일은 기록물에 따라 존재여부의 차이가 발생하므로 연구에서 배제하였다. 마지막으로 단위업무는 1999년 「공공기록물법」이 제정되며 기록물분류기준표에 따른 분류기준으로 기능하였지만 개정된 현행 기록관리체계에서는 분류기준으로 사용하지 않으므로 연구에 사용하지 않았다.

3.4 데이터 수집 및 분석

연구 데이터는 일차적으로 2023년 3월 20일 오전 12:00부터 같은날 01:00까지 크롤링 기법을 활용하여 정보공개포털에서 제공하는 2022년의 원문정보 데이터를 수집하였다. 작업 수행 결과 제목, 기관명, 담당부서명, 담당자명, 생산일자, 문서번호, 단위과제 등의 7개 메타데이터로 구성된 95,930건의 데이터를 수집할 수 있었다. 수집된 데이터의 자질별 데이터 수량을 살펴보면 〈표 2〉와 같이 40,597개의 제목명과 51개의 기관명, 2,093개의 담당부서명, 8,831건의 담당자명, 350건의 생산일자, 49,292건의 문서번호, 5,329건의 단위과제를 확인할 수 있었다. 그런데 문서번호는 개별문서에 부여되는 고유값이므로 전체 데이터에 미달하는 문서번호는 미달된 수량만큼 데이터가 중복되어 있다는 것을 의미한다. 즉, 문서번호가 확인된 49,292건의 기록물 외 46,638건의 데이터는 중복데이터라는 것이다.

중복데이터의 발생 원인을 살펴본 결과 정보공개포털에서 제공하는 웹페이지의 문제를 발견하였다. 예를 들어 2022년 1월 3일의 전체 원문정보 428개를 살펴본 결과 22페이지까지는 정상적인 서비스를 제공하고 있었으나, 절반을 조금 초과하는 페이지인 23페이지부터 마지막 페이지인 43페이지까지의 원문정보는 모두 동일한 리스트를 제공하고 있었다. 이러한 문제는 2022년 1월 3일의 원문정보뿐 아니라 다른 일자의 원문정보 대부분에서도 동일하게 발생하고 있는 문제였다. 따라서 파일럿 테스트 과정에 있어서는 문서번호를 기준으로 중복데이터 46,638건을 제거하고 49,292건의 데이터만 활용하여 연구를 진행하였다.

파일럿 테스트 진행 과정에서 중복데이터가 상당수 관찰된다는 것이 확인되었기 때문에 이 문제를 해결하기 위해 별도의 데이터수집 방법이 요구되었다. 이에 따라 이차 데이터 수집에 있어서는 웹페이지에서 발생하는 중복데이터 문제가 발생하지 않도록 정보공개 청구를 통해

〈표 2〉 수집 데이터 자질별 수량

	자질 명	수량(개)
1	제목	40,597
2	기관명	51
3	담당부서명	2,093
4	담당자명	8,831
5	생산일자	350
6	문서번호	49,292
7	단위과제	5,329

온전한 데이터의 수집을 진행하였다. 정보공개청구 결과 수집한 데이터는 총 97,064건이었는데, 1차 수집 데이터와 자질별 수량을 비교하면 〈표 3〉과 같다. 중복데이터 3건을 제외하고 97,061건의 자질별 데이터 수량을 살펴보면 78,814개의 제목명과 51개의 기관명, 2,424개의 담당부서명, 12,526개의 담당자명, 352개의 생산일자, 97,061개의 문서번호, 8,489개의 분류체계를 확인할 수 있었다.

단위과제가 직접 제시되던 1차 데이터 수집 결과와는 달리 2차 데이터 수집의 결과에서는 BRM에 따른 6단계 구조를 반영한 분류체계를 제공하고 있었다. 이 연구에서는 최하위 분류기준인 단위과제가 분류대상이므로 6단계 분류체계 중 단위과제를 선별하여 종속변수로 한정하였다. 그 결과 단위과제에서 5,136건의 결측치가 발생하였다. 이러한 문제는 분류체계가 기입되어 있지 않아 공란이거나, 6단계 구조를 지키지 않고 단위과제만 기입되어 있는 데이터에서 발생하였다. 분류체계를 지키지 않은 단위과제의 신뢰성을 담보할 수 없어 이 데이터들 또한 제외하고 91,925건의 데이터를 대상으로 본 연구를 진행하였다.

4. 연구결과

4.1 파일럿 테스트 모델 구축

자동분류 모델을 구축하기 위해서는 학습에 적합한 데이터를 선별하고 선별한 데이터를 학습에 적합한 형태로 변환한 뒤, 변환된 데이터를 분류 알고리즘에 적용하여야 한다. 먼저, 훈련에 적합한 데이터를 선정하기 위해 종속변수인 단위과제의 최소 데이터 개수를 설정한다. 단위과제의 최소 데이터 개수에 따라 분류할 범주인 단위과제의 개수가 달라지고 그에 따라 연구 대상이 되는 전체 데이터의 개수도 달라질 것이기 때문에 데이터 선별 과정이 가장 먼저 수행된다. 최소 데이터 기준 변경에 따른 성능변화는 연구문제 2에서 살펴볼 것이므로 여기서는 최소 데이터 기준을 100으로 지정하고 단위과제별 기록물 건이 100개 이상인 데이터만 연구에 활용한다. 1차 수집 데이터를 대상으로 선별작업을 수행한 결과 5,320개의 단위과제를 가진 49,292건의 데이터에서 55건의 단위과제를 지닌 24,702건의 데이터를 선별하여 연구에 활용할 수 있었다. 55개 단위과제는 연구

〈표 3〉 수집 데이터 자질별 수량

	자질 명	1차 수집 데이터 수량	2차 수집 데이터 수량
1	제목	40,597	78,814
2	기관명	51	51
3	담당부서명	2,093	2,424
4	담당자명	8,831	12,526
5	생산일자	350	352
6	문서번호	49,292	97,061
7	단위과제	5,329	-
8	분류체계	-	8,489

의 종속변수로 레이블인코딩을 통해 벡터화하였다.

연구의 주요 독립변수 중 하나인 제목은 '인사발령(근속승진) 보고(통보)', '「물관리위원회 지원단 구성 및 운영에 관한 세부규정」(환경부 훈령) 일부개정 보고'와 같이 대부분이 명사구로 이루어져 있었으며 세부내용을 표현하거나 강조하기 위해 '「 」, 「 」, ( )'와 같은 특수문자가 종종 사용되었다. 특수문자는 자연어처리 과정에서 자질로 사용하기에 적절하지 않으므로 정규표현식을 사용하여 제거하였다. 이 과정을 통해 숫자, 알파벳, 한글에 해당하는 문자만 선별하였다. 특수문자가 제거된 제목 자질은 형태소 분석과 문헌표현기법을 통해 벡터화되었다. 전체 성능과 특히 명사, 대명사, 수사를 포함하는 범주인 체언의 정답률이 가장 높은 것으로 밝혀진 komoran 형태소 분석기(김수연 외, 2022)를 활용하여 형태소 분석을 수행하였으며 문헌표현기법은 가장 기본적인 기법인 TF 기법을 사용하였다.

그 외 명목형 데이터인 기관명, 부서명, 담당자명은 원핫인코딩을 수행하고 시계열데이터인 생산일자는 연, 월, 일을 구분한 뒤 최소 단위인 일(日)로 변환하여 정수로 단위를 통일하였다. 또한, 자질 간의 스케일 차이를 줄이기 위해 변환된 값이 0~1 사이의 값을 갖도록 정규화(min-max scaling)하였다. 마지막으로 모델 구성을 위해 훈련 데이터와 테스트 데이터 비율을 7:3으로 설정하고 분류 알고리즘으로 텍스트 데이터 분류에서 우수한 성능을 보이고 있는 RF(김판준, 2019)를 사용하였다.

파일럿 테스트 모델 구축 결과 모델 구축 과정에서는 별도의 문제점이 발견되지 않았다. 테스트 결과 55개 범주를 0.8655의 정확도를 보이며 자동분류하는 다중분류 모델을 구성할 수 있었다. 중복데이터 발생 외에는 이상이 없었으므로 데이터 수집 방법만 변경하여 본 실험을 진행하였다.

4.2 기준선 모델 구축

이차 수집 데이터 91,925건을 대상으로 기준선 모델을 구축하기 위해 파일럿 테스트와 동일하게 먼저 단위과제의 최소 데이터 개수 설정하고 데이터를 선별한다. 최소 데이터 기준을 100으로 지정하고 데이터를 선별한 결과 〈표 4〉와 같은 데이터 자질별 수량을 파악할 수 있었다. 따라서 실제로 연구에 사용하는 데이터는

〈표 4〉 데이터 선별 전·후 자질별 수량 비교

	자질 명	선별 전 데이터 수량	선별 후 데이터 수량
1	제목	74,635	39,010
2	기관명	45	45
3	담당부서명	2,173	1,431
4	담당자명	11,498	5,580
5	생산일자	350	330
6	문서번호	91,925	52,435
7	단위과제	6,830	118

52,435건이었으며 118개 단위과제 범주로 분류하는 다중분류를 수행하였다. 기준선 모델에서는 가장 기본이 되는 자질로 내용적 정보를 담고 있는 제목과 맥락적 정보를 담고 있는 기관명 및 담당부서명만을 사용하여 분류모델을 구성하였다.

파일럿 테스트에서 수행한 프로세스와 같이 제목 자질에 정규표현식, 형태소분석, TF 기법을 적용하고 기관명 및 담당부서에는 원핫인코딩, 단위과제에는 레이블인코딩을 수행하여 데이터를 벡터화하였다. TF 기법은 기본적으로 모든 단어의 출현빈도를 DTM(Document Term Matrix)으로 나타내는데 파이썬 내부파라미터에서는 5회 미만의 출현 빈도를 가진 단어는 유의하지 않다고 보고 자질에서 배제한다. 이 기준의 변경에 따른 성능 비교는 연구문제 3에서 다룰 것이므로 기준선 모델에서는 기본값인 5번을 기준으로 제목 자질을 선정한다.

분류 알고리즘 또한 RF를 사용하여 테스트를 수행한 결과 정확도 0.8308, macro F1 0.8464와 weighted F1 0.8268의 성능을 확인할 수 있었다. 기준선 모델의 성능 도출을 통해 이후 연구문제에서 확인할 다양한 모델들의 성능을 비교할 수 있는 기준을 마련하였다.

4.3 분류 알고리즘별 모델 성능 비교

다양한 분류 알고리즘을 적용하여 그에 따른 성능을 비교한 결과는 〈표 5〉와 같다. SVM, kNN, NB, KT, RF, XGB 등 총 6개 분류 알고리즘을 적용하였으며 성능지표로 정확도, macro F1, weighted F1, 훈련시간, 예측시간 을 측정하였다. 비교 결과 기준선 모델에서 사용되었던 RF가 정확도 0.8308, macro F1 0.8464, weighted F1 0.8268로 가장 높은 성능을 보이는 것을 확인할 수 있었다. 훈련 및 예측시간은 분류 알고리즘의 특성에 따라 큰 차이를 보였는데, kNN, NB, DT등의 기법은 1분 내로 훈련 및 예측이 완료되는 결과를 보이는 반면 SVM과 XGB 는 각각 50분 및 100분 정도까지 걸리는 결과를 확인할 수 있었다. 종합하면 RF가 최고 정확도를 보이고 있으면서도 2분 내에 훈련 및 예측을 완료하고 있어 기록물 자동분류 알고리즘으로는 가장 적절한 것으로 나타났다. 또한, SVM, kNN, DT, RF, XGB 등 대부분의 모델에서 macro F1이 가장 높은 수치를 보이고 그 다음이 정확도, weighted F1은 가장 낮은 수치를 보이는 추이를 확인할 수 있었다. 분류 알고리즘별 모델 성능을 비교한 결과 RF가 가장 좋은 성능을 보였으므로 차후 연구문제에

〈표 5〉 모델별 성능 비교

	정확도	macro F1	weighted F1	훈련시간(초)	예측시간(초)
SVM	0.811	0.8218	0.8037	4047.97	2178.83
kNN	0.7611	0.7649	0.7562	1.54	12.97
NB	0.6872	0.6628	0.6681	1.06	0.51
DT	0.805	0.8202	0.8024	27.49	0.37
RF	**0.8308**	**0.8464**	**0.8268**	108.04	2.55
XGB	0.7734	0.8264	0.7716	3236.49	0.81

서는 RF를 사용하여 성능을 도출하도록 한다.

4.4 데이터 선정 기준별 모델 성능 비교

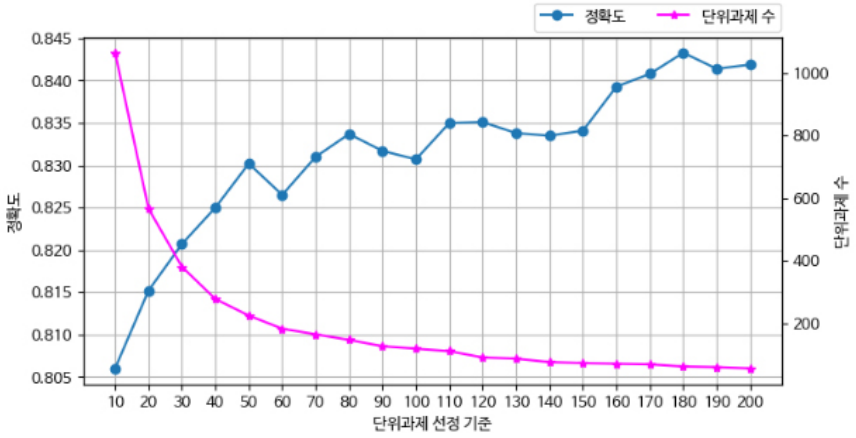
4.4.1 단위과제 선정기준 비교

기준선 모델에서는 단위과제별 최소 데이터 개수를 100개로 임의 지정하고 최소 데이터 개수를 충족하는 단위과제 118개의 데이터 52,435건만 선정하여 실험을 진행하였다. 그러나 단위과제 선정 기준에 따라 전체 데이터와 종속변수인 단위과제의 개수가 달라지고 이는 성능에 직결될 수 있는 만큼 단위과제 선정 기준 변경에 따른 성능 변화를 비교할 필요가 있다. 그런데 단위과제 선정 기준이 낮아질 경우, 훈련 데이터와 테스트 데이터 분리 시에 특정 데이터 셋에 단위과제가 편중되는 상황이 발생할 수 있다. 이는 올바른 훈련이 이루어지지 못하게 할 뿐더러 오류의 원인이 되기도 한다. 이러한 문제를 방지하기 층화추출을 사용하여 동일한 단위과제가 훈련 데이터와 테스트 데이터에 항상 지정된 비율로 존재할 수 있도록 한다.

단위과제 선정 기준을 단위과제당 최소 데이터 10개에서 200개까지 10 단위로 점진적으로 증가시키며 기준의 변화에 따른 전체 데이터 수, 단위과제 수, 성능(정확도)을 비교한 결과는 〈표 6〉과 같다. 또한 기준 변화에 따른 정확도와 단위과제 수의 변화를 그래프로 나타내면 〈그림 3〉과 같다. 단위과제 선정기준이 높아질

〈표 6〉 데이터 선정 기준 별 데이터, 단위과제, 정확도

기준	데이터 수	단위과제 수	정확도
n = 10	76,665	1,065	0.806
n = 20	69,941	566	0.8152
n = 30	65,488	380	0.8207
n = 40	62,043	277	0.825
n = 50	59,653	224	0.8302
n = 60	57,372	182	0.8265
n = 70	56,206	164	0.831
n = 80	54,893	146	0.8337
n = 90	53,184	126	0.8317
n = 100	52,435	118	0.8307
n = 110	51,605	110	0.835
n = 120	49,314	90	0.8351
n = 130	48,818	86	0.8338
n = 140	47,339	75	0.8335
n = 150	46,900	72	0.8341
n = 160	46,595	70	0.8393
n = 170	46,266	68	0.8408
n = 180	45,042	61	0.8433
n = 190	44,675	59	0.8414
n = 200	43,899	55	0.8419



〈그림 3〉 단위과제 선정 기준에 따른 정확도 및 단위과제 수

수록 단위과제 수는 감소하고 정확도는 불규칙적으로 상승하는 추세를 확인할 수 있다. 단위과제 선정기준이 30일 때의 정확도와 단위과제 수가 교차에 근사한 것을 확인할 수 있는데, 이때 정확도와 단위과제 수 간의 상호 상충 정도가 최소화되므로 단위과제 선정 기준은 30으로 삼는 것이 가장 적절한 것으로 나타났다. 또한, 단위과제별 최소 데이터 개수를 최저 10개로 설정하는 경우에 1,065개의 범주에 다중 분류하는 작업임에도 불구하고 0.806의 정확도를 보여 단위과제 별 최소 데이터가 적더라도 그 결과가 현저히 낮은 성능으로 직결되는 것은 아님을 확인할 수 있었다. 데이터 선정 기준별 성능을 측정한 결과 데이터 선정 기준을 30으로 둘 때 가장 적절한 것으로 나타났으므로 다음 연구문제에서는 단위과제의 최소 데이터 기준을 30으로 두고 성능을 도출하도록 한다.

4.4.2 독립변수 자질 성능 비교 및 영향력 분석

기준선 모델에서는 독립변수로 제목, 기관명, 담당부서명 자질만을 사용하였다. 그러나 각 자질이 모델 구성에 미치는 영향력을 확인할 수 없었고 사용되지 않은 자질 또한 존재했다. 따라서 제목 자질만 사용한 모델부터 순차적으로 자질을 추가해가며 성능변화를 관찰하고 각 자질이 모델에 미치는 영향력을 비교하였다. 기준선 모델에서 사용하지 않았던 담당자명과 생산일자는 파일럿 테스트 과정과 동일하게 각각 원핫인코딩과 정수값 변환 및 스케일링을 수행하여 벡터화하였다.

자질 구성에 변화를 주며 모델을 학습시키고 정확도와 각 자질별 영향력을 비교해본 결과는 〈표 7〉과 같다. 제목 자질만 사용한 모델 1의 경우 0.7146의 가장 낮은 정확도를 기록했고 이후 자질을 추가할수록 성능은 꾸준히 향상되어 모든 자질이 사용된 모델 5는 0.8659의 정확도를 보였다. 각 자질이 성능에 미치는 영향은 feature importances를 활용하여 파악할 수 있는데 주제정보를 담고 있는 '제목' 자질이 가장 큰 영향력을 지니고 있었다. 그러나 '기관명', '담당부서명', '담당자명', '생산일자' 등의 맥락정보를 담

〈표 7〉 자질 변경에 따른 정확도 및 영향력 비교

구분	정확도	자질별 영향력				
		제목	기관명	담당부서명	담당자명	생산일자
모델 1	0.7146	1	-	-	-	-
모델 2	0.7609	0.9113	0.0887	-	-	-
모델 3	0.8207	0.7544	0.0686	0.177	-	-
모델 4	0.8618	0.6132	0.0552	0.1382	0.1934	-
모델 5	0.8659	0.5929	0.0538	0.1371	0.1841	0.032

고 있는 자질이 추가될수록 '제목' 자질의 영향력이 줄어들어 모든 자질이 사용된 모델 5에서는 59%까지 영향력이 낮아지고 다른 자질들의 영향력이 높아진 것을 확인할 수 있었다.

자질을 추가할수록 영향력이 분산되고 성능이 향상되는 것은 명확하나 자동분류를 적용하는 맥락에 따라 각 상황에 알맞은 자질을 선정하는 것이 중요하다. 신규 업무담당자의 기록물 생산시 단위과제를 자동부여하는 맥락에서는 '담당자명' 자질을 배제하는 것이 바람직하지만 기록물 이관에 따라 재분류하는 맥락에서는 '담당자명' 자질을 추가하는 것이 더 좋은 성능을 보일 수 있다. 따라서 기록물 자동분류에 있어서도 업무맥락에 따른 적절한 자질 선택 과정이 분류 성능에 많은 영향을 줄 수 있음을 인지하고 이에 따른 신중한 선택이 요구된다.

4.5 문헌표현기법별 모델 성능 비교

기준선 모델에서는 문헌 내에서 유의한 의미를 담은 형태소의 최소 출현 빈도를 5로 두었지만 여기서는 1~10까지 다양하게 변경해보며 성능을 확인해본다. 또한, 가장 기본적인 문헌표현기법인 TF 기법 외에도 TF-IDF 기법, TF-ICF 기법 등을 적용해보고 그에 따른 성능을 비교해보았다. 이를 위해 독립변수 중 문헌표현기법의 영향을 받는 '제목' 자질만 사용하여 모델의 성능을 측정하였다. 형태소 출현 빈도와 문헌표현기법에 따른 성능을 비교해본 결과는 〈표 8〉과 같다.

〈표 8〉 형태소 최소 출현 빈도와 문헌표현기법에 따른 성능 변화

	TF	TF-IDF	TF-ICF
n=1	0.7178	0.7132	0.7128
n=2	0.7178	0.7153	**0.7165**
n=3	**0.7179**	0.7123	0.7124
n=4	0.7176	**0.7139**	0.7152
n=5	0.7168	0.7112	0.711
n=6	0.7146	0.712	0.7119,
n=7	0.7142	0.7112	0.7086
n=8	0.7137	0.7087	0.7109
n=9	0.716	0.7077	0.7086
n=10	0.7121	0.709	0.7097

TF 기법을 사용했을 경우 3번 이상 출현한 형태소를 자질로 삼았을 때 가장 성능이 높은 0.7179의 정확도를 볼 수 있었다. TF-IDF 기법을 사용한 경우에는 4번을 기준으로 삼을 때 0.7139의 정확도, TF-ICF 기법을 사용했을 경우 2번을 기준으로 삼을 때 0.7165의 정확도가 가장 높은 성능을 보였다. 종합하면 TF 기법을 사용하고 출현빈도 3번 이상인 형태소를 자질로 삼을 때 가장 좋은 성능을 보인다는 것을 확인할 수 있었다.

4.6 분석 결과 논의

이 연구의 결과는 기록물 자동분류에 있어 분류 알고리즘과 데이터 선정 기준, 문헌표현기법에 따라 사용가능한 데이터의 개수와 범주의 범위부터 분류모델의 성능까지 변화한다는 것을 보여준다. 연구 결과는 다음과 같은 시사하는 바가 있다.

첫째, 기록물을 대상으로 한 분류 알고리즘 중 RF가 가장 좋은 성능을 보인다는 것이다. 정확도, macro F1, weighted F1 등 모든 분류 성능에서 가장 높은 성능을 보였으며 훈련 및 예측시간 또한 2분 내의 준수한 속도를 보여주었다. 다만, 분류 알고리즘 별 성능비교에서 눈여겨 볼만한 지점은 대부분의 모델에서 macro F1이 가장 높고 정확도가 중간, weighted F1이 가장 낮은 성능을 보였다는 점이다. 이는 모델이 데이터가 더 적은 범주를 상대적으로 더 올바르게 분류하고 데이터가 더 많은 범주는 상대적으로 더 잘못 분류했다는 것을 의미한다. 이는 일반적으로 데이터가 많을수록 더 잘 분류한다고 여겨지는 것(Jo & Japkowicz, 2004)과는 상반된 결과이다.

데이터의 품질에 이상이 있을 경우 이러한 문제가 발생할 수 있는데 이 연구의 데이터 품질에 관해 분석해보면 다음과 같은 두 가지 예상 원인이 있다. 먼저, 데이터가 많은 범주가 그 자체의 포괄적인 특성에 기인하여 상대적으로 오분류된 데이터를 많이 지니고 있는 경우이다. 오분류된 데이터가 훈련부에 들어갈 경우 모델이 잘못된 패턴을 학습하기 때문에 간접적인 영향을 미치고 테스트부에 들어갈 경우 오분류된 결과가 그대로 반영되기 때문에 직접적인 영향을 미친다. 다음으로, 데이터가 적은 범주에 가중될만한 자질(형태소)이 명확한 경우이다. 데이터를 구체적으로 확인해본 결과 이러한 경우의 사례를 발견할 수 있었다. 예를 들어 단위과제 중 '교도관 회의'라는 단위과제에 108건 중 106건이 '교도관'과 '회의' 또는 '회의록'이라는 형태소를 함께 가지고 있었는데 특히 '교도관'이라는 단어가 일상적으로 잘 쓰이는 단어가 아니라는 점을 감안하면 해당 단어에 매우 높은 가중치가 부여되고 그에 따라 '교도관 회의' 단위과제로 분류될 수 있다. 이와 같은 경우에 범주당 데이터의 양은 적지만 오히려 높은 분류 성능을 보이는 현상이 설명될 수 있다. 이러한 문제를 극복하려면 데이터의 품질이 보장된 양질의 학습 데이터가 필요하다. 결국 기록물의 효과적인 자동분류를 위해서는 양질의 학습 데이터 구축이 요구된다.

둘째, 학습을 위한 종속변수의 범주(단위과제)의 최소 데이터를 낮게 설정하더라도 안정적인 성능을 보이는 것이 확인되었다. 단위과제 선정 기준을 높게 잡으면 범주의 개수가 감소하고 범주의 개수가 감소하면 오분류 할 수

있는 경우의 수가 감소하기 때문에 정확도가 높아지는 것이 일반적이다. 그런데 최소 데이터 기준의 증가에 따라 범주의 감소와 정확도가 높아지는 추세는 관찰되었지만 그 결과가 극적인 정확도의 향상으로 이어지지 않는다는 점을 확인할 수 있었다. 오히려 앞서 살펴본 바와 같이 특정 범주에 대하여 자질의 가중치가 명확하여 전체 정확도에 양의 영향을 미치는 범주와 데이터가 제거될 경우에는 오히려 정확도가 감소하는 사례도 관찰할 수 있었다. 또한 최소 데이터를 10개까지 설정하더라도 현저한 성능 저하는 없었는데 오히려 이 경우 최소 데이터를 100개로 설정한 기준선 모델에 비해 정확도는 3% 하락하였지만 1.5배 많은 데이터와 9배 많은 범주를 다루었다는 점을 감안하면 상황에 따라 더 낮은 최소 데이터 기준의 가능성을 시사한다. 한편 학습을 위한 독립변수의 자질 변경에 따라서는 뚜렷한 성능의 변화를 확인할 수 있었다. '제목' 자질만 사용한 모델의 오류율은 29%로 '제목', '기관명', '담당부서명', '담당자명', '생산일자' 등 모든 자질을 사용한 모델의 오류율 14%의 두 배에 달했다. 자질이 추가될수록 명확하게 성능이 향상되는 추이를 보여 성능향상을 위해서는 메타데이터 추가 수집을 통한 자질 추가도 고려해 볼 수 있을 것이다.

셋째, 형태소 출현빈도와 문헌표현기법에 따른 성능변화가 크지 않다는 것을 확인하였다. 형태소 출현빈도 10가지와 문헌표현기법 3가지, 총 30가지 모델의 성능을 비교한 결과 최고성능모델과 최저성능모델의 성능 차이는 1%에 지나지 않았다. 이는 사용된 문헌표현기법이 모두 통계적 기법이라는 한계에서 비롯된 것으로 추후 연구에서는 구조적 기법과 의미적 기법의 적용 및 비교를 통해 문헌표현기법에 따른 성능 변화를 엄밀히 따져볼 필요가 있다.

5. 결 론

이 연구에서는 문헌 자동분류에 사용된 기법을 기록물의 특성을 반영한 기록물 자동분류에 적용하고 기록물 자동분류를 위한 최적 성능요소를 탐색하고자 하였다. 이를 위해 선행연구를 분석하여 문헌 분류와 기록물 분류의 차이를 밝히고 기록물분류제도에 따라 분류대상으로 적합한 대상을 파악하였다. 또한, 2022년 중앙행정기관 공개원문정보를 대상으로 하여 분류 알고리즘, 데이터 선정 방법, 문헌표현기법 등의 실질적인 성능 요소 비교를 수행하였다.

먼저 최적의 분류 알고리즘 파악을 위해 SVM, kNN, NB, DT, RF, XGB 등의 다양한 알고리즘을 적용하였으며 정확도, macro F1, weighted F1, 훈련시간, 예측시간 등 다양한 성능을 종합적으로 비교하였다. 그 결과 Random Forest가 0.8308의 정확도, 0.8464의 macro F1, 0.8268의 weighted F1와 준수한 훈련시간 및 예측시간을 보여 최적의 분류 알고리즘으로 파악되었다.

다음으로 최적의 데이터 선정기준을 파악하기 위해 독립변수인 자질과 종속변수인 단위과제의 범주 기준을 다양하게 변경하며 성능을 측정하였다. 종속변수인 단위과제 선정기준을 최소 10개에서 200개까지 변경하며 성능을 측정한 결과 최소 0.806에서 최대 0.8419의 정확도를 보여 단위과제 선정기준의 변경에 따른 성능변화는 크지 않음을 확인할 수 있었다. 오히려 선정기준을 낮게 잡을 경우 많은 데이터

와 범주를 대상으로 삼을 수 있는 것이 입증되며 목적에 알맞은 단위과제 선정기준을 지정하는 것이 중요한 것임을 밝혔다. 한편 독립변수인 자질을 다양하게 변경하며 성능을 측정한 결과 최소 0.7146에서 최대 0.8659의 정확도를 보여 자질 추가 및 변경에 따른 성능변화가 명확함을 보였다. 특히 모든 자질을 사용할 경우 자질별 영향력이 '제목' 0.5929, '기관명' 0.0538, '담당부서명' 0.1371, '담당자명', 0.1841, '생산일자' 0.032와 같이 밝혀져 제목의 영향력이 가장 크고 이후 담당자명, 담당부서명, 기관명, 생산일자 순으로 분류에 영향을 주는 것이 확인되었다.

마지막으로 최적의 문헌표현기법을 파악하기 위해 TF-IDF 기법, TF-ICF 기법 등을 적용하고 자질로 삼을 수 있는 형태소의 최소 출현빈도를 1~10까지 다양하게 적용하여 성능을 비교해 보았다. 비교결과 TF 기법이 형태소의 최소 출현빈도가 3일 때 0.7179의 정확도로 가장 높은 성능을 보였으며 TF-IDF 기법은 형태소의 최소 출현빈도가 4일 때 0.7139로 가장 높은 성능을 보였고 TF-ICF 기법은 형태소의 최소출현빈도가 2일 때 0.7165로 가장 높은 성능을 보였다. 세 가지 기법들은 모두 가장 성능이 높은 최소출현빈도를 기준으로 기준에서 멀어질수록 성능이 저하되는 추이를 보였다. 또한 그 중 가장 높은 성능을 보인 것은 TF 기법이었으나 다른 기법과 현격한 정확도의 차이는 없었다. 이는 세 기법이 모두 통계적 기법을 기반으로 하고 있는 문제일 수 있어 추후 구조적 기법과 의미적 기법을 적용한 문헌표현기법의 적용에 따른 성능변화를 관찰할 필요성이 제기된다.

인공지능기술은 급속도로 발전하고 있지만 기록관리 분야는 자동분류 등의 개념접목이 늦어지고 있는 현실에서 이 연구는 기록물 자동분류를 위한 현실적인 자료가 될 것이다. 기록물 자동분류에 있어 분류 알고리즘으로는 RF를 사용하고 문헌표현기법으로는 TF 기법을 사용하는 것이 가장 좋은 성능을 낸다는 점을 밝혔다. 또한 범주별 최소 데이터 선정 기준이 낮더라도 분류성능에는 큰 영향을 주지 않지만 독립변수가 되는 자질은 명확한 성능변화를 보이게 한다는 점을 밝혀 추후 기록물 자동분류에 있어서는 그 목적과 환경에 따른 데이터선정기준을 다르게 하고 그 판단의 근거가 될 수 있는 수치들을 제시하였다.

본 연구에서는 word2vec나 BERT등의 최신 딥러닝 기술들을 접목하여 연구를 진행하지는 못했으므로 이를 활용한 추가 연구는 향후 과제가 될 것이다. 분류 문제는 여전히 많은 분야에서 요구되고 있는 만큼 이 연구가 기록관리계에서도 다양한 방법론이 시도되고 고도화되는데 기여하여 기록물 관리사들의 업무가 지능화 되는 지능형 아카이브 시스템이 갖춰지기를 기대한다.

참 고 문 헌

공공기관의 기록물관리에 관한 법률 시행령. 대통령령 제16609호.
공공기관의 기록물관리에 관한 법률. 법률 제5709호.

공공기록물 관리에 관한 법률 시행령. 대통령령 제19985호.

국가기록원 (2012). 대학 기록물 보존기간 책정기준 가이드(11-1311153-000250-01).

국가기록원 (2017). 차세대 기록관리 모델 재설계 연구(11-1741050-000008-01).

국가기록원 (2021). 기록관리 AI 기술적용을 위한 공통 학습데이터 세트 구축 연구 (11-1741050-000073-01).

김수연, 안석호, 김동현, 이의종, 서영덕 (2022). 형태소 분석기의 품사별 정확성 분석. 한국정보기술학회 종합학술발표논문집, 378-381.

김판준 (2016). 기계학습에 기초한 자동분류 성능 요소에 관한 연구. 정보관리학회지, 33(2), 33-59. https://doi.org/10.3743/KOSIM.2016.33.2.033

김판준 (2019). 랜덤포레스트를 이용한 국내 학술지 논문의 자동분류에 관한 연구. 정보관리학회지, 36(2), 57-77. https://doi.org/10.3743/KOSIM.2019.36.2.057

김해찬솔, 안대진, 임진희, 이해영 (2017). 기계학습을 이용한 기록 텍스트 자동분류 사례 연구. 정보관리학회지, 34(4), 321-344. https://doi.org/10.3743/KOSIM.2017.34.4.321

남은경, 안혜림, 송민 (2013). 공공사이트 게시판 자료의 기록관리를 위한 자동 분류 시스템. 정보관리학회 학술대회 논문집, 175-178.

방재현 (2018). 지능형 아카이브 시스템을 위한 기계학습 기술 적용 방안 연구. 박사학위논문, 한국외국어대학교 대학원 정보・기록관리학과.

설문원 (2012). 기록분류에 관한 국내 연구 동향과 과제. 한국기록관리학회지, 12(3), 203-232.

설문원 (2013). 단위과제 기반 공공기록물 평가제도의 문제점과 개선방안. 한국기록관리학회지, 13(3), 231-254. https://doi.org/10.14404/JKSARM.2013.13.3.231

설문원 (2013). 단위과제 기반 공공기록물 평가제도의 문제점과 개선방안. 한국기록관리학회지, 13(3), 231-254.

육지희, 송민 (2018). 토픽모델링과 딥 러닝을 활용한 생의학 문헌 자동 분류 기법 연구, 정보관리학회지, 35(2), 63-88. https://doi.org/10.3743/KOSIM.2018.35.2.063

이원영 (2000). 기록물분류의 원리: 문헌분류와의 비교. 기록학연구, 2, 103-128. https://doi.org/10.20923/kjas.2000.2.103

이현실, 한성국 (2006). 기록 관리 메타데이터의 개념 모델링. 정보관리학회지, 23(3), 23-48. https://doi.org/10.3743/KOSIM.2006.23.3.023

장지숙, 이해영 (2009). 맥락정보를 이용한 기록 자동분류시스템 설계. 한국기록관리학회지, 9(1), 151-173. https://doi.org/10.14404/JKSARM.2009.9.1.151

정부기능의 분류 및 관리에 관한 규정. 대통령훈령 제209호.

정지혜, 이젬마, 왕호성, 오효정 (2022). 이관기록물 분류 자동화를 위한 이상치 판별 학습데이터 구축. 한국기록관리학회지, 22(1), 43-59. https://doi.org/10.14404/JKSARM.2022.22.1.043

최윤수, 최성필 (2019). 기술용어 분산표현을 활용한 특허문헌 분류에 관한 연구. 한국문헌정보학회지, 53(2), 179-199. https://doi.org/10.4275/KSLIS.2019.53.2.179

Gates, W. H (2023. March 21). The Age of AI has begun. Available: https://www.gatesnotes.com/The-Age-of-AI-Has-Begun

Jo, T. & Japkowicz, N. (2004). Class imbalances versus small disjuncts. Association for Computing Machinery Special Interest Group on Knowledge Discovery in Data Exploration, 6, 40-49. https://doi.org/10.1145/1007730.1007737

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

Bang, Jae Hyun (2018). A Study on Application of Machine Learning for Intelligent Archive System: Focusing on Application of Deep Neural Network. Doctoral dissertation, Hankuk University of Foreign Studies.

Choi, Yunsoo & Choi, Sung-Pil (2019), A study on patent literature classification using distributed representation of technical terms. Journal of the Korean Society for Library and Information Science, 53(2), 179-199. https://doi.org/10.4275/KSLIS.2019.53.2.179

Enforcement Decree of Records Management of Public Institution Act. No.16609.

Enforcement Decree of the Public Records Management Act. No.19985.

Jang, Ji-Sook & Rieh, Hae-Young (2009). Design of automatic records classification system using contextual information. Journal of Korean Society of Archives and Records Management, 9(1), 151-173. https://doi.org/10.14404/JKSARM.2009.9.1.151

Jeong, Jihye, Lee, Gemma, Wang, Hosung, & Oh, Hyo-Jung (2022). Building the outlier candidate discrimination training data based on inventory for automatic classification of transferred records. Journal of Korean Society of Archives and Records Management, 22(1), 43-59. https://doi.org/10.14404/JKSARM.2022.22.1.043

Kim, Hae Chan Sol, Ahn, Dae Jin, Lim, Jin Hee, & Lee, Haeyoung (2017). A study on automatic classification of record text using machine learning. Journal of the Korean Society for Information Management, 34(4), 321-344. https://doi.org/10.3743/KOSIM.2017.34.4.321.

Kim, Pan Jun (2016). An analytical study on performance factors of automatic classification based on machine learning. Journal of the Korean Society for Information Management, 33(2), 33-59. https://doi.org/10.3743/KOSIM.2016.33.2.033

Kim, Pan Jun (2019). An analytical study on automatic classification of domestic journal articles

using random forest. Journal of the Korean Society for Information Management, 36(2), 57-77. https://doi.org/10.3743/KOSIM.2019.36.2.057

Kim, Suyeon, Ahn, Seokho, Kim, Donghyun, Lee, Euijong, & Seo, Young-Duk (2022). Accuracy analysis by part-of-speech of morpheme analyzers. Proceedings of Korean Istitute of Information Technology Conference, 378-381.

Lee, Hyunsil & Han, SungKook (2006). Conceptual modeling of record management metadata. Journal of the Korean Society for Information Management, 23(3), 23-48. https://doi.org/10.3743/KOSIM.2006.23.3.023

Lee, Wonyoung (2000). The principles of records classification: compares with librarian materials. The Korean Journal of Archival Studies, 2, 103-128. https://doi.org/10.20923/kjas.2000.2.103

Nam, Eunkyung, Ahn, Hye-Rim, & Song, Min (2013). Automatic classification system for record management of bulletin board on public website. Proceedings of the 20th Academic Conference of the Korean Society for Information Management, 175-178.

National Archives of Korea (2012). Guide for Determining the Retention Period of University Records(11-1311153-000250-01).

National Archives of Korea (2017). A Study on Designing a Next-Generation Records Management Model(11-1741050-000008-01).

National Archives of Korea (2021). Study on Common Training Dataset Construction for Appling AI Technology for Records Managements(11-1741050-000073-01).

Records Management of Public Institution Act. No. 5709.

Regulations on the Classification and Management of Government Functions. Presidential Decree No. 209.

Seol, Mun Won (2012). Research trends and issues of records and archives classification in Korea. Journal of the Korean Society for Archives and Records Management, 12(3), 203-232.

Seol, Mun Won (2013). A study on problems of the public records appraisal system based on the value of 'business transaction' and application of a multi-appraisal model. Journal of the Korean Society for Archives and Records Management, 13(3), 231-254. https://doi.org/10.14404/JKSARM.2013.13.3.231.

Yuk, JeeHee & Song, Min (2018). A study of research on methods of automated biomedical document classification using topic modeling and deep learning. Journal of the Korean Society for Information Management, 35(2), 63-88. https://doi.org/10.3743/KOSIM.2018.35.2.063