자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 대한 연구*

An Empirical Study on Improving the Performance of Text Categorization Considering the Relationships between Feature Selection Criteria and Weighting Methods

이 재 윤(Jae-Yun Lee)**

목 차

초 록

이 연구에서는 문서 자동분류에서 분류자질 선정과 가중치 할당을 위해서 일관된 전략을 채택하여 kNN분류기의 성능을 향상시킬 수 있는 방안을 모색하였다. 문서 자동 분류에서 분류자질 선정 방식과 자질 가중치 할당 방식은 자동분류 알고리즘과 함께 분류성능을 좌우하는 중요한 요소이다. 기존 연구에서는 이 두 방식을 결정할 때 상반된 전략을 사용해왔다. 이 연구에서는 색인파일 저장공간과 실행시간에 따른 분류성능을 기준으로 분류자질 선정 결과를 평가해서 기존 연구와 다른 결과를 얻었다. 상호정보량과 같은 저빈도 자질 선호 기준이나 심지어는 역문헌빈도를 이용해서 분류 자질을 선정하는 것이 kNN 분류기의 분류 효과와 효율 면에서 바람직한 것으로 나타났다. 자질 선정 기준으로 저빈도 자질 선호 척도를 자질 선정 및 자질 가중치 할당에 일관되게 이용한 결과 분류성능의 저하 없이 kNN 분류기의 처리 속도를 약 3배에서 5배 정도 향상시킬 수 있었다.

ABSTRACT

This study aims to find consistent strategies for feature selection and feature weighting methods, which can improve the effectiveness and efficiency of kNN text classifier. Feature selection criteria and feature weighting methods are as important factor as classification algorithms to achieve good performance of text categorization systems. Most of the former studies chose conflicting strategies for feature selection criteria and weighting methods. In this study, the performance of several feature selection criteria are measured considering the storage space for inverted index records and the classification time. The classification experiments in this study are conducted to examine the performance of IDF as feature selection criteria and the performance of conventional feature selection criteria, e.g. mutual information, as feature weighting methods. The results of these experiments suggest that using those measures which prefer low-frequency features as feature selection criterion and also as feature weighting method, we can increase the classification speed up to three or five times without loosing classification accuracy.

키워드: 문서범주화, 자동분류, 자질선정, 자질가중치, kNN 분류기
Text Categorization, Automatic Classification, Feature Selection, Feature Weighting Methods, kNN Classifier

* 본 연구는 2004학년도 경기대학교 학술연구비(신진연구과제) 지원에 의하여 수행되었음.
** 경기대학교 문헌정보학과 (memexlee@kgu.ac.kr)
논문접수일자 2005년 5월 15일
게재확정일자 2005년 6월 8일

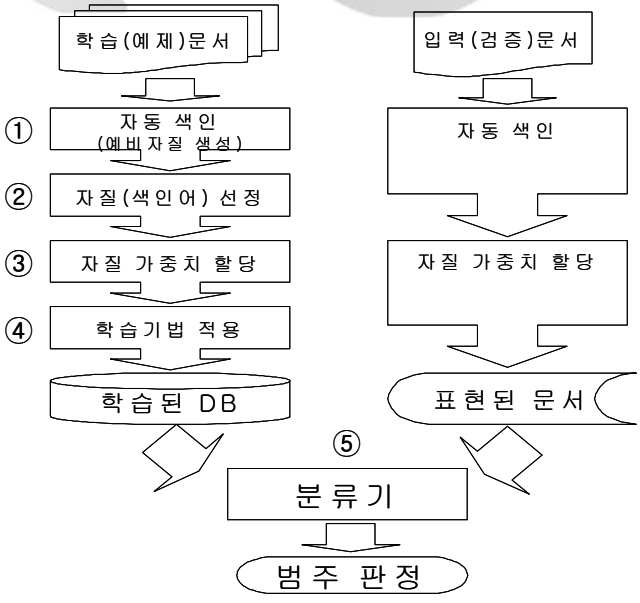
1. 서 론

인터넷의 대중화와 전자도서관의 활성화로 인하여 원문정보가 폭발적으로 증가하고 있다. 이런 대량의 정보 중에서 이용자에게 적합한 정보를 선별하여 제공하기 위해서는 효율적인 정보관리 및 검색이 필요하다. 특히 최근에는 문서의 내용을 바탕으로 미리 정의된 분류범주를 문서에 부여함으로써 문서를 자동분류하는 문서 범주화에 대한 연구가 활성화되고 있다.

문서 범주화 혹은 문서 자동분류는 크게 규칙기반 방식과 기계학습기반 방식으로 나눌 수 있는데, 최근에는 기계학습기반 문서 자동분류에 대한 연구가 활발하게 이루어지고 있다. 이 방식은 〈그림 1〉과 같이 사전에 분류가 되어있는 학습문서를 처리하는 학습과정과, 분류대상 문서를 입력받아서 적합범주를 판정하는 분류과정으로 이루어진다.

여기서 ④의 학습기법 적용 단계와 ⑤의 분류 실행 단계는 적용되는 분류기의 특성에 좌우되며, kNN(k-Nearest Neighbor)과 같은 분류기는 ④의 학습기법 적용단계가 없으므로 게으른 학습(lazy learning) 기법이라고 불리기도 한다. 이와 달리 단계 ②의 분류자질 선정과 단계 ③의 자질 가중치 할당은 적용되는 분류기법과는 별도로 분류 성능에 영향을 미치는 중요한 요소이다.

분류자질 선정과 가중치 할당의 목적은 똑같이 분류에 도움이 되는 자질—색인어—을 그렇지 못한 것과 구분하려는 것이다. 다만 분류자질 선정에서는 그 구분이 자질의 선정 혹은 제거라는 이원 판정으로 나타나는 반면에, 가중치 할당에서는 자질의 중요도에 비례하는 가중치의 높고 낮음으로 나타나는 것이 다르다.



〈그림 1〉 기계학습 기반 문서 자동분류 과정

그럼에도 불구하고 기본적으로는 문서의 분류에 어떤 자질이 더 도움이 되는가를 파악하려는 목적은 같다.

그런데 실제 분류자질 선정 단계에서는 문헌빈도가 높은 자질을 선정하는 것이 성능 면에서 유리하다고 알려져 있으며(Yang and Pederson 1997) 가중치 할당 단계에서는 역문헌빈도와 같이 저빈도 자질의 중요성을 강조하는 전략이 주로 사용되고 있다(Sebastiani 2002). 이 연구에서는 이와 같이 기본 목적을 공유하는 자질 선정 단계와 자질 가중치 할당 단계가 전략상으로는 상반되는 원인을 살펴보고, 일관된 전략을 적용하여 문서 자동분류의 성능과 효율을 향상시킬 수 있는 가능성을 찾고자 한다.

2. 분류자질 선정과 자질 가중치 할당

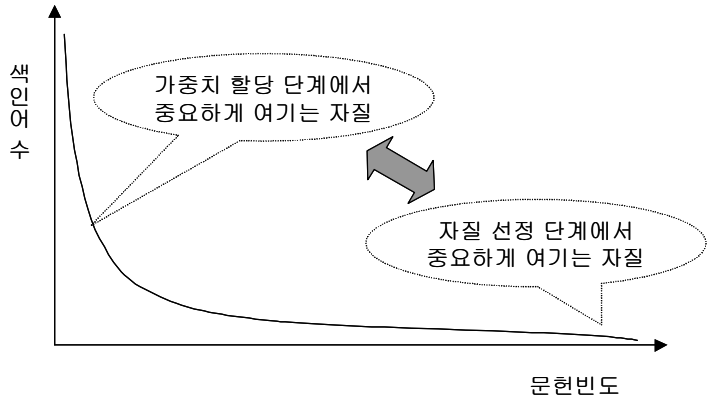
문서 자동 분류 분야에서는 흔히 분류자질 가중치 할당 전략으로 정보검색 분야와 마찬가지로 TFIDF 방식을 채택하고 있다. 일부 이진 가중치를 사용한 연구들도 가중치를 다루지 못하는 분류기의 한계 때문에 TFIDF를 채택하지 못한 경우가 대부분이라고 Sebastiani(2002)는 지적하였다. TFIDF방식의 가중치 할당에서는 역문헌빈도 때문에 문서집단 내에서 적은 수의 문서에 출현한 색인어가 높은 가중치를 부여받는다.

한편 자질 선정 단계에서는 자질 가중치 할당 방식처럼 저빈도 자질을 중요시하는 전략을 취하지 않는다. 오히려 저빈도 자질을 배척하는 정보획득량(Information Gain)이나 카이제곱통계량이 자질 선정 기준으로 주로 사용되고 있다. Yang과 Pederson(1997)은 문헌빈도를 기준으로 단순히 저빈도어를 제거하는 방식으로도 복잡한 자질 선정 기준을 이용한 경우에 견줄만한 성능을 얻었다고 보고하였다. 이들은 실제로 문헌빈도가 높은 자질일수록 정보획득량도 높아지는 상관관계가 있음을 알아내었고, 이를 근거로 일반적으로 정보검색에서 역문헌빈도 가중치의 근거로 사용되는 가정이 자동분류에서는 통하지 않는다고 주장하였다. 이들의 연구 이후에는 단순하다는 이유로 분류자질 선정 기준으로 문헌빈도를 채택하여 상대적인 고빈도어만을 분류자질로 삼는 연구도 다수 발표되었다(김제욱, 김한준, 이상구 2002; 이재윤, 유수현 2003).

결국 자질 선정 단계와 자질 가중치 할당 단계가 문서의 분류에 어떤 자질이 더 도움이 되는가를 파악하려는 목적을 공유함에도 불구하고 〈그림 2〉와 같이 실제로 채택하고 있는 전략은 상반되는 것이 현실이다.

이와 같은 두 단계의 전략 불일치는 자동분류 알고리즘 중에서도 성능이 우수한 것으로 알려진 kNN 분류기에서 더 두드러진다. kNN 분류기는 분류 대상 문서를 질의로 이용하여 학습문서 중에서 유사한 k개의 문서를 검색한 다음, 검색된 문서들의 소속 범주를 근거로 분류 대상 문서의 적정 범주를 결정하게 된다. 기본적인 형태는 정보검색과 같음에도 불구하고 자동분류에서는 고빈도어가 아닌 저빈도어를 마치 불용어인 것처럼 배제하는 상황인 것이다.

이 연구에서는 전략 불일치의 원인을 자질 선정 기준의 분류성능을 비교 평가하는 방법에서 찾아보고자 한다. 그 이유는 기존에 널리 사용되어온 비교 평가 방법이 특히 저빈도 자질



〈그림 2〉 자질 선정 단계와 가중치 할당 단계의 상반된 전략

을 중요시하는 선정 기준에게 불리한 측면이 있다고 생각되기 때문이다.

문서 분류만 아니라 일반적인 자동분류에 있어서 자질 선정은 주어진 자질 집합 중에서 수행 시간이나 저장 공간 면에서 효율적인 부분집합을 경험적으로 찾으려는 것이다(Alpaydin 2004). 이에 따라 자질 선정 기준의 성능을 비교하기 위해서는 각 자질 선정 기준에 따라서 동일한 차원(자질 종수)의 축소된 자질집합을 만들어서 이를 이용한 분류 성능을 측정한다. Yang과 Pederson(1997)도 이 방법으로 각 자질 선정 기준의 성능을 평가하였으며 최근까지도 많은 문서 자동분류 연구에서 채택하고 있는 기준이다(박부영 2004; Forman 2002; Fragoudis, Meretakis, and Likothanassis 2005; Galavotti, Sebastiani, and Simi 2000; Zu et al. 2003).

학습문서집단을 문서-용어 행렬로 표현하는 분류기에서는 자질 선정 기준별로 동일한 크기의 행렬을 입력데이터로 삼게 되므로 앞의 방법이 저장 공간 면에서 공정한 비교 기준이라고 할 수 있다. 반면에 도치색인파일로 학습문서집단을 표현하는 분류기에서는 자질 선정 기준에 따라서 색인어의 종수가 같더라도 색인파일의 규모는 다를 수 있다. 예를 들어 500개의 문서에 출현한 자질은 1개의 문서에 출현한 자질에 비해서 도치색인레코드의 수가 500배에 달하므로 저장공간 측면에서 공정한 비교 기준이 될 수 없다. 또한 이와 같은 저장 공간의 차이는 실행시간 측면에도 영향을 미친다.

*k*NN 분류기는 지지벡터기계(SVM; Support Vector Machine)과 함께 성능이 가장 좋은 문서분류 알고리즘의 하나로 알려져 있으며(Yang and Liu 1999), 단순한 알고리즘 때문에 구현하기가 용이하므로 문서분류에 대한 여러 연구에서 이용되었고 Verity나 InXight와 같은 여러 상업용 문서분류 시스템에서도 채택하고 있다(Blumberg and Atre 2003).

*k*NN 분류기를 구현할 때에는 알고리즘의 특성상 검색시스템과 마찬가지로 학습문서를 도치색인파일에 저장하는 경우가 많다(Cöster and Svensson 2002; Lim 2002; Yang

1999). Zhou 등(2003)은 도치색인파일을 이용해서 kNN 분류기를 구현함으로써 시스템 효율을 크게 향상시킬 수 있었다고 보고하였다. 따라서 도치색인파일을 사용하는 kNN 분류기에 대해서 자질 선정 기준을 비교 평가할 때에는 축소된 자질 집합을 저장하는 색인레코드의 수를 고려해야 할 것이다.

이를 감안하여 이 연구에서는 다음과 같은 세 가지 기준에 따라서 kNN 분류기에 대한 분류자질 선정 기준을 평가하고, 이를 통해서 효율적이고 효과적인 자질 선정 및 가중치 할당 전략을 알아보고자 한다.

(1) 총 자질 종수 기준 분류 성능 비교 - Yang과 Pederson(1997)의 실험과 마찬가지로 각 자질 선정 기준별로 동일한 종수의 축소자질 집합을 사용하여 분류 성능을 비교하는 방법이다.

(2) 저장공간 기준 분류 성능 비교 - 각 자질 선정 기준별로 학습문서집단을 저장한 색인레코드의 규모를 동일하게 하였을 경우의 분류 성능을 비교하는 방법이다. 이는 분류자질 선정의 목적을 저장공간 복잡도의 감소로 본 경우이다.

(3) 실행시간 기준 분류 성능 비교 - 각 자질 선정 기준별로 축소된 자질집합을 이용하여 문서를 분류하는데 소요되는 시간을 측정하여 이를 기준으로 분류성능을 비교하는 방법이다. 동일한 규모의 저장공간이 동일한 실행시간을 보장하지 않기 때문에 이와 같은 실행시간 비교도 필요하다. 이는 분류자질 선정의 목적을 실행시간 복잡도의 감소로 본 경우이다.

3. 실험 설계

앞에서 제시한 분류자질 선정 기준의 세 가지 비교 방법에 따라서 실제 문서 자동 분류 실험을 수행하고 여러 분류자질 선정 기준의 성능을 평가해 보았다. 실험을 위한 도치색인파일 기반 kNN 분류기는 Windows XP를 운영체제로 하는 펜티엄Ⅳ PC 상에서 Visual Foxpro로 구현하였다.

분류 실험을 위한 실험 문서 집단으로는 〈표 1〉과 같이 특성이 다른 두 집단을 사용하였다. KFCM-896 분류실험집단은 1992년 국제 및 경제분야 신문기사 896건으로 구성되어 있으며 각 범주별로 게재 시기가 늦은 기사 20%(178건)를 검증집단(분류기의 성능 실험을 위한 분류대상 문서집단)으로, 그보다 게재 시기가 이른 기사 80%(718건)를 학습집단(분류기의 학습 데이터 집단)으로 구분하고 있다. 기사의 분류는 1992년판 『전국언론사 기사자료 표준 분류표』에 따라 이루어져 있으며 이

〈표 1〉 실험에 사용된 문서 집단

명칭	내용	문서의 수(학습/검증)	범주의 수
KFCM-896	신문기사	896(718/178)	17
TREND-12746	해외과학기술문헌속보	12,746(7,466/5,280)	18

실험에서는 분류의 깊이를 두 번째 중분류 수준까지 적용한 결과 17개 범주로 구성되었다. TREND-12746 분류실험집단은 정보검색용 실험집단인 HANTEC v.2.0(김지영 외 2000) 중에서 분류정보가 포함된 해외과학기술문헌속보 문서만을 추출한 것이다. 이 실험에서는 1997년과 1998년에 등록된 문서 12,746건 중에서 1997년에 등록된 7,466건을 학습집단으로, 1998년에 등록된 5,280건을 검증집단으로 사용하였다. 해외과학기술문헌속보 문서에는 KISTI(작성 당시에는 KORDIC)에서 분류번호를 부여하였는데 이 실험에서는 18개의 대분류 범주를 적용하였다.

이 실험문서들을 색인할 때 학습문서 집단에서 문헌빈도 2 이하인 색인어는 추후 입력되는 분류대상 문서에 나타날 가능성이 희박하여 시스템 효율만 저하시킬 가능성이 크므로 여러 선행연구(Forman 2002; Rogati and Yang 2003)와 마찬가지로 제외하였다. 만일을 위해서 사전 실험으로 문헌빈도 2 이하인 색인어를 포함한 경우와 제외한 경우의 분류성능을 비교한 결과 차이가 없음이 확인되었다. 또한 색인어 가중치 $W(t_i)$는 사전 실험에서 좋은 성능을 보인 것으로 확인된 이진TF에 IDF를 곱한 공식을 사용하였다.

$$W(t_i) = TF \times \log\frac{N}{df}\ ,\ (if\ \ tf > 0,$$

$$TF = 1\ \ else\ \ TF = 0\ )$$

비교를 위한 분류자질 선정 기준으로는 우선 Yang과 Pederson(1997)이 비교한 문헌빈도(DF), 정보획득량(IG), 카이제곱통계량(CHI), 상호정보량(MI)을 포함하였다. 또한 저빈도 자질보다는 고빈도 자질을 우선적으로 채택하는 자질 선정 기준이 자동분류에 유리하다는 Yang과 Pederson(1997)의 주장을 검증하기 위해서 다른 자질 선정 기준을 추가하였다. 정보획득량과 같이 고빈도 자질에 유리한 기준으로는 자카드 계수(JAC), 코사인 계수(COS), GSS 계수(GSS)의 세 가지 척도를 추가하였다. 반면에 상호정보량처럼 저빈도 자질에 유리한 기준으로는 로그승산비(LOR), 상대적 상호정보량 J(RMIJ), 역문헌빈도(IDF)를 추가하였다. 또한 6장의 가중치 실험에서는 로그승산비와 동일한 자질값 순위를 보여주면서 값의 범위가 -1에서 1로 제한되는 율의 Y(YULE)를 추가로 사용하였다.

전체 학습문서 수가 N이고 단어 k의 출현확률을 $P(t_k)$, 범주 i의 출현확률을 $P(c_i)$라고 할 때 각 자질 선정 기준을 공식으로 나타내면 〈표 2〉와 같다. 이 표에서 공식을 표기한 방식은 문서자동분류에 대해 폭넓게 소개한 논문으로서 자주 인용되는 Sebastiani(2002)의 확률표기방식을 따랐으며, 각 공식의 특징은 다음 장의 실험결과에서 확인된 사실에 따라서 적었다.

분류실험 결과의 성능은 마이크로 평균 정확률 척도로 평가하였다. 마이크로 평균 정확률 척도는 전체 범주 할당 건수 중에서 옳게 분류된 경우의 비율을 산출하는 것이다. 이 연구에서 사용한 실험집단에는 각 문서마다 분류기호가 하나씩만 할당되어 있으므로 마이크로 평균 정확률과 마이크로 평균 재현율, 그리고 이를 결합한 마이크로 평균 F1 척도는 같은 값을 가진다. 분류실험의 성능 평가를 위한 척도로는 이밖에도 각 범주별로 정확률과 재현율을 구해

〈표 2〉 분류자질 선정 기준

명칭	약호	공 식	특징
문헌빈도	DF	$P(t_k)$	고빈도 자질 우선
정보획득량	IG	$P(t_k,c_i)\log\frac{P(t_k,c_i)}{P(c_i)P(t_k)}+P(\overline{t_k},c_i)\log\frac{P(\overline{t_k},c_i)}{P(c_i)P(\overline{t_k})}$	고빈도 자질 우선
GSS 계수	GSS	$P(t_k,c_i)P(\overline{t_k},\overline{c_i})-P(t_k,\overline{c_i})P(\overline{t_k},c_i)$	고빈도 자질 우선
자카드 계수	JAC	$\frac{P(t_k,c_i)}{P(t_k)+P(c_i)-P(t_k,c_i)}$	고빈도 자질 우선
코사인 계수	COS	$\frac{P(t_k,c_i)}{\sqrt{P(t_k)P(c_i)}}$	고빈도 자질 우선
카이제곱통계량	CHI	$\frac{N\left(P(t_k,c_i)P(\overline{t_k},\overline{c_i})-P(t_k,\overline{c_i})P(\overline{t_k},c_i)\right)^2}{P(t_k)P(\overline{t_k})P(c_i)P(\overline{c_i})}$	고/중간빈도 자질 우선
상대적 상호정보량 J	RMIJ	$\frac{\log_2P(t_k)+\log_2P(c_i)-\log_2P(t_k,c_i)}{\log_2P(t_k,c_i)}$	저/중간빈도 자질 우선
로그승산비	LOR	$\log\frac{P(t_k,c_i)P(\overline{t_k},\overline{c_i})}{P(t_k,\overline{c_i})P(\overline{t_k},c_i)}$	저빈도 자질 우선
율의 Y	YULE	$\frac{\sqrt{P(t_k,c_i)P(\overline{t_k},\overline{c_i})}-\sqrt{P(t_k,\overline{c_i})P(\overline{t_k},c_i)}}{\sqrt{P(t_k,c_i)P(\overline{t_k},\overline{c_i})}+\sqrt{P(t_k,\overline{c_i})P(\overline{t_k},c_i)}}$	저빈도 자질 우선
상호정보량	MI	$\log_2\frac{P(t_k,c_i)}{P(t_k)P(c_i)}$	저빈도 자질 우선
역문헌빈도	IDF	$-\log_2P(t_k)$	저빈도 자질 우선

서 평균을 산출하는 매크로 평균 정확률 및 재현율 척도가 있다. 매크로 평균 척도는 크기가 매우 작은 범주의 성능에 지나치게 영향을 받으므로 문서자동분류 실험의 평가를 위해서는 마이크로 평균 척도가 더 널리 사용된다(Yang and Liu 1999).

4. 총 자질 종수 기준 분류 성능 비교

4. 1 자질 선정 기준별 분류 실험 결과

선행 연구에서 사용되었던 분류자질의 성능 평가 방식과 마찬가지로 총 자질 종수(자질 차원)를 동일하게 한 상태의 분류 성능을 먼

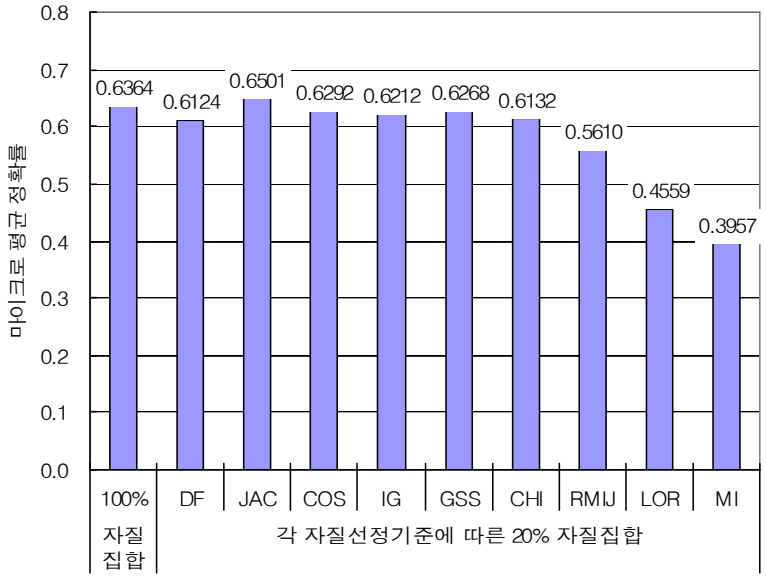
저 평가해 보았다. 이 평가의 목적은 이 연구에서 사용된 실험문서 집단에 대해서 선행 연구의 결과를 재확인하는 것과 함께, 각 자질 선정 기준의 빈도수준 선호 특성을 확인하려는 것이다.

Yang과 Pederson(1997) 의 연구에서는 저빈도 선호 성향을 가진 상호정보량이 매우 낮은 성능을 보였고, 문헌빈도, 정보획득량, 카이제곱통계량은 비슷하게 성능이 좋은 것으로 나타났다. 선정 기준으로 사용되는 연관성 척도의 빈도수준 선호 성향이 이런 성능 차이에 영향을 미쳤는지 여부를 확인하기 위해서 이 실험에서는 다른 척도를 추가하였다. 고빈도 선호 척도로는 자카드 계수, 코사인 계수, GSS 계수를 추가하였고, 저빈도 선호 척도로는 로그승산비와 상대적 상호정보량 J를 추가하였다. 추가된 척도들은 카이제곱통계량이나 상호정보량과 마찬가지로 모두 각 분류자질과 범주 사이의 연관성을 측정하여 각 척도별로 연관성이 높은 자질을 채택하게 된다.

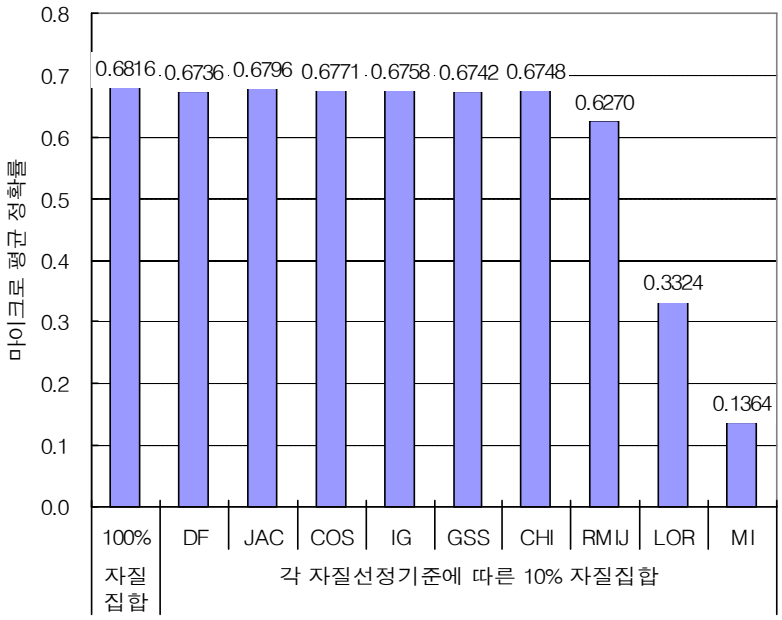
이들 9가지 선정 기준에 따라 축소된 자질집합을 이용한 경우의 분류성능을 측정하여 전체 자질을 모두 사용한 경우의 분류성능과 함께 〈그림 3〉과 〈그림 4〉에 제시하였다. 규모가 작은 KFCM-896 실험집단에서는 원래의 20%로 축소한 집합을, 규모가 큰 TREND-12746 실험집단에서는 원래의 10%로 축소한 집합을 사용하였다.

실험 결과에서 얻어진 사실은 다음과 같다.

첫째, Yang과 Pederson(1997)의 결과와 마찬가지로 상호정보량은 매우 나쁜 성능을 보였으며 문헌빈도, 정보획득량, 카이제곱통계량은 이보다 훨씬 좋은 성능을 보였지만 전체 자질을 사용한 경우보다는 약간 성능이 뒤쳐졌다.



〈그림 3〉 20%로 축소된 자질집합의 분류 성능 비교 (KFCM-896)



〈그림 4〉 10%로 축소된 자질집합의 분류 성능 비교 (TREND-12746)

둘째, 추가로 적용해 본 척도 중에서 고빈도 선호 성향을 가진 자카드 계수, 코사인 계수, GSS 계수는 모두 좋은 성능을 보였고, 저빈도 선호 성향을 가진 상대적 상호정보량 J, 로그승산비는 상호정보량보다는 좋으나 대체적으로 나쁜 분류 성능을 보였다. 전체적으로는 두 실험집단에서 모두 자카드 계수로 자질을 선정한 경우에 가장 좋은 결과가 나타났으며 KFCM-896 실험집단에 대해서는 전체 자질을 사용한 것보다도 자카드 계수에 의한 20% 축소 자질집합을 사용한 경우가 더 좋은 성능을 보였다.

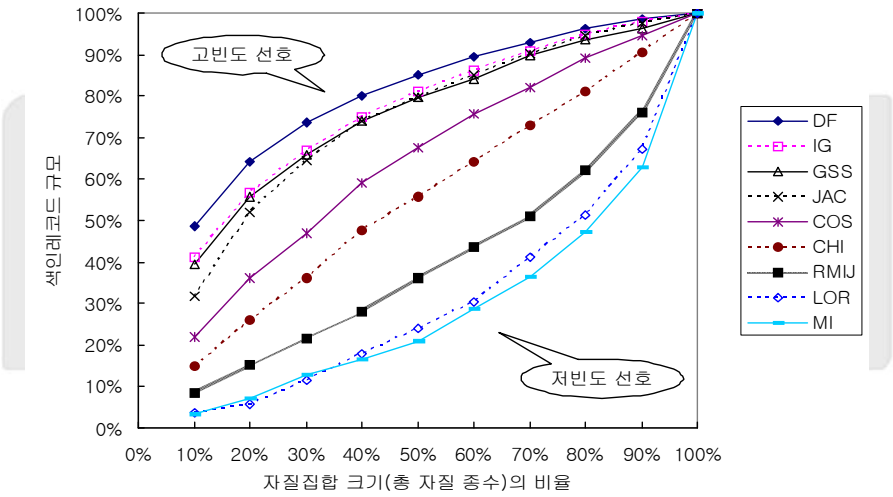
이상의 사실에서 총 자질 종수를 일치시켜서 평가할 경우에는 자질집합 선정기준으로 저빈도 선호 기준을 사용하는 것에 비해서 고빈도 선호 기준을 사용하는 것이 분류 성능 면에서 유리하다는 점이 확인되었다.

4.2 총 자질 종수 기준의 문제점

위와 같이 총 자질 종수, 즉 자질 차원을 동일하게 한 상태에서 각 자질 선정 기준의 성능을 비교하는 것은 도치색인파일을 저장구조로 하는 kNN 분류기에서는 불공정한 평가기준이 될 수가 있다. 동일한 종수의 자질이라 하더라도 저빈도 자질과 고빈도 자질이 각각 차지하는 저장공간의 크기에는 큰 차이가 있기 때문이다. 자질 선정 기준에 따라서 실제 저장공간이 어느 정도 차이가 나는지 알아보기 위해서는 도치색인레코드의 수를 비교해 보아야 한다. 두 실험집단에 대해서 자질집합의 크기를 전체 집합에서부터 10% 포인트씩 줄여나가면서 실제 저장되는 도치색인레코드의 수를 파악하여 〈표 3〉과 〈그림 5〉, 〈표 4〉와 〈그림 6〉에 각

〈표 3〉 축소된 도치색인레코드의 규모 비교(KFCM-896)

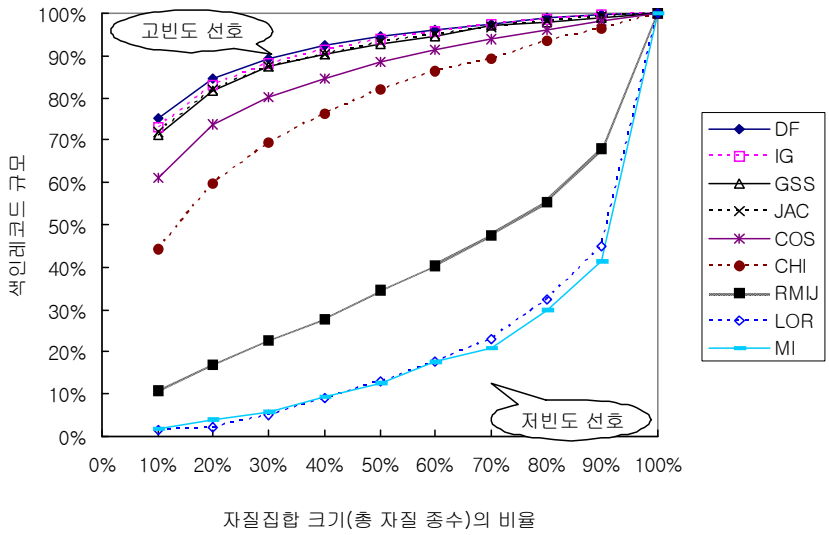
자질집합 크기 (%)	자질 선정 기준								
	DF	IG	GSS	JAC	COS	CHI	RMIJ	LOR	MI
10%	48.7%	41.1%	39.5%	31.6%	21.8%	14.7%	8.5%	3.7%	3.4%
20%	64.1%	56.7%	55.8%	52.0%	36.1%	26.1%	15.1%	5.9%	7.0%
30%	73.5%	66.9%	65.9%	64.7%	47.0%	36.0%	21.8%	11.5%	12.7%
40%	80.1%	74.8%	73.9%	74.0%	59.3%	47.5%	28.0%	18.0%	16.5%
50%	85.2%	81.0%	79.7%	79.8%	67.7%	55.6%	36.0%	24.0%	20.8%
60%	89.5%	86.0%	84.2%	85.1%	75.8%	64.2%	43.5%	30.3%	28.7%
70%	92.9%	90.9%	89.9%	90.1%	82.2%	73.0%	51.1%	41.1%	36.6%
80%	96.3%	94.8%	93.4%	94.6%	89.3%	81.2%	62.1%	51.5%	47.4%
90%	98.5%	98.1%	96.2%	97.8%	94.4%	90.7%	75.9%	67.4%	62.8%
100%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%



〈그림 5〉 축소된 도치색인레코드의 규모 비교(KFCM-896)

〈표 5〉 축소된 도치색인레코드의 규모 비교 (TREND-12746)

자질집합 크기 (%)	자질 선정 기준								
	DF	IG	GSS	JAC	COS	CHI	RMIJ	LOR	MI
10%	75.2%	73.2%	71.4%	72.0%	61.2%	44.4%	10.7%	1.5%	1.7%
20%	84.6%	83.1%	81.8%	82.0%	73.8%	59.6%	17.0%	2.3%	4.1%
30%	89.3%	88.1%	87.3%	87.3%	80.1%	69.6%	22.6%	5.1%	5.6%
40%	92.3%	91.4%	90.2%	90.3%	84.7%	76.4%	27.9%	9.1%	9.3%
50%	94.5%	93.8%	92.8%	93.0%	88.7%	82.0%	34.6%	13.0%	12.5%
60%	96.2%	95.7%	94.7%	95.1%	91.5%	86.3%	40.4%	17.5%	17.8%
70%	97.6%	97.3%	97.0%	96.7%	93.8%	89.3%	47.6%	23.1%	20.9%
80%	98.7%	98.6%	97.7%	98.2%	96.1%	93.4%	55.5%	32.3%	29.7%
90%	99.5%	99.5%	98.8%	99.3%	98.2%	96.6%	68.1%	44.8%	41.2%
100%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%



〈그림 6〉 축소된 도치색인레코드의 규모 비교 (TREND-12746)

각 제시하였다.

〈그림 5〉와 〈그림 6〉에서 저빈도 자질을 선호하는 선정기준은 그래프의 곡선이 오른쪽 아래로, 고빈도 자질을 선호하는 선정기준은 그래프의 곡선이 왼쪽 상단으로 치우쳐 나타난다. 만약 고빈도나 저빈도 자질 중 어느 쪽에도 치우치지 않는 중립적인 선정기준이 있다면 왼쪽 아래에서 오른쪽 위로 완전한 사선이 그어질 것이다. 규모가 작은 KFCM-896 실험집단에서는 카이제곱통계량이 비교적 중립에 가깝게 나타났다. 반면에 규모가 큰 TREND-12746 실험집단에서는 카이제곱통계량도 고빈도 자질 선호경향이 큰 것으로 나타났으며 오히려 상대적 상호정보량 J가 중립에 조금 가까운 것으로 보인다.

두 실험집단 중에서는 실험문서의 수가 훨씬 많은 TREND-12746 집단에서 저빈도 선호 성향과 고빈도 선호 성향의 차이가 더 뚜렷하게 나타났다. 분석 결과 드러난 특성에 따라서 고빈도 자질을 선호하는 정도를 기준으로 각 자질 선정 기준을 나열하면 다음과 같다.

DF 〉 IG 〉 GSS ≒ JAC 〉 COS 〉 CHI 〉 RMIJ 〉 LOR 〉= MI

고빈도 자질을 선호하는 기준인 정보획득량으로 선정된 10% 자질 집합은 저빈도 자질을 선호하는 기준인 상호정보량으로 선정된 10% 자질 집합에 비해서 KFCM-896 실험집단에서는 12배(41.1% 대 3.4%), TREND-12746 실험집단에서는 무려 43배(73.2% 대 1.7%)의 저장공간을 차지한다. TREND-12746 실험집단의 경우에는 상호정보량을 기준으로 90%의 자질을 포함하더라도 색인파일의 규모는 정보획득량으로 10%의 자질만 포함한 경우보다 더 작다. 도치색인파일을 사용하는 kNN

분류기에서는 자질의 종수를 기준으로 자질집합을 축소하는 것이 각 자질 선정 기준을 공평하게 비교하는 방법이라고 할 수 없는 것이다.

5. 저장공간 및 실행시간 기준 분류 성능 비교

각 자질 선정 기준에 따라 축소 생성된 부분 자질집합의 분류성능을 실제 저장공간 및 실제 처리시간을 기준으로 평가하는 실험을 수행하였다. 이 실험에서는 자질 종수를 일치시키지 않고 색인파일의 규모, 즉 도치색인레코드의 수를 일치시켰을 때의 각 자질 선정 기준별 분류성능을 비교하였다. 도치색인레코드의 수를 학습문서의 수로 나누면 학습문서당 자질 종수의 평균이 된다. 이전 실험에서 전체 문서집단을 표현하는 자질 종수를 기준으로 평가한 것에 반해 이번 실험에서는 한 문서를 표현하는 자질 종수의 평균을 기준으로 하도록 바꾼 셈이다.

5.1 저장공간 기준 분류 성능 비교

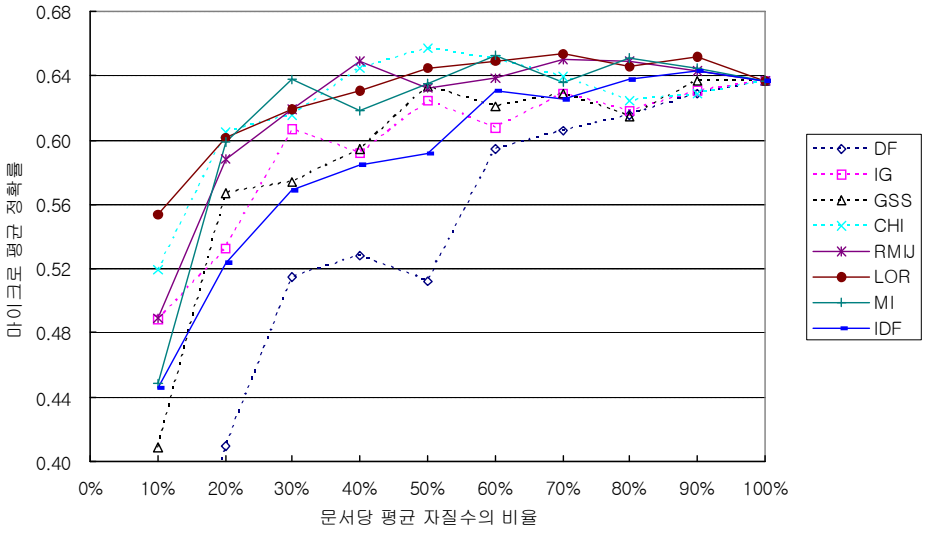
이 실험에서는 앞에서 검토했던 자질 선정 기준 중에서 고빈도 자질 선호 기준으로 확인된 문헌빈도, 정보획득량, GSS 계수, 카이제곱 통계량, 그리고 저빈도 자질 선호 기준으로 확인된 상대적 상호정보량 J, 로그승산비, 상호정보량 등과 함께 역문헌빈도를 추가하여 실험하였다. 역문헌빈도는 문헌빈도와 반대로 극단적으로 저빈도 자질만을 선정하는 기준이 된다. 앞에서 총 자질 종수를 기준으로 평가하는 것이 저빈도 자질 선호기준에 불리하였다고 확인된 만큼, 자질 가중치 할당에 사용되어온 역문헌빈도를 저빈도 자질 선호 기준의 하나로써 가능성을 검토해 보고자 한다.

도치색인파일의 규모(색인레코드의 총 수)를 100%에서부터 10% 포인트씩 줄이면서 각 분류자질 선정 기준에 따른 부분 자질집합의 분류성능을 비교한 결과를 〈표 5〉와 〈그림 7〉, 〈표 6〉과 〈그림 8〉에 제시하였다.

〈표 6〉 축소된 색인파일의 분류성능 비교 (KFCM-896, 마이크로 평균 정확률)

색인파일규모	자질 선정 기준							
	DF	IG	GSS	CHI	RMIJ	LOR	MI	IDF
10%	0.2360	0.4888	0.4085	0.5193	0.4896	0.5538	0.4486	0.4462
20%	0.4101	0.5321	0.5666	0.6051	0.5883	0.6011	0.5987	0.5233
30%	0.5153	0.6067	0.5738	0.6156	0.6188	0.6188	0.6380	0.5690
40%	0.5281	0.5915	0.5947	0.6445	0.6493	0.6308	0.6180	0.5843
50%	0.5120	0.6244	0.6332	0.6573	0.6324	0.6445	0.6348	0.5915
60%	0.5947	0.6076	0.6212	0.6501	0.6388	0.6493	0.6525	0.6308
70%	0.6059	0.6284	0.6284	0.6397	0.6501	0.6533	0.6356	0.6252
80%	0.6156	0.6180	0.6148	0.6244	0.6493	0.6453	0.6509	0.6380
90%	0.6292	0.6308	0.6364	0.6292	0.6429	0.6517	0.6445	0.6429
100%	0.6364	0.6364	0.6364	0.6364	0.6364	0.6364	0.6364	0.6364

* 음영 부분은 동일 규모의 색인파일(가로줄)을 기준으로 가장 성능이 좋은 경우



〈그림 7〉 축소된 색인파일의 분류성능 비교 (KFCM-896)

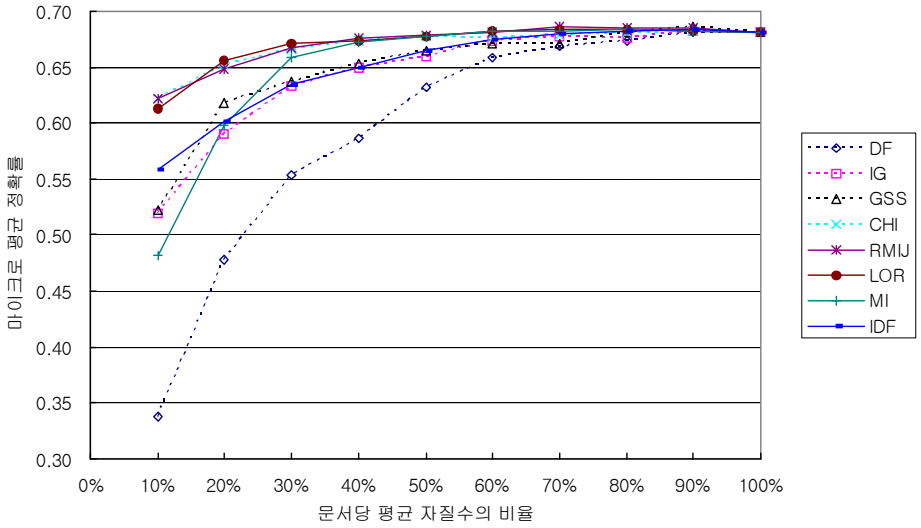
〈표 7〉 축소된 색인파일의 분류성능 비교 (TREND-12746, 마이크로 평균 정확률)

색인파일규모	자질 선정 기준							
	DF	IG	GSS	CHI	RMIJ	LOR	MI	IDF
10%	0.3374	0.5190	0.5215	0.6220	0.6213	0.6124	0.4818	0.5593
20%	0.4778	0.5898	0.6183	0.6515	0.6488	0.6554	0.5978	0.6013
30%	0.5534	0.6334	0.6363	0.6657	0.6673	0.6710	0.6579	0.6343
40%	0.5865	0.6491	0.6538	0.6731	0.6755	0.6737	0.6721	0.6499
50%	0.6323	0.6597	0.6652	0.6769	0.6780	0.6773	0.6778	0.6652
60%	0.6590	0.6732	0.6709	0.6756	0.6813	0.6825	0.6825	0.6746
70%	0.6685	0.6771	0.6712	0.6789	0.6860	0.6835	0.6829	0.6803
80%	0.6739	0.6764	0.6812	0.6811	0.6853	0.6837	0.6842	0.6823
90%	0.6807	0.6827	0.6858	0.6820	0.6844	0.6829	0.6821	0.6833
100%	0.6816	0.6816	0.6816	0.6816	0.6816	0.6816	0.6816	0.6816

* 음영 부분은 동일 규모의 색인파일(가로줄)을 기준으로 가장 성능이 좋은 경우

실험 결과에서 관찰된 사실은 다음과 같다.

첫째, 동일한 규모의 색인파일을 사용하였을 때 가장 좋은 성능을 보이는 기준은 대부분 상대적 상호정보량 J, 로그승산비, 상호정보량과 같은 저빈도 자질 선호 기준인 것으로 나타났다. 고빈도 자질을 선호하는 정보획득량과 저빈도 자질을 선호하는 상호정보량을 비교해 보면, 전체의 10% 수준 자질집합으로 과도하게 축소한 경우를 제외한 모든 경우에서 상호정보량을 기준으로 추출한 자질집합의 성능이 더 좋았다. 이는 총 자질 종수를 기준으로 분석한 앞의 실험에서 상호정보량을 이용한 결과가 가



〈그림 8〉 축소된 색인파일의 분류성능 비교 (TREND-12746)

장 나빴던 것과는 전혀 다른 결과이다.

둘째, 고빈도 자질을 선호하는 선정 기준은 색인파일의 규모를 줄여나갈 수록 분류 성능이 뚜렷하게 저하되었다. 대표적으로 정보획득량은 KFCM-896 실험집단에서는 모든 경우에, TREND-12746 실험집단에서는 90% 규모를 제외한 모든 경우에 전체 자질을 사용한 것보다 성능이 나쁘게 나타났으며 GSS 계수도 비슷한 성능을 보였다. 상대적으로 중립에 가까운 카이제곱통계량은 색인파일 규모의 축소에 따른 분류성능의 저하가 덜하였다.

셋째, 저빈도 자질을 선호하는 선정 기준은 색인파일 규모의 축소에 따른 분류 성능의 저하가 크지 않았다. KFCM-896 실험집단에서 상호정보량은 전체의 30% 규모, 상대적 상호정보량 J는 40% 규모, 로그승산비는 50% 규모에 불가한 색인파일을 사용하고도 전체 자질을 모두 사용하는 경우보다 성능이 좋았다. TREND-12746 실험집단의 경우에 상호정보량과 로그승산비는 60% 규모, 상대적 상호정보량 J는 70% 규모의 색인파일을 사용하여 전체 자질을 모두 사용한 것보다 좋은 성능을 얻었다. 다만 상호정보량은 10% 내지 20% 로 매우 작은 규모의 색인파일만을 사용하는 경우에는 성능이 뚜렷하게 떨어지는 것으로 나타났다. 규모가 큰 TREND-12746 실험집단이 KFCM-896 실험집단에 비해서 자질축소의 효과가 적은 듯 보이는 이유는, KFCM-896 집단이 신문기사이므로 분류에 도움이 되지 않는 일반어휘가 많은 반면에 과학기술정보를 다루는 TREND-12746 실험집단에는 분류에 도움이 되는 주제어의 비율이 더 높기 때문이다.

넷째, 문헌빈도는 90%에서부터 10%에 이르기까지 모든 색인파일 규모 수준에서 가장 나쁜 분류 성능을 보인 반면에, 역문헌빈도는 자질 선정 기준으로서 정보획득량과 같은 고빈

도 자질 선호 기준을 적용한 경우와 대등한 성능을 보였다.

결론적으로 자질 선정이 저장 공간의 절약을 목표로 하는 경우라면 도치색인파일에 기반한 kNN 분류기에서는 고빈도 선호 기준보다 저빈도 선호 기준을 적용하는 것이 더 좋은 결과를 얻는 것으로 나타났다.

5. 2 실행시간 기준 분류 성능 비교

kNN 분류기는 기계학습 기반의 자동분류 알고리즘 중에서 분류속도가 가장 느린 편에 속하는 것으로 알려져 있다(Manning and Schütze 1999). 이런 단점을 극복하기 위해서 kNN 분류기의 문서분류속도를 향상시키기 위한 시도도 계속되고 있다(이재문 2002; 이재윤, 유수현 2003; Bell and Moffat 1996; Persin 1994; Zhou et al. 2003). 따라서 동일한 실행시간을 기준으로 각 자질 선정 기준의 분류 성능을 비교하는 것은 kNN 분류기의 실용화 측면에서 큰 의미가 있다.

이를 위해서는 동일한 실행시간을 기준으로 분류성능을 평가하는 것이 이상적이지만 실행시간을 정확히 일치시키는 것은 현실적으로 어렵다. 따라서 앞의 실험에서 색인파일의 규모를 일치시켜서 10%에서부터 90% 규모에 이르기까지 실행하였을 때의 실행시간을 측정하여 〈표 7〉과 〈표 8〉에 제시하였다.

실행시간 측정에서 드러난 사실은 다음과 같다.

첫째, 고빈도 자질을 선호하는 분류자질 선정 기준은 색인파일의 규모를 축소하더라도 실행시간의 단축 효과는 적게 나타났다. 정보획득량이나 GSS 계수의 경우에는 색인파일의 규모를 전체의 10%까지 줄이더라도 실행 시간은 KFCM-896 집단에서는 30%, TREND-12746 집단에서는 40%대로 감소되는데 그쳤다. 색인파일의 규모를 50%까지만 줄이는 경우에는 실행시간은 90% 내외로 소요되어 단축효과가 미미하였다.

둘째, 저빈도 자질을 선호하는 분류자질 선정 기준은 색인파일의 규모를 축소함에 따라서 실행시간의 단축 효과가 큰 것으로 나타났다.

〈표 8〉 축소된 색인파일의 분류 실행시간 비교 (KFCM-896)

색인파일규모	자질 선정 기준							
	DF	IG	GSS	CHI	RMIJ	LOR	MI	IDF
10%	45.6%	38.3%	33.9%	18.3%	13.6%	8.1%	8.0%	6.7%
20%	64.7%	57.3%	56.6%	25.7%	17.8%	12.6%	9.2%	7.5%
30%	77.5%	69.3%	68.2%	41.1%	23.5%	16.2%	12.2%	8.8%
40%	85.0%	78.5%	77.9%	50.7%	28.3%	22.4%	16.4%	11.0%
50%	90.1%	84.8%	86.3%	60.4%	36.9%	29.1%	22.4%	14.6%
60%	93.7%	89.3%	90.6%	69.8%	43.0%	37.8%	30.4%	19.8%
70%	95.8%	92.5%	93.5%	76.0%	53.7%	46.5%	39.4%	27.5%
80%	97.1%	95.0%	95.9%	86.1%	63.5%	57.3%	51.7%	40.3%
90%	97.8%	95.7%	97.6%	93.2%	75.9%	75.8%	65.5%	59.6%
100%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

* 축소하지 않은 색인파일을 사용한 경우의 실행시간을 100%로 하였음

〈표 9〉 축소된 색인파일의 분류 실행시간 비교 (TREND-12746)

색인파일규모	자질 선정 기준							
	DF	IG	GSS	CHI	RMIJ	LOR	MI	IDF
10%	47.8%	42.2%	44.9%	14.3%	5.4%	1.7%	1.4%	0.9%
20%	68.2%	64.5%	67.8%	33.0%	10.3%	3.8%	2.4%	1.2%
30%	80.9%	77.1%	78.1%	52.1%	15.9%	6.4%	4.5%	1.8%
40%	88.3%	86.0%	82.2%	63.3%	20.4%	12.2%	8.0%	3.0%
50%	92.9%	91.6%	89.5%	73.0%	25.8%	18.4%	11.6%	5.5%
60%	95.6%	94.8%	93.2%	83.2%	33.4%	25.6%	17.9%	10.1%
70%	97.0%	96.7%	95.8%	91.4%	45.1%	41.5%	28.1%	17.7%
80%	97.6%	97.1%	97.5%	96.6%	57.4%	57.5%	39.7%	36.7%
90%	97.9%	97.5%	97.8%	98.8%	74.9%	92.3%	61.0%	51.0%
100%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%

* 축소하지 않은 색인파일을 사용한 경우의 실행시간을 100%로 하였음

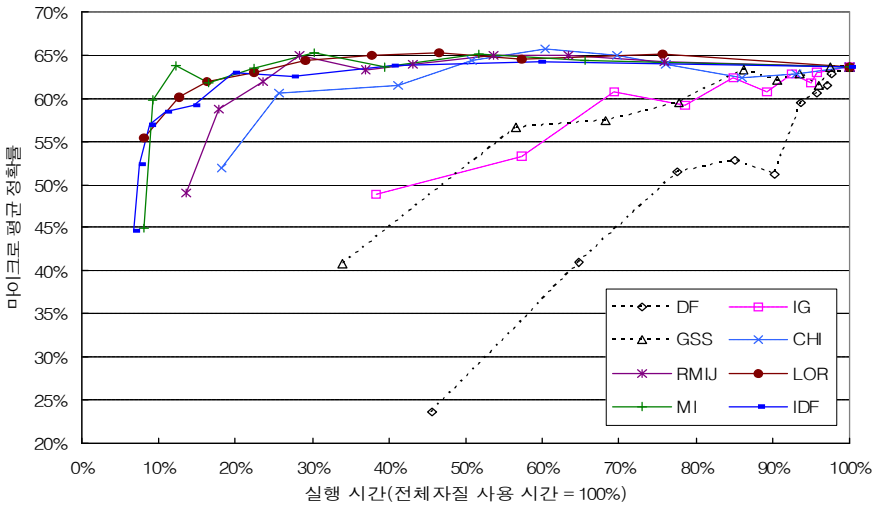
상호정보량이나 로그승산비와 같은 척도를 자질 선정 기준으로 사용하면 TREND-12746 집단의 경우에는 실행시간이 2% 이하로 대폭 단축되었다. 색인파일의 규모를 50% 까지만 줄이는 경우에도 실행시간은 20%대 내지 10%대로 크게 감소하였다.

상호정보량과 같이 저빈도 자질을 선호하는 자질 선정 기준을 적용하였을 때 실행시간 단축의 효과가 큰 이유는 kNN분류기에서 비교횟수의 감소효과가 크게 나타나기 때문이다. 분류대상 문서에 문헌빈도 500인 자질이 포함되어 있으면 500개의 학습문서가 비교대상에 포함되는데 반해서, 문헌빈도 5인 자질은 5개의 학습문서하고만 비교하게 된다. 실제로는 분류대상 문서에 복수의 자질이 포함될 것이고 고빈도 색인어들이 출현한 문서들은 상당수 겹치기 마련이므로 문헌빈도와 비교횟수가 단순 정비례하는 것은 아니다.

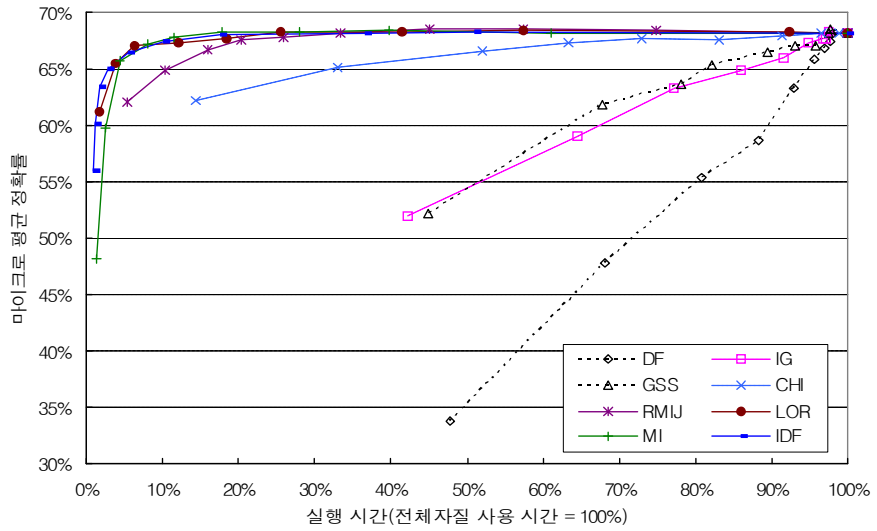
실행시간과 분류성능을 함께 살펴보기 위해서 실행시간을 가로축으로, 분류성능을 세로축으로 하는 〈그림 9〉와 〈그림 10〉을 작성하였다. 이 그림들은 〈그림 7〉과 〈그림 8〉에서 가로축을 색인파일의 규모로 했던 것을 실행시간으로 바꾼 것만 다르다. 이중에서 TREND-12746을 대상으로 한 〈그림 10〉의 일부분을 〈그림 11〉로 확대하였다.

두 실험문서집단 중에서 규모가 매우 작은 KFCM-896의 경우는 실행시간 단축이 큰 의미가 없으므로 규모가 큰 TREND-12746 실험집단을 다룬 〈그림 11〉을 중심으로 결과를 분석하면 다음과 같다.

첫째, 고빈도 자질을 선호하는 분류자질 선정 기준은 실행시간을 단축시키기 위해서는 분류 성능의 심각한 저하를 피할 수 없는 것으로 나타났다. 예를 들어 정보획득량을 적용해서 실행시간을 80% 이하로 단축시킨 결과 분류성능은 68%대에서 63%대로 5% 포인트 가량 저하되었다. 고빈도 자질 선호경향이 덜한 카이제곱통계량은 실행시간을 70% 정도로 단축시켰을때 분류성능의 저하는 1% 포인트 이내로 나타나서 양호한 편이었다. 그러나 이보다 더 시간을 단축시키면 분류성능의 저하가 뚜



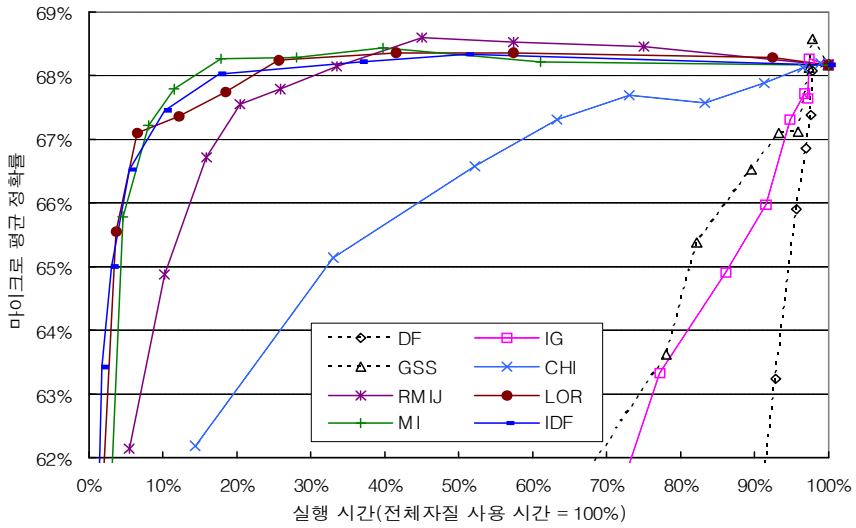
〈그림 9〉 축소된 자질집합의 실행 시간 비교 (KFCM-896)



〈그림 10〉 축소된 자질집합의 실행 시간 비교 (TREND-12746)

렷해졌다. 문헌빈도의 경우에는 실행시간을 조금이라도 단축시키면 분류성능이 급격하게 저하되므로, 처리시간을 고려한다면 도치색인파일 기반의 kNN 분류기를 위한 자질 선정 기준으로는 맞지 않다.

둘째, 저빈도 자질을 선호하는 분류자질 선정 기준은 실행시간이 크게 단축되더라도 분류성능의 저하가 거의 나타나지 않았다. 상호정



〈그림 11〉 축소된 자질집합의 실행 시간 비교 (TREND-12746) - 부분 확대

보량은 10%대로, 로그승산비는 20%대로, 상대적 상호정보량 J의 경우에는 30%내지 40%대로 각각 실행시간이 단축되는 경우에도 분류성능은 오히려 약간 향상되었다.

셋째, 역문헌빈도를 자질 선정 기준으로 사용하였을 때에도 분류성능의 저하 없이 실행시간을 30%대까지 감소시킬 수 있는 것으로 나타났다. 실행시간이 30% 이하로 단축되는 경우에도 역문헌빈도를 이용한 경우의 성능이 다른 저빈도 자질 선호 기준을 이용한 경우와 별 차이가 없었다.

넷째, 전체적으로 실행시간이 50%, 즉 절반 정도로 단축되는 경우에는 상대적 상호정보량 J를 이용한 경우의 성능이 가장 좋고, 40%에서 10%대까지 실행시간이 더 단축되는 경우에는 상호정보량을 이용한 경우의 성능이 가장 좋게 나타났다. 10% 이하까지로 실행시간이 크게 단축되면 어느 선정 기준을 이용하더라도 분류성능의 저하가 1% 포인트를 넘어서면서 뚜렷해졌다.

분석 결과 동일한 실행시간을 기준으로 한다면 분류 성능은 저빈도 자질 선호 기준이 고빈도 자질 선호 기준에 비해서 월등한 것으로 나타났다. 저빈도 자질 선호 기준을 적용하여 분류 성능의 저하 없이 얻을 수 있는 최단 실행시간은 대개 원래의 20% 전후까지였다. 이는 처리속도로 따지면 5배 내외로 빠른 분류가 가능해지는 셈이다. 빠른 처리속도를 분류자질 선정의 목표로 삼는 경우라면 도치색인파일 기반 kNN 분류기에서는 반드시 저빈도 선호 기준을 채택해야 할 것이다.

이상의 결과에 의하면 가중치 할당에 적용되어온 역문헌빈도가 자질 선정 기준으로 사용되더라도 매우 양호한 분류성능을 얻을 수 있다는 사실이 드러났다. 색인레코드의 저장공간을 기준으로 비교했을 때에는 정보획득량과 같은

고빈도 자질 선호 기준과 대등한 성능을 보였으며, 분류 실행시간을 기준으로 비교한 결과에서는 상호정보량이나 로그승산비와 같이 뛰어난 자질 선정 기준과도 견줄만한 성능이 얻어졌다. 다음 장에서는 역으로 자질 선정 기준으로 사용되어온 척도를 자질 가중치 할당에 사용할 수 있는지 여부를 살펴보고자 한다.

6. 자질 선정 기준의 자질 가중치로의 활용

앞의 실험에서 전통적으로 자질 가중치로 사용되어온 역문헌빈도가 자질 선정 기준으로서도 도치색인파일 기반 kNN 분류기에서는 무난한 것으로 나타났다. 즉 분류자질 선정 단계와 가중치 할당 단계의 전략을 상반되지 않게 채택할 여지가 있는 것이다. 이를 다른 측면에서 검토하기 위하여 상호정보량을 비롯한 자질 선정 기준값을 그대로 가중치로 사용하는 실험을 수행하였다. 이를 통해서 만약 좋은 자질 선정 기준이 좋은 가중치가 될 수도 있는 것으로 나타난다면, 자질 선정과 가중치 할당 단계에서 좋은 자질을 판단하는 기준을 일관되게 사용할 수 있을 것이다.

자질 선정에 사용되던 각 자질의 값을 가중치로 사용하려면 자질값의 범위 때문에 두 가지 사항을 고려해야 한다. 첫째로는 자질값의 음수값 문제가 있다. 자질 선정 시에는 자질의 순서를 정하는 것이므로 값의 우열이 중요한 반면에, 가중치로 사용할 때에는 값의 대소만이 아닌 값 자체가 이용된다. 상호정보량이나 로그승산비와 같은 척도는 0 미만의 음수가 있을 수 있으므로 가중치로 사용할 때에는 실제 값의 범위를 확인해야 한다. 이 연구에서는 전역적 자질 선정 방식을 적용하면서 각 범주와의 자질 값 중에서 가장 큰 값을 채택하므로 다행히 모든 자질은 0보다 큰 양수가 된다. 둘째로 고려할 사항은 자질값의 상한이다. 대부분의 연관성 척도는 상한이 1이지만, 로그승산비와 같은 척도는 수십, 카이제곱은 수백 이상의 값을 가질 수도 있다. 따라서 자질 가중치의 순위 차이에 비해서 값의 차이가 지나치게 큰 것이 문제가 될 수 있다. 마침 로그승산비는 이와 동일한 자질 순위를 산출하면서 값의 상한이 1까지로 제한되는 율의 Y 척도가 있으므로 이를 자질값 가중치 실험에 추가로 적용해 보았다.

이번 실험에서는 분류자질 선정을 하지 않고 모든 자질을 사용하되, 자질 가중치로 역문헌빈도 대신 다른 분류자질 선정 기준에 의한 값을 사용해서 문서를 표현한 다음 kNN 분류기에 의한 분류 성능을 비교하였다. 실험 결과는 〈표 9〉와 〈표 10〉에 제시하였으며 역문헌빈도를 사용한 경우의 성능과 대비한 결과를 별도로 〈그림 12〉와 〈그림 13〉에 제시하였다.

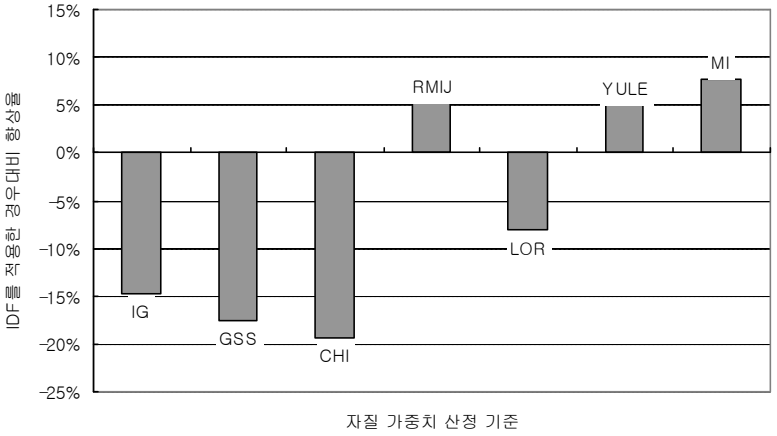
분류자질 선정 기준을 자질 가중치로 사용하는 실험 결과 드러난 사실은 다음과 같다.

첫째, 고빈도 자질을 선호하는 기준에 비해서 저빈도 자질을 선호하는 기준이 산출한 값을 가중치로 사용하는 경우에 더 좋은 분류성능을 얻었다. 이는 앞에서 저빈도 자질을 선호하는 선정기준이 저장공간과 처리시간 면에서 더 좋은 결과를 보였던 것과 일치한다.

둘째, 저빈도 자질을 선호하는 기준이 산출한 값을 가중치로 사용한 결과, 역문헌빈도를 가중치로 사용하는 경우보다 KFCM-896 실험

〈표 10〉 분류자질 선정 기준값을 가중치로 적용한 결과 (KFCM-896)

	자질 가중치 산정 기준							
	IG	GSS	CHI	RMIJ	LOR	YULE	MI	IDF
마이크로 평균 정확률	0.5425	0.5249	0.5136	0.6693	0.5851	0.6685	0.6854	0.6364
IDF를 적용한 경우대비 향상율	-14.75%	-17.53%	-19.29%	5.17%	-8.07%	5.04%	7.69%	—



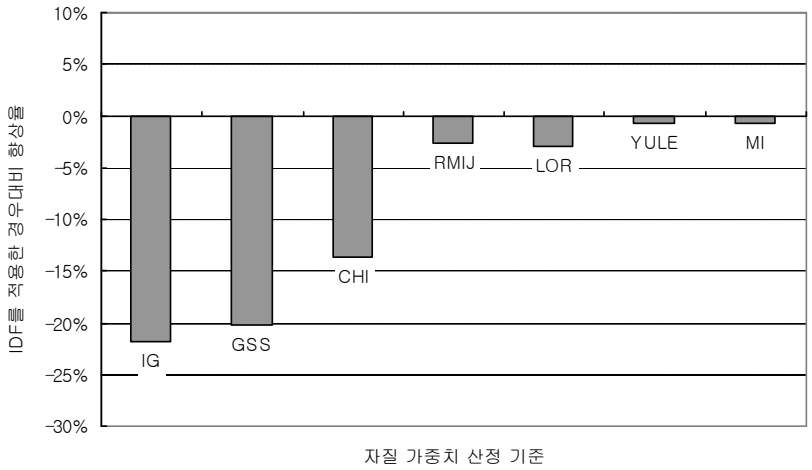
〈그림 12〉 분류자질 선정 기준값을 가중치로 적용한 결과(KFCM-896)

〈표 11〉 분류자질 선정 기준값을 가중치로 적용한 결과 (TREND-12746)

	자질 가중치 산정 기준							
	IG	GSS	CHI	RMIJ	LOR	YULE	MI	IDF
마이크로 평균 정확률	0.5329	0.5435	0.5880	0.6634	0.6619	0.6773	0.6772	0.6816
IDF를 적용한 경우대비 향상율	-21.82%	-20.26%	-13.73%	-2.68%	-2.90%	-0.64%	-0.65%	—

집단에서는 더 좋은 결과를, TREND-12746 실험집단에서는 약간 낮은 결과를 얻었다. 상호정보량의 경우에는 KFCM-896 실험집단에서 역문헌빈도를 가중치로 사용한 것보다 분류성능이 7.69% 향상되었고, TREND-12746 실험집단에서는 약간 성능이 저하되긴 하였으나 그 비율은 -0.65%에 그쳤다.

결론적으로 분류자질 선정 기준으로서 좋은 성능을 보였던 상호정보량과 같은 척도값은 자질 가중치로 사용해도 역문헌빈도에 견줄만한 성능을 보였다. 역문헌빈도가 분류자질 선정 기준으로서 나쁘지 않았던 점을 함께 감안하면 이는 분류자질 선정 기준과 가중치 할당 방식이 kNN 분류기에서는 별개의 문제가 아님을 뜻한다.



〈그림 13〉 분류자질 선정 기준값을 가중치로 적용한 결과 (TREND-12746)

7. 결 론

이상의 실험에서 상호정보량과 같은 저빈도 선호 기준이 자동분류를 위한 자질 선정 기법으로 적합하지 않다는 Yang과 Pederson(1997)의 주장이 도치색인파일 기반 kNN 분류기에서는 사실이 아닌 것으로 나타났다. kNN 분류기는 도치색인파일로 구현하는 것이 효율적임을 감안할 때 기존의 관례에서처럼 고빈도 자질을 선호하는 문헌빈도나 정보획득량을 기준으로 분류자질을 선정하는 것은 바람직하지 않다.

저장공간의 절약을 자질 선정의 목적으로 보아서 동일한 색인파일의 규모를 기준으로 분류 성능을 비교하였을 때, 저빈도 선호 기준이 고빈도 선호 기준에 비해서 대등하거나 더 나은 성능을 보였다. 실행시간의 절약을 자질 선정의 목적으로 보고 실행시간의 감소에 따른 분류 성능의 추이를 분석한 결과, 고빈도 선호 기준은 실행 시간을 조금만 감소시키려고 해도 분류 성능의 심각한 저하를 피할 수 없는 것으로 나타났다. TREND-12746 실험집단에서 정보획득량으로 자질축소를 하면서 처리시간을 원래의 80%로 줄이기 위해서는 분류정확률이 5% 포인트 이상 저하되었다. 반면에 저빈도 선호 기준은 실행 시간을 원래의 30% 내지 20% 내외로 크게 감소시키면서도 성능이 떨어지지 않는 것으로 나타났다. 즉, 성능 저하 없이 처리 속도가 3배에서 5배까지 향상된 것이다. 이재문(2002)의 연구에서 다양한 휴리스틱을 적용하고서도 kNN 분류기의 문서 분류 속도를 약 1.4배 정도 향상시키는 것에 그쳤던 것과 비교하면 저빈도 자질 선호 기준을 이용한 분류자질의 선정은 매우 중요한 의미를 지닌다.

역문헌빈도를 분류자질 선정 단계의 기준으로 사용하는 실험과, 그 반대로 상호정보량과 같은 분류자질 선정 기준을 가중치 할당 단계

에서 가중치로 삼는 실험을 통해서는 두 단계가 밀접하게 관련되어 있다는 결과를 얻었다. 즉, 분류자질 선정 기준으로서 좋은 척도는 자질 가중치로 사용하더라도 좋은 성능을 보인 것이다. 따라서 kNN분류기를 이용하면서 효과적인 분류정확률과 함께 효율적인 처리시간 및 저장공간을 동시에 보장하기 위해서는 일관되게 저빈도 선호 척도를 기준으로 하여 자질 선정과 가중치 할당에 적용해야 할 것이다.

일반적으로 문서를 분류해서 저장한 다음에는 검색의 대상으로 삼는다는 점을 감안하면 자질 선정 기준의 비교를 위해서는 저장공간보다는 실행속도가 더 중요한 고려사항이라고 할 수 있다. 어차피 검색을 위해서는 학습문서의 모든 자질을 색인해서 저장해두어야 하기 때문이다. 따라서 학습문서에는 자질 선정을 적용하지 않고 분류대상 문서를 표현할 때에만 색인어를 선택적으로 이용하는 방식도 현실적으로 유용하리라고 본다. 이에 대해서는 후속 연구에서 살펴볼 계획이다.

참 고 문 헌

김제욱, 김한준, 이상구. 2002. "베이지언 문서분류시스템을 위한 능동적 학습기반의 학습문서집합 구성방법." 『정보과학회 논문지 : 소프트웨어 및 응용』, 29(11/12): 966-978.

박부영. 2004. 『잠재의미색인(LSI) 기법을 이용한 kNN 분류기의 자질 선정에 관한 연구』. 연세대학교 석사학위논문.

이재문. 2002. "휴리스틱을 이용한 kNN의 효율성 개선." 『한국정보처리학회 논문지 B』, 10(6): 719-724.

이재윤, 유수현. 2003. "대표용어를 이용한 kNN 분류기의 처리속도 개선." 『제10회 한국정보관리학회 학술대회 논문집』, pp.65-72.

Alpaydin, Ethem. 2004. *Introduction to Machine Learning.* MIT Press.

Bell, T. A. H., and A. Moffat. 1996. "The design of a high performance information filtering system." *Proceedings of the 19th Annual ACM Conference on Research and Development in Information Retrieval*, pp. 12-20.

Blumberg, Robert, and Shaku Atre. 2003. "Automatic classification: moving to the mainstream." *DM Review Magazine*, (April, 2003). [cited 2005. 3.10] 〈view.com/article_sub.cfm?articleId=6501〉

Cöster, Rickard, and Martin Svensson. 2002. "Inverted file search algorithms for collaborative filtering." *Proceedings of the 25th Annual ACM Conference on Research and Development in Information Retrieval*, pp.246-252.

Forman, George. 2002. "An extensive empirical study of feature selection metrics for text classification." *Journal of Machine Learning Research*, 3: 1289-1305.

Fragoudis, D., D. Meretakis, and S. Likothanassis. 2005. "Best terms: an efficient feature-selection algorithm for text categorization". *Knowledge and Information Systems*, 8(1): 16-33.

Galavotti, L., F. Sebastiani, and M. Simi. 2000. "Experiments on the use of feature selection and negative evidence in automated text categorization." *Proceedings of the 4th European Conference on Research and Advanced Technology for Digital Libraries*, 59-68.

Lim, Heui-Seok. 2002. "An improved kNN learning based korean text classifier with heuristic information." *Proceedings of the 9th International Conference on Neural Information Processing(ICONIP'02)*, Vol.2, pp.731-734.

Manning, C. D., and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge, Mass.: MIT Press.

Persin, M. 1994. "Document filtering for fast ranking." *Proceedings of the 17th Annual ACM Conference on Research and Development in Information Retrieval*, pp. 341 - 348.

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

Yang, Y., and J. P. Pederson. 1997. "A comparative study on feature selection in text categorization." *Proceedings of the Fourteenth International Conference on Machine Learning*, 412-420.

Yang, Y., and Xin Liu. 1999. "A re-examination of text categorizationi methods." *Proceedings of the 22nd Annual ACM Conference on Research and Development in Information Retrieval*, pp.42-49.

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

Zhou, S., T. W. Ling, J. Guan, J. Hu, and A. Zhou. 2003. "Fast text classification: A training-corpus pruning based approach." *Proceedings of the 8th International Conference on Database Systems for Advanced Applications (DASFAA'2003)*, pp. 127-136.

Zu, G., W. Ohyama, T. Wakabayashi, and F. Kimura. 2003. "Accuracy improvement of automatic text

classification based on feature transformation." *Proceedings of the 2003 ACM Symposium on Document Engineering*, pp.118-120.