기계학습에 기초한 자동분류의 성능 요소에 관한 연구

An Analytical Study on Performance Factors of Automatic Classification based on Machine Learning

김판준 (Pan Jun Kim)*

초 록

국내 학술회의 논문으로 구성된 문헌집합을 대상으로 기계학습에 기초한 자동분류의 성능에 영향을 미치는 요소들을 검토하였다. 특히 구현이 쉽고 컴퓨터 처리 속도가 빠른 로치오 알고리즘을 사용하여 『한국정보관리학회 학술대회 논문집』의 논문에 주제 범주를 자동 할당하는 분류 성능 측면에서 분류기 생성 방법, 학습집합 규모, 가중치부여 기법, 범주 할당 방법 등 주요 요소들의 특성을 다각적인 실험을 통해 살펴보았다. 결과적으로 분류 환경 및 문헌집합의 특성에 따라 파라미터(β, λ)와 학습집합의 크기(5년 이상)를 적절하게 적용하는 것이 효과적이며, 동등한 성능 수준이라면 보다 단순한 단일 가중치부여 기법을 사용하여 분류의 효율성을 높일 수 있음을 발견하였다. 또한 국내 학술회의 논문의 분류는 특정 논문에 하나 이상의 범주가 부여되는 복수-범주 분류(multi-label classification)가 실제 환경에 부합한다고 할 수 있으므로, 이러한 환경을 고려하여 주요 성능 요소들의 특성에 기초한 최적의 분류 모델을 개발할 필요가 있다.

ABSTRACT

This study examined the factors affecting the performance of automatic classification for the domestic conference papers based on machine learning techniques. In particular, In view of the classification performance that assigning automatically the class labels to the papers in *Proceedings of the Conference of Korean Society for Information Management* using Rocchio algorithm, I investigated the characteristics of the key factors (classifier formation methods, training set size, weighting schemes, label assigning methods) through the diversified experiments. Consequently, It is more effective that apply proper parameters (β, λ) and training set size (more than 5 years) according to the classification environments and properties of the document set. and If the performance is equivalent, I discovered that the use of the more simple methods (single weighting schemes) is very efficient. Also, because the classification of domestic papers is corresponding with multi-label classification which assigning more than one label to an article, it is necessary to develop the optimum classification model based on the characteristics of the key factors in consideration of this environment.

키워드: 자동분류, 텍스트 범주화, 성능 요소, 학술회의 논문, 로치오 알고리즘, 복수-범주 분류, 기계학습
automatic classification, text categorization, performance factors, conference paper, rocchio algorithm, multi-label classification, machine learning

* 신라대학교 문헌정보학과 조교수(pjkim@silla.ac.kr)

■ 논문접수일자 : 2016년 5월 13일 ■ 최초심사일자 : 2016년 5월 25일 ■ 게재확정일자 : 2016년 6월 2일
■ 정보관리학회지, 33(2), 33-59, 2016. 〔http://dx.doi.org/10.3743/KOSIM.2016.33.2.033〕

1. 서 론

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 정보조직을 통한 분류 정보가 필수적으로 필요하다. 그러나 이러한 정보가 기본 항목으로 제공되고 있는 해외 학술데이터베이스와는 달리, 국내에서는 학술논문(학술회의 및 학술지 논문)에 대한 분류 정보가 제대로 제공되지 않고 있다(김판준, 이재윤, 2014). 예를 들면, 정부기관의 학술데이터베이스로서 KCI(한국연구재단의 한국학술지인용색인)와 RISS(한국교육학술정보원의 학술연구정보서비스)는 각각 8개와 10개의 학문분야로 구분한 대분류와 중분류 정보를 학술지(또는 발행기관)에 기초하여 제공하고 있을 뿐이다. 또한 민간 업체인 DBpia(누리미디어)와 KISS(한국학술정보(주))에서도 대분류와 중분류 정보를 학술지(또는 발행기관)에 기초하여 제공하고 있다. 따라서 현재 국내에서 논문 단위의 분류는 이루어지지 않고 있으며, 논문이 수록된 학술지(또는 발행기관)에 따라 대분류 또는 중분류까지만 제공하고 있는 실정이다.

학문분야별로 과거와 현재 연구의 흐름과 동향을 체계적으로 파악하여 향후의 연구 방향을 설정하기 위해서는 논문 단위의 분류정보 제공 및 활용이 필수적이다. 그러나 지난 수십 년간 국내에서 출판된 모든 학술논문에 대한 분류작업을 전문 인력을 활용하여 수작업으로 단기간에 추진하기는 불가능하다. 한 예로 2016년 4월 현재 KCI 통계에 따르면 5,134종의 학술지에 수록된 총 1,212,575건의 학술 논문이 KCI 웹사이트를 통해 서비스되고 있다(한국학술지인용색인, 2016). 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 전문인력의 부족은 물론, 예산상의 문제를 극복할 수 있는 대안으로 기계학습에 기초한 국내 학술논문의 자동분류를 적극적으로 모색할 필요가 있다.

본 연구의 목적은 국내 학술회의 논문집(『한국정보관리학회 학술대회 논문집』)에 수록된 논문을 대상으로 기계학습에 기초한 자동분류의 성능에 영향을 미치는 요소들의 특성을 규명하는 것이다. 국내 학술데이터베이스에서 학술회의 논문은 학술지 논문과 함께 대표적인 연구성과물로 서비스되고 있으나 논문 단위의 분류정보가 제공되지 않고 있다는 공통점이 있다. 따라서 국내 학술회의 논문에 대한 자동분류 실험을 통해 도출한 주요 성능 요소들의 특성은 향후 전체 학술논문의 자동분류를 추진하는 과정에서 의미 있는 시사점을 제공할 수 있을 것이다.

2. 이론적 배경

2.1 문헌의 자동분류

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 datasets, 20-Newsgroups datasets, OSUMED, TREC datasets, Ling-Spam datasets 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, 2009, 2013; 이재윤, 2005a, 2005b; Foulds & Frank, 2010; Khan, Baharudin, & Lee, 2010).

기계학습 기반의 자동분류는 다양한 분류 알고리즘과 이들을 서로 조합한 하이브리드 방식에 기초한 연구들이 빠른 성장과 확산의 양상을 보이고 있다(Chen, Lin, Xiong, Luo, & Ma, 2011; Jiang, Pang, Wu, & Kuang, 2012; Kumar & Gopal, 2010; Wu, 2009; Yu, Xu, & Li, 2008; Zhang, Yoshida, & Tang, 2011). 이외에도 기계학습 알고리즘과 전문가시스템의 조합(Li & Park, 2009; Villena-Román, Collada-Pérez, Lana-Serrano, & González-Cristóbal, 2011), 미분류 문헌의 활용(김판준, 이재윤, 2007; Torii, Yin, Nguyen, Mazumdar, Liu, Hartley, & Nelson, 2011), WordNet이나 Wikipedia와 같은 외부 정보의 활용(김용환, 정영미, 2012; 정은경, 2009) 등 자동분류의 성능 향상을 위한 다양하고도 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), 복수-범주 분류(multi-class classification 또는 multi-label classification) 등에 관한 논의가 활발하게 진행되고 있다(김종민, 유창동, 2014; AI-Salemi, Aziz, Juzaiddin, & Noah, 2015; Devi, Suganya, & Abirami, 2015; Read, Pfahringer, Holmes, & Frank, 2011; Schapire & Singer, 2000; Tarragó, Cornelis, Bello, & Herrera, 2014; Tsoumakas & Katakis, 2007).

학술논문을 대상으로 기계학습에 기초한 자동분류(또는 텍스트 범주화)를 적용한 국내의 연구는 많지 않다. 이러한 연구로는 김성희와 엄재은(2008), 김판준과 이재윤(김판준, 2006a, 2006b, 2008; 김판준, 이재윤, 2012, 2014), 심경 등(심경, 2006; 심경, 정영미, 2006)의 연구가 있다. 이외에 이용구(2009)가 국내 학술지 논문으로 구성된 KTSET을 대상으로 서로 다른 언어로 작성된 문헌의 자동분류를 위하여 여러 교차언어 텍스트 범주화(CLTC: Cross-Language Text Categorization) 방법들을 적용한 결과에서 학습집단 번역방법의 분류 성능이 비교적 좋은 것으로 보고한 바 있다. 그러나 지금까지 국내에서 실제 출판된 학술회의 논문을 대상으로 기계학습 기반의 알고리즘을 사용하여 자동분류의 성능에 영향을 미치는 요소들의 특성을 규명한 연구는 찾아볼 수 없다.

2.2 로치오 알고리즘

웹상에서 대규모 정보처리를 수행하는 실제 환경에서는 속도가 빠른 분류기가 필수적이며, 이에 따라 자동분류의 최근 트렌드는 속도가 빠른 분류기의 설계이다. 그러나 텍스트에 출현한 단어집합(Bag of Words)에 기초하는 문헌의 자동분류는 본질적으로 컴퓨터 처리 측면의 효율성(efficiency) 문제를 내재하고 있으며, 범주별 학습집합의 편차가 큰 실제 환경의 문헌집합(imbalanced document set)에 대한 분류 성능(effectiveness)이 낮게 나타나는 문제가 있다. 따라서 문헌의 자동분류에서 구현이 용이하고 컴퓨터 처리 속도 측면에서 상당한 장점을 가지고 있는 로치오 알고리즘의 적용에 대한 지속적인 연구와 재검토가 이어지고 있다(김판준, 2006b, 2008; Joachims, 1997; Khan, Baharudin, & Lee, 2010; Moschitti, 2003; Schapire, Singer, & Singhal, 1998; Tan, 2008). 특히, 텍스트 처리에도 적합한 로치오 알고리즘은 최근까지 다양한 측면에서의 개선 및 보완을 통해 다른 분류기와 거의 동등하거나 더 나은 성능을 보이는 것으로 보고되었다(김판준, 2006b, 2008; Tarragó et al., 2014; Zeng & Huang, 2011).

로치오 알고리즘의 기본 공식은 다음과 같다.

$$w_{ik} = \beta \frac{1}{|R_c|} \sum_{i \in R_c} w_{ik} - \gamma \frac{1}{|\overline{R_c}|} \sum_{i \in \overline{R_c}} w_{ik} \quad \text{(공식 1)}$$

여기서 R_c는 긍정문헌의 수, $\overline{R_c}$는 부정문헌의 수이며, 각 항의 센트로이드를 산출하는 방법(평균, 합, 최대값)에 따라 범주 프로파일을 다르게 생성할 수 있다. 또한, 실제 환경의 문헌집합에서는 대부분의 범주가 적은 수의 긍정문헌과 상당히 많은 수의 부정문헌을 가지는 경우가 많으므로 두 개의 항에 서로 다른 파라미터(β, γ) 값을 곱해주어 범주 프로파일을 생성한다.

2.3 자동분류의 성능 요소

문헌의 자동분류에서 성능에 영향을 미치는 특정 요소에 중점을 두어 분류 성능의 향상을 모색한 연구는 다양한 측면에서 시도되었다. 지금까지 이러한 연구들에서 주로 다루어진 대표적인 성능 요소는 사전처리, 문헌표현, 학습집합의 규모, 자질선정 방법, 가중치부여 기법, 분류 알고리즘 등이다(김판준, 2008; AI-Salemi et al., 2015; Aphinyanaphongs, Fu, Li, & Peskin, 2014; Forman, 2003; Forman & Kirshenbaum, 2008; Genkin, Lewis, & Madigan, 2007; Harish, Guru, & Manjunath, 2010; Jain & Nitin, 2015; Khan, Baharudin, & Lee, 2010; Korde & Mahender, 2012; Patra & Singh, 2013; Rogati & Yang, 2002; Yang & Pedersen, 1997). 특히 로치오 알고리즘을 이용한 문헌의 자동분류에서 분류 성능에 상당한 영향을 미치는 것으로 알려진 요소는 프로파일 생성 방법, 학습집합 규모, 자질선정 방법, 가중치부여 기법, 범주 할당 방법(단일-범주 분류, 복수-범주 분류) 등이다(김판준, 2006b, 2008; Cohen & Singer, 1999; Hull, 1994; Ittner, Lewis, & Ahn, 1995; Joachims, 1997; Kohavi & John, 1997; Moschitti, 2003; Pang & Jiang, 2013; Schapire, Singer, & Singhal, 1998; Tarragó

et al., 2014; Zeng & Huang, 2011).

로치오 알고리즘에서 범주 프로파일 생성 방법은 로치오 공식(공식 1)에서 각 자질의 값을 결정하는 센트로이드 산출 방법과 파라미터 설정의 두 가지 측면으로 구분할 수 있다. 첫째, 센트로이드 산출 방법은 대표적으로 세 가지(평균, 합, 최대값)가 있는데, 일반적으로 평균을 취하는 방식을 많이 사용한다. 둘째, 파라미터(β, γ)의 설정은 학습집합에서 긍정 및 부정문헌의 포함 여부와 비율을 조정하는 것으로, 크게 세 가지 방법(긍정문헌만 사용($\beta=1, \gamma=0$), 긍정 및 부정문헌을 동등하게 사용($\beta=\gamma=1$), 긍정 및 부정문헌을 함께 사용하되 양자 간의 비율을 조정($\beta=16, \gamma=4$ 등))이 있다(김판준, 2006b, 2008; Hull, 1994; Ittner, Lewis, & Ahn, 1995; Joachims, 1997; Schapire, Singer, & Singhal, 1998; Singhal, Mitra, & Buckley, 1997; Tarragó et al., 2014; Yang, 1999). 지금까지 범주 프로파일 생성 방법과 관련하여 다양한 응용과 문헌집합에 일반적으로 적용될 수 있는 최적의 조합은 제안된 바 없으며(Cohen & Singer, 1999; Moschitti, 2003), 연구자에 따라 응용 분야 및 분류 환경을 고려하여 적절한 방법을 채택하고 있다.

로치오 알고리즘을 이용한 문헌의 자동분류에서 학습집합의 크기는 대부분 범주 프로파일의 학습 과정에서 학습집합에 포함되는 문헌 수(또는 비율)를 의미한다. 일반적으로 학습에 사용된 문헌의 규모가 증가할수록 분류 성능이 향상되는 경향을 보이지만, 일정 수준으로 증가한 이후 또는 전체 학습집합을 모두 사용하는 경우에 분류 성능이 정체되거나 오히려 하락하는 사례도 보고되었다(김판준, 2006b; Yang & Pedersen, 1997; Zeng & Huang, 2011).

자질선정은 문헌의 자동분류에서 가장 많이 다루어진 성능 요소 중의 하나이다. 가장 단순한 빈도 기반의 자질선정에서부터 통계학적 기법이나 언어학적 기법 등을 활용한 자질선정에 이르기까지 다양한 방법들이 분류 성능의 향상을 목적으로 시도되었다(김판준, 2008; Al-Salemi et al., 2015; Chen & Chen, 2011; Jain & Nitin, 2015; Rogati & Yang, 2002). 이들 연구에서 통계학적 기법(카이제곱(χ^2), 정보획득량(Information Gain), 상호정보량(Mutual Information) 등)에 기초한 자질선정의 성능이 우수한 것으로 보고되었지만, 단순한 빈도 기반(TF, DF)의 자질선정으로도 비교적 좋은 성능을 기대할 수 있는 것으로 나타났다(김판준, 2006b; Forman, 2003; Yang & Pedersen, 1997). 또한, 최근의 연구에서는 자질선정이 기존의 분류 성능을 유지하는 수준에 머물거나 오히려 떨어뜨린다는 주장도 제기되었다(Aliferis, Statnikov, Tsamardinos, Mani, & Koutsoukos, 2010a, 2010b; Aphinyanaphongs et al., 2014).

문헌의 자동분류에서 일반적으로 사용되는 용어가중치는 정보검색 분야와 동일한 것으로 출현빈도와 역문헌빈도의 조합가중치이다. 이러한 출현빈도와 역문헌빈도의 조합 가중치부여 기법(tfidf)은 특정 문헌 내 출현빈도(tf)가 높으면서, 전체 문헌집단 내 출현빈도(df)는 낮은 용어가 더 중요하다는 가정에 기초한 것이다. 자동분류의 성능 요소로서 가중치부여 기법은 학습문헌에 부여된 범주 정보의 사용 여부에 따라 비지도 방법과 지도 방법으로 구분하기도 한다. 즉, 문헌 또는 문헌집합 내 출

현 정보만을 사용하는 경우를 비지도 가중치부여 방법(unsupervised weighting methods)이라 하고, 용어가 출현한 문헌에 부여된 범주 정보를 사용하는 경우는 지도 가중치부여 방법(supervised weighting methods)이라 한다. 또한, 가중치를 산출하는 정보원에 따라 문헌 요소(TF), 문헌집합 요소(DF), 범주 요소(카이제곱(χ^2), 유사계수(cosine, jaccard, jaccard), 적합성 가중치(relevance weight) 등)로 구분할 수도 있다. 여기서 범주 요소는 용어가 출현한 문헌에 부여된 범주 정보에 기초한 것으로, 이러한 범주 정보를 이용하여 용어의 중요성을 나타내는 가중치 값을 산출하는 여러 가중치부여 방법들이 제안되었다(Debole & Sebastiani, 2003; Liu, Loh, Yousef-Toumi, & Tor, 2007). 또한, 이들 세 가지 가중치 요소 가운데 하나의 요소만을 사용하는 단일 가중치부여 방법과 두 개 이상의 요소를 조합하는 조합 가중치부여 방법의 성능을 비교한 연구에서는, 서로 다른 문헌집단에 대한 가중치부여 기법의 실제 적용에 있어 분류 대상이 되는 문헌집단과 범주들의 특성을 함께 고려해야 할 필요성을 제기하였다(김판준, 2008).

지도학습 기반의 자동분류는 분류 대상 문헌에 여러 개의 범주 중 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 전자의 성능이 낮은 것으로 보고되었다(AI-Salemi et al., 2015; Khan, Baharudin, & Lee, 2010; Tsoumakas & Katakis, 2007). 실제 환경에서 문헌의 자동분류는 단일-범주 분류와 복수-범주 분류가 모두 발생할 수 있다. 특히, 학술논문은 특정 논문이 하나의 주제에 관련된 경우도 있지만, 여러 주제에 걸친 내용을 함께 다루는 경우도 적지 않다. 따라서 논문을 대상으로 하는 자동분류의 성능 요소로서 범주 할당 방법(단일-범주 분류와 복수-범주 분류)을 검토해 볼 필요가 있다.

3. 실험 설계

3.1 연구문제

본 연구는 컴퓨터 처리의 효율성(efficiency)과 분류 성능(effectiveness) 양자에서 강점을 가진 로치오 알고리즘을 사용하여 국내 학술회의 논문의 자동분류에 영향을 미치는 주요 요소들의 특성을 검토하였다. 특히, 실제 환경의 불균형 데이터 세트(imbalanced data set)라 할 수 있는 『한국 정보관리학회 학술대회 논문집』에 수록된 논문을 대상으로 로치오 알고리즘을 사용한 자동분류 실험을 수행하여 분류 성능에 영향을 미치는 요소들의 특성을 살펴보았다. 이를 위해 로치오 분류기의 성능에 상당한 영향을 미치는 프로파일 생성 방법, 학습집합의 규모, 가중치부여 기법, 범주 할당 방법(단일-범주 분류, 복수-범주 분류)의 네 가지 요소를 중심으로 분류 성능을 다각적으로 검토하였다. 이에 따라 본 연구에서 자동분류 실험을 통해 규명하고자 하는 연구문제는 다음과 같다.

첫째, 로치오 분류기는 프로파일 생성 방법에 따라 분류 성능에 차이가 있는가?
둘째, 로치오 분류기는 학습집합의 규모가 증가할수록 분류 성능이 향상되는가?
셋째, 로치오 분류기는 가중치부여 기법에 따라 분류 성능에 차이가 있는가?
넷째, 로치오 분류기는 범주 할당 방법에 따라 분류 성능에 차이가 있는가?

연구문제의 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(김판준, 2006b, 2008; 이재윤, 2005a; Moschitti, 2003; Uysal & Gunal, 2014; Yang & Liu, 1999; Zeng & Huang, 2011), 매크로 평균 F1은 빈도에 상관없이 모든 범주에 동일한 가중치를 주는 것으로 범주 당 평균을 산출한다(Jain & Nitin, 2015). 선행연구에서 전자는 고빈도 범주의 영향이 큰 반면, 후자는 저빈도 범주의 영향을 더 크게 받는 것으로 알려져 있다(이재윤, 2005b; Aphinyanaphongs et al., 2014; Sokolova & Lapalme, 2009). 본 연구는 실제 환경의 불균형 데이터세트(imbalanced data set)를 대상으로 자동분류의 성능에 영향을 주는 네 가지 주요 성능 요소를 다각적으로 검토하기 위하여, 서로 다른 특성을 가진 척도인 마이크로 평균 F1과 매크로 평균 F1을 함께 사용하였다(AI-Salemi et al., 2015; Devi, Suganya, & Abirami, 2015; Jain & Nitin, 2015; Pang & Jiang, 2013; Tan, 2008).

3.2 실험 문헌집합

본 연구에서 실험 문헌집합은 『한국정보관리학회 학술대회 논문집』에 수록된 최근 11년(2005년~2015년) 동안의 논문 349편을 대상으로, 이전 9년(2005년~2013년)의 281편(80.5%)을 학습집합, 최근 2년(2014년~2015년)의 68편(19.5%)은 검증집합으로 구분하였다. 현재 대부분의 국내 학술회의 논문에는 분류 범주가 부여되어 있지 않으므로, 한국정보관리학회 학술대회 논문 제출 시에 작성하는 "논문 발표 신청서"에 명시된 문헌정보학의 하위분야(22개)를 각 논문의 주제 범주로 간주하여 수작업 분류하였다(〈표 1〉 참조). 이러한 문헌정보학의 22개 주제범주는 한국연구재단의 "학술연구분야 분류표(2015)"상의 분류명과 대부분 일치하고 있으나 일부 항목에서 차이가 있다. 즉, 한국정보관리학회의 범주명에는 '정보공학'과 '정보교육'이 없는 반면, 한국연구재단의 분류명에는 '인공지능/전문가시스템', '네트워크/인터넷', '장서관리', '지식관리'가 없는 것이 차이점이다. 수작업 분류 과정에서 분류의 정확성을 확보하기 위하여, 각 논문이 발표된 학술대회 섹션명과 저자 및 소속, 본문의 내용을 종합적으로 검토하였다. 또한, 각 논문에 하나 이상의 분류 범주를 부여하여 단일-범주 분류와 복수-범주 분류를 비교할 수 있도록 하였다. 그 결과, 전체 22개 범주 중에서 '인공지능/전문가시스템', '장서관리', '서지학', '기타'를 제외한 18개 범주가 부여되었다. 이 중에서 학습집합에 최소 3개 이상의 논문이 존재하면서 검증이 가능한 10개 범주('계량정보학', '기록관리', '도서관/정보센터 경영', '디지털도서관/도서관자동화', '목록/메타

데이터', '문헌정보학일반', '이용자연구', '정보검색', '정보서비스', '정보시스템/데이터베이스')를 대상으로 자동분류 실험을 수행하였다. 이러한 문헌집합은 〈표 1〉에서 보는 바와 같이 대부분 학습문헌의 수가 비교적 적은 저빈도 범주로 구성되어 있으며, 각 범주별 학습집합의 편차가 큰 불균형 데이터(imbalanced data)라고 할 수 있다(AI-Salemi et al., 2015).

실험 문헌집합을 구성하는 국내 학술회의 논문집(『한국정보관리학회 학술대회 논문집』)의 논문은 대부분 4쪽의 분량이며, 초록의 길이도 학술지 논문에 비해 상대적으로 짧은 편이다. 이들 각 논문의 제목과 초록을 대상으로 단어 분리(tokenization), 형태소 분석, 불용어 제거 등 자동색인과 동일한 사전처리 과정을 거쳐 실험을 위한 키워드 집합을 구성하였다. 그 결과, 학습 및 검증에 사용되는 문헌의 길이가 평균 51.96개로 비교적 짧은 편에 속하며, 전체 키워드의 종수도 5,636개로 많지 않았다. 이는 신문 기사에 기초한 Reuters 데이터세트나 학술논문으로 구성된 OHSUMED 실험 문헌집합에 비해 낮은 수준에 있어 분류를 위한 자질집합이 상대적으로 소규모라고 할 수 있다(Yang & Pedersen, 1997). 따라서 본 연구에서는 최대한의 분류 자질을 확보하기 위해 자질선정을 수행하지 않고 전체 자질집합을 모두 사용하였

〈표 1〉 실험 문헌집합의 분류 범주 통계

번호	범주명 (한국정보관리학회)	분류명 (한국연구재단)	학습 문헌 수	검증 문헌 수	합계	비율
1	정보검색	정보검색	28	5	33	9.46
2	정보시스템/데이터베이스	데이터베이스	22	2	24	6.88
3	색인/자동색인	색인/초록; 자동색인/요약	1	0	1	0.29
4	초록/자동요약	색인/초록; 자동색인/요약	2	1	3	0.86
5	인공지능/전문가시스템	-	0	0	0	0.00
6	자동분류	자동분류/클러스터링	6	0	6	1.72
7	네트워크/인터넷	-	1	0	1	0.29
8	계량정보학	계량정보학	52	12	64	18.3
9	디지털도서관/도서관자동화	디지털도서관	15	5	20	5.73
10	시소러스, 온톨로지	전문용어/시소러스	7	0	7	2.01
11	목록/메타데이터	편목/메타데이터	8	2	10	2.87
12	분류	분류	4	0	4	1.15
13	도서관/정보센터경영	도서관/정보센터경영	46	13	59	16.9
14	장서관리	-	0	0	0	0.00
15	지식관리	-	3	0	3	0.86
16	기록관리	기록관리/보존	19	2	21	6.02
17	정보서비스	정보서비스	35	17	52	14.9
18	정보정책	정보/도서관정책	1	2	3	0.86
19	이용자연구	-	26	6	32	9.17
20	서지학	서지학	0	0	0	0.00
21	문헌정보학일반	문헌정보학일반	5	1	6	1.72
22	기타	기타문헌정보학	0	0	0	0.00

다. 실험 문헌집합에 대한 사전처리와 로치오 분류기의 구현은 파이썬 언어로 작성한 프로그램을 사용하였고, 추가적으로 한글 형태소 분석기(강승식, 2002: KLT, 2015), 마이크로소프트 엑셀 등을 사용하였다.

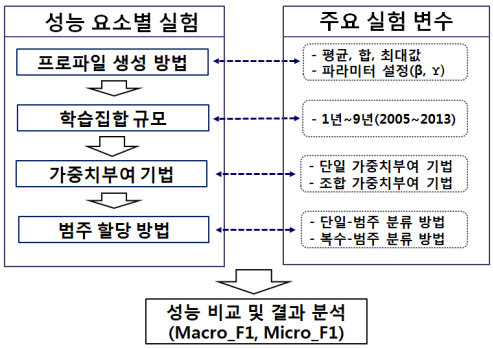
본 연구의 실험에 사용된 실험 문헌집합의 통계는 〈표 2〉와 같다.

〈표 2〉 실험 문헌집합 통계

항 목	내 역
전체 문헌 수/학습문헌 수/검증문헌 수	349/281/68
전체 범주 수/학습집합 범주 수/ 검증집합 범주 수	22/18/12
단일-범주 문헌빈도(최대/최소/평균)	1/1/1
복수-범주 문헌빈도(최대/최소/평균)	4/1/1.47
범주 당 긍정문헌 수(최대/최소/평균)	52/1/12.8
키워드 문헌빈도(최대/최소/평균)	183/1/3.51
학습문헌 당 키워드 종수(최대/최소/평균)	99/22/51.96
학습문헌 당 키워드 수(최대/최소/평균)	278/32/92.68
키워드 종수	5,636

3.3 실험 단계

〈그림 1〉에서 보는 바와 같이 분류 성능에 영향을 주는 네 가지 주요 요소별로 실험을 수행한 결과를 서로 다른 특성을 가진 두 가지 성능척도(mac_F1, mic_F1)에 기초하여 분석하였다. 첫째, 프로파일 생성 방법에 따른 실험은 로치오 기본 공식에서 긍정 문헌(β)과 부정 문헌(ɣ)의 포함 여부 및 비율에 관한 것으로 파라미터의 설정에 따른 분류 성능의 변화를 살펴보았다. 둘째, 학습집합의 규모에 따른 실험은 최근 1년(2013년)부터 전체 9년(2005년~2013년)까지 연차적으로 학습집합을 증가하는 경우에 분류 성능의 변화를 살펴보았다. 셋째, 가중치부여 기법에 따른 실험은 분류 자질에 여러 가중치부여 기법을 적용한 결과에 따른 성능을 살펴보았다. 넷째, 범주 할당 방법에 따른 실험은 각 문헌에 단일-범주를 할당하는 경우와 복수-범주를 할당하는 경우로 구분하여 분류 성능을 비교하였다.



〈그림 1〉 실험 단계

4. 실험 결과 및 분석

4.1 프로파일 생성 방법

로치오 분류기에서 프로파일의 생성 방법은 기본 공식에서 센트로이드 산출 방법(평균, 합, 최대값)과 각 항에 대한 파라미터 설정(β, ɣ)의 두 가지로 구분할 수 있다. 첫째, 사전실험에서 세 가지 센트로이드 산출 방법에 따른 성능에 유의한 차이가 없었기 때문에, 일반적으로 사용되는 평균을 센트로이드 산출 방법으로 채택하였다. 둘째, 파라미터 설정(β, ɣ)과 관련하여 긍정문헌과 부정문헌의 사용 여부 및 비율이라는 두 가지 측면에서 실험을 수행한 결과에서는 부정문헌의 사용 여부가 분류 성능에 상당한 영향을 미치는 반면(β=ɣ=1 vs. β=1,

〈표 3〉 파라미터 설정에 따른 로치오 기본형(baseline)의 성능: 긍정문헌만 사용(p1n0) vs. 긍정/부정문헌 사용(p1n1)

구분	mac_precision	mac_recall	mac_F1	mic_precision	mic_recall	mic_F1
p1n0	0.7045	0.6602	**0.6816**	0.6	0.6	**0.6**
p1n1	0.5847	0.6466	**0.6141**	0.5692	0.5692	**0.5692**

γ=0), 긍정문헌과 부정문헌의 비율을 조정하는 것은 성능에 큰 영향을 주지 않았다(β=16, γ=4 등). 따라서 로치오 분류기의 생성 방법에서 파라미터 설정은 긍정문헌만을 사용한 경우(β=1, γ=0)와 긍정문헌과 부정문헌을 함께 사용한 경우(β=γ=1)의 두 가지로 구분하여 실험 결과를 분석하였다.

〈표 3〉은 로치오 기본형(baseline)의 성능을 긍정문헌만 사용한 경우(p1n0: β=1, γ=0)와 긍정문헌과 부정문헌을 함께 사용한 경우(p1n1: β=γ=1)로 구분하여 비교한 것이다. 로치오 기본형(baseline)은 학습집합 9년(2005년~2013년)을 모두 사용하고, 가중치부여 기법으로 정보검색 및 자동분류에서 많이 사용되는 ltfidf(Log_tf*idf)를 적용하였다(김판준, 2008). 여기서 부정문헌의 사용 여부에 따라 두 가지 성능 척도에서 모두 상당한 차이가 있음을 알 수 있다. 즉, 매크로 F1(mac_F1)와 마이크로 평균 F1(mic_F1) 양자에서 부정문헌을 포함하는 경우보다 긍정문헌만을 사용한 경우의 성능이 더 높게 나타났다.

4.2 학습집합 규모

분류 성능에 영향을 미치는 주요 요소로서 학습집합의 규모에 따른 성능 변화는 전체 학습집합(9년: 2005년~2013년)을 가장 최근의 연도부터 연차적으로 1년씩 추가하여 살펴보았다. 〈표 4〉는 학습집합의 크기를 최근 1년(2013년)부터 전체 9년(2005년~2013년)까지 연차적으로 증가시킨 로치오 기본형(baseline)의 성능이다. 매크로와 마이크로 평균 F1 모두 5년(2009년~2013년) 이상의 학습집합을 사용한 경우에 일정 수준 이상으로 성능이 안정화되었으며, 이후 7년(2007년~2013년) 이상의 학습집합을 사용한 경우에 최고 성능을 보였다. 따라서 국내 학술회의 논문의 자동분류에서는 최소 5년 이상의 학습집합을 사용하는 경우에 지속적으로 일정 수준 이상의 안정적인 성능을 기대할 수 있는 것으로 나타났다. 그러나 전체 9년의 학습집합을 모두 사용한 결과가 항상 최고 성능을 보장하는 것은 아니었다. 이러한 결과는 학습에 사용된 자질의 규모가 늘어날수록 성능이 향상되지만, 일정 수준(3,000~4,000개) 이상으로 증가한 이후에는 정체되는 현상을 보고한 연구와 동일한 맥락에서 이해할 수 있다(Tan, 2008; Zeng & Huang, 2011). 또한, 매크로 평균 F1 측면에서는 긍정문헌만을 사용(p1n0)하는 것이 더 좋은 성능이지만, 마이크로 평균 F1 측면에서는 긍정문헌과 부정문헌을 함께 사용(p1n1)하는 것이 더 나은 성능을 보였다. 이에 따라 이후의 실험에서는 프로파일 생성 방법(p1n0 vs. p1n1)과 학습집합의 크기(5년~9년)를 동일하게 적용한 환경에서 다

〈표 4〉 학습집합의 연차적 증가에 따른 로치오 기본형(baseline)의 성능

구분	1년	2년	3년	4년	5년	6년	7년	8년	9년
p1n0_mac_F1	0.4825	0.4402	0.4827	0.394	0.5223	0.5201	0.6524	0.5398	**0.6816**
p1n1_mac_F1	0.4416	0.4616	0.4798	0.4995	0.542	0.5555	0.642	**0.6446**	0.6141
p1n0_mic_F1	0.5385	0.4923	0.5692	0.5538	0.5846	0.5692	**0.6**	**0.6**	**0.6**
p1n1_mic_F1	0.4615	0.4923	0.5538	0.5538	0.5846	0.5846	0.5846	**0.6308**	0.5692

른 성능 요소의 영향을 살펴보았다.

4.3 가중치부여 기법

본 연구에서 가중치부여 기법에 관한 실험은 크게 단일 가중치부여 기법(single weighting schemes)과 조합 가중치부여 기법(combined weighting schemes)으로 구분하여 수행하였다. 첫째, 단일 가중치부여 기법은 문헌 및 문헌집합 내 출현정보에 기초한 용어빈도(tf, ltf/log_tf, otf/okapi_tf)와 역문헌빈도(idf), 문헌에 부여된 범주정보에 기초한 공식(chi/카이제곱, jac/자카드, mi/상호정보량)을 사용하였다. 둘째, 조합 가중치부여 기법은 2개 이상의 단일 가중치를 조합하여 사용하였다.

〈표 5〉는 단일 가중치부여 기법의 성능을 매크로 평균 F1(mac_F1)으로 산출한 결과이다. 여기서 가장 좋은 성능은 프로파일 생성 방법으로 긍정문헌만을 사용(p1n0)하고 7년의 학습집합과 단일 가중치(otf)를 적용한 경우이며, 그 다음은 긍정문헌과 부정문헌을 함께 사용하고, 전체 학습집합(9년)과 단일 가중치(idf)를 사용한 것이다. 전반적으로 단일 가중치부여 기법 중에서 출현정보에 기초한 기법들이 범주정보에 기초한 기법들보다 높은 성능을 보였으며, 상호정보량(mi)은 다른 기법들에 비해 상당히 낮은 수준의 성능을 보였다. 그리고 동일한 가중치부여 기법을 적용하였을 때, 긍정문헌만을 사용한 경우(p1n0)가 긍정문헌과 부정문헌을 함께 사용(p1n1)한 것보다 다소 나은 성능이었다. 학습집합의 규모 측면에서는 각 가중치부여 기법별로 최고 성능이 6년(2008년~2013년)에서 9년(2005년~2013년)까지 다양하게 나타났다. 또한, 〈표 6〉은 단일 가중치부여 기법의 성능을 마이크로 평균 F1(mic_F1)으로 산출한 결과이다. 가장 좋은 성능은 프로파일 생성 방법으로 긍정문헌만을 사용(p1n0)하고 8년의 학습집합과 출현정보에 기초한 단일 가중치(ltf, otf)를 적용한 경우이며(0.6615), 그 다음은 긍정문헌과 부정문헌(p1n1)을 함께 사용하고, 8년의 학습집합과 단일 가중치(otf) 그리고 전체 학습집합과 가중치(idf)를 적용한 것이다(0.6308). 여기서 〈표 5〉와 동일하게 출현정보에 기초한 기법들의 성능이 범주정보에 기초한 기법들보다 더 높은 수준이었고, 상호정보량(mi)이 다른 기법들에 비해 상당히 낮은 수준을 보였다. 동일한 가중치부여 기법을 적용한 경우에 긍정문헌만을 사용한 경우(p1n0)가 긍정문헌과 부정문헌을 함께 사용(p1n1)한 것보다 조금 더 나은 성능을 보이며, 학습집합의 크기 측면에서 각 가중치부여 기법별로 최고 성능이 6년(2008년~2013년)에서 9년(2005년~2013년)까지 다양하게 나타나 〈표 5〉와 유사한 양상을 보였다.

〈표 5〉 단일 가중치부여 기법의 성능: 매크로 평균 F1(mac_F1)

구분	가중치	5년	6년	7년	8년	9년
긍정only (p1n0)	tf	0.4657	0.4682	**0.601**	0.5227	0.564
	ltf	0.5248	0.5284	0.6355	0.6283	**0.6479**
	otf	0.5494	0.5368	**0.6638**	0.6541	0.657
	idf	0.5295	0.4903	0.6467	0.4995	**0.6619**
	chi	0.4545	**0.5973**	0.5642	0.5454	0.4838
	jac	0.4545	**0.5973**	0.5642	0.5454	0.4406
	mi	0.2372	0.289	**0.2938**	0.2736	0.291
긍정&부정 (p1n1)	tf	0.37	0.4149	0.4711	0.4797	**0.4818**
	ltf	0.54	0.5072	**0.6228**	0.5996	0.5871
	otf	0.4433	0.4476	0.6083	**0.632**	0.5813
	idf	0.4927	0.5111	0.6392	0.6043	**0.6521**
	chi	0.3715	0.4873	0.5058	0.4848	**0.5974**
	jac	0.3715	0.4873	0.5058	0.4848	**0.5355**
	mi	0.2325	0.2869	**0.3149**	0.3024	0.2875

〈표 6〉 단일 가중치부여 기법의 성능: 마이크로 평균 F1(mic_F1)

구분	가중치	5년	6년	7년	8년	9년
긍정only (p1n0)	tf	0.5538	0.5538	**0.6**	0.4769	0.5538
	ltf	0.5846	0.5692	0.5846	**0.6615**	0.6
	otf	0.6154	0.5846	0.6154	**0.6615**	0.6308
	idf	0.5692	0.5385	**0.6**	0.5538	**0.6**
	chi	0.5231	**0.6**	0.5231	0.5231	0.4615
	jac	0.5231	**0.6**	0.5231	0.5231	0.4308
	mi	0.2923	0.3077	0.2923	0.2769	0.1846
긍정&부정 (p1n1)	tf	0.5231	**0.5077**	0.4923	0.4462	0.4615
	ltf	0.6	0.5692	0.6	**0.6154**	0.5538
	otf	0.5231	0.5385	0.5846	**0.6308**	0.5692
	idf	0.5692	0.5692	0.6	0.6	**0.6308**
	chi	0.4769	0.5231	0.5385	0.5077	**0.5692**
	jac	0.4873	0.5231	0.5385	0.5077	**0.5538**
	mi	0.2769	**0.3077**	**0.3077**	**0.3077**	0.1846

〈표 7〉은 단일 가중치부여 기법에서 성능이 좋았던 출현정보 기반의 가중치(otf, idf)와 범주정보 기반의 가중치(chi, jac)를 조합한 여러 가중치부여 기법의 성능을 매크로 평균 F1(mac_F1)으로 산출한 결과이다. 여기서 가장 좋은 성능은 프로파일 생성 방법으로 긍정문헌만을 사용(p1n0)하고 8년의 학습집합과 조합가중치(otfidfjac/otf*idf*jac)를 적용한 경우이며(0.6675), 그 다음은 긍정문헌과 부정문헌(p1n1)을 함께 사용하고, 7년의 학습집합과 조합가중

〈표 7〉 조합 가중치부여 기법의 성능: 매크로 평균 F1(mac_F1)

구분	가중치	5년	6년	7년	8년	9년
긍정only (p1n0)	otfidf	**0.6**	0.5538	0.5846	**0.6**	0.5846
	otfchi	0.4896	0.4465	**0.6003**	0.5882	0.5952
	otfjac	0.4596	0.5304	0.6105	0.5792	0.5401
	idfchi	0.5274	0.5948	**0.6511**	0.5981	0.5746
	idfjac	0.5578	0.5396	**0.6645**	0.6183	0.5418
	otfidfchi	0.4753	0.4757	**0.6393**	0.5948	0.5766
	otfidfjac	0.5074	0.5402	0.6061	**0.6675**	0.5514
긍정&부정 (p1n1)	otfidf	0.5427	0.4969	**0.6498**	0.5301	0.6366
	otfchi	0.3692	0.3774	0.4557	**0.459**	0.4532
	otfjac	0.4805	0.4958	0.5788	**0.5831**	0.5222
	idfchi	0.4919	0.5031	0.5283	0.5613	**0.6251**
	idfjac	0.4838	0.5554	**0.5862**	0.5606	0.5614
	otfidfchi	0.4207	0.4929	0.5215	0.46	0.5719
	otfidfjac	0.4091	0.4594	0.5371	**0.5933**	0.54

치(otfidf/otf*idf)를 적용한 것이다(0.6498). 전반적으로 단일 가중치부여 기법보다는 다소 나은 성능을 보이고 있지만, 현저한 성능 차이는 없는 것으로 나타났다. 동일한 조합 가중치부여 기법을 적용하였을 때 긍정문헌만을 사용한 경우(p1n0)가 긍정문헌과 부정문헌을 함께 사용(p1n1)한 것보다 높은 성능을 보이며, 학습집합의 크기 측면에서는 최고 성능이 7년(2007년~2013년)에서 9년(2005년~2013년)까지 다양하게 나타났다. 한편, 〈표 8〉은 이러한 조합 가중치

〈표 8〉 조합 가중치부여 기법의 성능: 마이크로 평균 F1(mic_F1)

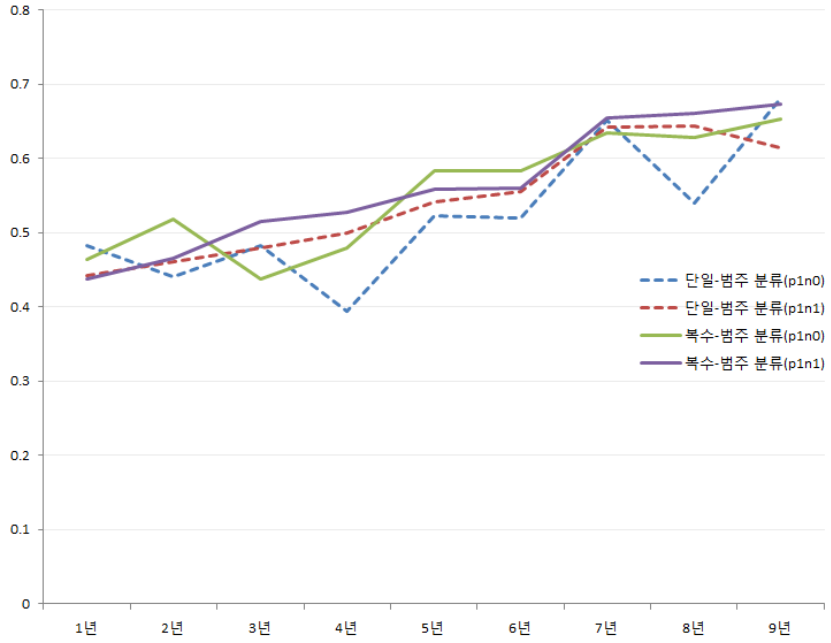
구분	가중치	5년	6년	7년	8년	9년
긍정only (p1n0)	otfidf	0.5846	0.5538	0.5692	**0.6**	**0.6**
	otfchi	0.5077	0.5231	**0.5692**	0.5538	0.5231
	otfjac	0.5385	0.5846	**0.6**	0.5692	0.4923
	idfchi	0.6	**0.6154**	0.6	0.5692	0.5385
	idfjac	0.6	0.6	**0.6154**	0.6	0.5077
	otfidfchi	0.5385	0.5538	**0.6**	0.5846	0.5846
	otfidfjac	0.5846	0.5692	0.6	**0.6308**	0.5231
긍정&부정 (p1n1)	otfidf	0.5075	0.4969	0.6229	0.5301	**0.6594**
	otfchi	0.5231	0.5077	**0.5385**	**0.5385**	0.5231
	otfjac	0.5692	0.5538	**0.6**	**0.6**	0.5385
	idfchi	0.5538	0.5692	0.5692	**0.5846**	**0.5846**
	idfjac	0.5538	0.5692	**0.5846**	**0.5846**	0.5538
	otfidfchi	0.5385	**0.5692**	0.5538	0.5385	0.5231
	otfidfjac	0.5385	0.5538	0.5692	**0.6**	0.5538

부여 기법의 성능을 마이크로 평균 F1(mic_F1)으로 산출한 결과이다. 최고 성능은 프로파일 생성 방법으로 긍정문헌만을 사용(p1n0)하고 8년의 학습집합과 조합가중치(otfidfjac)를 적용한 경우이며(0.6308), 그 다음이 긍정문헌과 부정문헌(p1n1)을 함께 사용하고, 전체 9년의 학습집합과 조합가중치(otfidf)를 적용한 것이다(0.6594). 또한 동일한 조합 가중치부여 기법을 적용하였을 때, 긍정문헌만을 사용한 경우(p1n0)보다 긍정문헌과 부정문헌을 함께 사용(p1n1)한 경우에 성능이 더 높지만, 학습집합의 크기 측면에서는 최고 성능이 6년(2008년~2013년)에서 9년(2005년~2013년)까지 다양하게 나타나 이전 실험의 결과와 유사한 양상을 보였다.

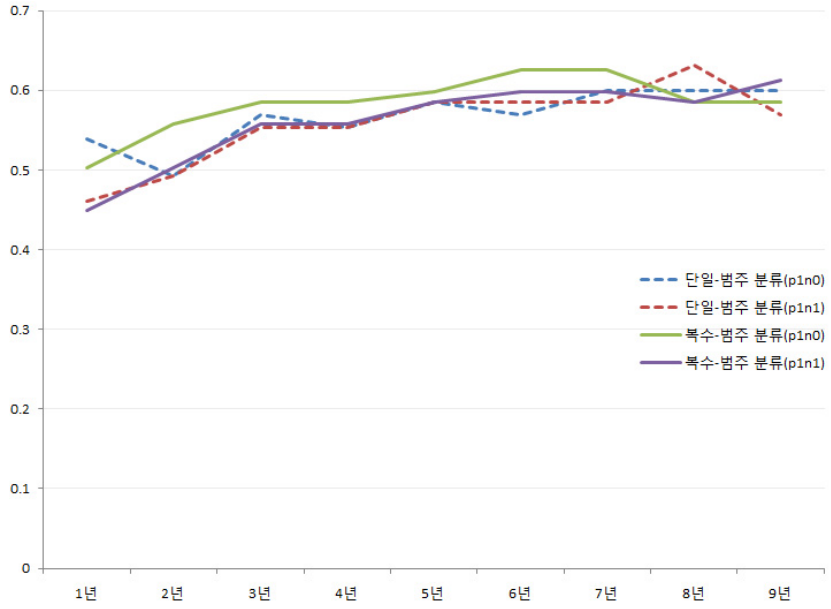
4.4 범주 할당 방법

문헌의 자동분류는 범주를 할당하는 방법에 따라 단일-범주 분류 방법(single-label classification)과 복수-범주 분류(multi-label classification)로 구분할 수 있다. 단일-범주 분류 방법은 입력 문헌에 다수의 범주 중에서 하나의 범주를 할당하는 반면, 복수-범주 분류 방법은 해당 문헌에 여러 개의 범주를 할당한다. 최근의 연구에서는 단일-범주 분류에 비해 복수-범주 분류의 성능이 낮은 것으로 보고된 바 있다(AI-Salemi et al., 2015). 실제 환경의 국내 학술회의 논문의 분류에서 이러한 두 가지 유형의 분류가 모두 발생할 수 있으므로, 지금까지 단일-범주 분류 방법을 전제로 수행한 실험 환경을 복수-범주 분류 방법에 동일하게 적용하여 양자의 결과를 비교하였다. 먼저, 로치오 기본형(baseline)을 사용하여 단일-범주 분류 방법과 복수-범주 분류 방법의 성능을 비교한 결과는 〈그림 2〉, 〈그림 3〉과 같다. 〈그림 2〉는 매크로 평균 F1(mac_F1) 기준으로 단일-범주 분류와 복수-범주 분류의 최고 성능 간에 큰 차이가 없음을 보여준다(0.6816 vs. 0.6737). 그러나 긍정문헌만을 사용하는 경우(p1n0)와 긍정문헌과 부정문헌을 모두 사용하는 경우(p1n1)를 비교하면, 단일-범주 분류와 복수-범주 분류 간에 상당한 차이가 발생하였다(0.6816 vs. 0.6526 ; 0.6446 vs. 0.6737). 마이크로 평균 F1(mic_F1) 기준의 〈그림 3〉에서도 단일-범주 분류와 복수-범주 분류의 최고 성능 간에는 큰 차이가 없었다(0.6308 vs. 0.6259). 그러나 긍정문헌만을 사용하는 경우(p1n0)와 긍정문헌과 부정문헌을 모두 사용하는 경우(p1n1)를 비교하면, 단일-범주 분류와 복수-범주 분류 간에 다소 차이가 있었다(0.6 vs. 0.6259 ; 0.6308 vs. 0.6122). 따라서 분류 성능에 영향을 미치는 다른 요소를 함께 고려하지 않고, 복수-범주 분류의 성능이 단일-범주 분류보다 하위에 있다고 판단하는 것은 문제가 있다. 또한, 이전의 단일-범주 분류를 전제로 한 실험에서와 마찬가지로 5년 이상의 학습집합을 사용하는 경우에는 분류 성능이 일정 수준 이상으로 유지되는 것으로 나타났다.

다음으로, 프로파일 생성 방법(p1n0: $\beta=1$, $\gamma=0$ vs. p1n1: $\beta=\gamma=1$)과 학습집합 크기(5년 이상), 그리고 가중치부여 기법 중에서 최고 성능을 보인 가중치들을 사용하여, 단일-범주 분류와 복수-범주 분류의 성능을 비교하였다. 〈표 9〉는 긍정문헌만으로 생성된 프로파일(p1n0)과 5년 이상의 학습집합을 사용한 경우에 상위 5개 가중치부여 기법별로 단일-범주 분류와 복수-범주 분



〈그림 2〉 로치오 기본형(baseline)을 사용한 단일-범주 분류와 복수-범주 분류의 성능: 매크로 평균 F1(mac_F1)



〈그림 3〉 로치오 기본형(baseline)을 사용한 단일-범주 분류와 복수-범주 분류의 성능: 마이크로 평균 F1(mic_F1)

〈표 9〉 단일-범주 분류와 복수-범주 분류의 성능 비교: 긍정문헌과 부정문헌 함께 사용(p1n1), 가중치부여 기법별 성능: 매크로 평균 F1(mac_F1)

구분	가중치	5년	6년	7년	8년	9년
단일-범주분류 (p1n1)	baseline	0.542	0.5555	0.642	**0.6446**	0.6141
	ltf	0.54	0.5072	**0.6228**	0.5996	0.5871
	otf	0.4433	0.4476	0.6083	**0.632**	0.5813
	idf	0.4927	0.5111	0.6392	0.6043	**0.6521**
	otfidf	0.5427	0.4969	**0.6498**	0.5301	0.6366
	otfidfjac	0.4091	0.4594	0.5371	**0.5933**	0.54
복수-범주분류 (p1n1)	baseline	0.5591	0.5595	0.655	0.6607	**0.6737**
	ltf	0.5681	0.5207	0.6232	**0.6798**	0.6278
	otf	0.5071	0.5326	0.6836	**0.6737**	0.6579
	idf	0.5292	0.5495	0.6502	**0.6828**	0.6447
	otfidf	0.5332	0.5454	0.6618	**0.6849**	0.6795
	otfidfjac	0.4643	0.4506	0.5587	**0.6139**	0.5569

류의 성능을 매크로 평균 F1(macro_F1) 척도로 제시한 것이다. 여기서 긍정문헌만을 사용하는 경우에는 단일-범주 분류보다 복수-범주 분류가 더 나은 성능을 보였다(otf(0.6638) 〈 otf(0.7017)). 또한 〈표 10〉은 긍정문헌과 부정문헌을 동등하게 사용하여 생성된 프로파일(p1n1)과 5년 이상의 학습집합을 사용한 경우에 상위 5개 가중치부여 기법별로 단일-범주 분류와 복수-범주 분류의 성능을 매크로 평균 F1(macro_F1) 척도로 제시한 것이다. 이 경우에도 단일-범주 분류보다 복수-범주 분류가 더 나은 성능을 보였다(idf(0.6521) 〈 otfidf(0.6849)). 따라서 매크로 평균 F1 기준으로 5년 이상의 학습집합을 사용하면 부정문

〈표 10〉 단일-범주 분류와 복수-범주 분류의 성능 비교: 긍정문헌만 사용(p1n0), 가중치부여 기법별 성능: 매크로 평균 F1(mac_F1)

구분	가중치	5년	6년	7년	8년	9년
단일-범주분류 (p1n0)	baseline	0.5223	0.5201	0.6524	0.5398	**0.6816**
	ltf	0.5248	0.5284	0.6355	0.6283	**0.6479**
	otf	0.5494	0.5368	**0.6638**	0.6541	0.657
	idf	0.5295	0.4903	0.6467	0.4995	**0.6619**
	otfidf	0.5075	0.4969	0.6229	0.5301	**0.6594**
	otfidfjac	0.4091	0.4594	0.5371	**0.5933**	0.54
복수-범주분류 (p1n0)	baseline	0.5834	0.5837	0.6346	0.6284	**0.6526**
	ltf	0.5415	0.5479	0.6698	**0.6927**	0.669
	otf	0.5313	0.553	**0.7017**	0.6862	0.6536
	idf	0.4868	0.5223	0.5812	0.5976	**0.6393**
	otfidf	0.546	0.5196	0.626	**0.6329**	0.6237
	otfidfjac	0.5182	0.5753	0.6768	**0.6792**	0.6348

헌의 포함 여부에 상관없이 단일-범주 분류보다 복수-범주 분류의 성능이 더 높게 나타나며, 보다 단순한 단일 가중치부여 기법을 사용하여 최고 성능을 도출할 수 있음을 발견하였다((otf(0.7017)).

〈표 11〉과 〈표 12〉는 프로파일 생성 방법 두 가지(p1n0, p1n1)와 5년 이상의 학습집합을 사용한 경우에 상위 5개 가중치부여 기법별로 단일-범주 분류와 복수-범주 분류의 성능을 마이크로 평균 F1(mic_F1) 척도로 제시한 것이다. 매크로 평균 F1 척도로 본 결과와 달리, 긍정문헌만을 사용한 결과인 〈표 11〉에서는 최고 성능 기준으로 단일-범주 분류가 복수-범주 분류보다 나은 성능을 보였다(ltf, otf(0.6615) 〉 otf,

〈표 11〉 단일-범주 분류와 복수-범주 분류의 성능 비교: 긍정문헌만 사용(p1n0), 가중치부여 기법별 성능: 마이크로 평균 F1(mic_F1)

구분	가중치	5년	6년	7년	8년	9년
단일-범주분류 (p1n0)	baseline	0.5846	0.5692	**0.6**	**0.6**	**0.6**
	ltf	0.5846	0.5692	0.5846	**0.6615**	0.6
	otf	0.6154	0.5846	0.6154	**0.6615**	0.6308
	idf	0.5692	0.5385	**0.6**	0.5538	**0.6**
	otfidf	0.5846	0.5538	0.5692	**0.6**	**0.6**
	otfidfjac	0.5846	0.5692	0.6	0.6308	0.5231
복수-범주분류 (p1n0)	baseline	0.5986	**0.6259**	**0.6259**	0.585	0.585
	ltf	0.5986	0.5714	0.585	**0.6122**	0.585
	otf	0.585	0.585	**0.6259**	0.6122	0.585
	idf	0.585	0.5986	0.585	0.5714	**0.6122**
	otfidf	0.5986	**0.6122**	**0.6122**	0.5986	0.5714
	otfidfjac	0.5714	0.585	**0.6259**	0.5986	0.5578

〈표 12〉 단일-범주 분류와 복수-범주 분류의 성능 비교: 긍정문헌과 부정문헌 함께 사용(p1n1), 가중치부여 기법별 성능: 마이크로 평균 F1(mic_F1)

구분	가중치	5년	6년	7년	8년	9년
단일-범주분류 (p1n1)	baseline	0.5846	0.5846	0.5846	**0.6308**	0.5692
	ltf	0.6	0.5692	0.6	**0.6154**	0.5538
	otf	0.5231	0.5385	0.5846	**0.6308**	0.5692
	idf	0.5692	0.5692	0.6	**0.6**	**0.6308**
	otfidf	**0.6**	0.5538	0.5846	**0.6**	0.5846
	otfidfjac	0.5385	0.5538	0.5692	**0.6**	0.5538
복수-범주분류 (p1n1)	baseline	0.585	0.5986	0.5986	0.585	**0.6122**
	ltf	**0.5986**	**0.5986**	**0.5986**	0.585	**0.5986**
	otf	0.5714	0.5578	**0.6122**	0.585	0.5714
	idf	0.5442	0.585	0.585	**0.6259**	0.6122
	otfidf	0.585	0.5714	**0.5986**	**0.5986**	**0.5986**
	otfidfjac	**0.5442**	0.517	**0.5442**	**0.5442**	0.5034

otfidfjac(0.6259)). 부정문헌을 포함하는 경우인 〈표 12〉에서도 단일-범주 분류가 복수-범주 분류보다 조금 나은 성능을 보였다(otf, idf(0.6308) 〉 idf(0.6259)). 결과적으로 마이크로 평균 F1 기준으로 5년 이상의 학습집합을 사용하는 경우에는 〈표 9〉, 〈표 10〉과는 달리 단일-범주 분류가 복수-범주 분류보다 성능이 더 높게 나타나는 차이가 있는 반면, 단순한 단일 가중치 기법으로 최고 성능을 도출할 수 있다는 점은 일치하였다((otf=idf(0.6308)).

4.5 종합 분석

국내 학술회의 논문을 대상으로 로치오 분류기의 성능에 영향을 주는 네 가지 주요 요소별로 실험을 수행한 결과를 서로 다른 특성을 갖는 두 가지 성능척도(mac_F1, mic_F1)에 기초하여 분석하였다. 먼저, 로치오 기본형(baseline)을 비롯한 대부분의 실험 결과에서 매크로 평균 F1(mac_F1)의 성능 수준이 마이크로 평균 F1(mic_F1)보다 전반적으로 높게 나타나, 전자가 범주 중심의 척도로서 상대적으로 저빈도 범주에 유리한 경향이 있음을 확인하였다(Tan, 2008; Yang & Liu, 1999).

다음으로, 본 연구에서 설정한 연구문제를 중심으로 분석한 결과는 다음과 같다.

첫째, 프로파일 생성 방법에 따라 성능에 차이가 있는 가를 로치오 기본형(baseline)을 사용하여 두 가지 측면에서 검토하였다. 먼저, 센트로이드 산출 방법 측면에서는 세 가지 방법(평균, 합, 최대값) 간에 큰 차이가 없었다. 다음으로 파라미터 설정(β, ɣ)과 관련하여 긍정문헌과 부정문헌의 사용 여부 및 비율에 관한 실험에서는 부정문헌의 사용 여부가 분류 성능에 상당한 영향을 미치는 반면(β=ɣ=1 vs. β=1, ɣ=0), 긍정문헌과 부정문헌의 비율을 조정하는 것은 성능에 큰 영향을 주지 않았다(β=16, ɣ=4 등). 이에 따라 긍정문헌만을 사용하는 경우와 긍정문헌과 부정문헌을 함께 사용하는 경우의 두 가지(β=ɣ=1 vs. β=1, ɣ=0)로 구분하여 다른 성능 요소를 적용해 본 대부분의 실험에서, 양자가 최고 성능을 보이는 경우가 번갈아서 나타났다. 결과적으로 로치오 분류기는 프로파일 생성 방법에서 부정문헌의 사용 여부에 따라 상당한 성능 차이가 발생하므로, 실험 문헌집합의 특성 및 다른 성능 요소와의 조합을 고려하여 부정문헌의 사용 여부를 결정하여야 할 것이다.

둘째, 학습집합의 크기가 증가함에 따라 분류 성능이 향상되는 가를 알아보기 위해 로치오 기본형(baseline)을 사용하여 학습집합의 크기를 최근 1년(2013년)부터 전체 9년(2005년~2013년)까지 연차적으로 증가시킨 성능을 살펴보았다. 국내 학술회의 논문을 대상으로 하는 자동분류는 5년 이상의 학습집합을 사용하면 일정 수준 이상의 성능이 유지되며, 7년(2007년~2013년)부터 전체 9년까지의 학습집합을 사용한 경우에 최고 성능을 보였다. 따라서 로치오 분류기를 사용한 국내 학술회의 논문의 자동분류는 최소 5년 이상의 학습집합을 사용하는 경우에 일정 수준 이상의 안정적인 성능을 기대할 수 있으며, 최고 성능을 도출하기 위해서는 다른 성능 요소와의 조합을 고려하여 적절한 학습집합의 규모를 설정할 필요가 있다.

셋째, 가중치부여 방법에 따라 로치오 분류

기의 성능에 차이가 있는 가를 알아보기 위해 단일 가중치와 조합 가중치로 구분하여 여러 가중치부여 기법을 적용한 성능을 살펴보았다. 단일 가중치 중에서는 매크로와 마이크로 평균 F1 척도 양자에서 문헌(집합) 내 출현정보에 기초한 가중치(ltf, otf, idf)가 범주정보 기반의 가중치(chi, jac, mi)보다 높은 성능 수준을 보였다. 또한, 조합 가중치 실험에서는 두 가지 기법(otfidf, otfidfjac)이 매크로와 마이크로 평균 F1 척도 양자에서 최고 성능을 보였다. 그러나 이들 단일 가중치와 조합 가중치의 최고 성능 간에는 현저한 차이가 없는 것으로 나타나, 컴퓨터 처리의 효율성 측면에서 보다 단순한 단일 가중치부여 기법을 사용하는 것이 효율적임을 발견하였다.

넷째, 문헌에 범주를 할당하는 방법(단일-범주 분류, 복수-범주 분류)에 따라 분류 성능에 차이가 있는 지를 알아보았다. 이를 위해, 단일-범주 분류를 전제로 한 이전의 실험 환경을 동일하게 적용하여 복수-범주 분류 실험을 수행하고 양자의 성능을 비교하였다. 먼저, 로치오 기본형(baseline)을 사용한 실험에서는 매크로 평균 F1과 마이크로 평균 F1 척도 양자에서 단일-범주 분류와 복수-범주 분류의 최고 성능 간에 큰 차이가 없는 것으로 나타났다. 그러나 프로파일 생성 방법(β, γ)과 학습집합의 크기를 다르게 적용하는 경우에는 양자 간에 성능 차이가 있었다. 다음으로, 단일-범주 분류 실험에서 좋은 성능을 보인 성능 요소들을 적용하여 복수-범주 분류 실험을 수행하여 이들 두 가지 방법의 성능을 비교하였다. 즉, 이전 실험에서 좋은 성능을 보인 프로파일 생성 방법(p1n0: $\beta=1, \gamma=0$ vs. p1n1: $\beta=\gamma=1$), 학습집합 5년 이상, 그리고 가중치부여 기법(ltf, otf, idf, otfidf, otfidfjac)을 동일하게 적용하여, 두 가지 범주 할당 방법의 성능을 비교하였다. 여기서 매크로 평균 F1의 최고 성능 기준으로는 단일-범주 분류보다 복수-범주 분류가 더 높은 성능을 보였다. 반면, 마이크로 평균 F1의 최고 성능 기준으로는 단일-범주 분류가 복수-범주 분류보다 조금 나은 수준이지만, 양자 간의 성능 차이는 크지 않았다. 따라서 두 가지 방법 중 어느 하나의 성능이 더 높다고 단정할 수는 없으며, 분류 환경과 다른 성능 요소를 함께 고려하여 최적의 방법을 모색하여야 할 것이다.

5. 결 론

국내 학술회의 논문집의 논문을 대상으로 기계학습에 기초한 자동분류의 성능에 영향을 미치는 요소들의 특성을 살펴보았다. 특히 구현이 쉽고 컴퓨터 처리 속도가 빠른 로치오 분류기를 사용하여 『한국정보관리학회 학술대회 논문집』의 논문에 주제 범주를 자동 할당하는 분류 성능 측면에서 주요 요소들의 특성을 분석한 결과는 다음과 같다.

첫째, 로치오 분류기는 프로파일 생성 방법에 따라 분류 성능에 차이가 있다. 특히, 파라미터 설정에서 부정문헌의 사용 여부($\beta=\gamma=1$ or $\beta=1, \gamma=0$)가 성능에 상당한 영향을 미치는 것으로 나타났다. 둘째, 로치오 분류기는 학습집합의 크기가 증가할수록 분류 성능이 향상된다. 즉, 학습집합의 규모를 연차적으로 증가시키는 경우에 최소 5년 이상의 학습집합을 사용하면 일정 수준 이상의 성능 향상을 기대할 수

있다. 그러나 전체 학습집합(9년)을 모두 사용하는 경우에 항상 최고 성능을 보이는 것은 아니며, 7년에서 9년까지 최고 성능인 경우가 다양하게 나타났다. 셋째, 로치오 분류기는 가중치부여 기법에 따라 분류 성능에 차이가 있다. 단일 가중치와 조합 가중치로 구분하여 실험한 결과에서 세 개의 단일 가중치부여 기법(ltf, otf, idf)과 두 개의 조합 가중치부여 기법(otfidf, otfidfjac)이 좋은 성능을 보였다. 따라서 동등한 성능 수준이라면, 컴퓨터 처리 속도 측면에서 보다 단순한 단일 가중치를 사용하는 것이 효율적이다. 넷째, 로치오 분류기는 범주 할당 방법(단일-범주 분류, 복수-범주 분류)에 따라 분류 성능에 차이가 있다. 구체적으로 매크로 평균 F1(mac_F1) 기준의 최고 성능에서는 복수-범주 분류가 단일-범주 분류가 더 나은 결과를 보인 반면, 마이크로 평균(mic_F1) 기준으로는 단일-범주 분류가 더 좋은 성능을 보였다. 특히, 매크로 평균 F1(mac_F1) 기준의 최고 성능은 복수-범주 분류(0.7017: 긍정문헌만 사용($\beta=1$, $\gamma=0$), 학습집합 7년, 단일 가중치 otf)인 반면, 마이크로 평균 F(mic_F) 기준의 최고 성능은 단일-범주 분류(0.6615: 긍정문헌만 사용($\beta=1$, $\gamma=0$), 학습집합 8년, 단일 가중치 ltf, otf)인 것으로 나타났다. 결과적으로, 국내 학술회의 논문의 분류는 특정 논문에 하나 이상의 범주가 부여되는 복수-범주 분류가 실제 환경에 더 부합한다고 볼 수 있으므로, 이러한 환경을 고려하여 주요 성능 요소들의 특성을 종합적으로 검토한 결과로서 최적의 분류 모델을 개발할 필요가 있다.

본 연구는 국내 학술데이터베이스에서 학술회의 논문이 학술지 논문과 함께 대표적인 연구성과물로 서비스되고 있으나, 기본적인 분류정보가 제공되지 않고 있다는 현실적인 문제에서 출발하였다. 따라서 본 연구에서 실제 출판된 학술회의 논문으로 구성된 문헌집합을 대상으로 자동분류 실험을 수행한 결과로 규명한 주요 성능 요소들의 특성은 향후 전체 학술논문의 자동분류를 추진하는 과정에서 의미 있는 시사점을 제공할 수 있을 것이다.

본 연구의 제한점은 특정 분야(정보학)의 학술회의 논문으로 구성한 문헌집합을 대상으로 수행한 실험 결과를 전체 학문분야로 일반화하기 어렵다는 것이다. 따라서 여러 학문분야 또는 학술지 논문을 포함하는 전체 학술논문으로 문헌집합을 확장하는 연구가 필요하다. 또한, 분류기의 성능에 영향을 주는 개별 요소들 간의 연관관계를 심층적으로 검토하여 최적의 성능을 도출하기 위한 연구도 필요할 것이다.

참 고 문 헌

강승식 (2002). 한국어 형태소 분석과 정보검색. 서울: 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, 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

심 경 (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 학술연구분야 분류표 (2015). Retrieved from http://www.nrf.re.kr

한국학술지인용색인 웹사이트 (2016). 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

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., & Koutsoukos, X. D. (2010a). Local causal and markov blanket induction for causal discovery and feature selection for classification. Part I: Algorithms and empirical evaluation. Journal of Machine Learning Research, 11, 171-234.

Aliferis, C. F., Statnikov, A., Tsamardinos, I., Mani, S., & Koutsoukos, X. D. (2010b). Local causal and markov blanket induction for causal discovery and feature selection for classification. Part II: Analysis and extensions. Journal of Machine Learning Research, 11, 235-284.

Aphinyanaphongs, Y., Fu, L., Li, Z., & Peskin, E. R. (2014). A comprehensive empirical comparison of modern supervised classification and feature selection methods for text categorization. Journal of the Association for Information Science and Technology, 65(10), 1964-1987. http://dx.doi.org/10.1002/asi.23110

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

Cohen, W. W., & Singer, Y. (1999). Context-sensitive learning methods for text categorization. ACM Transactions on Information Systems, 17(2), 141-173. http://dx.doi.org/10.1145/306686.306688

Debole, F., & Sebastiani, F. (2003). Supervised term weighting for automated text categorization. Proceedings of the 18th ACM Symposium on Applied Computing (SAC) 2003, 784-788. http://dx.doi.org/10.1145/952532.952688

Devi, P. R., Suganya, B. R., & Abirami, S. (2015). Multi-label learning with class-based features using extended centroid-based classification technique (CCBF). Procedia Computer Science, 54, 405-411. http://dx.doi.org/10.1016/j.procs.2015.06.047

Forman, G. (2003). An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research, 3, 1289-1305.

Forman, G., & Kirshenbaum, E. (2008). Extremely fast text feature extraction for classification and indexing. Proceedings of the 17th ACM Conference on Information and Knowledge Mining (CIKM) 2008, 26-30. http://dx.doi.org/10.1145/1458082.1458243

Foulds, J., & Frank, E. (2010). A review of multi-instance learning assumptions. Knowledge Engineering Review, 25(1), 1-25. http://dx.doi.org/10.1017/s026988890999035x

Genkin, A., Lewis, D. D., & Madigan, D. (2007). Large-scale bayesian logistic regression for text categorization. Technometrics, 49(3), 291-304. http://dx.doi.org/10.1198/004017007000000245

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

Hull, D. A. (1994). Improving text retrieval for the routing problem using latent semantic indexing. SIGIR-94, 282-291. http://dx.doi.org/10.1007/978-1-4471-2099-5_29

Ittner, J. D., Lewis, D. D., & Ahn, D. D. (1995). Text categorization of low quality images. Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval (SDAIR) 1995, 301-315.

Jain, R., & Nitin, P. (2015). Feature selection for effective text classification using semantic information. International Journal of Computer Applications, 113(10), 18-25. http://dx.doi.org/10.5120/19861-1818

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. http://dx.doi.org/10.1016/j.eswa.2011.08.040

Joachims, T. (1997). A probabilistic analysis of the rocchio algorithm with tdf for text categorization. Proceedings of the International Conference on Machine Learning (ICML) 1997, 143-151.

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. http://dx.doi.org/10.4304/jait.1.1.4-20

Kohavi, R., & John, G. H. (1997). Wrappers for feature subset selection. Artificial Intelligence, 97(1-2), 273-324. http://dx.doi.org/10.1016/s0004-3702(97)00043-x

Korde, V., & Mahender, C. N. (2012). Text classification and classifiers: A survey. International Journal of Artificial Intelligence & Applications (IJAIA), 3(2), 85-99.

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

Moschitti, A. (2003). Study on optimal parameter tuning for rocchio text classifier. Lecture Notes in Computer Science, (2633), 420-435. http://dx.doi.org/10.1007/3-540-36618-0_30

Pang, G., & Jiang, S. (2013). A generalized cluster centroid based classifier for text categorization. Information Processing and Management, 49(2), 576-586. http://dx.doi.org/10.1016/j.ipm.2012.10.003

Patra, A., & Singh, D. (2013). A survey report on text classification with different term weighing methods and comparison between classification algorithms. International Journal of Computer Applications, 75(7), 14-18. http://dx.doi.org/10.5120/13122-0472

Read, J., Pfahringer, B., Holmes, G., & Frank, E. (2011). Classifier chains for multi-label classification. Machine Learning, 85(3), 333-359. http://dx.doi.org/10.1007/s10994-011-5256-5

Rogati, M., & Yang, Y. (2002). High-performing feature selection for text classification. Proceedings of the 11th International Conference on Information and knowledge management (CIKM) 2002, 4-9. http://dx.doi.org/10.1145/584792.584911

Schapire, R. E., & Singer, Y. (2000). BoosTexter: A boosting-based system for text categorization. Machine Learning, 39(2-3), 135-168.

Schapire, R. E., Singer, Y., & Singhal, A. (1998). Boosting and rocchio applied to text filtering. Proceedings of the 21st Annual International ACM SIGIR conference on research and development in information retrieval (SIGIR) 1998, 215-223. http://dx.doi.org/10.1145/290941.290996

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

Singhal, A., Mitra, M., & Buckley, C. (1997). Learning routing queries in a query zone. Proceedings of the 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) 1997, 25-32. http://dx.doi.org/10.1145/258525.258530

Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing and Management, 45(4), 427-437. http://dx.doi.org/10.1016/j.ipm.2009.03.002

Tan, S. (2008). An improved centroid classifier for text categorization. Expert Systems with Applications, 35(1-2), 279-285. http://dx.doi.org/10.1016/j.eswa.2007.06.028

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. http://dx.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. http://dx.doi.org/10.1016/j.ijmedinf.2010.10.015

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: An overview. International Journal of Data Warehousing and Mining (IJDWM), 3(3), 1-13. http://dx.doi.org/10.4018/jdwm.2007070101

Uysal, A. K., & Gunal, S. (2014). The impact of preprocessing on text classification. Information Processing and Management, 50(1), 104-112. http://dx.doi.org/10.1016/j.ipm.2013.08.006

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 24th International Florida Artificial Intelligence Research Society Conference, 323-328.

Wu, C. (2009). Behavior-based spam detection using a hybrid method of rule-based techniques and neural networks. Expert Systems with Applications, 36(3), 4321-4330. http://dx.doi.org/10.1016/j.eswa.2008.03.002

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

Yang, Y., & Liu, X. (1999). A re-examination for text categorization methods. Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR) 1999, 42-49. http://dx.doi.org/10.1145/312624.312647

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) 1997, 412-420.

Yu, B., Xu, Z., & Li, C. (2008). Latent semantic analysis for text categorization using neural network. Knowledge-Based Systems, 21(8), 900-904. http://dx.doi.org/10.1016/j.knosys.2008.03.045

Zeng, A., & Huang, Y. (2011). A text classification algorithm based on rocchio and hierarchical

clustering. Lecture Notes in Computer Science, 432-439. http://dx.doi.org/10.1007/978-3-642-24728-6_59

Zhang, W., Yoshida, T., & Tang, X. (2011). A comparative study of TF*IDF, LSI and multi-words for text classification. Expert Systems with Applications, 38(3), 2758-2765. http://dx.doi.org/10.1016/j.eswa.2010.08.066

• 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. 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, & 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 (2008). A study on the documents' 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

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

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