# 자질선정을 통한 국내 학술지 논문의 자동분류에 관한 연구

## An Experimental Study on the Automatic Classification of Korean Journal Articles through Feature Selection

김판준 (Pan Jun Kim)*


### 초 록

국내 학술연구의 동향을 구체적으로 파악하여 연구개발 활동의 체계적인 지원 및 평가는 물론 현재와 미래의 연구 방향을 설정할 수 있는 기초 데이터로서, 개별 학술지 논문에 표준화된 주제 범주(통제키워드)를 부여할 수 있는 효율적인 방안을 모색하였다. 이를 위해 한국연구재단「학술연구분야분류표」상의 분류 범주를 국내 학술지 논문에 자동 할당하는 과정에서, 자질선정 기법을 중심으로 자동분류의 성능에 영향을 미치는 주요 요소들에 대한 다각적인 실험을 수행하였다. 그 결과, 실제 환경의 불균형 데이터세트(imbalanced dataset)인 국내 학술지 논문의 자동분류에서는 보다 단순한 분류기와 자질선정 기법, 그리고 비교적 소규모의 학습집합을 사용하여 상당히 좋은 수준의 성능을 기대할 수 있는 것으로 나타났다.

### ABSTRACT

As basic data that can systematically support and evaluate R&D activities as well as set current and future research directions by grasping specific trends in domestic academic research, I sought efficient ways to assign standardized subject categories (control keywords) to individual journal papers. To this end, I conducted various experiments on major factors affecting the performance of automatic classification, focusing on feature selection techniques, for the purpose of automatically allocating the classification categories on the National Research Foundation of Korea's Academic Research Classification Scheme to domestic journal papers. As a result, the automatic classification of domestic journal papers, which are imbalanced datasets of the real environment, showed that a fairly good level of performance can be expected using more simple classifiers, feature selection techniques, and relatively small training sets.


키워드: 자동분류, 텍스트 범주화, 자질선정, 필터, 학술지 논문
automatic classification, text categorization, feature selection, filter, journal articles

* 신라대학교 문헌정보학과 부교수(pjkim@silla.ac.kr)

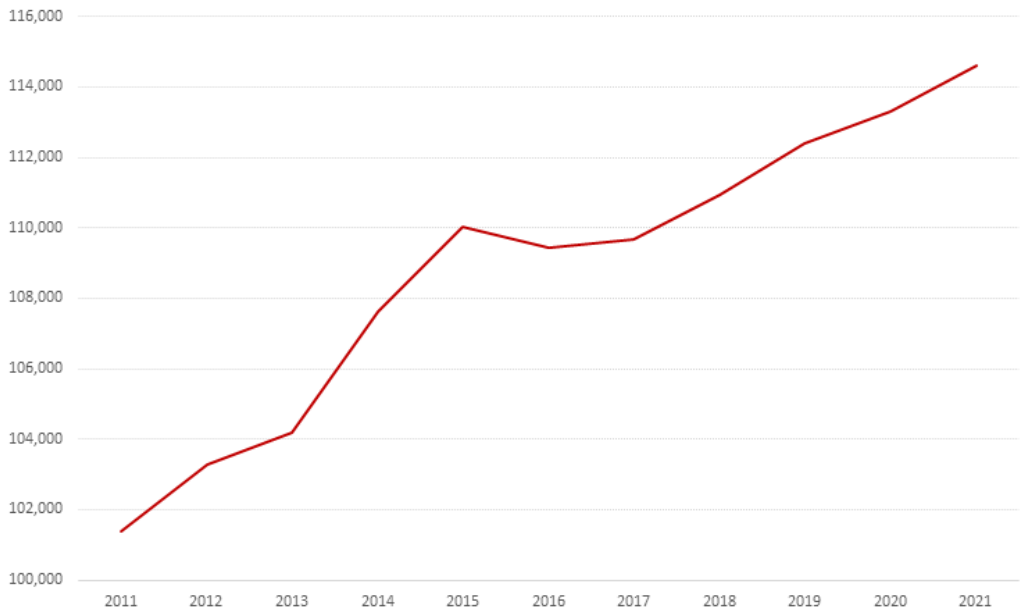
■ 논문접수일자 : 2022년 2월 14일 ■ 최초심사일자 : 2022년 2월 24일 ■ 게재확정일자 : 2022년 3월 4일
■ 정보관리학회지, 39(1), 69-90, 2022. http://dx.doi.org/10.3743/KOSIM.2022.39.1.069

# 1. 서 론

## 1.1 연구의 필요성 및 목적

2000년대 이후 인터넷의 확산과 함께 학술정보의 생산 및 유통이 꾸준히 증가하고 있으며, 학문 분야별로 연구동향을 효과적으로 파악할 수 있는 구조화된 데이터의 필요성이 커지고 있다. 또한 디지털 문헌의 수가 크게 증가함에 따라 이를 효율적으로 정확하게 자동분류하는 기능이 점점 더 중요하면서도 어려워지고 있는 실정이다(Mengle & Goharian, 2009; Uysal, 2016). 한국학술지인용색인(KCI)(2022) 통계에 따르면 국내에서 2022년 2월 현재 5,900여 종의 학술지에 수록된 약 188만 건의 논문이 서비스되고 있으며, 최근 10여 년간 출판 논문 수는 지속적으로 증가하고 있다(〈그림 1〉 참조).

국내외 학술데이터베이스에 축적된 메타데이터는 과거와 현재의 연구를 다양한 측면에서 분석할 수 있는 구조화된 데이터이다. 특정 분야의 연구 동향을 다각적으로 파악할 수 있는 방법으로 많이 사용되고 있는 지적구조 분석은 이러한 메타데이터 중에서 개별 자료가 다루고 있는 내용을 표준화된 형식으로 표현할 수 있는 통제키워드(디스크립터, 주제어 등)를 필요로 한다(김판준, 2021a, 2021b). 그러나 국외 학술데이터베이스(WoS, SCOPUS, LISTA 등)와 달리, 대부분의 국내 학술데이터베이스는 개별 학술지 논문의 주제를 파악할 수 있는 표준화된 범주 정보(통제키워드)를 거의 제공하지 않고 있다(김판준, 이재윤, 2014). 이처럼 꾸준히 증가하고 있는 현재와 미래의 논문은 물론 과거에 출판된 모든 논문에 대하여 표준화된 주제 정보를 제공할 수 있는 수작업 분류(전문가




〈그림 1〉 국내 학술지 논문 수: 2011년~2021년

에 의한 통제키워드 부여)를 단기간에 제한된 예산으로 추진하는 것은 현실적으로 거의 불가능하다. 따라서 수작업 분류에 소요되는 막대한 시간과 전문 인력의 부족은 물론 예산상의 문제를 극복할 수 있는 효과적인 대안으로 기계학습에 기초한 국내 학술지 논문의 자동분류를 적극적으로 모색할 필요가 있다. 특히, 수작업 분류의 장점을 최대한 유지하면서 전문가의 분류 작업에 소요되는 시간과 노력을 최소화하는 방식으로 지원할 수 있도록 기계학습 기반의 자동분류를 적극적으로 추진할 필요가 있다.

본 연구의 목적은 국내 학술연구의 동향을 구체적으로 파악하여 연구개발 활동의 체계적인 지원 및 평가는 물론 현재와 미래의 연구 방향을 설정할 수 있는 기초 데이터로서, 개별 학술지 논문에 표준화된 주제 범주(통제키워드)를 부여할 수 있는 실제적인 방안을 모색하는 것이다. 이를 위해 기계학습에 기초한 자동분류 과정에 자질선정 기법을 적용하여 한국연구재단(2016)「학술연구분야 분류표」상의 분류 범주(소분류명/세분류명)를 국내 학술지 논문에 자동 할당하는 방안을 검토하였다. 특히, 자질선정 기법을 중심으로 기계학습에 기초한 자동분류의 성능에 영향을 미치는 주요 요소들에 대한 다각적인 실험을 수행한 결과를 종합적으로 검토하여, 국내 학술지 논문의 자동분류를 실제적으로 추진하기 위한 효율적인 방안을 제시하였다.

### 1.2 문헌의 자동분류와 자질선정

1960년대에 시작된 문헌의 자동분류(또는 텍스트 범주화)에 관한 연구는 1980년대에 전문가시스템 기반을 중심으로 이루어졌고(Harish & Revanasiddappa, 2017; Salton & Buckley, 1988), 1990년대에 와서 기계학습의 도입으로 분류기의 성능이 크게 향상되었다(Sebastiani, 2002). 이에 따라 최근 학술 텍스트(학술지, 학위논문 등) 분류와 추천에 대한 관심이 커지고 있다(Kragelj & Kljajić, 2021). 대부분의 텍스트 범주화 연구는 문헌을 BoW(Bag of Word) 형식으로 표현하는데(Joachims, 1997; Manning, Raghavan, & Schutze, 2008), 특정 문헌에 출현한 단어 또는 용어가 이러한 문헌벡터의 구성요소가 된다. 비교적 소규모의 문헌집합에서도 흔히 수만 개의 단어가 출현하기 때문에 문헌벡터의 고차원성과 희소성(high dimensionality and sparseness)은 컴퓨터 저장 공간과 실행시간의 증가와 함께 분류 성능 저하의 근본적인 원인이 되고 있다(Joachims, 2002; Rehman et al., 2018; Wang et al., 2014; Wang et al., 2016; Wu & Zhang, 2004). 따라서 자질공간의 차원을 줄이고 분류기의 성능을 향상시킬 수 있는 자질선정 방법이 텍스트 범주화에서 널리 사용되고 있다(Chang et al., 2015).

모든 단어를 분류 자질로 이용하는 것보다 문헌의 내용을 대표할 수 있는 단어들로 축소된 자질집합을 사용할 때 분류 성능이 향상된다는 것은 선행연구를 통하여 잘 알려진 사실이다(김판준, 2016; Gutkin, Shamir, & Dror, 2009; Yang & Pedersen, 1997). 자질선정의 주요 목적은 예측 정확도를 저하시키지 않으면서 차원을 줄이고 관련 기본 자질을 식별하여 데이터세트를 단순화하는 것이며(Kashef, Nezamabadi-pour, & Nikpour, 2018; Pintas, Fernandes, & Garcia, 2021), 분류기와의 상호작용에 따라 filter, wrapper, embedded 방법

으로 구분할 수 있다(Chandrashekar & Sahin, 2014; Drotár, Gazda, & Vokorokos, 2019; Guyon & Elisseeff, 2003; Kumar & Minz, 2014; Venkatesh & Anuradha, 2019). 이 중에서 filter는 분류기와의 상호작용이 필요하지 않기 때문에 가장 대중적이고 빠른 접근법이면서도 좋은 성능으로 wrapper와 embedded 방법보다 선호되고 있다(Chang et al., 2015; Drotár, Gazda, & Smékal, 2015; Fragoudis, Meretakis, & Likothanassis, 2005; Günal, 2012; Pinheiro, Cavalcanti, & Ren, 2015; Talavera, 2005; Uysal, 2016). 또한 filter는 대규모 데이터집합에서 빠른 처리시간과 과적합(overfitting) 문제에 강한 장점으로 인해(Abiodun et al., 2021), 최근 문헌의 자동분류를 위한 자질선정 방법으로 가장 많이 사용되고 있다(Pintas, Fernandes, & Garcia, 2021).

문헌의 자동분류(텍스트 범주화)를 위한 자질선정은 최근 상당한 이슈가 되고 있어 여러 학자들에 의해 다수의 리뷰 논문이 발표되었다. Chandrashekar와 Sahin(2014), Kumar와 Minz(2014)는 자질선정 방법에 대한 일반적인 개요를 제공하였고, Deng et al.(2019)은 텍스트 분류를 위한 자질선정에 초점을 맞추어 문헌표현, 유사성 척도, 분류기 등의 성능 요소를 검토한 후에 자질선정 방법을 네 가지 유형(filter, wrapper, embedded, hybrid)으로 구분하여 각각의 장단점을 제시하였다. 텍스트 범주화를 위한 자질선정 방법에 대한 실증적 연구로는 Harish와 Revanasiddappa(2017)가 많이 사용되고 있는 5개 자질선정 기법과 5개 분류기를 적용한 실험 결과를 보고하였다. 또한, Iqbal et al.(2020)은 텍스트 범주화를 위한 자질선정에 대한 리뷰에서 텍스트 범주화에 대한 개요와 함께 널리 사용되고 있는 자질선정 방법들을 검토한 결과를 제시하였다. 이외에도 복수-범주명 분류를 위한 자질선정에 대한 리뷰에서는 복수-범주명 분류 및 기계학습 알고리즘에 대한 포괄적인 검토와 함께 복수-범주명 자질선정 방법의 분류를 위한 다양한 관점을 제시하였다(Kashef, Nezamabadi-pour, & Nikpour, 2018). 그리고 Pereira et al.(2018)은 복수-범주명(multi-label) 분류에 중점을 둔 자질선정에 대한 포괄적인 리뷰와 함께 여러 자질선정 방법에 대한 새로운 범주화를 제시하였다.

가장 최근에는 Abiodun et al.(2021)이 200건 이상의 논문에 대한 체계적인 문헌 리뷰(SLR: systematic literature review)를 통해 텍스트 범주화를 위한 자질선정의 최적화 방법을 검토하였고, Pintas, Fernandes, Garcia(2021)는 최근 8년간 출판된 175개 논문에 대한 체계적인 문헌 리뷰의 결과로 자질선정 방법을 범주화하는 새로운 스키마를 제시하고, 이에 기초하여 자동분류 실험의 주요 요소들에 대한 매핑을 제공하였다.

지금까지 국내에서 학술지 논문의 자동분류에 자질선정을 적용한 연구는 많지 않으며(김판준, 2006; 김판준, 이재윤, 2012; 육지희, 송민, 2018; 정은경, 2009), 실제 사용되고 있는 분류체계의 범주명(분류명)을 자동 할당하는 방안을 모색한 연구는 더욱 찾아보기 힘들다(김선우 외, 2018; 김판준, 2018; 2019). 따라서 본 연구는 특정 학문분야의 학술지 논문으로 구성된 문헌집단을 대상으로 「학술연구분야 분류표」의 소분류 및 세분류명(한국연구재단, 2016)을 자동 할당하는 목적으로 자질선정 방법을 적용한

결과를 다각적으로 검토하였다.

### 1.3 문헌의 자동분류를 위한 자질선정 기법

본 연구에서 적용한 자질선정은 문헌의 자동분류(텍스트 범주화)에서 가장 많이 사용되고 있는 filter 방법에 해당하는 자질 순위화 기법이라 할 수 있다. 여기서 사용된 8개 기법은 문헌빈도(df: document frequency), 자카드 계수(jac: jaccard coefficient), 카이제곱 통계량(chi: chi-square), 상대적 상호정보량 J(rmij: relative mutual information J), GSS 계수(gss: gss coefficient), 피어슨 계수(pcc: pearson coefficient), 상호정보량(mi: mutual information), 로그승산비(lor: log odds ratio)이다. 대부분의 자질 순위화 기법들은 문헌빈도에 기반하고 있으며(Forman, 2003), 텍스트 분류의 경우 용어에 대하여 네 가지 문헌빈도를 정의할 수 있다(Rehman et al., 2018). 따라서 이러한 기법들을 〈표 1〉의 2*2 분할표에 기초하여 〈표 2〉의 자질 순위화 공식으로 표현하였다.

〈표 1〉 문헌빈도에 기초한 2*2 분할표

| | 자질($t_i$) 출현 | 자질($t_i$) 미출현 |
|---|---|---|
| 긍정 범주($c_j$) | tp(true positive) → a | fn(false negative) → c |
| 부정 범주(not $c_j$) | fp(false positive) → b | tn(true negative) → d |

- tp(true positive) → a: 범주 $c_j$에 속하고 자질 $t_i$가 출현한 문헌 수
- fp(false positive) → b: 범주 $c_j$에 속하지 않고 자질 $t_i$가 출현한 문헌 수
- fn(false negative) → c: 범주 $c_j$에 속하고 자질 $t_i$가 출현하지 않은 문헌 수
- tn(true negative) → d: 범주 $c_j$에 속하지 않고 자질 $t_i$가 출현하지 않은 문헌 수
- N(전체 문헌 수) = a+b+c+d

〈표 2〉 자질선정 기법과 자질 순위화 공식

| 번호 | 자질선정 기법 | 자질 순위화 공식 | 출처 |
|---|---|---|---|
| 1 | df | $a+b$ | (Forman, 2003) |
| 2 | jac | $\frac{a}{(a+b+c)}$ | (이재윤, 2005) |
| 3 | chi | $\frac{N(ad-cb)^2}{(a+c)(b+d)(a+b)(c+d)}$ | (Chang et al., 2015) |
| 4 | rmij | $\log\frac{(a+b+c)}{a}$ | (이재윤, 2005) |
| 5 | gss | $\frac{(ad-bc)}{N^2}$ | (이재윤, 2005) |
| 6 | pcc | $\frac{(ad-bc)}{((a+b)(a+c)(b+d)(c+d))^2}$ | (Cai et al., 2018) |
| 7 | mi | $\log\frac{aN}{(a+c)(a+b)}$ | (Chang et al., 2015) |
| 8 | lor | $\log\frac{ad}{bc}$ | (이재윤, 2005) |

## 2. 국내 학술지 논문의 자동분류 실험

### 2.1 연구문제

문헌의 자동분류(또는 텍스트 범주화)는 분류 대상 문헌에 하나의 범주명 또는 복수의 범주명을 할당하는 가에 따라 단일-범주명 분류(single-label classification)와 복수-범주명 분류(multi-label classification)로 구분할 수 있다.[1] 학술지 논문은 특정 논문이 단일 주제에 관련된 경우도 있지만, 여러 주제에 걸친 내용을 함께 다루는 경우가 많기 때문에 학술지 논문을 대상으로 하는 텍스트 범주화는 범주명 부여 방법(단일-범주명 분류와 복수-범주명 분류)에 따른 성능을 고려하여야 한다(Kashef, Nezamabadi-pour, & Nikpour, 2018; Mironczuk & Protasiewicz, 2018; Pintas, Fernandes, & Garcia, 2021). 따라서 국내 학술지 논문의 자동분류에서 자질선정 기법을 중심으로 다음과 같이 연구 문제를 설정하였다.

- 연구문제1. 국내 학술지 논문의 자동분류에서 자질선정 기법과 주요 성능 요소(범주 부여 방법, 분류기)에 따른 분류 성능에 차이가 있는가?
- 연구문제2. 국내 학술지 논문의 자동분류에서 자질선정 기법과 주요 성능 요소(범주 부여 방법, 분류기, 학습집합의 크기(연차별))에 따른 분류 성능에 차이가 있는가?

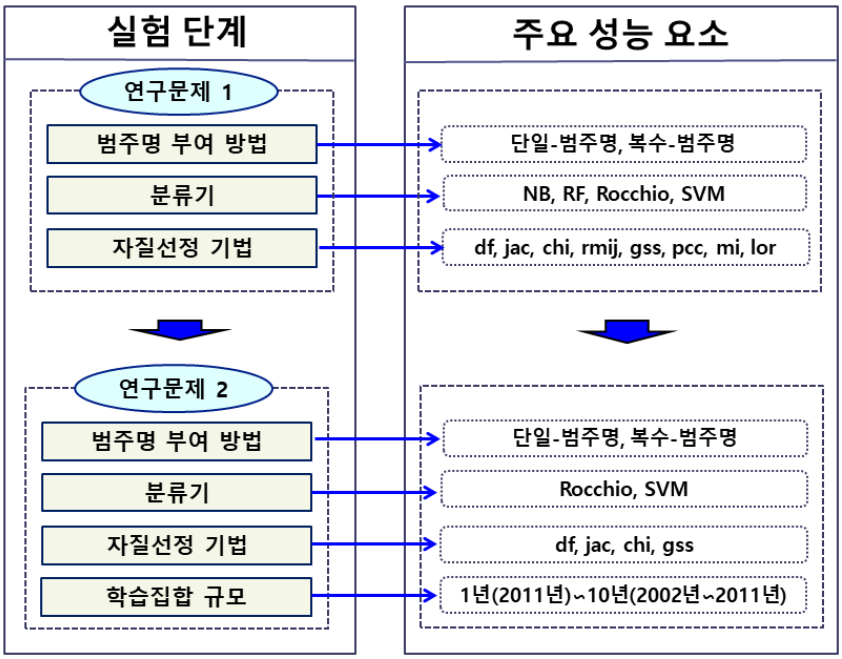
### 2.2 실험 설계

본 연구의 실험 문헌집단은 문헌정보학 분야의 『정보관리학회지』에 수록된 14년(2002년~2015년) 동안의 학술논문 651편이다. 이 중에서 이전 10년(2002년~2011년)의 453편(70%)은 학습집합, 이후 4년(2012년~2015년)의 198편(30%)을 검증집합으로 사용하였다. 자동분류 실험을 위한 범주명은 「정보관리학회지(2002년~2015년)」에 수록된 논문을 대상으로, 한국연구재단(2016) 「학술연구분야 분류표」의 분류명(소분류명, 세분류명)을 전문가 3인(문헌정보학 교수)이 직접 수작업으로 부여하였다. 개별 논문에 대하여 대표 주제를 단일 범주명(single-label)으로 부여하고, 복수의 주제가 포함된 논문에 대해서는 해당 논문의 내용에서 다루어진 비중에 따라 최대 3개까지 복수의 범주명(multi-labels)을 부여하였다. 또한 실험 문헌집단에 대한 사전처리를 통해 문헌벡터를 생성하고, 벡터를 구성하는 각 용어에 가중치(tfidf)를 부여하였다(김판준, 2018; 2019).

본 연구의 연구문제에 기초한 실험 단계와 주요 성능 요소는 〈그림 2〉와 같다.

첫째, 연구문제1에 대한 자동분류 실험은 주요 성능 요소로서 두 가지 범주명 부여 방법(단일-범주명, 복수-범주명)과 4개 분류기(로치오/Rocchio, 나이브 베이즈/NB, 랜덤 포레스트/RF, 지지

1) 분류 대상 문헌에 여러 범주 중에서 하나의 범주를 할당하는 복수-범주 분류(multi-class classification)와 복수의 범주명을 할당하는 복수-범주명 분류(multi-label classification)를 구분하기 위하여, '복수-범주명 분류(multi-label classification)'로 표기하였다.




〈그림 2〉 실험 단계와 주요 성능 요소

벡터기계/SVM)[2]에 대하여 8개의 자질선정 기법(df, jac, chi, rmij, gss, pcc, mi, lor)을 적용하였다. 둘째, 연구문제2에 대한 자동분류 실험은 주요 성능 요소로서 두 가지 범주 부여 방법(단일-범주명, 복수-범주명)에 대하여 이전 실험에서 좋은 성능을 보인 2개의 분류기(Rocchio, SVM)와 4개 자질선정 기법(df, jac, chi, gss)을 사용하였다. 또한, 학습집합의 크기를 최근 년도부터 연차별(1년(2011년)~10년(2011년~2002년))로 구성하여 사용하였다. 여기서 학습집합의 규모를 기존의 자동분류 실험에서 많이 적용되었던 비율 대신 연차적으로 적용한 것은, 실제 환경에서 수작업 분류작업이 가장 최근의 논문부터 소급하여 추진되는 경향을 반영하고자 한 것이다. 셋째, 각 단계별로 분류 성능을 서로 다른 특성을 가진 매크로 평균 F1(mac_F1)과 마이크로 평균 F1(mic_F1)으로 구분하여 산출하였다(김판준, 2016; 2019; Chang et al., 2015). 또한, 분류 성능을 최고성능과 평균성능의 두 가지 측면에서 검토하였다. 여기서 최고성능은 주요 성능 요소를 적용한 결과에서 가장 높은 값이며, 평균성능은 학습집합의 크기를 연차적으로 변화시킨 결과를 평균한 값이다.

2) 본 연구에 사용된 4개 분류기는 로치오(Rocchio), 나이브 베이즈(NB: Naive Bayes), 랜덤 포레스트(RF: Random Forest), 지지벡터기계(SVM: Support Vector Machine)이다. 이후 각 분류기를 Rocchio, NB, RF, SVM으로 표기한다.

## 3. 국내 학술지 논문의 자동분류 실험 결과

### 3.1 연구문제1

국내 학술지 논문의 자동분류에서 8개 자질선정 기법을 중심으로 주요 요소(범주명 부여 방법, 분류기)에 따른 성능 차이를 살펴보기 위한 실험을 수행하였다. 〈표 3〉은 논문에 가장 적합한 하나의 범주를 부여하는 단일-범주명 분류 환경에서, 4개의 분류기(NB, RF, Rocchio, SVM)에 8개 자질선정 기법을 적용한 분류 성능을 매크로 F1 척도(mac_F1)로 산출한 것이다. 여기서 최고성능은 SVM 분류기에 자질선정 기법(jac/20%)을 적용한 경우이며(0.7079), Rocchio 분류기에 자질선정 기법(df/20%)을 적용한 경우에도 거의 동등한 성능을 보였다(0.7018). 또한, 〈표 4〉는 단일-범주명 분류 환경에서 4개의 분류 알고리즘에 8개 자질선정 기법을 적용한 분류 성능을 마이크로 F1(mic_F1) 척도로 산출한 것이다. 마이크로 F1(mic_F1)으로 산출한 최고성능은 Rocchio 분류기에 자질선정 기법(gss/70%)을 적용한 경우였으며(0.7176), SVM 분류기에 다른 2개의 자질선정 기법(df, rmij/90%)을 적용한 경우에 이와 거의 동등한 수준이었다(0.7139).

〈표 3〉 자질선정 기법을 적용한 4개 분류기의 분류 성능: 단일-범주명, mac_F1

| 구분 | NB | RF | Rocchio | SVM |
|---|---|---|---|---|
| df | 0.6523(100%) | 0.5094(60%) | **0.7018(20%)** | 0.6880(20%) |
| jac | 0.6523(100%) | 0.5052(50%) | 0.7014(40%) | **0.7079(20%)** |
| chi | 0.6523(100%) | **0.5189(10%)** | 0.6867(100%) | 0.6711(20%) |
| rmij | 0.6523(100%) | 0.5084(60%) | 0.6949(40%) | 0.6673(90%) |
| gss | 0.6523(100%) | 0.4978(100%) | 0.6970(90%) | 0.6846(50%) |
| pcc | 0.6523(100%) | 0.5122(40%) | 0.6897(100%) | 0.6799(70%) |
| mi | 0.6523(100%) | 0.5084(90%) | 0.6867(100%) | 0.6656(100%) |
| lor | 0.6523(100%) | 0.4981(100%) | 0.6930(30%) | 0.6656(100%) |
| 최고성능 | 0.6523(100%) | 0.5189(10%) | 0.7018(20%) | **0.7079(20%)** |

〈표 4〉 자질선정 기법을 적용한 4개 분류기의 분류 성능: 단일-범주명, mic_F1

| 구분 | NB | RF | Rocchio | SVM |
|---|---|---|---|---|
| df | 0.6768(100%) | 0.6485(50%) | 0.7056(100%) | **0.7139(90%)** |
| jac | 0.6768(100%) | 0.6478(60%) | 0.7107(40%) | 0.7038(100%) |
| chi | 0.6768(100%) | 0.6358(30%) | 0.7056(100%) | 0.7038(100%) |
| rmij | 0.6768(100%) | **0.6510(50%)** | 0.7056(100%) | **0.7139(90%)** |
| gss | 0.6768(100%) | 0.6359(40%) | **0.7176(70%)** | 0.7038(100%) |
| pcc | 0.6768(100%) | 0.6356(40%) | 0.7056(100%) | 0.7038(100%) |
| mi | 0.6768(100%) | 0.6346(100%) | 0.7056(100%) | 0.7038(100%) |
| lor | 0.6768(100%) | 0.6166(100%) | 0.7056(100%) | 0.7038(100%) |
| 최고성능 | 0.6768(100%) | 0.6510(50%) | **0.7176(70%)** | 0.7139(90%) |

논문에서 다루어진 주제의 비중에 따라 최대 3개까지 복수의 범주를 할당하는 복수-범주명 분류 환경에서, 4개의 분류기(NB, RF, Rocchio, SVM)에 8개 자질선정 기법을 적용한 분류 성능을 매크로 F1 척도(mac_F1)로 산출한 결과는 〈표 5〉와 같다. 여기서 최고성능은 SVM 분류기에 자질선정 기법(chi/60%)을 적용한 것이며(0.8582), Rocchio 분류기에 자질선정 기법(df/40%)을 적용한 것이 그 다음이었다(0.8011). 또한, 〈표 6〉은 복수-범주명 분류 환경에서 4개의 분류기에 8개 자질선정 기법을 적용한 분류 성능을 마이크로 F1(mic_F1)으로 산출한 것이다. 〈표 5〉의 매크로 F1으로 산출한 결과와 동일하게 최고성능은 SVM 분류기에 자질선정 기법(chi/70%)을 적용한 것이었고(0.8137), Rocchio 분류기에 자질선정 기법(df/60%)을 적용한 경우에도 이에 근접한 성능이었다(0.7916). RF 분류기에 자질선정 기법(chi/10%)을 사용한 경우에도 좋은 성능을 보였으나(0.7983), 다른 3개 분류기에 비해 처리 속도가 현저히 느린 문제가 있었다.

### 3.2 연구문제2

이전 실험(연구문제1)에서 좋은 성능을 보인 4개 자질선정 기법(df, jac, chi, gss)과 2개

〈표 5〉 자질선정 기법을 적용한 4개 분류기의 분류 성능: 복수-범주명, mac_F1

| 구분 | NB | RF | Rocchio | SVM |
|---|---|---|---|---|
| df | 0.7779(10%) | 0.7214(30%) | **0.8111(40%)** | 0.8122(70%) |
| jac | **0.7923(10%)** | 0.7210(20%) | 0.8066(10%) | 0.8502(30%) |
| chi | 0.7503(40%) | **0.7819(10%)** | 0.8038(60%) | **0.8582(60%)** |
| rmij | 0.7316(30%) | 0.7511(10%) | 0.8016(100%) | 0.8003(100%) |
| gss | 0.6893(100%) | 0.6917(100%) | 0.8016(100%) | 0.8003(100%) |
| pcc | 0.6893(100%) | 0.6942(100%) | 0.8016(100%) | 0.8003(100%) |
| mi | 0.6893(100%) | 0.7019(90%) | 0.8016(100%) | 0.8003(100%) |
| lor | 0.6893(100%) | 0.6913(100%) | 0.8016(100%) | 0.8003(100%) |
| 최고성능 | 0.7923(10%) | 0.7819(10%) | 0.8111(40%) | **0.8582(60%)** |

〈표 6〉 자질선정 기법을 적용한 4개 분류기의 분류 성능: 복수-범주명, mic_F1

| 구분 | NB | RF | Rocchio | SVM |
|---|---|---|---|---|
| df | 0.7500(100%) | 0.7836(10%) | **0.7916(60%)** | 0.7831(80%) |
| jac | 0.7606(10%) | 0.7835(10%) | 0.7854(30%) | 0.8130(30%) |
| chi | 0.7538(90%) | **0.7983(10%)** | 0.7908(50%) | **0.8137(70%)** |
| rmij | 0.7500(100%) | 0.7814(50%) | 0.7854(100%) | 0.8069(90%) |
| gss | 0.7500(100%) | 0.7528(40%) | 0.7854(100%) | 0.7678(100%) |
| pcc | 0.7500(100%) | 0.7672(40%) | 0.7854(100%) | 0.7678(100%) |
| mi | 0.7500(100%) | 0.7402(100%) | 0.7854(100%) | 0.7678(100%) |
| lor | 0.7500(100%) | 0.7371(100%) | 0.7854(100%) | 0.7678(100%) |
| 최고성능 | 0.7615(40%) | 0.7983(10%) | 0.7916(60%) | **0.8137(70%)** |

분류기(SVM, Rocchio)에 대하여 두 가지 범주명 부여 방법과 학습집합의 규모(연차별)에 따른 성능 변화를 살펴보았다.

### 3.2.1 Rocchio 분류기

단일-범주명 분류 방법으로 Rocchio 분류기에 4개 자질선정 기법을 적용하면서 학습집합을 연차적으로 증가시킨 분류 성능을 매크로 F1(mac_F1)척도로 산출한 결과는 〈표 7〉과 같다. 여기서 최고성능은 자질선정 기법(gss/70%)과 학습집합(8년)을 적용한 것이었지만(0.7172), 자질선정 기법(df/20%)과 학습집합(8년)을 적용한 경우에도 거의 유사한 성능이었다(0.7135). 또한 평균성능은 자질선정 기법 중에서 jac(0.6414)와 df(0.6409)가 가장 높은 수준이었고, 그 다음은 gss(0.6398), chi(0.6248) 순이었다.

단일-범주명 분류 환경에서 Rocchio 분류기에 대하여 4개 자질선정 기법과 연차별 학습집합을 적용한 마이크로 F1(mic_F1) 성능은 〈표 8〉이다. 여기서 Rocchio 분류기의 최고성능은 자질선정 기법(gss/70%)과 학습집합(10년)을 적용한 것이었으며(0.7176), 다른 3개의 자질선정 기법들을 적용하면서 8년 이상의 학습집합을 사용하는 경우에도 거의 대등한 수준이었다(df/0.7139, jac/0.7139, chi/0.7139). 또한 평균성능 측면에서도 자질선정 기법에 따른 성능 차이가 크지 않은 것으로 나타났다.

복수-범주명 분류 방법으로 Rocchio 분류기에 4개 자질선정 기법을 적용하면서 학습집합을 연차적으로 증가시킨 분류 성능을 매크로 F1(mac_F1)척도로 산출한 결과는 〈표 9〉이다. 여기서 최고성능은 자질선정 기법(df/90%)과 소규모의 학습집합(3년)을 적용한 것이 가장 높았고(0.8343), 다른 기법들은 이보다 낮은 수준으로 서로 큰 차이가 없었다. 한편, 평균성능은 자질선정 기법 jac(0.7815)와 df(0.7799)이 가장 높았고, 그 다음은 gss(0.7696), chi(0.7673) 순이었다.

복수-범주명 분류 환경에서 Rocchio 분류기에 4개 자질선정 기법과 연차별 학습집합을 적용한 마이크로 F1(mic_F1) 성능은 〈표 10〉이다. 여기서 최고성능은 자질선정 기법(jac/70%)

〈표 7〉 자질선정 기법을 적용한 Rocchio 분류기의 분류 성능: 단일-범주명, mac_F1

| 학습집합 구분 | df | jac | chi | gss |
|---|---|---|---|---|
| 1년 | 0.3945(80%) | 0.3943(80%) | 0.4027(60%) | 0.3787(80%) |
| 2년 | 0.5484(70%) | 0.5533(40%) | 0.4936(100%) | 0.5315(80%) |
| 3년 | 0.6714(90%) | 0.6659(50%) | 0.6531(50%) | 0.6807(90%) |
| 4년 | 0.6475(50%) | 0.6579(20%) | 0.6348(100%) | 0.6702(90%) |
| 5년 | 0.6695(50%) | 0.6794(70%) | 0.6578(100%) | 0.6578(100%) |
| 6년 | 0.6910(30%) | 0.6883(30%) | 0.6629(20%) | 0.6782(60%) |
| 7년 | 0.6812(20%) | 0.6906(20%) | 0.6811(90%) | 0.6934(80%) |
| 8년 | **0.7135(20%)** | 0.7003(70%) | **0.6951(90%)** | **0.7172(70%)** |
| 9년 | 0.6902(20%) | 0.6827(30%) | 0.6804(80%) | 0.6930(70%) |
| 10년 | 0.7018(20%) | **0.7014(40%)** | 0.6867(100%) | 0.6970(90%) |
| 최고성능 | 0.7135(20%) | 0.7014(40%) | 0.6951(90%) | **0.7172(70%)** |
| 평균성능 | 0.6409 | **0.6414** | 0.6248 | 0.6398 |

〈표 8〉 자질선정 기법을 적용한 Rocchio 분류기의 분류 성능: 단일-범주명, mic_F1

| 학습집합 구분 | df | jac | chi | gss |
|---|---|---|---|---|
| 1년 | 0.5303(80%) | 0.5303(80%) | 0.5303(60%) | 0.5101(100%) |
| 2년 | 0.6162(40%) | 0.6263(40%) | 0.5606(90%) | 0.6010(90%) |
| 3년 | 0.6650(90%) | 0.6701(50%) | 0.6532(100%) | 0.6701(90%) |
| 4년 | 0.6616(100%) | 0.6633(20%) | 0.6616(100%) | 0.6616(100%) |
| 5년 | 0.6701(50%) | 0.6785(70%) | 0.6684(20%) | 0.6684(100%) |
| 6년 | 0.6954(30%) | 0.6904(30%) | 0.6835(40%) | 0.6751(50%) |
| 7년 | 0.7038(70%) | 0.6987(60%) | 0.7038(90%) | 0.6886(100%) |
| 8년 | **0.7139(100%)** | **0.7139(100%)** | **0.7139(100%)** | 0.7157(70%) |
| 9년 | 0.6987(100%) | 0.6987(100%) | 0.6987(100%) | 0.7107(70%) |
| 10년 | 0.7056(100%) | 0.7107(90%) | 0.7056(100%) | **0.7176(70%)** |
| 최고성능 | 0.7139(100%) | 0.7139(100%) | 0.7139(100%) | **0.7176(70%)** |
| 평균성능 | 0.6661 | **0.6681** | 0.6580 | 0.6619 |

〈표 9〉 자질선정 기법을 적용한 Rocchio 분류기의 분류 성능: 복수-범주명, mac_F1

| 학습집합 구분 | df | jac | chi | gss |
|---|---|---|---|---|
| 1년 | 0.6456(30%) | 0.6581(20%) | 0.6482(90%) | 0.6444(100%) |
| 2년 | 0.7321(80%) | 0.7175(50%) | 0.6871(80%) | 0.7060(80%) |
| 3년 | **0.8343(90%)** | **0.8272(90%)** | 0.7669(40%) | 0.7908(90%) |
| 4년 | 0.7738(100%) | 0.7738(100%) | 0.7738(100%) | 0.7826(90%) |
| 5년 | 0.7892(90%) | 0.7880(30%) | 0.7820(100%) | 0.7820(100%) |
| 6년 | 0.7908(80%) | 0.7981(70%) | 0.7752(100%) | 0.7785(60%) |
| 7년 | 0.7854(20%) | 0.8096(30%) | 0.8054(70%) | 0.7810(100%) |
| 8년 | 0.8224(100%) | 0.8224(100%) | **0.8224(100%)** | **0.8224(100%)** |
| 9년 | 0.8139(40%) | 0.8132(80%) | 0.8079(60%) | 0.8070(100%) |
| 10년 | 0.8111(40%) | 0.8066(10%) | 0.8038(60%) | 0.8016(100%) |
| 최고성능 | **0.8343(90%)** | 0.8272(90%) | 0.8224(100%) | 0.8224(100%) |
| 평균성능 | 0.7799 | **0.7815** | 0.7673 | 0.7696 |

〈표 10〉 자질선정 기법을 적용한 Rocchio 분류기의 분류 성능: 복수-범주명, mic_F1

| 학습집합 구분 | df | jac | chi | gss |
|---|---|---|---|---|
| 1년 | 0.6731(100%) | 0.6731(100%) | 0.6731(90%) | 0.6731(100%) |
| 2년 | 0.7505(80%) | 0.7380(50%) | 0.7189(80%) | 0.7266(80%) |
| 3년 | 0.8008(90%) | 0.8008(90%) | 0.7731(100%) | 0.7793()90%) |
| 4년 | 0.7816(100%) | 0.7816(100%) | 0.7816(100%) | 0.7816(100%) |
| 5년 | 0.7939(10%) | 0.7969(10%) | 0.7908(100%) | 0.7908(100%) |
| 6년 | **0.8031(70%)** | **0.8145(70%)** | 0.7885(100%) | 0.7885(100%) |
| 7년 | 0.7931(100%) | 0.7954(10%) | 0.7992(70%) | 0.7931(100%) |
| 8년 | 0.8008(100%) | 0.8008(100%) | **0.8008(100%)** | **0.8008(100%)** |
| 9년 | 0.7931(40%) | 0.7931(80%) | 0.7931(50%) | 0.7916(100%) |
| 10년 | 0.7916(20%) | 0.7854(100%) | 0.7908(50%) | 0.7854(100%) |
| 최고성능 | 0.8031(70%) | **0.8145(70%)** | 0.8008(100%) | 0.8008(100%) |
| 평균성능 | **0.7782** | 0.7780 | 0.7710 | 0.7711 |

과 학습집합(6년)을 사용한 것이었고(0.8145), 다른 자질선정 기법들은 6년 또는 8년의 학습집합을 사용하는 경우에 거의 대등한 성능이었다(df/70%, 6년(0.8031), chi/100%, 8년(0.8008), gss/100%, 8년(0.8008)). 평균성능 측면에서는 df가 가장 좋은 성능이었고(0.7782), 그 다음이 jac(0.7780), ig(0.7772)순이었지만 기법들 간의 성능 차이는 크지 않았다.

### 3.2.2 SVM 분류기

단일-범주명 분류 환경에서 SVM 분류기의 4개 자질선정 기법과 학습집합의 연차별 증가에 따른 분류 성능을 매크로 F1(mac_F1)척도로 산출한 결과는 〈표 11〉과 같다. 여기서 최고성능은 자질선정 기법(jac/20%)과 학습집합(10년)을 적용한 것이었고(0.7079), 평균성능은 자질선정 기법(df)을 적용한 경우에 가장 좋았다(0.6063).

〈표 12〉는 단일-범주명 분류 환경에서 SVM 분류기에 4개 자질선정 기법과 연차별 학습집합을 사용한 마이크로 F1(mic_F1) 성능이다. 여기서 SVM 분류기의 최고성능은 가장 단순한 자질선정 기법(df/70%)과 학습집합(7년)을 적용한 것이었으며(0.7172), 평균성능도 동일한 자질선정 기법(df)을 사용한 경우가 가장 높았다(0.6568).

복수-범주명 분류 환경에서 SVM 분류기에 4개 자질선정 기법과 연차별 학습집합을 적용한 매크로 F1(mac_F1) 성능은 〈표 13〉이다. 여기서 최고성능이 가장 높은 경우는 자질선정 기법(chi/60%)과 전체 학습집합(10년)을 적용한 것이었고(0.8582), 이외에 자질선정 기법(jac/30%)과 학습집합(9년)을 사용한 경우에도 상당히 높은 수준이었다(0.8520). 그러나 평균성능은 가장 단순한 기법인 문헌빈도(df)를 사용한 경우에 가장 좋았다(0.7704).

〈표 14〉는 복수-범주 분류 환경에서 SVM 분류기에 4개 자질선정 기법과 연차별 학습집합을 적용한 마이크로 F1(mic_F1) 성능이다. 여기서 최고성능이 가장 높은 것은 자질선정 기법(chi/70%)과 학습집합(10년)을 적용한 것이며(0.8137), 다른 기법(jac/30%)도 전체 학

〈표 11〉 자질선정 기법을 적용한 SVM 분류기의 분류 성능: 단일-범주명, mac_F1

| 학습집합 구분 | df | jac | chi | gss |
|---|---|---|---|---|
| 1년 | 0.3939(70%) | 0.3939(70%) | 0.3731(80%) | 0.3736(90%) |
| 2년 | 0.4793(70%) | 0.4257(30%) | 0.3808(90%) | 0.4583(60%) |
| 3년 | 0.6331(90%) | 0.6117(80%) | 0.5981(100%) | 0.5981(100%) |
| 4년 | 0.5537(20%) | 0.5624(90%) | 0.5718(90%) | 0.5995(80%) |
| 5년 | 0.6555(90%) | 0.6528(70%) | 0.6451(100%) | 0.6451(100%) |
| 6년 | 0.6539(100%) | 0.6539(100%) | 0.6539(100%) | 0.6539(100%) |
| 7년 | 0.6634(70%) | 0.6274(70%) | 0.6113(70%) | 0.6226(80%) |
| 8년 | 0.6518(70%) | 0.6571(30%) | 0.6298(20%) | 0.6515(80%) |
| 9년 | **0.6908(20%)** | 0.6952(20%) | 0.6635(100%) | **0.7045(50%)** |
| 10년 | 0.6880(20%) | **0.7079(20%)** | **0.6711(20%)** | 0.6846(50%) |
| 최고성능 | 0.6908(20%) | **0.7079(20%)** | **0.6711(20%)** | **0.7045(50%)** |
| 평균성능 | **0.6063** | 0.5988 | 0.5799 | 0.5992 |

〈표 12〉 자질선정 기법을 적용한 SVM 분류기의 분류 성능: 단일-범주명, mic_F1

| 학습집합 구분 | df | jac | chi | gss |
|---|---|---|---|---|
| 1년 | 0.5202(100%) | 0.5202(100%) | 0.5202(100%) | 0.5253(90%) |
| 2년 | 0.5909(100%) | 0.5707(40%) | 0.5354(90%) | 0.5657(60%) |
| 3년 | 0.6414(80%) | 0.6465(80%) | 0.6263(100%) | 0.6263(100%) |
| 4년 | 0.6263(100%) | 0.6263(100%) | 0.6263(100%) | 0.6263(100%) |
| 5년 | 0.6818(90%) | 0.6515(100%) | 0.6515(100%) | 0.6515(100%) |
| 6년 | 0.6616(100%) | 0.6616(100%) | 0.6616(100%) | 0.6616(100%) |
| 7년 | 0.7172(70%) | 0.6768(100%) | 0.6768(90%) | 0.6768(100%) |
| 8년 | 0.7071(90%) | 0.7020(80%) | **0.7071(90%)** | 0.6919(100%) |
| 9년 | 0.7071(100%) | **0.7071(100%)** | **0.7071(100%)** | **0.7071(100%)** |
| 10년 | 0.7139(90%) | 0.7038(100%) | 0.7038(100%) | 0.7038(100%) |
| 최고성능 | 0.7172(70%) | 0.7071(100%) | 0.7071(100%) | 0.7071(100%) |
| 평균성능 | 0.6568 | 0.6467 | 0.6416 | 0.6436 |

〈표 13〉 자질선정 기법을 적용한 SVM 분류기의 분류 성능: 복수-범주명, mac_F1

| 학습집합 구분 | df | jac | chi | gss |
|---|---|---|---|---|
| 1년 | 0.6476(50%) | 0.6500(20%) | 0.5889(90%) | 0.5869(100%) |
| 2년 | 0.7172(90%) | 0.7092(90%) | 0.6848(80%) | 0.7109(80%) |
| 3년 | 0.8110(80%) | 0.7354(80%) | 0.7591(90%) | 0.7298(90%) |
| 4년 | 0.7418(40%) | 0.7334(90%) | 0.7517(60%) | 0.7124(80%) |
| 5년 | 0.7734(90%) | 0.8006(60%) | 0.7772(50%) | 0.7676(100%) |
| 6년 | 0.8021(90%) | 0.7732(90%) | 0.7996(60%) | 0.7732(100%) |
| 7년 | 0.7305(90%) | 0.7793(100%) | 0.7907(90%) | 0.7194(100%) |
| 8년 | **0.8387(80%)** | 0.7847(80%) | 0.8270 (60%) | 0.7585(100%) |
| 9년 | 0.8290(100%) | **0.8520(30%)** | 0.8290(100%) | **0.8290(100%)** |
| 10년 | 0.8122(70%) | 0.8502(30%) | 0.8582(60%) | 0.8003(100%) |
| 최고성능 | 0.8387(80%) | 0.8520(30%) | 0.8582(60%) | 0.8290(100%) |
| 평균성능 | 0.7704 | 0.7668 | 0.7666 | 0.7388 |

〈표 14〉 자질선정 기법을 적용한 SVM 분류기의 분류 성능: 복수-범주명, mic_F1

| 학습집합 구분 | df | jac | chi | gss |
|---|---|---|---|---|
| 1년 | 0.6667(50%) | 0.6602(20%) | 0.6448(100%) | 0.6448(100%) |
| 2년 | 0.7414(40%) | 0.7290(80%) | 0.7262(100%) | 0.7414(80%) |
| 3년 | 0.7657(30%) | 0.7648(80%) | 0.7619(80%) | 0.7447(90%) |
| 4년 | 0.7801(100%) | 0.7801(100%) | 0.7801(100%) | 0.7801(100%) |
| 5년 | 0.7548(20%) | 0.7924(80%) | 0.7763(50%) | 0.7447(80%) |
| 6년 | 0.7710(50%) | 0.7572(20%) | 0.8068(60%) | 0.7533(50%) |
| 7년 | 0.7939(50%) | 0.7947(60%) | 0.7924(80%) | **0.7924(100%)** |
| 8년 | **0.8083(80%)** | 0.7908(100%) | 0.7908(100%) | 0.7908(100%) |
| 9년 | 0.7954(40%) | 0.8023(30%) | 0.7947(60%) | 0.7916(100%) |
| 10년 | 0.7831(80%) | **0.8130(30%)** | 0.8137(70%) | 0.7678(100%) |
| 최고성능 | 0.8083(80%) | 0.8130(30%) | 0.8137(70%) | 0.7924(100%) |
| 평균성능 | 0.7660 | 0.7685 | 0.7688 | 0.7552 |

습집합(10년)을 사용하는 경우에 이와 유사한 수준이었다(0.8130). 한편, 평균성능은 3개 자질선정 기법이 거의 대등하게 좋은 성능을 보였다(chi/0.7688 ≧ jac/0.7685 ≧ df/0.7660).

## 4. 논 의

국내 학술지 논문의 자동분류에서 자질선정을 중심으로 주요 요소들의 영향을 살펴보기 위하여 다각적인 실험을 수행한 결과를 연구문제를 중심으로 검토하였다. 먼저, 연구문제1에 대한 결과는 다음과 같다. 첫째, 단일-범주명을 부여하는 환경에서 매크로 F1(mac_F1) 성능이 가장 높은 것은 SVM 분류기에 자질선정 기법(jac/20%)을 사용한 것이었다. 또한, 마이크로 F1(mic-F1) 최고성능은 Rocchio 분류기에 자질선정 기법(gss/70%)을 적용하는 것이었다. 둘째, 복수-범주명을 부여하는 환경에서는 매크로 F1(mac_F1) 최고성능은 SVM 분류기에 자질선정 기법(chi/60%)을 적용하는 것이었고, 마이크로 F1(mic_F1)의 경우에는 SVM 분류기에 자질선정 기법(chi/70%)를 사용하는 것이 가장 좋은 성능을 보였다. 따라서 국내 학술지 논문의 자동분류에서 8개 자질선정 기법과 주요 성능요소(범주 부여 방법, 분류기)에 따른 분류 성능은 차이가 있었다(연구문제1).

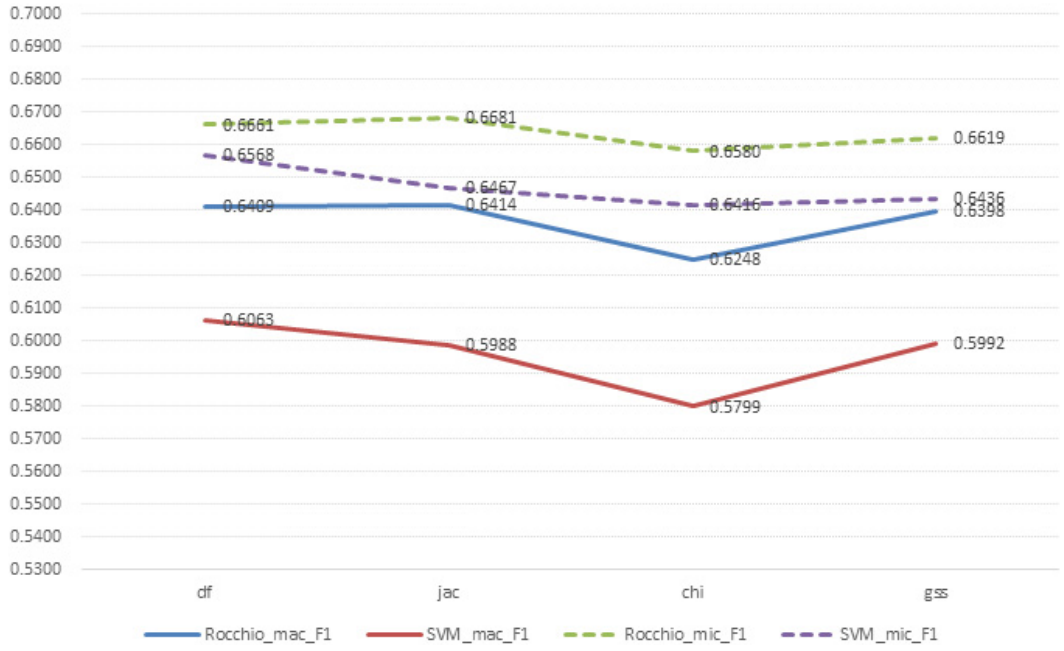
다음으로, 연구문제2에 대한 결과는 다음과 같다. 첫째, Rocchio 분류기는 단일-범주명을 부여하는 환경에서 두 가지 성능 척도 모두 자질선정 기법(gss/70%)과 비교적 많은 학습집합(8년, 10년)을 사용했을 때 성능이 가장 높았다. 그러나 복수-범주명 부여 환경에서 매크로 F1(mac_F1)으로 본 최고성능은 자질선정 기법(df/90%)과 소규모의 학습집합(3년)을 사용하는 것이 가장 좋았고, 마이크로 F1(mic_F1)으로는 자질선정 기법(jac/70%)과 중간 크기의 학습집합(6년)이 가장 좋았다. 둘째, SVM 분류기는 단일-범주명의 부여 환경에서 매크로 F1(mac_F1)으로 보면 자질선정 기법(jac/20%)과 전체 학습집합(10년)을 적용한 경우, 그리고 마이크로 F1(mic_F1) 측면에서는 자질선정 기법(df/70%)과 학습집합(7년)을 적용한 것이 가장 좋은 수준이었다. 또한, 복수-범주명 부여 환경의 경우에는 두 가지 척도 모두 자질선정 기법(chi/60%, 70%)과 전체 학습집합(10년)을 사용한 것이 가장 좋은 성능이었다. 따라서 국내 학술지 논문의 자동분류에서 4개 자질선정 기법과 주요 성능요소(범주 부여 방법, 분류기, 학습집합(연차별))에 따른 분류 성능에 차이가 있었다(연구문제2).

국내 학술지 논문의 자동분류를 위한 자질선정 기법을 중심으로 주요 성능 요소(범주 부여 방법, 분류기, 학습집합)에 따른 영향을 최고성능 측면에서 종합적으로 검토한 결과는 다음과 같다. 첫째, 국내 학술지 논문의 자동분류를 위한 분류기는 Rocchio 또는 SVM을 사용하는 것이 가장 좋은 것으로 나타났다. 둘째, 국내 학술지 논문의 자동분류를 위한 자질선정 기법으로는 4개 기법(df, jac, chi, gss)이 서로 다른 조건(범주 부여 방법, 분류기, 학습집합)에서 가장 좋은 성능을 보였다. 특히 문헌빈도(df)는 가장 단순한 기법임에도 불구하고 다른 3개의 자질선정 기법(chi, gss, jac)보다 우위에 있거나 거의 동등한 수준의 성능을 보였다. 셋째, 국내 학술

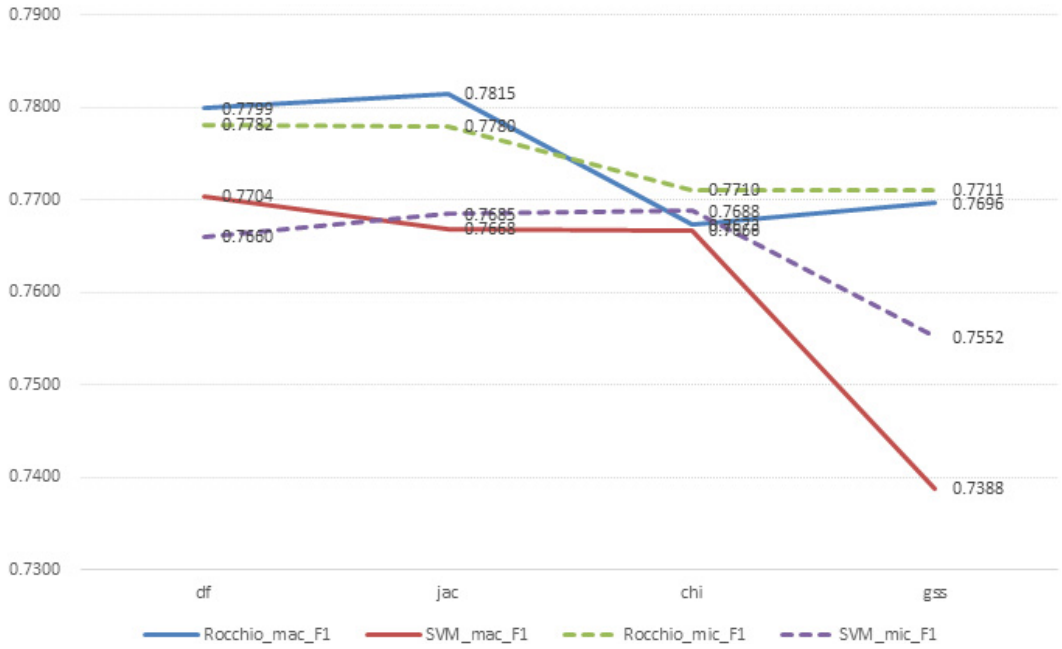
지 논문의 자동분류에서는 학습집합이 연차적으로 증가할수록 대체로 성능이 향상되는 경향을 보였다. 그러나 전체 학습집합을 사용하지 않아도 성능이 가장 좋은 경우가 다수 있었으며(3년~7년), 전반적으로 최근 3년 이상의 학습집합을 사용하면 성능이 크게 향상되는 경향이 나타났다. 한편, 국내 학술지 논문의 자동분류를 위한 자질선정 기법을 중심으로 평균성능 측면에서 종합적으로 살펴본 결과는 다음과 같다(〈그림 3〉, 〈그림 4〉 참조). 첫째, 두 가지 척도(mac_F1, mic_F1)로 산출한 평균성능에서 Rocchio가 SVM보다 근소하게 더 나은 수준이었다. 둘째, 두 가지 척도 모두 2개의 자질선정 기법(df 또는 jac)이 더 좋은 평균성능을 보였다.

## 5. 결 론

국내 학술연구의 동향을 구체적으로 파악할 수 있는 기초 데이터로서 개별 학술지 논문의 표준화된 주제 범주(통제키워드)를 제공할 수 있는 실제적인 방안을 모색하였다. 이를 위해 문헌정보학 분야의 대표적인 학술지인 '정보관리학회지'에 수록된 논문에 한국연구재단의 학문분야분류표의 주제 범주를 자동 할당하는 분류 성능 측면에서, 자질선정 기법을 중심으로 주요 요소들의 영향을 검토하기 위한 다각적인 실험을 수행하였다. 이러한 실험 결과를 최고성능과 평균성능의 두 가지 측면에서 종합적으로 분석한 결과는 다음과 같다. 첫째, 최고성능 측면에서는 2개 분류기(Rocchio, SVM)에 4개 자질선정 기법(df, jac, chi, gss)을 적용하면서




〈그림 3〉 자질선정을 적용한 분류기의 평균성능: 단일-범주명, Rocchio vs. SVM




〈그림 4〉 자질선정을 적용한 분류기의 평균성능: 복수-범주명, Rocchio vs. SVM

학습집합의 크기를 연차적으로 증가시키는 환경에서 대체로 성능이 향상되는 경향을 보였다. 주목할 것은 가장 단순한 분류기(Rocchio)와 자질선정 기법(df), 소규모의 학습집합(3년~6년)을 사용하는 경우에도 상당히 좋은 수준의 성능을 보였다는 점이다. 둘째, 평균성능 측면에서는 2개 분류기(Rocchio, SVM)에 2개 자질선정 기법(df, jac)을 적용하는 것이 가장 좋은 수준이었다. 결론적으로 국내 학술지 논문의 자동분류는 단순한 분류기와 자질선정 기법, 비교적 소규모의 학습집합을 사용하여 상당히 좋은 수준의 성능을 기대할 수 있는 것으로 나타났다. 따라서 국내 학술지 논문에 대한 자동분류는 보다 단순한 분류기와 자질선정 기법을 적용하면서, 최근 출판된 논문들부터 연차적으로 학습집합을 구성하여 추진하는 것이 가장 효율적인 방안이라 할 수 있다.

본 연구의 의의는 현재 국내 대부분의 학술 데이터베이스에서 개별 논문에 대한 표준화된 주제 범주(통제키워드)가 거의 제공되고 있지 않은 환경에서, 보다 적은 노력과 예산을 투입하여 실제적으로 자동분류를 추진할 수 있는 효율적인 방안을 제시한 것이다. 그러나 이러한 연구 결과는 특정 분야의 학술지에 수록된 논문집합을 대상으로 실험한 결과이므로 전체 학문분야로 일반화하기에는 어려움이 있다. 따라서 동일 분야의 전체 학술지 또는 다른 학문분야 등으로 실험 문헌집합을 확장하여 일반화하는 지속적인 연구가 필요하다. 또한, 다른 문헌유형(신문기사, 특허 등)에 대한 자동분류로 확장하는 연구도 필요할 것이다.

# 참 고 문 헌

김선우, 고건우, 최원준, 정희석, 윤화묵, 최성필 (2018). 기술과학 분야 학술문헌에 대한 학습집합 반자동 구축 및 자동 분류 통합 연구. 정보관리학회지, 35(4), 141-164. http://dx.doi.org/10.3743/KOSIM.2018.35.4.141

김판준 (2006). 기계학습을 통한 디스크립터 자동부여에 관한 연구. 정보관리학회지, 23(1), 279-299. https://doi.org/10.3743/KOSIM.2006.23.1.279

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

김판준 (2018). 기계학습에 기초한 국내 학술지 논문의 자동분류에 관한 연구. 정보관리학회지, 35(2), 37-62. https://doi.org/10.3743/KOSIM.2018.35.2.037

김판준 (2019). 랜덤포레스트를 이용한 국내 학술지 논문의 자동분류에 관한 연구. 정보관리학회지, 36(2), 57-77. http://dx.doi.org/10.4275/10.3743/KOSIM.2019.36.2.057

김판준 (2021a). 동시출현단어 분석에 기초한 지적구조 분석에서 키워드 유형별 특성에 관한 연구: 국외 오픈액세스 분야를 중심으로. 한국문헌정보학회지, 55(3), 103-129. http://dx.doi.org/10.4275/KSLIS.2021.55.3.103

김판준 (2021b). 프로파일링에 기초한 키워드 유형별 지적구조 분석에 관한 연구: 국외 오픈액세스 분야를 중심으로. 한국문헌정보학회지, 55(4), 115-140. http://dx.doi.org/10.4275/KSLIS.2021.55.4.115

김판준, 이재윤 (2012). 디스크립터 자동 할당을 위한 저자키워드의 재분류에 관한 실험적 연구. 정보관리학회지, 29(2), 225-246. http://dx.doi.org/10.3743/KOSIM.2012.29.2.225

김판준, 이재윤 (2014). 해외 데이터베이스의 통제키워드에 기초한 국내 학술지 논문의 자동분류 성능 향상에 관한 실험적 연구. 한국문헌정보학회지, 48(3), 491-510. http://dx.doi.org/10.4275/KSLIS.2014.48.3.491

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

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

정은경 (2009). 문서범주화 성능 향상을 위한 의미기반 자질확장에 관한 연구. 정보관리학회지, 26(3), 261-278. http://dx.doi.org/10.3743/KOSIM.2009.26.3.261

한국연구재단 (2016). 학술연구분야분류표. Available: https://www.nrf.re.kr/biz/doc/class/view?menu_no=323

한국학술지인용색인(KCI) (2022). Data 구축 통계. 한국연구재단. Available: https://www.kci.go.kr/kciportal/po/statistics/poStatisticsMain.kci?tab_code=Tab3

Abiodun, E. O., Alabdulatif, A., Abiodun, O. I., Alawida, M., Alabdulatif, A., & Alkhawaldeh, R. S. (2021). A systematic review of emerging feature selection optimization methods for optimal text classification: the present state and prospective opportunities. Neural Computing & Applications, 33(4), 1-28. https://doi.org/10.1007/s00521-021-06406-8

Cai, J., Luo, J., Wang, S., & Yang, S. (2018). Feature selection in machine learning: a new perspective. Neurocomputing, 300, 70-79. https://doi.org/10.1016/j.neucom.2017.11.077

Chandrashekar, G. & Sahin, F. (2014) A survey on feature selection methods. Computers & Electrical Engineering, 40(1), 16-28. https://doi.org/10.1016/j.compeleceng.2013.11.024

Chang, F., Guo, J., Xu, W., & Yao, K. (2015). A feature selection method to handle imbalanced data in text classification. Journal of Digital Information Management, 13, 169-175. Available: https://www.dline.info/fpaper/jdim/v13i3/v13i3_6.pdf

Deng, X., Li, Y., Weng, J., & Zhang, J. (2019). Feature selection for text classification: a review. Multimedia Tools and Applications, 78, 3797-3816. https://doi.org/10.1007/s11042-018-6083-5

Drotár, P., Gazda, J., & Smékal, Z. (2015). An experimental comparison of feature selection methods on two-class biomedical datasets. Computers in Biology and Medicine, 66, 1-10. https://doi.org/10.1016/j.compbiomed.2015.08.010

Drotár, P., Gazda, M., & Vokorokos, L. (2019). Ensemble feature selection using election methods and ranker clustering. Information Sciences, 480, 365-380. https://doi.org/10.1016/j.ins.2018.12.033

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. The Journal of Machine Learning Research, 3, 1289-1305. Available: https://www.jmlr.org/papers/volume3/forman03a/forman03a_full.pdf

Fragoudis, D., Meretakis, D., & Likothanassis, S. (2005). Best terms: an efficient feature-selection algorithm for text categorization. Knowledge and Information Systems, 8(1), 16-33. https://doi.org/10.1007/s10115-004-0177-2

Günal, S. (2012). Hybrid feature selection for text classification. Turkish Journal of Electrical Engineering and Computer Science, 20(Sup.2), 1296-1311. Available: https://dergipark.org.tr/en/pub/tbtkelektrik/issue/12058/144170

Gutkin, M., Shamir, R., & Dror, G. (2009). SlimPLS: a method for feature selection in gene expression-based disease classification. PloS One, 4(7), e6416.

https://doi.org/10.1371/journal.pone.0006416

Guyon, I. & Elisseeff, A. (2003). An introduction to variable and feature selection. The Journal of Machine Learning Research, 3, 1157-1182. Available: https://dl.acm.org/doi/pdf/10.5555/944919.944968

Harish, B. & Revanasiddappa, M. (2017). A comprehensive survey on various feature selection methods to categorize text documents. International Journal of Computer Applications, 164, 1-7. http://doi.org/10.5120/ijca2017913711

Iqbal, M., Abid, M. M., Khalid, M. N., & Manzoor, A. (2020). Review of feature selection methods for text classification. International Journal of Advanced Computer Research, 10(49), 138-152. http://dx.doi.org/10.19101/IJACR.2020.1048037

Joachims, T. (1997). A probabilistic analysis of the Rocchio algorithm with TFIDF for text categorization. Proceedings of the Fourteenth International Conference on Machine Learning (ICML '97), 143-151. Available: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.6977&rep=rep1&type=pdf

Joachims, T. (2002). Learning to Classify Text Using Support Vector Machines: Methods, theory and algorithms. USA: Kluwer Academic Publishers.

Kashef, S., Nezamabadi-pour, H., & Nikpour, B. (2018). Multi-label feature selection: a comprehensive review and guiding experiments. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 8(2), e1240. https://doi.org/10.1002/widm.1240

Kragelj, M. & Kljajić Borštnar, M. (2021). Automatic classification of older electronic texts into the Universal Decimal Classification-UDC. Journal of Documentation, 77(3), 755-776. https://doi.org/10.1108/JD-06-2020-0092

Kumar, V. & Minz, S. (2014). Feature selection: a literature review. Smart Computing Review, 4(3), 211-229. Available: https://faculty.cc.gatech.edu/~hic/CS7616/Papers/Kumar-Minz-2014.pdf

Manning, C., Raghavan, P., & Schutze, H. (2008). Introduction to information retrieval. NY, USA: Cambridge University Press.

Mengle, S. S. R. & Goharian, N. (2009). Ambiguity measure feature-selection algorithm. Journal of the American Society for Information Science & Technology, 60(5), 1037-1050. https://doi.org/10.1002/asi.21023

Mironczuk, M. & Protasiewicz, J. (2018). A recent overview of the state-of-the-art elements of text classification. Expert Systems with Applications, 106, 36-54. https://doi.org/10.1016/j.eswa.2018.03.058

Pereira, R. B., Plastino, A., Zadrozny, B., & Merschmann, L. H. (2018). Correlation analysis of performance measures for multi-label classification. Information Processing & Management, 54(3), 359-369. https://doi.org/10.1016/j.ipm.2018.01.002

Pinheiro, R. H. W., Cavalcanti, G. D. C., & Ren, T. I. (2015). Data-driven global-ranking local feature selection methods for text categorization, Expert Systems with Applications, 42 (4), 1941-1949. https://doi.org/10.1016/j.eswa.2014.10.011

Pintas, J. T., Fernandes, L. A. F., & Garcia, A. C. B. (2021). Feature selection methods for text classification: a systematic literature review. Artificial Intelligence Review, 54, 6149-6200. https://doi.org/10.1007/s10462-021-09970-6

Rehman, A., Javed, K., Babri, H. A., & Asim, N. (2018). Selection of the most relevant terms based on a max-min ratio metric for text classification. Expert Systems with Applications, 114, 78-96. https://doi.org/10.1016/j.eswa.2018.07.028

Salton, G. & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. Information Processing & Management, 24(5), 513-523. https://doi.org/10.1016/0306-4573(88)90021-0

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys (CSUR), 34(1), 1-47. https://doi.org/10.1145/505282.505283

Talavera, L. (2005). An evaluation of filter and wrapper methods for feature selection in categorical clustering. In International Symposium on Intelligent Data Analysis. Springer, Berlin, Heidelberg, 440-451. https://doi.org/10.1007/11552253_40

Uysal, A. K. (2016). An improved global feature selection scheme for text classification. Expert Systems with Applications, 43(1), 82-92, https://doi.org/10.1016/j.eswa.2015.08.050

Venkatesh, B. & Anuradha, J. (2019). A review of feature selection and its methods. Cybernetics and Information Technologies, 19(1), 3-26. https://doi.org/10.2478//cait-2019-0001

Wang, D., Zhang, H., Liu, R., Liu, X., & Wang, J. (2016). Unsupervised feature selection through gram-Schmidt orthogonalization-A word co-occurrence perspective. Neurocomputing, 173(P3), 845-854. https://doi.org/10.1016/j.neucom.2015.08.038

Wang, D., Zhang, H., Liu, R., Lv, W., & Wang, D. (2014). t-test feature selection approach based on term frequency for text categorization. Pattern Recognition Letters, 45, 1-10. https://doi.org/10.1016/j.patrec.2014.02.013

Wu, Y. & Zhang, A. (2004). Feature selection for classifying high-dimensional numerical data. Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004, CVPR 2004, 2, 251-258.

http://doi.org/10.1109/CVPR.2004.1315171

Yang, Y. & Pedersen. J. O. (1997). A comparative study on feature selection in text categorization. In Proceedings of the Fourteenth International Conference on Machine Learning, July 08-12, 412-420. Available: http://nyc.lti.cs.cmu.edu/yiming/Publications/yang-icml97.pdf

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

Chung, Eunkyung (2009). A semantic-based feature expansion approach for improving the effectiveness of text categorization by using WordNet. Journal of the Korean Society for information Management, 26(3), 261-278. https://doi.org/10.3743/KOSIM.2009.26.3.261

KCI(Korea Citation Index) (2022). Data Statistics. National Research Foundation of Korea. Available: https://www.kci.go.kr/kciportal/po/statistics/poStatisticsMain.kci?tab_code=Tab3

Kim, Pan Jun & Lee, Jae Yun (2012). A study on the reclassification of author keywords for automatic assignment of descriptors. Journal of the Korean Society for Information Management, 29(2), 225-246. https://doi.org/10.3743/KOSIM.2012.29.2.225

Kim, Pan Jun & Lee, Jae Yun (2018). An experimental study on the performance improvement of automatic classification for the articles of Korean journals based on controlled keywords in international database. Journal of the Korean Library and Information Science, 48-3, 491-510. https://doi.org/10.4275/KSLIS.2014.48.3.491

Kim, Pan Jun (2006). A study on automatic assignment of descriptors using machine learning. Journal of the Korean Society for Information Management, 23(1), 279-299. https://doi.org/10.3743/KOSIM.2006.23.1.279

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. http://dx.doi.org/10.3743/KOSIM.2016.33.2.033

Kim, Pan Jun (2018). An analytical study on automatic classification of domestic journal articles based on machine learning. Journal of the Korean Society for Information Management, 35(2), 37-62. https://doi.org/10.3743/KOSIM.2018.35.2.037

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), 37-62. https://doi.org/10.3743/KOSIM.2019.36.2.057

Kim, Pan Jun (2021a). A study on the characteristics by keyword types in the intellectual

structure analysis based on co-word analysis: focusing on overseas open access field. Journal of the Korean Library and Information Science, 55-3, 103-129. http://dx.doi.org/10.4275/KSLIS.2021.55.3.103

Kim, Pan Jun (2021b). A study on the intellectual structure analysis by keyword type based on profiling: focusing on overseas open access field. Journal of the Korean Library and Information Science, 55-4, 115-140. http://dx.doi.org/10.4275/KSLIS.2021.55.4.115

Kim, Seon-Wu, Ko, Gun-Woo, Choi, Won-Jun, Jeong, Hee-Seok, Yoon, Hwa-Mook, & Choi, Sung-Pil (2018). Semi-automatic construction of learning set and integration of automatic classification for academic literature in technical sciences. Journal of the Korean Society for Information Management, 35(4), 141-164. http://dx.doi.org/10.3743/KOSIM.2018.35.4.141

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

National Research Foundation of Korea (2016). Academic Research Classification Scheme. Available: https://www.nrf.re.kr/biz/doc/class/view?menu_no=323

Yuk, Jee Hee & 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