디스크립터 자동 할당을 위한 저자키워드의 재분류에 관한 실험적 연구

A Study on the Reclassification of Author Keywords for Automatic Assignment of Descriptors

김판준(Pan Jun Kim)*
이재윤(Jae Yun Lee)**

초 록

본 연구는 국내 주요 학술 DB의 검색서비스에서 제공되고 있는 저자키워드(비통제키워드)의 재분류를 통하여 디스크립터(통제키워드)를 자동 할당할 수 있는 가능성을 모색하였다. 먼저 기계학습에 기반한 주요 분류기들의 특성을 비교하는 실험을 수행하여 재분류를 위한 최적 분류기와 파라미터를 선정하였다. 다음으로, 국내 독서 분야 학술지 논문들에 부여된 저자키워드를 학습한 결과에 따라 해당 논문들을 재분류함으로써 키워드를 추가로 할당하는 실험을 수행하였다. 또한 이러한 재분류 결과에 따라 새롭게 추가된 문헌들에 대하여 통제키워드인 디스크립터와 마찬가지로 동일 주제의 논문들을 모아주는 어휘통제 효과가 있는지를 살펴보았다. 그 결과, 저자키워드의 재분류를 통하여 디스크립터를 자동 할당하는 효과를 얻을 수 있음을 확인하였다.

ABSTRACT

This study purported to investigate the possibility of automatic descriptor assignment using the reclassification of author keywords in domestic scholarly databases. In the first stage, we selected optimal classifiers and parameters for the reclassification by comparing the characteristics of machine learning classifiers. In the next stage, learning the author keywords that were assigned to the selected articles on readings, the author keywords were automatically added to another set of relevant articles. We examined whether the author keyword reclassifications had the effect of vocabulary control just as descriptors collocate the documents on the same topic. The results showed the author keyword reclassification had the capability of the automatic descriptor assignment.

키워드: 자동분류, 텍스트 범주화, 재분류, 어휘통제, 디스크립터, 저자키워드
automatic classification, text categorization, reclassification, vocabulary control, descriptors, author keywords

* 신라대학교 문헌정보학과 전임강사(pjkim@silla.ac.kr) (제1저자)
** 경기대학교 문헌정보학과 부교수(memexlee@kgu.ac.kr) (교신저자)

■ 논문접수일자 : 2012년 6월 2일 ■ 최초심사일자 : 2012년 6월 2일 ■ 게재확정일자 : 2012년 6월 21일
■ 정보관리학회지. 29(2). 225-246, 2012. 〔http://dx.doi.org/10.3743/KOSIM.2012.29.2.225〕

1. 서 론

현재 대부분의 국내 학술 데이터베이스 검색에서 색인전문가 및 통제어휘표의 부재라는 근본적인 문제로 인하여 학술지 논문에 대한 통제키워드(디스크립터, 주제명 등)가 전혀 제공되지 않고 있는 상황이다. 일반적으로 자연언어 색인을 채택한 검색 시스템에서는 동일한 개념을 표현하는 모든 동의어들과 단수/복수 명사 등 형태가 다른 용어들을 다 검색어로 사용해야만 적합한 정보자료를 모두 검색할 수 있다. 반면 통제언어 색인은 특정한 개념은 항상 같은 용어에 의해 색인이 가능하므로 검색 시 이 용어를 검색어로 사용하면 관련된 모든 정보자료를 검색할 수 있다. Dialog, ProQuest, CSA Illumina 등 대부분의 온라인 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, p. 38).

인터넷의 발전으로 누구나 대규모의 정보에 언제라도 접근할 수 있는 현재의 환경과는 달리, 비교적 소수의 전문가 또는 학자들 간에 정보가 유통되었던 초기의 온라인 데이터베이스 환경에서는 수록된 정보의 유형에 따라 비통제어휘(자연언어)와 통제어휘를 사용하는 색인은 각기 다른 측면에서 발전되었다. 첫째로 미국 국립의학도서관(National Library of Medicine), 국방부(Department of Defense), 미 항공우주국(National Aeronautics an Space Administration)과 같은 정부기관이 주도한 서지 데이터베이스를 중심으로 한 색인전문가에 의한 통제언어 색인의 발전과, 둘째로 법률분야 등의 민간부문에서 주도한 전문 데이터베이스를 중심으로 컴퓨터에 의해 본문의 단어를 그대로 추출하는 비통제언어 색인의 발전이 그것이다(Lancaster, 2003). 1960년대와 1970년대에 걸친 이러한 두 가지 색인언어 간의 우위에 대한 논의는, 1980년대에 와서 양자를 함께 사용하는 복합 시스템이 가장 이상적이라는 합의에 이르러 현재 대부분의 정보검색시스템은 이러한 두 가지 유형의 색인을 상호보완적으로 제공하는 추세이다(김판준, 2006b). 따라서 학술정보서비스를 목적으로 하는 데이터베이스의 경우 색인작업은 크게 두 가지 경로로 이루어진다고 할 수 있다. 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(김판준, 2006a).

그러나 현재 국내 학술 데이터베이스들은 기계적으로 추출하는 키워드 이외에 논문 제출 시 저자가 직접 제시한 저자키워드를 자연언어 색인어로 제공하고 있을 뿐, 색인전문가에 의한 통제키워드로서 통제언어 색인어를 전혀 제공하지 않고 있다. 실제로 국내 학술 데이터베이스들에서 전체 필드를 대상으로 하여 '독서'를 탐색어로 검색한 결과를 살펴본 결과인 〈표 1〉에 따르면, 색인어 필드에서 제공하고 있는 색인어는 모두 저자키워드로서 자연언어 색인어에 해당한다.

자연언어 색인으로서 저자키워드는 동일한

〈표 1〉 국내 학술 데이터베이스들의 검색 결과 및 색인어 필드 특성: 독서

DB	URL	검색건수	색인어 필드명	비고
DBPIA	http://www.dbpia.co.kr	1685	키워드	저자키워드
RISS	http://www.riss4u.net	3368	주제어	저자키워드
KCI	http://www.kci.go.kr	1652	키워드	저자키워드
KISS	http://kiss.kstudy.com	888	주제키워드	저자키워드
NDSL	http://scholar.ndsl.kr	1569	주제어	저자키워드

개념을 저자에 따라 여러 개의 다른 용어로 표현하고 있는 경우가 많아, 동일한 개념을 표현하는 모든 용어들과 어형이 다른 용어들을 모두 검색어로 사용해야만 관련된 모든 정보자료의 검색이 가능한 문제가 있다. 예를 들면, 독서에 의한 인간형성이라는 의미로 흔히 혼용되고 있는 '독서교육'과 '독서지도'는 연구자들의 배경 및 성향에 따라 '독서교육', '독서지도', '讀書敎育', '讀書指導', 'reading education', 'reading guidance', 'reading instruction' 등 다양한 용어로 표현되어 저자키워드로 제공되고 있다.[1] 여기에 각 용어들의 영문 대소문자 및 띄어쓰기 이형까지 포함한다면 '독서교육' 또는 '독서지도'라는 거의 동일한 개념에 대한 표현 형식은 더욱 다양해진다. 따라서 이용자는 동일한 개념에 대한 서로 다른 용어들은 물론 관련 이형들까지 모두 검색어로 사용하는 경우에만 관련 자료를 모두 검색할 수 있다는 큰 부담을 갖게 된다.

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 통제언어 자동색인시스템에 대한 필요성이 커지고 있다. 그러나 대부분의 국내 학술 데이터베이스에서는 색인전문가 및 통제어휘표가 부재한 이유로 통제언어 색인어를 전혀 제공하지 못하고 있는 현실에서, 지금까지 색인전문가에 의해 전적으로 수행되어 온 이러한 역할을 일부 대체하거나 효과적으로 지원할 수 있는 방법을 적극적으로 모색할 필요가 있다. 이러한 측면에서 본 연구는 '독서' 분야 학술지 논문을 대상으로 저자가 직접 자신의 논문에 부여한 자연언어 색인어인 저자키워드를 활용하여 통제언어 색인어인 디스크립터를 자동할당할 수 있는 가능성을 기계학습 접근법을 중심으로 검토하여 보는 것을 목적으로 한다.

2. 이론적 배경

2.1 텍스트 범주화와 통제언어 자동색인

시소러스, 주제명표의 용어 또는 분류체계의 엔트리로서 주제를 표현하는 통제언어 색인어 또는 분류기호의 부여는 텍스트 범주화라고도 부른다(Moen, 2002). 이러한 텍스트 범주화(또는 문헌의 자동분류)는 사전 정의된 범주들

1) DBPIA 검색 결과에서 저자키워드 필드에 있는 용어들을 확인하여 제시한 것임. 이외에도 '독서치료'는 '독서요법', '독서치유', '讀書治療', '讀書療法', 'bibliotherapy' 등 다양한 용어로 표현되어 있음.

로 문헌을 분류하는 것을 목적으로 하고 있는데(Joachims, 1998), 여기서 사전 정의된 범주들을 문헌에 부여된 디스크립터(통제언어 색인어)로 본다면 통제언어에 기초한 자동색인과 거의 동일한 접근법을 사용할 수 있다. 또한 문헌에 대한 통제색인어의 자동부여는 "특정한 처리절차에 따라 시소러스 또는 주제명표와 같은 통제어휘집에서 색인어를 개개의 문헌에 자동적으로 부여하기 위한 것"(윤구호, 1999)으로, 이전에는 색인전문가에 의해 수작업으로 이루어졌던 색인작업을 보다 효율적으로 수행하기 위하여 기계적인 색인방법으로 대체 또는 지원하는 것이라 할 수 있다. 즉, 통제언어에 의한 자동색인은 통제어휘에서 용어를 사용하는 것을 의미하며, 문서 범주화와 동의어라고 할 수 있다(Sebastiani, 2002).

텍스트 범주화 또는 통제언어를 사용한 자동색인에 관한 연구는 1980년대의 지식공학적 접근법에서 1990년대 이후 기계학습 기반 접근법(Chung, Pottenger, & Schatz, 1998; Joachim, 1998; Khan, Baharudin, & Lee, 2010; Lauser & Hotho, 2003; Lewis, 1996; Ruiz & Srinivasan, 2002; Yang, 1997)으로 중심이 바뀌어 다양한 응용과 시도가 이루어지고 있다. 특히, 최근 몇 년 동안에는 Naive Bayes(NB), k-Nearest Neighbor(kNN), Support Vector Machine(SVM), Neural Networks(NN), Decision Tree(DT), Logistic Regression(LR), Rocchio' Algorithm(Rocchio), Fuzzy Correlation(FC), Genetic Algorithms(GA) 등 다양한 알고리즘과 이들을 서로 조합한 하이브리드 방식의 기계학습에 기초한 연구들이 빠른 성장과 확산의 양상을 보이고 있다(김판준, 2006a; 김판준, 2006b; 김판준, 2008; 이재윤, 2005a; 이재윤, 2005b; Aseervatham et al., 2011; Chen & Chen, 2011; Chen et al., 2011; Jiang et al., 2012; Kumar & Gopal, 2010; McCallum & Nigam, 2003; Miao & Kamel, 2011; Ûğuz, 2011; Vasuki & Cohen, 2010; Wang & Chiang, 2007; Wu, 2009; Yu, Xu, & Li, 2008; Zhang et al., 2011). 이외에도 기계학습 알고리즘과 전문가시스템의 조합(Li & Park, 2009; Villena-román et al., 2011), 미분류 문헌의 활용(김판준, 이재윤, 2007; Torii et al., 2011), WordNet이나 Wikipedia 등의 외부 정보의 활용(김용환, 정영미, 2012; 정은경, 2009) 등 텍스트 범주화 성능 향상을 위한 다양하고도 새로운 기법들이 지속적으로 개발 및 적용되고 있다.

최근 Hurt(2010)의 연구에서 저자키워드와 기계적으로 자동 생성된 키워드를 비교한 결과에 따르면, 양자 간에 통계학적으로 유의한 차이가 없는 것으로 나타났다. 따라서 현재 국내 대부분의 학술 DB에서 제공하고 있는 비통제 색인어로서 저자키워드는 통제 색인어의 대안이 될 수 없으며 텍스트에서 기계적으로 자동 추출된 키워드(자연언어 색인어)와 동일한 특성을 갖는다고 할 수 있다. 한편, Gil-Leiva와 Alonso-Arroyo(2007)의 연구에서는 4개 학술정보 데이터베이스(INSPEC, CAB Abstract, ISTA, LISA)를 대상으로 학술지 논문의 레코드에 포함된 저자키워드와 디스크립터를 비교하였다. 그 결과, 저자키워드의 상당수(약 46%)가 디스크립터에 포함되어 있는 것으로 나타났으며, 그 중 약 25%는 양자가 그 자체로 동일한 형태였고 나머지 21%는 어형통제(normalization)를 통해 동일한 형태로 전환되었다. 즉, 학술지

논문에 부여된 저자키워드의 거의 절반이 원형 그대로 또는 어형통제를 통해 디스크립터와 동일한 형태로 출현한다는 것이다. 따라서 저자가 직접 문헌에 부여한 저자키워드는 그 자체로는 자연언어 색인어로서의 특성을 갖고 있으나, 학술정보 데이터베이스에서 통제어휘인 디스크립터를 할당할 때 이를 참고하고 있음을 알 수 있다. 즉, 저자키워드를 그대로 통제어휘로 사용할 수는 없으나 디스크립터를 자동 할당하는데 활용할 수 있는 유용한 자원이 될 수 있다. 그럼에도 불구하고 지금까지 이러한 저자키워드를 활용하여 통제키워드를 자동 할당하려는 노력은 찾아볼 수 없었다.

본 연구에서는 이러한 사실에 기초하여 비통제 색인어인 저자키워드를 활용하여 통제 색인어로서 디스크립터를 자동 할당할 수 있는 가능성을 모색하고자 한다. 특히, 현재 대부분의 국내 학술지 논문 검색서비스에서 제공하고 있는 저자키워드에 대한 기계학습 기반의 재분류를 통해 통제키워드로서 디스크립터의 자동 할당 효과를 얻을 수 있는지를 검토해 보고자 한다.

2.2 기계학습 기반 분류기

기계학습 기반의 텍스트 범주화에서 문헌 및 자질집합과 함께 가장 중요한 요소는 분류기이다. 본 연구의 목적에 적합한 분류기를 선정하기 위해 지금까지 선행연구들에서 주로 사용되어 온 7개 기본 분류기들(NB, kNN, RBF, SVM, VPT, ADT, J4.8)을 검토하였다. 통제키워드(디스크립터)의 자동 할당을 위한 저자키워드의 재분류라는 목적에 부합하기 위하여 분류기는 다음과 같은 특성을 가져야 한다.

첫째, 분류기가 지나치게 엄격하거나 너그럽지 않아서 재분류에 의한 새로운 키워드의 할당이 가능하여야 한다.

둘째, 재분류 결과 중에서 제외보다는 추가가 더 많이 발생하는 것이 바람직하다.

셋째, 재분류 결과 추가되는 문헌이 너무 과도하지 않아야 한다. 즉, 추가 문헌 수가 재분류 이전에 비해 절반을 넘기지 않아야 한다.

2.2.1 NB(Naive Bayesian)

NB 분류기는 베이즈 정리(Bayes theorem)에 근거한 확률적 분류기로서 학습문헌을 이용하여 각 단어가 특정 범주를 대표할 확률을 계산한 다음, 분류할 입력문헌에 출현한 단어들을 단서어로 하여 이 문헌의 범주를 예측한다. NB 분류기의 장점은 속도가 빠르고 구현이 쉬우며 비교적 성능이 좋은 분류기라는 점이다. 반면, NB 분류기의 단점은 자질집합의 특성(품질 및 규모)에 상당한 영향을 받으며 기본적인 독립성 가설이 지켜지지 않는 경우에는 성능이 크게 저하된다는 것이다. 또한, 분류 결정이 다소 너그러운 경향이 있어 다른 엄격한 분류기들(SVM, VPT 등)에 비하여 각 범주에 상대적으로 많은 수의 문헌을 할당하는 문제가 있다(김판준, 2006a).

2.2.2 kNN(k-Nearesrt Neighbors)

kNN 분류기는 학습문헌을 저장하는 사례기반 학습 방법을 채택한 대표적인 분류기로서, 입력문헌과 유사도가 가장 높은 k개의 최근접 이웃문헌을 학습문헌 집합으로부터 찾아 이들에 배정된 범주들에 근거하여 입력문헌을 분류할 하나 이상의 범주를 선정한다. kNN은 잘못

분류된 사례의 제거(noisy exemplar pruning)와 가중치부여 기법을 조합하는 경우에 좋은 성능을 기대할 수 있다. 또한 중요한 파라미터로서 k는 사전실험을 통해 경험적으로 설정하거나 특정 값으로 설정된 상한이 되도록 leave-one-out cross-validation을 이용하여 자동적으로 결정될 수 있다(Witten & Frank, 2005).

kNN 분류기의 장점은 구현이 쉽고 비교적 성능이 좋은 분류기라는 점이다. 반면 kNN의 단점은 대규모의 학습집합에 대하여 처리속도가 느리고, 잘못 분류된 학습사례 및 부적합한 자질집합에 의한 성능 저하가 크다는 것이다.

2.2.3 RBF Network(Radial Basis Function Network): 신경망(Neural Network)

신경망(neural network)을 이용한 문헌 분류기는 단위(unit) 노드들의 네트워크로서, 입력노드는 문헌벡터를 구성하는 각 용어를, 출력노드는 범주를 나타낸다. 노드 간 링크가 갖는 가중치는 의존관계를 나타내며 학습을 통해 결정된다(Sebastiani, 2002). RBF Network는 대표적인 순방향 네트워크(feedforward network)로서 입력층 이외에 두 개의 은닉층(hidden layer)을 갖는다. 각 은닉단위(hidden unit)는 입력공간 내 특정 지점을 나타내며 특정 사례에 대한 출력 또는 활성화(activation)는 해당 지점과 사례 간의 유클리드 거리에 따라 발생하고, 이들 두 지점이 근접할수록 활성화가 더 강하게 발생한다. 여기서 특정 은닉단위가 동일한 활성화를 산출하는 사례 공간 내 지점들이 결정구(hypersphere: 다층 퍼셉트론에서의 결정면(hyperplane))를 형성하기 때문에, 이러한 은닉단위(hidden units)들을 RBFs라 한다. RBF Network는 첫 번째 파라미터 집합을 두 번째 파라미터 집합과 독립적으로 결정할 수 있으면서 정확한 분류기를 생성할 수 있다는 것이 다층 퍼셉트론보다 큰 장점이다. 반면, RBF Network의 단점은 거리 계산에서 동등하게 취급되기 때문에 모든 속성에 동일한 가중치를 부여하므로, 다층 퍼셉트론과는 대조적으로 부적합한 속성들을 효과적으로 다룰 수 없다는 점이다.

2.2.4 SVM(Support Vector Machine)

단순한 SVM은 긍정예제 집합과 부정예제 집합을 최대의 마진(margin)을 갖고 분리하는 결정면(hyperplane)을 찾아내어, 이와 가장 가까운 문헌들을 지지벡터(support vectors)로 사용하여 학습하는 분류기이다. 본 연구에서 사용한 SVM 분류기는 WEKA 프로그램에서 제공하는 SMO(Sequential minimal optimization algorithm for support vector classification)로서 polynomial 또는 가우시안 커널(Gaussian kernels)을 사용한다(Keerthi et al., 2001; Platt, 1998).

SVM 분류기의 장점은 자질집합의 특성(품질, 규모 등)에 크게 영향을 받지 않으며 일반적으로 다른 분류기들에 비해 높은 성능을 기대할 수 있다(김판준, 2006a; Isa et al., 2008; Joachims, 1998; Yang & Liu, 1999). 반면, SVM 분류기의 단점은 구현이 어렵고 자질집합의 규모가 커질수록 처리속도가 느려진다는 것이다. 또한 분류 결정이 엄격하여 다른 분류기들에 비하여 각 범주에 비교적 적은 수의 문헌을 할당하는 경향이 있다(김판준, 2006a).

2.2.5 VPT(Voted PercepTron)

VPT는 퍼셉트론 알고리즘의 변형으로서 투표 방식의 퍼셉트론 분류기이다(Witten & Frank, 2005). 퍼셉트론 알고리즘은 사례별로 데이터를 반복적으로 학습하고 이들 사례 중 하나가 지금까지 학습된 가중치에 기초하여 잘못 분류될 때마다 가중치 벡터를 갱신한다. 여기서 가중치 벡터는 사례의 속성 값을 더하거나 빼서 갱신되므로 최종 가중치 벡터는 잘못 분류된 사례들의 합이 된다. 이러한 퍼셉트론은 학습 과정 동안 잘못 분류된 사례들을 저장(추적)하고 각 예측의 형성에 이러한 표현을 사용함으로써 비선형 분류기를 학습할 수 있다는 측면에서 커널 퍼셉트론이라고 부른다.

퍼셉트론 알고리즘에 의해 발견된 솔루션 벡터는 사례들이 입력되는 순서에 크게 영향을 받으며, 벡터들이 하나의 예측에 대하여 투표하도록 학습하는 동안 마지막 벡터가 아니라 학습에 사용된 모든 가중치 벡터들을 사용하는 것이 좋다. 이들 각 가중치 벡터의 정확성(correctness)은 사례들이 정확하게 분류되어 변경되지 않는 시점 이후의 연속적인 시도 횟수로 측정할 수 있으며, VPT는 이러한 측정값을 각 가중치 벡터에 부여된 투표 횟수로 사용한다. VPT는 비교적 처리속도가 빠르면서도 SVM과 유사하거나 이에 조금 못 미치는 성능을 보여주는 것으로 알려져 있다(김판준, 2006a).

2.2.6 ADT(Alternating Decision Tree)

ADT는 부스팅(boosting)[2]을 사용하는 교차 결정트리(alternating decision trees) 분류기이다. 부스팅 알고리즘을 이용하여 노드들을 추가한 결과 트리를 교차 결정트리라고 하고, 이러한 트리 내에서 결정 노드는 '분리 노드(splitter nodes)', 조건 노드는 '예측 노드(prediction nodes)'라고 부르며, 예측 노드는 더 이상 어떠한 분리 노드도 추가되지 않을 때 잎(leaves)이 된다. 일반적으로 교차 결정트리는 2-범주 문제에 적용되며, 각 예측 노드는 긍정 또는 부정 값에 연계된다. 특정 사례(입력문헌)의 범주는 모든 적용가능한 가지들로 진행하면서 경로 상에 있는 예측 노드들의 값을 합산하고, 그 값이 양수 또는 음수 인지에 따라 결정된다. 부스팅의 반복 횟수가 문제의 복잡성과 분류 성능 간의 tradeoff를 위한 파라미터가 될 수 있으며, 각각의 반복에서 노드들이 통합될 수 없을 경우에는 트리에 3개의 노드(하나의 분리 노드와 두 개의 예측 노드)를 추가한다. WEKA 프로그램에서 제공한 디폴트 탐색 방법은 망라적 탐색이다(Witten & Frank, 2005).

결정트리 분류기의 장점은 잘못 분류된 사례와 학습집합의 규모에 대한 영향을 적게 받으며, 결과의 이해와 해석이 쉽다는 것이다. 반면, 결정트리 분류기의 단점은 지역적인 각 노드에서 결정이 이루어지는 greedy 알고리즘에 기초하므로 전역적인 최적 결정트리를 보장할 수 없으며, 일부 가지들이 학습 데이터에 너무 특정하

2) 부스팅(boosting)이란 주어진 학습집합을 적절히 조작하여 서로 다른 여러 개의 분류함수들을 만들고, 이들을 합쳐서 하나의 성능이 좋은 분류함수를 만드는 것이다. 따라서 부스팅 알고리즘에서는 주어진 학습집합에 속한 각 학습문서의 가중치(weight)를 달리하여 여러 개의 학습집합을 만든다. 그 후 이들 각 학습집합을 이용하여 여러 개의 분류함수를 만들고 투표(voting)방법을 이용하여 이들을 합친 결과로 하나의 분류함수를 만든다.

게 됨으로써 과적합(overfitting)이 발생할 수 있다. 따라서 기본적으로 이러한 문제를 방지하기 위한 방법을 필요로 하며, 대부분의 결정트리 학습 방법은 트리의 성장과 너무 특정한 가지들을 제거하기 위한 트리의 가지치기(pruning) 방법을 포함한다(Mitchell, 1997).

2.2.7 J4.8: 결정트리(Decision Tree)

J4.8은 C4.5의 개선된 버전으로서 Weka에서 제공하는 결정트리 분류기이다. 결정트리는 이용 가능한 데이터의 여러 속성 값에 기초하여 새로운 사례(문헌)의 목표값(종속변수)를 결정하는 기계학습 방법이다. 결정트리의 내부 노드들은 서로 다른 속성들이 되고 노드들 간의 가지는 이들 속성이 가질 수 있는 값들을 나타내며, 최종적인 노드는 종속변수의 범주가 된다. J4.8 결정트리 분류기는 새로운 문헌을 분류하기 위해 학습집합의 속성 값에 기초한 결정트리를 생성한다. 따라서 새로운 문헌(학습집합)을 만날 때마다 여러 사례들을 가장 명확하게 구분하는 속성을 식별한다. 또한, 최적의 결과를 산출할 수 있도록 사례에 대한 정보를 가장 많이 갖는 자질이 가장 높은 가중치(Information Gain: 정보획득량)를 갖는다. 이러한 자질 값 중에서 모호성이 없는 값이 있으면, 해당 범주에 속한 사례는 목표 변수와 동일한 값을 갖게 되어 그 가지에서 멈추고 목표값을 할당한다.

결정트리에서는 모든 개별 속성과 이들의 값을 확인하고 새로운 문헌의 범주를 할당하거나 예측할 수 있다. 결정트리 알고리즘에서 가장 기본적인 파라미터는 트리 절단 옵션이다. J4.8은 트리의 절단을 위한 두 가지 방법을 제공한다. 첫 번째는 결정트리 내 노드들을 하나의 잎으로 대체하는 것으로 특정 경로 상에서 검토해야 할 노드 수를 감소시킨다. 두 번째는 경로 상에 있는 다른 노드들을 대체하면서, 특정 노드를 트리의 뿌리 노드(root node)방향으로 이동할 수 있다(Witten & Frank, 2005).

3. 실험 설계

3.1 실험 데이터

실험집단에 대한 사전처리와 자질선정을 위한 프로그램은 Python 및 Visual FoxPro로 구현된 프로그램을 사용하였고, 저자키워드 자동부여 실험을 위한 프로그램(분류기)은 공개된

〈표 2〉 실험 문헌집단

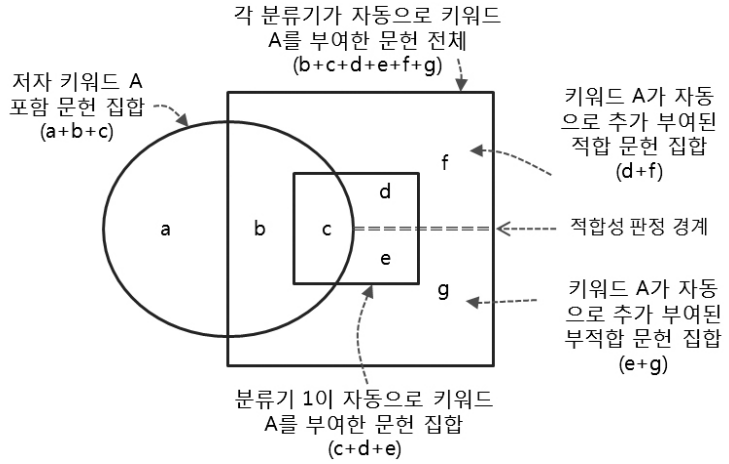
항 목	내 역
전체 문헌 수(학습・검증집합)	652
저자키워드 종수	10
저자키워드 당 학습・검증문헌 수(최대/최소/평균)	37/9/17.7
문헌 당 저자키워드 수(최대/최소/평균)	31/2/10.8
학습・검증집합의 키워드 종수(전체/자질(비율))	27279/1909(7%)
학습・검증집합의 자질 종수(최대/최소/평균)	299/6/96.4
학습・검증집합의 자질 수(최대/최소/평균)	2631/6/391.6

기계학습 실험 패키지인 WEKA Version 3.6 (Witten & Frank, 2005)[3]을 사용하였다. 본 실험은 WEKA 프로그램의 기능 중에서 학습 집합을 그대로 검증집합으로 삼아 재분류하는 기능을 사용하므로 학습집합과 검증집합이 동일하다. 또한, 자질선정 측면에서는 〈표 2〉와 같이 학술지 논문의 표제와 초록에 출현한 단어 중에서 전체 문헌집단에서 6개 이상의 논문에 출현한 키워드(df≥6)를 자질로 선정하여 문헌벡터를 구성하였다.

3.2 성능 평가 방법

분류기의 성능 평가를 위해서 이 연구에서는 TREC(Voorhees & Harman, 2005)을 비롯한 정보검색 평가에서 흔히 사용하는 풀링(pooling) 기법을 채택하였다. 정보검색에서 풀링 기법은 모든 검색질의에 대해서 실험 집합의 모든 문헌을 대상으로 적합 여부를 판정하지 않는다. 그 대신 실험에 참여한 여러 검색시스템이 특정 검색질의에 대해서 검색해 낸 문헌들을 모은 다음 그 문헌집합에 대해서만 해당 검색질의에 대해 적합한지 여부를 판정한다. 이와 유사한 방법으로 이 연구에서는 특정 저자키워드에 대해서 실험에 사용한 여러 분류기가 자동으로 분류한 문헌들을 모아서 이 문헌집합에 대해서만 해당 키워드가 적합한지 여부를 판정하였다. 이와 같은 풀링 기법을 이용하여 정확률과 재현율, 그리고 결합 척도인 F1을 계산하는 과정은 〈그림 1〉을 활용해서 설명하기로 한다.

〈그림 1〉에서 원은 각 문헌의 저자가 키워드 A를 할당한 문헌 집합을 의미하며, 작은 사각형은 분류기 1이 키워드 A를 자동 부여한 문헌 집합, 큰 사각형은 어느 한 분류기에 의해서라도 키워드 A가 자동으로 부여된 모든 문헌 집합을 의미한다. 가로로 그어진 점선은 저자가 키워드 A를 할당하지 않았으나 분류기에 의해 자동으로 키워드 A가 부여된 문헌에 대해서 적



〈그림 1〉 분류기 1의 키워드 자동 분류 결과 평가를 위한 벤다이어그램

3) 〈http://www.cs.waikato.ac.nz/~ml/weka/〉.

합 여부를 판정한 구분선이다. a부터 g까지 분할된 각 영역은 다음을 뜻한다.

a : 저자키워드 A를 포함하고 있으나 자동 분류에서 키워드 A가 부여되지 않은 문헌
b : 저자키워드 A를 포함하고 있으며 자동 분류에서 분류기 1이 아닌 다른 분류기에 의해 키워드 A가 부여된 문헌
c : 저자키워드 A를 포함하고 있으며 자동 분류에서 분류기 1에 의해 키워드 A가 부여된 문헌
d : 저자키워드 A를 포함하지 않으나 자동 분류에서 분류기 1에 의해 키워드 A가 부여되었으며, 적합하다고 판정된 문헌
e : 저자키워드 A를 포함하지 않으며 자동 분류에서 분류기 1에 의해 키워드 A가 부여되었지만 부적합하다고 판정된 문헌
f : 저자키워드 A를 포함하지 않으나 자동 분류에서 분류기 1이 아닌 다른 분류기에 의해 키워드 A가 부여되었으며, 적합하다고 판정된 문헌
g : 저자키워드 A를 포함하지 않으며 자동 분류에서 분류기 1이 아닌 다른 분류기에 의해 키워드 A가 부여되었지만 부적합하다고 판정된 문헌

저자가 키워드 A를 부여한 문헌은 일단 모두 적합하다고 가정한다. 물론 관점에 따라서는 다소 부적절하다고 간주할 수 있는 경우도 포함되어 있으나 실제 이용자에게 서비스하는 경우에 저자키워드를 선별적으로 제시하는 것은 무리이기 때문이다. 이 연구에서는 특정 저자키워드와 관련된 문헌을 추가로 파악하는 것을 목표로 하기에 저자가 부여한 키워드는 모두 적절한 것으로 간주하였다.

키워드 A에 대한 분류기 1의 검색 성능은 다음과 같이 정확률과 재현율 그리고 복합 척도인 F1으로 산출한다.

$$\text{정확률} = \frac{a+b+c+d}{a+b+c+d+e}$$

$$\text{재현율} = \frac{a+b+c+d}{a+b+c+d+f}$$

$$F_1 = \frac{2\times\text{정확률}\times\text{재현율}}{\text{정확률}+\text{재현율}}$$

전체 키워드에 대한 종합 성능은 각 척도값의 평균을 구하여 평가한다. 아무것도 추가로 분류하지 않았을 경우, 즉 원래 저자키워드가 부여된 경우의 성능을 평가하면 평균 정확률은 1.0이고 평균 재현율은 0.807, 그리고 F1값은 0.892이다. 평균 정확률이 1.0인 이유는, 저자키워드가 부여된 경우를 모두 적합한 것으로 가정했기 때문이며, 평균 재현율이 1.0이 아닌 이유는 각 분류기가 추가로 파악한 적합 문헌이 저자키워드 부여 집합에는 포함되지 않았기 때문이다.

각 분류기의 성능을 정확률과 재현율, 그리고 F1으로 측정한 후 저자키워드만 부여된 경우를 기준으로 각 지표의 향상율을 산출하였다. 정확률은 저자키워드만 부여된 경우가 1.0으로 완벽한 경우였으므로 이보다 가급적 적게 하락하는 것이 목표이며, 재현율은 저자키워드만 부여된 경우보다 향상되도록 하는 것이 목표이다. 재현율이 향상되더라도 정확률이 하락한 비율보다 더 높게 향상되지 않으면 F1값은 하락하게 된다. 정확률은 저자키워드만 부여한 경우가 최고일 수밖에 없다는 점을 감안하면, 결국 이 실험에서

는 정확률의 손실을 최소화하면서 재현율을 최대로 높이는 방법을 찾아내는 것이 관건이다.

4. 실험 결과 및 분석

4.1 최적 분류기 선정을 위한 비교 실험

4.1.1 1차 실험: 극단적인 분류기 배제

1차적으로 실험에 적용할 분류기를 선정하기 위해서 각 분류기가 얼마나 많은 문헌에 추가로 키워드를 할당하는지를 파악해보았다. 평가한 분류기는 나이브베이즈 분류기(NB), k=5인 kNN 분류기(IBK5), k=10인 kNN 분류기(IBK10), 신경망 분류기(RBF), 지지벡터기계(SVM), 투표형 퍼셉트론 분류기(VPT), 5회 반복 의사결정트리 분류기(ADT5), 10회 반복 의사결정트리 분류기(ADT10), 의사결정트리 분류기(J4.8)의 9종이다. 다른 분류기와 달리 kNN 분류기와 의사결정트리(ADT) 분류기는 성능에 큰 영향을 끼치는 파라미터가 있으므로 이를 고려하여 파라미터가 다른 두 가지 경우를 설정하여 실험에 포함하였다.

이 실험은 학습집합과 검증집합이 동일한 상황, 즉 학습집합에 대해서 분류기가 학습한 결과를 동일한 집단에 다시 적용하는 것이다. 따라서 분류기가 지나치게 엄격하여 완벽하게 학습한다면 원래 분류된 문헌 이외에 추가하는 문헌이 아예 없을 것이므로 재분류에 의한 색인어 추가 할당 문제에 적용할 수가 없다. 또한 너그러운 기준을 적용하여 지나치게 많은 문헌을 추가하는 분류기도 적절하지 않다. 추가되는 문헌이 지나치게 많으면 최초 분류된 문헌 집합의 주제에서 벗어나게 되는 주제 표류(topic drift) 현상이 나타날 가능성이 높기 때문이다. 추가 문헌의 수가 원래 분류된 문헌의 수에 비해서 절반 이상이 된다면 저자 집단의 판단 중 반 이상을 신뢰하지 않게 된다는 뜻이므로 과도한 수정이 될 수 있다. 따라서 지나치게 엄격한 분류기나 반대로 너무 너그러운 분류기는 저자키워드의 재분류(확장)라는 목적에 적합하지 않다. 분류기의 학습에 따른 특성을 확인해보기 위해서 1차적으로 각 분류기가 학습집합에 대해서 학습한 이후, 재분류를 수행한 결과 어느 정도의 추가 문헌을 제시하는지를 분석하였다. 결과적으로 매우 엄격한 분류기라면 저자가 키워드를 부여한 문헌 이외에 추가로 해당 키워드를 부여하는 문헌은 드물 것이다. 반대로 너그러운 분류기라면 추가로 키워드를 부여하는 문헌이 매우 많을 것이다.

분류기에 의해서 해당 키워드 범주에 새로 추가된 문헌 수가 원래 문헌 수에 비해서 얼마나 되는지를 나타내는 추가율을 측정해본 결과는 〈표 3〉과 같다. 추가율은 대부분 20% 미만이었으나 NB는 평균 99.1%로 나타나서 원래 문헌 수와 거의 같은 수의 문헌에 키워드를 추가로 부여하는 것으로 나타났다. 지나치게 많은

〈표 3〉 각 분류기의 추가 문헌 수 비율 비교

분류기	NB	IBK5	IBK10	RBF	SVM	VPT	ADT5	ADT10	J4.8
추가율 평균	99.1%	1.8%	1.1%	7.0%	0.0%	18.7%	16.0%	3.7%	6.3%
추가율 최고	323.1%	10.0%	10.8%	23.1%	0.0%	38.5%	42.9%	12.5%	33.3%

문헌에 키워드를 추가로 부여하게 되면 문헌의 내용과 맞지 않는 경우 또한 많아지므로 이용자에게 도움이 되지 않는다. 반면에 SVM은 추가하는 문헌이 1건도 없어서 추가율이 0%로 나타났다. 추가 문헌을 한 건도 찾아주지 못하는 분류기는 저자 집단의 판단을 완벽하게 학습한 결과이기 때문에 적용하는 문제에 따라서는 좋은 것으로 판단할 수도 있지만, 저자키워드에 근거한 재분류(색인어 자동 추가 할당)라고 하는 문제에 있어서는 추가 문헌이 전혀 없으므로 적합하지 않다. 따라서 키워드 추가 부여 문제에 적합하지 않은 극단적인 성향의 NB, SVM 분류기를 제외하면 kNN 분류기 2종, ADT 분류기 2종, RBF, VTP, J4.8의 7개 분류기를 1차 실험을 통해 선정하였다.

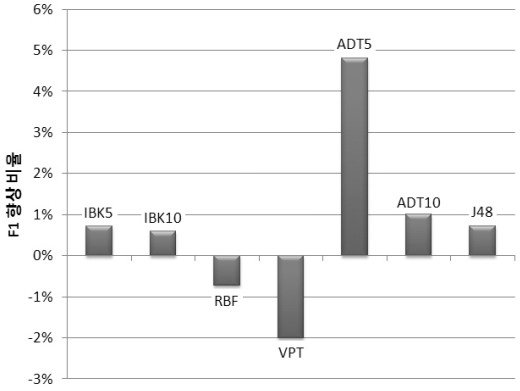
4.1.2 2차 실험: 분류기별 성능 비교

1차 실험을 거쳐 선정된 7개 분류기를 사용하여 10개 저자키워드에 대한 색인어 자동 할당을 수행한 결과에 대해서 정확률, 재현율, F1을 〈표 4〉와 같이 측정해보았다. 한편 색인어 자동 할당 없이 저자가 부여한 경우의 성능을 기준으로 하였을 때 각 분류기의 상대적인 F1의 향상 비율을 〈그림 2〉에 제시하였다.

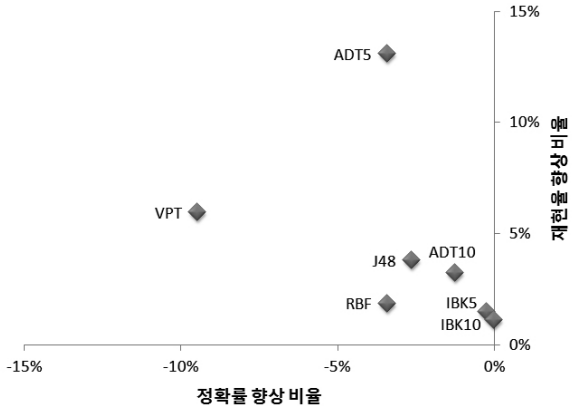
F1 성능을 비교해보면 가장 좋은 경우는 5회 반복한 ADT 분류기인 ADT5로서 자동 키워드

〈표 4〉 분류기별 성능 비교

		저자 키워드	IBK5	IBK10	RBF	VPT	ADT5	ADT10	J4.8
평균 성능	추가 적합문헌 수	-	0.3	0.4	0.3	0.9	2.3	0.6	0.5
	추가 문헌 수	-	0.4	0.4	0.9	2.7	3.3	0.8	0.9
	평균 정확률	1.000	0.998	1.000	0.966	0.905	0.966	0.987	0.972
	평균 재현율	0.807	0.794	0.792	0.798	0.855	0.913	0.833	0.813
	평균 F1	0.892	0.883	0.882	0.870	0.874	0.935	0.901	0.883
향상 비율	평균 정확률 향상	-	-0.3%	0.0%	-3.4%	-9.5%	-3.4%	-1.3%	-2.7%
	평균 재현율 향상	-	1.5%	1.1%	1.9%	6.0%	13.1%	3.2%	3.8%
	평균 F1 향상	-	0.7%	0.6%	-0.7%	-2.0%	4.8%	1.0%	0.7%



〈그림 2〉 분류기별 평균 F1 향상 비율



〈그림 3〉 분류기별 평균 정확률과 평균 재현율 향상 비율

추가 할당을 수행하기 이전 저자키워드만 할당된 경우의 F1인 0.892보다 4.8% 향상된 0.935로 나타났다. 한 키워드 당 평균 3.3개의 추가 문헌을 찾아주며 이중에서 적합 문헌은 평균 2.3개, 부적합 문헌은 평균 1개로 나타났다. 10회 반복 ADT 분류기를 비롯한 4종의 분류기는 F1이 1% 이내로 미미하게 향상되었으며, RBF와 VPT는 재현율 향상 정도에 비해서 정확률 하락 정도가 더 커서 F1이 오히려 저하되는 것으로 나타났다.

일반적으로 정확률과 재현율은 반비례하며 상호보완적인 성향이 있다. 저자가 키워드를 부여한 경우에 비해서 정확률과 재현율이 향상된 비율을 〈그림 3〉에 제시해보았다. 그림을 보면 향상비율이 정확률과 재현율 모두 0%인 원점을 기준으로 왼쪽 위로 45도 선을 가상으로 그려볼 때 선 위쪽에 점이 위치하면 정확률이 하락한 정도보다 재현율이 향상된 정도가 더 크다는 것을 나타낸다. 즉, 정확률의 일부 하락을 무릅쓰고 재현율을 높일 가치가 있다는 뜻이다. 이 그림에서도 5회 반복한 ADT 분류기와 10회 반복한 ADT 분류기가 정확률 하락 정도에 비해서 재현율 상승 정도가 높다는 것을 알 수 있다. 따라서 여러 분류기 중에서 ADT 분류기가 저자키워드의 추가 할당 문제에 대해서 가장 적합한 분류기임을 알 수 있다.

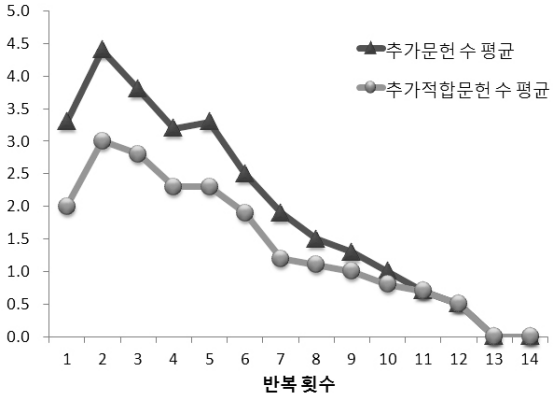
4.2 3차 실험: ADT 분류기의 최적 파라미터 파악

7개 분류기 간 비교 실험에서는 ADT 분류기가 가장 좋은 성능을 보였다. 그중에서도 부스팅을 5회 반복한 경우가 10회 반복한 것보다 더 좋은 성능을 보였다. ADT 분류기에서 부스팅 반복 횟수가 적을수록 분류 수행 시간이 짧으므로 반복 횟수가 적으면서 성능이 더 좋은 경우를 이상적인 분류기라고 할 수 있다. 그러나 더 많은 수행시간이 소요되더라도 성능이 더 향상되는 분류기의 경우에도 실시간 분류가 아닌 상황에서는 더 적합하다. 따라서 ADT 분류기를 사용하면서 5회보다 반복을 더 적게 하거나 10회보다 반복을 더 많이 하였을 때의 성능을 알아보는 실험을 수행하였다.

ADT 분류기의 부스팅 반복 횟수를 기존에

수행했던 5회와 10회에서 각각 4회씩 더 적거나 많게 설정하여 최소 1회에서 최대 14회까지 바꾸면서 재분류를 수행하였다. 반복 횟수에 따라 추가되는 문헌 수는 〈그림 4〉와 같이 2회 반복하였을 때 실험한 10개 키워드 당 평균 약 4.5개로 가장 많았으며, 13개 이후로는 앞의 실험에서 SVM 분류기의 경우처럼 추가되는 문헌이 전혀 없었다. 따라서 ADT 분류기의 부스팅 반복 횟수를 크게 늘리면 학습집합에서 저자키워드 할당 정보를 완벽하게 학습하므로 추가 키워드 할당이 더 이상 발생하지 않는다는 것을 알 수 있다.

1회부터 14회까지 부스팅 반복 횟수를 달리하였을 때의 색인어 추가 할당 성능은 〈표 5〉와

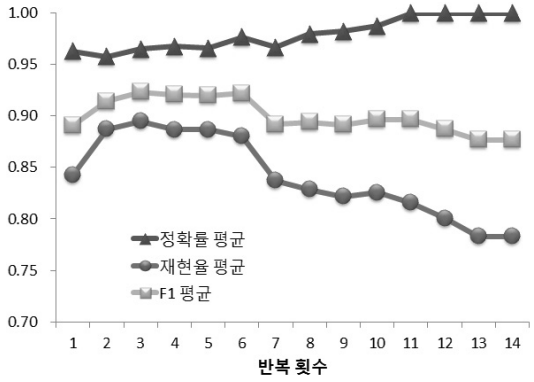


〈그림 4〉 ADT 분류기의 반복횟수에 따른 키워드별 평균 추가 문헌 수

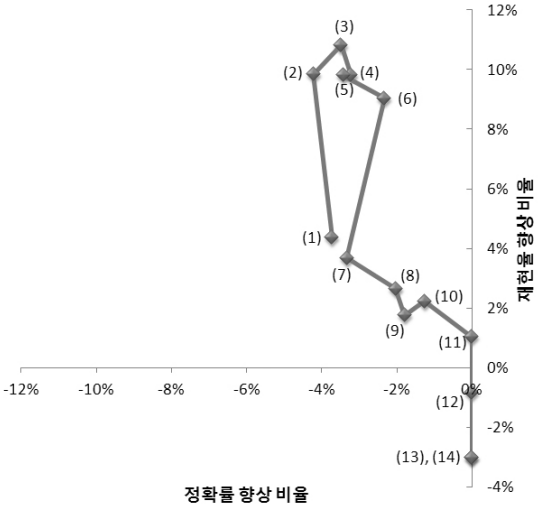
〈표 5〉 ADT 분류기의 부스팅 반복 횟수에 따른 평균 수치와 성능 향상 비율

분류기	반복횟수	평균 성능					성능 향상 비율		
		추가적합 문헌 수	추가 문헌 수	정확률	재현율	F1	정확률 향상	재현율 향상	F1 향상
ADT	1회	2.0	3.3	0.9628	0.8426	0.8911	-3.7%	4.4%	-0.1%
	2회	3.0	4.4	0.9578	0.8869	0.9141	-4.2%	9.8%	2.5%
	3회	2.8	3.8	0.9651	0.8948	0.9232	-3.5%	10.8%	3.5%
	4회	2.3	3.2	0.9677	0.8866	0.9206	-3.2%	9.8%	3.2%
	5회	2.3	3.3	0.9657	0.8866	0.9196	-3.4%	9.8%	3.1%
	6회	1.9	2.5	0.9767	0.8804	0.9221	-2.3%	9.0%	3.4%
	7회	1.2	1.9	0.9667	0.8371	0.8919	-3.3%	3.7%	0.0%
	8회	1.1	1.5	0.9796	0.8287	0.8940	-2.0%	2.6%	0.3%
	9회	1.0	1.3	0.9821	0.8216	0.8913	-1.8%	1.8%	0.0%
	10회	0.8	1.0	0.9874	0.8254	0.8966	-1.3%	2.2%	0.5%
	11회	0.7	0.7	1.0000	0.8157	0.8965	0.0%	1.0%	0.5%
	12회	0.5	0.5	1.0000	0.8006	0.8869	0.0%	-0.8%	-0.5%
	13회	0.0	0.0	1.0000	0.7830	0.8769	0.0%	-3.0%	-1.7%
	14회	0.0	0.0	1.0000	0.7830	0.8769	0.0%	-3.0%	-1.7%

같으며, F1과 정확률 및 재현율 향상 정도는 별도로 〈그림 5〉에 제시하였다. F1 성능은 2회에서 6회 사이가 0.9 이상으로 높게 나타났으며 3회 반복하였을 때가 가장 좋았다. 정확률 향상 비율과 재현율 향상 비율을 각각 가로축과 세로축으로 설정한 〈그림 6〉을 보면 반복 횟수가 2회부터 6회까지는 원점에서 먼 위쪽에 위치하다가 7회부터 원점에 가까운 위치로 내려오는 것을 볼 수 있다. 12회를 넘어서면 원점 아래쪽으로 이동하므로 ADT 분류기는 지나치게 많은 반복 수행이 오히려 키워드 추가 할당 성능을 떨어뜨린다는 것이 확인되었다. 따라서 ADT 분류기로 안정적인 좋은 성능을 얻기 위해서는 부스팅을 3회에서 6회 사이로 반복시키는 것이 가장 적절한 것으로 나타났다.



〈그림 5〉 ADT 분류기의 부스팅 반복 횟수에 따른 성능 변화



〈그림 6〉 부스팅 반복횟수에 따른 ADT 분류기의 정확률과 재현율 향상 비율 변화(괄호 안은 반복 횟수)

4.3 디스크립터 자동 할당 효과 분석

본 연구에서 제시한 기계학습 기법 기반의 저자키워드 재분류를 통하여 새로 추가된 적합 문헌들이 실제로 통제키워드(디스크립터)에 의한 어휘통제[4] 효과가 있는 지를 확인해 보았다. 먼저, 〈표 6〉은 본 연구에서 사용된 저자키워드들과 이들이 원래 할당되어 검색된 문헌 수, 그리고 해당 저자키워드의 재분류에 따라 새롭게 추가된 적합문헌 수를 제시한 것이다. 이에 따르면, 10개 저자키워드에 대하여 처음에는 최소 9개에서 최대 37개 문헌이 검색되었지만, 저자키워드의 재분류 결과로는 평균 4.4개의 적합문헌이 추가로 검색되는 것으로 나타났다.

다음으로, 〈표 7〉에서는 새로 추가된 적합문헌들이 실제 디스크립터를 할당한 경우와 유사한 어휘통제 효과를 보다 구체적으로 살펴보기 위하여, 각 저자키워드 당 추가된 적합문헌을 최대 3개까지 제시하였다. 여기서 디스크립터는 한글 표현으로 문헌집단 내에서 가장 많이 할당된 용어로 선정한 것이며, 추가 적합문헌의 키워드는 해당 논문의 서명, 저자키워드, 초록(또는 목차) 필드에 출현한 것으로 자동색인된 결과 그대로 추출하였다. 특히 추가 적합문헌의 키워드를 살펴보면 각 저자키워드의 동의어(또는 유사어)로서 동일한 개념에 대한 서로 다른 표현들이 논문의 표제, 저자키워드, 초록 필드에 출현하고 있으며, 이들은 대부분 원래의 저자키워드가 할당된 논문이 다루고 있는 주제(개념)를 나타내고 있다. 따라서 저자키워드의 재분류를 통하여 동일한 개념에 대한 다양한 표현들을 키워드로 갖고 있는 논문들을 적합문헌 집합에 추가함으로써, 같은 주제를 다루는 문헌들이 하나의 색인어 아래 모이게 하는 디스크립터의 '어휘통제' 효과를 얻을 수 있다는 것을 알 수 있다. 즉, 기계학습에 기초한 저자키워드의

〈표 6〉 저자키워드의 최초 할당문헌 수와 추가 적합문헌 수

번호	저자키워드	최초 할당문헌 수	추가 적합문헌 수
1	communication	9	2
2	독서교육	28	4
3	독서지도	14	6
4	학교도서관	24	4
5	독자	13	1
6	reading program	12	3
7	독서치료	31	10
8	공공도서관	37	9
9	도서관	9	4
10	정체성	9	1

4) 검색의 효율을 높이기 위하여 단어의 어형 통제, 동의어 통제, 단수/복수형 통제 등을 통하여 가능한 한 같은 개념을 나타내는 다양한 단어들을 하나의 용어로 통일함으로써 특정한 개념은 항상 동일한 색인어로 표현되도록 하는 것을 말한다. 따라서 같은 개념을 다루는 문헌은 하나의 색인어 아래 모이는 효과를 갖는다(문헌정보학용어사전, 2010).

〈표 7〉 저자키워드 재분류에 따른 추가 적합문헌의 키워드와 서명

저자키워드	디스크립터	추가 적합논문의 키워드	추가 적합논문의 서명
communication	의사소통	의사소통	공학교육에서 문식성 학습목표 달성을 위한 글쓰기 수업모형
		의사소통능력; Communication ability	지식정보화사회에서 리더십능력 개발을 위한 대학생 교양교육의 사례 연구
독서교육	독서교육	독서교육	전문가 협력을 통한 어린이 독서교육 프로그램 개발 및 운영
		독서교육	독서교육지원시스템의 구성요소 설정에 관한 연구
		독서교육; 일본의 독서교육; Reading Education Policy in Japan	일본의 독서교육에서 학교도서관의 의미
독서지도	독서지도	독서지도, 독서교육, 독서지도(론); 독서교육(론); Reading Education	독서교육 교과목의 내용 구성에 관한 연구
		독서교육, Reading Education	디지털 스토리텔링을 이용한 독서지도 방안 연구
		독서교육; 독서지도; reading education	효과적인 독서교육 방향 정립을 위한 학생 독서실태 조사연구
학교도서관	학교도서관	학교도서관; 중학교도서관	학교도서관의 교육적 효과에 대한 중학생의 인지도 분석
		학교도서관; 일본의 학교도서관; School Library in Japan	일본의 독서교육에서 학교도서관의 의미
		학교도서관; 중학교도서관; Middle School Library	중학교도서관 프로그램에 나타난 파트너십에 대한 연구
독자	독자	독자; 근대 문학 독자; reader; modern reader of literature	김동인 초기소설에 나타난 근대 '문학 독자'의 형성 연구
reading program	독서 프로그램	프로그램; 청소년 프로그램; Young Adults Programs	공공도서관 청소년프로그램 분석과 활성화 방안 연구
		독서 프로그램	토크·뮤직·영상 쇼의 개념을 적용한 북 토크 쇼 개발 연구
		프로그램; 문화프로그램; Cultural Program	공공도서관 문화프로그램 현황 분석과 활성화 방안
독서치료	독서치료	독서요법;Bibliotherapy	독서요법이 경증 치매노인에게 미치는 효과
		독서치료; 치유	문학 독서와 치유독자
		독서치료; Bibliotherapy; bibliotherapy	국어교과서에 기초한 독서치료 프로그램이 중학생의 자아존중감에 미치는 효과
공공도서관	공공도서관	공공도서관	공공도서관 아동교육프로그램의 사서역할 모델설계 및 사례분석
		평생교육; 평생학습; 도서관; life time learning; library	평생 학습의 장으로서 도서관 독서프로그램 고찰
		공동도서관; Public Library	국내외 도서관의 아동서비스와 직무사례 분석
도서관	도서관	공공도서관; Public Library	공공도서관 청소년 독서프로그램의 현황과 과제
		圖書館	1980年代 中後半期 讀書大衆化 運動 硏究
		도서관	장애인 독서환경 개선 방안 Ⅰ
정체성	정체성	정체성; 문화적 정체성; cultural identity	다문화 소설에 형상화된 유목적 존재들의 삶 이해를 통한 소설교육

재분류를 통하여 특정 디스크립터가 대표하는 동일한 개념을 다루고 있지만, 서로 다른 용어(비통제키워드)로 색인되어 검색되지 않은 새로운 문헌들을 추가적으로 검색할 수 있는 것이다.

5. 결 론

본 연구는 국내 주요 학술 DB의 검색서비스에서 제공되고 있는 저자키워드(비통제키워드)의 재분류를 통하여 디스크립터(통제키워드)를 자동 할당할 수 있는 가능성을 모색하였다. 이를 위해 먼저 기계학습에 기반한 주요 분류기들의 특성을 검토하는 실험을 수행하여 재분류를 위한 최적 분류기와 파라미터를 선정하였다. 다음으로, 국내 독서 분야 학술지 논문들에 부여된 저자키워드를 학습한 결과에 따라, 해당 논문들을 재분류함으로써 키워드를 추가로 할당하는 실험을 수행하였다. 결과적으로 실험 목적에 적합한 분류기 선정을 위한 1차 실험에서 극단적인 성향을 보이는 2개 분류기(NB, SVM)를 제외하였고, 나머지 7개 분류기 간의 성능을 비교한 2차 실험에서는 ADT 분류기가 가장 좋은 성능을 보였다. 또한, 3차 실험은 ADT 분류기의 최적 파라미터를 찾기 위한 것으로, 안정적으로 좋은 성능을 얻기 위해서는 부스팅을 3회에서 6회 사이로 반복시키는 것이 가장 적절한 것으로 나타났다.

이에 따라 가장 좋은 성능을 보인 ADT 분류기 및 파라미터를 적용하여 저자키워드의 재분류를 수행한 결과로 새롭게 추가된 적합문헌들에 대하여, 실제로 디스크립터의 어휘통제 효과가 발생하는 지를 살펴보았다. 그 결과 저자키워드의 재분류에 의해 특정 디스크립터가 대표하는 동일한 개념을 다루고 있지만, 서로 다른 용어(비통제키워드)로 색인된 관계로 이전에는 검색되지 않았던 적합문헌들을 추가적으로 검색할 수 있음을 확인하였다.

이 연구는 실험적인 연구로서 연구 대상 분야를 전체 학문 분야가 아닌 '독서'로 제한하였으며, 전체 저자키워드 집합을 모두 사용하지 않고 상위의 저자키워드만을 대상으로 하였다는 제한점이 있다. 물론 실제 통제어휘에 적합할만한 저자키워드 후보를 선정하기 위해서는 사용빈도가 일정 수준 이상인 키워드만을 채택하는 것이 무난한 방법이지만, 후속 연구에서는 통제어휘에 어울리는 저자키워드를 선정하는 방법이나 기준도 개발해야 할 것이다. 또한 적용 범위를 확대하여 다른 학문분야나 문헌데이터베이스 전체에 대해서 적용해보는 실험도 필요하다.

참 고 문 헌

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

김판준 (2006a). 기계학습을 통한 디스크립터 자동부여에 관한 연구. 정보관리학회지, 23(1), 279-299.

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

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

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

윤구호 (1999). 색인·초록. 서울: 한국도서관협회.

이재윤 (2005a). 문헌간 유사도를 이용한 SVM 분류기의 문헌분류성능 향상에 관한 연구. 정보관리학회지, 22(3), 261-287.

이재윤 (2005b). 자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 대한 연구. 한국문헌정보학회지, 39(2), 123-146.

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

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

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. http://dx.doi.org/10.1016/j.ipm.2010.07.003

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. http://dx.doi.org/10.1016/j.eswa.2010.08.100

Chung, Y., Pottenger, W. M., & Schatz, B. R. (1998). Automatic subject indexing using an associative neural network. Proceedings of the 3rd ACM International Conference on Digital Libraries (DL '98), ACM Press, 59-68.

Gil-Leiva, I., & Alonso-Arroyo, A. (2007). Keywords given by authors of scientific articles in database descriptors. Journal of the American Society for Information Science and Technology, 58(8), 1175-1187.

Harish, B. S., Guru, D. S., & Manjunath, S. (2010). Representation and classification of text documents: A brief review. IJCA Special Issue on "Recent Trends in Image Processing and Pattern Recognition" RTIPPR, 2010, 110-119.

Hurt, C. D. (2010). Automatically generated keywords: A comparison to author-generated keywords in the sciences. Journal of Information and Organizational Sciences, 34(1), 81-88. Retrieved from https://jios.foi.hr/index.php/jios/article/view/158

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.

Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. Proceedings of the 10th European Conference on Machine Learning, 137-142.

Khan, A., Baharudin, B., & Lee, Lam Hong (2010). A review of machine learning algorithms for text-documents classification. Journal of Advances in Information Technology, 1(1), 4-20.

Kumar, M. Arun, & Gopal, M. (2010). A comparison study on multiple binary-class SVM methods for unilabel text categorization. Pattern Recognition Letters, 31(11), 1437-1444.

Lauser, B., & Hotho, A. (2003). Automatic multi-label subject indexing in a multilingual environment. Proceedings of the 7th European Conference in Research and Adavanced Technology for Digital Libraries (ECDL '03), 140-151.

Lewis, D. D., Schapire, R. E., Callan, J. P., & Papka, R. (1996). Training algorithms for linear text classfiers. Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '96), 298-306.

Li, Cheng Hua, & Park, Soon Choel (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.

Li, Xiangdong, & Sun, Qin (2011). The review of text categorization research over Chinese Library Classification. American Journal of Engineering and Technology Research, 11(9), 2729-2734.

Miao, Yun-Qian, & Kamel, M. (2011). Pairwise optimized Rocchio algorithm for text categorization. Pattern Recognition, 32(2), 375-382.

Mitchell, T. M. (1997). Machine learning. New York, NY: McGraw-Hill.

Moens, Marie-Francine (2000). Automatic indexing and abstracting of document texts. Boston: Kluwer Academic Publishers.

Nidhi, & Gupta, V. (2011). Recent trends in text classification techniques. International Journal of Computer Applications, 35(6), 45-51.

Ruiz, M. E., & Srinivasan, P. (2002). Hierarchical text categorization using neural networks. Information Retrieval, 5(1), 87-118.

Sebastiani, F. (2002). Machine learning in automated text categorization. ACM Computing Surveys, 34(1), 1-47.

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.

Uĝuz, H. (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.

Vasuki, V. & Cohen, T. (2010). Reflective random indexing for semi-automatic indexing of the biomedical literature. Journal of Biomedical Informatics, 43(5), 694-700.

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. Proceedings of the Twenty-Fourth International Florida Artificial Intelligence Research Society Conference, 323-328.

Voorhees, E. M., & Harman, D. K. (2005). TREC: Experiment and evaluation in information retrieval. Cambridge, Mass.: MIT Press.

Wang, Tai-Yue, & Chiang, Huei-Min (2007). Fuzzy support vector machine for multi-class text categorization. Information Processing and Management, 43(4), 914-929.

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.

Yang, Y. (1999). An evaluation of statistical approaches to text categorization. Information Retrieval, 1(1-2), 69-90.

Yang, Y., & Pedersen, J. O. (1997). A comparative study on feature selection in text categorization. Proceedings of the 14th International Conference on Machine Learning (ICML '97), 412-420.

Yang, Y., & Liu, Xin (1999). A re-examination for text categorization methods. Proceedings of the 22th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval ('SIGIR 99), 42-49.

Yu, Bo, Xu, Zong-ben, & Li, Cheng-hua (2008). Latent semantic analysis for text categorization using neural network. Knowledge-Based Systems, 21(8), 900-904.

Zhang, J., & Yang, Y. (2003). Robustness of regularized linear classification methods in text categorization. Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '03), 190-197.

Zhang, Y., Tsai, F. S., & Kwee, A. T. (2011). Multilingual sentence categorization and novelty mining. Information Processing and Management, 47(5), 667-675.

• 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

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

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.

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.

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, & 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, 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

Lee, Jae Yun (2005a). Improving the performance of SVM text categorization with inter-document similarities. Journal of the Korean Society for Information Management, 22(3), 261-287.

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.

Yoon, Koo-ho (1999). Index & abstract. Seoul: Korean Library Association.