문헌빈도와 장서빈도를 이용한 kNN 분류기의 자질선정에 관한 연구*

A Study on Feature Selection for kNN Classifier using Document Frequency and Collection Frequency

이 용 구(Yong-Gu Lee)**

〈목 차〉

초 록

이 연구에서는 자동 색인을 통해 쉽게 얻을 수 있는 자질의 문헌빈도와 장서빈도를 이용하여 자동분류에서 자질 선정 기법을 kNN 분류기에 적용하였을 때, 어떠한 분류성능을 보이는지 알아보고자 하였다. 실험집단으로 한국일보-20000(HKIB-20000)의 일부를 이용하였다. 실험 결과 첫째, 장서빈도를 이용하여 고빈도 자질을 선정하고 저빈도 자질을 제거한 자질선정 방법이 문헌빈도보다 더 좋은 성능을 가져오는 것으로 나타났다. 둘째, 문헌빈도와 장서빈도 모두 저빈도 자질을 우선으로 선정하는 방법은 좋은 분류성능을 가져오지 못했다. 셋째, 장서빈도와 같은 단순빈도에서 자질 선정 구간을 조정하는 것이 문헌빈도와 장서빈도의 조합보다 더 좋은 성능을 가져오는 것으로 나타났다.

키워드: 자동분류, 자질 선정, kNN 분류기, 문헌빈도, 장서빈도

ABSTRACT

This study investigated the classification performance of a kNN classifier using the feature selection methods based on document frequency(DF) and collection frequency(CF). The results of the experiments, which used HKIB-20000 data, were as follows. First, the feature selection methods that used high-frequency terms and removed low-frequency terms by the CF criterion achieved better classification performance than those using the DF criterion. Second, neither DF nor CF methods performed well when low-frequency terms were selected first in the feature selection process. Last, combining CF and DF criteria did not result in better classification performance than using the single feature selection criterion of DF or CF.

Keywords: Automatic classification, Feature selection, kNN classifier, Document frequency, Collection frequency

* 본 연구는 2010년도 계명대학교 비사연구기금으로 이루어졌으며, 2012년 추계 학술대회 발표를 추가・보완하였음.
** 계명대학교 문헌정보학과 조교수(yonggulee@kmu.ac.kr)
• 접수일: 2013년 3월 5일 • 최초심사일: 2013년 3월 12일 • 최종심사일: 2013년 3월 26일

Ⅰ. 서 론

기계에 의해 색인이 이루어지는 것을 자동색인이라 하듯이, 기계에 의해 이루어지는 분류를 자동분류라 한다. 분류의 대상으로 다양한 정보원을 이용할 수 있다. 그 대상으로 문헌이 될 수도 있으며, 웹에 존재하거나 생산되는 각종 데이터가 될 수도 있다. 자동분류의 경우 크게 그 대상에 대해 범주 정보가 있느냐, 없느냐에 따라 지도학습에 해당하는 범주화(categorization)과 비지도학습에 속하는 클러스터링으로 나뉠 수 있다.

특히 범주화의 경우 클러스터링에 비해 그 성능이 우수하여 다양한 분야에서 실제 시스템에 많이 적용되고 있다. 문헌 또는 문서 범주화부터 의학 정보 관련 범주화, 더 나아가 데이터 마이닝에 핵심적인 기법으로 인식이 되고 있다.

범주화를 적용하기 위해서는 그 대상이 갖는 특성 또는 속성을 추출하여 이를 사용한다. 문헌의 경우 정보검색과 마찬가지로 문헌에 출현한 단어 또는 자동색인으로부터 추출된 키워드 등이 속성을 이루며, 이와 같은 속성 내지 특성을 자질(feature)이라고 부른다. 자질은 분류 대상이 무엇이냐에 따라 달라진다. 예를 들어 분류 대상이 이용자라면 성별, 나이, 주소 등과 같은 이용자의 속성이 자질이 된다.

문헌 범주화에서 자질은 그 문헌에 출현한 키워드로 자질의 좋고 나쁨에 따라 범주화 결과인 자동분류에 많은 영향을 미친다. 어떤 자질은 최종 성능에 영향을 미치지 않거나 심지어 악영향을 미쳐 분류성능을 저하시키기도 한다. 이러한 자질은 잡음(noise)에 해당하여 분류시에 도움이 되지 않아 자질에서 제외시키며 이러한 방법론을 자질선정(feature selection)이라 한다.

일반적으로 문헌을 대상으로 자동색인을 수행하면, 불용어 처리 기법을 적용하더라도 한 문헌으로부터 많은 키워드인 자질이 생성된다. 그리고 이러한 자질 모두가 앞서 얘기하였듯이 문헌분류에 도움이 되질 않아 자질선정을 수행하게 된다. 문헌을 대상으로 한 자질선정에 핵심적인 연구결과의 하나인 Yang과 Pederson의 논문에서는 자질선정을 통해 좀 더 좋은 분류성능을 가져왔는데[1], 이는 잡음에 해당하는 자질을 제거함으로 가능한 것이다.

범주화에서 어떤 자질을 사용할 것인지, 또는 사용하지 않을 것인지를 결정하기 위한 자질선정 기준이 필요하다. 이러한 기준으로 단어의 출현빈도와 같은 통계량이나 정보이론에 기반한 다수의 자질선정 방법들이 있는데, 이들 방법 중에 몇몇은 자질선정을 하지 않을 때보다 우수한 성능을 가져오는 것으로 알려져 있다.[2]

---

1) Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization," *Proceedings of the 14th International Conference on Machine Learning*(1997), pp.412-420.

2) *Ibid.*

자질선정 방법 중에는 다양한 통계정보를 이용하거나 경우에 따라서는 매우 복잡한 계산을 수행하여 그 기준값을 얻는 경우가 있다. 종종 이는 자동분류보다 더 오래 걸리거나 더 많은 전산 자원을 요구하는 경우도 있다.

실험문헌 집단에서 특정 색인어가 출현한 문헌의 수인 문헌빈도(document frequency: DF)나 그 색인어가 출현한 총 빈도인 장서빈도(collection frequency: CF)는 검색 시스템을 구축하기 위해 필요하며, 자질을 추출하기 위해 문헌을 자동 색인하는 과정에 쉽게 구축되는 정보에 해당한다. 즉 자질선정 기법을 적용하고자 할 때 이러한 정보들은 쉽게 생성할 수 있거나 이미 생성되어 있는 경우가 많다. 따라서 이들 정보를 자질선정에 활용한다면 보다 쉬우며, 앞서 말한 전산자원에 대한 부담도 없을 것이다.

현재 자동분류에서는 다양한 분류기가 존재한다. 특히 SVM(support vector machine)이나 VP(voted perceptron)와 같은 분류기는 그 분류성능이 kNN(*k*-nearest neighbors) 분류기에 비해 더 좋은 것으로 나타난다. 하지만, kNN 분류기는 정보검색 시스템에서 검색 결과(상위 k개)를 추출하는 것과 같은 환경[3] 하에서 쉽게 구현할 수 있다는 장점이 있다.[4] 특히, 문헌이 벡터 형태로 표현하는 벡터공간 검색모형에서는 더욱 더 그러하다.

따라서 이 연구에서는 실재 검색 시스템이 이미 갖추어진 환경에서 쉽게 구현할 수 있는 kNN 분류기를 대상으로 하였으며, 이러한 환경에서 이용가능한 자동색인의 결과물인 문헌빈도와 장서빈도를 적용하여 자질선정을 수행하여 분류성능을 알아보고자 하였다. 다만, 분류기 자체 보다는 좋은 성능을 보이는 효율적인 자질선정 방법을 찾기 위한 기초적인 연구로 하나의 분류기와 문헌빈도와 장서빈도인 두 빈도만을 적용하는 것으로 제한하여, 추후 다양한 분류기나 다른 자질선정 방법과의 비교 등의 연구가 필요하다.

Ⅱ. 이론적 배경

1. kNN 분류기

kNN 분류기는 단순히 학습문헌을 저장하는 사례 기반 학습(instance-based learning) 방법을 채택한 대표적인 분류기이다. kNN 분류기는 입력문헌과 유사도가 가장 높은 k 개의 최근접 이웃

3) 유사한 문헌을 필요로 하는 상황에서, 한번으로 문헌을 색인하여 표현하고 해당 문헌에 다수의 범주가 부여되는 문헌 중심(document-centric)의 환경을 뜻한다.

4) P. Jackson and I. Moulinier, *Natural Language Processing for Online Applications - Text Retrieval, Extraction and Categorization*(Amsterdam : Benjamins Publishing Co., 2002), pp.148-149.

문헌을 학습문헌 집합으로부터 찾아낸 다음, 이 이웃문헌들에 배정된 범주들에 근거하여 입력문헌을 분류할 하나 이상의 범주를 선정한다. kNN 분류기는 각 범주를 구체적인 용어벡터로 표현하지 않고 학습문헌에 배정된 범주 표시만을 이용하여 분류한다.[5] 이러한 분류기는 학습 기법을 적용하는 단계가 없이 입력문헌이 들어왔을 때 분류를 하기에 게으른 학습(lazy learning) 기법이라고도 불린다.

입력문헌과 학습문헌 간의 유사도를 산출하기 위해 다양한 유사계수 공식이나 거리계수가 사용될 수 있다. 일반적으로 코사인 유사계수를 많이 사용하지만, 이 연구에서는 가장 기본적인 거리계수인 유클리드 거리계수를 사용하였으며[6] 이를 공식으로 표현하면 〈식 1〉과 같다.

$$d(x,y)=\sqrt{\sum_{i=1}^{n}(w_{xi}-w_{yi})^2} \qquad \langle 식\ 1 \rangle$$

이 공식에서 x와 y는 입력문헌과 학습문헌 중 하나이며, 이들이 갖는 전체 색인어는 n개이며, 색인어의 가중치 w_{xi}, w_{yi}는 각각 x 문헌의 i 번째 색인어의 가중치이며, y문헌의 i 번째 색인어의 가중치이다. 이 공식이 거리계수이므로 값이 작을수록, 즉 거리가 가까울수록 두 문헌을 더 유사하다고 볼 수 있다.

kNN에서 고려해야 할 변수 중에 하나는 k 값을 생각할 수 있다. 대개 이 값은 어떠한 실험문헌 집단을 사용하는지, 어떤 분야의 데이터를 사용하는지 등 다양한 경험적 사례에 기반하여 달라진다. 일반적으로 k 값을 선정할 때 고려사항은 다음과 같다.[7]

- 자질 공간에서 범주들이 얼마나 가까이 존재하는가. 즉 범주가 서로 가까울수록 k 값은 더 작은 값을 가질 것이다.
- 특정 범주에서 학습문헌들이 얼마나 고유(typical)하게 존재하는가. 즉 학습문헌들이 이질적이라면 보다 큰 k 값이 대표적인 사례를 확보하기 위해 적당할 것이다.

최종적으로 입력문헌에 대해 특정 범주를 부여하기 위해서는 유사 문헌으로 선정된 k개의 최근접 이웃 문헌들이 부여된 범주 중에 가장 많은 것을 선정하거나 이들 문헌들과 입력문헌 간의 유사도 가중치의 총합(total sum)이 가장 큰 범주를 선정한다.

kNN 분류기는 학습기법을 적용하지 않으므로 입력문헌이 들어오기까지는 빠르게 진행할 수 있지만, 입력문헌과 모든 학습문헌을 비교해야 하므로 학습문헌의 규모에 따라 처리 속도가 떨어지는 문제점을 갖고 있다. 다만 그 성능이 우수하고, 문헌 중심의 분류기라서 범주의 수가 크다면 고려해

5) 정영미, 정보검색연구(서울 : 구미무역 출판부, 2005), p.154.
6) 상게서, p.121.
7) P. Jackson and I. Moulinier, *op. cit.*, p.149.

볼 만한 분류기이다.

문헌 범주화에서 kNN 분류기를 많이 이용한다.[8][9] 특히, 텍스트를 기반으로 하는 문헌들을 분류할 때 마주하는 현상 중에 하나는 한 범주에 많은 문헌이 부여되는 경향이 많다는 것이다. 이러한 상황에 대해 적은 수가 할당된 범주에 포함된 이웃문헌에 대해 높은 가중치를 주고, 큰 규모의 범주에 포함된 이웃문헌에 대해 적은 가중치를 부여하여 방식으로 kNN 분류기에 적용하여 의미있는 성능향상을 가져오기 한다.[10]

문헌을 분류하기 위해 kNN 분류기를 이용한 국내 선행연구를 살펴보면, 심경은 kNN 분류기를 이용하여 실재 시스템 환경에서 문헌 분류할 때 적정한 문헌범주화 성능을 이루기 위해 범주당 가장 이상적인 학습문헌집합의 규모를 100개로 규정하였다. 하지만, 실재 시스템이기에 분류 성능이 기존의 연구들에 비해 상당히 낮은 값을 가져왔다.[11]

2. 자질 선정

앞서 기술하였듯이 학습집단으로부터 분류기를 구축하는데 필요한 자질을 추출하는데, 분류 성능에 모든 자질이 긍정적인 영향을 미치는 것은 아니다. 어떤 자질은 분류 성능에 도움이 되지만, 어떤 자질은 방해가 되기도 한다. 따라서 불필요한 자질을 제거하거나 적절한 자질의 선정을 통해 자질집합 또는 자질공간의 축소를 가져올 수 있다.[12]

자질 선정 또는 자질 축소의 혜택은 데이터 시각화와 데이터 이해의 촉진, 측정과 저장공간의 감소, 학습과 활용 시간의 감소, 예측 성능의 향상시키기 위한 차원의 저주(Curse of dimensionality)에 도전 등을 들 수 있다.[13] 즉 다시 말하면, 자질 선정은 분류 성능의 잡음 내지 도움이 되지 않는 자질들을 제거함으로써 분류 성능이 향상이 가능할 수 있으며, 다양한 이점이 있는 자질 공간의 축소를 가져오게 할 수 있다.

자질 선정 방법 중에 가장 흔히 쓰이는 자질 하부 집합 선정은 기존의 자질 집합으로부터 좋은 분류 성능을 가져오는 유용한 자질들을 선택하고, 반대로 중복되거나 부가적인 자질들의 제외시키

---

8) Y. Yang, "An evaluation of statistical approaches to text categorization," *Information Retrieval*, Vol.1, No.1-2(1999), pp.69-90.

9) Y. Yang and X. Lin, "A re-examination of text categorization methods," In: *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in the information retrieval*(1999), pp.42-49.

10) S. Tan, "Neighbor-weighted K-nearest Neighbor for Unbalanced Text Corpus," *Expert Systems with Applications*, Vol.28, No.4(2005), pp.667-671.

11) 심경, "문헌범주화에서 학습문헌수 최적화에 관한 연구," 정보관리학회지, 제23권, 제4호(2006. 12), pp.277-294.

12) 이용구, "단어 중의성 해소를 위한 지도학습 방법의 통계적 자질선정에 관한 연구," 한국비블리아학회지, 제22권, 제2호(2011. 6), pp.5-25.

13) I. Guyon and A. Elisseeff, "An Introduction to Variable and Feature Selection," *Journal of Machine Learning Research*, 3(2002), pp.1157-1182.

는 것을 말한다. 새로운 자질의 구축은 원래의 자질로부터 조합과 변형을 통한 새로운 자질 생성하는 것을 말한다. 이 방법에는 어간이 동일한 단어들의 접사나 어미를 제거하는 스테밍(stemming), 용어 클러스터링, 잠재의미색인(LSI) 방법 등을 통해 여러 자질을 하나의 자질로 변화시킨다.[14]

자질의 하부집합을 선정하기 위한 방법으로 문헌빈도와 같은 단어의 출현빈도를 이용하는 방법이 있다. 또한 상호정보량이나 정보획득량(information gain)과 같은 정보이론적 기준, 카이제곱 통계량(χ^2 statistic), 상관계수, 적합성 점수 승산비와 같은 기준을 적용할 수 있다.[15] 이 중에서 정보획득량과 카이제곱이 가장 효과적이었으며, 문헌빈도도 비슷한 성능을 보인 것으로 나타났다.

Yang과 Pedersen의 연구를 보면, 로이터(Reutors) 말뭉치에서 kNN 분류기를 정보획득량을 이용하여 자질의 98%를 축소하였음에도 약간의 분류 성능이 향상되는 것을 보여주었다. 또한 문헌빈도는 정보획득량이나 카이제곱 기준보다 계산측면에서 가장 낮은 비용을 가지면서 유사한 성능을 보였다.[16]

국내에서는 문헌 자동분류에서 분류자질 선정과 가중치 할당을 위해서 일관된 전략을 채택하여 kNN 분류기의 성능을 향상시킬 수 있는 연구가 수행되었다.[17] 이 연구에서는 색인파일 저장 공간과 실행 시간에 따른 분류성능을 기준으로 분류자질 선정 결과를 평가하였는데, 상호정보량과 같은 저빈도 자질 선호 기준이나 심지어는 역문헌빈도를 이용해서 분류자질을 선정하는 것이 kNN 분류기의 분류 효과와 효율 면에서 바람직한 것으로 나타났다.

문헌 자동분류에서 자질 선정이 중요함을 많은 연구에서 제시하였는데, 특히 용어의 단어빈도와 문헌빈도를 기반으로 자질선정을 수행한 연구에서는 용어간의 상대적 중요도로 이들 척도를 사용하였다. 로이터 데이터를 이용하여 실험한 결과, GINI와 같은 척도를 응용한 단어빈도가 보다 작은 자질 집합에서 유용하였다.[18][19]

경우에 따라서는 자질선정은 자질을 축소하는 것에서 자질을 확장하는 방법으로 가능하다. 학술지 수록 논문의 필수적인 구성요소인 저자 제공 키워드와 논문제목에 나타난 용어를 대상으로 WordNet 동의어 관계를 활용하여 자질확장을 수행하였으며, 그 결과 많은 분류성능을 가져왔다. 특히, 용어의 중의성 해소 적용과 비적용 모두 성능이 향상된 것으로 파악되었다.[20]

---

14) F. Sebastiani, "Machine Learning in Automated Text Categorization," *ACM Computing Surveys*, Vol.34, No.1 (2002), pp.1-47.

15) 정영미, 전게서, p.148.

16) Yang and Pedersen, *op. cit.*, pp.412-420.

17) 이재윤, "자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 대한 연구," 한국문헌정보학회지, 제39권, 제2호(2005. 6), pp.123-146.

18) W. Shang et al., "A novel feature selection algorithm for text categorization," *Expert Systems with Applications*, Vol.33, No.1(July, 2007), pp.1-5.

19) N. Azam and J. Yao, "Comparison of term frequency and document frequency based feature selection metrics in text categorization," *Expert Systems with Applications*, Vol.39, No.5(2012), pp.4760-4768.

20) 정은경, "문서범주화 성능 향상을 위한 의미기반 자질확장에 관한 연구," 정보관리학회지, 제26권, 제3호(2009. 9),

다만, 텍스트 범주화에서 자질 선정 방법으로 단어빈도를 사용하거나 문헌빈도를 사용하는 사례는 많지만[21], 실제 장서빈도를 적용한 사례를 거의 없다. 이 연구에서는 장서빈도를 중심으로 자질 선정 방법의 가능성을 살펴보고, 이를 문헌빈도와 비교하고자 하였다.

Ⅲ. 실험설계 및 데이터

문헌 범주화 실험을 하기 위해서는 범주가 부여된 실험문헌집단이 필요하며, 이 문헌집단을 대개 학습집단(train set)과 검증집단(test set)으로 나눈다. 이들 중의 학습집단을 이용하여 분류기를 구축하고, 검증집단을 사용하여 분류기의 성능을 평가한다.

이 연구에서 문헌 범주화 실험을 하기 위해서 한국일보-20000 실험문헌집단(HKIB-20000)의 일부를 이용하였다. 이 실험집단은 1998~1999년의 2년간 신문기사를 바탕으로 각 문헌별로 3단계 분류체계의 말단 범주를 부여하여 구축하였다.[22] 이 실험집단의 전체 규모는 다섯 개의 묶음 파일에 총 20,000개의 문헌이 분류되어 있으며 이들 문헌을 5겹 교차 검증(5-fold cross validation)을 위해 대략 4000개 씩 한 파일에 담겨있다. 이 실험에 쓰인 파일은 첫 번째 파일로 총 3988개의 문헌을 포함하고 있으며, 이들의 범주 정보는 2003년 버전과 2007년 버전이 있으며, 이 실험에서는 2003 버전을 사용하였다. 2003 버전도 계층적인 대중소 분류 체계를 가지고 있으며 문헌당 하나의 범주가 부여되어 있다. 실험집단의 분석 결과를 보면, 다음 〈표 1〉과 같다.

〈표 1〉 실험문헌집단의 범주와 문헌수

번호	분류명	전체 문헌수	학습문헌 수	검증문헌 수
1	건강과 의학	112	93	19
2	경제	1,206	1,100	106
3	과학	139	127	12
4	교육	123	109	14
5	문화와 종교	613	547	66
6	사회	899	816	83
7	산업	822	730	92
8	여가생활	74	68	6
합계		3,988	3,590	398

pp.261-278.

21) Azam and Yao, *op. cit.*, p.4760.

22) http://www.kristalinfo.com/TestCollections/#hkib

이 실험집단의 경우 자동분류나 다른 분야에 사용하기 위해 몇 가지 특징적인 측면을 가지고 있다. 신문기사의 단어가 강제로 줄바꿈된 사례들이 보이며, 중간에 키워드와 같은 내용(KW로 구분되어진 내용)이 있는데, 이는 한국일보에서 제공된 것을 그대로 가져온 것으로 본문의 일부로 취급하거나 원문 그대로 반영한 결과이다. 실험집단 구축기관은 실험집합을 구축할 때 원문은 그대로 두어야만 이러한 오류까지도 수용하는 기술을 개발할 수 있는 여지를 남겨둘 수 있다는 취지하에 그대로 수용했다고 하였다.[23] 또한 이 실험집단을 이용하여 kNN 분류기의 분류성능을 측정한 선행연구 결과, 기존의 외국 유명 실험집단보다 낮은 분류성능을 보였다.[24]

이 연구에서는 실험집단을 색인하기 위하여 HAM을 이용하여 명사, 고유명사, 영어 등의 품사를 추출하였다.[25] 다만 앞서 이야기하였듯이 강제 줄바꿈된 명사나 기타 색인어 대상에 대해 오류를 교정하기 위한 추가적인 텍스트 처리를 하지 않았다. 이는 형태소 분석기의 자체 오류와는 별개로 텍스트 자체에서 기인하는 형태소 분석 오류를 초래하였다. 또한 HAM의 키워드 추출 모듈을 사용하여 잘못 분석된 키워드를 모두 그냥 색인어로 선정하여 추출하였다. 따라서 실제 이러한 오류는 자동분류 성능에 영향을 미칠 수 있을 것으로 생각된다.

이 연구에서는 해당 실험집단에 대해 학습문헌 대 검증문헌을 90%:10% 비율로 임의(random)로 나누었다. 분류기를 구축하기 위해 90%의 학습문헌을 이용하였으며, 10%의 검증문헌을 사용하여 분류기의 성능을 평가하였다.

이 연구에 쓰인 실험집단의 기본적인 정보에 대해 설명하면 다음과 같다. 우선 전체 3,988개의 문헌에 대해 추출된 색인어의 수는 620,153개로 한 문헌당 색인어의 수는 155.5개 이며, 전체 고유색인어 수(total type)는 38,132개 이다. 이 중 문헌빈도가 1인 색인어, 즉 하나의 문헌에만 출현한 색인어 수가 18,537개로 이들은 약 48.6% 정도를 차지한다. 이들 단어는 하나의 문헌에만 나타났기에 분류 자질로써의 의미가 없어 학습단계나 검증단계에서 제거하였다.

다만 학습집단에 출현한 총 고유 색인어 수는 36,262개 이고 문헌빈도가 2 이상인 색인어 수는 18,498개 이다. 검증집단에 출현한 고유 색인어 수는 11,437개 이다.

문헌 자동분류를 위한 분류기로는 Orange 마이닝 툴(http://orange.biolab.si/) 패키지를 사용하였다. 이 패키지의 경우 파이선(python)과 연동이 가능하며 다양한 기계학습 알고리즘과 분류기 알고리즘을 포함하고 있다.

보통 분류기를 이용하여 범주화 실험을 할 때에 다양한 파라미터를 설정할 필요가 있다. 이는 분류기를 구축하거나 성능을 최적화할 때 필요하다. 이 연구에서는 특정 분류기의 최적화 성능을

23) 2012년 5월 24일부터 27일까지 해당 실험집단을 관리하는 KISTI의 담당자와 이메일 인터뷰를 통해 얻은 내용임.
24) J. Kim et al., "HKIB-2000 & HKIB-40075: Hangul Benchmark Collections for Text Categorization Research," *Journal of Computing Science and Engineering*, Vol.3, No.3(Sep. 2009), pp.165-180.
25) http://nlp.kookmin.ac.kr/HAM/kor/

보기 위한 것이 아니기에 이 패키지의 기본 설정을 그대로 사용하였다. 따라서 다른 실험과의 직접적인 비교는 문제가 될 수 있다.

kNN 분류기를 이용하여 자동분류 하기 위해 자질의 가중치는 ltc를 적용하였다. 즉 자질에 대해 로그를 취한 단어빈도와 로그를 취한 역문헌빈도 요소를 적용하였으며, 해당 문헌에 출현한 모든 자질의 가중치에 대해 문헌길이로 정규화 하였다.

이 연구에서 분류기의 성능평가를 위해 정확도(accuracy)를 적용하였다. 정확도에 대한 정의는 전체 검증 대상이 되는 문헌 수에 대해 올바르게 범주를 부여한 문헌 수로 나눈 비율을 뜻한다. 즉 전체 검증집단인 398개 중에 분류기가 올바르게 범주를 부여한 문헌 수가 200개가 된다면, 정확도의 값은 0.5025로 50.25%가 된다.

Ⅳ. 자질선정 실험결과

1. 문헌빈도를 이용한 자질선정

성능이 좋은 자질선정 기법으로 정보획득량과 카이제곱 통계량이 있다. 문헌빈도의 경우 이들보다 약간 낮은 성능을 보이나 적용하기 쉽다는 큰 장점을 가진다.[26)]

정보검색에서는 높은 문헌빈도를 갖는 키워드는 많은 문헌들에 출현하여 문헌 식별력이 약하다고 가정하고, 이를 전체 문헌 수에 대해 나눈 역문헌빈도(inverse document frequency)를 많이 이용한다. 반대로 역문헌빈도는 낮은 문헌빈도를 갖는 색인어에 높은 중요도를 주는 방식이다.

하지만 자동분류에서 자질선정은 이와는 반대로 높은 문헌빈도를 갖는 자질을 선택하고 낮은 문헌빈도의 자질은 제거한다. 따라서 이 실험에서는 문헌빈도와 장서빈도를 기준으로 저빈도 자질을 먼저 제거하는 방법과 저빈도 자질을 먼저 사용하는 방법을 적용하였다. 즉 두 가지 기준에 두 가지 방법으로 총 네 가지 방식을 수행하여 kNN 분류기의 성능을 측정하였다.

저빈도 자질을 먼저 제거하는 방법은 많은 선행연구들에서 이루어진 일반적인 방법이다. 이는 낮은 빈도를 갖는 자질을 제거하여 최종적으로 높은 빈도를 갖는 자질만 남게 된다. 이는 뒤집어 놓으면, 높은 빈도를 갖는 자질을 먼저 사용하고 차츰 낮은 빈도를 갖는 자질을 포함하여 분류 성능을 측정하는 것과 같다.

이 연구의 주요 목적 중에 하나로 문헌빈도를 이용하여 자질선정을 할 때 어떤 방법을 효율적인지 알아보기 위함이다. 먼저, 18,498개의 자질을 갖는 첫 번째 HKIB-20000 실험문헌집단을 대상으

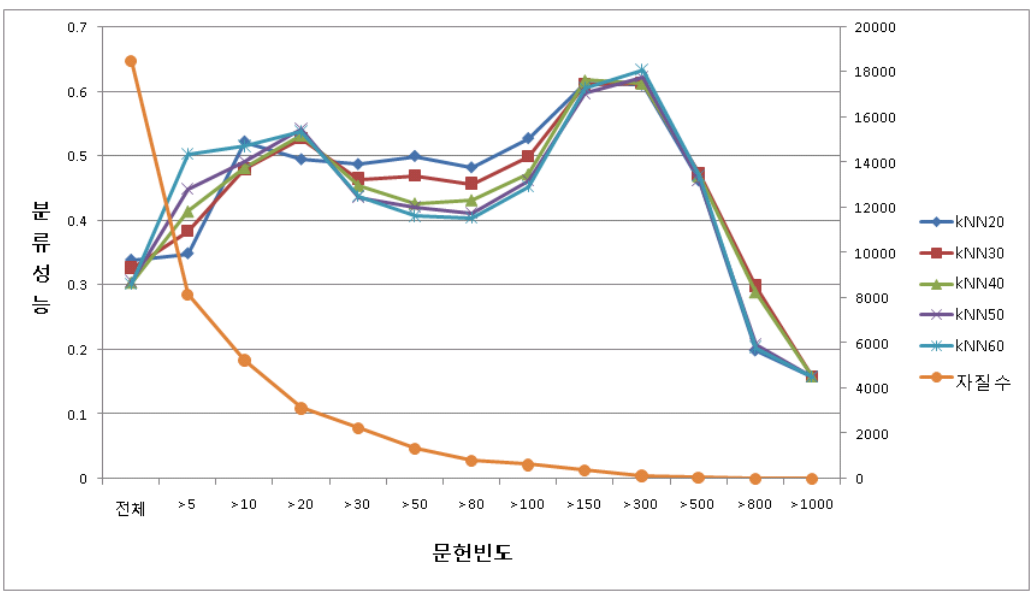
26) Yang and Pedersen, *op. cit.*, pp.412-420.

로 낮은 문헌빈도를 우선 제거하는 자질선정 기법을 적용하여 〈표 2〉와 〈그림 1〉과 같은 결과를 얻었다.

다만 kNN 분류기의 경우 앞서 설명한 k 값에 따라 분류성능이 달라지므로, 다른 연구를 참조하여 분류기의 k 값에 대해 20부터 60까지 변화시켜 분류성능을 알아보고자 하였다. 따라서 k 값이 20인 경우는 kNN20으로, k 값이 30인 경우는 kNN30 등등으로 표기하였다.

〈표 2〉 문헌빈도에서 저빈도 우선 제거에 따른 분류 성능

DF 크기	전체	>5	>10	>20	>30	>50	>80	>100	>150	>300	>500	>800	>1000
자질 수	18,498	8,152	5,252	3,121	2,222	1,337	804	615	363	133	40	8	2
kNN20(k=20)	0.3392	0.3492	0.5226	0.4950	0.4874	0.5000	0.4824	0.5276	0.6131	0.6106	0.4623	0.1985	0.1583
kNN30(k=30)	0.3266	0.3844	0.4799	0.5276	0.4648	0.4698	0.4573	0.5000	0.6106	0.6131	0.4724	0.2990	0.1583
kNN40(k=40)	0.3040	0.4146	0.4824	0.5327	0.4548	0.4271	0.4322	0.4724	0.6181	0.6131	0.4698	0.2889	0.1583
kNN50(k=50)	0.3065	0.4497	0.4925	0.5427	0.4372	0.4221	0.4121	0.4623	0.5980	0.6231	0.4623	0.2085	0.1583
kNN60(k=60)	0.3015	0.5025	0.5151	0.5377	0.4372	0.4070	0.4045	0.4523	0.6055	0.6332	0.4698	0.2035	0.1583
평균(AVG)	0.3156	0.4201	0.4985	0.5271	0.4563	0.4452	0.4377	0.4829	0.6090	0.6186	0.4673	0.2397	0.1583



〈그림 1〉 문헌빈도에서 저빈도 우선 제거에 따른 성능비교

〈표 2〉에서 k값을 20, 30, 40, 50, 60으로 주었을 때 각각의 분류 성능을 보여주며, 두 번째 칼럼(DF 〉 1인 전체 자질의 평균)의 경우 문헌빈도가 1인 경우가 자질에서 삭제되었으며, 그 결과 자질의 수가 18,498개로 되었다. 즉 이 실험의 대상인 전체 자질을 뜻한다. 계속하여 세 번째 칼럼의 경우 DF 〉 5인 경우로 문헌빈도 5이하인 색인어들이 제거되었으며 그 결과 자질의 수가 18,498개에서 10,346개가 제거되어 8,152개로 줄었다. 즉 세 번째 칼럼은 DF 〉 5인 8,152개의 자질만을 대상으로 분류성능을 측정한 결과이다.

요약된 정보로 〈그림 1〉에서 보면, 자질의 수가 문헌빈도 값에 따라 줄어들면서 문헌빈도가 10 또는 20에서 분류성능이 좋아졌다가 다시 낮아지고, 150 또는 300 이상의 문헌빈도에 해당하는 자질을 사용하였을 때 가장 좋은 성능을 보였다. 특히 k값이 60(kNN60)이며 DF 〉 300일 때 가장 좋은 성능인 0.6332(63.32%)를 보였다. 이는 Yang과 Pederson의 연구[27]에서와 같이 전체 자질 집합의 2%, 즉 이 실험에서는 DF 〉 150인 353개와 DF 〉 300인 133개에서 가장 좋은 성능을 보이고 있음을 알 수 있다.

또한 문헌빈도가 500 이상인 경우 그 성능이 매우 급격하게 낮아지는 것을 알 수 있다. 이와 같은 가장 큰 이유는 자질의 수가 40개 밖에 안 되는 것이 그 원인이며, 또 다른 측면으로 문헌빈도의 특성상 전체 학습집단의 문헌 수인 3,590개에서 500개 이상의 문헌에서 출현하는 단어라는 의미로 너무 문헌 식별력이 낮고, 분류 시에 다양한 범주에서 출현하기 때문에 당연한 결과라고 생각된다. 즉 이들 자질들은 범주와 무관한 다양한 문헌에서 출현하므로 분류성능에 좋은 영향을 미치지 않는 것으로 보인다.

전반적으로 살펴보면, 전체 자질을 사용하는 경우(DF 〉 1인 전체 자질의 평균)에서 가장 좋은 분류성능을 보이는 경우(DF 〉 300의 평균)를 비교해 보면, 0.3156에서 0.6186으로 0.3030(95.9%)의 향상을 보였다. 또한 고빈도(150 또는 300)에서는 k 값에 상관없이 모두 높은 분류성능을 보였으며, 저빈도 내지 중빈도(10에서 100까지)에서는 k=20인 경우가 비교적 높은 성능을 보였다.

기존의 선행연구들에서 문헌빈도가 높은 구간이 좋은 분류성능을 내는 것으로 되어 있어 150 구간과 300 구간에 좋은 성능을 가져온 것은 당연한 것으로 생각된다. 하지만, 문헌빈도가 10~20인 구간에서 좋은 분류성능을 보인 것은 이 실험에서는 가장 주목할 부분이다(〈표 3〉과 〈그림 2〉 참조). 즉 kNN 분류기는 저빈도 자질에서도 비교적 좋은 성능을 보이는 것으로 생각된다.

고빈도를 기준으로 문헌빈도의 일정 자질 선정 구간의 분류성능 기여 정도를 알아보기 위해 〈표 2〉에서 앞의 문헌빈도 구간과 연이은 뒷부분의 문헌빈도 구간의 성능차이를 구하였다. 이는 일반적으로 문헌빈도를 이용하여 자질 선정을 할 때, 고빈도부터 자질을 추가하는 것이 비교적 간단하고 쉬운 방법이라 생각하였기 때문이다.[28] 예로 20≦ DF 〈29 자질 선정 구간의 성능(0.5271)은 DF 〉

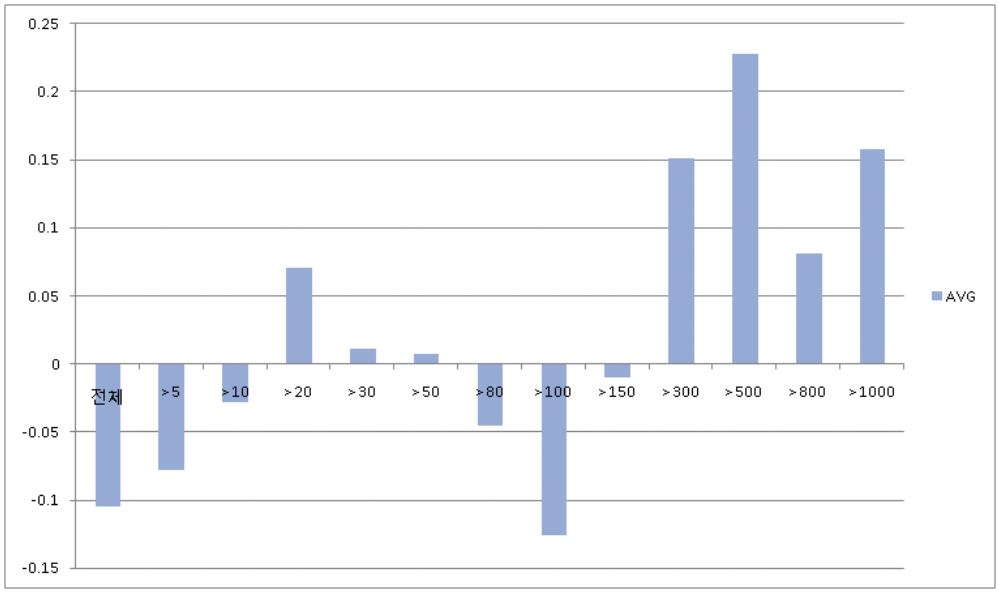
27) *Ibid.*

28) 좀 더 문헌빈도 구간별 성능의 차이를 알 수 있는 방법으로 자질 구간을 똑같이 등분하거나, 해당 구간(예로

30인 자질을 통한 성능(0.4452)에서 해당 구간의 자질이 포함되어 만들어진 성능이므로, 이들의 차로 해당 구간의 기여정도를 파악하고자 하였다. 그 결과는 〈표 3〉과 같았다. 이 표에서 평균(AVG)의 성능차이를 그림으로 표현하면, 〈그림 2〉와 같았다. 표와 그림을 보면, 가장 많은 기여를 하는 문헌빈도 구간은 DF 〉 500과 DF 〉 300으로 나타났다.

〈표 3〉 문헌빈도를 이용한 자질 구간의 분류성능 차이

DF	전체	>5	>10	>20	>30	>50	>80	>100	>150	>300	>500	>800	>1000
kNN20	−0.0101	−0.1734	0.0276	0.0075	−0.0126	0.0176	−0.0452	−0.0854	0.0025	0.1482	0.2638	0.0402	-
kNN30	−0.0578	−0.0955	−0.0477	0.0628	−0.0050	0.0126	−0.0427	−0.1106	−0.0025	0.1407	0.1734	0.1407	-
kNN40	−0.1106	−0.0678	−0.0503	0.0779	0.0276	−0.0050	−0.0402	−0.1457	0.0050	0.1432	0.1809	0.1307	-
kNN50	−0.1432	−0.0427	−0.0503	0.1055	0.0151	0.0101	−0.0503	−0.1357	−0.0251	0.1608	0.2538	0.0503	-
kNN60	−0.2010	−0.0126	−0.0226	0.1005	0.0302	0.0025	−0.0477	−0.1533	−0.0276	0.1633	0.2663	0.0452	-
AVG	−0.1045	−0.0784	−0.0286	0.0709	0.0111	0.0075	−0.0452	−0.1261	−0.0095	0.1513	0.2276	0.0814	-



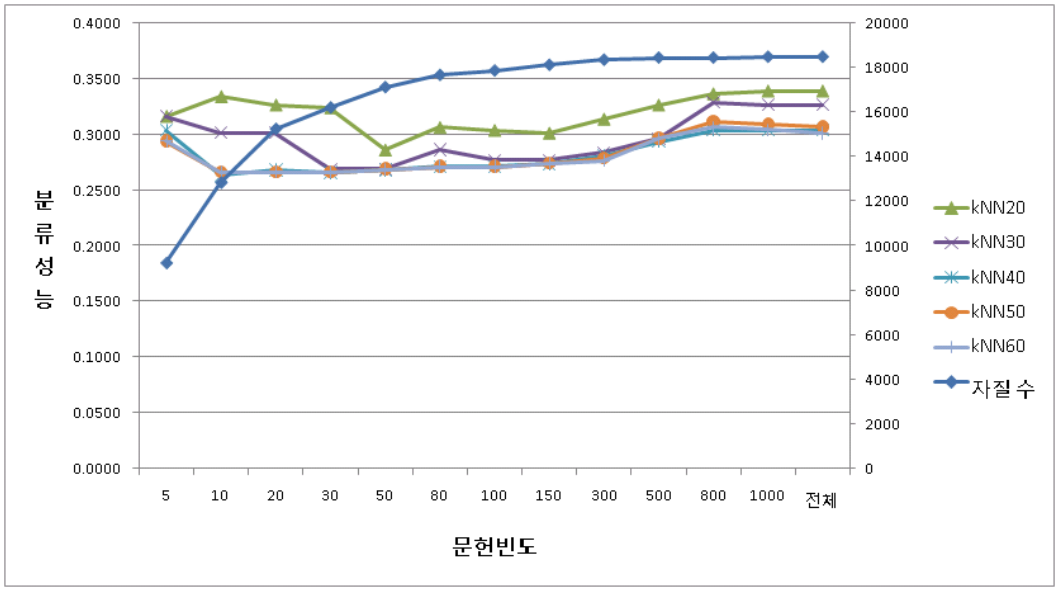
〈그림 2〉 문헌빈도 구간별 평균의 성능차이

DF=3)의 자질만 사용하여 실험할 수 있으나 구간마다 자질 규모의 편차가 심하여 어렵거나 적절하지 않다고 판단하였다. 예로 문헌빈도의 최소값으로부터 최대값까지 10 등분한 경우, 7번째 구간까지 DF가 10미만에 해당하며, 마지막 구간은 DF가 37부터 시작하여 상세한 분석이 어려웠다.

다음으로 자질선정 기준으로 문헌빈도의 저빈도를 우선적으로 사용하는 기법을 적용하여 분류성능을 알아보고자 하였다. 자질을 선정할 때 낮은 문헌빈도를 먼저 자질로 사용하는 것을 의미한다. 이 기법을 적용하여 〈표 4〉와 〈그림 3〉과 같은 결과를 얻었다.

〈표 4〉 문헌빈도에서 저빈도 우선 사용에 따른 분류 성능

DF	<5	<10	<20	<30	<50	<80	<100	<150	<300	<500	<800	<1000	전체
자질 수	9,224	12,839	15,253	16,216	17,131	17,680	17,877	18,134	18,363	18,458	18,458	18,496	18,498
kNN20	0.3166	0.3342	0.3266	0.3241	0.2864	0.3065	0.3040	0.3015	0.3141	0.3266	0.3367	0.3392	0.3392
kNN30	0.3166	0.3015	0.3015	0.2688	0.2688	0.2864	0.2764	0.2764	0.2839	0.2965	0.3291	0.3266	0.3266
kNN40	0.3040	0.2638	0.2688	0.2663	0.2688	0.2714	0.2714	0.2739	0.2814	0.2940	0.3040	0.3040	0.3040
kNN50	0.2940	0.2663	0.2663	0.2663	0.2688	0.2714	0.2714	0.2739	0.2789	0.2965	0.3116	0.3090	0.3065
kNN60	0.2940	0.2663	0.2663	0.2663	0.2688	0.2714	0.2714	0.2739	0.2764	0.2965	0.3065	0.3040	0.3015
AVG	0.3050	0.2864	0.2859	0.2784	0.2724	0.2814	0.2789	0.2799	0.2869	0.3020	0.3176	0.3166	0.3156



〈그림 3〉 문헌빈도에서 저빈도 우선 사용에 따른 성능비교

〈표 4〉와 〈그림 3〉를 보면, 저빈도 중심으로 자질을 선정하여 분류성능을 측정하였을 때, 지속적인 자질의 추가에도 대부분의 k 값에 따른 kNN 분류기의 성능을 크게 좋아지지 않는 것으로 나타

났다. 이는 몇 가지 이유로 해서 분류성능이 좋지 않은 것으로 볼 수 있다. 우선 처음 시작하는 자질 선정 기준이 문헌빈도가 5 미만으로 자질 개수가 9,224개로, 앞서 분석되었던 최고의 분류성능을 보였던 2% 내외의 자질보다 훨씬 많았다. 또한 이렇게 많은 수의 자질 속에는 분류 성능과는 무관한 자질들이 계속해서 남아 있기 때문인 것으로 보인다. 예로 문헌빈도가 5 미만과 같은 작은 자질은 극소수의 문헌에만 출현하기에 분류에 주요한 영향을 미치지 못하는 것으로 보인다.

이상에서 kNN 분류기의 성능을 높이기 위한 문헌빈도를 이용한 두 가지 자질선정 방법에서 적절하게 저빈도를 먼저 제거하고 고빈도의 문헌빈도를 갖는 자질들을 선택하였을 때 좋은 분류성능을 보이는 것을 알 수 있다. 또한 좋은 분류성능을 보이는 저빈도를 선택하는 방법을 같이 병행하는 방안도 필요하다고 생각된다.

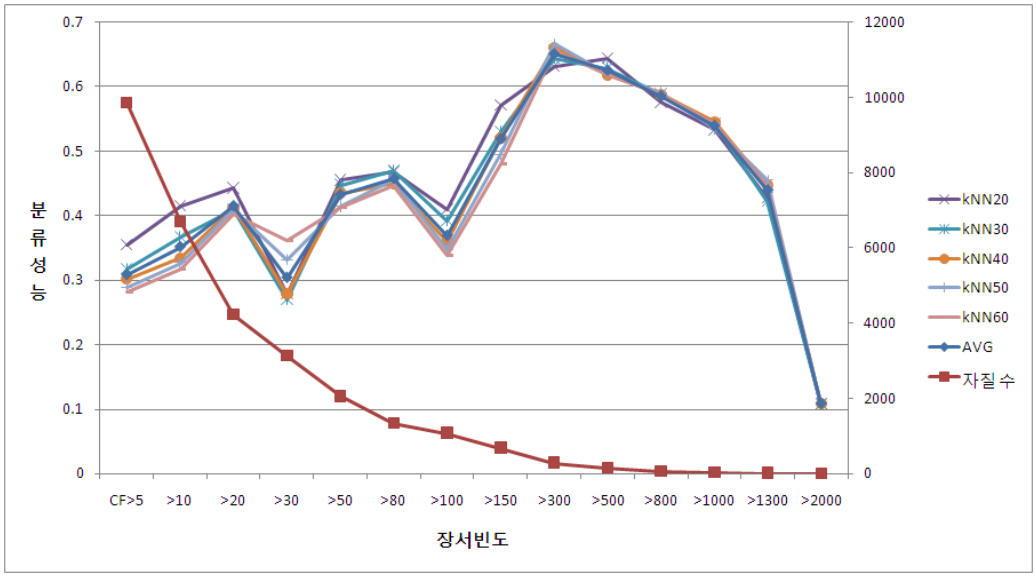
2. 장서빈도를 이용한 자질선정

장서빈도는 지금까지 자질선정 방법으로 많이 사용되지 않았다. 이 빈도가 크다는 의미는 일반적으로 한 문헌 내에서 많은 단어 빈도를 갖는 색인어이거나, 많은 문헌에 출현한 경우의 색인어 이다.

장서빈도를 이용한 자질선정 방법도 문헌빈도와 마찬가지 두 가지 방법을 적용하였다. 먼저 앞의 분석처럼 낮은 장서빈도를 제거하여 고빈도를 남겨놓게 자질 선정을 하였다. 즉 고빈도의 자질을 중심으로 분류성능을 파악하였다. 그 결과 분류성능은 〈표 5〉, 〈그림 4〉와 같았다.

〈표 5〉 장서빈도에서 저빈도 우선 제거에 따른 성능비교

CF	>5	>10	>20	>30	>50	>80	>100	>150	>300	>500	>800	>1000	>1300	>2000
자질 수	9,852	6,688	4,228	3,138	2,056	1,343	1,077	683	292	145	66	38	15	4
kNN20	0.3543	0.4146	0.4422	0.2789	0.4548	0.4673	0.4095	0.5704	0.6307	0.6432	0.5754	0.5327	0.4271	0.1080
kNN30	0.3166	0.3668	0.4095	0.2688	0.4472	0.4698	0.3920	0.5302	0.6432	0.6281	0.5879	0.5427	0.4221	0.1080
kNN40	0.3015	0.3342	0.4121	0.2789	0.4347	0.4497	0.3593	0.5201	0.6608	0.6181	0.5879	0.5452	0.4472	0.1080
kNN50	0.2889	0.3266	0.4095	0.3317	0.4146	0.4548	0.3467	0.4950	0.6658	0.6206	0.5905	0.5352	0.4548	0.1080
kNN60	0.2814	0.3166	0.4020	0.3618	0.4121	0.4472	0.3392	0.4799	0.6558	0.6181	0.5879	0.5427	0.4472	0.1080
AVG	0.3085	0.3518	0.4151	0.3040	0.4327	0.4578	0.3693	0.5191	0.6513	0.6256	0.5859	0.5397	0.4397	0.1080



〈그림 4〉 장서빈도에서 저빈도 우선 제거에 따른 성능비교

〈표 5〉와 〈그림 4〉의 결과를 살펴보면, 저빈도 장서빈도를 우선 제거하여 고빈도 중심의 장서빈도를 자질로 사용한 경우, 계속하여 저빈도 장서빈도를 제거하여 지속적으로 분류성능을 향상되다가 장서빈도 300이상(CF〉300) 이후로 감소되는 것으로 나타났다. 특히 주목해야 할 부분은 장서빈도를 이용한 자질 축소의 최고의 분류성능이 문헌빈도를 이용한 경우보다 더 높은 것을 알 수 있다. 실제 문헌빈도를 이용한 방법에서는 0.6332(DF〉300, k=60)였지만, 장서빈도에서는 0.6658(CF〉300, k=50)로 0.0326(3.26%) 더 높은 것을 알 수 있다.

다만, 〈그림 4〉을 보면 장서빈도가 30 이상일 때나 100 이상일 매우 급격한 성능 저하가 나타났다. 이러한 변화가 장서빈도의 특성인지 이 실험에 쓰인 집단에서 기인하는 것인지는 좀 더 추가적인 실험이 필요한 것으로 보인다.

장서빈도를 자질 선정 기준으로 사용할 때에도 문헌빈도와 비슷하게 자질의 크기가 줄어드는 것을 알 수 있다. 이는 장서빈도가 자질 선정이나 축소에서 문헌빈도와 같은 역할을 할 수 있음을 의미한다. 다만, 같은 기준 구간이면 대체로 장서빈도가 더 큰 자질을 갖고 있다. 가장 좋은 분류성능을 보이는 구간을 보면, 두 빈도 모두 300개 보다 큰 지점이었다. 이때 자질의 수는 문헌빈도가 133개인 반면, 장서빈도는 292개로 나타났다. 292개의 자질은 전체 자질 18,498개에 대해 1.58%에 해당하는 비율이다.

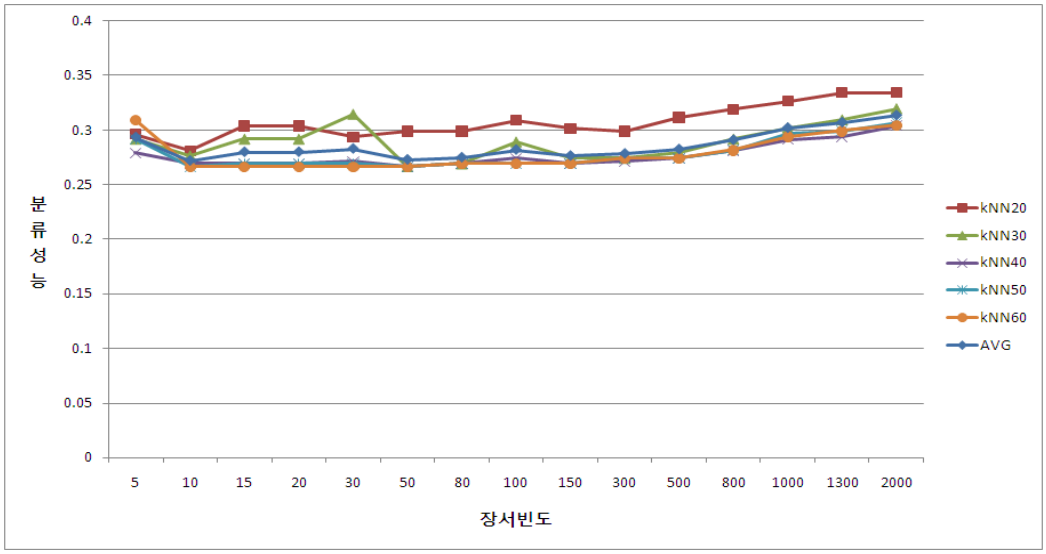
장서빈도를 이용한 자질 선정기준에서 k 값을 달리하여 구축된 분류기들의 성능이 그리 큰 편차를 보이지도 않았다. 특히 장서빈도가 커지는 구간에서는 5개의 분류기가 거의 같은 분류성능을 보였다.

장서빈도 실험에 대해 전반적으로 살펴보면, 전체 자질을 사용하는 경우(CF 〉 5의 평균)에서 가장 좋은 분류성능을 보이는 경우(CF 〉 300의 평균)를 비교해 보면, 0.3085에서 0.6513으로 0.3428(111.1%)의 많은 향상을 보였다. 문헌빈도와 마찬가지로 고빈도(CF 〉 300)에서는 k 값에 상관없이 모두 높은 분류성능을 보였다.

장서빈도를 저빈도 우선으로 하여 자질을 사용하는 방법을 적용한 결과는 〈표 6〉과 〈그림 5〉와 같았다. 다만 문헌빈도와 마찬가지로 각 경우에 따른 눈에 띄는 성능의 향상은 보이지 않았다. 처음으로 선정된 자질의 수가 7,549개(CF 〈 5)로 여전히 많은 자질을 포함하고 있다.

〈표 6〉 장서빈도에서 저빈도 우선 사용에 따른 분류 성능

CF	<5	<10	<15	<20	<30	<50	<80	<100	<150	<300	<500	<800	<1000	<1300	<2000
자질 수	7,549	11,349	13,100	14,127	15,278	16,406	17,140	17,410	17,806	18,205	18,353	18,432	18,460	18,483	18,494
kNN20	0.2965	0.2814	0.3040	0.3040	0.2940	0.2990	0.2990	0.3090	0.3015	0.2990	0.3116	0.3191	0.3266	0.3342	0.3342
kNN30	0.2915	0.2764	0.2915	0.2915	0.3141	0.2663	0.2688	0.2889	0.2739	0.2739	0.2789	0.2915	0.3015	0.3090	0.3191
kNN40	0.2789	0.2688	0.2688	0.2688	0.2714	0.2663	0.2688	0.2739	0.2688	0.2714	0.2739	0.2814	0.2915	0.2940	0.3040
kNN50	0.2915	0.2663	0.2688	0.2688	0.2688	0.2663	0.2688	0.2688	0.2688	0.2739	0.2739	0.2814	0.2965	0.2990	0.3065
kNN60	0.3090	0.2663	0.2663	0.2663	0.2663	0.2663	0.2688	0.2688	0.2688	0.2739	0.2739	0.2814	0.2940	0.2990	0.3040
AVG	0.2935	0.2719	0.2799	0.2799	0.2829	0.2729	0.2749	0.2819	0.2764	0.2784	0.2824	0.2910	0.3020	0.3070	0.3136



〈그림 5〉 장서빈도에서 저빈도 우선 사용에 따른 성능비교

두 개의 빈도를 이용하여 자질 선정 기준을 적용한 결과, HKIB-2000의 첫 번째 실험집단은 자질 선정을 통해 최대 111.1%의 분류성능의 증가를 가져왔다. 이는 장서빈도를 고빈도 위주로 선정한 조건에서 그러하였다. 물론 실험집단의 특성과 학습집단의 무작위 추출에 따라 달라질 수 있다.

두 빈도 모두 저빈도를 우선으로 사용하여 자질 선정하는 방법은 좋은 성능을 가져오지 못했다. 주로 저빈도 자질들의 수가 많아 자질의 크기를 줄이려면, 다른 효과적인 방법을 연구하는 것을 좋을 듯하다.

3. 두 빈도의 조합

앞서 사용한 문헌빈도와 장서빈도를 조합하여 자질 선정을 하였을 때 더 좋은 성능이 나오는지 알아보기 위해 두 빈도에서 좋은 구간에 해당하는 값의 자질들을 대상으로 분류 실험을 하였다.

조합하기 위한 방법을 크게 두 가지로 나누었다. 문헌빈도와 장서빈도를 각각 따로 독자적으로 구간을 조합하여 분류성능을 측정하였으며, 다음에는 문헌빈도와 장서빈도의 구간을 서로 조합하여 분류성능을 측정하였다.

먼저 〈표 2〉에서 문헌빈도를 기준으로 분류성능이 좋은 구간을 선택하면, 문헌빈도가 150보다 크고 1,000보다 작은 구간과 20 보다 크고 30보다 작은 구간으로 볼 수 있다. 또한 이들을 조합하는 구간도 고려해볼 수 있다. 이렇게 선별된 구간들(Case1, Case2, Case3)의 자질의 분류성능은 〈표 7〉과 같았다. Case2은 앞서 설명한 성능 차이를 보여주기 위해 Case1을 바탕으로 구간을 더 미세하게 조정하였다. 이 세 가지 조합 중에 가장 좋은 성능을 가져온 것은 Case2로 0.6101의 값을 가져왔다.

〈표 7〉 문헌빈도와 장서빈도의 독자적인 조합

	Case1	Case2	Case3	Case4	Case5
조합 조건	150 < DF < 1000	300 < DF < 1000	20 < DF < 30 or DF > 300	300 < CF < 2000	250 < CF < 2000
자질 수	361	131	1,032	288	314
kNN20	0.5955	0.5955	0.5528	0.6256	0.6281
kNN30	0.6005	0.6131	0.5829	0.6332	0.6382
kNN40	0.6080	0.6030	0.6005	0.6508	0.6457
kNN50	0.5955	0.6131	0.6005	0.6457	0.6508
kNN60	0.5930	0.6256	0.5905	0.6533	0.6407
AVG	0.5985	0.6101	0.5854	0.6417	0.6407

〈표 5〉에서 장서빈도를 기준으로 분류성능이 좋은 구간을 선택하면, 장서빈도가 300 보다 큰 구간(Case4)이다. 이 구간을 더 확장하였다(Case5). 문헌빈도나 장서빈도만을 이용하여 구간을 지정하고 그에 따른 분류성능에서 가장 좋은 경우는 Case4로 0.6417의 값을 가져왔다. 이는 〈표 5〉에서 가장 좋은 평균 분류성능인 0.6513 보다 약간 낮은 값이다. 두 사례의 차이는 장서빈도가 2,000이 넘는 4개의 자질을 포함하느냐 아니냐의 차이이다. 즉 가장 높은 장서빈도를 갖는 자질도 분류성능에 영향을 미치는 것을 의미한다. 이러한 현상은 문헌빈도에서도 마찬가지 이다. 고빈도의 자질이 분류성능에 영향을 미치고 있음을 확인할 수 있다.

다음으로 문헌빈도와 장서빈도를 서로 조합하는 사례의 분류성능을 파악하여 〈표 8〉과 같은 결과를 얻었다. 첫 번째 조합(Case6)은 DF 〉 150이거나 CF 〉 300인 경우로 자질의 크기는 387개 이다. 두 번째 조합(Case7)은 DF 〉 150에 해당하는 자질이면서 CF 〉 300인 자질로 총 268개의 자질이 그 대상이었다. 그리고 Case8에서 Case11까지 다른 다양한 조합방식을 이용하여 자질을 선정하고 분류성능을 평가하였다. 실험 결과는 〈표 8〉와 나와 있듯이 그리 좋은 성능을 보이지 않았다.

〈표 8〉 문헌빈도와 장서빈도의 구간별 조합

	Case6	Case7	Case8	Case9	Case10	Case11
조합 조건	DF > 150 or CF > 300	DF > 150 and CF > 300	100< DF <500 and 100< CF <1300	100< DF <500 or 100< CF <1300	150< DF <300 and 300< CF <500	150< DF <300 or 300< CF <500
자질 수	387	268	575	1,076	135	347
kNN20	0.6206	0.6256	0.4799	0.4070	0.5578	0.4774
kNN30	0.6080	0.6332	0.4623	0.3894	0.5352	0.5251
kNN40	0.6005	0.6357	0.4422	0.3618	0.5126	0.5553
kNN50	0.5930	0.6281	0.4196	0.3492	0.5050	0.5628
kNN60	0.5779	0.6307	0.4020	0.3367	0.4849	0.5477
AVG	0.6000	0.6307	0.4412	0.3688	0.5191	0.5337

두 빈도의 조합을 통한 자질 선정 실험의 결과를 요약하면, 이 실험문헌집단에서는 단순히 하나의 빈도(장서빈도)를 이용하여 구간을 적절히 조정하는 것이 더 좋은 성능을 가져오는 것으로 나타났다. 다만, 장서빈도를 단순히 사용하는 것보다 더 좋은 성능을 가져오지는 못했다.

또한 빈도와 구간을 조합하여 자질의 수가 커질수록 분류성능이 낮아지는 경향을 보였으며, 이는 적절한 자질의 수를 확보하는 것을 고려해야 함을 의미한다.

Ⅴ. 결 론

이 연구에서는 자동 색인을 통해 쉽게 얻을 수 있는 자질의 문헌빈도와 장서빈도를 이용하여 자동분류에서 자질선정 기법을 kNN 분류기에 적용하였을 때, 어떠한 분류성능을 보이는지 알아보고자 하였다.

실험에 사용된 실험집단으로 한국일보-20000 실험문헌집단의 첫 번째 파일을 이용하였다. 이 파일은 총 3,988개의 문헌을 포함하고 있으며, 범주 정보는 2003년 버전의 대분류를 사용하였다. 이 실험집단에 대해 분류기를 구축하기 위해 임의로 90%의 학습문헌을 이용하였으며, 나머지 10%는 검증문헌으로 사용하여 분류기의 성능을 평가하였다.

실험 결과, 첫째로 문헌 범주화와 같은 자동분류에서는 고빈도를 사용하고 저빈도를 제거하는 것이 더 좋은 분류성능을 가져오는 것으로 나타났다. 특히 기존의 연구에서 자질 선정 기법으로 문헌빈도를 많이 사용하였으나, 장서빈도를 적용한 경우가 더 좋은 성능을 가져오는 것으로 나타났다. 장서빈도의 경우 이 실험에서는 111.1%의 분류성능의 향상을 가져왔다. 따라서 장서빈도를 이용한 방법도 문헌빈도 보다 더 좋은 성능을 보였기에 자질선정 기준으로 고려해볼 가치가 있음을 알 수 있었다.

두 번째, 문헌빈도와 장서빈도 모두 저빈도 자질을 우선으로 선정하는 방법은 좋은 성능을 가져오지 못했다. 이는 주로 저빈도 자질들의 수가 많아 자질의 크기가 줄어들지 않았기 때문인 것으로 생각되며, 저빈도 자질을 효과적으로 줄이는 방법을 연구할 필요가 있다.

세 번째, 자질 선정 기준으로 문헌빈도와 장서빈도를 결합 내지 조합하였는데, 이 연구에서는 단순히 하나의 빈도(장서빈도)를 이용하여 자질 선정 구간을 적절히 조정하는 것이 더 좋은 성능을 가져오는 것으로 나타났다. 다만, 장서빈도를 단순히 사용하는 것보다 더 좋은 성능을 가져오지는 못했다.

이 연구의 제한점으로는 하나의 실험문헌집단과 분류기를 사용하였으므로 연구결과의 보편성을 위해 보다 폭넓고 다양한 자동분류 실험을 수행할 필요가 있다. 또한 실험집단의 특성과 학습집단의 무작위 추출에 따라 실험의 결과가 달라질 수 있어 추가적인 실험이나 연구가 필요하다.

참고문헌

심경. “문헌범주화에서 학습문헌수 최적화에 관한 연구.” 정보관리학회지, 제23권, 제4호(2006. 12), pp.277-294.

이용구. "단어 중의성 해소를 위한 지도학습 방법의 통계적 자질선정에 관한 연구." 한국비블리아학회지, 제22권, 제2호(2011. 6), pp.5-25.

이재윤. "자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 대한 연구." 한국문헌정보학회지, 제39권, 제2호(2005. 6), pp.123-146.

정영미. 정보검색연구. 서울 : 구미무역 출판부, 2005.

정은경. "문서범주화 성능 향상을 위한 의미기반 자질확장에 관한 연구." 정보관리학회지, 제26권, 제3호(2009. 9), pp.261-278.

Azam, N. and J. Yao. "Comparison of term frequency and document frequency based feature selection metrics in text categorization." *Expert Systems with Applications*, Vol.39, No.5(2012), pp.4760-4768.

Guyon, I. and A. Elisseeff. "An Introduction to Variable and Feature Selection." *Journal of Machine Learning Research*, 3(2002), pp.1157-1182.

Jackson, P. and I. Moulinier. *Natural Language Processing for Online Applications - Text Retrieval, Extraction and Categorization*. Amsterdam : Benjamins Publishing Co., 2002.

Kim, J. et al. "HKIB-2000 & HKIB-40075: Hangul Benchmark Collections for Text Categorization Research." *Journal of Computing Science and Engineering*, Vol.3, No.3(Sep. 2009), pp.165-180.

Sebastiani, F. "Machine Learning in Automated Text Categorization." *ACM Computing Surveys*, Vol.34, No.1(2002), pp.1-47.

Shang, W. et al. "A novel feature selection algorithm for text categorization." *Expert Systems with Applications*, Vol.33, No.1(July. 2007), pp.1-5.

Tan, S. "Neighbor-weighted K-nearest Neighbor for Unbalanced Text Corpus." *Expert Systems with Applications*, Vol.28, No.4(2005), pp.667-671.

Yang, Y. and J.O. Pedersen. "A comparative study on feature selection in text categorization." In: *Proceedings of the 14th International Conference on Machine Learning*(1997), pp.412-420.

Yang, Y. and X. Lin. "A re-examination of text categorization methods." In: *Proceedings of the 22nd annual international ACM SIGIR conference on research and development in the information retrieval*(1999), pp.42-49.

HKIB 실험집단. 〈http://www.kristalinfo.com/TestCollections/#hkib〉 [cited 2012. 7. 10].

HAM 형태소 분석기. 〈http://nlp.kookmin.ac.kr/HAM/kor/〉 [cited 2012. 7. 15].

국한문 참고문헌의 영어 표기
(English translation / Romanization of references originally written in Korean)

Eun-Kyung, Chung. "A Semantic-Based Feature Expansion Approach for Improving the Effectiveness of Text Categorization by Using WordNet." *Journal of the Korean Society for information Management*, Vol.26, No.3(Sep. 2009), pp.261-278.

Jae-Yun, Lee. "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*, Vol.39, No.2(Jun. 2005), pp.123-146.

Kyung, Shim. "Optimization of Number of Training Documents in Text Categorization." *Journal of the Korean Society for information Management*, Vol.23, No.4(Dec. 2006), pp.277-294.

Yong-Gu, Lee. A Study on Statistical Feature Selection with Supervised Learning for Word Sense Disambiguation. *Journal of the Korean BIBLIA Society for library and Information Science*, Vol.22, No.2(Jun. 2011), pp.5-25.

Young-Mee, Chung. *Research in Information Retrieval*. Seoul : Gumi Trading, 2005.