해외 데이터베이스의 통제키워드에 기초한 국내 학술지 논문의 자동분류 성능 향상에 관한 실험적 연구

An Experimental Study on the Performance Improvement of Automatic Classification for the Articles of Korean Journals Based on Controlled Keywords in International Database

김 판 준 (Pan Jun Kim)*
이 재 윤 (Jae Yun Lee)**

목 차

1. 서 론
1.1 연구의 필요성 및 목적
1.2 선행 연구
2. 분류 실험 설계
2.1 실험 데이터
2.2 분류기와 성능 척도
2.3 실험 단계
3. 단일 분류기 실험
3.1 마이크로 분류 성능 평가
3.2 매크로 분류 성능 평가
4. 분류기 결합 실험
4.1 분류기 결합 전략
4.2 분류기 결합 실험 결과
5. 결 론

초 록

학술지 논문의 효율적인 관리 및 검색을 위한 주요 요소인 키워드는 통제키워드와 비통제키워드로 구분할 수 있다. 그러나 현재 국내 데이터베이스에서 대부분의 학술지 논문에는 비통제키워드인 저자키워드만이 부여되어 있을 뿐, 망라적인 탐색을 돕는 통제키워드로서 디스크립터는 제공되지 않고 있다. 이 연구에서는 해외 데이터베이스의 학술지 논문에 부여된 통제키워드를 학습한 분류기를 사용하여, 국내 학술지 논문에 디스크립터를 자동 할당하는 실험을 수행하였다. 그 결과, 국외 데이터베이스의 디스크립터 학습을 통해 영문 초록이 있는 국내 학술지 논문에 통제키워드를 자동 할당할 수 있는 가능성을 확인하였다. 또한, 다양한 분류기 및 분류기 결합을 통하여 이러한 디스크립터 자동 할당의 성능 향상을 모색하였다.

ABSTRACT

As a major factor for efficient management and retrieval of the articles in databases, keywords are classified into uncontrolled keywords and controlled keywords. Most of Korean scholarly databases fail to provide controlled vocabularies to indexing research articles which help users to retrieve relevant papers exhaustively. In this paper, we carried out automatic descriptor assignment experiments to Korean articles using automatic classifiers learned with descriptors in international database. The results of the experiments show that the classifier learning with descriptors in international database can potentially offer controlled vocabularies to Korean scholarly articles having English abstracts. Also, we sought to improve the performance of automatic descriptor assignment using various classifiers and combination of them.

키워드: 디스크립터, 자동분류, 텍스트 범주화, 분류기 결합, 통제어휘
Descriptors, Automatic Classification, Text Categorization, Classifier Fusion, Controlled Vocabulary

* 신라대학교 문헌정보학과 조교수(pjkim@silla.ac.kr) (제1저자)
** 명지대학교 문헌정보학과 부교수(memexlee@mju.ac.kr) (교신저자)
논문접수일자: 2014년 7월 22일 최초심사일자: 2014년 8월 13일 게재확정일자: 2014년 8월 20일
한국문헌정보학회지, 48(3): 491-510, 2014. [http://dx.doi.org/10.4275/KSLIS.2014.48.3.491]

1. 서 론

1.1 연구의 필요성 및 목적

1960년대에 시작된 비통제색인(자연언어색인)과 통제색인(통제언어색인 또는 주제색인) 간의 우위에 대한 논의는 1980년대에 와서 양자를 함께 사용하는 복합시스템이 가장 이상적이라는 합의에 도달하였다. 그러나 1990년대 중반 이후 인터넷의 등장으로 상용 검색엔진을 통한 검색이 보편화되면서 새로운 문제가 제기되었다. 다양한 언어와 주제, 유형들로 존재하는 엄청난 규모의 정보자원을 대상으로 검색이 이루어지는 웹 환경에서는 적절한 통제어휘표(시소러스, 주제명표목표, 온톨로지 등)를 이용하여 모든 자원을 색인하는 것은 거의 불가능에 가깝다. 이에 따라 웹 환경에서는 기계적으로 자동추출된 비통제키워드를 사용한 검색이 보편화되었고, 막대한 시간과 비용 문제로 인하여 통제키워드에 기초한 검색은 상대적으로 소홀하게 다루어지고 있다.

그러나 인터넷이 대중화된 1990년대 이후에도 여러 학자들이 통제언어와 비통제언어(자연언어)를 사용하는 검색을 서로 비교하는 연구를 지속적으로 수행한 결과에 따르면, 두 가지 유형의 언어를 모두 사용하는 것이 더 효과적이라는 결론에 이르고 있다(Gross and Taylor 2005; Kipp 2005; McCutcheon 2009; Rowley 1994; Tillostson 1995; Voorbij 1998). 특히, 학술 데이터베이스의 탐색과 활용 측면에서 통제키워드의 필요성과 유용성은 의심의 여지가 없는 사실이다. 통제키워드는 특정한 개념에 대한 다양한 표현들을 대표 용어로 일관성 있게 표현하므로 특정 주제에 관한 정보자료를 망라적으로 검색할 수 있다는 고유의 장점을 갖는다.

따라서 웹 환경에서도 대부분의 해외 학술 데이터베이스들은 이러한 두 가지 유형의 언어로 작성된 키워드를 상호보완적으로 제공하여, 양자의 장단점을 보완하여 검색을 수행할 수 있도록 하고 있다. 특히, CSA Illumina, DIALOG, EBSCO, ProQuest 등 대표적인 해외 데이터베이스 서비스들은 자연언어 색인과 통제언어 색인을 함께 채택하고 있다. 이렇게 함으로써 이용자가 익숙한 언어를 선택할 수 있도록 하고 있으며, 또한 두 가지 언어를 모두 사용하여 망라적인 검색이 가능하도록 한다(정영미, 2012, 38). 해외 학술 데이터베이스의 경우 학술지 논문의 색인작업은 크게 두 가지 경로로 이루어진다. 첫째, 컴퓨터가 입력문헌의 텍스트를 분석하여 문헌의 내용을 대표하는 키워드(자연언어 색인어)를 일정한 기준에 의해 기계적으로 추출한다. 둘째, 색인전문가는 해당 문헌의 내용을 분석하여 다루고 있는 주제를 판단한 다음, 통제어휘집에서 이를 표현할 수 있는 적절한 디스크립터(통제언어 색인어)를 부여한다.

그러나 국내 학술 데이터베이스들은 색인전문가 및 통제어휘표의 부재라는 근본적인 문제로 인하여 학술지 논문에 대한 통제키워드(디스크립터, 주제명 등)가 전혀 제공되지 않고 있는 상황이다. 국내 학술 데이터베이스의 색인어 필드에서 제공하고 있는 색인어는 모두 저자키워드로서 비통제 색인어이다. 자연언어 색인으로서 저자키워드는 동일한 개념을 저자에 따라 여러 개의 다른 용어로 표현하고 있는 경우가 많아, 동일한 개념을 표현하는 모든 용어들과 어형이 다른 용어들을 모두 검색어로 사용해야

만 관련된 모든 정보자료의 검색이 가능한 문제가 있다(김판준, 이재윤 2012). 이러한 상황에서 색인전문가 측면에서는 보다 효율적으로 많은 문헌을 빠른 시간 내에 일관성 있게 색인하고, 이용자 측면에서는 원하는 정보를 주제 또는 개념 측면에서 접근할 수 있도록 하기 위한 통제키워드의 제공에 대한 필요성이 커지고 있다. 그러나 대부분의 국내 학술 데이터베이스에서 통제키워드를 전혀 제공하지 않고 있는 상황에서, 지금까지 색인전문가에 의해 전적으로 수행되어 온 이러한 역할을 일부 대체하거나 효과적으로 지원할 수 있는 방법을 적극적으로 모색할 필요가 있다. 이러한 측면에서 본 연구는 '독서(reading)' 분야를 대상으로 해외 데이터베이스의 학술지 논문에 부여된 통제키워드를 자동으로 학습하여, 영문 초록이 있는 국내 학술지 논문에 디스크립터를 자동 할당할 수 있는 가능성을 확인해 보고자 한다. 또한, 다양한 분류기와 이들 분류기의 결합을 통해 이러한 디스크립터 자동 할당의 성능을 향상시킬 수 있는 방안을 제시하고자 한다.

1.2 선행 연구

웹 환경에서도 이미 학술정보의 검색에서 비통제키워드와 통제키워드의 상호보완적인 사용이 가장 효율적이라는 여러 학자들의 연구결과가 보고되었음에도 불구하고(Chan 2000; Gross and Taylor 2005; McCutcheon 2009; Voorbij 1998), 현재까지 국내 학술 데이터베이스의 학술지 논문 검색환경에서 통제키워드를 제공하기 위한 시도는 거의 없었다. 더구나, 색인전문가와 적절한 통제어휘표의 부재라는 근본적인 문제를 해소하면서 국내 학술 데이터베이스의 검색 환경에서 통제키워드를 제공할 수 있는 방안을 제시한 연구는 찾아보기 힘든 상황이다. 본 연구에서는 국내외에서 수행된 학술지 논문에 대한 통제키워드의 자동 할당을 위한 연구들을 크게 두 가지 유형으로 구분하였다. 먼저 국내에서 수행된 텍스트 범주화 기법을 사용하여 학술지 논문에 통제키워드(디스크립터, 주제명, 범주명)를 자동 할당하기 위한 연구들이 있다. 다음으로 국외에서 수행된 텍스트 범주화 기법에 기초한 문헌(신문기사, 국제기구 문서 등)에 통제키워드를 자동 할당하기 위한 연구들이 있다.

국내 학술지 논문을 대상으로 통제키워드의 자동 할당을 모색한 연구는 2006년부터 시작되었다고 볼 수 있다. 김판준은 2006년 국내 최초로 기계학습 기반의 접근법을 사용하여 국내 학술지 논문에 통제키워드(디스크립터)를 자동 할당하기 위한 효율적인 방안을 자질선정, 학습집합의 규모, 분류기 등 다양한 측면에서 검토하였다(김판준 2006a; 2006b; 2008). 또한, 김판준과 이재윤(2007; 2012)은 국내 학술지 논문에 디스크립터를 자동 할당하는 과정에서 실용적으로 활용이 가능한 미분류 문헌을 활용하는 방안을 제시하였고, 최근에는 현재 국내 대부분의 학술 데이터베이스에서 통제키워드 대신 제공되고 있는 저자키워드(비통제키워드)의 재분류를 통하여 디스크립터(통제키워드)를 자동 할당하는 방안을 제안한 바 있다. 이처럼 김판준과 이재윤은 텍스트 범주화(문헌의 자동분류)를 위한 기계학습 접근법에 기반하여 색인전문가와 통제어휘표 없이도 학술지 논문에 통제키워드로서 디스크립터를 자동 할당할 수 있는 방

안을 여러 측면에서 모색하였다. 이외에 이용구(2012)는 국내 학술지 논문으로 구성된 KTSET(김성혁 외 1994)을 대상으로 서로 다른 언어로 작성된 문헌의 자동분류를 위하여 여러 교차언어 텍스트 범주화(CLTC: Cross-Language Text Categorization) 방법들을 적용한 결과, 학습집단 번역방법의 분류 성능이 비교적 좋은 것으로 보고하였다.

국외에서 텍스트 범주화 기법에 기반하여 문헌에 통제키워드를 자동 할당하기 위한 연구는 1990년대 이후 기계학습 알고리즘을 기반으로 전문가시스템, 미분류 문헌, 외부 정보(WordNet, Wikipedia 등)를 활용하는 등의 다양한 응용과 시도가 이루어졌다(김판준, 이재윤 2012, 4-5). 또한, 최근에는 교차언어 정보검색(Cross-Language Information Retrieval: CLIR) 접근법에 기초하여 문헌에 통제키워드를 자동 할당하는 교차언어 텍스트 범주화에 관한 연구가 활발히 수행되고 있다(Bel, Koster and Villegas 2003; Olsson, Oard and Hajic 2005; Rigutini, Maggini and Liu 2005; Wei et al. 2014; Wu and Oard 2008). 기존의 텍스트 범주화 기법들은 단일언어로 작성된 문헌을 대상으로 하지만, 교차-언어 텍스트 범주화는 서로 다른 언어로 작성된 문헌을 대상으로 텍스트 범주화를 수행한다. 즉, 하나의 언어(주로 영어)로 작성된 문헌에 부여된 범주를 학습한 결과로서 범주가 부여되지 않은 다른 언어(이탈리아어, 스페인어, 체코어, 중국어 등)로 작성된 문헌에 적합한 범주를 할당하는 것이다.

그러나 교차-언어 텍스트 범주화의 문제점은 기계번역, 사전-기반, 말뭉치-기반 등의 방법에 기초하여 미분류된 문헌을 번역하는 과정에서 요구되는 시간과 비용에 비하여, 언어와 문화의 차이로 인한 오류와 정보 손실(information loss)이 상당히 크다는 것이다(Guo and Xiao 2012; Wei, Lin and Yang 2011). 지금까지 다양한 측면에서 이러한 문제를 해소하기 위한 방안을 제시하는 많은 연구들이 보고되고 있지만 여전히 서로 다른 언어를 번역하는데 막대한 시간과 비용을 필요로 한다는 문제점을 내재하고 있다. 또한, 대부분의 교차-언어 텍스트 범주화 연구에서는 분류 대상을 뉴스 기사로 구성된 실험문헌 집합으로 하고 있어, 아직까지 학술지 논문을 대상으로 하는 연구는 찾아보기 힘들다.

따라서 본 연구는 학술지 논문을 대상으로 하나의 언어(영어)로 작성된 문헌에 부여된 통제키워드(범주명)를 학습하여 다른 언어(한국어)로 작성된 문헌에 디스크립터를 자동 할당하는 실험을 수행하였다. 특히, 이러한 실험은 색인전문가 및 통제어휘표의 부재로 인해 통제키워드는 없지만 대부분의 학술지 논문에 대한 영어 제목과 초록 필드가 제공되고 있는 국내 데이터베이스의 현실을 반영하여, 투자 대비 효율이 떨어지는 번역 과정을 필요로 하지 않는 기계학습 접근법에 기초하였다.

언어뿐만 아니라 출처가 다양한 해외 데이터베이스의 논문들을 학습집합으로 사용하여 국내 논문을 자동으로 분류하는 문제는 일반적인 자동분류에 비해서 좋은 성능을 얻기가 쉽지 않다. 따라서 분류 성능을 향상시키기 위한 추가 전략이 필요하다. 기계학습을 통한 자동분류의 성능을 향상시키기 위한 전략은 용어 가중치 활용이나 자질 선정(김판준 2008; 이재윤 2005), WordNet이나 Wikipedia와 같은 외부 자원의 활용(김용환, 정영미 2012; 정은경 2007), 미분

류 학습문헌의 활용(김판준, 이재윤 2007), 여러 분류기의 판정 결과를 조합하는 분류기 결합(송성전, 정영미 2012; 유호현, 정영미 2008) 등이 사용된다. 이중에서 외부 자원을 추가로 필요로 하지 않고 실험에 사용된 자원과 조건만으로 수행이 가능한 분류기 결합 방법을 채택하여 성능 향상 실험을 추가로 실시해보았다.

2. 분류 실험 설계

본 연구에서 실험집단에 대한 사전처리와 자질선정을 위한 프로그램은 Python 및 Visual FoxPro로 구현된 프로그램을 사용하였고, 국내 학술지 논문에 대한 통제키워드 자동 부여 실험을 위한 프로그램(분류기)은 공개된 기계학습 실험 패키지인 WEKA Version 3.6(Witten, Frank and Hall 2011)을 사용하였다.

2.1 실험 데이터

디스크립터 자동부여를 위한 데이터 집합의 구성에서 세 가지 기본적인 요소는 문헌집합(collection), 디스크립터(descriptors), 분류자질이다(김판준 2006a). 본 연구의 실험을 위해 준비한 세 요소의 구체적인 내용은 다음과 같다.

첫째, 문헌집합은 '독서' 영역을 대상으로 2013년 7월에 국내외의 대표적인 학술 데이터베이스에서 직접 검색한 결과를 사용하였다. 국외 학술 데이터베이스(LISTA)와 국내 학술 데이터베이스(RISS)의 검색문헌 집합에 기초하여 구성한 실험문헌 집합의 세부 내용은 〈표 1〉과 같다. 즉 학습집합은 국외 학술 데이터베이스(LISTA)에서 검색된 1,809건의 제목과 초록, 통제키워드 필드에 기초하여 구성하고, 검증집합은 국내 학술 데이터베이스(RISS)에서 검색된 한국어 학술지 논문 798건의 영문 제목과 영문 초록 필드로 구성하였다.

둘째, 통제키워드로서 디스크립터는 국외 학술 데이터베이스(LISTA)의 검색 결과에서 'SU (Subject Terms)' 필드에 출현한 통제키워드를 사용하였다. 본 실험에서 사용된 상위 10개의 통제키워드는 'WRITING', 'TEACHER-librarians', 'SCHOOL libraries', 'PUBLIC libraries', 'READING comprehension', 'READERS', LITERACY, 'CURRICULA (Courses of study)', 'BIBLIOTHERAPY', 'BOOKS'이다. 이들 통제키워드가 할당된 문헌은 검색어인 'reading'이 출현한 학습문헌 1,809건 중에서 479건이었으며 나머지 1,330건은 이들이 할당되지 않았다. 10종 디스크립터의 할당 교차표는 〈표 2〉와 같다.

이들 통제 키워드 10개를 검증집합 798건에 대해 수작업으로 부여하였다. 연구자 2인이 각자 부여한 후 의견이 불일치된 경우는 상의하여 조정한 결과로 〈표 3〉과 같이 할당되었다. 전체

〈표 1〉 실험문헌 집합

구분	DB	질의어	검색문헌수	발행연도	대상 필드
학습집합	LISTA	reading	1,809	2000년 - 2013년	제목, 초록, 통제키워드
검증집합	RISS	독서	798	2000년 - 2013년	영문 제목, 영문 초록

〈표 2〉 학습문헌의 10종 디스크립터 할당 교차표

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	UC	계
C1 - Writing	25	0	1	1	1	0	1	0	0	3	0	32
C2 - Teacher-Librarians	0	60	10	1	7	1	12	1	0	0	0	92
C3 - School Libraries	1	10	62	3	1	0	10	0	0	6	0	93
C4 - Public Libraries	1	1	3	128	1	1	7	0	4	2	0	148
C5 - Reading Comprehension	1	7	1	1	41	1	4	1	0	1	0	58
C6 - Readers	0	1	0	1	1	32	3	0	1	3	0	42
C7 - Literacy	1	12	10	7	4	3	111	1	0	6	0	155
C8 - Curricula	0	1	0	0	1	0	1	17	1	0	0	21
C9 - Bibliotherapy	0	0	0	4	0	1	0	1	17	0	0	23
C10 - Books	3	0	6	2	1	3	6	0	0	68	0	89
UC - 미할당	0	0	0	0	0	0	0	0	0	0	1330	1330
계	32	92	93	148	58	42	155	21	23	89	1330	1809

〈표 3〉 검증문헌에 대한 디스크립터 수작업 할당 결과 교차표

	C1	C2	C3	C4	C5	C6	C7	C8	C9	C10	UC	계
C1 - Writing	22	0	0	0	9	0	11	3	1	0	0	46
C2 - Teacher-Librarians	0	37	13	5	3	2	4	5	0	1	0	70
C3 - School Libraries	0	13	35	2	1	0	4	4	1	1	0	61
C4 - Public Libraries	0	5	2	65	0	4	1	0	0	4	0	81
C5 - Reading Comprehension	9	3	1	0	76	7	31	10	0	0	0	137
C6 - Readers	0	2	0	4	7	60	6	1	1	2	0	83
C7 - Literacy	11	4	4	1	31	6	76	9	1	0	0	143
C8 - Curricula	3	5	4	0	10	1	9	36	0	0	0	68
C9 - Bibliotherapy	1	0	1	0	0	1	1	0	20	0	0	24
C10 - Books	0	1	1	4	0	2	0	0	0	27	0	35
UC - 미할당	0	0	0	0	0	0	0	0	0	0	466	466
계	46	70	61	81	137	83	143	68	24	35	466	798

798개 문헌 중에서 가장 많은 143개 문헌에 할당된 디스크립터는 C7인 'Literacy'였으며, 가장 적은 24개 문헌에 할당된 디스크립터는 C9인 'Bibliotherapy'였다. 가장 많은 문헌에 할당된 'Literacy'는 다른 디스크립터와 중복 할당되는 경우도 많아서 C5인 'Reading Comprehension'과는 31건, C1인 'Writing'과는 11건의 문헌에 각각 중복 할당되었다. 466건의 문헌은 실험에 사용한 10종의 디스크립터가 할당되지 않았으며 분류 실험에서 이들 문헌에 디스크립터가 할당되면 부정 오류(FN: False Negative)로 판정된다. 미할당 문헌을 실험 집합에서 제외할 수도 있으나 실제 디스크립터 자동 할당 상황과 최대한 유사하도록 실험 환경을 설정하기 위해서 그대로 포함하였다. 디스크립터를 자동 부여하는 실제 상황에서는 10종 디스크립터 중 어느 하나가 부여되는지 여부를 미리 알 수가 없기 때문에 아무 것도 부여되지 않을 문헌도 분류기

에 투입될 수밖에 없기 때문이다.

셋째, 자질선정 측면에서는 실험집합에 속한 학술지 논문의 영어 표제와 초록에 출현한 단어 중에서 Porter 스태머를 이용한 어근추출과 불용어(전치사, 조사, 숫자 등) 제거 절차를 거친 이후, 전체 문헌집단에서 6개 이상의 논문에 출현한 키워드를 자질로 선정하여 출현빈도에 기반한 가중치(logTF × IDF)를 부여한 문헌벡터를 구성하였다.

2.2 분류기와 성능 척도

기계학습 기반의 텍스트 범주화를 통한 디스크립터 자동 할당에서 문헌 및 자질집합과 함께 가장 중요한 요소는 분류기이다. 본 연구의 목적에 적합한 분류기를 선정하기 위해 지금까지 선행연구들에서 주로 사용되어 온 기본 분류기들을 검토하여, 본 연구의 목적에 적합한 것으로 WEKA Version 3.6(Witten, Frank and Hall 2011)에서 제공하는 9개 분류기(NB, SVM, VPT, RBF, KNN1, ADT10, J48, OneR, Ridor)를 선정하였다(김판준, 이재윤 2012).

각 분류기의 성능은 분류 정확률과 분류 재현율, 그리고 이들에 기초한 복합 척도인 매크로 F1 및 마이크로 F1과 함께, 재현율보다 정확률에 두 배 중요도를 부여하는 매크로 F0.5와 마이크로 F0.5로도 평가하였다. 정확률을 재현율보다 중요하게 반영하는 이유는 디스크립터 할당의 경우에 하나의 논문에 여러 개의 디스크립터가 할당되므로 소수의 디스크립터가 누락되는 경우는 눈에 잘 띄지 않지만, 잘못된 디스크립터가 할당된 것은 논문 정보에서 바로 드러나기 때문이다(김판준, 이재윤 2012). 또한 분류실험의 경우에는 재현율을 향상시키는 것에 비해서 상대적으로 정확률을 향상시키는 것이 어려운 문제이다. 따라서 디스크립터 할당의 경우에는 할당된 디스크립터 중 올바른 디스크립터의 비율인 정확률이 재현율보다 더 중요한 기준이라고 판단하여 F0.5 척도를 평가에 함께 사용하였다.

2.3 실험 단계

9개 분류기의 성능을 비교한 1차 실험에서는 각 분류기마다 10개 디스크립터가 할당된 영어 학술지 논문의 제목과 초록에 출현한 단어 자질로 학습단계를 수행한 후, 798건의 국내 논문집합을 대상으로 디스크립터 자동 할당을 수행하여 성능을 비교해보았다. 모든 분류는 각 문헌에 각 디스크립터의 할당 여부를 판정하는 이원분류 방식으로 수행하였다.

분류기 결합을 시도한 2차 실험에서는 1차 실험에서 좋은 성능을 보인 분류기를 중심으로 여러 분류기의 판정 결과를 결합하는 분류기 결합 실험을 수행해보았다. 이를 통해 단일 분류기를 사용한 경우에 비해 더 좋은 성능과 더 안정적인 성능을 얻을 수 있는지 여부를 살펴보았다.

3. 단일 분류기 실험

3.1 마이크로 분류 성능 평가

1차 실험에서는 WEKA가 제공하는 9종의 분류기를 사용하여 10개 디스크립터를 798건의

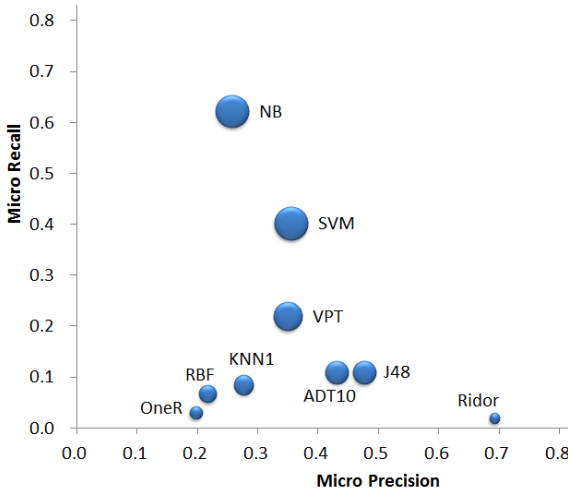
논문에 할당하는 실험을 수행하였다. 실험 결과 분류기 9종의 마이크로 분류 성능을 〈표 4〉와 〈그림 1〉에 제시하였다.

마이크로 분류 성능은 각 문헌마다 살펴보면서 할당된 디스크립터를 평가하는 입장에 가깝다고 할 수 있다. 분류 횟수 대비 성능을 기준으로 하는 마이크로 분류 성능은 다른 디스크립터에 비해 월등하게 많은 문헌에 부여된 'Reading Comprehension'이나 'Literacy'의 성능에 상당히 좌우된다. 마이크로 분류 성능 기준으로는 정확률은 Ridor가 가장 좋았고 재현율은 NB 분류기가 가장 좋았다. 정확률과 재현율을 같은 비율로 종합한 마이크로 F1 성능은 SVM이 가장 좋았으며 NB 분류기, VPT가 그 다음이었다. 정확률에 두 배 가중치를 두고 종합한 마이크로 F0.5 성능도 SVM이 가장 좋았으며 VPT가 2위, NB 분류기가 3위를 차지했다. 학습문헌과 검증문헌을 동일 데이터베이스의 문헌으로 구성했던 선행 연구인 김판준(2007)의 실험 결과에서는 학습문헌대 검증문헌 비율이 이 연구와 유사한 2:1인 경우에 SVM과 NB가 0.3대 중후반의 성능을 보이고 SVM이 NB보다 약간 더 나은 성능을 보인 것으로 보고되었다(김판준 2007, 그림 6). 본 연구의 실험 결과는 학습

〈표 4〉 분류기 9종의 마이크로 분류 성능 비교

	NB	SVM	VPT	RBF	KNN1	ADT10	J48	OneR	Ridor
마이크로 P	0.2573	0.3547	0.3497	0.2168	0.2766	0.4310	0.4762	0.1972	**0.6923**
마이크로 R	**0.6233**	0.4031	0.2203	0.0683	0.0859	0.1101	0.1101	0.0308	0.0198
마이크로 F1	0.3642	**0.3773**	0.2703	0.1039	0.1311	0.1754	0.1789	0.0533	0.0385
마이크로 F0.5	0.2915	**0.3634**	0.3129	0.1511	0.1916	0.2723	0.2860	0.0949	0.0889

(밑줄 친 부분은 해당 평가 기준의 최고 성능)



(원의 면적은 마이크로 F1에 비례)

〈그림 1〉 분류기 9종의 마이크로 분류 성능 비교

문헌과 검증문헌이 상이한 데이터베이스로 구성되었음에도 불구하고 거의 유사한 결과를 보였다. 다만 VPT의 경우 김판준(2007)의 연구에서보다 상당히 낮은 성능을 보인 것으로 나타났다.

마이크로 정확률과 마이크로 재현율을 가로축과 세로축으로 표현한 〈그림 1〉을 보면 SVM 분류기는 정확률과 재현율이 비교적 균형잡힌 성능을 보여주며, NB 분류기는 재현율이 월등하게 높은 성능을 보인다. 나머지 분류기는 모두 재현율보다 정확률이 훨씬 높은 성능을 보인다. 마이크로 정확률이 가장 좋았던 Ridor는 마이크로 재현율이 최하위로 나타났고 마이크로 F1과 F0.5도 가장 낮게 나타났는데, 이는 극소수의 문헌에만 디스크립터를 할당하였기 때문이다. 이로 미루어볼 때 마이크로 분류 성능을 기준으로 보면 균형잡인 성능을 얻기 위해서는 SVM이 가장 바람직하며, 부적합 문헌이 다수 포함되더라도 디스크립터와 관련된 문헌을 가급적 많이 찾을 수 있게 하려면 NB 분류기도 좋은 선택임을 알 수 있다. 반면에 각 문헌에 관계없는 디스크립터가 할당되는 것을 최대한 방지하고 비교적 정확한 디스크립터만 할당하기를 원할 때에는 VPT도 SVM에 버금가게 좋은 분류기인 것으로 나타났다.

3.2 매크로 분류 성능 평가

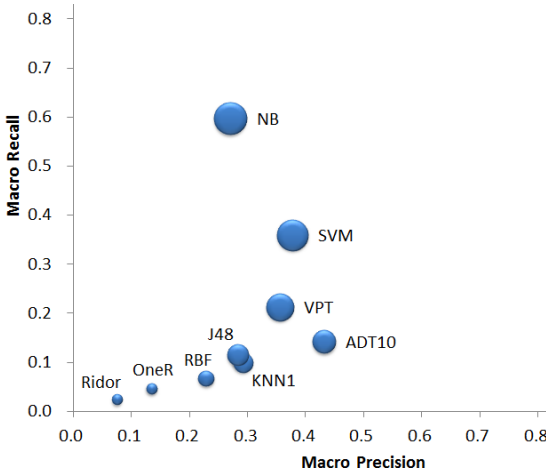
소속 문헌이 많은 디스크립터의 성능에 좌우되는 마이크로 분류 성능과 달리 매크로 분류 성능은 각 디스크립터의 분류 성능을 균등하게 평균하여 산출하는 것이므로 문헌을 탐색하는 입장에서 각 탐색 행위별로 평가하는 입장에 가깝다고 할 수 있다. 9종 분류기의 1차 실험 결과 분류기 9종의 매크로 분류 성능은 〈표 5〉와 〈그림 2〉에 제시하였다.

매크로 정확률은 ADT10이 가장 좋았고 매크로 재현율은 마이크로 재현율과 마찬가지로 NB 분류기가 가장 좋았다. 정확률과 재현율을 같은 비율로 종합한 매크로 F1 성능도 NB 분류기가 가장 좋았으며 SVM과 VPT가 그 다음이었다. 정확률에 두 배 가중치를 두고 종합한 매크로 F0.5 성능에서는 마이크로 F0.5의 경우와 마찬가지로 SVM이 가장 좋았으며 NB 분류기와 VPT가 근소한 차이로 2위와 3위를 차지했다. 매크로 정확률이 가장 좋은 ADT10은 매크로 재현율이 매우 낮아서 F1과 F0.5 성능에서는 4위로 밀려났다. 이는 ADT10이 소극적으로 적은 수의 문헌에만 디스크립터를 할당했기 때문에 높은 정확률에 비해서 낮은 재현율을 보인 것으로 해석할 수 있다.

〈표 5〉 분류기 9종의 매크로 분류 성능 비교

	NB	SVM	VPT	RBF	KNN1	ADT10	J48	OneR	Ridor
매크로 P	0.2698	0.3774	0.3555	0.2284	0.2920	**0.4319**	0.2835	0.1349	0.0750
매크로 R	**0.5977**	0.3598	0.2124	0.0678	0.0993	0.1418	0.1150	0.0465	0.0243
매크로 F1	**0.3537**	0.3252	0.2470	0.0824	0.1252	0.1796	0.1486	0.0417	0.0367
매크로 F0.5	0.2963	**0.3327**	0.2929	0.1058	0.1586	0.2453	0.2023	0.0578	0.0529

(밑줄 친 부분은 해당 평가 기준의 최고 성능)



(원의 면적은 매크로 F1에 비례)

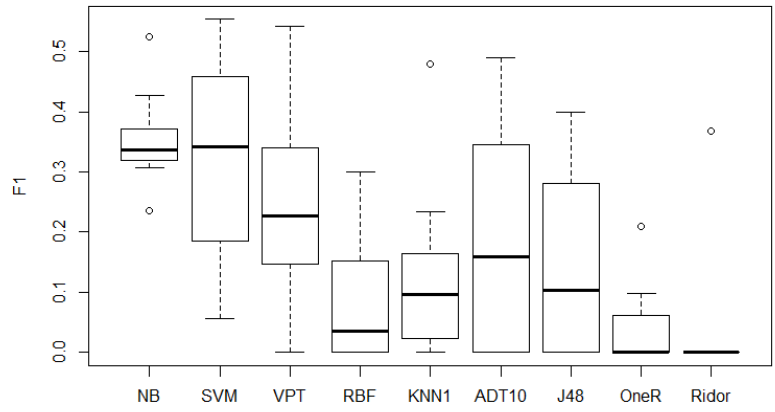
〈그림 2〉 분류기 9종의 매크로 분류 성능 비교

매크로 정확률과 매크로 재현율을 가로축과 세로축으로 표현한 〈그림 2〉를 보면 〈그림 1〉의 마이크로 분류 성능에서와 마찬가지로 SVM 분류기는 정확률과 재현율이 비교적 균형잡힌 성능을 보여주며, NB 분류기는 재현율이 월등하게 높은 성능을 보인다. 나머지 분류기는 ADT10과 마찬가지로 모두 재현율보다 정확률이 훨씬 높은 성능을 보이므로 디스크립터 할당을 매우 적게 하는 소극적인 분류기인 것으로 판단된다.

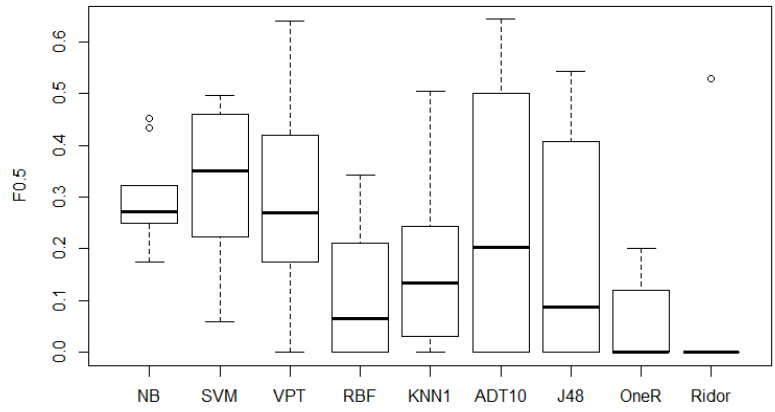
각 디스크립터에 대해 측정한 분류 성능의 평균에 해당하는 매크로 분류 성능을 기준으로 하더라도 균형잡인 성능을 얻기 위해서는 마이크로 분류 성능을 기준으로 살펴보았을 때와 마찬가지로 SVM이 가장 바람직한 것으로 나타났다. 부적합 문헌이 다수 포함되더라도 디스크립터와 관련된 문헌을 가급적 많이 찾을 수 있게 하려면 역시 NB 분류기도 좋은 선택임을 알 수 있다. 반면에 각 문헌에 관계없는 디스크립터가 할당되는 것을 최대한 방지하고 비교적 정확한 디스크립터만 할당하기를 원할 때에는 VPT나 ADT10도 고려할 수 있다.

10종 디스크립터에 대한 평균 성능으로 각 분류기의 성능을 비교하는 것 이외에, 각 디스크립터에 대한 성능이 어떻게 분포되었는가를 살펴보는 것도 필요하다. 평균 성능이 아무리 좋더라도 각 디스크립터마다의 성능 편차가 심해서 일부 디스크립터에 대해서는 매우 나쁜 성능을 보이는 분류기라면 실제로 사용하기에는 적절하지 않기 때문이다.

〈그림 3〉과 〈그림 4〉는 10개 디스크립터에 대한 9종 분류기의 F1 성능과 F0.5 성능 분포를 상자 도표로 그린 것이다. 이를 보면 굵은 실선으로 표시된 중앙값이 0에 해당하여 매우 저조한 성능을 보이는 OneR과 Ridor는 고려 대상에서 제외해야 한다. 중앙값이 0보다 큰 나머지 7종의 분류기 중에서 상자의 몸통, 즉 50%의 값이 집중 분포된 영역이 가장 작은 분류기는 NB인 것으로 나타났다. 이는 NB 분류기가 각 디스크립터를 할당하는 성능이 가장 안정적임



〈그림 3〉 10개 디스크립터에 대한 9종 분류기의 F1 성능 분포



〈그림 4〉 10개 디스크립터에 대한 9종 분류기의 F0.5 성능 분포

을 의미한다. SVM 분류기는 최고값인 성능 분포의 윗부분이 NB 분류기보다 높고 중앙값도 F0.5에서는 뚜렷하게 높게 나타나서 더 좋은 성능을 기대할 수 있지만, 최하값인 아랫 수염이 매우 낮으므로 디스크립터에 따른 성능 편차가 심하다는 것을 알 수 있다. 매크로 분류 성능 평균에서 높은 정확률 덕분에 F1과 F0.5 기준으로 3위와 4위를 차지한 VPT와 ADT10은 최고 성능과 최저 성능의 차이가 매우 크고 분류 성능이 0인 경우도 있을 정도로 기복이 심한 것으로 나타났다.

전체적으로 매크로 분류 성능 평균과 분포를 고려하였을 때 SVM 분류기는 정확률과 재현율이 고르게 높게 나타났으며, NB 분류기는 재현율이 두드러지게 높으면서 안정적인 성능을 보였다. 따라서 문헌마다 부여된 디스크립터를 살펴보는 경우와 같이 정확률이 중요하다면 SVM을 선택해야 하고, 특정 주제에 관한 문헌을 망

라적으로 찾기 위해 재현율도 정확률만큼 중요하게 고려한다면 NB 분류기를 선택하는 것이 바람직하다. 이밖에 VPT도 정확률을 중요하게 보는 관점인 F0.5 평균 기준으로는 NB 분류기에 가까운 좋은 성능을 보였고 디스크립터별 F0.5 성능의 최고점은 가장 높았지만, 디스크립터마다 성능 편차가 심하다는 단점이 두드러졌다.

4. 분류기 결합 실험

4.1 분류기 결합 전략

단일 분류기를 사용한 3장의 실험에서는 SVM 분류기와 NB 분류기가 좋은 성능을 보였으며 VPT도 정확률 면에서 가능성이 있는 것으로 나타났다. 그러나 SVM 분류기는 디스크립터별 분류 성능의 편차가 다소 크게 나타났으며, NB 분류기는 정확률이 낮다는 단점이 있었다. 따라서 자동분류의 성능 향상을 위한 여러 전략 중에서 별도의 추가 자원을 사용하지 않고 실험에 사용된 여러 분류기를 그대로 활용하는 투표 방식의 분류기 결합을 시도해보았다.

이 연구에서 시도한 결합 분류기는 사전 실험을 거쳐 〈표 6〉과 같은 다섯 종류를 채택하였다. Any_TOP2는 가장 성능이 좋은 NB와 SVM 분류기를 결합하여 둘 중 하나라도 특정 문헌에 디스크립터를 할당하면 최종 할당으로 판정하는 결합 분류기이다. Any_TOP3는 개별 성능 3위 이내에 속하는 NB, SVM, VPT의 세 분류기 중 하나라도 특정 문헌에 디스크립터를 할당하면 최종 할당으로 판정하는 결합 분류기이다. Any_MID5는 최상위 성능을 가진 두 분류기를 제외하고 중간 성능의 분류기 5종(VPT, RBF, KNN1, ADT10, J48)을 결합하여 이중 하나라도 특정 문헌에 디스크립터를 할당하면 최종 할당으로 판정하는 결합 분류기이다. Two_TOP3는 개별 성능 3위 이내에 속하는 NB, SVM, VPT의 세 분류기 중 둘 이상의 분류기가 공통으로 특정 문헌에 디스크립터를 할당하면 최종 할당으로 판정하는 결합 분류기이다. Two_ALL은 전체 9종 분류기 중 둘 이상의 분류기가 공통으로 특정 문헌에 디스크립터를 할당하면 최종 할당으로 판정하는 결합 분류기이다.

〈표 6〉 2차 실험에서 시도한 결합 분류기 5종

결합된 분류기	약칭	설명
{NB+SVM}	Any_TOP2	• 1, 2위 성능 분류기 2종 결합 • 둘 중 하나라도 해당 문헌에 디스크립터를 할당하면 최종 할당
{NB+SVM+VPT}	Any_TOP3	• 1, 2, 3위 성능 분류기 3종 결합 • 셋 중 하나라도 해당 문헌에 디스크립터를 할당하면 최종 할당
{VPT+RBF+KNN1+ADT10+J48}	Any_MID5	• 중간 성능 분류기 5종 결합 • 5종 중 하나라도 해당 문헌에 디스크립터를 할당하면 최종 할당
2 of {NB+SVM+VPT}	Two_TOP3	• 1, 2, 3위 성능 분류기 3종 결합 • 셋 중 둘 이상이 해당 문헌에 디스크립터를 할당하면 최종 할당
2 of {ALL}	Two_ALL	• 전체 분류기 9종 결합 • 9종 분류기 중 둘 이상이 해당 문헌에 디스크립터를 할당하면 최종 할당

4.2 분류기 결합 실험 결과

결합 분류기 5종의 성능은 단일 분류기 중에서 성능이 가장 좋은 NB 분류기, SVM, VPT와 비교하여 살펴보았다. 단일 분류기 최상위 3종과 결합 분류기 5종의 마이크로 평균 성능은 〈표 7〉에 제시하였고, 매크로 평균 성능은 〈표 8〉에 제시하였다. 마이크로 평균 성능과 매크로 평균 성능 양쪽 모두 가장 성능이 좋은 분류기는 일치하게 나타났으며 대체로 유사한 순위를 보여준다. 따라서 정확률과 재현율을 각각 가로축과 세로축으로 반영하는 그림은 〈그림 5〉와 같이 매크로 평균 성능만 제시하였다. 정확률의 경우 최상위 분류기 3종의 판정을 다수결로 반영하는 Two_TOP3가 가장 좋아서 단일 분류기 중 가장 좋은 SVM 분류기의 정확률과 비교했을 때 마이크로 정확률은 27.0% 향상되었으며 매크로 정확률은 16.4% 향상되었다. 재현율의 경우 최상위 3종 분류기 중 1종이라도 디스크립터 할당으로 판정하는 경우에 인정하는 Any_TOP3가 가장 좋았으며 단일 분류기 중 가장 재현율이 좋은 NB 분류기의 성능 대비 마이크로 재현율은 42.0%, 매크로 재현율은 43.0% 향상되었다. 종합 성능인 F1은 최상위 2종 분류기인 NB와 SVM을 결합한 Any_TOP2가 가장 좋았으며 단일 분류기 중 마이크로 F1이 가장 좋은 SVM에 비해서는 7.9% 향상되었고 매크로 F1이 가장 좋은 NB 분류기에 비해서는 13.2% 향상되었다. 정확률에 두 배의 가중치를 두는 F0.5 기준으로는 모든 분류기의 판정을 결합한 Two_ALL이 가장 좋은 성능을 보였으며 단일 분류기 중 성능이 가장 좋은 SVM과 비교했을 때 마이크로 F0.5는 11.6% 향상되었고 매크로 F0.5는 11.0% 향상되었다. 중간 성능의 분류기 5종을 결합한 Any_MID5는 정확률과 재현율 양 측면 모두 단일 분류기의 최고 성능에는 미치지 못하였다.

정확률과 재현율 면에서 각각 단일 분류기 최고 성능보다 정확률이 향상된 결합 분류기를 보면 일정한 패턴이 있는 것이 확인된다. 정확률 면에서 단일 분류기보다 성능이 향상된 결합 분류기는 Two_TOP3와 Two_ALL로서 셋 이상의 분류기 중에서 둘 이상이 긍정 판정을 하는 경우에 디스크립터를 할당하는 방식의 결합이 정확률을 향상시키는데 효과적인 것으로 나타났다. 재현율 면에서 단일 분류기보다 성능이 향

〈표 7〉 단일 분류기 성능 최상위 3종과 결합 분류기의 마이크로 평균 성능 비교

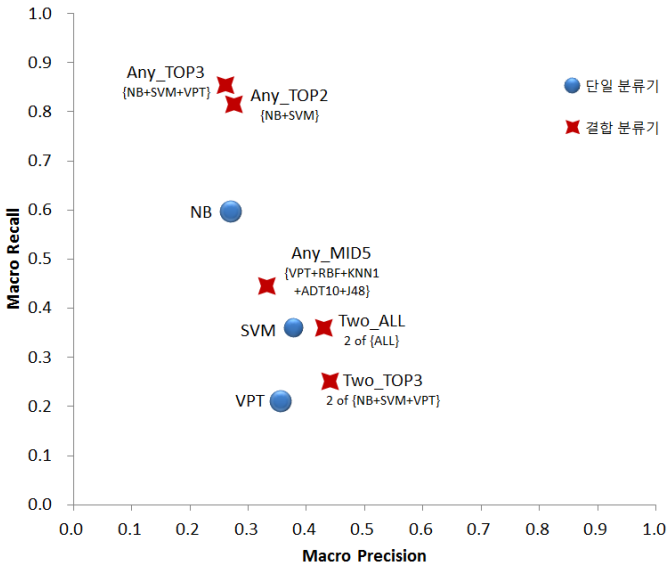
분류기		마이크로 P	마이크로 R	마이크로 F1	마이크로 F0.5
단일 분류기	NB	0.2573	0.6233	0.3642	0.2915
	SVM	0.3547	0.4031	0.3773	0.3634
	VPT	0.3497	0.2203	0.2703	0.3129
결합 분류기	Any_TOP2	0.2674	**0.8524**	**0.4072**	0.3100
	Any_TOP3	0.2575	**0.8855**	**0.3990**	0.3001
	Any_MID5	0.3082	0.4317	0.3596	0.3269
	Two_TOP3	**0.4504**	0.2797	0.3451	**0.4014**
	Two_ALL	**0.4103**	0.3877	0.3986	**0.4055**

(밑줄 친 부분은 해당 평가 기준의 단일 분류기 최고 성능보다 좋은 경우)

〈표 8〉 단일 분류기 성능 최상위 3종과 결합 분류기의 매크로 평균 성능 비교

분류기		매크로 P	매크로 R	매크로 F1	매크로 F0.5
단일 분류기	NB	0.2698	0.5977	0.3537	0.2963
	SVM	0.3774	0.3598	0.3252	0.3327
	VPT	0.3555	0.2124	0.2470	0.2929
결합 분류기	Any_TOP2	0.2752	**0.8164**	**0.4006**	0.3139
	Any_TOP3	0.2604	**0.8549**	**0.3897**	0.2998
	Any_MID5	0.3318	0.4459	0.3475	0.3245
	Two_TOP3	**0.4391**	0.2519	0.2917	**0.3412**
	Two_ALL	**0.4293**	0.3603	0.3488	**0.3691**

(밑줄 친 부분은 해당 평가 기준의 단일 분류기 최고 성능보다 좋은 경우)



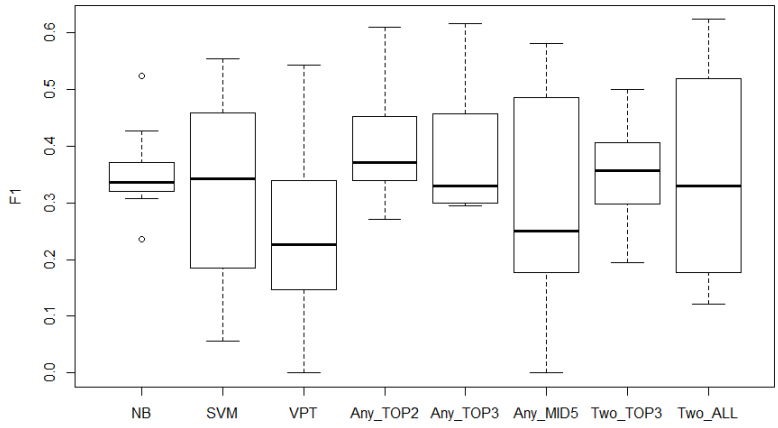
〈그림 5〉 단일 분류기와 결합 분류기의 매크로 정확률과 매크로 재현율 비교

상된 결합 분류기는 Any_TOP3와 Any_TOP2로서, 최상위 성능을 보이는 복수의 분류기 중 하나라도 긍정 판정을 하면 디스크립터를 할당하는 방식의 결합이 재현율을 향상시키는데 효과적인 것으로 나타났다. 결국 정확률을 향상시키기에는 셋 이상의 분류기 중 둘 이상의 긍정 판정에 따르는 보수적인 'Two_' 방식의 결합 분류기가 유용하며, 재현율을 향상시키기에는 좋은 성능을 가진 복수의 단일 분류기 중 하나 이상의 긍정 판정에 따르는 적극적인 'Any_' 방식의 결합 분류기가 유용하였다.

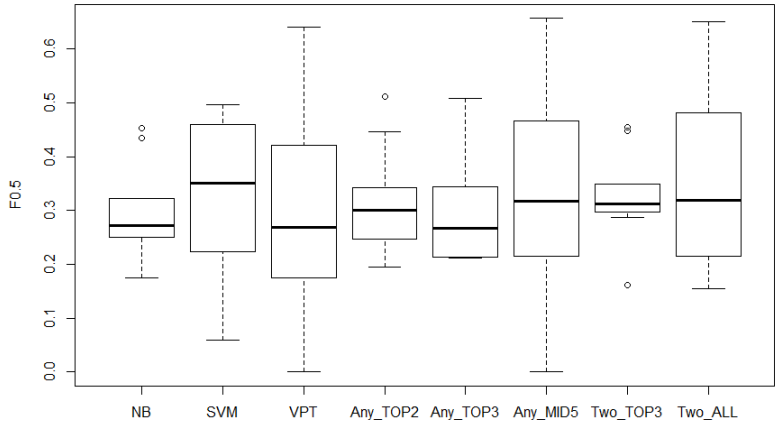
종합 성능인 F척도를 살펴보면 Two_TOP3와 Two_ALL은 F1 기준으로 단일 분류기 최고 성능보다 좋은 결과를 보였고, 적극적인 'Any_' 방식의 결합 분류기인 Any_TOP2와 Any_TOP3는 정확률을 더 강조하는 F0.5 기

준에서 단일 분류기 최고 성능보다 좋은 결과를 보였다. 분류기를 결합하는 방식 중에서 보수적인 'Two_' 방식의 결합은 재현율을 유지하면서 정확률을 향상시키는데 효과가 있고, 적극적인 'Any_' 방식의 결합은 정확률을 유지하면서 재현율을 향상시키는데 효과가 있기 때문에 종합 성능인 F척도에서 이런 성향이 나타난 것으로 해석된다.

결합 분류기의 평균 성능 이외에 각 디스크립터에 대한 성능이 어떻게 분포되었는가를 살펴보기 위해서 10종 디스크립터에 대한 F1 성능 분포를 〈그림 6〉에, F0.5 성능 분포를 〈그림 7〉에 상자 도표로 제시하였다. 앞의 〈표 8〉에서 매크로 F1 평균은 결합 분류기인 Any_TOP2와



〈그림 6〉 10개 디스크립터에 대한 최상위 단일 분류기 3종과 결합 분류기 5종의 F1 분포



〈그림 7〉 10개 디스크립터에 대한 최상위 단일 분류기 3종과 결합 분류기 5종의 F0.5 성능 분포

Any_TOP3가 가장 좋았는데, 〈그림 6〉의 분포를 보면 두 결합 분류기의 상자 길이가 NB 분류기보다는 길지만 SVM보다는 짧게 나타나서 디스크립터별 성능 편차가 심하지 않음을 알 수 있다. 정확률을 강조한 매크로 F0.5 평균이 가장 좋았던 Two_ALL과 두 번째로 좋았던 Two_TOP3는 〈그림 7〉에서 보듯이 디스크립터별 F0.5 성능 편차가 다르게 나타났다. 9종 분류기를 모두 결합하는 Two_ALL은 가장 좋은 평균 성능을 보였지만 상자의 길이가 매우 길어서 성능 편차가 큰 것으로 나타났는데, 세 분류기만 결합한 Two_TOP3는 성능 편차가 상대적으로 작게 나타났다. 대체로 결합에 동원된 분류기의 수가 많을수록 성능의 편차가 큼을 알 수 있다. 〈그림 6〉과 〈그림 7〉을 보면 분류기를 둘이나 셋 결합한 경우보다 다섯 혹은 아홉 개 전체를 결합한 경우에 최고 성능과 최저 성능의 차이가 크고 상자의 길이도 길어서 디스크립터별 성능 차이가 크다. 따라서 안정적인 성능을 얻으려면 너무 많은 분류기를 결합하는 것은 좋지 않은 전략이라고 할 수 있다.

전체적으로 매크로 분류 성능 평균과 분포를 고려하였을 때 재현율을 저하시키지 않으면서 정확률을 향상시키려면 두 종 이상의 분류기가 긍정 판정하는 경우에 디스크립터를 할당하는 Two_ALL이나 Two_TOP3를 사용하는 것이 바람직하며, 이중에서 Two_TOP3가 더 안정적인 성능을 보였다. 반면에 정확률을 저하시키지 않으면서 재현율을 향상시키기 위해서는 Any_TOP2나 Any_TOP3를 사용할 수 있으며, 이중에서도 더 높은 F1과 F0.5 성능을 보이는 Any_TOP2를 사용하는 것이 바람직하다.

5. 결 론

최근 국내 학술 데이터베이스는 공공이나 민간에서 양적으로 크게 발전하고 있다. 데이터베이스의 유형도 다양하게 분화하여 복수의 국내 연구데이터 아카이브를 비교하거나(신영란, 정연경 2012) 국내 인용색인 데이터베이스를 비교하는 연구(박상근 2013)도 수행되고 있다. 그러나 가장 기본이 되는 학술논문 데이터베이스는 양적인 성장에도 불구하고 통제어휘색인과 같은 핵심 요소를 아직까지 제공하지 못하고 있다. 이런 한계를 극복하기 위해서 본 연구에서는 해외 학술데이터베이스로부터 통제어휘색인 정보를 학습하여 국내 학술논문에 통제언어 색인어로서 디스크립터를 부여하는 실험을 수행해 보았다. 또한, 다양한 분류기와 이들 분류기의 결합을 통하여 이러한 디스크립터 자동 할당의 성능을 향상하는 방안을 모색하였다.

실험집합은 독서 영역을 대상으로 국외 학술 데이터베이스인 LISTA로부터 검색된 1,809건의 영어 학술 논문과 10개 디스크립터를 학습집합으로 하고, 국내 학술 데이터베이스인 RISS에서 검색된 798건의 영문 제목과 영문 초록을 검증 집합으로 하였다. WEKA에서 제공하는 9종 분류기를 이용한 1차 실험에서는 SVM 분류기가 정확률과 재현율이 고르게 높게 나타났으며, NB 분류기는 재현율이 두드러지게 높으면서 안정적인 성능을 보였다. 따라서 문헌마다 부여된 디스크립터를 살펴보는 경우와 같이 정확률이 중요하다면 SVM을 선택해야 하고, 특정 주제에 관한 문헌을 망라적으로 찾기 위해 재현율도 정확률만큼 중요하게 고려한다면 NB 분류기를 선택하는 것이 더 바람직하다.

투표 방식의 분류기 결합을 통해 성능 향상을 시도한 2차 실험에서는 매크로 정확률과 매크로 재현율이 각각 최고 16.4%와 43.0% 향상되는 효과를 얻었다. 정확률을 향상시키기에는 셋 이상의 분류기 중 둘 이상의 긍정 판정에 따르는 보수적인 방식의 결합 분류기가 유용했으며, 재현율을 향상시키기에는 좋은 성능을 가진 복수의 단일 분류기 중 하나 이상의 긍정 판정에 따르는 적극적인 방식의 결합 분류기가 유용하였다. 종합 성능인 F척도를 살펴보면 보수적인 방식의 Two_TOP3와 Two_ALL은 F1 기준으로 단일 분류기 최고 성능보다 좋은 결과를 보였고, 적극적인 방식의 Any_TOP2와 Any_TOP3는 정확률을 더 강조하는 F0.5 기준에서 단일 분류기 최고 성능보다 좋은 결과를 보였다. 이를 통해 종합성능인 매크로 F1은 13.2%, 매크로 F0.5는 11.0% 향상시킬 수 있었다.

실험 결과에서 단일 분류기인 NB 분류기를 이용할 경우에 0.6 이상의 매크로 재현율을 얻었고, 보수적인 결합 분류기인 Any_TOP2를 이용할 경우 정확률 저하 없이 0.8 이상의 매우 높은 재현율을 얻었다. 그러나 정확률의 경우에는 단일 분류기 중 가장 좋은 SVM의 매크로 정확률이 0.3774에 머물렀고 결합 분류기인 Two_ALL을 사용하더라도 0.4293 정도로 나타나서 절반을 넘기기가 어려웠다. 물론 통제색인어휘의 사용이 재현율을 향상시키는 것에 주된 목적이 있으므로 이와 같은 결과가 기대에 크게 어긋나는 것은 아니지만, 추후에는 정확률의 향상을 도모할 수 있는 다른 전략의 개발도 필요할 것이다.

참 고 문 헌

[1] 김용환, 정영미. 2012. 위키피디아를 이용한 분류자질 선정에 관한 연구.『정보관리학회지』, 29(2): 155-171.

[2] 김판준. 2006a. 기계학습을 통한 디스크립터 자동부여에 관한 연구.『정보관리학회지』, 23(1): 279-299.

[3] 김판준. 2006b. 로치오 알고리즘을 이용한 학술지 논문의 디스크립터 자동부여에 관한 연구.『정보관리학회지』, 23(3): 69-90.

[4] 김판준. 2008. 용어 가중치부여 방법을 이용한 로치오 분류기의 성능 향상에 관한 연구.『정보관리학회지』, 25(1), 211-233.

[5] 김판준, 이재윤. 2007. 문헌간 유사도를 이용한 자동분류에서 미분류 문헌의 활용에 관한 연구.『정보관리학회지』, 24(1): 251-271.

[6] 김판준, 이재윤. 2012. 디스크립터 자동 할당을 위한 저자키워드의 재분류에 관한 실험적 연구.『정보관리학회지』, 29(2): 225-246.

[7] 박상근. 2013. 인문학 분야의 인용 데이터정보원 비교 분석: 네이버 전문정보, KCI.『정보관리학회지』, 30(1): 33-50.

[8] 송성전, 정영미. 2012. 용어의 문맥활용을 통한 문헌 자동 분류의 성능 향상에 관한 연구.『정보관리학회지』, 29(2): 205-224.

[9] 신영란, 정연경. 2012. 국내 인문사회 연구데이터 아카이브의 개선방안에 관한 연구.『한국기록관리학회지』, 12(3): 93-115.

[10] 유호현, 정영미. 2008. 분류기 조합을 통한 신경망 분류기의 성능 향상 실험.『제15회 한국정보관리학회 학술대회 논문집』, 207-214.

[11] 이용구. 2009. 기계번역을 이용한 교차언어 문서 범주화의 분류 성능 분석.『한국문헌정보학회지』, 43(1): 313-332.

[12] 이재윤. 2005. 자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 대한 연구.『한국문헌정보학회지』, 39(2): 123-146.

[13] 정영미. 2012.『정보검색연구』. 증보판. 서울: 연세대학교 출판문화원.

[14] 정은경. 2009. 문서범주화 성능 향상을 위한 의미기반 자질확장에 관한 연구.『정보관리학회지』, 26(3): 261-278.

[15] Amini, B. M. and Goutte, C. 2010. "A Co-classification Approach to Learning from Multilingual Corpora." *Machine Learning*, 79: 105-121.

[16] Bel, N., Koster, C. H. A. and Villegas, M. (2003). "Cross-lingual Text Categorization." In *Proceedings of the 7th European Conference on Research and Advanced Technology for Digital Libraries*, 126-139.

[17] Chan, L. M. 2000. "Exploiting LCSH, LCC and DDC to Retrieve Networked Resources: Issues and Challenges." In *Proceedings of the Bicentennial Conference on Bibliographic Control for the New Millennium*, sponsored by the Library of Congress Cataloging Directorate. Retrived from http://www.loc.gov/catdir/bibcontrol/chan_paper.html

[18] Gross, T. and Taylor, A. G. 2005. "What Have We Got to Lose? The Effect of Controlled Vocabulary on Keyword Searching Results." *College and Research Libraries*, 66(3): 212-230.

[19] Kipp, M. E. I. 2005. "Complementary or Discrete Contexts in Online Indexing: A Comparison of User, Creator, and Intermediary Keywords." *Canadian Journal of Information and Library Science*, 29(4): 419-436.

[20] McCutcheon, S. 2009. "Keyword vs Controlled Vocabulary Searching: The One with the Most Tools Wins." *Indexer*, 27(2): 62-65.

[21] Olsson, J. S., Oard, D. W. and Hajic, J. 2005. Cross-language Text Classification. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 645-646.

[22] Rigutini, L., Maggini, M. and Liu, B. 2005. "An EM Based Training Algorithm for Cross-language

Text Categorization." In *Proceedings of the 2005 IEEE/WIC/ACM International Conference on Web Intelligence*, 529-535.

[23] Rowley, J. 1994. "The Controlled versus Natural Indexing Languages Debate Revisited: A Perspective on Information Retrieval Practice and Research." *Journal of Information Science*, 20(2): 108-119.

[24] Tillotson, J. 1995. "Is Keyword Searching the Answer?" *College & Research Libraries*, 56: 199-206.

[25] Voorbij, H. J. 1998. "Title Keywords and Subject Descriptors: A Comparison of Subject Search Entries of Books in the Humanities and Social Sciences." *Journal of Documentation*, 54(4): 466-476.

[26] Wei, Chih-Ping et al. 2014. "Exploiting Poly-Lingual Documents for Improving Text Categorization Effectiveness." *Decision Support Systems*, 57: 64-76.

[27] Wei, Chih-Ping, Lin, Yen-Ting and Yang, C. C. 2011. "Cross-lingual Text Categorization: Conquering Language Boundaries in Globalized Environments." *Information Processing and Management*, 47: 786-804.

[28] Wu, Y. and Oard, D. W. 2008. "Bilingual Topic Aspect Classification with a Few Training Examples." In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 203-210.

[29] Witten, I. H., Frank, E. and Hall, M. A. 2011. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd ed. Burlington, MA: Morgan Kaufmann.

• 국문 참고자료의 영어 표기
(English translation / romanization of references originally written in Korean)

[1] Kim, Yong-Hwan and 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.

[2] 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.

[3] Kim, Pan Jun. 2006b. "A Study on the Automatic Descriptor Assignment for Scientific Journal Articles Using Rocchio Algorithm." *Journal of the Korean Society for Information Management*, 23(3): 69-90.

[4] 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.

[5] Kim, Pan Jun and 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.

[6] Kim, Pan Jun and 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.

[7] Park, Sang-Keun. 2013. "A Comparative Analysis of the Humanities Citation Tools: NAVER Scholar and KCI." *Journal of the Korean Society for Information Management*, 30(1): 33-50.

[8] Song, Sung-Jeon and 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.

[9] Shin, Young-Ran and Chung, Yeon-Kyoung. 2012. "A Study on the Improvement Plans of the Humanities and Social Sciences Research Data Archives in Korea." *Journal of Records Management & Archives Society of Korea*, 12(3): 93-115.

[10] Ryu, Hohyun and Chung, Young-Mee. 2008. "Combining Classifiers to Improved the Performance of a Neural Network Classifier." In *Proceedings of the 15th Conference of the Korean Society for Information Management*, 207-214.

[11] 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.

[12] Lee, Jae Yun. 2005. "Empirical Study on Improving the Performance of Text Categorization Considering the Relationships between Feature Selection Criteria and Weighting Methods." *Journal of the Korean Library and Information Science Society*, 39(2): 123-146.

[13] Chung, Young-Mee. 2012. *Information Retrieval Research*. 2nd ed. Seoul: Yonsei University Publishing.

[14] 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.