문헌간 유사도를 이용한 SVM 분류기의 문헌분류성능 향상에 관한 연구

Improving the Performance of SVM Text Categorization with Inter-document Similarities

이 재 윤 (Jae-Yun Lee) *

초 록

이 논문의 목적은 SVM(지지벡터기계) 분류기의 성능을 문헌간 유사도를 이용해서 향상시키는 것이다. SVM은 효과적인 기계학습 시스템으로서 최고 수준의 문헌자동분류 기술로 인정받고 있다. 이 연구에서는 문헌 벡터 자질 표현에 기반한 SVM 문헌자동분류를 제안하였다. 제안한 방식은 분류 자질로 색인어 대신 문헌 벡터를, 자질값으로 가중치 대신 벡터유사도를 사용한다. 제안한 방식에 대한 실험 결과, SVM 분류기의 성능을 향상시킬 수 있었다. 실행 효율 향상을 위해서 문헌 벡터 자질 선정 방안과 범주 센트로이드 벡터를 사용하는 방안을 제안하였다. 실험 결과 소규모의 벡터 자질 집합만으로도 색인어 자질을 사용하는 기존 방식보다 나은 성능을 얻을 수 있었다.

ABSTRACT

The purpose of this paper is to explore the ways to improve the performance of SVM(Support Vector Machines) text classifier using inter-document similarities. SVMs are powerful machine learning systems, which are considered as the state-of-the-art technique for automatic document classification. In this paper text categorization via SVMs approach based on feature representation with document vectors is suggested. In this approach, document vectors instead of index terms are used as features, and vector similarities instead of term weights are used as feature values. Experiments show that SVM classifier with document vector features can improve the document classification performance. For the sake of run-time efficiency, two methods are developed: One is to select document vector features, and the other is to use category centroid vector features instead. Experiments on these two methods show that we can get improved performance with small vector feature set than the performance of conventional methods with index term features.

키워드: 문헌자동분류, 문서범주화, SVM 분류기, 분류자질, 문헌유사도
automatic document classification, text categorization, SVM classifier, classification features, document similarity

---

* 경기대학교 문헌정보학과(memexlee@kgu.ac.kr)
■ 논문접수일자 : 2005년 8월 20일
■ 게재확정일자 : 2005년 9월 6일

1. 서 론

문헌자동분류 또는 문서범주화는 주어진 문헌의 내용에 근거해서 소속 범주를 지정하는 기법이다. 이에 대한 접근은 정보검색 분야만큼 오래 전부터 있었지만, 1990년대에 기계학습 이론이 본격적으로 적용되기 시작하면서 연구가 활성화 되었다고 할 수 있다. 로치오 분류기, 나이브베이즈 분류기, 의사결정트리, 신경망, kNN 분류기 등의 다양한 기계학습 방식을 문헌분류 문제에 적용한 여러 연구가 발표되었다. 그러나 1990년대 후반에 지지벡터기계(Support Vector Machines; SVM으로 약칭)가 문헌분류 문제에 도입된 이후에는 상황이 달라지게 되었다.

V. Vapnik이 개발한 SVM은 1979년에 제안되었으나 한동안 주목을 받지 못하다가, 1990년대 중반에 와서 뒤늦게 뛰어난 성능과 안정성이 입증되어 여러 기계학습 알고리즘 중에서 가장 선호되는 방식중 하나가 되었다(Vapnik 1995). 문헌분류 영역에서도 이를 처음으로 적용한 Joachims(1998) 이후 SVM 문헌 자동분류에 대한 많은 논문이 발표되었다(정영미, 임혜영 2000; Caldas & Soibelman 2003; Dumais et al. 1998; Hirotoshi & Masahiko 1999; Yang & Liu 1999). 나아가서 정보검색에서 적합/부적합 문헌을 학습하는 용도로 적합성피드백 검색에 응용한 연구(Drucker et al. 2002)도 발표되는 등 SVM 분류기의 응용영역도 넓어지고 있다.

현재는 SVM 분류기가 문헌자동분류 문제에 대한 최고수준 기술(state-of-the-art)로 인정받고 있다(Cristianini & Shawe-Taylor 2000). 이에 따라 문헌자동분류에 대한 최근 연구는 어떻게 하면 다른 분류기를 써서 SVM 분류기 이상의 성능을 얻을 수 있는지, SVM 분류기의 성능이나 효율을 향상시킬 수 있는 방안은 무엇인지를 과제로 삼은 경우가 많다.

이 연구에서는 SVM 분류기의 문헌분류성능 향상을 위해서 문헌을 표현하는 자질(feature)을 기존과 다른 방식으로 처리하고자 한다. 목적이 검색이든 분류든 간에 전통적으로 문헌을 표현하는 자질로는 색인어가 사용되어왔다. 이 연구에서 검토하는 방식은 색인어가 아니라 문헌벡터를 자질로 하고, 색인어 가중치가 아닌 문헌간 벡터유사도를 자질값으로 사용하여 문헌을 표현하는 것이다. 즉, 다른 문헌과의 유사한 정도를 통해서 한 문헌을 표현하려는 시도이다. 모든 문헌간 벡터유사도를 구하는 것이 실용적인 측면에서 불리할 수가 있으므로, 이를 보완할 수 있는 방법도 함께 제안하고 실험을 통해 검증하였다.

2. SVM 분류기와 분류자질

2. 1 SVM 분류기와 자질 선정

SVM 분류기는 구조적 위험 최소화 원리에 근거한 것으로서 부정예제로부터 긍정예제를 분리해낼 수 있는 결정면을 찾아내는 알고리즘이다(Vapnik 1995). SVM이 아닌 다른 분류기, 예를 들어 나이브베이즈 분류기나 kNN 분류기를 문헌자동분류에 적용할 때에는 분류자질 선정 등의 방법으로 분류성능과 실행효율을 모두 향상시킬 수 있다. 반면에 SVM 분류기의 분류성능을 자질 선정을 통해 향상시키는 것은 상당히 어렵다고 알려져 있다. SVM 분류기를 이용한 문헌자동분류에서 자질 선정실험을 수행한 주요 사례

를 살펴보면 다음과 같다.

Yang과 Liu(1999)는 SVM 분류기와 다른 네 가지 분류기의 성능을 비교하면서 다른 분류기는 자질을 1천 내지 2천개 정도로 축소했을 때를 가장 좋은 성능으로 제시한 반면에 SVM분류기만은 자질을 1만개 사용했을 때를 최고 성능으로 제시하였다.

Taira와 Haruno(1999)의 실험에서는 비록 각 범주별로는 자질 선정을 통해서 성능이 향상되거나 저하되는 경우가 있었지만, 평균적인 전체 성능은 자질 선정을 하지 않는 경우가 SVM 분류기에서는 가장 좋았다고 보고하였다.

Rogati와 Yang(2002)은 나이브베이즈 분류기, 로치오 분류기, kNN 분류기, SVM 분류기의 네 가지 시스템을 사용해서 다양한 자질 선정 기법의 조합 적용을 시험해보았다. 이들은 자질 축소 결과를 25% 이내인 경우에만 보고하여 전체를 확인할 수는 없지만, 나머지 세 분류기는 자질 축소를 10% 내지 3%까지로 많이 한 경우에 좋은 성능을 얻었다. 그러나 SVM 분류기에 대해서만은 어떤 자질 선정 기법을 조합하더라도 보고한 범위 내에서는 자질축소를 덜하여 최대한 많은 자질을 사용한 경우가 성능이 좋은 것으로 나타났다.

Caldas와 Soibelman(2003)은 SVM 분류기를 이용해서 계층적 문헌분류 실험을 수행하였는데, 역시 자질 선정을 하지 않은 경우가 자질 선정을 한 경우에 비해서 약간 좋은 것으로 나타나서 자질 선정으로는 성능향상 효과를 얻지 못했다고 보고하였다.

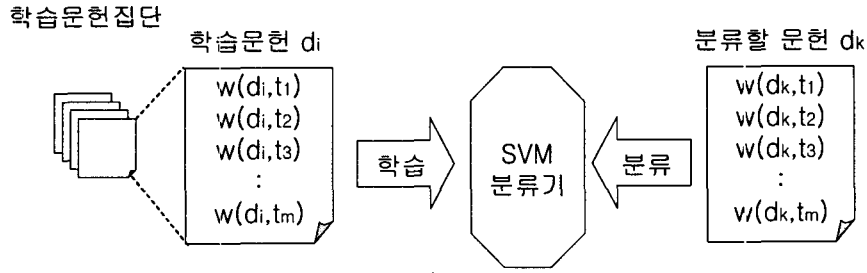
이와 같이 SVM 분류기로 문헌자동분류를 할 경우에 자질 선정을 통해 성능향상을 얻기 어려운 이유는 SVM 분류기 자체의 성능이 워낙 좋기 때문이기도 하지만, 사용된 자질 중에서 유용하지 않은 것의 영향을 어느 정도는 줄여서 학습하는 특성을 SVM 분류기가 가지고 있기 때문이다. 문헌자동분류에 처음으로 SVM을 적용한 Joachims(1998)도, SVM은 고차원 데이터를 일반화하는 능력이 뛰어나기 때문에 문헌집단과 같이 자질의 종류는 많고, 대부분의 자질이 나름대로 쓸모가 있으며, 개별 자료에는 한정된 수의 자질만 출현하는 대상을 분류할 때에는 자질 선정의 필요성을 없애준다고 지적하였다.

2. 2 벡터 자질 표현

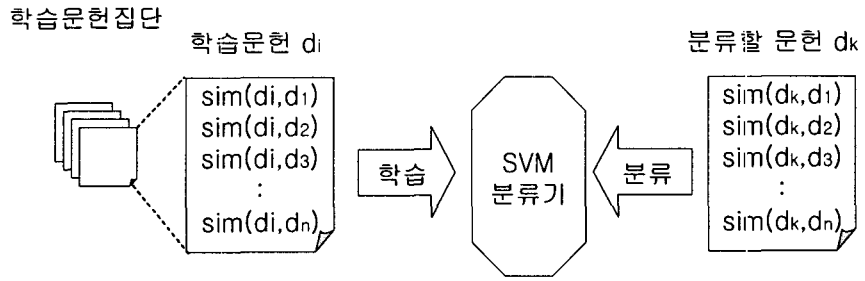
이 연구에서는 SVM 문헌분류를 위한 새로운 자질표현방법으로 벡터 자질 표현 방식을 제안하고 이의 효율화를 위한 자질 선정 방법을 개발하였다.

벡터 자질 표현 방식의 핵심은 벡터유사도이다. 이 방식에서는 각 문헌(학습문헌, 분류대상문헌 모두)을 학습문헌들과의 벡터유사도로 표현한다. 여기서 사용하는 벡터유사도는 코사인유사도이다(Salton & McGill 1983). 기존의 색인어 자질을 이용한 문헌 표현에서는 색인어가 자질, 색인어 가중치가 자질값으로 사용되었다. 제안하는 방식에서는 자질은 벡터, 자질값은 벡터유사도를 사용한다. 즉, 기존 방식에서는 분류기에 입력되는 자료가 문헌-용어 행렬이었는데 반해서 제안하는 방식에서는 문헌-문헌 유사도 행렬이 된다.

〈그림 1〉에 두 표현 방식을 그림으로 비교하였다. 색인어 자질로 문헌을 표현하는 경우에 자질값 $w(d_i,t_m)$은 문헌 d_i에서 색인어 t_m의 가중치를 뜻한다. 벡터 자질로 문헌을 표현하는 경우에



(a) 자동분류에서 색인어 자질을 이용한 문헌 표현



(b) 자동분류에서 벡터 자질을 이용한 문헌 표현

〈그림 1〉 색인어 자질과 벡터 자질의 비교

자질값 $sim(d_k,d_n)$은 문헌 d_k와 학습문헌 d_n과의 벡터유사도를 뜻한다.

학습문헌을 벡터 자질로 표현하기 위해서는 일단 먼저 색인어 자질로 표현한 다음 문헌간의 벡터유사도를 구하는 과정을 거쳐야 한다. 이 과정에서 학습시간이 상당히 소요되기는 하지만, 일반적으로 학습문헌의 수가 색인어 종수보다 상당히 적으므로 자질 차원이 축소되는 효과를 얻는다. 만약 학습문헌이 색인어 종수에 비해서 그리 적지 않거나 오히려 많을 경우에는 학습문헌 벡터 중에서 일부만 선정하여 벡터 자질로 사용하면 된다.

벡터 자질의 축소 필요성은 학습할 때보다 분류할 때 더 크다. 분류대상 문헌을 벡터 자질로 표현하기 위해서는 kNN분류기처럼 모든 벡터 자질과의 유사도를 구해야 한다. 그 결과 분류속도가 상당히 느리다고 알려진 kNN 분류기보다도 더 느린 분류기가 된다. 결국 벡터 자질 표현이 아무리 좋은 분류 성능을 보이더라도 실용적인 측면에서는 속도 향상을 위한 방안이 추가로 필요하게 된다.

이 연구에서는 벡터 자질 표현의 속도 향상을 위해서 두 가지 방식으로 접근하였다. 첫째는 자질 선정이다. 색인어 자질과 마찬가지로 벡터 자질도 분류에 매우 도움이 되는 문헌 벡터가 있는가 하면, 그 정도가 덜하거나 오히려 도움이 못되는 것도 있을 수 있다. 일정한 척도를 사용해서 분류에 도움이 되는 벡터 자질을 선정하는 방안을 제시하고 실험을 통해 성능을 검증하였다.

두 번째 속도 향상 방안은 범주 센트로이드 벡

터를 자질로 사용하는 방안이다. 범주 센트로이드 벡터는 범주에 속한 문헌 벡터의 평균 벡터이다. 색인어나 문헌의 수가 아무리 많더라도 분류범주의 수는 상당히 한정된 경우가 많다. 대개는 분류범주가 몇 개에서 몇 십 개 정도이다. 따라서 표현력은 감소되더라도 적은 수의 범주 센트로이드 벡터를 벡터 자질로 삼고, 이 센트로이드 벡터 자질과의 유사도를 자질값으로 삼으면 분류실행 속도는 별 문제가 되지 않는다. 다만, 몇 개에 불과한 센트로이드 벡터를 자질로 사용하였을 때의 성능이 어느 정도인지가 관건이 된다. 범주 센트로이드 벡터를 자질로 사용하는 실험을 통해 성능을 검증해보았다.

3. 실험 설계

3. 1 실험문헌집단과 소프트웨어

이 연구에서는 〈표 1〉와 같이 두 가지 분류실험용 문헌집단을 이용하였다.

KFCM-896 분류실험집단은 KFCM-CL 1020(정영미, 이재윤 2001) 실험집단에서 주요 대분류 항목인 정치, 경제, 산업, 국제분야에 속한 기사 896건만 추출한 것이다. 각 범주별로 게재 시기가 늦은 기사 20%(178건)를 검증집단(분류기의 성능 실험을 위한 분류대상 문헌집단)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 실험에서는 분류의 깊이를 두 번째 중분류 수준까지 적용한 결과 17개 범주로 구성되었다.

TREND-2287 분류실험집단은 정보검색용 실험집단인 HANTEC v.2.0(김지영 외 2000)에서 분류정보가 포함된 해외과학기술문헌속보 문헌 2,287건을 추출한 것이다. 2,287건 중에서 1997년 4/4분기 3개월간 등록된 문헌 1,178건을 학습문헌집단으로, 1998년 1/4분기 3개월간 등

〈표 1〉 실험에 사용된 문헌 집단

실험문헌집단	KFCM-896	TREND-2287
내용	신문기사	해외과학기술문헌속보
문헌의 수 [전체/학습/검증]	[896/718/178]	[2,287/1,178/1,109]
범주의 수	17	8
범주별 학습문헌 수 [평균/최대/최소]	[42.2/81/17]	[147.3/424/27]
(저빈도어 제거 후) 학습문헌집단의 색인어 종수	7,261	7,544
(저빈도어 제거 후) 학습문헌의 평균 색인어 종수	88.3	97.9
(저빈도어 제거 후) 학습문헌의 평균 색인어 수	151.4	161.8

록된 문헌 1,109건을 검증문헌집단으로 이용하였다. 문헌의 분류는 현재 해외과학기술동향 홈페이지(http://techtrend.kisti.re.kr/)에서 구분한 9개 대분류 중 '과학기술 일반'을 제외한 8개 대분류를 적용하였다. 원래 해외과학기술문헌속보 문헌에는 KISTI(작성 당시에는 KORDIC)에서 부여한 분류번호가 18개 범주로 나누어져 있으나, 이 실험에서는 현재 서비스에서 규정된 8개 대분류 범주별로 18개 중에서 한 분야씩 선정하고 소속 문헌을 추출하였다.

각 문헌은 제목과 본문을 대상으로 자동색인하였고 규모가 작은 KFCM-896은 추출된 색인어 중에서 CF가 2 이하인 경우를, 이보다 규모가 큰 TREND-2287은 DF가 2 이하인 경우를 전처리 단계에서 제거하였다.

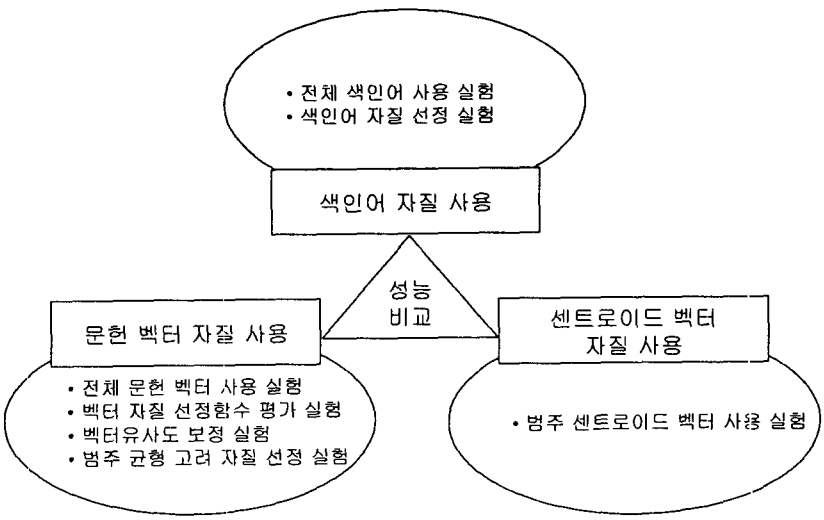
벡터 자질의 추출과 유사도 산출, 자질 선정을 위한 프로그램은 Visual FoxPro로 구현하였고, 성능비교를 위한 SVM 분류기는 공개용 기계학습 실험 패키지인 WEKA version 3.4(Witten & Frank 2000)를 사용하였다.

3. 2 실험 구성

전통적인 색인어 자질을 사용한 경우를 기준 성능으로 삼고, 문헌 벡터 자질과 범주 센트로이드 벡터 자질을 사용한 경우의 성능을 SVM 분류기로 확인하는 실험을 〈그림 2〉와 같이 구성하였다.

우선 색인어 자질 사용 실험에서는 전통적인 방법대로 색인어 자질로 문헌을 표현한 경우의 SVM 분류기 성능을 확인하였다. 또한 자질 선정을 적용하였을 때의 성능도 알아보았다. 이를 확인한 이유는 색인어 자질 전체를 사용한 경우와, 색인어 자질 선정을 적용한 경우의 성능을 이후의 실험에서 비교 기준으로 삼기 위해서이다.

문헌 벡터 자질 사용 실험에서는 문헌 벡터 표현의 성능과 벡터 자질 선정의 효과가 어느 정도인가를 확인하는 실험을 수행하였다. 먼저 문헌



〈그림 2〉 자질 표현 방식에 따른 SVM 분류기 성능 비교 실험 구성

벡터간 유사도를 산출한 다음 각 문헌의 자질과 자질값을 문헌 벡터와 벡터유사도로 표현하였을 때의 성능을 검증하였다. 전체 문헌 벡터를 모두 사용한 실험 이후에는 일부 문헌 벡터를 선정하기 위한 기준 함수를 제안하고 이를 이용한 벡터 자질 선정 실험을 수행하였다. 이 실험까지는 벡터유사도를 그대로 자질값으로 사용하였으나, 벡터 자질을 선정하였을 때의 성능 개선을 위해서 벡터유사도를 보정하여 자질값으로 사용하는 실험을 그 다음으로 수행하였다. 또한 벡터 자질 선정이 범주별로 불균등한 점을 개선하기 위한 방안도 마련해보았다.

마지막 센트로이드 벡터 자질 사용 실험에서는, 빠른 처리가 가능한 센트로이드 벡터 표현이 색인어 자질이나 문헌 벡터 자질 표현에 비해서 어느 정도의 분류성능을 보이는 가를 알아보았다. 각 범주를 대표하는 센트로이드 벡터를 산출한 다음 학습문헌과 검증문헌을 이 센트로이드 벡터와 비교한 벡터유사도로 표현하여 SVM 분류기를 실행하였다.

각 방식에 따른 분류성능의 측정은 마이크로 평균 정확률 척도로 평가하였다. 마이크로 평균 정확률 척도는 전체 범주 할당 건수 중에서 옳게 분류된 경우의 비율을 산출하는 것이다. 이 연구에서 사용한 실험문서에는 각 문서마다 분류기호가 하나씩만 할당되어 있으므로 마이크로 평균 정확률과 마이크로 평균 재현율, 그리고 이를 결합한 마이크로 평균 F1 척도는 같은 값을 가진다. 마이크로 평균 정확률 척도의 공식은 다음과 같다.

$$\text{마이크로 평균 정확률} = \frac{\text{검증문헌이 적합 범주에 할당된 횟수}}{\text{검증문헌의 총 할당 횟수}}$$

분류실험의 성능 평가를 위한 척도로는 이밖에도 각 범주별로 정확률과 재현율을 구해서 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 & Liu 1999).

4. SVM 분류기를 이용한 분류 실험 결과

4. 1 색인어 자질 사용 실험

이 연구에서 실험대상으로 삼은 문헌집단에 대해서 기존의 방법과 같은 색인어 자질을 사용해서 학습문헌과 검증문헌을 표현하고 이를 SVM 분류기에 입력하여 성능을 확인하였다. 이때 색인어 자질 선정의 효과를 알아보기 위해서 가장 많이 사용되고 성능이 좋은 것으로 알려진 카이제곱 통계량과 정보획득량 공식에 따른 색인어 자질 선정 실험도 함께 수행하였다. 카이제곱 통계량과 정보획득량 공식은 k번째 색인어를 t_k, i번째 범주를 c_i, 범주의 수를 m, 총 학습문헌 수를 N이라고 할 때 각각 다음과 같다(Yang & Pederson 1997).

$$IG(t_k) = -\sum_{i=1}^{m} P_r(c_i)\log P_r(c_i) + P_r(t_k)\sum_{i=1}^{m} P_r(c_i|\, t_k)\log P_r(c_i|\, t_k) + P_r(\overline{t_k})\sum_{i=1}^{m} P_r(c_i|\, \overline{t_k})\log P_r(c_i|\, \overline{t_k})$$

$$\chi^2(t_k, c_i) = \frac{N[P(t_k, c_i)P(\bar{t}_k, \bar{c}_i) - P(t_k, \bar{c}_i)P(\bar{t}_k, c_i)]^2}{P(t_k)P(\bar{t}_k)P(c_i)P(\bar{c}_i)}$$

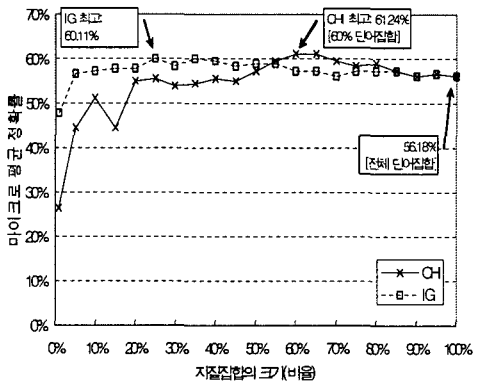
$$\chi^2_{\max}(t_k) = \max_{i=1}^{m} \{\chi^2(t_k, c_i)\}$$

각 색인어의 가중치는 문헌내 빈도에 로그를 취하고 1을 더한 로그TF에 역문헌빈도를 곱한 로그TFIDF 가중치 방식을 사용하였다. 이후의 모든 실험에서는 색인어의 가중치를 이와 동일하게 설정하였다.

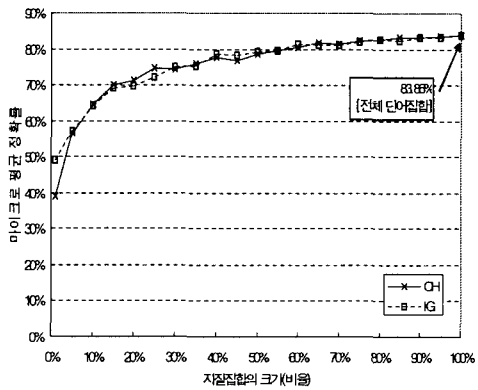
색인어 자질 선정에서는 전체 자질을 모두 사용한 경우의 자질집합 크기를 100%로 보았을 때 5% 포인트씩 크기를 줄여가면서 SVM 분류기에 적용해보았다. 두 문헌집단 KFCM-896과 TREND-2287에 대한 실험 결과는 〈표 2〉와 〈그림 3〉, 〈그림 4〉에 제시하였다.

〈표 2〉 색인어 자질 선정에 따른 분류성능 비교

자질집합의 크기 (비율)	KFCM-896			TREND-2287		
	색인어 종수	자질 선정 방법		색인어 종수	자질 선정 방법	
		CHI 기준	IG 기준		CHI 기준	IG 기준
1%	73	26.40%	47.75%	75	39.04%	49.14%
5%	363	44.38%	56.74%	377	56.54%	57.08%
10%	726	51.12%	57.30%	754	64.47%	64.11%
15%	1089	44.38%	57.87%	1132	70.06%	69.07%
20%	1452	55.06%	57.87%	1509	71.33%	69.79%
25%	1815	55.62%	60.11%	1886	74.66%	72.32%
30%	2178	53.93%	58.43%	2263	74.30%	74.93%
35%	2541	54.49%	60.11%	2640	75.92%	75.29%
40%	2904	55.62%	59.55%	3018	77.73%	78.72%
45%	3267	55.06%	58.43%	3395	76.92%	78.27%
50%	3631	57.30%	58.99%	3772	78.54%	79.26%
55%	3994	59.55%	58.99%	4149	79.44%	79.62%
60%	4357	61.24%	57.30%	4526	80.61%	81.51%
65%	4720	61.24%	57.30%	4904	81.61%	81.15%
70%	5083	59.55%	56.18%	5281	81.51%	81.06%
75%	5446	58.43%	57.30%	5658	82.42%	81.97%
80%	5809	58.99%	57.30%	6035	82.60%	82.69%
85%	6172	57.30%	57.30%	6412	83.32%	82.33%
90%	6535	56.18%	56.18%	6790	83.32%	83.14%
95%	6898	56.74%	56.74%	7167	83.59%	83.14%
100%	7261	56.18%	56.18%	7544	83.86%	83.86%



〈그림 3〉 색인어 자질 선정에 따른 분류성능 비교 (KFCM-896)



〈그림 4〉 색인어 자질 선정에 따른 분류성능 비교 (TREND-2287)

이 결과를 보면 KFCM-896 실험집단에서는 전체 자질을 모두 사용한 경우에 56.18%, 카이제곱통계량(CHI)을 기준으로 자질 선정하였을 때 60% 자질집합을 사용해서 61.24%, 정보획득량(IG)을 기준으로 자질 선정하였을 때 25% 자질집합을 사용해서 60.11%의 최고성능을 얻었다. 따라서 카이제곱통계량을 기준으로 자질 선정한 경우에 전체 자질을 모두 사용한 경우보다 최고 9.0% 향상된 성능을 얻을 수 있었다.

TREND-2287 실험집단에서는 전체 자질을 모두 사용한 경우에 83.86%의 성능을 보였고 자질 선정을 한 경우에는 카이제곱통계량이나 정보획득량 어느 기준을 적용하더라도 모두 성능이 저하되는 것으로 나타났다.

이후의 실험에서는 KFCM-896에서 56.18%(전체 자질 사용 성능)와 61.24%(자질 선정 후 최고 성능), TREND-2287 실험집단에서 83.86%(전체 자질 사용 성능 겸 최고 성능)를 기준 성능으로 삼아서 비교하였다.

KFCM-896 실험집단에서만 자질 선정으로 성능이 향상된 이유는 과학기술동향속보인 TREND-2287 실험집단과 달리 일반 신문기사이기 때문에, 분류에 도움이 안되는 불필요한 색인어가 많아서 자질 선정의 효과가 나타난 것이라고 판단된다.

4. 2 문헌 벡터 자질 사용 실험

4.2.1 전체 문헌 벡터 사용

각 학습문헌 벡터를 자질로 삼고 학습문헌간 벡터유사도를 자질값으로 하여 학습문헌을 표현하였다. 검증문헌도 개별 학습문헌과의 코사인유사도를 산출하여 이를 자질값으로 삼아서 표현하였다. 벡터유사도를 자질값으로 하는 벡터 자질의 수는 학습문헌의 수와 같으므로 KFCM-896 실험집단에서는 718개, TREND-2287 실험집단에서는 1,178개가 되었다. 벡터 자질로 표현한 문헌을 SVM 분류기로 분류한 결과를 색인어 자질을 사용한 결과와 비교하면 〈표 3〉, 〈표 4〉와 같다.

〈표 3〉 분류자질 표현에 따른 분류성능 비교 (KFCM-896)

	색인어 자질 사용			문헌 벡터 자질 사용
	색인어 전체	CHI 기준 자질 선정 후 최고 성능	IG 기준 자질 선정 후 최고 성능	
자질 수	7,261	4,357	1,815	718
마이크로 평균 정확률	56.18%	61.24%	60.11%	65.73%
성능 향상율 (색인어 전체 사용 대비)	-	9.0%	7.0%	17.0%

〈표 4〉 분류자질 표현에 따른 분류성능 비교 (TREND-2287)

	색인어 자질 사용			문헌 벡터 자질 사용
	색인어 전체	CHI 기준 자질 선정 후 최고 성능	IG 기준 자질 선정 후 최고 성능	
자질 수	7,544	7,167	7,167	1,178
마이크로 평균 정확률	83.86%	83.59%	83.14%	87.29%
성능 향상율 (색인어 전체 사용 대비)	-	-0.3%	-0.9%	4.1%

결과를 보면, 문헌 벡터 자질을 사용하는 경우에 두 실험집단 모두에서 색인어 자질을 사용하는 것보다 좋은 성능을 얻을 수 있었다. KFCM-896 실험집단에서는 색인어 자질을 사용한 경우에 비해서 17.0%, TREND-2287 실험집단에서는 4.1% 향상된 성능을 얻었다.

색인어 자질을 일부 선정한 경우에 비해서도 문헌 벡터 자질을 사용한 경우가 더 좋았다. 특히 색인어 자질 선정으로는 성능을 향상시키지 못했던 TREND-2287 실험집단에서도 벡터 자질을 사용한 결과 성능이 향상된 것은 고무적인 결과이다.

4.2.2 문헌 벡터 자질 축소

앞에서와 같이 문헌 벡터 자질을 이용하여 SVM분류기의 문헌분류 성능을 높일 수는 있으나 문제는 처리 속도에 있다. kNN 분류기처럼 분류대상 문헌과 모든 학습문헌과의 유사도를 구해야 하기 때문이다. 그 결과 분류성능은 좋아지더라도 실행시간은 kNN 분류기보다도 더 걸리게 된다.

이에 대한 해결책으로 이 절에서는 색인어 자질 표현에서와 마찬가지로 자질 선정을 적용해보기로 한다. 학습문헌 중에서 일정한 기준에 의해 선정된 일부 문헌만을 벡터 자질로 채택하고 이 문헌 벡터와의 유사도만 산출하는 것이다. 이를 위해서 학습문헌 중에서 벡터 자질로 사용할 부분집합을 선정하기 위한 평가 함수로 다음의 두 가지를 검토하였다.

① 문헌 벡터 자질 평가함수 M (범주내 유사도 기준) : 학습집단에서 소속 범주내 타 문헌과의 유사도가 가장 높은 문헌을 우선적으로 선정하는 방식이다. 소속범주에 대한 충성도가 높은

문헌에 유리한 기준이 된다. 어떤 문헌이 소속된 주제의 다른 문헌과 유사하여 해당 주제를 잘 대변한다면 분류자질로서의 자격을 갖추었다는 판단에 근거한 것이다. 이 방식은 색인어 자질 선정에서 특정 범주와의 연관성이 높은 자질을 우선적으로 선정하는 경우와 마찬가지 발상이다. 범주 c_j에 속한 문헌 d_k의 자질로서의 가치를 평가하기 위한 평가함수 M을 공식으로 표현하면 다음과 같다.

$$f_M(d_k) = \frac{\sum_{i=1}^{|c_j|} sim(d_k, d_i)}{|c_j|} \quad , \ i \neq k$$

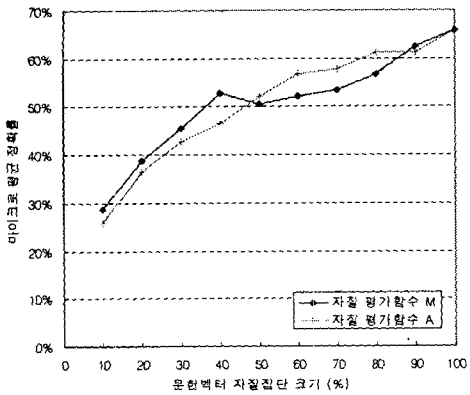
② 문헌 벡터 자질 평가함수 A (전체 유사도 기준) : 학습집단에 속한 모든 타 문헌과의 유사도 평균이 가장 높은 문헌을 우선적으로 선정하는 방식이다. 이 방식은 색인어 자질 선정에서 출현한 문헌 수가 많은 고빈도어를 선정하는 경우에 대응되는 방법이다. 되도록이면 많은 학습문헌과 비교적 강한 관계를 맺고 있는 문헌을 선정하는 결과가 된다. 즉, 산출된 문헌유사도가 가급적 학습집단의 다수 문헌에 대한 정보를 반영할 수 있도록 배려하는 방법이다. 전체 학습문헌이 n개일 때 문헌 d_k의 자질로서의 가치를 평가하기 위한 평가함수 A를 공식으로 표현하면 다음과 같다.

$$f_A(d_k) = \frac{\sum_{i=1}^{n} sim(d_k, d_i)}{n} \quad , \ i \neq k$$

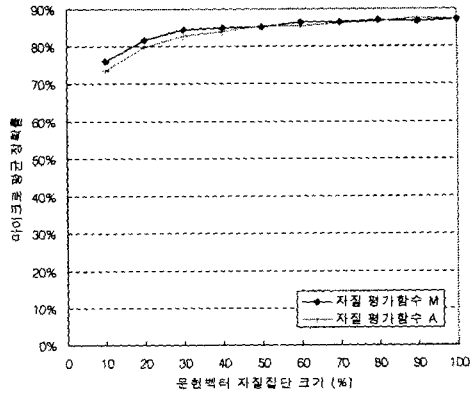
두 가지 평가함수 M과 A를 기준으로 하여 함수값이 높은 문헌을 우선적으로 벡터 자질로 채택하였다. 문헌 벡터 자질을 10% 포인트씩 줄여가면서 SVM 분류기로 분류실험한 결과를 〈표 5〉와 〈그림 5〉, 〈그림 6〉에 제시하였다.

〈표 5〉 문헌 벡터 자질 선정기준에 따른 분류 성능 비교

자질집합의 크기(비율)	KFCM-896			TREND-2287		
	문헌 벡터 자질 수	분류 성능		문헌 벡터 자질 수	분류 성능	
		평가함수 M	평가함수 A		평가함수 M	평가함수 A
10%	72	28.65%	25.84%	118	75.92%	73.58%
20%	144	38.76%	36.52%	236	81.79%	79.89%
30%	215	45.51%	42.70%	353	84.49%	82.69%
40%	287	52.81%	46.63%	471	84.94%	83.95%
50%	359	50.56%	52.25%	589	85.21%	85.48%
60%	431	52.25%	56.74%	707	86.56%	85.48%
70%	503	53.37%	57.87%	825	86.56%	86.11%
80%	574	56.74%	61.24%	942	87.02%	86.74%
90%	646	62.36%	61.24%	1060	86.74%	87.47%
100%	718	65.73%	65.73%	1178	87.29%	87.29%



〈그림 5〉 문헌 벡터 자질 선정기준에 따른 분류 성능 (KFCM-896)



〈그림 6〉 문헌 벡터 자질 선정기준에 따른 분류 성능 (TREND-2287)

문헌 벡터 자질 선정 실험결과는 거의 모든 경우(TREND-2287 집단에서 평가함수 A으로 90% 자질 선정한 경우만 제외)에서 문헌 벡터를 모두 사용한 것에 비해 성능 향상을 얻지 못하는 것으로 나타났다. 두 평가함수를 이용한 결과를 비교하면 자질집합을 50% 축소할 때까지는 평가함수 A가 약간 더 좋거나 비슷한데, 그 이하로 더 작게 축소하면 두 실험집단 모두에서 평가함수 M이 더 좋은 성능을 보인다. 이는 문헌 벡터 자질이 너무 적으면 함수 A와 같은 경우에는 큰 범주에 속한 문헌들 위주로 선정될 가능성이 더 높으므로 선정된 문헌 벡터 자질이 여러 범주를 고르게 반영하지 못하기 때문으로 생각된다.

색인어 자질을 사용한 경우와 비교해보면, 평가함수 A에 따라 자질 선정한 경우에 절반정도 (KFCM-896에서는 60% 이상, TREND-2287에서는 50%) 축소했을 때까지는 전체 색인어 자질을 사용한 경우(56.18%와 83.86%)보다 성능이 좋으나, 그보다 더 작게 축소하면 색인어 자질을 모두 사용한 경우보다 성능은 낮아진다. 실험집단에 따른 차이는 흥미롭게도 색인어 자질 선정의 경우와 상반되게 나타났다. KFCM-896 실험집단에서는 문헌 벡터 자질을 줄여나갈 수록 성능이 급격하게 저하되는 반면에 TREND-2287 실험집단에서는 완만하게 성능이 저하되었다. 색인어 자질을 사용해서 얻을 수 있는 최고 성능(61.24%와 83.86%)과 비교해도 KFCM-896 실험집단에서는 80%로만 문헌 벡터 자질 집합을 줄이더라도 색인어 자질의 경우보다 같거나 낮아지는 반면에, TREND-2287 실험집단에서는 평가함수 M으로는 30%, 평가함수 A로는 40%로까지 줄이더라도 더 좋게 나타났다. 이는 일반 신문기사인 KFCM-896 실험집단에서는 한 범주에 속한 문헌들의 다양성이 큰 반면에 TREND-2287 실험집단의 문헌들은 같은 범주에 속하면 내용도 비슷하기 때문으로 판단된다. 주제 내 문헌의 다양성이 큰 경우에는 문헌을 일부만 제외하더라도 해당 범주의 한 측면이 배제되기 때문이다.

결론적으로 문헌 벡터를 자질로, 벡터유사도를

자질값으로 사용한 경우에는 자질 선정이 분류성능을 향상시키는 데 도움이 되지 못하는 것으로 나타났다. 다만 주제적으로 동질성이 강한 TREND-2287 실험집단에서는 30%내지 40%로까지 문헌 벡터 자질을 줄이더라도 색인어 자질로 얻을 수 있는 최고 성능보다는 높은 성능을 얻을 수 있었다.

4. 3 벡터 유사도 보정 실험

앞 절의 실험 결과는 비록 문헌 벡터 자질 표현이 색인어 자질 표현에 비해서 좋은 성능을 보이긴 하지만, 실행 효율을 높이기 위한 자질 선정 적용은 성능 저하를 피할 수 없는 것으로 나타났다. 이를 보완하기 위한 방법으로 이 절에서는 자질값으로 사용한 벡터유사도를 그대로 사용하지 않고 보정하는 방안을 검토하였다.

벡터유사도를 보정할 필요가 있다고 판단한 것은 자질 선정 결과에서 벡터유사도의 분포가 학습문헌에 따라서 심하게 차이가 나기 때문이다. 자질 선정 전에는 SVM 분류기에 입력하는 학습문헌 자료가 문헌-문헌 유사도 행렬이면서 대칭행렬이었던 반면에, 자질 선정 이후에는 비대칭 행렬이 된다. 전체가 아닌 선택된 일부 학습문헌들과의 벡터유사도만 포함하게 되면 선정된 학습문헌의 벡터에는 1.0인 값(자기 자신과의 유사도)이 있고 나머지 학습문헌의 벡터에는 1.0인 값이 없는 상황이 발생한다. 대부분의 문헌간 벡터유사도는 1보다 훨씬 낮은 값이 된다는 점을 고려하면 선정/비선정 학습문헌 벡터의 값 분포가 심하게 차이날 수 있다.

이와 같은 유사도 분포의 차이를 보정하기 위해서 벡터유사도의 분포를 변환하는 함수를 적용하였다. 유사도 분포 변환의 기본 방향은 1보다는 0에 훨씬 가까운 대부분의 유사도를 1에 가깝게 되도록 올려주는 것이다(<그림 8> 참조). 이 연구에서 적용한 유사도 분포 변환 함수는 다음의 두 가지이다.

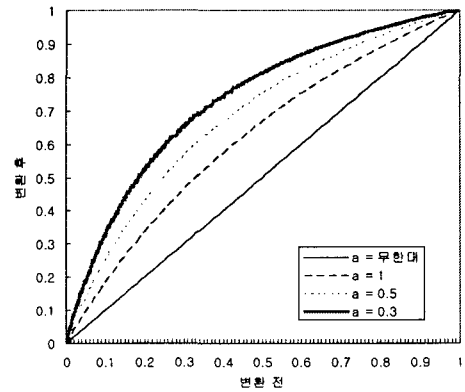
① 멱변환 (power transformation) : 학습문헌과의 코사인 유사도의 제곱근을 자질가중치로 사용하는 방식이다. 0에서 1사이의 값에 대해서 제곱근을 취하면 전체적인 유사도값(자질값)의 분포가 정규분포와 비슷하게 변환된다(<그림 8> 참조). 이를 통해 0에 가까운 값이 지나치게 많은 경우를 피할 수 있다. 멱지수를 2가 아닌 다른 값을 사용할 수도 있으며 이와 비슷하게 제곱근이 아닌 로그를 취하는 로그 변환도 있다 (Fukunaga 1990). 멱변환 공식은 다음과 같으며 이 연구에서 사용한 매개변수 a는 제곱근이므로 2이다.

$$y = x^{\frac{1}{a}}$$

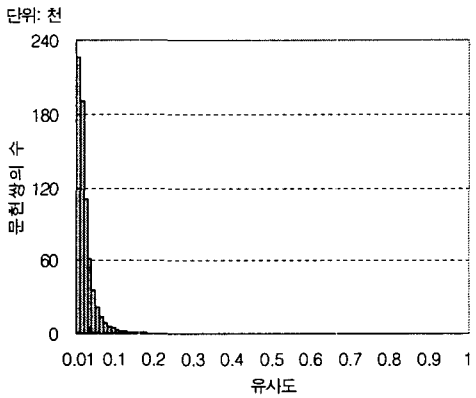
② 굽이변환(bend transformation) : 이 연구에서 제안하는 새로운 변환 방식으로서 다음의 공식을 적용하는 방법이다.

$$y = \frac{x + ax}{x + a}$$

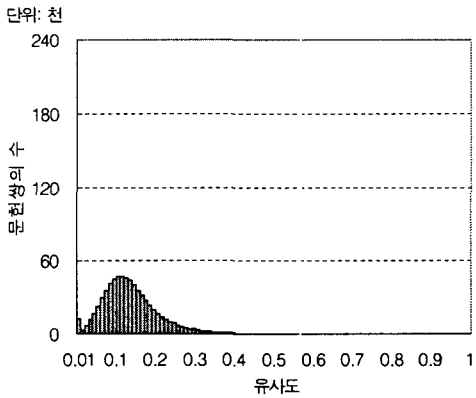
여기서 매개변수 a가 0이면 변환값은 항상 1이다. a가 무한대이면 y=x가 되어 변환하지 않는 것과 같다. 원래 유사도가 0.5(1/2)이면 a가 1일 때 0.6667(2/3), a가 0.5일때 0.75(3/4), a가 0.25일때 0.8333(5/6)이 된다. 즉 매개변수 a는 <그림 8>에서 보듯이 0에 가깝게 작을수록 분포



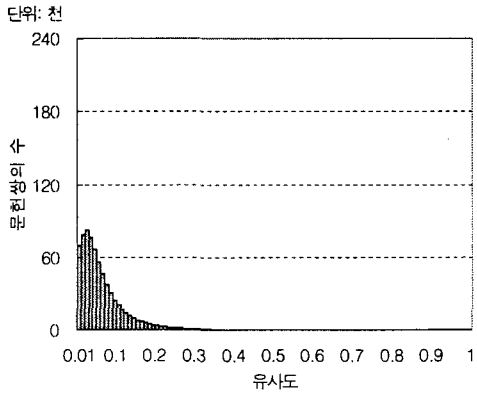
〈그림 7〉 굽이변환에서 매개변수 a에 따른 변환값의 차이



(a) 학습문헌 쌍의 벡터유사도 분포



(b) 학습문헌 쌍의 벡터유사도 멱변환 (a=0. 5) 값 분포



(c) 학습문헌 쌍의 벡터유사도 굽이변환(a=0. 5) 값 분포

〈그림 8〉 0. 01 단위로 구간을 나누었을 때 학습문헌 쌍 유사도의 구간별 빈도 분포 (TREND-2287)

변환이 심하고, 반대로 커질수록 분포 변환이 미미한 특성을 가진다. 굽이변환은 〈그림 8〉에서 보듯이 분포를 변환하였을 때 멱변환과 달리 값이 낮은 경우가 급격히 감소하지 않는다.

학습문헌간 벡터유사도를 멱변환과 굽이변환을 거쳐서 자질값으로 사용하였을 때의 SVM 분류기의 성능을 〈표 6〉-〈표 9〉, 〈그림 9〉-〈그림 12〉에 제시하였다. 실험 결과를 보면 평가함수 M을 사용한 경우보다 평가함수 A를 사용한 경우에 자질 선정을 통한 성능개선의 효과가 실험집단별로 안정적으로 나타나므로 이후 성능에 대한 분석은 평가함수 A를 적용하여 자질 선정한 경우에 대해서만 하기로 한다.

일단 자질 선정을 하지 않고 전체 벡터 자질을 사용한 경우를 보면 굽이변환을 한 경우에 벡터유사도를 그대로 자질값으로 사용한 경우보다 두 실험집단에서 모두 성능이 약간 향상된 것으로 나타났다(a=0.5일때, 65.73%→68.54; 87.29%→87.38%). 멱변환을 한 경우에는 TREND-2287 실험집단에서는 벡터유사도를 그대로 자질값으로 사용한 경우보다 저하되었고 KFCM-896 실험집단에서는 약간 향상되었지만 굽이변환을 한 경우보다는 낮았다.

평가함수 A를 기준으로 자질선정하였을 때, 유사도를 굽이변환(a=0.5)하여 자질값으로 사용한 경우에는 두 실험집단에서 모두 성능을 향상시킬 수 있는 것으로 나타났다. KFCM-896 실험집단에서는 〈표 7〉과 같이 벡터 자질을 70%만 사용하더라도 자질 선정 없이 벡터유사도를 자질값으로 사용한 경우(65.73%)와 같은 성능을 얻었으며 90% 사용하였을 때에는 69.10%로 상당히 향상된 성능을 얻었다. TREND-2289 실험집단에서는 〈표 9〉와 같이 벡터 자질을 60%만 사용하더라도 벡터유사도를 사용한 경우(87.29%) 보다 좋은 성능(87.56%)을 얻었다. 이와 달리 멱변환을 한 경우에는 자질 선정을 하더라도 〈표 9〉와 같이 TREND-2289 실험집단에서는 벡터유사도를 사용한 경우보다 좋은 성능을 얻지 못했다.

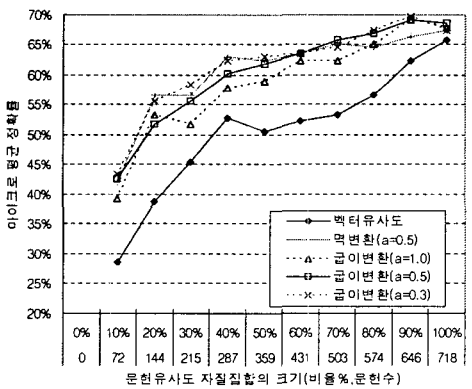
이상과 같이 벡터유사도를 그대로 자질값으로 사용하지 않고 분포변환하여 사용한 결과 자질선정을 한 경우와 하지 않은 경우에 모두 분류 성능을 향상시킬 수 있었다. 이로써 문헌 벡터 자질을 사용한 경우에도 자질값인 벡터유사도를 분포변환하면 자질 선정으로 성능을 향상시킬 수 있는 것으로 나타났다.

〈표 6〉 문헌유사도 자질값 보정 결과 (KFCM-896, 평가함수 M 기준 자질 선정)

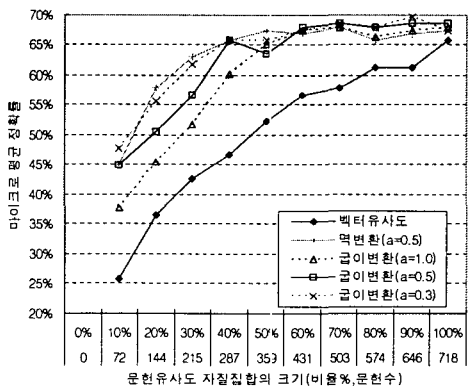
자질집합 크기		문헌 벡터 유사도	멱변환 (a=0.5)	굽이변환 (a=1.0)	굽이변환 (a=0.5)	굽이변환 (a=0.3)
문헌 수	비율					
72	10%	28.65%	42.13%	39.33%	42.70%	43.26%
144	20%	38.76%	56.74%	53.37%	51.69%	55.62%
215	30%	45.51%	56.74%	51.69%	55.62%	58.43%
287	40%	52.81%	62.92%	57.87%	60.11%	62.36%
359	50%	50.56%	62.36%	58.99%	61.80%	62.92%
431	60%	52.25%	63.48%	62.36%	63.48%	63.48%
503	70%	53.37%	65.17%	62.36%	65.73%	64.61%
574	80%	56.74%	64.61%	65.17%	66.85%	67.42%
646	90%	62.36%	66.29%	69.10%	69.10%	69.66%
718	100%	65.73%	67.42%	67.98%	68.54%	67.42%

〈표 7〉 문헌유사도 자질값 보정 결과 (KFCM-896, 평가함수 A 기준 자질 선정)

자질집합 크기		문헌 벡터 유사도	멱변환 (a=0.5)	굽이변환 (a=1.0)	굽이변환 (a=0.5)	굽이변환 (a=0.3)
문헌 수	비율					
72	10%	25.84%	44.94%	37.64%	44.94%	47.75%
144	20%	36.52%	57.87%	45.51%	50.56%	55.62%
215	30%	42.70%	62.92%	51.69%	56.74%	61.80%
287	40%	46.63%	65.73%	60.11%	65.73%	65.73%
359	50%	52.25%	67.42%	65.17%	63.48%	65.73%
431	60%	56.74%	66.85%	67.42%	67.98%	67.42%
503	70%	57.87%	67.98%	67.98%	68.54%	68.54%
574	80%	61.24%	65.73%	66.29%	67.98%	67.98%
646	90%	61.24%	66.85%	67.42%	68.54%	69.66%
718	100%	65.73%	67.42%	67.98%	68.54%	67.42%



〈그림 9〉 문헌유사도 자질값 보정 결과
(KFCM-896, 평가함수 M 기준 자질 선정)



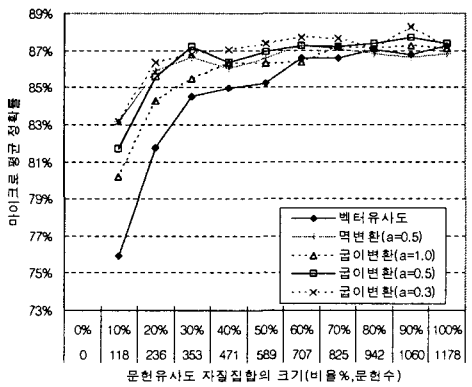
〈그림 10〉 문헌유사도 자질값 보정 결과
(KFCM-896, 평가함수 A기준 자질 선정)

〈표 8〉 문헌유사도 자질값 보정 결과 (TREND-2287, 평가함수 M 기준 자질 선정)

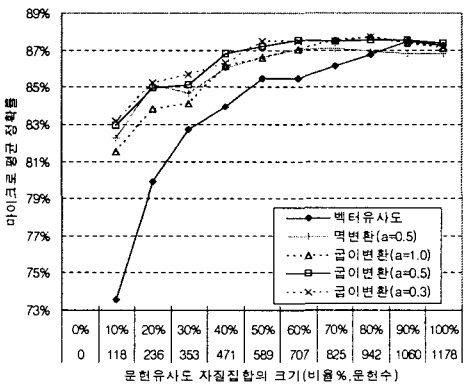
자질집합 크기		문헌 벡터 유사도	멱변환 (a=0.5)	굽이변환 (a=1.0)	굽이변환 (a=0.5)	굽이변환 (a=0.3)
문헌 수	비율					
118	10%	75.92%	83.14%	80.16%	81.70%	83.14%
236	20%	81.79%	85.93%	84.31%	85.57%	86.38%
353	30%	84.49%	86.65%	85.48%	87.20%	86.83%
471	40%	84.94%	86.02%	86.29%	86.38%	87.02%
589	50%	85.21%	86.56%	86.29%	86.93%	87.38%
707	60%	86.56%	87.29%	86.38%	87.29%	87.74%
825	70%	86.56%	87.11%	87.20%	87.20%	87.65%
942	80%	87.02%	86.83%	87.11%	87.38%	87.11%
1060	90%	86.74%	86.65%	87.29%	87.74%	88.28%
1178	100%	87.29%	86.83%	87.11%	87.38%	87.20%

〈표 9〉 문헌유사도 자질값 보정 결과 (TREND-2287, 평가함수 A 기준 자질 선정)

자질집합 크기		문헌 벡터 유사도	멱변환 (a=0.5)	굽이변환 (a=1.0)	굽이변환 (a=0.5)	굽이변환 (a=0.3)
문헌 수	비율					
118	10%	73.58%	82.24%	81.51%	82.96%	83.14%
236	20%	79.89%	85.12%	83.86%	84.94%	85.21%
353	30%	82.69%	84.67%	84.13%	85.12%	85.66%
471	40%	83.95%	86.02%	86.11%	86.83%	86.29%
589	50%	85.48%	86.65%	86.56%	87.20%	87.47%
707	60%	85.48%	87.02%	87.02%	87.56%	87.47%
825	70%	86.11%	87.11%	87.56%	87.47%	87.47%
942	80%	86.74%	86.93%	87.74%	87.56%	87.74%
1060	90%	87.47%	86.83%	87.38%	87.56%	87.38%
1178	100%	87.29%	86.83%	87.11%	87.38%	87.20%



〈그림 11〉 문헌유사도 자질값 보정 결과
(TREND-2287, 평가함수 M 기준 자질 선정)



〈그림 12〉 문헌유사도 자질값 보정 결과
(TREND-2287, 평가함수 A 기준 자질 선정)

4. 4 범주별 균형 자질 선정 실험

문헌 벡터 자질 표현 방식을 실용화하기 위해서는 실행 효율을 높이기 위해서 가급적 소수의 학습문헌만을 벡터 자질로 사용해서도 높은 성능을 얻을 수 있어야 한다. 앞 절의 실험 결과에서 보듯이 자질가중치로 사용하는 벡터유사도의 분포변환을 통해서 문헌 벡터 자질 집합을 60% 크기로 축소할 때까지는 성능 저하를 피할 수 있었다(TREND-2287 실험집단의 경우). 이를 통해 어느 정도 실행 효율 향상은 가능하지만 아직까지 벡터 자질을 절반 이상 제거할 경우에는 분류 성능의 저하를 피할 수가 없는 것으로 나타났다. 더군다나 KFCM-896 실험집단에서는 여전히 80% 이하로 벡터 자질 집합을 줄이면 성능이 저하되었다.

이 절에서는 문헌 벡터 자질을 절반 이하로 적게 선정하는 경우에 SVM 분류기의 성능을 개선하기 위해서, 선정된 벡터 자질이 각 범주를 공평하게 반영하도록 하는 방안을 찾아보기로 한다. 이에 착안한 이유는 〈표 10〉과 같이 문헌 벡터 자질을 모두 사용한 경우와 20%만 선정해서 사용한 경우의 성능을 범주별로 비교해보면 범주 C4나 C5와 같은 큰 범주의 성능은 조금밖에 저하되지 않는 반면에 크기가 제일 작은 범주 C8의 성능은 54.1%에서 7.7%로 대폭 저하되기 때문이다.

이와 같은 결과는 문헌유사도 자질을 선정하는 기준인 평가함수 A나 M을 그대로 적용하면 선정된 학습문헌이 원래의 비율을 그대로 유지하지 못하고 있음을 시사한다. 만약 규모가 큰 범주에 속한 학습문헌이 선정된 문헌 벡터 자질의 대부분을 차지하는 경우라면 소수 범주를 대변하는 문헌은 드물게 될 것이다. 과연 그런지 확인하기 위해서 실제로 학습문헌집단을 축소하였을 때, 각 분류범주의 구성비율이 어떻게 달라지는지를 알아보았다.

〈표 10〉 벡터 자질 선정 전후의 범주별 성능 비교

범주	크기 (백분율)	전체 문헌 벡터 자질 사용 성능*	20% 문헌 벡터 자질 사용 성능** (괄호 안은 성능차이***)
C1	4.7%	73.4%	73.6%(+ 0.2%)
C2	4.1%	32.1%	24.6%(- 7.5%)
C3	5.0%	76.8%	73.6%(- 3.2%)
C4	26.8%	92.7%	92.0%(- 0.7%)
C5	39.0%	96.0%	95.8%(- 0.2%)
C6	7.2%	73.4%	65.4%(- 8.0%)
C7	10.9%	81.5%	73.4%(- 8.1%)
C8	2.3%	54.1%	7.7%(-46.4%)
매크로 평균		72.5%	63.3%(- 9.2%)

* 성능은 각 범주별 F1척도로 측정함.

** 벡터 자질 선정은 평가함수 A 기준, 자질값은 벡터유사도의 굽이변환(a=0.5) 값임.

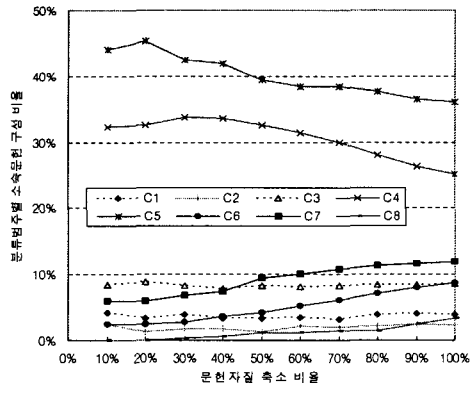
*** 성능 차이의 단위는 %포인트임.

〈표 11〉 축소된 문헌 벡터 자질 집합에서 문헌의 소속범주별 구성비율 (TREND-2287, 평가함수 A)

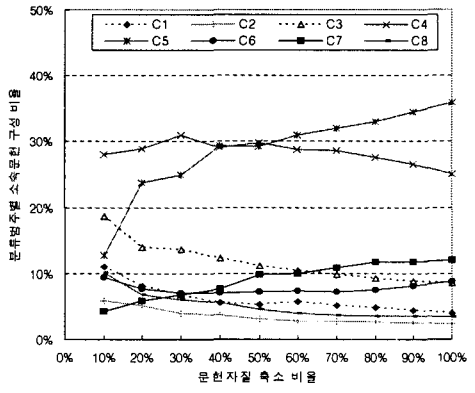
문헌자질축소 비율	분류 범주							
	C1	C2	C3	C4	C5	C6	C7	C8
10%	4.2%	2.5%	8.5%	32.2%	44.1%	2.5%	5.9%	0.0%
20%	3.4%	1.3%	8.9%	32.6%	45.3%	2.5%	5.9%	0.0%
30%	4.0%	1.7%	8.2%	33.7%	42.5%	2.8%	6.8%	0.3%
40%	3.4%	1.7%	7.9%	33.5%	41.8%	3.6%	7.4%	0.6%
50%	3.4%	1.4%	8.3%	32.6%	39.4%	4.2%	9.5%	1.2%
60%	3.5%	2.1%	8.2%	31.4%	38.3%	5.2%	10.0%	1.1%
70%	3.2%	2.1%	8.2%	29.9%	38.3%	6.1%	10.8%	1.5%
80%	3.9%	2.1%	8.4%	28.0%	37.6%	7.1%	11.4%	1.5%
90%	4.1%	2.5%	8.5%	26.3%	36.5%	8.0%	11.7%	2.5%
100%	4.0%	2.3%	8.6%	25.1%	36.0%	8.7%	12.0%	3.3%

〈표 12〉 축소된 문헌자질 집합에서 문헌의 소속범주별 구성비율 (TREND-2287, 평가함수 M)

문헌자질축소 비율	분류 범주							
	C1	C2	C3	C4	C5	C6	C7	C8
10%	11.0%	5.9%	18.6%	28.0%	12.7%	9.3%	4.2%	10.2%
20%	8.1%	5.1%	14.0%	28.8%	23.7%	7.6%	5.9%	6.8%
30%	6.8%	4.0%	13.6%	30.9%	24.9%	7.1%	6.8%	5.9%
40%	5.5%	3.6%	12.3%	29.1%	29.3%	7.0%	7.6%	5.5%
50%	5.3%	3.1%	11.2%	29.7%	29.2%	7.1%	9.8%	4.6%
60%	5.7%	2.8%	10.5%	28.7%	31.0%	7.4%	10.0%	4.0%
70%	5.1%	2.8%	9.8%	28.6%	32.0%	7.2%	10.9%	3.6%
80%	4.9%	2.7%	9.2%	27.5%	33.0%	7.5%	11.7%	3.5%
90%	4.4%	2.5%	8.8%	26.5%	34.4%	8.1%	11.7%	3.6%
100%	4.0%	2.3%	8.6%	25.1%	36.0%	8.7%	12.0%	3.3%



〈그림 13〉 자질 축소에 따른 각 범주의 크기 비율 변화 (TREND-2287, 평가함수 A)



〈그림 14〉 자질 축소에 따른 각 범주의 크기 비율 변화 (TREND-2287, 평가함수 M)

평가함수 A과 M에 따라서 TREND-2287 실험집단에서 학습문헌집단을 축소하면서, 선정된 문헌 벡터의 각 분류범주별 구성 비율을 살펴보면 〈표 11〉, 〈표 12〉, 〈그림 13〉, 〈그림 14〉과 같다.

분석 결과를 보면 평가함수 A를 자질 선정 기준으로 하였을 때 원래 규모가 큰 범주인 C4와 C5는 20%로 축소된 학습집단에서 더 큰 몫(36.0%->45.3%, 25.1%->32.6%)을 차지하는데 반해서, 규모가 작은 C6이나 C8과 같은 범주는 20%로 축소된 학습집단에서 구성비가 훨씬 줄어드는 것(8.7%→2.5%, 3.3%→0.0%)을 볼 수 있다. 평가함수 M에서는 이와 다소 상반된 결과가 나타나서 규모가 큰 범주 C5는 20%로 축소된 학습집단에서 구성비가 감소하고(36.0%→23.7%), 규모가 작은 범주 C6은 구성비가 증가하였다(3.3%→6.8%).

어느 경우든 간에 원래 학습문헌의 구성비가 학습집단의 축소과정에서 상당히 달라진다는 것이 뚜렷하게 나타났다. 최초의 구성비율을 기준으로 하여, 축소된 학습집단에서도 같은 구성비율이 유지될 수 있도록 문헌 벡터 자질 축소방식을 개선해보았다.

개선된 방식에서는 범주별 균형을 고려하여, 전체 학습집단 단위로 문헌 벡터 자질을 선택하지 않고 각 범주별로 문헌 벡터 자질을 선택하였다. 즉, 50% 축소라면 각 범주마다 원래 학습문헌의 절반씩만 남기도록 한 것이다. 이 방법을 사용하면 원래 학습문헌의 범주별 구성비율은 항상 유지될 수 있다. 다만 소속범주가 다르다는 이유로 선정에 사용하는 척도값이 더 낮은 문헌이 선정되고 더 높은 문헌이 배제되는 경우도 있을 것이다. 범주별 균형 유지가 중요한지, 학습문헌 선정을 위한 척도값의 절대적인 기준 유지가 중요한지는 실험 결과에서 확인될 것이다.

범주별 균형 자질 선정 실험 결과는 〈표 13〉-〈표 16〉, 〈그림 15〉, 〈그림 16〉과 같다. 평가함수 A를 기준으로 자질 선정하였을 때, 벡터유사도를 굽이변환(a=0.3)하여 자질값으로 사용한 경우에는 두 실험집단에서 모두 벡터 자질을 50%

〈표 13〉 범주별 균형 자질 선정 결과 (KFCM-896, 평가함수 M 기준 자질 선정)

자질집합 크기		문헌 벡터 유사도	멱변환 (a=0.5)	굽이변환 (a=1.0)	굽이변환 (a=0.5)	굽이변환 (a=0.3)
문헌 수	비율					
72	10%	31.46%	50.56%	46.63%	50.00%	52.81%
144	20%	42.13%	55.62%	51.69%	55.06%	54.49%
215	30%	47.19%	56.18%	52.81%	55.62%	55.06%
287	40%	48.88%	58.99%	55.62%	57.87%	58.99%
359	50%	56.18%	62.36%	61.80%	64.04%	62.92%
431	60%	56.18%	62.36%	64.04%	64.61%	63.48%
503	70%	57.30%	65.73%	66.29%	67.98%	65.17%
574	80%	60.11%	64.61%	69.10%	69.66%	70.22%
646	90%	64.04%	64.61%	69.10%	69.66%	68.54%
718	100%	65.73%	65.73%	67.98%	68.54%	67.42%

〈표 14〉 범주별 균형 자질 선정 결과 (KFCM-896, 평가함수 A 기준 자질 선정)

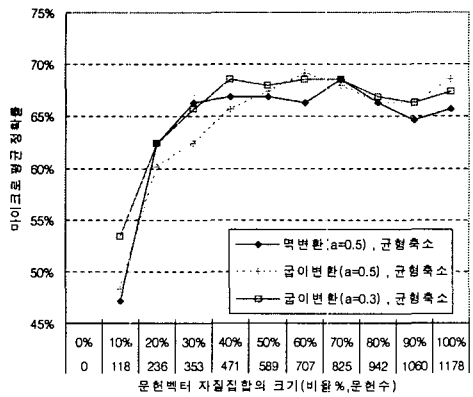
자질집합 크기		문헌 벡터 유사도	멱변환 (a=0.5)	굽이변환 (a=1.0)	굽이변환 (a=0.5)	굽이변환 (a=0.3)
문헌 수	비율					
72	10%	29.21%	47.19%	38.76%	48.31%	53.37%
144	20%	38.76%	62.36%	55.62%	60.11%	62.36%
215	30%	45.51%	66.29%	57.87%	62.36%	65.73%
287	40%	55.06%	66.85%	63.48%	65.73%	68.54%
359	50%	55.62%	66.85%	65.73%	67.42%	67.98%
431	60%	55.62%	66.29%	66.85%	69.10%	68.54%
503	70%	56.18%	68.54%	64.61%	67.98%	68.54%
574	80%	61.80%	66.29%	66.29%	66.29%	66.85%
646	90%	61.80%	64.61%	66.85%	66.29%	66.29%
718	100%	65.73%	65.73%	67.98%	68.54%	67.42%

〈표 15〉 범주별 균형 자질 선정 결과 (TREND-2287, 평가함수 M 기준 자질 선정)

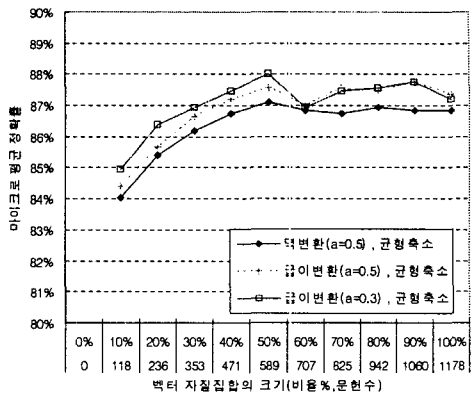
자질집합 크기		문헌 벡터 유사도	멱변환 (a=0.5)	굽이변환 (a=1.0)	굽이변환 (a=0.5)	굽이변환 (a=0.3)
문헌 수	비율					
118	10%	75.56%	83.50%	81.97%	82.96%	83.68%
236	20%	80.07%	85.48%	84.31%	85.12%	85.84%
353	30%	81.79%	86.47%	85.39%	86.29%	86.93%
471	40%	83.59%	86.56%	86.47%	86.74%	86.65%
589	50%	84.58%	86.74%	86.20%	86.74%	87.20%
707	60%	85.48%	86.74%	86.74%	87.11%	87.29%
825	70%	85.93%	87.02%	86.93%	87.29%	87.02%
942	80%	86.74%	86.93%	87.38%	87.65%	87.65%
1060	90%	86.74%	86.93%	87.38%	87.38%	87.83%
1178	100%	87.29%	86.83%	87.11%	87.38%	87.20%

〈표 16〉 범주별 균형 자질 선정 결과 (TREND-2287, 평가함수 A 기준 자질 선정)

자질집합 크기		문헌 벡터 유사도	멱변환 (a=0.5)	굽이변환 (a=1.0)	굽이변환 (a=0.5)	굽이변환 (a=0.3)
문헌 수	비율					
118	10%	75.83%	84.04%	82.87%	84.40%	84.94%
236	20%	80.07%	85.39%	84.58%	85.66%	86.38%
353	30%	82.87%	86.20%	85.75%	86.65%	86.93%
471	40%	84.94%	86.74%	86.74%	87.20%	87.47%
589	50%	85.48%	87.11%	86.83%	87.56%	88.01%
707	60%	85.66%	86.83%	87.20%	87.02%	86.93%
825	70%	86.38%	86.74%	87.65%	87.56%	87.47%
942	80%	86.83%	86.93%	87.47%	87.47%	87.56%
1060	90%	87.20%	86.83%	87.38%	87.74%	87.74%
1178	100%	87.29%	86.83%	87.11%	87.38%	87.20%



〈그림 15〉 범주별 균형 자질 선정 결과
(KFCM-896, 평가함수 A 기준 자질 선정)



〈그림 16〉범주별 균형 자질 선정 결과
(TREND-2287, 평가함수 A 기준 자질 선정)

〈표 17〉 벡터 자질을 선정할 때 범주별 균형 고려 여부에 따른 성능 비교

범주	크기 (백분율)	전체 문헌 벡터 자질 사용 성능*	20% 문헌 벡터 자질 사용 성능** (괄호 안은 성능차이***)	
			균형 고려 안함	균형 고려함
C1	4.7%	73.4%	73.6%(+0.2%)	71.7%(-1.7%)
C2	4.1%	32.1%	24.6%(-7.5%)	31.0%(-1.1%)
C3	5.0%	76.8%	73.6%(-3.2%)	73.9%(-2.9%)
C4	26.8%	92.7%	92.0%(-0.7%)	92.1%(-0.6%)
C5	39.0%	96.0%	95.8%(-0.2%)	95.1%(-0.9%)
C6	7.2%	73.4%	65.4%(-8.0%)	67.1%(-6.3%)
C7	10.9%	81.5%	73.4%(-8.1%)	77.0%(-4.5%)
C8	2.3%	54.1%	7.7%(-46.4%)	51.4%(-2.7%)
매크로 평균		72.5%	63.3%(-9.2%)	69.9%(-2.6%)

* 성능은 각 범주별 F1척도로 측정함.
** 벡터 자질 선정은 평가함수 A 기준, 자질값은 벡터유사도의 굽이변환(a=0.5) 값임.
*** 성능 차이의 단위는 %포인트임.

이하로 줄이더라도 성능을 향상시킬 수 있는 것으로 나타났다. KFCM-896 실험집단에서는 〈표 14〉와 같이 벡터 자질을 30%만 사용하더라도, 벡터유사도를 자질값으로 사용한 경우(65.73%)와 같은 성능을 얻었다. TREND-2289 실험집단에서는 〈표 16〉과 같이 벡터 자질을 40%만 사용하더라도 벡터유사도를 사용한 경우(87.29%)보다 좋은 성능(87.47%)을 얻었다.

앞에서 범주별 균형을 고려하지 않았던 실험에서는 문헌 벡터 자질을 두 실험집단에서 각각 90% 이상, 60%이상 사용하여야 자질 선정으로 인한 성능 저하를 피할 수 있었다. 범주별 균형을

고려해서 자질 선정한 결과, 두 실험집단에서 각각 30%, 40%까지 자질 집합의 규모를 줄여도 성능이 저하되지 않았다. 이로써 문헌 벡터 자질 선정에 있어서 범주별 균형을 고려하면 자질 집합을 절반 이하로 줄여도 성능 저하를 막을 수 있음이 확인되었다.

균형 자질 선정이 앞서의 〈표 10〉에서처럼 작은 범주의 성능이 심하게 저하되는 현상을 방지하는가를 알아보기 위해서 〈표 17〉을 제시하였다. 이 표를 보면 균형을 고려하지 않고 자질 선정하였을 때 범주의 성능이 -46.4%포인트나 저하되었던 최소규모 범주 C8이, 균형을 고려한 자질 선정에서는 -2.7%포인트 저하되는데 그쳤음을 알 수 있다. 전체적으로 범주별 균형을 고려하지 않은 경우와 달리, 각 범주의 성능이 저하된 정도가 특정 범주에 치우치지 않고 대체로 고르게 나타났다.

4. 5 센트로이드 벡터 자질 사용 실험

벡터 자질 표현 방식이 실용화되기 위해서는 분류 대상 문헌과 비교해야 하는 벡터 자질의 수를 줄이는 것이 관건이다. 앞에서는 문헌 벡터 자질 선정을 통해 이를 해결하는 방안을 모색하였다. 이 장에서는 문헌 벡터가 아닌 범주 센트로이드 벡터를 자질로 사용하는 방법으로 비교 대상을 줄이는 방안을 제시하고자 한다.

범주 센트로이드 벡터는 한 범주에 소속된 모든 문헌의 벡터를 평균한 벡터이다. 대부분의 문헌집합에는 범주의 수가 문헌의 수보다 훨씬 적다. 예를 들어 KFCM-896에서는 범주가 17개, TREND-2287에서는 범주가 8개이므로 각각 학습문헌 수의 2.4%(17/718)와 0.7%(8/1178)에 불과하다. 따라서 비교 대상을 줄이고자 하는 목표는 확실하게 달성된다.

오히려 너무 적은 수의 자질로 문헌을 표현하게 되는 만큼 분류성능이 그다지 좋지 않을 수도 있다. 예를 들면 규모가 크고 주제 범위가 넓은 군집은 센트로이드 벡터에 너무 다양한 문헌이 반영되어 주제가 모호해지므로 자질로서의 자격이 부족할 수도 있다. 센트로이드 벡터 수를 늘리기 위해서 범주의 수를 늘릴 수는 없으므로 다른 방법이 필요하다. 이런 경우에는 큰 범주에 속한 학습문헌을 클러스터링하여 여러 군집으로 나눈 다음, 각 군집의 센트로이드 벡터를 자질로 삼는 것도 가능할 것이다.

이 실험에서는 두 학습집단의 각 분류범주에서 범주 센트로이드 벡터를 각각 구하여 자질로 삼아서 학습문헌과 분류대상 문헌을 표현하였다. KFCM-896 실험집단에서는 센트로이드 벡터 17개와, TREND-2287 실험집단에서는 센트로이드 벡터 8개와 비교해서 벡터유사도 자질값을 산출한 다음 표현된 문헌을 SVM 분류기에 입력하여 학습과 분류 과정을 거쳤다. 실험 결과 분류성능은 KFCM-896 실험집단에서 62.36%, TREND-2287 실험집단에서 86.29%로 나타났다. 이는 모두 문헌 벡터 자질을 이용한 경우보다는 낮지만 색인어 자질을 사용한 최고 성능보다는 각각 1.8%와 2.9% 향상된 결과이다.

4. 6 자질 유형별 성능 종합 분석

이상의 각 실험, 즉 색인어 자질 표현, 문헌 벡터 자질 표현, 범주 센트로이드 벡터 자질 표현 실험의 성능을 종합적으로 비교하기 위해서 〈표 18〉과 〈표 19〉를 제시하였다. 비교 결과 SVM

〈표 18〉 각 자질 유형별 SVM 분류 성능 비교 (KFCM-896 실험집단)

자질 유형	자질 가중치	자질 선정 시 고려사항	전체 자질 사용 성능 (%)	자질 선정 포함 최고 성능 (%)	Ⓐ보다 좋은 성능을 보이는 최소 자질 수*	Ⓑ보다 좋은 성능을 보이는 최소 자질 수*	Ⓒ 이상의 성능을 보이는 최소 자질 수*
색인어	색인어 가중치		56.18 Ⓐ	61.24 Ⓑ	—	—	—
문헌 벡터**	벡터 유사도		65.73 Ⓒ	65.73	—	—	—
	굽이변환 (a=0.5)	범주 균형 무시	68.54	68.54	215(30%)	287(40%)	503(70%)
	굽이변환 (a=0.5)	범주 균형 고려	68.54	69.10	144(20%)	215(30%)	287(40%)
	굽이변환 (a=0.3)	범주 균형 고려	67.42	68.54	144(20%)	144(20%)	215(30%)
센트로이드 벡터	벡터 유사도		62.36	—	17	17	없음

* 괄호 안은 전체 자질 집합 대비 선정된 자질 집합의 크기임.
** 문헌 벡터에서 자질 선정 기준으로는 평가 함수 A를 적용함.

〈표 19〉 각 자질 유형별 SVM 분류 성능 비교 (TREND-2287 실험집단)

자질 유형	자질 가중치	자질 선정 시 고려사항	전체 자질 사용 성능(%)	자질 선정 포함 최고 성능 (%)	Ⓐ보다 좋은 성능을 보이는 최소 자질 수*	Ⓑ보다 좋은 성능을 보이는 최소 자질 수*	Ⓒ 이상의 성능을 보이는 최소 자질 수*
색인어	색인어 가중치		83.86 Ⓐ	83.86 Ⓑ	—	—	—
문헌 벡터**	벡터 유사도		87.29 Ⓒ	87.47	—	—	—
	굽이변환 (a=0.5)	범주 균형 무시	87.38	87.56	236(20%)	236(20%)	707(60%)
	굽이변환 (a=0.5)	범주 균형 고려	87.38	87.56	118(10%)	118(10%)	589(50%)
	굽이변환 (a=0.3)	범주 균형 고려	87.20	87.74	118(10%)	118(10%)	471(40%)
센트로이드 벡터	벡터 유사도		86.29	—	8	8	없음

* 괄호 안은 전체 자질 집합 대비 선정된 자질 집합의 크기임.
** 문헌 벡터에서 자질 선정 기준으로는 평가 함수 A를 적용함.

분류기의 성능은 두 실험집단 모두에서 문헌 벡터 자질을 사용한 경우가 가장 좋았고, 범주 센트로이드 벡터 자질을 사용한 경우가 그 다음, 그리고 색인어 자질을 사용한 경우가 가장 낮았다.

실행 효율을 높이기 위해 문헌 벡터 자질을 일부만 선정하여 사용하는 경우에 범주별 균형을 고려하고 자질 가중치를 변환하는 등의 조치를 추가하였다. 그 결과 문헌 벡터 자질 전체를 사용하는 것 이상의 성능을 낼 수 있는 자질집합의 규모는, KFCM-896 실험집단에서 30%, TREND-2287 실험집단에서 40%로 나타났다. 즉, 절반 이하의 자질만 사용하고도 전체 문헌 자질 벡터를 모두 사용한 것 이상의 성능을 얻을 수 있었다.

또한 색인어 자질을 사용한 경우의 최고 성능보다 더 좋은 성능을 낼 수 있는 자질집합의 규모는, KFCM-896 실험집단에서 20%, TREND-2287 실험집단에서 10%로 나타났다. 즉, 10% 내지 20%의 학습문헌 벡터 자질만 가지고도 기존의 색인어 자질 표현보다 좋은 성능을 얻을 수 있으므로 실용화에 큰 문제는 없다고 판단된다.

범주 센트로이드 벡터를 자질로 사용한 경우는 비록 문헌 벡터 자질을 사용한 경우보다 성능이 낮게 나타나긴 했지만, 매우 적은 수의 자질과 복잡하지 않은 처리만으로도 색인어 자질을 사용한 것보다 성능이 좋게 나타난 점은 실용화 측면에서 고무적인 결과이다.

5. 결 론

SVM 분류기의 성능을 향상시키기 위해서 색인어 자질이 아닌 벡터 자질을 사용하는 방안을 제안하고 이의 실용화가 가능한지를 판단하기 위한 실험을 수행하였다. 문헌 벡터 자질을 이용한 SVM분류기는 색인어 자질을 이용하는 것보다 성능이 뚜렷하게 좋았으나, 벡터유사도 산출을 위한 처리 시간이 오래 걸린다는 단점이 있다. 이 문제를 해결하기 위하여 문헌 벡터 자질 선정 방안과 범주 센트로이드 벡터 자질을 사용하는 방안을 제안하고 실험을 통해 성능을 검증하였다. 실험 결과는 다음과 같다.

첫째, 문헌 벡터 자질을 사용하는 경우에 두 실험집단 모두에서 색인어 자질을 사용하는 것보다 좋은 성능을 얻을 수 있었다. KFCM-896 실험집단에서는 색인어 자질을 사용한 경우에 비해서 17.0%, TREND-2287 실험집단에서는 4.1% 향상된 성능을 얻었다.

둘째, 실행효율 향상을 위해서 문헌 벡터 자질을 일부만 선정해서 사용한 실험결과는 거의 모든 경우에서 문헌 벡터를 모두 사용한 것에 비해 성능 향상을 얻지 못하는 것으로 나타났다. 문헌 벡터를 자질로, 벡터유사도를 자질값으로 사용한 경우에는 자질 선정 결과 분류성능은 뚜렷하게 저하되었다.

셋째, 문헌 벡터 자질을 사용한 경우에도 자질값인 벡터유사도를 분포변환하고 범주별 균형을 고려해서 자질 선정하면 자질 집합의 규모를 상당히 줄여도 성능이 저하되지 않았다. 두 실험집단에서 각각 30%, 40%까지 문헌 벡터 자질 집합의 규모를 줄여도 전체 문헌 벡터를 모두 사용한 것과 동일하거나 더 좋은 성능을 얻었다. 전통적인 색인어 자질 표현을 사용한 경우의 성능과 비교하면 10% 내지 20%의 학습문헌 벡터 자질만 가지고도 더 좋은 성능을 얻을 수 있는 것으로 나타났다.

넷째, 범주 센트로이드 벡터를 자질로 사용한 경우는 비록 문헌 벡터 자질을 사용한 경우보다 성능이 낮게 나타나긴 했지만, 매우 적은 수의 자질과 복잡하지 않은 처리만으로도 색인어 자질을 사용한 것보다 성능이 좋게 나타났다.

이상의 결과로 미루어볼 때, SVM 분류기에서 가급적 높은 성능을 필요로 하는 경우에는 문헌 벡터 자질을 사용하고, 실행 효율을 높여야 하는 경우에는 범주 센트로이드 벡터 자질을 이용하는 것이 바람직할 것으로 판단된다.

이 연구에서 제안된 벡터 자질 표현 방식은, 분류된 학습문헌이 적은 경우에 특히 유용하다. 자동분류 시스템을 도입하려고 검토하면서 분류해야할 문헌은 많이 쌓여있지만 이미 분류해놓은 문헌은 적은 경우가 이에 해당한다. 일반적인 방식에서는 소수의 분류된 문헌만 학습정보로 활용하게 되지만, 이 연구에서 제안된 방식에 따르면 분류되지 않은 다수의 문헌도 자질로 사용할 수 있으므로 시스템 도입 초기의 분류 성능을 높일 수 있다. 향후 연구에서는 Nigam(2001)과 같이 소수의 분류된 문헌과 다수의 미분류 문헌을 가지고 자동분류를 위한 학습을 수행하는 문제에 벡터 자질 표현 방식을 적용해볼 계획이다.

참 고 문 헌

김지영, 장동현, 맹성현, 이석훈, 서정현, 김현. 2000. 한국어 테스트 컬렉션 HANTEC의 확장 및 보완『제12회 한글 및 한국어 정보처리 학술대회 논문집』, 210-215.

정영미, 이재윤. 2001. 지식 분류의 자동화를 위한 클러스터링 모형 연구.『정보관리학회지』, 18(2): 203-230.

정영미, 임혜영. 2000. SVM 분류기를 이용한 문서 범주화 연구.『정보관리학회지』, 17(4): 229-248.

Basu, A., C. Watters, and M. Shepherd. 2003. "Support vector machines for text categorization." *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03).*

Caldas, Carlos H., and Lucio Soibelman. 2003. "Automating hierarchical document classification for construction management information systems." *Automation in Construction*, 12(4): 395-406.

Cristianini, N., and J. Shawe-Taylor. 2000. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods.* Cambridge University Press.

Drucker, H., B. Shahrary, and D. C. Gibbon. 2002. "Support vector machines: relevance feedback and information retrieval." *Information Processing & Management*, 38(3): 305-323.

Dumais, S., J. Platt, D. Heckerman, and M. Sahami. 1998. "Inductive learning algorithms and representations for text

categorization." *Proceedings of the Seventh International Conference on Information and Knowledge Management*, pp. 148-155.

Fukunaga, Keinosuke. 1990. *Introduction to Statistical Pattern Recognition*. 2nd ed. San Diego, CA: Academic Press.

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

Nigam, Kamal. 2001. *Using Unlabeled Data to Improve Text Classification*. Doctoral Dissertation, Computer Science Department, Carnegie Mellon University

Rogati, Monica, and Yiming Yang. 2002. "High-performing feature selection for text classification." *Proceedings of the Eleventh International Conference on Information and Knowledge Management (CIKM'02)*, pp. 659-661.

Salton, Gerard, and Michael J. McGill. 1983. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill.

Taira, H., and M. Haruno. 1999. "Feature selection in SVM text categorization." *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*, pp. 480-486.

Vapnik, V. N. 1995. *The Nature of Statistical Learning Theory*. New York: Springer.

Witten, Ian H., and Eibe Frank. 2000. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann.

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*, pp. 412-420.

Yang, Y., and X. Liu. 1999. "A re-examination of text categorization methods." *Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'99)*, pp. 42-49.