기계학습에 기초한 국내 학술지 논문의 자동분류에 관한 연구*

An Analytical Study on Automatic Classification of Domestic Journal articles Based on Machine Learning

김판준 (Pan Jun Kim)**

초 록

문헌정보학 분야의 국내 학술지 논문으로 구성된 문헌집합을 대상으로 기계학습에 기초한 자동분류의 성능에 영향을 미치는 요소들을 검토하였다. 특히, 「정보관리학회지」에 수록된 논문에 주제 범주를 자동 할당하는 분류 성능 측면에서 용어 가중치부여 기법, 학습집합 크기, 분류 알고리즘, 범주 할당 방법 등 주요 요소들의 특성을 다각적인 실험을 통해 살펴보았다. 결과적으로 분류 환경 및 문헌집합의 특성에 따라 각 요소를 적절하게 적용하는 것이 효과적이며, 보다 단순한 모델의 사용으로 상당히 좋은 수준의 성능을 도출할 수 있었다. 또한, 국내 학술지 논문의 분류는 특정 논문에 하나 이상의 범주를 할당하는 복수-범주 분류(multi-label classification)가 실제 환경에 부합한다고 할 수 있다. 따라서 이러한 환경을 고려하여 단순하고 빠른 분류 알고리즘과 소규모의 학습집합을 사용하는 최적의 분류 모델을 제안하였다.

ABSTRACT

This study examined the factors affecting the performance of automatic classification based on machine learning for domestic journal articles in the field of LIS. In particular, In view of the classification performance that assigning automatically the class labels to the articles in 「Journal of the Korean Society for Information Management」, I investigated the characteristics of the key factors(weighting schemes, training set size, classification algorithms, label assigning methods) through the diversified experiments. Consequently, It is effective to apply each element appropriately according to the classification environment and the characteristics of the document set, and a fairly good performance can be obtained by using a simpler model. In addition, the classification of domestic journals can be considered as a multi-label classification that assigns more than one category to a specific article. Therefore, I proposed an optimal classification model using simple and fast classification algorithm and small learning set considering this environment.

키워드: 자동분류, 텍스트 범주화, 성능 요소, 학술지 논문, 로치오, 지지벡터기계, 나이브 베이즈, 단일-범주 분류, 복수-범주 분류, 기계학습
automatic classification, text categorization, performance factors, Journal articles, Rocchio, SVM (Support Vector Machine), NB (Naïve Bayes), single-label classification, multi-label classification, machine learning

* 이 논문은 2016년 대한민국 교육부와 한국연구재단의 지원을 받아 수행된 연구임 (NRF-2016S1A5A2A01021902).

** 신라대학교 문헌정보학과 부교수(pjkim@silla.ac.kr)

■ 논문접수일자 : 2018년 5월 17일 ■ 최초심사일자 : 2018년 6월 17일 ■ 게재확정일자 : 2018년 6월 19일
■ 정보관리학회지, 35(2), 37-62, 2018. 〔http://dx.doi.org/10.3743/KOSIM.2018.35.2.037〕

1. 서 론

1.1 연구의 필요성 및 목적

21세기 디지털 시대의 도래와 함께 전 세계적으로 학술정보의 생산 및 유통이 폭발적으로 증가하였다. 이에 따라 학문분야별로 연구의 흐름과 동향을 체계적으로 파악하여 효율적인 연구개발 활동의 지원 및 평가와 함께 미래의 연구 방향을 설정하기 위한 기초 데이터의 필요성이 날이 갈수록 증대하고 있다. 이러한 측면에서 국내외 학술데이터베이스에 축적된 서지레코드와 메타데이터는 과거와 현재의 연구 동향을 다양한 측면에서 체계적으로 분석할 수 있는 대표적인 기초 데이터이다. 이 중에서도 학문분야별로 세부적인 연구의 양상을 구체적으로 파악하기 위해서는 개별 자료의 분류 정보가 필수적이다. 그러나 이러한 정보가 기본 항목으로 제공되고 있는 해외 학술데이터베이스와는 달리, 국내에서는 학술지 논문에 대한 분류 정보가 제대로 제공되지 않고 있다(김판준, 이재윤, 2014).

현재 국내 학술지 논문의 분류에 사용할 수 있는 한국연구재단(2016)의 「학술연구분야 분류표」가 마련되어 있으나, 국내에서 출판된 학술지 논문에는 이에 기초한 논문 단위의 세부 분류명이 부여되어 있지 않다. 지난 수십 년간 출판된 모든 학술지 논문에 대한 분류작업을 전문 인력을 활용하여 수작업으로 단기간에 추진하기는 불가능하다. 한 예로 2018년 5월 현재 KCI 통계에 따르면 5,445종의 학술지에 수록된 총 1,444,977건의 국내 학술지 논문이 KCI 서비스를 통해 서비스되고 있으며, 이 수치는 매년 지속적으로 증가하고 있다(www.kci.go.kr). 이러한 대규모 문헌집단의 모든 문헌을 특정 시점에 일괄적으로 수작업 분류하는 것은 막대한 시간과 인력, 비용이 소요되므로 사실상 불가능다고 할 수 있다. 따라서 기존의 수작업 분류에서 필연적인 시간과 전문 인력의 부족은 물론 예산상의 문제를 극복할 수 있는 효율적인 대안으로 기계학습에 기초한 국내 학술지 논문의 자동분류를 적극적으로 모색할 필요가 있다.

본 연구의 목적은 국내 학술활동의 흐름과 동향을 실질적으로 파악하여 연구개발 활동의 체계적인 지원 및 평가는 물론 미래의 연구 방향을 설정할 수 있는 기초 데이터로서 학술지 논문의 분류정보를 제공할 수 있는 효율적인 방안을 제안하는 것이다. 구체적으로 본 연구는 기계학습에 기초한 자동분류 기법을 활용하여 한국연구재단(2016) 「학술연구분야 분류표」 상의 분류 범주(소분류명/세분류명)를 국내 학술지 논문에 자동 할당하는 효율적인 방안을 검토하였다. 이를 위해 기계학습에 기초한 자동분류의 성능에 영향을 미치는 주요 요소들에 대한 다각적인 실험을 수행하고, 그 결과를 분석하여 국내 학술지 논문의 자동분류를 위한 최적의 분류모형을 제안하고자 한다.

1.2 선행 연구

1960년대에 시작된 문헌의 자동분류(또는 텍스트 범주화)에 관한 연구는 1990년대에 기계학습 이론이 도입되면서 활성화되었고, 이에 따라 텍스트를 대상으로 하는 분류기의 성능이 크게 향상되었다(Sebastiani, 2002). 국내외의

관련 연구는 대부분 실험 환경에서 표준 데이터세트(Reuters, 20-Newsgroups, OSUMED, TREC, Ling-Spam 등)를 대상으로 분류 성능의 향상을 도모하는 것이었다. 이러한 연구들은 대부분 다양한 응용분야에서 분류 성능의 개선을 위해 특정 또는 일부 영향 요소들에 중점을 두어 실험을 수행한 결과를 보고하였다(송성전, 정영미, 2012; 이용구, 2013; 이용구, 2009; 이재윤, 2005a; 이재윤, 2005b; Joorabchi & Mahdi, 2011; Foulds, 2010; Khan, Baharudin, & Lee, 2010).

기계학습 기반의 자동분류는 다양한 분류 알고리즘과 이들을 서로 조합한 하이브리드 방식에 기초한 연구들이 빠른 성장과 확산의 양상을 보이고 있다(Pawar & Gawande, 2012; Jiang et al., 2012; Chen & Chen, 2011; Chen et al., 2011; Miao & Kamel, 2011; Uğuz, 2011; Kumar & Gopal, 2010; Vasuki & Cohen, 2010; Wu 2009; Yu, Xu, & Li, 2008; Wang & Chiang, 2007). 이외에도 기계학습 알고리즘과 전문가시스템의 조합(Villena-román et al., 2011; Li & Park, 2009), 미분류 문헌의 활용(김판준, 이재윤, 2007; Torii et al., 2011), WordNet이나 Wikipedia와 같은 외부 정보의 활용(김용환, 정영미, 2012; 정은경, 2009) 등 자동분류의 성능 향상을 위한 다양하고 새로운 기법들이 지속적으로 개발 및 적용되었다. 최근에는 다중-사례 학습(multi-instance learning)이나 다중-관점(multi-view)의 분류자질, 토픽 모델링 기반의 문헌표현, 앙상블 방식에 기초한 새로운 분류 알고리즘(random forest, AdaBoost.MH, deep blue 등)과 함께 복수-범주 분류(multi-class classification 또는 multi-label classification) 등에 관한 논의가 활발하게 진행되고 있다(김종민, 유창동, 2014; AI-Salemi et al., 2015; Jindal, Malhotra, & Jain, 2015; Tarragó et al., 2014; Read et al., 2011; Schapire & Singer, 2000).

지도학습 기반의 자동분류는 분류 대상 문헌에 하나의 범주 또는 복수의 범주를 할당하는 가에 따라 단일-범주 분류(single-label classification)와 복수-범주 분류(multi-label classification)로 구분할 수 있다. 최근 복수-범주 분류 관련 연구가 많이 수행되고 있으며, 단일-범주 분류에 비해서 성능이 낮은 것으로 보고되었다(Shehab et al., 2016; AI-Salemi et al., 2015; Vogrinčič & Bosnić, 2011; Khan, Baharudin, & Lee, 2010). 학술지 논문은 특정 논문이 하나의 주제에 관련된 경우도 있지만, 여러 주제에 걸친 내용을 함께 다루는 경우도 적지 않다. 따라서 학술지 논문을 대상으로 하는 자동분류의 성능 요소로서 범주 부여 방법(단일-범주 분류와 복수-범주 분류)을 검토할 필요가 있다.

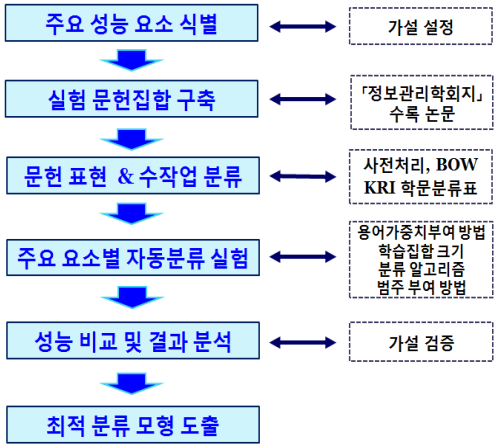
국내에서 학술지 논문을 대상으로 기계학습에 기초한 자동분류(또는 텍스트 범주화)를 적용한 연구는 많지 않다. 2000년대 중반 이후 지금까지 김판준과 이재윤(김판준, 2016; 김판준, 이재윤, 2014; 김판준, 이재윤, 2012; 김판준, 2008; 김판준, 이재윤, 2007; 김판준, 2006a; 김판준, 2006b)이 이러한 연구를 주도하고 있으며, 이외에 이용구(2013; 2009), 김성희와 엄재은(2008), 심경(2006a; 2006b), 정은경(2009)의 연구가 있다. 그러나 이들 연구 중에서 국내에서 현재 출판되고 있는 학술지 논문을 대상으로, 실제 사용되고 있는 분류체계(한국연구재단, 2016)의 범주(분류명)를 자동 할당하는 방안을 모색

한 연구는 찾아볼 수 없다. 따라서 본 연구는 특정 학문분야의 학술지 논문으로 구성된 문헌집단을 대상으로, 「학술연구분야 분류표」의 소분류 및 세분류명을 자동 할당하는데 영향을 미치는 주요 요소들을 다각적으로 검토하여 국내 학술지 논문의 자동분류에 최적화된 분류모형을 제안하였다.

2. 연구 방법

2.1 연구 단계

본 연구의 단계별 내용과 방법은 〈그림 1〉과 같다.



〈그림 1〉 연구 단계

첫째, 국내 학술지 논문의 자동분류를 위한 최적의 분류모형을 개발하기 위하여 자동분류의 성능에 영향을 미치는 요소들을 식별하고 이에 기초하여 연구 가설을 설정하였다. 특히, 선행연구를 통해 식별된 주요 요소로서 용어가중치부여 기법, 학습집합 크기, 분류 알고리즘, 범주 부여 방법(단일-범주 분류, 복수-범주 분류)에 대한 가설을 설정하였다.

둘째, 본 연구의 실험을 위한 실험 문헌집단을 구축하였다. 전체 학문분야의 모든 학술지 논문에 대한 분류작업을 일시에 추진하기는 불가능하므로 최근에 출판된 논문부터 단계적으로 추진하는 것이 실제적이라 할 수 있다. 따라서 연구기간을 고려하여 문헌정보학 분야의 학술지 「정보관리학회지」에 수록된 최근 14년간의 논문집합(2002년~2015년)을 대상으로 실험 문헌집단을 구성하였다.

셋째, 실험 문헌집단에 속한 논문의 텍스트(제목, 초록, 저자키워드)에 기초하여 전체 문헌집합을 BOW(Bag of Word) 형식으로 변환하고 논문벡터($\vec{d}$)를 생성하였다. 또한, 문헌정보학 분야 현직 교수 3인이 각 논문에 대하여 한국연구재단 「학술연구분야 분류표」상의 분류명(소분류명, 세분류명)을 수작업으로 직접 부여하고, 적절한 범주가 부여되었는지에 대한 교차검증을 실시하여 분류작업의 신뢰도를 확보하였다.

넷째, 국내 학술지 논문에 대한 자동분류의 성능에 영향을 미치는 주요 요소별로 성능을 비교하는 실험을 수행하였다. 실험 문헌집단을 학습집합(10년)과 검증집합(4년)으로 구분한 다음, 이전에 출판된 논문(training instances)에 부여된 범주를 학습하여 이후의 입력 논문(test instances)에 자동 할당한 결과를 전문인력의 수작업 분류 결과와 비교하여 성능을 평가하였다.

다섯째, 주요 요소별로 성능을 비교하는 실험 결과에 기초하여 가설의 검증 여부를 판정

하였다. 즉, 국내 학술지 논문의 자동분류에서 용어 가중치부여 기법, 학습집합 크기, 분류 알고리즘, 범주 부여 방법(단일-범주 분류, 복수-범주 분류)을 다각적으로 적용한 결과를 분석하여 가설을 검증하였다.

여섯째, 실험 및 가설 검증의 결과를 종합적으로 분석한 결과에 따라 국내 학술지 논문에 한국연구재단(2016) 「학술연구분야 분류표」 상의 분류 범주(소분류명/세분류명)를 자동 할당하기 위한 최적의 분류모형을 도출하였다.

2.2 연구 가설

문헌의 자동분류에서 성능에 영향을 미치는 특정 요소에 중점을 두어 분류 성능의 향상을 모색한 연구는 다양한 측면에서 시도되었다. 이러한 연구들에서 주로 다루어진 대표적인 성능 요소는 용어 가중치부여 기법, 학습집합 크기, 자질선정 방법, 분류 알고리즘, 범주 부여 방법 등이다(김판준, 2016). 그러나 본 연구는 자질 정보를 최대한 활용하기 위하여 성능 요소로 별도의 자질선정 방법을 적용하지 않고 전체 자질집합을 모두 사용하였다. 따라서 국내 학술지 논문의 자동분류를 위한 최적의 분류모형을 도출하기 위하여, 자동분류의 성능에 영향을 미치는 주요 요소들에 대한 가설을 다음과 같이 크게 네 가지로 설정하였다.

- 가설 1. 용어 가중치부여 방법은 국내 학술지 논문의 자동분류 성능에 영향을 미친다.
- 가설 2. 학습집합의 크기가 국내 학술지 논문의 자동분류 성능에 영향을 미친다.
- 가설 3. 분류 알고리즘이 국내 학술지 논문의 자동분류 성능에 영향을 미친다.
- 가설 4. 범주 부여 방법(단일-범주 분류와 복수-범주 분류)은 국내 학술지 논문의 자동분류 성능에 영향을 미친다.

2.3 실험 문헌집단

본 연구의 실험 문헌집단은 문헌정보학 분야의 『정보관리학회지』에 수록된 최근 14년(2002년~2015년)의 논문 중에서 한글로 작성되고 저자 키워드와 초록이 있는 논문 651편으로 구성하였다. 총 651편의 논문 중에서 이전 10년(2002년~2011년)의 453편(70%)을 학습집합, 이후 4년(2012년~2015년)의 198편(30%)은 검증집합으로 사용하였다. 선행 연구에서 이러한 7대3 분할은 좋은 분류 성능을 보이는 것으로 보고된 바 있다(Dalal & Zaveri, 2012; Dalal & Zaveri, 2013). 〈표 1〉은 본 연구에서 사용한 실험 문헌집단의 통계이다.

실험 문헌집합을 구성하는 국내 학술지(「정보관리학회지(2002년~2015년」) 수록 논문의 텍스트(제목, 초록, 저자키워드)를 대상으로 단어분리(tokenization), 형태소분석(morphological analysis), 불용어제거(stopword removing) 등 자동색인과 유사한 사전처리를 통해 실험을 위한 자질집합을 구성하였다. 그 결과, 학습 및 검증에 사용되는 문헌의 길이는 평균 119.5개였으며, 전체 키워드의 종수는 9,430개였다. 사전처리는 파이썬 언어로 구현한 프로그램과 한글 형태소 분석기(강승식, 2002: KLT2015), 마이크로소프트 엑셀 등을 사용하였다.

〈표 1〉 실험 문헌집단 통계

번호	항목	내역
1	전체 문헌 수/학습문헌 수/검증문헌 수	651/453(69.6%)/198(30.4%)
2	전체 범주 수/단일범주 수/복수범주 수	18/16/18
3	단일-범주 문헌빈도(최대/최소/평균)	137/1/36.2
4	복수-범주 문헌빈도(최대/최소/평균)	177/3/47.7
5	단일-범주 당 긍정문헌 수(최대/최소/평균)	95/1/25.2
6	복수-범주 당 긍정문헌 수(최대/최소/평균)	120/2/62
7	키워드 문헌빈도(최대/최소/평균)	282/1/4.35
8	학습문헌 당 키워드 수(최대/최소/평균)	264/42/119.5
9	키워드 종수	9,430

2.4 분류 범주

현재 국내 학술지 논문에는 논문 단위의 분류 범주가 부여되어 있지 않으므로, 문헌정보학 분야의 학술지인 「정보관리학회지(2002년~2015년)」에 수록된 논문을 대상으로 한국연구재단(2016) 「학술연구분야 분류표」상의 분류명(소분류명, 세분류명)을 수작업으로 부여하였다. 먼저, 문헌정보학과 교수 3인이 각자 각 논문에 대표 주제를 단일 범주로 부여하되, 복수의 주제가 포함된 논문에 대해서는 해당 논문의 내용에서 다루어진 비중에 따라 복수의 범주를 3개까지 순서대로 부여하였다. 다음으로 각 논문에 적절한 범주(분류명)가 부여되었는지에 대한 교차검증 작업을 통해 최종 분류 범주를 결정하여 분류작업의 일관성과 신뢰도를 확보하였다. 그 결과, 실험 문헌집합에 최종 부여된 분류 범주의 통계는 〈표 2〉와 같다. 여기서 실험 문헌집합이 대부분 학습문헌의 수가 비교적 적은 저빈도 범주로 구성되어 있으며, 각 범주별 학습집합의 편차가 큰 불균형 데이터(imbalanced data)임을 알 수 있다(AI-Salemi et al., 2015; Eriksson, 2013; Joorabchi & Mahdi, 2011; Liu et al., 2007).

2.5 분류 알고리즘

국내 학술지 논문의 자동분류 성능에 영향을 미치는 주요 요소로서 3개의 분류 알고리즘(로치오(Rocchio), 나이브 베이즈(NB: Naive Bayes), 지지벡터기계(SVM: Support Vector Machine))에 기초한 분류기를 사용하였다. 이후 기술(記述) 상의 편의를 위해 논문의 나머지 부분에서 로치오(Rocchio) 분류기, 나이브 베이즈(NB: Naive Bayes) 분류기, 지지벡터기계(SVM: Support Vector Machine) 분류기를 각각 Rocchio, NB, SVM으로 표기하였다. 이러한 3개 분류 알고리즘에 기초한 분류기의 구현 방법과 주요 파라미터는 다음과 같다.

첫째, Rocchio 분류기는 Python 프로그래밍을 통해 구현한 로치오 분류기를 이용하였다. 각 범주별로 분류기를 각각 구성하였고, 이들 분류기와 입력문헌 간 코사인 유사도(cosine similarity) 값이 가장 큰 범주에 할당하는 방식을 취하였다. 또한, 본 연구에서 사용한 Rocchio 분류기는 긍정문헌만을 사용한 경우(Rocchio_긍정)와 긍정문헌과 부정문헌을 함께 사용한 경우(Rocchio_긍정+부정)의 두 가지로 구분하였다(김판준, 2016).

〈표 2〉 분류 범주 통계

번호	범주명	학습문헌 수		검증문헌 수		합계	
		단일	복수	단일	복수	단일	복수
1	도서관/정보센터경영	95	120	42	57	137	177
2	정보서비스	58	79	46	55	104	134
3	계량정보학	38	40	35	40	73	80
4	기록관리/보존	36	38	8	10	44	48
5	편목/메타데이터	37	59	7	14	44	73
6	정보검색	30	40	10	15	40	55
7	전문용어/시소러스	29	35	5	5	34	40
8	디지털도서관	24	35	5	8	29	43
9	정보/도서관정책	16	18	9	10	25	28
10	정보자료/미디어	15	22	7	13	22	35
11	자동분류/클러스터링	17	21	5	8	22	29
12	문헌정보학일반	14	26	6	15	20	41
13	정보교육	11	14	5	5	16	19
14	분류	10	13	4	4	14	17
15	검색모형/기법	11	11	2	2	13	13
16	자동색인/요약	9	9	2	3	11	12
17	데이터베이스	2	7	0	4	2	11
18	도서관사	1	2	0	1	1	3
	합계	453	589	198	269	651	858

둘째, NB 분류기는 Python scikit-learn 라이브러리의 MultinomialNB 모듈을 사용하였다(Pedregosa et al., 2011). 이는 다항분포 나이브 베이즈(Multinomial Naive Bayes)를 적용한 것으로 베이즈 정리를 통한 변환에서 우도(likelihood)에 해당하는 부분을 다항분포에 근거해서 계산하기 때문에 정해진 몇 개의 값 중 하나로 할당하는 상황에 적합하다. 이로 인해 나이브 베이즈 분류를 통한 문헌 분류에서 이 모델이 주로 사용된다. 스무딩(Smoothing) 기법으로는 가짜 수(pseudocount) 1.0을 부여하는 라플라스 스무딩(Laplace smoothing)을 적용하였다.

셋째, SVM 분류기는 Python scikit-learn 라이브러리의 LinearSVC 모듈을 사용하였다(Pedregosa et al., 2011). 또한 OvR(One-vs-the-Rest) 방식으로 분류를 수행하였고, 선형(linear) 모델을 적용하였다. 이는 각 주제 범주에 따라 이진 분류기(binary classifier)를 생성하고, 이들 분류기와 입력문헌 간 산출 값이 가장 큰 범주로 할당하는 방식이다. 이 때 사용된 손실함수는 hinge_squared이고, 최고 반복(max iteration)은 1,000회로 설정하였다.

2.6 성능 평가 척도

연구 가설의 판정을 위한 성능 척도로는 자동분류 연구에서 많이 사용되는 매크로 평균 F1(mac_F1)과 마이크로 평균 F1(mic_F1)을 사용하였다. 마이크로 평균 F1은 모든 문헌에 동일한 가중치를 주어 문헌 당 평균을 산출하는 반면, 매크로 평균 F1은 빈도에 상관없이

모든 범주에 동일한 가중치를 주는 것으로 범주 당 평균을 산출한다. 따라서 단일-범주 분류의 성능 평가에서 실제 환경의 불균형 데이터(imbalanced data)를 대상으로 자동분류의 성능에 영향을 주는 네 가지 주요 성능 요소를 다각적으로 검토하기 위하여, 서로 다른 특성을 가진 마이크로 평균 F1과 매크로 평균 F1을 함께 산출하였다(김판준, 2006a; 김판준, 이재윤, 2012; Hmeidi et al., 2015).

복수-범주의 성능 평가는 기존의 척도(Read, 2010; Tsoumakas, Katakis, & Vlahavas, 2010)에 완화된 기준을 적용하여 조정한 복수-범주 매크로 평균 F1과 복수-범주 마이크로 평균 F1을 사용하였다. 즉, 특정 문헌에 수작업으로 부여된 복수의 분류명 중 어느 것이든 자동분류에 의해 할당이 되면 성공한 것으로 간주하였다. 그 이유는 수작업 분류 과정에서 3명의 전문가가 부여한 개별적으로 분류명에 차이가 있는 경우, 전문가의 전공이나 관점에 따라 우선순위에 차이가 있을 뿐이며 부여된 범주들이 모두 해당 논문에서 다루고 있는 주제와 부합하였기 때문이다. 예를 들면, '공공도서관 교육문화프로그램 참여와 도서관 이용의 관계 연구(이혜윤, 이지연, 2014)'는 3명의 전문가가 부여한 단일-범주가 "도서관/정보센터경영"과 "정보서비스"로 서로 달랐고, 복수-범주 분류에서의 우선순위도 이견이 있었다. 또한, 이강산다정(2015)의 논문 'Lubetzky'의 목록법 사상 연구'의 경우는 단일-범주로 '편목/메타데이터'를 부여하는 것에는 3명이 일치하였으나, 복수-범주로 부여된 '편목/메타데이터'와 '도서관사'의 순서에는 차이가 있었다.

따라서 복수-범주 분류의 경우에는 특정 논문이 다루고 있는 주제에 대한 분류자의 전공이나 관점의 차이로 인한 누락을 방지하기 위하여, 완화된 기준을 적용한 평가 척도(복수-범주 매크로 F1: ml_mac_F1, 복수-범주 마이크로 F1: ml_mic_F1)를 사용하였다. 즉, 특정 논문 i에 자동으로 할당된 하나의 범주 $\hat{y}_i$가 해당 논문 i의 정답 범주집합인 Y_i에 속하는지 여부를 판별하여, 이에 포함되는 경우($\hat{y}_i \in Y_i$)에는 성공한 것으로 판정하였다. 이에 따라 각 범주별로 정확률(TP/(TP+FP)), 재현율(TP/(TP+FP)), F1((2*Precision*Recall)/(Precision+Recall))을 산출하였다. 특히, 자동 할당이 올바르게 된 사례의 개수를 나타내는 TP(True Positive)에 완화된 기준을 적용하였고, 범주 t에 대한 TP(True Positive) 등의 공식은 다음과 같다. 여기서 $\hat{y}_i$는 학술지 논문 i에 대한 범주 자동할당 결과이고, Y_i는 논문 i의 정답 범주집합이다.

$$\text{True Positive}(t) = |\{\hat{y}_i | \hat{y}_i \in Y_i \text{ and } t \in Y_i\}|$$

$$\text{False Positive}(t) = |\{\hat{y}_i | \hat{y}_i \notin Y_i \text{ and } \hat{y}_i = t\}|$$

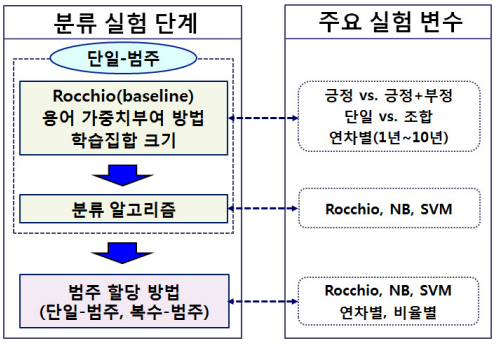
$$\text{False Negative}(t) = |\{\hat{y}_i | \hat{y}_i \notin Y_i \text{ and } t \in Y_i\}|$$

그러나 본 연구의 실험 문헌집단과 같이 저빈도 범주가 많은 소규모의 불균형 문헌집합이 아닌, 고빈도 범주가 대다수인 대규모 문헌집합의 경우에는 보다 엄격한 기준을 적용하는 복수-범주 분류 성능 평가 척도(Exact-Match, Hammig-Loss 등)의 적용을 검토할 필요가 있다(Read, 2010; Tsoumakas, Katakis, & Vlahavas, 2010).

3. 실험 및 결과

3.1 실험 단계

국내 학술지 논문에 대한 자동분류의 성능에 영향을 미치는 네 가지 주요 요소별로 연구 가설의 검증을 위하여 〈그림 2〉와 같이 단계별로 다각적인 분류 실험을 수행하였다.



〈그림 2〉 실험 단계

첫째, 단일-범주 분류 환경에서 Rocchio 분류기를 사용하여 용어 가중치부여 기법과 학습집합 크기에 따른 실험을 수행하였다. 용어 가중치부여 기법은 단일 가중치와 조합 가중치로 구분하여 다양한 가중치를 적용하였고, 학습집합 크기는 최근 1년(2012년)부터 전체 10년(2002년~2012년)까지 연차적으로 학습집합을 증가하는 경우에 성능의 변화를 살펴보았다.

둘째, 단일-범주 분류 환경에서 가장 좋은 성능을 보인 용어 가중치부여 기법을 적용하여 3개 분류기(Rocchio, NB, SVM)의 성능을 살펴보았다.

셋째, 범주 할당 방법에 따라 각 문헌에 단일-범주를 할당하는 경우와 복수-범주를 할당하는 경우로 구분하여 분류 성능을 비교하였다. 또한, 실제 환경에 부합하는 복수-범주 분류 환경에서 이전 실험의 주요 변수들을 다각적으로 검토하는 실험(3개 분류기: Rocchio, NB, SVM, 학습집합 크기: 연차별, 비율별)을 수행하였다.

3.2 단일-범주 분류 방법

3.2.1 용어 가중치부여 기법과 학습집합 크기

단일-범주 환경에서 Rocchio 분류기를 사용하여 용어 가중치부여 기법과 학습집합의 크기에 따른 성능을 살펴보았다. 첫째, Rocchio 분류기는 긍정문헌만을 사용한 경우(Rocchio_긍정)와 긍정문헌과 부정문헌을 함께 사용한 경우(Rocchio_긍정+부정)로 구분하였다. 둘째, 용어 가중치부여 기법은 크게 단일 가중치와 조합 가중치로 구분하여 수행하였다. 먼저, 단일 가중치부여 기법은 총 9개로 문헌 및 문헌집합 내 출현정보에 기초한 용어빈도(tf, ltf/log_tf, otf/okapi_tf)와 역문헌빈도(idf), 문헌에 부여된 범주정보에 기초한 공식(ig/정보획득량, jac/자카드 상관계수, chi/카이제곱통계량, lor/로그승비, mi/상호정보량)을 사용하였다. 다음으로, 조합 가중치부여 기법은 단일 가중치를 적용한 실험에서 좋은 성능을 보인 단일 가중치 2개 또는 3개를 조합하여 사용하였다. 셋째, 전체 실험에서 학습집합의 크기를 최근 1년부터 10년(2015년~2002년)까지 1년씩 연차별로 추가하여 적용하였다. 넷째, 분류 성능을 서로 다른 특성을 가진 매크로 평균 F1(mac_F1)과 마이크로 평균 F1(mic_F1)으로 구분하여 제시하였다(김판준, 2016).

〈표 3〉은 Rocchio 분류기를 긍정문헌만을

〈표 3〉 단일 가중치부여 방법의 연차별 성능: Rocchio, mac_F1

구분	가중치	1년	2년	3년	4년	5년	6년	7년	8년	9년	10년
Rocchio _긍정	tf	0.2832	0.377	0.4795	0.4517	**0.5231**	0.4955	0.4793	0.4895	0.4954	0.5136
	ltf	0.4146	0.4843	0.5194	0.5948	0.6204	**0.6401**	0.6221	0.6157	0.5679	0.6252
	otf	0.3414	0.3996	0.5559	0.5525	0.5681	0.5669	0.5857	0.5712	0.588	**0.5965**
	ig	0.3158	0.4198	0.4214	0.4546	0.4112	0.4619	**0.4653**	0.427	0.4186	0.4259
	jac	0.2157	0.351	0.3157	0.3046	0.3935	**0.3979**	0.3834	0.3837	0.3846	0.3911
	chi	0.3507	0.3176	0.3859	0.4115	0.4364	**0.4376**	0.4329	0.3807	0.3943	0.3923
	lor	0.1915	0.3245	0.3843	0.475	**0.5282**	0.5093	0.432	0.4425	0.4353	0.4362
	mi	0.2442	0.3282	0.3862	0.4666	0.4897	**0.5135**	0.5096	0.4889	0.4445	0.4801
	idf	0.3891	0.4658	0.5702	0.6204	0.6558	**0.6654**	0.6131	0.6043	0.6596	0.6542
Rocchio _긍정+부정	tf	0.273	0.3882	0.4748	0.4475	**0.5139**	0.4869	0.4661	0.4929	0.477	0.5045
	ltf	0.3818	0.5194	0.5293	0.5664	0.5878	0.6206	0.6058	0.605	**0.6224**	0.5762
	otf	0.3161	0.4276	0.5413	0.488	0.5454	**0.5606**	0.5547	0.536	0.5565	0.5588
	ig	0.3004	0.3746	0.4542	0.4184	**0.4585**	0.4492	0.4574	0.4402	0.4292	0.4239
	jac	0.1774	0.2827	0.2963	0.3187	0.3539	0.3238	0.3405	0.3632	**0.3653**	0.3547
	chi	0.3602	0.3161	0.3511	0.3712	**0.4364**	0.3524	0.3658	0.3642	0.3546	0.3513
	lor	0.1653	0.3223	0.4163	0.4311	**0.5155**	0.4973	0.4617	0.4243	0.4385	0.436
	mi	0.237	0.3291	0.3863	0.472	0.4873	0.5073	**0.512**	0.5083	0.4081	0.4854
	idf	0.3311	0.485	0.597	0.5873	**0.6362**	0.6202	0.6234	0.6249	0.6274	0.6274

사용한 경우와 긍정문헌과 부정문헌을 함께 사용한 경우로 구분한 다음, 단일 가중치부여 기법 9개를 적용하고 학습집합의 크기를 최근 1년부터 10년까지 증가시킨 실험의 결과를 매크로 평균 F1(mac_F1)으로 산출한 것이다. 여기서 가장 좋은 성능은 긍정문헌만을 사용하고 6년의 학습집합과 단일 가중치(idf)를 적용한 경우(0.6654/Rocchio_긍정, 6년, idf)이며, 그 다음은 긍정문헌을 사용하고 6년의 학습집합과 단일 가중치(ltf)를 사용한 것(0.6401/Rocchio_긍정, 6년, ltf)이다. 대체로 단일 가중치부여 기법 중에서 출현정보에 기초한 기법들이 범주정보에 기초한 기법들보다 높은 성능을 보였다. 또한 동일한 단일 가중치부여 기법을 적용하였을 때, 긍정문헌만을 사용한 경우가 긍정문헌과 부정문헌을 함께 사용한 것보다 나은 성능이었다.

〈표 4〉는 Rocchio 분류기에서 단일 가중치 9개의 성능을 마이크로 평균 F1(mic_F1)으로 산출한 결과이다. 여기서 최고 성능은 긍정문헌만을 사용하고 전체 10년의 학습집합과 단일 가중치(ltf)를 적용한 경우(0.6734/Rocchio_긍정, 10년, ltf)이며, 그 다음은 긍정문헌과 부정문헌을 함께 사용하고 9년의 학습집합과 단일 가중치(ltf)를 사용한 것(0.6684/Rocchio_긍정+부정, 9년, ltf)이다. 〈표 3〉과 마찬가지로 단일 가중치부여 기법 중에서 출현정보에 기초한 기법들이 범주정보에 기초한 기법들보다 대체로 좋은 성능을 보였다.

Rocchio 분류기를 사용한 단일 가중치 실험에서 좋은 성능을 보인 단일 가중치들을 조합한 것으로, 여러 조합 가중치부여 기법의 성능을 매크로 평균 F1(mac_F1)으로 산출한 결과는 〈표 5〉이다. 여기서 가장 좋은 성능은 긍정문헌만을 사용하고 8년의 학습집합과 단순한 조합 가중치를 적용한 경우(0.6944/Rocchio_긍정, 8년, ltfidf)이며, 그 다음은 긍정문헌과 부정문

〈표 4〉 단일 가중치부여 방법의 연차별 성능: Rocchio, mic_F1

구분	가중치	1년	2년	3년	4년	5년	6년	7년	8년	9년	10년
Rocchio _긍정	tf	0.4697	0.4646	0.5316	0.5266	**0.5772**	0.5367	0.5215	0.4962	0.5077	0.5231
	ltf	0.5455	0.6162	0.6162	0.6667	0.6667	0.6667	0.6616	0.6566	0.638	**0.6734**
	otf	0.5152	0.5455	0.6111	0.6111	0.6162	0.6061	0.6162	0.6111	0.6138	**0.626**
	ig	0.4848	0.5606	0.5354	0.5505	0.5165	0.5354	**0.5657**	0.5202	0.5114	0.5114
	jac	0.4091	0.4848	0.4658	0.4619	0.5228	0.533	0.5381	**0.5431**	0.5394	0.5293
	chi	0.4343	0.3889	0.4112	0.4405	**0.491**	0.4859	0.4388	0.4416	0.4619	0.4529
	lor	0.2727	0.399	0.4694	0.4721	**0.5482**	0.533	0.4861	0.4848	0.4772	0.467
	mi	0.298	0.399	0.4377	0.5025	0.5063	0.5135	0.5505	0.5152	0.5367	**0.5657**
	idf	0.5253	0.5808	0.6162	0.6111	**0.6667**	**0.6667**	0.6414	0.6616	**0.6667**	0.6633
Rocchio _긍정+부정	tf	0.4646	0.4747	0.5228	0.5367	**0.562**	0.5367	0.5381	0.5114	0.5181	0.5426
	ltf	0.5152	0.6162	0.6263	0.6313	0.6414	0.6515	0.6515	0.6515	**0.6684**	0.6345
	otf	0.4798	0.5455	0.6025	0.5873	0.5975	0.5924	0.5924	0.6025	**0.6134**	0.6015
	ig	0.3687	0.4293	0.4658	0.4557	**0.481**	0.4456	0.4444	0.4405	0.4061	0.4061
	jac	0.3586	0.4798	0.4619	0.4784	0.5178	0.5025	0.5191	**0.5394**	0.5102	0.5038
	chi	0.3939	0.3586	0.3687	0.3838	0.491	0.4091	0.3848	0.4	0.3858	0.3756
	lor	0.2677	0.3687	0.4473	0.491	0.491	0.5025	0.5127	0.5076	0.4847	**0.5127**
	mi	0.2424	0.3788	0.473	0.4377	**0.5293**	0.5102	0.4721	0.4608	0.473	0.4643
	idf	0.399	0.5303	0.5939	0.5736	0.6278	0.6177	0.6329	**0.6481**	0.6447	0.6447

〈표 5〉 조합 가중치부여 방법의 연차별 성능: Rocchio, mac_F1

구분	가중치	1년	2년	3년	4년	5년	6년	7년	8년	9년	10년
Rocchio _긍정	ltfidf	0.3666	0.4936	0.6449	0.6348	0.6578	0.6502	0.6681	**0.6944**	0.6722	0.6867
	ltfig	0.3224	0.4185	**0.5208**	0.5193	0.4938	0.5007	0.4952	0.5074	0.5074	0.5029
	ltfjac	0.2669	0.3322	0.4183	0.42	**0.5049**	0.4739	0.4567	0.4536	0.4685	0.4671
	ltfchi	0.3808	0.3768	0.4905	0.4327	0.4591	0.4348	0.4601	0.5402	0.5424	**0.5617**
	ltflor	0.1823	0.2886	0.3678	0.4394	0.4659	**0.4987**	0.4095	0.4334	0.4311	0.4219
	ltfmi	0.185	0.3165	0.396	0.4535	0.4763	**0.4951**	0.4345	0.4923	0.4256	0.424
	idfig	0.3883	0.4436	0.5867	**0.5883**	0.5605	0.5648	0.5457	0.5517	0.5519	0.5545
	idfjac	0.313	0.3603	0.4032	0.4756	0.5635	0.5281	0.5355	0.5334	**0.5757**	0.5739
	idfchi	0.44	0.4146	0.5162	0.5195	**0.5275**	0.5224	0.4953	0.4836	0.4801	0.5044
	idflor	0.192	0.3704	0.4251	0.3836	0.4433	0.4681	0.4287	0.4853	**0.4898**	0.4891
	idfmi	0.2631	0.3711	0.4225	0.4149	0.4841	0.5256	0.5456	**0.5893**	0.566	0.5692
	ltfidfig	0.3259	0.4736	0.5869	0.6018	0.5549	0.6031	**0.6155**	0.5983	0.6133	0.598
	ltfidfjac	0.3521	0.3565	0.3849	0.4626	0.5711	**0.5901**	0.5695	0.5613	0.5624	0.5413
	ltfidfchi	0.4601	0.4134	0.5121	0.513	0.5546	0.5453	**0.5655**	0.5514	0.5444	0.5545
	ltfidflor	0.1752	0.3322	0.4165	0.4679	0.5047	**0.5469**	0.4432	0.4975	0.4605	0.4547
	ltfidfmi	0.2275	0.3076	0.3985	0.4617	0.5179	**0.5756**	0.5115	0.5263	0.5329	0.5246
Rocchio _긍정+부정	ltfidf	0.3741	0.4824	0.6286	0.6083	0.6512	0.6349	0.6606	0.6671	0.6602	**0.673**
	ltfig	0.3324	0.4238	**0.5296**	0.4557	0.4844	0.4727	0.4527	0.4319	0.4243	0.4167
	ltfjac	0.274	0.2845	0.2835	0.3292	0.3887	0.368	0.3922	**0.4485**	0.4431	0.4385
	ltfchi	0.353	0.3436	0.3988	0.41	0.4332	0.4415	0.4288	**0.4918**	0.4389	0.4483
	ltflor	0.1919	0.3043	0.3802	0.4652	0.4609	0.483	0.4068	**0.4871**	0.4141	0.4167
	ltfmi	0.1821	0.2998	0.3897	0.4454	0.4689	0.4817	0.4105	**0.5297**	0.4675	0.4123
	idfig	0.4112	0.5209	0.6057	0.5843	**0.6116**	0.5785	0.5395	0.5451	0.528	0.5148
	idfjac	0.2782	0.3939	0.3912	0.4281	0.4649	0.5	0.4576	**0.5185**	0.491	0.4624
	idfchi	0.4142	0.4358	0.5234	0.518	0.5304	0.5322	0.5238	0.5431	0.524	**0.5502**
	idflor	0.1901	0.3645	0.4366	0.3933	0.4448	0.4735	0.4495	**0.4939**	0.487	0.4921
	idfmi	0.2601	0.3503	0.4246	0.4311	0.4774	0.516	0.5372	0.5743	0.5759	**0.5886**
	ltfidfig	0.394	0.5066	0.5588	**0.5848**	0.5755	0.5726	0.5591	0.5473	0.5292	0.5406
	ltfidfjac	0.3607	0.412	0.4045	0.4471	0.6015	0.5917	0.5389	**0.6115**	0.5991	0.5815
	ltfidfchi	0.4469	0.4075	0.5114	0.5015	0.5351	0.542	0.5446	0.5512	0.5395	**0.5628**
	ltfidflor	0.1674	0.3415	0.4162	0.4613	0.5131	**0.5443**	0.413	0.4897	0.4532	0.4644
	ltfidfmi	0.2121	0.3016	0.3923	0.4762	0.5114	**0.5499**	0.4986	0.5008	0.5066	0.4968

헌을 함께 사용하고 전체 10년의 학습집합과 조합 가중치를 적용한 것(0.673/Rocchio_긍정+부정, 10년, ltfidf)이다. 결과적으로 출현정보에 기초한 2개의 단일 가중치를 사용한 단순한 조합 가중치를 적용한 것이 다른 복잡한 조합 가중치보다 훨씬 좋은 성능이었다. 또한, 전반적으로 학습집합의 규모가 커질수록 성능이 향상되는 경향을 보이지만 전체 학습집합(10년)을 사용한 것이 항상 최고 성능은 아니며, 범주정보에 기초한 단일 가중치를 조합한 경우(ltfig, idfig, ltfidfig)에는 오히려 학습집합의 규모가 작은 것이 더 좋은 성능이었다.

한편, Rocchio 분류기를 사용한 여러 조합 가중치부여 기법의 성능을 마이크로 평균 F1(mic_F1)으로 산출한 결과는 〈표 6〉이다. 여기서 최고 성능은 긍정문헌만을 사용하고 8년의 학습집합과 조합 가중치를 적용한 경우(0.7139/Rocchio_긍정, 8년, ltfidf)이며, 그 다음은 긍정문헌과 부정문헌을 함께 사용하고 8년의 학습집합과 조합 가중치를 적용한 것(0.6853/Rocchio_긍

〈표 6〉 조합 가중치부여 방법의 연차별 성능: Rocchio, mic_F1

구분	가중치	1년	2년	3년	4년	5년	6년	7년	8년	9년	10년
Rocchio _긍정	ltfidf	0.5101	0.5606	0.6532	0.6616	0.6684	0.6734	0.6886	**0.7139**	0.6987	0.7056
	ltfig	0.5202	0.5707	0.5808	0.596	0.601	0.5909	0.5758	0.5859	**0.6091**	0.6091
	ltfjac	0.4192	0.4949	0.5063	0.5013	0.5367	0.5418	0.5215	**0.5418**	0.5381	0.532
	ltfchi	0.4192	0.404	0.4365	0.4354	0.4822	0.467	0.4518	**0.5013**	0.4835	0.4987
	ltflor	0.2626	0.3182	0.4194	0.4365	0.5051	**0.5253**	0.5	0.4899	0.4897	0.4847
	ltfmi	0.202	0.3586	0.4275	0.4518	0.4495	0.5505	0.5303	**0.5758**	0.5685	0.5671
	idfig	0.4949	0.5859	0.6061	**0.6313**	0.6162	0.6162	0.6162	0.596	0.5859	0.5859
	idfjac	0.4646	0.5	0.5253	0.5505	**0.6061**	**0.6061**	0.596	0.601	0.596	0.596
	idfchi	0.5556	0.5051	0.5152	0.5404	**0.5808**	0.5606	0.5	0.4949	0.4949	0.5152
	idflor	0.2879	0.4343	0.4733	0.4315	0.4861	0.5	0.4798	**0.5152**	0.5114	0.5101
	idfmi	0.3535	0.4394	0.4564	0.4784	0.5191	0.5127	0.5431	**0.5888**	0.5575	0.5663
	ltfidfig	0.4949	0.5909	0.6414	**0.6667**	0.6313	0.6515	0.6566	0.6313	0.6481	0.638
	ltfidfjac	0.5051	0.5303	0.5367	0.557	0.5975	**0.6127**	0.5823	0.5975	0.599	0.5751
	ltfidfchi	0.5556	0.4747	0.4697	0.4949	**0.5606**	0.5505	0.5505	0.5303	0.5316	0.5316
	ltfidflor	0.2626	0.3788	0.4619	0.4697	0.5101	**0.5354**	0.4848	**0.5354**	0.5038	0.5114
	ltfidfmi	0.298	0.3384	0.4071	0.4478	0.4975	0.5635	0.5584	0.5635	0.5678	0.5649
Rocchio _긍정+부정	ltfidf	0.4697	0.5303	0.6209	0.5939	0.626	0.6244	0.6599	**0.6853**	0.6701	0.6819
	ltfig	0.4848	0.4545	0.5152	0.4899	**0.5303**	0.4646	0.4242	0.4444	0.4365	0.401
	ltfjac	0.404	0.5	0.4949	0.5114	0.5404	0.5505	0.5556	**0.5657**	0.5533	0.5445
	ltfchi	0.4141	0.3889	0.4192	0.4091	0.4343	0.4192	0.3889	**0.4343**	0.401	0.4112
	ltflor	0.2677	0.3232	0.4246	0.4388	0.4911	0.5076	0.4899	**0.5404**	0.4607	0.4757
	ltfmi	0.202	0.3182	0.4184	0.4213	0.4173	0.5202	0.5303	**0.5758**	0.5496	0.5367
	idfig	0.4798	0.5455	0.5859	0.5859	**0.6061**	0.5707	0.5303	0.5455	0.5215	0.5127
	idfjac	0.4091	0.5051	0.5316	0.5556	0.5671	0.5909	0.601	**0.6212**	0.6025	0.596
	idfchi	0.4949	0.5	0.5505	0.5202	0.5455	0.5152	0.4848	0.5556	0.5455	**0.5606**
	idflor	0.2374	0.399	0.4552	0.4082	0.4733	0.4745	0.4604	**0.5051**	0.5	0.5
	idfmi	0.3384	0.404	0.4576	0.4715	0.5052	0.4987	0.5217	0.5575	0.5389	**0.5633**
	ltfidfig	0.5354	0.5606	0.601	**0.6061**	0.5808	0.5808	0.5152	0.5303	0.5127	0.4975
	ltfidfjac	0.5	0.5556	0.5671	0.5975	0.6294	0.6177	0.6177	**0.638**	0.6294	0.6056
	ltfidfchi	**0.5556**	0.4545	0.4697	0.4444	0.4798	0.4596	0.4646	0.481	0.4873	0.481
	ltfidflor	0.2374	0.3586	0.4552	0.4365	0.5114	**0.5165**	0.4596	0.5152	0.4767	0.5
	ltfidfmi	0.2677	0.3283	0.3939	0.4439	0.4847	0.5344	0.5242	**0.5459**	0.5371	0.5385

정+부정, 8년, ltfidf)이다. 이는 전체적으로 매크로 평균 F1으로 산출한 〈표 5〉의 결과와 유사한 패턴이라고 할 수 있다. 이러한 결과를 종합하면, 용어 가중치부여 방법은 국내 학술지 논문의 자동분류 성능에 영향을 미친다(가설 1). 단일 가중치보다는 조합 가중치의 성능이 더 좋고, 특히 조합 가중치 중에서는 출현정보에 기초한 단순한 조합 가중치부여 방법(ltfidf)이 가장 좋은 것으로 나타났다. 또한, 학습집합의 크기도 국내 학술지 논문의 자동분류 성능에 상당한 영향을 미치는 것으로 나타났다(가설 2).

3.2.2 분류 알고리즘

Rocchio 분류기 결과와 비교하기 위하여 최근까지 문헌의 자동분류에 많이 사용되고 있는 NB, SVM 2개의 분류기를 사용한 실험을 수행하였다. 또한 이전 실험에서 가장 좋은 성능을 보인 요소로서 단순한 조합 가중치(ltfidf)와 학습집합 크기(최근 1년~10년)를 이들 분류기에 동일하게 적용하였다. 〈표 7〉은 3개 분류기에 전체 10년의 학습집합을 최근 1년부터 연차별로 적용하여 매크로 평균 F1(mac_F1)으로 산출한 결과이다. 여기서 Rocchio 분류기는 이전 실험에서 성능이 더 좋았던 긍정문헌만을 사용한 것이다. 3개 분류기 중에서 Rocchio (0.6944/Rocchio_긍정, 8년)가 최고 성능이었으며, 특히 다른 2개 분류기가 학습집합 10년(0.6310/NB, 0.6656/SVM)을 모두 사용한 것이 최고 성능인데 반해, Rocchio는 이보다 적은 8년의 학습집합을 사용한 경우에 가장 좋은 성능을 보였다.

3개 분류기의 연차별 성능을 마이크로 평균 F1(mic_F1)으로 산출한 결과는 〈표 8〉이다. 〈표 7〉과 마찬가지로 3개 분류기 중에서 Rocchio 분류기(0.7139/Rocchio_긍정, 8년)의 성능이 가장 좋았다. 그러나 〈표 8〉에서는 3개 분류기 모두 전체 학습집합 10년을 사용한 것보다 8년(0.7139/Rocchio_긍정, 0.6294/NB) 또는 9년(0.7071/SVM)의 학습집합이 더 좋은 성능을 보였다. 이러한 결과를 종합하면, 분류 알고리즘은 국내 학술지 논문의 자동분류 성능에 영향을 미친다(가설 3). 특히, 국내 학술지 논문의 자동분류에서는 텍스트 분류에서 가장 좋은 성능을 보이는 것으로 알려진 SVM보다 단순하고 빠른 Rocchio 분류기의 성능이 더 우수한 것으로 나타났다.

〈표 7〉 단일-범주 분류, 3개 분류기의 연차별 성능: mac_F1

구분	1년	2년	3년	4년	5년	6년	7년	8년	9년	10년
Rocchio_긍정	0.3666	0.4936	0.6449	0.6348	0.6578	0.6502	0.6681	**0.6944**	0.6722	0.6867
NB	0.3384	0.3848	0.5179	0.5174	0.5828	0.5812	0.5943	0.6233	0.6161	**0.6310**
SVM	0.3646	0.3504	0.5981	0.5515	0.6451	0.6539	0.6000	0.6263	0.6635	**0.6656**

〈표 8〉 단일-범주 분류, 3개 분류기의 연차별 성능: mic_F1

구분	1년	2년	3년	4년	5년	6년	7년	8년	9년	10년
Rocchio_긍정	0.5101	0.5606	0.6532	0.6616	0.6684	0.6734	0.6886	**0.7139**	0.6987	0.7056
NB	0.4848	0.5303	0.5671	0.5707	0.5960	0.5873	0.6228	**0.6294**	0.6142	0.6193
SVM	0.5202	0.5202	0.6263	0.6263	0.6515	0.6616	0.6768	0.6919	**0.7071**	0.7038

3.3 복수-범주 분류 방법

3.3.1 단일-범주 분류와 복수-범주 분류

단일-범주 분류 방법에서 최고 성능을 나타낸 요소들의 특성에 기초하여 단일-범주 분류와 복수-범주 분류 간에 성능 차이가 있는 지를 검토하였다. 그 이유는 최근 14년간 「정보관리학회지」에 수록된 논문들에 대한 수작업 분류의 결과, 단일 범주가 부여된 논문과 복수 범주가 부여된 논문의 비중이 거의 동일하였기 때문이다(〈표 9〉 참조). 이에 따라 특정 논문에 하나의 분류명만을 부여하는 단일-범주 분류와 함께 하나 이상의 분류명을 부여하는 복수-범주 분류를 검토하였다. 또한, 수작업 분류에 참여한 3명의 전문가 간에도 특정 논문의 대표 분류명(단일-범주)과 주제가 다루어진 비중에 따른 분류명 부여의 순서(복수-범주)에 대한 이견이 다수 있었기 때문에, 하나의 논문에 복수의 분류명(복수-범주)을 부여하는 것이 보다 실제적이라 할 것이다.

〈표 9〉 단일-범주와 복수-범주 분류의 수작업 부여 문헌 수

구분	단일-범주	복수-범주	합계
문헌 수	320	331	651
비율	49.2	50.8	100

〈표 10〉은 이전 실험에서 가장 좋은 성능을 보인 Rocchio 분류기(Rocchio_긍정)에 대하여 단일-범주 분류와 복수-범주 분류의 성능을 매크로 평균 F1(mac_F1, ml_mac_F1)으로 산출한 것이다. 여기서 단일범주 분류와 복수범주 분류 모두 학습집합의 크기를 연차적으로 증가할수록 성능이 향상되는 경향을 보였고, 두 가지 방법에서 동일하게 8년의 학습집합을 사용한 경우가 최고 성능이었다. 또한, 〈표 11〉은 긍정문헌만을 학습집합으로 사용한 로치오 분류기(Rocchio_긍정, ltfidf)에 대한 단일-범주 분류와 복수-범주 분류의 성능을 마이크로 평균 F1(mic_F1, ml_mic_F1)으로 산출한 것이다. 〈표 10〉과 마찬가지로 두 가지 범주 부여 방법 모두 학습집합의 크기를 연차적으로 증가하면 성능이 향상되는 경향이 보이는 가운데, 8년의 학습집합을 사용한 경우가 최고 성능이었다. 또한, 매크로와 마이크로 평균 F1 양자에서 모두 단일-범주 분류보다 복수-범주 분류의 성능이 더 우수한 것으로 나타났다. 이를 통해 Rocchio 분류기를 사용하였을 때, 범주 부여 방법(단일-범주 분류, 복수-범주 분류)에 따른 국내 학술지 논문의 자동분류 성능에 뚜렷한 차이가 있다는 것을 확인하였다(가설 4).

〈표 10〉 단일-범주와 복수-범주 분류의 성능 비교(Rocchio_긍정): mac_F1, ml_mac_F1

구분	1년	2년	3년	4년	5년	6년	7년	8년	9년	10년
단일-범주	0.3666	0.4936	0.6449	0.6348	0.6578	0.6502	0.6681	**0.6944**	0.6722	0.6867
복수-범주	0.6444	0.6805	0.7632	0.7738	0.7820	0.7752	0.7810	**0.8224**	0.8070	0.8016

〈표 11〉 단일-범주와 복수-범주 분류의 성능 비교(Rocchio_긍정): mic_F1, ml_mic_F1

구분	1년	2년	3년	4년	5년	6년	7년	8년	9년	10년
단일-범주	0.5101	0.5606	0.6532	0.6616	0.6684	0.6734	0.6886	**0.7139**	0.6987	0.7056
복수-범주	0.6731	0.7063	0.7731	0.7908	0.7908	0.7885	0.7931	**0.8008**	0.7916	0.7854

3.3.2 분류 알고리즘과 학습집합 크기: 연차별

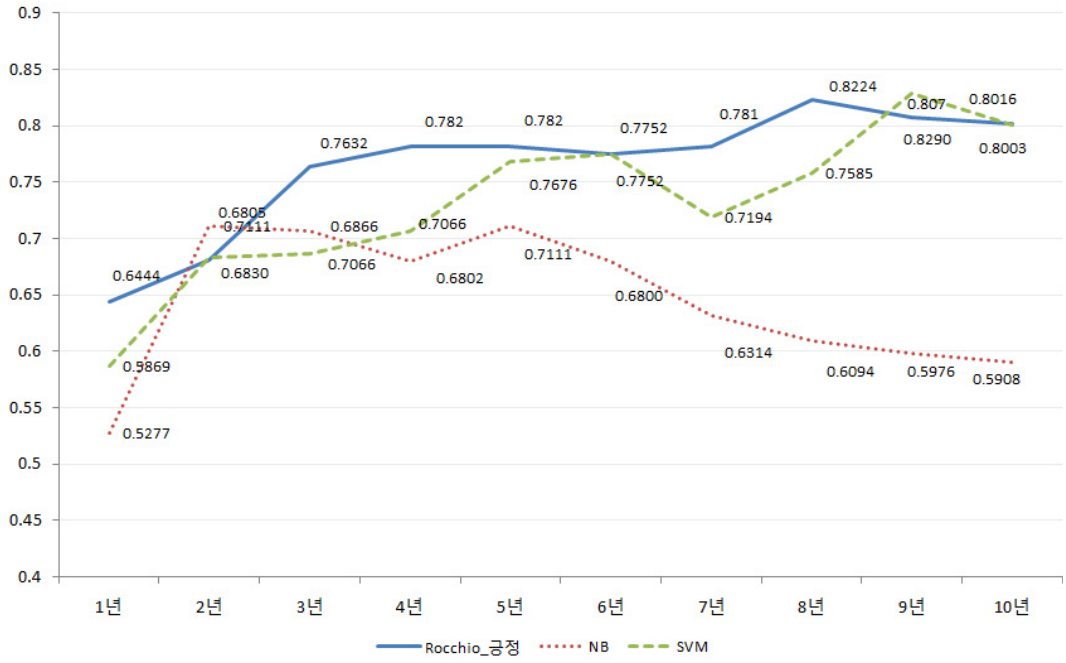
복수-범주 분류 환경에서 이전 실험에서 좋은 성능을 보인 조합가중치(ltfidf)를 적용하여, 3개 분류기(Rocchio_긍정, NB, SVM)의 연차별 성능을 비교하였다. 〈그림 3〉은 3개 분류기의 성능을 복수-범주 매크로 평균 F1(ml_mac_F1)으로 나타낸 것이다. 여기서 최고 성능을 보인 분류기는 SVM(0.8290, 5년)이었지만 Rocchio(0.8224, 8년)와 성능 차이가 거의 없었다. 전반적으로는 Rocchio의 성능이 가장 좋았고 SVM이 그 다음이며, NB가 상대적으로 가장 낮은 성능을 보였다. 또한, Rocchio와 SVM이 학습집합의 크기가 증가함에 따라 성능이 향상되는 경향을 보이는 반면, NB는 학습집합의 증가(5년 이상)에 따라 오히려 성능이 지속적으로 하락하였다. 특히, Rocchio는 3년 이상 소규모의 학습집합을 사용한 이후에는 안정적으로 좋은 분류 성능을 보였다. 또한, 〈그림 4〉는 복수-범주 마이크로 평균 F1(ml_mic_F1)으로 산출한 3개 분류기의 성능이다. 여기서 최고 성능을 보인 분류기는 Rocchio(0.8008, 8년)이었지만 SVM(0.7924, 7년)와 성능 차이가 거의 없었다. 〈그림 3〉과 마찬가지로 전반적으로 Rocchio의 성능이 가장 좋았고, SVM이 그 다음이며, NB가 가장 낮은 성능을 보였다. 특히, NB는 3년 이상의 학습집합을 사용하는 경우에 성능 하락을 보이는 반면, Rocchio는 3년 이상, SVM은 6년 이상의 학습집합을 사용한 경우에 안정적으로 좋은 성능을 보여 서로 상반된 경향을 나타냈다. 이러한 결과는 선행연구에서 Rocchio의 성능이 SVM과 거의 동등한 수준인 반면, NB의 성능이 상대적으로 낮은 것과 일치한다(김판준, 이재윤, 2007; 김판준, 2006b). 여기서 두 가지 성능 척도 모두에서 Rocchio가 SVM보다 소규모의 학습집합을 사용하는 경우에도 안정적으로 좋은 성능을 보이는 점에 주목할 필요가 있다.

3.3.3 분류 알고리즘과 학습집합 크기: 비율별

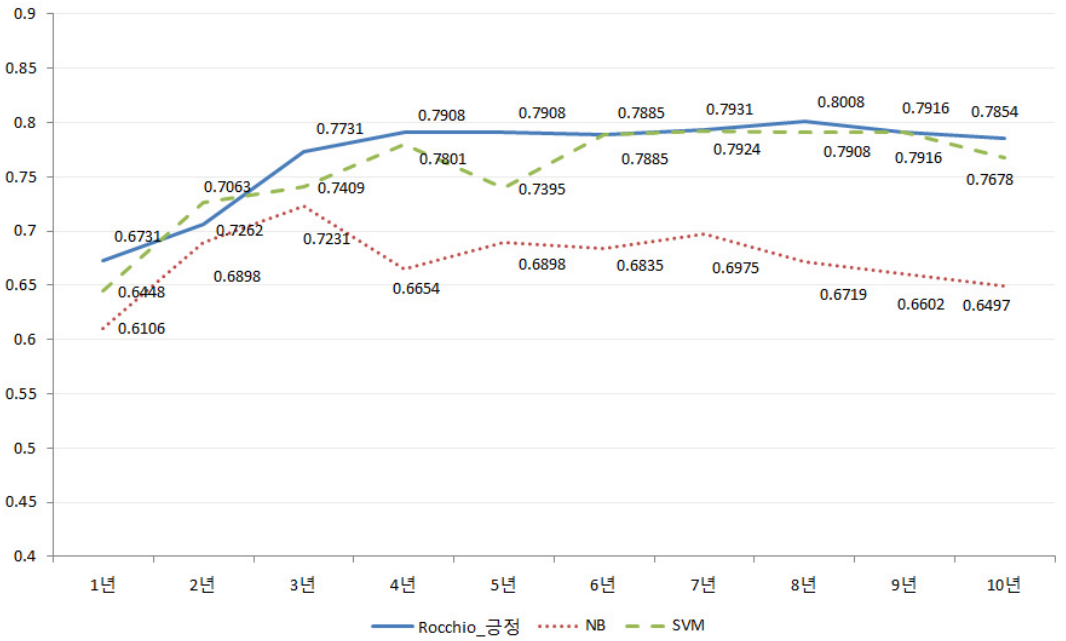
복수-범주 분류 환경에서 학습집합의 구성 방법을 달리하여 3개 분류기(Rocchio_긍정, NB, SVM)의 성능을 비교해 보았다. 즉, 자동분류 관련 연구에서 주로 사용되어 온 학습집합 구성 방법으로 전체 학습집합을 비율별로 랜덤하게 구분하여 10%부터 100%까지 적용하였다.

〈그림 5〉는 이전 실험의 주요 요소를 동일하게 적용한 3개 분류기의 비율별 성능을 복수-범주 매크로 평균 F1(ml_mac_F1)으로 나타낸 것이다. 여기서 최고 성능의 분류기는 Rocchio(0.8018, 70%)로서, 그 다음인 SVM(0.8003, 100%)과 성능 차이가 크지 않았다. 전반적으로는 Rocchio의 성능이 가장 좋았고 SVM이 그 다음이며, NB가 상대적으로 가장 낮은 성능 수준을 보였다. 또한, Rocchio와 SVM은 학습집합의 크기가 증가함에 따라 성능이 향상되는 경향을 보였지만, NB는 학습집합의 증가에 따른 성능 향상이 거의 없었다.

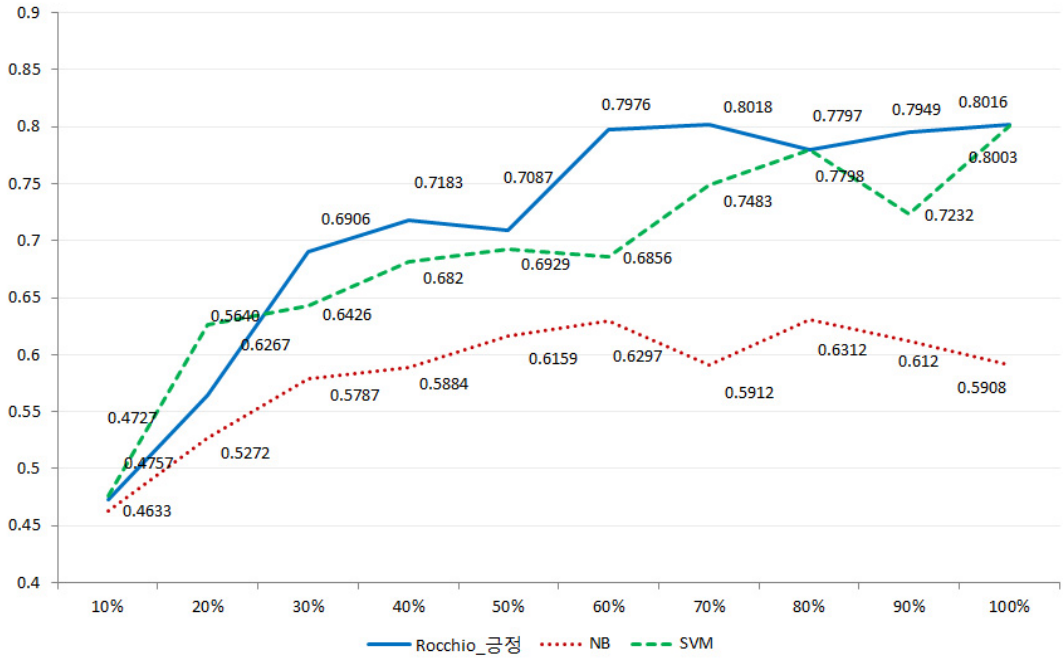
또한, 〈그림 6〉은 복수-범주 마이크로 평균 F1(ml_mic_F1)으로 산출한 3개 분류기의 비율별 성능이다. 〈그림 5〉와 마찬가지로 최고 성능인 Rocchio(0.8018, 70%)와 그 다음의 SVM(0.8003, 100%) 간에 성능 차이가 거의 없었다. 또한 Rocchio와 SVM의 성능이 좋은 반면, NB는 상대적으로 낮은 성능을 보였다. Rocchio와 SVM은 30% 이상의 학습집합을 사용하면 지속적으로 성능이 향상되는 경향을 보인 반면, NB는



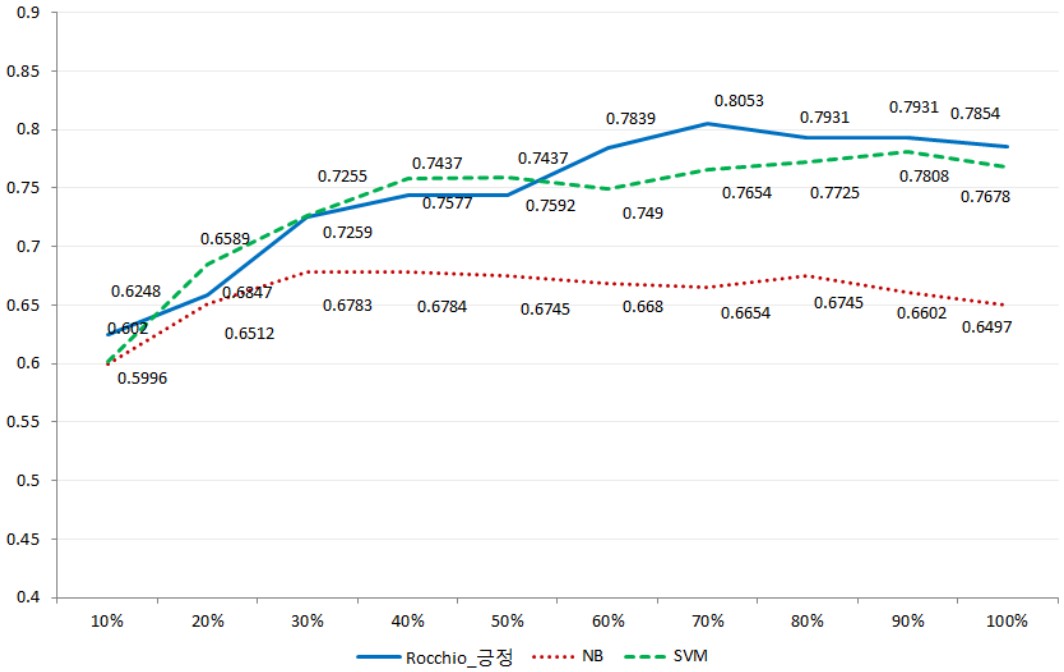
〈그림 3〉 복수-범주 분류, 3개 분류기의 연차별 성능: ml_mac_F1



〈그림 4〉 복수-범주 분류, 3개 분류기의 연차별 성능: ml_mic_F1



〈그림 5〉 복수-범주 분류, 3개 분류기의 비율별 성능: ml_mac_F1



〈그림 6〉 복수-범주 분류, 3개 분류기의 비율별 성능: ml_mic_F1

오히려 40% 이상의 학습집합을 사용하는 경우에 성능이 하락하는 경향을 보여주었다. 특히, Rocchio는 60% 이상의 학습집합이 사용되는 경우에 지속적으로 SVM보다 나은 성능을 보였다.

4. 분석 및 논의

국내 학술지 논문의 자동분류를 위한 최적의 분류 모형을 도출하기 위하여 기계학습에 기초한 자동분류의 성능에 영향을 미치는 주요 요소들에 대한 가설을 설정하고 실험을 수행하였다. 먼저, 단일-범주 분류 환경에서 분류 성능에 영향을 미치는 주요 요소인 용어 가중치부여 기법, 학습집합의 크기, 분류 알고리즘에 대한 가설을 검증하는 실험을 수행하였다. 다음으로, 이전 실험에서 최고의 성능을 보인 요소들을 적용하여 단일-범주 분류와 복수-범주 분류의 성능을 비교한 다음, 복수-범주 분류 환경에서 주요 요소에 따른 성능을 다각적으로 검토하였다. 최종적으로, 이러한 실험 결과와 가설 검증 과정에서 최고 성능을 보인 요소를 종합적으로 분석하여 최적의 분류 모형을 도출하였다.

첫째, 용어 가중치부여 방법은 국내 학술지 논문의 자동분류 성능에 영향을 미친다(가설 1). 단일-범주 분류 환경에서 Rocchio 분류기에 기초하여 다양한 가중치부여 방법을 적용한 분류 실험을 수행한 결과, 단순한 조합 가중치(ltfidf)가 가장 성능이 더 좋은 것으로 나타났다. 특히 Rocchio 분류기에서 긍정문헌만 사용한 경우(Rocchio_긍정)와 긍정문헌과 부정문헌을 사용한 경우(Rocchio_긍정+부정) 양자에서 용어의 출현정보에 기초한 조합 가중치의 성능이 가장 우수하였다. 이는 자동분류 연구에서 일반적으로 사용되어 온 조합가중치 부여 방법(ltfidf)의 성능을 확인한 결과라고 할 수 있다.

둘째, 학습집합의 크기는 국내 학술지 논문의 자동분류 성능에 영향을 미친다(가설 2). 단일-범주 분류 환경에서 Rocchio 분류기를 사용하여 학습집합의 크기를 최근 1년부터 10년까지 연차적으로 추가한 실험에서 대체로 분류 성능이 향상되는 경향을 보였다. 그러나 전체 학습집합(10년)을 모두 사용하는 경우가 항상 최고 성능은 아닌 것은 선행연구와 유사한 결과이다(김판준, 2016; 김판준, 2006b).

셋째, 분류 알고리즘은 국내 학술지 논문의 자동분류 성능에 영향을 미친다(가설 3). 단일-범주 분류 환경에서 Rocchio, NB, SVM의 3개 분류기를 사용하여 실험한 결과, 전반적으로 Rocchio 분류기의 성능이 가장 우수하였다. 특히 기존에 가장 좋은 성능인 것으로 알려진 SVM과 성능이 거의 동등하거나 더 좋을 뿐 아니라 NB보다는 크게 나은 성능이었다.

넷째, 범주 부여 방법(단일-범주 분류, 복수-범주 분류)은 국내 학술지 논문의 자동분류 성능에 영향을 미친다(가설 4). 먼저 Rocchio 분류기로 이전 실험에서 가장 좋은 성능을 나타낸 요소들을 적용하여 단일-범주 분류와 복수-범주 분류 방법 간에 성능 차이가 있는 지를 비교하였다. 그 결과, 단일-범주 분류 방법보다 복수-범주 분류 방법이 더 나은 성능을 보였다. 다음으로 복수-범주 분류 환경에서 3개 분류기의 연차별 성능을 비교한 결과, Rocchio와 SVM이 거의 동등하게 좋은 성능을 보였고 NB는 성능은 상

대적으로 낮은 수준이었다. 또한, 학습집합의 구성 방법을 달리하여 3개 분류기의 비율별 성능을 살펴본 결과도 이와 유사한 양상을 보였다. 따라서 단일-범주 분류와 마찬가지로 복수-범주 분류 환경에서도 주요 요소들은 국내 학술지 논문의 자동분류 성능에 상당한 영향을 미치는 것으로 나타났다(가설 1, 가설 2, 가설 3).

다섯째, 지금까지의 실험 및 가설 검증 과정에서 최고 성능을 보인 요소를 종합적으로 검토하여, 국내 학술지 논문에 대한 한국연구재단「학술연구분야 분류표」상의 분류 범주(소분류명/세분류명)를 자동으로 할당하기 위한 최적의 분류모형을 도출하였다. 단일-범주 분류와 복수-범주 분류 모두에서 단순한 조합 가중치(ltfidf)를 사용하고 학습집합의 크기를 증가하면 성능이 향상되는 경향을 보였다. 그러나 모든 학습집합을 사용하는 것이 항상 최고 성능은 아니었으며, 분류 대상이 되는 문헌집합 또는 분야의 특성에 따라 어느 정도 규모의 학습집합이 가장 효과적인 가를 파악하는 것이 중요하였다(김판준 2016). 또한, Rocchio와 SVM 분류기가 거의 동등한 수준으로 좋은 성능을 보인 반면, NB의 성능은 상대적으로 낮게 나타났다. 실제 수작업 분류 작업의 결과와 환경을 고려하면, 국내 학술지 논문의 자동분류는 복수-범주 분류 방법을 적용하는 것이 보다 실제적인 접근이라 판단된다. 따라서 복수-범주 분류 방법에서 국내 학술지 논문의 자동분류를 위한 최적의 모형은 단순한 조합 가중치(ltfidf)와 소규모의 학습집합(연차별 3년 이상, 비율별 60%)을 적용한 Rocchio 분류기(Rocchio_긍정)이다. 이러한 Rocchio 분류기는 단순하고 빠르며 소규모의 학습집합을 사용하면서도 SVM과 동등하거나 더 나은 성능을 도출할 수 있는 것으로 나타났다.

5. 결 론

기계학습에 기초한 자동분류 기법을 활용하여 자동분류의 성능에 영향을 미치는 주요 요소들에 대한 가설 설정 및 실험을 수행하고, 그 결과를 분석하여 국내 학술지 논문의 자동분류를 위한 최적의 분류 모형을 제안하였다. 먼저, 단일-범주 분류 환경에서 분류 성능에 영향을 미치는 주요 요소인 용어 가중치부여 기법, 학습집합 크기, 분류 알고리즘에 대한 가설을 검증하는 실험을 수행하였다. 다음으로 이전 실험에서 최고의 성능을 보인 요소들을 적용하여 단일-범주 분류와 복수-범주 분류의 성능을 비교한 다음, 복수-범주 분류 환경에서 주요 요소에 따른 성능을 다각적으로 검토하였다. 이러한 실험 결과와 가설 검증 과정에서 최고 성능을 보인 요소를 종합적으로 분석하여 최적의 분류 모형을 도출하였다. 결과적으로 단일-범주 분류와 복수-범주 분류 모두에서 단순한 조합 가중치(ltfidf)를 사용하고 학습집합의 크기를 증가하면 성능이 향상되지만, 전체 학습집합을 사용하는 것이 항상 최고 성능은 아니며 분류 대상이 되는 문헌집합 또는 분야의 특성에 따라 어느 정도 규모의 학습집합이 가장 효과적인 가를 파악하는 것이 매우 중요한 것으로 나타났다. 본 연구에서 사용된 국내 학술지의 논문들은 서로 동질적(categories linked thematically)이면서 저빈도 범주가 많고 학습집합의 편차가 큰 불균형(imbalanced) 문헌집

합이라 할 수 있다. 따라서 복수-범주 분류 환경을 전제로 할 때, 이러한 국내 학술지 논문의 자동분류를 위한 최적의 모형은 단순한 조합 가중치(ltfidf)와 소규모의 학습집합(연차별 3년 이상, 비율별 60% 이상)을 사용하는 Rocchio 분류기이다.

본 연구의 의의는 다음과 같다. 첫째, 문헌정보학 분야의 학술지인 「정보관리학회지」에 수록된 14년간의 논문집합(2002년~2015년)을 대상으로 한국연구재단(2016) "학술연구분야 분류표" 상의 분류명을 자동 할당하는 것으로, 실제 적용이 가능한 문헌집합과 분류체계에 기초한 실질적인 연구 결과이다. 둘째, 단일-범주 분류 및 복수-범주 분류 양 측면에서 국내 학술지 논문의 자동분류에 핵심적인 요소들을 다각적으로 검토한 결과를 제시하였다. 셋째, 단순하고 빠른 Rocchio 분류기가 소규모의 학습집합을 사용하는 경우에도 상당히 좋은 분류 성능을 안정적으로 보여준다는 것을 확인하였다.

본 연구는 국내 학술지 논문의 자동분류에서 최대한의 자질정보를 활용하기 위해 주요 요소 중 하나인 자질선정을 적용하지 않았으므로 이에 대한 추가적인 연구가 필요하다. 또한, 실험 데이터 구축에 필수적인 수작업 분류작업에 상당한 시간과 전문성이 요구되는 관계로 특정 분야의 학술지에 수록된 논문을 대상으로 하였다. 따라서 향후 다수의 학술지 또는 학문분야로 실험 문헌집합을 확장하는 연구가 필요할 것이다.

참 고 문 헌

강승식 (2002). 한국어 형태소 분석과 정보검색. 홍릉출판사.

김성희, 엄재은 (2008). 기계학습을 이용한 문서 자동분류에 관한 연구. 정보관리연구, 39(4), 47-66. http://dx.doi.org/10.1633/JIM.2008.39.4.047

김용환, 정영미 (2012). 위키피디아를 이용한 분류자질 선정에 관한 연구. 정보관리학회지, 29(2), 155-171. http://dx.doi.Org/10.3743/KOSIM.2012.29.2.155

김종민, 유창동 (2014). 특징 추출 비용에 민감한 분류를 위한 선형 분류기 최적화 알고리즘. 2014년도 대한전자공학회 하계학술대회 논문집, 37(1), 2021-2024.

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

김판준 (2006b). 로치오 알고리즘을 이용한 학술지 논문의 디스크립터 자동부여에 관한 연구. 정보관리학회지, 23(3), 69-89. http://dx.doi.org/10.3743/KOSIM.2006.23.3.069

김판준 (2008). 용어 가중치부여 기법을 이용한 로치오 분류기의 성능 향상에 관한 연구. 정보관리학회지, 25(1), 211-233. http://dx.doi.org/10.3743/KOSIM.2008.25.1.211

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, 이재윤 (2007). 문헌간 유사도를 이용한 자동분류에서 미분류 문헌의 활용에 관한 연구. 정보관리학회지, 24(1), 251-271. http://dx.doi.org/10.3743/KOSIM.2007.24.1.251

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

송성전, 정영미 (2012). 용어의 문맥활용을 통한 문헌 자동 분류의 성능 향상에 관한 연구. 정보관리학회지, 29(2), 205-224. http://dx.doi.Org/10.3743/KOSIM.2012.29.2.205

심경 (2006). 문헌범주화에서 학습문헌수 최적화에 관한 연구. 정보관리학회지, 23(4), 277-294. http://dx.doi.org/10.3743/KOSIM.2006.23.4.277

심경, 정영미 (2006). 학습문헌집합에 기 부여된 범주의 정확성과 문헌 범주화 성능. 정보관리학회지, 23(2), 265-285. http://dx.doi.org/10.3743/KOSIM.2006.23.2.265

이용구 (2009). 기계번역을 이용한 교차언어 문서 범주화의 분류 성능 분석. 한국문헌정보학회지, 43(1), 313-332. http://dx.doi.org/10.4275/kslis.2009.43.1.313

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

이재윤 (2005a). 문서측 자질선정을 이용한 고속 문서분류기의 성능향상에 관한 연구. 정보관리연구, 36(4), 51-69. http://dx.doi.org/10.1633/jim.2005.36.4.051

이재윤 (2005b). 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). 학술연구분야 분류표. Retrieved from http://www.nrf.re.kr

한국학술지인용색인 웹사이트 (2018). Retrieved from https://www.kci.go.kr

AI-Salemi, B., Aziz, M., Juzaiddin, A., & Noah, S. (2015). Boosting algorithms with topic modeling for multi-label text categorization: A comparative empirical study. Journal of Information Science, 41(5), 732-746. http://dx.doi.Org/10.1177/0165551515590079

Chen, E., Lin, Y., Xiong, H., Luo, Q., & Ma, H. (2011). Exploiting probabilistic topic models to improve text categorization under class imbalance. Information Processing and Management, 47(2), 202-214.

Chen, Yao-Tsung, & Chen, Meng Chang (2011). Using chi-square statistics to measure similarities for text categorization. Expert Systems with Application, 38(4), 3085-3090.

Dalal, M. K., & Zaveri, M. A. (2012). Automatic text classification of sports blog data, proceedings of the ieee international conference on computing, communications and applications (ComComAp 2012), Hong Kong, 11-13 January 2012, 219-222.

Dalal, M. K., & Zaveri, M. A. (2013). Automatic classification of unstructured blog text. Journal of Intelligent Learning Systems and Applications, 5(2), 108-114. http://dx.doi.Org/10.4236/jilsa.2013.52012.

Eriksson, Tobias (2013). Automatic web page categorization using text classification methods. Master's Degree Project in Computer Science CSC School of Computer Science and Communication.

Foulds, J., & Frank, E. (2010). A review of multi-instance learning assumptions. Knowl. Eng. Rev., 25(1), 1-25.

Hmeidi, I., Al-Ayyoub, M., Abdulla, N. A., Almodawar, A. A., Abooraig, R., & Mahyoub, N. A. (2015). Automatic arabic text categorization: A comprehensive comparative study. Journal of Information Science, 41(1), 114-124. https://doi.org/10.1177/0165551514558172

Jiang, S., Pang, G., Wu, M., & Kuang, L. (2012). An improved k-nearest-neighbor algorithm for text categorization. Expert Systems with Applications, 39(1), 1503-1509. https://doi.org/10.1016/j.eswa.2011.08.040

Jindal, Rajni, Malhotra, Ruchika, & Jain, Abha. (2015). Techniques for text classification: Literature review and current trends. Webology, 12(2), 2-28.

Joorabchi, A., & Mahdi, A. E. (2011). An unsupervised approach to automatic classification of scientific literature utilizing bibliographic metadata. Journal of Information Science, 37(5), 499-514. https://doi.org/10.1177/0165551511417785

Khan, A., Baharudin, B., & Lee, L. H. (2010). A review of machine learning algorithms for text-documents classification. Journal of Advances in Information Technology, 1(1), 4-20. https://doi.org/10.4304/jait.1.1.4-20

Kumar, M. A., & Gopal, M. (2010). A comparison study on multiple binary-class SVM methods for unilabel text categorization. Pattern Recognition Letters, 31(11), 1437-1444. https://doi.org/10.1016/j.patrec.2010.02.015

Li, C. H., & Park, S. C. (2009). An efficient document classification model using an improved back propagation neural network and singular value decomposition. Expert Systems with Applications, 36(2), 3208-3215. https://doi.org/10.1016/j.eswa.2008.01.014

Liu, Y., Loh, H. T., Yousef-Toumi, K., & Tor, S. B. (2007). Handling of imbalanced data in text classification: category-based term weights. In Kao, A., & Poteet, S. R. eds. Natural

Language Processing and Text Mining. Springer, 171-192. https://doi.org/10.1007/978-1-84628-754-1_10

Miao, Yun-Qian, & Kamel, Mohamed (2011). Pairwise optimized rocchio algorithm for text categorization. Pattern Recognition, 32(2), 375-382. https://doi.org/10.1016/j.patrec.2010.09.018

Pawar, P. Y., & Gawande, S. H. (2012). Comparative study on different types of approaches to text categorization. International Journal of Machine Learning and Computing, 2(4), 423-426. https://doi.org/10.7763/ijmlc.2012.v2.158

Pedregosa, F. et al. (2011). Scikit-learn: Machine learning in python. Journal of Machine Learning Research, 12, 2825-2830.

Read, J. (2010). Scalable Multi-label Classification (Thesis, Doctor of Philosophy (PhD)). University of Waikato, Hamilton, New Zealand. Retrieved from https://hdl.handle.net/10289/4645

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. Machine Learning, 85, 333-359.

Schapire, R. E., & Singer, Y. (2000). BoosTexter: A boosting-based system for text categorization. Machine Learning, 39, 135-168.

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

Shehab, M. A., Badarneh, O., Al-Ayyoub, M., & Jararweh, Y. (2016). A supervised approach for multi-label classification of Arabic news articles, 7th International Conference on Computer Science and Information Technology (CSIT), Amman, 2016, 1-6. http://dx.doi.Org/10.1109/CSIT.2016.7549465

Tarragó, D. S., Cornelis, C., Bello, R., & Herrera, F. (2014). A multi-instance learning wrapper based on the Rocchio classifier for web index recommendation. Knowledge-Based Systems, 59, 173-181. https://doi.org/10.1016/j.knosys.2014.01.008

Torii, M., Yin, L., Nguyen, T., Mazumdar, C. T., Liu, H., Hartley, D. M., & Nelson, N. P. (2011). An exploratory study of a text classification framework for Internet-based surveillance of emerging epidemics. International Journal of Medical Informatics, 80(1), 56-66. https://doi.org/10.1016/j.ijmedinf.2010.10.015

Tsoumakas G, Katakis I., & Vlahavas I. (2010). Mining multi-label data. In: Data mining and knowledge discovery handbook. Berlin: Springer, 667-685.

Uğuz, Harun. (2011). A two-stage feature selection methods for text categorization by using information gain, principal component analysis and genetic algorithm. Knowledge-Based

Systems, 24(7), 1024-1032. https://doi.org/10.1016/j.knosys.2011.04.014

Vasuki, Vidya, & Cohen, Trevor (2010). Reflective random indexing for semi-automatic indexing of the biomedical literature. Journal of Biomedical Informatics, 43(5), 694-700. https://doi.org/10.1016/j.jbi.2010.04.001

Villena-Román, J., Collada-Pérez, S., Lana-Serrano, S., & González-Cristóbal, J. C. (2011). Hybrid approach combining machine learning and a rule-based expert system for text categorization. In Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, 323-328.

Vogrinčič, Sergeja, & Bosnić, Zoran (2011). Ontology-based multi-label classification of economic articles. ComSIS, 8(1), 101-119. https://doi.org/10.2298/csis100420034v

Wang, Tai-Yue, & Chiang, Huei-Min (2007). Fuzzy support vector machine for multi-class text categorization. Information Processing and Management, 43(4), 914-929. https://doi.org/10.1016/j.ipm.2006.09.011

Wu, Chih-Hung (2009). Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. Expert Systems with Applications, 36(1), 4321-4330. https://doi.org/10.1016/j.eswa.2008.03.002

Yu, B., Xu, Zong-ben, & Li, Cheng-hua (2008). Latent semantic analysis for text categorization using neural network. Knowledge-Based Systems, 21(8), 900-904. https://doi.org/10.1016/j.knosys.2008.03.045

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

Chung, Eun-Kyung (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. http://dx.doi.Org/10.3743/KOSIM.2009.26.3.261

Kang, Seung-Shik (2002). Korean Morphology and Information Retrieval. Hongrung Publishing Company.

Kim, Jong-Min, & Yoo, Chang D. (2014). Linear classifier optimization for feature acquisition cost-sensitive classification. In Proceedings of the IEEK Conference, 37(1), 2021-2024.

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

Kim, Pan Jun (2006b). A study on the automatic descriptor assignment for scientific journal articles uing rocchio algorithm. Journal of the Korean Society for Information Management, 23(3), 69-89. http://dx.doi.org/10.3743/KOSIM.2006.23.3.069

Kim, Pan Jun (2008). A study on the performance improvement of rocchio classifier with term weighting methods. Journal of the Korean Society for Information Management, 25(1), 211-233. http://dx.doi.org/10.3743/KOSIM.2008.25.1.211

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, & Lee, Jae Yun (2007). Utilizing unlabeled documents in automatic classification with inter-document similarities. Journal of the Korean Society for Information Management, 24(1), 251-271. http://dx.doi.org/10.3743/KOSIM.2007.24.1.251

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. http://dx.doi.org/10.3743/KOSIM.2012.29.2.225

Kim, Pan Jun, & Lee, Jae Yun (2014). 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 Society for Library and Information Science, 48(3), 491-510. http://dx.doi.org/10.4275/KSLIS.2014.48.3.491

Kim, Seong-Hee, & Eom, Jae-Eun (2012). A study on the documents's automatic classification using machine learning. Journal of Information Management, 39(4), 47-66. http://dx.doi.org/10.1633/JIM.2008.39.4.047

Kim, Yong-Hwan, & Chung, Young-Mee (2012). An experimental study on feature selection using wikipedia for text categorization. Journal of the Korean Society for information Management, 29(2), 155-171. http://dx.doi.Org/10.3743/KOSIM.2012.29.2.155

Korea Citation Index (2018). Retrieved from https://www.kci.go.kr

Lee, Jae Yun (2005a). Improving the performance of a fast text classifier with document-side feature selection. Journal of Information Management, 36(4), 51-69. http://dx.doi.org/10.1633/jim.2005.36.4.051

Lee, Jae Yun (2005b). 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 (2009). Classification performance analysis of cross-language text categorization using machine translation. Journal of the Korean Society for Library and Information Science, 43(1), 313-332. http://dx.doi.org/10.4275/kslis.2009.43.1.313

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

National Research Foundation of Korea (2016). Research Field Classification Scheme. Retrieved from http://www.nrf.re.kr

Shim, Kyung (2006). Optimization of number of training documents in text categorization. Journal of the Korean Society for Information Management, 23(4), 277-294. http://dx.doi.org/10.3743/KOSIM.2006.23.4.277

Shim, Kyung, & Chung, Young-Mee (2006). The effect of the quality of pre-assigned subject categories on the text categorization performance. Journal of the Korean Society for Information Management, 23(2), 265-285. http://dx.doi.org/10.3743/KOSIM.2006.23.2.265

Song, Sung-Jeon, & Chung, Young-Mee (2012). A study on improving the performance of document classification using the context of terms. Journal of the Korean Society for Information Management, 29(2), 205-224. http://dx.doi.Org/10.3743/KOSIM.2012.29.2.205