문헌범주화에서 학습문헌수 최적화에 관한 연구

Optimization of Number of Training Documents in Text Categorization

심 경(Kyung Shim)*

초 록

본 연구는 실재 시스템 환경에서 문헌 분류를 위해 범주화 기법을 적용할 경우, 범주화 성능이 어느 정도이며, 적정한 문헌범주화 성능의 달성을 위하여 분류기 학습에 필요한 범주당 가장 이상적인 학습문헌집합의 규모는 무엇인가를 파악하기 위하여 *k*NN 분류기를 사용하여 실험하였다. 실험문헌집단으로15만 여건의 실제 서비스되는 데이터베이스에서 2,556건 이상의 문헌을 가진 8개 범주를 선정하였다. 이들을 대상으로 범주당 학습문헌수 20개(*Tr*-20)에서 2,000개(*Tr*-2000)까지 단계별로 증가시키며 8개 학습문헌집합 규모를 갖도록 하위문헌집단을 구성한 후, 학습문헌집합 규모에 따른 하위문헌집단 간 범주화 성능을 비교하였다. 8개 하위문헌집단의 거시평균 성능은 F_1 값 30%로 선행연구에서 발견된 *k*NN 분류기의 일반적인 성능에 미치지 못하는 낮은 성능을 보였다. 실험을 수행한 8개 대상문헌집단 중 학습문헌수가 100개인 *Tr*-100 문헌집단이 F_1 값 31%로 비용대 효과면에서 분류기 학습에 필요한 최적정의 실험문헌집합수로 판단되었다. 또한, 실험문헌집단에 부여된 주제범주 정확도를 수작업 재분류를 통하여 확인한 후, 이들의 범주별 범주화 성능과 관련성을 기반으로 위 결론의 신빙성을 높였다.

ABSTRACT

This paper examines a level of categorization performance in a real-life collection of abstract articles in the fields of science and technology, and tests the optimal size of documents per category in a training set using a *k*NN classifier. The corpus is built by choosing categories that hold more than 2,556 documents first, and then 2,556 documents per category are randomly selected. It is further divided into eight subsets of different size of training documents : each set is randomly selected to build training documents ranging from 20 documents (*Tr*-20) to 2,000 documents (*Tr*-2000) per category. The categorization performances of the 8 subsets are compared. The average performance of the eight subsets is 30% in F_1 measure which is relatively poor compared to the findings of previous studies. The experimental results suggest that among the eight subsets the *Tr*-100 appears to be the most optimal size for training a *k*NN classifier. In addition, the correctness of subject categories assigned to the training sets is probed by manually reclassifying the training sets in order to support the above conclusion by establishing a relation between and the correctness and categorization performance.

키워드 : 문헌범주화, 텍스트 범주화, 실험문헌집단, 학습문헌집합의 규모, *k*NN 분류기
text categorization, KNN classifier, test collections, size of training documents

* 아이리스닷넷 시스템개발연구소장 (shim@irisnet.co.kr)

■ 논문접수일자 : 2006년 11월 25일
■ 게재확정일자 : 2006년 12월 11일

1. 서 론

문헌범주화란 문헌에 미리 정해진 일련의 범주 중 하나 또는 그 이상을 자동 부여하는 것이다. 문헌범주화의 가설은 사람이 어떤 글의 내용을 충분히 이해하지 않고도 그 글의 범주(즉, 주제)를 쉽게 알아내는 능력을 기계에 학습시켜 분류업무를 자동화할 수 있다는 것이다. 문헌범주화의 역사는 1960년대 초기로 거슬러 올라가지만, 인터넷의 확산과 하드웨어의 발달로 1990년 초부터 정보검색 분야의 주요 과제로 등장하였다. 특히 1980년대 말까지도 문헌범주화의 주류를 이루는 방법은 지식공학(knowledge engineering)에 기반한 것이었으나 1990년대에 들어와 기계학습(machine learning) 접근방식이 주류를 이루고 있다 (Sebastiani 2002).

문헌범주화가 개별 문헌을 특정 범주에 할당한다는 원리적 측면에서는 정보검색의 문헌 클러스링 기법과 유사하게 보이나, 문헌범주화는 학습데이터가 이미 분류되어있으며 분류마다 범주표시(labels)가 결정되어 있어다는 의미에서 지도학습(supervised learning)에 속하고, 문헌 클러스터링은 학습데이터의 분류가 결정되어있지 않다는 의미에서 자율학습(unsupervised learning)에 속하며, 특히 후자는 전자에 비하여 해당 분류에 범주표시가 없는 것이 차이점이다. 또한 정보검색도 문헌을 적합문헌과 부적합문헌으로 구분한다는 점에서는 그 업무가 문헌범주화 업무의 하나라고 볼 수 있다. 그러나 이 둘의 활동을 분류라는 관점에서 보았을 때 차이점은 정보검색이 일반적으로 어느 시점의 특정한 정보요구에 집중하는 반면 문헌범주화는 장기적인 관심분야(long-term interest)에 대한 분류를 다룬다 (Jackson & Moulinier 2002).

문헌범주화 기법의 적용분야는 스팸메일 필터링, 뉴스기사 필터링, 뉴스기사 등의 주제선정(topic spotting), 웹 문헌 분류, 웹 에이전트 (예. WebAce, Iwingz 등), 단어의 중의성 해소 등 다양하며 (정영미 2005), 근래에는 웹 사이트의 계층목록(hierarchical catalog)에 적용되어 인터넷 검색엔진에 질의문을 입력하는 대신 범주의 계층구조를 네비게이트하고 검색은 관심 범주에 제한할 수 있다 (Sebastinai 2002).

1990년대에 시작된 기계학습(machine learning)에 의한 문헌범주화 연구와 실험은 특정 색인기법(text representation), 자질선정(feature selection), 자질추출(feature extraction), 또는 분류기(classifiers) 등과 관련된 특정 처리 기법의 우월성 증명에 초점을 두고 있다. 이와 같은 연구결과 간의 비교가 의미를 갖기 위하여는 실험대상 문헌집단을 포함한 문헌범주화 과정에 관련된 모든 변수들의 통제가 필요하다. 따라서 선행연구들은 비교를 위한 조건통제가 용이하도록 연구실 실험(laboratory experiments)에 치중되었다. 그 결과, 우리는 분류기 학습에 기반이 되는 문헌집단의 속성이 문헌범주화 성능에 미치는 영향과 실제 문헌집단의 문헌범주화 성능에 대하여 극히 제한적인 지식만을 가지고 있다.

본 연구에서는 현실 세계에서 실제 서비스되고 있는 문헌집단을 대상으로 평균 범주화 성능을 평가하고, 문헌집단의 속성 중 학습문헌의 범주당 문헌건수와 범주화 성능과의 관계를

kNN 분류기를 적용하여 실험하였다. 또한, 선행연구에 보고된 수치보다 낮은 평균 범주화 성능이 실험문헌집단에 기부여된 주제범주 정확도와 관련이 있는지 수작업 재분류를 통하여 확인하고, 범주별 성능과 정확도 간의 패턴을 살펴 보았다.

2. 범주당 문헌건수와 범주화 성능

문헌집단의 속성이란 문헌집단을 일반적으로 표현하는 문헌수와 범주 수뿐만 아니라, 이들의 분포까지를 포함하는 정량적 성격은 물론, 문헌범주화 과정의 출발점인 주제범주 부여의 정확성과 같은 정성적 요소까지를 포함한다. 하지만 앞서 지적한 바와 같이, 그 동안 범주화 기법 자체에 대한 연구는 많으나, 실제로 그 대상이 되는 문헌집단의 속성, 또는 이들 속성과 문헌범주화 성능과의 관계를 조직적으로 연구 관찰한 결과는 거의 존재하지 않는다.

문헌집단 속성과 문헌범주화 성능과의 관계 중 우리에게 알려진 것은 첫째, 범주 수의 증가는 범주화 성능에 영향을 미치며 (Apté, Damerau, and Weiss (1994)와 Yang (1996) cited in D'Alessio, Murray, and Schiaffino 1998), 둘째, 범주별 긍정예제가 너무 적으면 결과 해석에 문제가 있고 (Bennett 2003; Lewis, et al. 1996), 분류기에 따라서는 소수부정 분류기(trivial rejector)를 만들며, 셋째, 문헌집단의 속성 중 범주별 문헌수의 분포는 실제 문헌집단에서 전혀 균등하지 않다는 것이다 (Bennett 2003 ; Brank, et al. 2002 ; Lewis, et al. 1996 ; Ruiz, and Srinivasan 2002). 하지만, 이러한 현상들에 대한 설명은 실험결과의 해석 수준이어서 이들을 초래하는 변수 간의 정확한 상관관계를 알 수 없고, 결과적으로 이를 우회하거나 해결할 방법에 대한 이해가 절대적으로 부족하다.

일반적으로 모든 범주화 기법의 성능 연구는 확인되지 않은 범주당 적정 문헌수를 유지하기 위하여 실험문헌집단을 대상으로 두 가지 방법을 사용하였다. 실험문헌집단의 범주 중 범주당 문헌수가 상위에 속한 특정수의 범주만을 실험대상으로 하거나 (Bennet 2003 ; Cai & Hoffman 2004 ; Joachims 1999) 범주당 문헌수가 특정 건수 이상인 범주만을 선정하여 실험하였다 (Brank et al. 2002; D'Alessio et al 1998 & 2000 ; Gao et al. 2003 ; Kotcz, Prabakarmurthi & Kalita 2001 ; Lewis et al. 1996 ; Ruiz & Srinivasan 1999a & 1999b ; Shanahan & Roma 2003). 특히 후자의 경우, 범주당 1건에서 (Gao et al. 2003 ; Kotcz, Prabakarmurthi & Kalita 2001 ; Shanahan & Roma 2003) 최고 1,600건까지 (Brank et al. 2002) 각 연구마다 범주당 적정 건수로 간주된 문헌수는 다양하다.

이러한 선행연구에서 발견되는 학습문헌집합 선별규모의 기준치를 좀더 상세히 살펴보면 , Gao et al. (2003), Kotcz, Prabakarmurthi, and Kalita (2001)와 Shanahan, and Roma (2003)는 범주별 문헌수가 최소한 2건 이상, Weigend, Wiener, and Pedersen (1999)은 최소 16 건 이상, D'Alessio, Murray, and Schiaffino (1998 ;

2000)와 Ruiz, and Srinivasan (1999a)은 최소한 20 건 이상, Lewis et al. (1996)은 문헌수 75 건 이상이며, Brank et al. (2002)은 저자가 관련문헌에서 발견한 가장 많은 수인 범주당 문헌수 1,600 건을 사용하였다. 위 연구 중, Ruiz, and Srinivasan (1999a)이 신경망 분류기를 적절히 학습시키기 위한 최소한의 수치라고 밝힌 것을 제외하고는 이러한 기준치 선정의 이유는 명확하지 않다.

결론적으로 선행연구에서 사용된 범주당 최소 문헌수는 이론에 의하여 결정되기 보다는 경험 또는 짐작에 의하여 추정된 것이다. 이와 같이 분류기의 학습을 위하여 가장 이상적인 범주당 문헌수의 규모가 존재하는지 또한, 존재한다면 몇 건이 가장 적정한 규모인지에 대한 정확한 지식이 없어 범주화 성능 실험결과 간의 비교가 어려울 뿐 아니라, 향후 실험에서도 최적의 성능을 얻기 위한 범주당 최소 문헌수에 대한 지침이 없는 것이 문제점이다.

3. 문헌범주화 실험 설계

3.1 실험 개요

본 연구에서는 실제 서비스되고 있는 문헌집단을 사용하여 평균 범주화 성능을 살펴보고, 범주당 학습문헌수를 점차적으로 증가시키며, 학습문헌규모별 범주화 성능을 비교하여, 최적의 성능을 보이는 범주당 학습문헌수를 실험을 통하여 파악하였다.

대상 문헌의 표현을 위한 용어추출은 전처리 과정을 거쳐 형태소 분석을 한 후, 불용어를 제거하였다. 용어추출의 위치는 초록을 대상으로 수행하였고, 불용어 제거는 SMART 불용어 리스트를 사용하였다.

형태소분석 후, 자질선정을 위하여 추출된 용어는 문헌빈도(DF)의 내림차순으로 정렬하여, DF≥3의 자질만을 대상으로 전역적 자질축소를 실행하였다. 선정된 용어수는 학습문헌집단에 따라 다르나, 축소강도는 평균 74%를 보인다 (표 1 참조). 이 과정을 통하여 선정된

〈표 1〉 학습문헌집단의 추출용어 수 및 자질축소 강도

학습문헌집단명	총 용어 수	DF≥3 용어 수	자질축소 강도
Tr-2000	64,989	16,891	74%
Tr-1500	54,460	14,300	74%
Tr-1000	42,383	11,407	73%
Tr-500	27,649	7,601	73%
Tr-300	19,943	5,569	72%
Tr-100	10,384	2,865	72%
Tr-50	6,908	1,816	74%
Tr-20	3,889	887	77%

* 위 학습문헌집단명에서 *Tr* 뒤에 있는 숫자는 해당 문헌집단의 범주당 학습문헌수를 나타냄.

자질은 TF * IDF를 가중치로 부여하여 문헌 벡터 표현에 사용하였다.

문헌범주화를 위하여 kNN 분류기를 사용하였으며, 본 연구에 사용한 k값은 검증문헌집단을 대상으로 k=1~50까지를 실험하여 최적의 k 값을 선정한 후, 해당 k값의 범주화 성능을 학습문헌집합의 규모별로 실험하였다.

범주화 성능은 전통적 평가척도인 재현율과 정확률로 측정한 후, 범주 전체에 대한 평균을 계산하기 위하여 거시평균(macro-averaging) 방법을 채택하였다. 실험결과의 제시는 재현율과 정확률의 두 가지 값을 비교해야 하는 불편함을 해결하기 위하기 위하여 이들을 통합한 단일척도로 F_1값을 사용하였다. F_1 척도는 정확률과 재현율에 동일한 중요도를 부여한 것으로 다음과 같이 정의된다.

$$F_1 = \frac{2PR}{P+R}$$

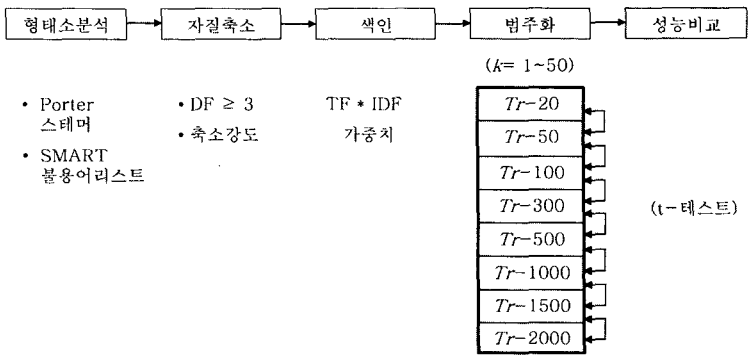
여기서 P는 정확률, R은 재현율을 의미한다.

이 실험에서는 위에 기술한 조건을 적용하여 학습문헌집합의 규모가 제일 작은 Tr-20부터 차 상위 문헌집단 간의 성능비교를 Tr-2000까지 실험하였다 (그림 1 참조).

3.2 실험문헌집단

이 실험에서는 실제 서비스가 되고 있는 한국과학기술정보연구원의 JAFO 데이터베이스를 대상으로 하였다. JAFO는 2004년 현재 과학기술분야의 14백만 건 이상의 기사색인정보를 포함하고 있으며, 초록은 선별적으로 보유하고 있다. 실험결과의 신뢰성과 타당성 확보를 위하여 다음과 같은 실험문헌집단의 선정작업을 수행하였다. 첫째, 대상문헌 중 영어로 구성되었으며, 초록을 보유한 문헌만을 레코드의 '언어' 및 '초록보유' 필드를 점검하여 추출하였고, 둘째, 1차 추출된 레코드를 대상으로 컴퓨터 프로그램을 사용하여 그들이 실제로 영어자료이며 보유한 초록이 영문초록인지를 재확인 하였으며, 셋째, 문헌에 주제범주가 부여되지 않았거나 정의된 범주세트에 명시되지 않은 범주가 부여된 레코드는 제거하였다.

위에 기술한 레코드 선별 및 유효성 점검과



〈그림 1〉 실험문헌 규모에 따른 성능 실험 설계

〈표 2〉 JAFO 중 문헌수 2,556 건 이상의 범주 및 추출된 문헌수

번호	주제코드	주제명	추출문헌수
1	BM0100	의학 : 내과	4,967 건
2	BM0200	의학 : 외과	6,427 건
3	BM0700	의학 : 피부비뇨기과	3,204 건
4	BM0800	의학 : 재활의학	3,439 건
5	BM1000	의학 : 방사선의학	3,161 건
6	BM1100	의학 : 마취과	3,167 건
7	BM1200	의학 : 의학일반	14,485 건
8	ET0204	전자공학 : 고체 디바이스 및 집적회로	2,556 건
총 레코드 수			38,239 건

정을 거쳐 모집단 14,096,672 건의 문헌 중 155, 266건(1.1%)이 실험대상 문헌집단으로 최종 선별되었다.

본 연구에서는 선행연구에서 사용된 학습문헌수를 참고하여 범주별 20건의 문헌수를 시작으로 2,000건까지 단계적으로 8개의 집합을 구성하여 실험하기로 결정하였다. 따라서 위에 기술된 추출문헌집단을 대상으로 범주당 문헌수가 2,556 건 이상의 8개 범주를 선정하였다 (표 2 참조).

동일한 실험조건을 부여하기 위하여 위 8개 범주에서 문헌수가 2,556건인 ET0204를 제외한 나머지 범주에서는 2,556건의 문헌만을 무작위로 골라 실험문헌집단을 구성하였다.

위 8개 범주를 대상으로 하위학습문헌집합을 규모에 따라 8개를 구성하였으며, 이렇게 구성된 8개의 하위문헌집단은 각각 *Tr-20, Tr-50, Tr-100, Tr-300, Tr-500, Tr-1000, Tr-1500*과 *Tr-2000*으로 명명하였다. 문헌집단 명칭에서 *Tr* 뒤에 있는 숫자는 해당 문헌집단의 범주당 학습문헌수를 나타낸다. 이들의 구성방법은 앞서 기술한 바와 같이 범주마다 표 2의 '추출문헌수'를 대상으로 무작위 표본 추출기법을 이용하여 추출된 2,556건 중 먼저 *Tr-2000*을 추출하고, 하위집합들은 *Tr-2000*에서 역시 무작위 표본 추출로 문헌집합 규모를 축소하였다.

먼저 학습문헌집합을 구성하고 개별 범주에 남아있는 문헌을 실험문헌집합으로 사용하였다. 이때, 학습문헌집합과 실험문헌집합은 8대 2의 비율을 유지하였다. 이들 문헌집합의 구성도 범주마다 역시 무작위 표본 추출방식을 사용하였다.

마지막으로 이 실험에서는 kNN 분류기를 사용하므로 먼저 최적의 k 값 계산을 위한 검증문헌집합(tuning set)을 필요로 한다. 이는 개별 범주에서 학습/실험문헌집합을 구성하고 남은 문헌을56 건씩을 역시 학습/ 실험문헌집합이 8 대 2가 되도록 나누었다. 결과적으로 각 범주마다 학습문헌, 실험문헌, 검증문헌의

〈표 3〉 문헌규모에 따른 범주당 학습/실험문헌 및 검증문헌수

번호	문헌집단명	범주당 학습문헌수 (80%)	범주당 실험문헌수 (20%)	범주당 검증문헌수	
				학습문헌	실험문헌
1	*Tr*-20	20 건	5 건	45 건 (80%)	11 건 (20%)
2	*Tr*-50	50 건	12 건		
3	*Tr*-100	100 건	25 건		
4	*Tr*-300	300 건	75 건		
5	*Tr*-500	500 건	125 건		
6	*Tr*-1000	1,000 건	250 건		
7	*Tr*-1500	1,500 건	375 건		
8	*Tr*-2000	2,000 건	500 건		

구성을 위하여 제일 규모가 큰 *Tr-2000* 문헌집단은 총 2,556 건의 문헌이 사용되었다.

이들 8개 실험문헌집단의 구성을 학습/실험 및 검증문헌수 별로 요약하면 〈표 3〉과 같다.

JAFO 데이터베이스의1995년에서 2004년까지 문헌 중, 범주당 2,556 건 이상의 문헌을 보유한 범주를 선별하는 과정에서 조건을 만족하는 범주 수는 8개뿐으로 실험대상 범주의 수가 대규모로 축소되었다. 따라서 적은 수의 범주만을 대상으로 수행된 실험결과의 타당성에 대하여 언급할 필요가 있다.

실험에 포함된 범주 수와 범주화 성능과의 관계를 D'Alessio et al. (1998)는 Apté, Damerau, and Weiss (1994)와 Yang (1996)의 연구결과를 인용하여 현재까지의 경험으로는 범주 수가 증가함에 따라 정확률과 재현율 모두가 감소하는 것으로 나타났다고 한다. 그 이유 중 하나는 실험문헌집단의 범위가 증가하면서 용어들이 점차적으로 다의적(polysemous)이 되기 때문이며, 특히 두문자어(acronyms)에 명백히 나타난다. 이들 두문자어는 3 또는 4자 조합으로 주로 한정되며, 주제분야에 따라 동일한 문자열이 다른 의미로 재사용되기 때문이기도 하다. 그러나 이 실험의 목적은 분류기 또는 범주화 과정에서 거치는 특정기법의 최대성능을 측정하기 위한 것이 아니고, 범주별 문헌수와 범주화 성능과의 관계를 관찰하는 것이 목적이다. 그러므로 이 실험은 사용된 범주 수와는 무관하거나, 관련이 있어도 이 실험은 동일한 문헌집단을 대상으로 규모만을 조정한 것이므로 모든 실험문헌집단에 동일한 영향을 미쳐 이 실험이 목적하는 결과에는 영향을 미치지 않는다.

3.3 범주화 분류기

이 실험에서는 *k*NN 분류기를 사용하였다. 이는 *k*NN 알고리즘이 기본 사상은 매우 단순하고 이해하기 쉬우나, 선행실험에서 매우 성능이 좋은 것으로 보고되었기 때문이다

(Joachims 1998 ; Lam, and Ho 1998 ; Yang and Liu 1999).

kNN 알고리즘의 기본 사상은 신규문헌과 가장 유사한 최근접 문헌을 학습문헌집합에서 찾아 신규문헌에 해당 학습문헌이 가진 것과 동일한 범주를 부여한다는 원리이다. 이 경우 유사한 문헌이 많을수록 범주화 결정이 정확할 수 있으므로 하나 이상의 최근접 문헌을 사용하는 것이 일반적이다.

학습문헌집합에서 신규문헌과 가까운 k개의 최근접 문헌을 찾아내기 위하여는 유사도를 산출하는 유사도 계수를 정의하여야 한다. 이 실험에서는 신규문헌과 학습문헌집합에 있는 문헌과의 유사도 계산을 위하여 다음 코싸인 유사계수공식을 사용하였다.

$$S(d_i, d_j) = \frac{\sum_{k=1}^{t} w_{ki} \cdot w_{kj}}{\sqrt{\sum_{k=1}^{t} w_{ki}^2} \cdot \sqrt{\sum_{k=1}^{t} w_{kj}^2}}$$

위 식에서 d_i는 신규문헌, d_j 학습문헌, w_{ki}와 w_{kj}는 d_i와 d_j의 벡터로 표시되는 가중치를 나타낸다.

kNN을 이용한 범주할당을 위하여 이 실험에 사용된 알고리즘은 신규문헌이 특정 범주에 할당될 조건확률을 고려한 것이다. 이는 각 범주의 적합성 점수를 조건확률의 가중치 합으로 산출하며, 다음 공식으로 표현된다 (Yang 1994).

$$rel(c_k, d_i) = \sum_{d_j \in (k\_top\_ranking\_documents)} sim(d_i, d_j) \times P(c_k \mid d_j)$$

위 식에서 이웃문헌 d_j가 범주 c_k에 할당된다면 $P(c_k \mid d_j)$=1, 그렇지 않으면 $P(c_k \mid d_j)$=0이 된다.

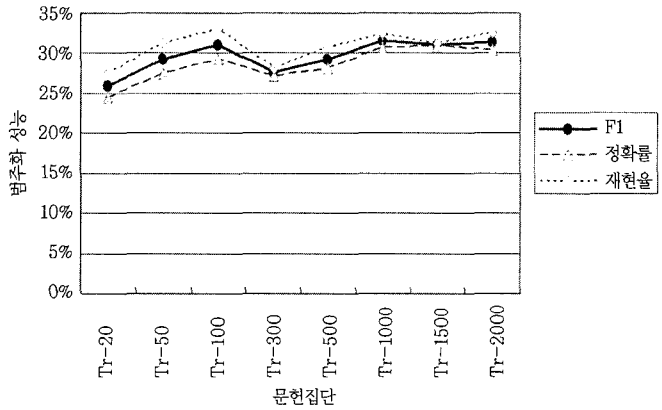
위 공식들을 사용하여 검증문헌집단을 대상으로 k=1~50까지를 실험한 결과, 최적의 k 값은 F_1 척도로 k=13에서 33%로 최대 성능을 보여 k=13을 이 실험의 범주화 성능 평가를 위하여 사용하였다.

4. 실험결과 및 분석

학습문헌집합의 최소 규모인 Tr-20에서 Tr-2000까지 범주당 학습문헌수를 증가시키며 8개의 Tr-문헌집단을 대상으로 k=13, 즉, 최근접 문헌수 13개를 가지고 범주화 성능을 실험하였다. 실험결과는 먼저 8개 범주의 성능에 대한 거시평균을 기준으로 학습문헌수 증가에 따른 범주화 성능 변화를 기술하고 이들을 평균하여 실제 서비스되는 문헌집단의 평균 범주화 성능으로 제시하였다. 이 때 분류기의 성능 변화는 분류기 성능의 최적화에 적절한 학습문헌 규모를 알아보기 위하여 제일 작은 규모인 Tr-20에서 Tr-2000까지의 성능변화를 관찰하였다. 세부 분석으로 학습문헌집단 규모에 따른 성능변화를 개별 범주별로도 제시하였다. 마지막으로 개별 범주에서 학습문헌수 증가에 따른 성능변화가 개별 범주의 기부여된 범주의 정확성과 연관성이 있는가도 학습문헌집합을 수작업 재분류를 한 후 관계를 살펴보았다.

〈표 4〉 *Tr*-문헌집단의 거시평균 범주화 성능

	Tr-20	*Tr*-50	*Tr*-100	*Tr*-300	*Tr*-500	*Tr*-1000	*Tr*-1500	*Tr*-2000	평균
F_1	26%	29%	31%	28%	29%	32%	31%	31%	30%
정확률	25%	28%	29%	27%	28%	31%	31%	30%	29%
재현율	28%	31%	33%	28%	31%	32%	31%	33%	31%



〈그림 2〉 *Tr*-문헌집단의 거시평균 범주화 성능

4.1 실험문헌집단의 평균 범주화 성능

〈표 4〉와 〈그림 2〉에 나타난 범주화 성능은 BM0100, BM0200, BM0700, BM0800, BM1000, BM1100, BM1200, ET0204를 포함한 8개 범주별 성능의 거시평균을 F_1, 정확률, 재현율의 세 가지 성능 척도로 표시한 것이다. 〈표 4〉에서 보듯이, 학습문헌규모별 거시평균을 모두 합하여 평균을 계산한 결과, 본 실험에 사용된 문헌집단의 평균 범주화 성능은 F_1값 30%라는 낮은 수치를 보였다. 이 범주화 성능은 심경과 정영미(2006)가 실험한 동일 모집단에서 추출된 다른 문헌집단의 평균 범주화 성능인 17% 보다는 높지만 표준실험문헌집단을 사용한 선행연구에서 보고된 80~90% 사이의 범주화 성능에는 역시 크게 미치지 못하였다.

4.2 학습문헌집단의 규모에 따른 범주화 실험결과

학습문헌집합의 규모 변화에 따른 범주화 성능을 살펴보면 *Tr*-20에서 *Tr*-2000까지 가장 규모가 작은 *Tr*-20과 모든 규모의 문헌집단, 또는 특정 문헌집단과 차상위 문헌집단 간 성능 비교는 6% 이하의 성능 변화만을 모든 성능 척도에서 보였으며 이들 범주화 성능 변화에 규칙성이 나타나지는 않았다 (표 4과 그림 2 참조).

이들을 개별 척도별로 살펴보면, 〈표 4〉에 제시한 것처럼 F_1의 경우, *Tr*-1000에서 가장 높은 32%를 보였으며, 문헌집단이 *Tr*-20에서 *Tr*-50, *Tr*-500에서 *Tr*-1000으로 변화할 때만 3%씩 향상하여, 가장 큰 폭의 성능 향상을 보였다. 반면 문헌 수 1,000 이상에서는 문헌 수가 증가함에 따라 작은 폭이지만 오히려 성능 저하를 보였다. 다시 말하여, 이 실험에서 기준척도로 삼은 F_1 척도는 *Tr*-20에서 *Tr*-100까지는 5%의 성능 향상을 보이며 증가 추세를 나타냈으나, *Tr*-100에서 *Tr*-2000까지 범주별 학습문헌수가 1,900 건이나 증가되었음에도 불구하고 범주화 성능은 감소하거나 증가해도 그 폭이 아주 작았다.

정확률에서는 *Tr*-1000과 *Tr*-1500에서 각각 최고 성능인 31%의 효과를 보였다. 전체 문헌집단에 대한 거시 평균 정확률 변화의 관찰에서 특이할 점은 문헌집단 간 범주화 성능 변화가 *Tr*-20에서 *Tr*-50, *Tr*-500에서 *Tr*-1000으로 증가할 때만 3%씩의 증가를 보이고, 나머지는 2% 이하의 증감만을 보였으며, 또한, *Tr*-1000에서 *Tr*-1500으로 문헌 수가 증가하였을 때는 전혀 성능 변화가 없었다. 재현율은 *Tr*-100과 *Tr*-2000에서 각각 33%로 가장 높은 성능을 보였으며, 문헌 수 증가에 따른 문헌집단 간 범주화 성능 변화는 *Tr*-100에서 *Tr*-300은 5%의 감소를 보였고, 나머지는 1~3% 범위 내에서 증가하였다.

세 가지 성능 척도 모두를 종합적으로 살펴보면 학습문헌수 100(*Tr*-100)까지는 문헌수의 증가에 따라 범주화 성능이 일괄적으로 향상을 보이는 반면, 학습문헌수를 범주당 200건이나 증가시킨 *Tr*-300에서는 성능이 오히려 감소하고, 그 이후 학습문헌수가 증가함에 따라 다시 범주화 성능 증가를 지속하나 매우 완만한 패턴을 보인다. 다시 말하여, F_1 척도를 기준으로 *Tr*-20에서 *Tr*-100까지는 5%의 성능 향상을 보이며 증가 추세를 나타냈으나, *Tr*-100에서 *Tr*-2000까지 범주당 학습문헌수가 1,900 건이나 증가되었음에도 불구하고 성능향상은 거의 일어나지 않았다.

결론적으로, F_1척도로 거시평균은*Tr*-1000에서 가장 높으나, 좀더 상세히 보면 실제 *Tr*-100과의 범주화 성능 차이는 1%이며, 범주별 문헌수 증가에 따른 성능 변화는 아주 완만하게 나타나고 있음을 알 수 있다. 따라서 이 실험에서 밝히고자 한 최적의 범주화 성능을 보이는 학습문헌 규모는 최고성능 문헌집단만으로는 결론을 내릴 수 없으며, 작은 수의 문헌으로 동일 또는 유사한 범주화 성능을 달성할 수 있다면 작은 수의 문헌규모가 바람직할 것이므로 100개의 학습문헌수가 *k*NN 분류기 학습에는 최적의 학습문헌수로 판단된다.

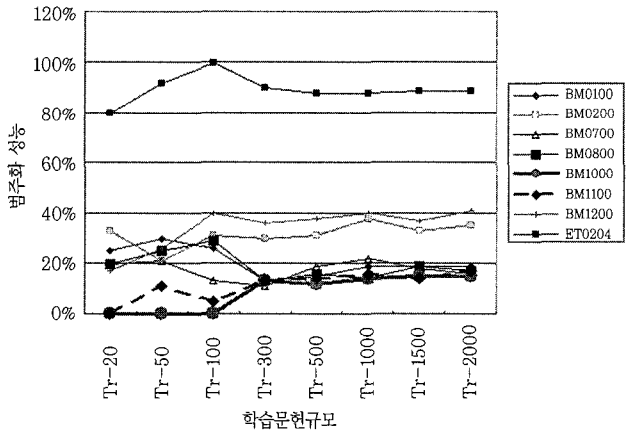
4.3 범주별 *Tr*-문헌집단의 범주화 성능

앞 절에서 관찰된 학습문헌수 증가에 따른 평균 범주화 성능의 변화는 규칙성이 결여되어 현 실험에 사용된 개별 범주 중 특정 범주의 영향이 있는가를 살펴보기 위하여 학습문헌수 증가에 따른 범주별 성능변화를 관찰하였다. 범주별 문헌수 증가에 따른 성능 변화는 F_1 척도만을 사용하여 분석하였다.

〈표 5〉와 〈그림 3〉에 보인 것처럼 개별 범주의 성능변화는 거시평균의 성능변화와는 차

〈표 5〉 개별범주의 학습문헌집합 규모에 따른 범주화 성능 (F_1 척도)

	Tr-20	*Tr*-50	*Tr*-100	*Tr*-300	*Tr*-500	*Tr*-1000	*Tr*-1500	*Tr*-2000	평균
BM0100	25%	30%	26%	13%	15%	19%	19%	19%	21%
BM0200	33%	21%	31%	30%	31%	38%	33%	35%	32%
BM0700	20%	21%	13%	11%	19%	22%	18%	16%	18%
BM0800	20%	25%	29%	13%	16%	14%	19%	17%	19%
BM1000	0%	0%	0%	13%	12%	14%	15%	15%	9%
BM1100	0%	11%	5%	13%	14%	16%	14%	17%	11%
BM1200	17%	26%	40%	36%	38%	40%	37%	41%	34%
ET0204	80%	92%	100%	90%	88%	88%	89%	89%	90%



〈그림 3〉 학습문헌집합 규모에 따른 범주별 범주화 성능

이를 보인다. 개별 범주의 평균 성능면에서 모든 규모의 문헌집단에 대하여 9%에서 32%의 일반적으로 아주 낮은 성능을 보이는 6 범주 (BM0100, BM0200, BM0700, BM0800, BM1000, BM1100)는 문헌 수 증가에 따른 성능 변화에 일정한 패턴이 없으나, 상대적으로 34%와 90%라는 높은 성능을 보이는 2 범주 (BM1200와 ET0204)는 서로 아주 유사한 성능 변화의 일정한 양상을 공유한다. 평균 범주화 성능이 34%~90%로 타 범주에 비하여 상대적으로 높은 BM1200, ET0204의 두 범주는 앞서 내린 결론을 뒷받침해 주고 있음을 확인할 수 있다.

위 결과 중 주목할 점은 선행연구에서 보고된 *k*NN 분류기의 성능과 가장 유사한 성능인 F_1 90%를 보인 ET0204 (전자공학 : 고체디바이스 및 집적회로) 범주는 *Tr*-20에서 *Tr*-100까지 범주화 성능이 명백하게 향상되었지만 그

이후 학습문헌수가 증가함에 따라 성능저하가 나타나고 20배에 달하는 학습문헌수 증가에도 특이할 만한 범주화 성능 증가는 보이지 않는다. 반면 평균 범주화 성능이 9%~18%로 낮은 편에 속하는 BM0700, BM1000, BM1100의 세 범주는 학습문헌수 증가에 따른 범주화 성능 변화에 매우 불규칙한 양상을 보인다. 평균 범주화 성능이 중간에 속하는 BM0100(21%), BM0200(32%)과 BM0800(19%)은 혼합된 양상이기는 하나 일반적으로 높은 성능을 보인 범주들과 유사한 양상을 보여준다.

범주화 성능 변화 중 가장 높은 범주화 성능을 보인 *Tr*-문헌집단만을 범주별로 나열하면 〈표 6〉과 같다. 이는 앞서 〈표 4〉와 〈그림 2〉에 보인 거시평균 범주화 성능 변화에서 본 것과 흡사하게 단편적으로는 학습문헌수가 많은 문헌집단에서 높은 성능이 주로 관찰된다. 하지만 〈표 4〉의 F_1거시평균은 *Tr*-1000에서 가장 높으나, 좀더 상세히 보면 실제 *Tr*-100과 *Tr*-1000 두 문헌집단에서 보인 성능 차이는 1%이며, 범주별 문헌 수 증가에 따른 성능 변화는 아주 완만하게 나타나고 있음을 알 수 있다.

범주별 최고 범주화 성능을 보인 *Tr*-문헌집단을 요약한 〈표 6〉에서 *Tr*-100에서 최대 범주화 성능을 보인 범주는 BM0800, ET0204 2범주뿐이다. BM0200 범주는 *Tr*-20에서 33%의 성능을 보이고 *Tr*-50에서 급격히 성능이 감소한 후 다시 서서히 증가하여 *Tr*-1000에서 최대 성능인 37%를 보이나 성능 변화가 매우 불규칙적이다. BM0700 범주는 *Tr*-1000에서 최대 성능인 22%를 나타내나, *Tr*-50의 21%와는 불과 1% 차이만을 보이며, BM1000 범주는 *Tr*-20, *Tr*-50, *Tr*-100에서 모두 F_1 척도가 0%이며, BM1100 범주는 *Tr*-300 미만의 범주에서 기복이 심하여 정확한 분석이 어렵다. BM1200 범주는 *Tr*-100과 *Tr*-1000이 각각 40%라는 높은 성능을 보이지만 그 사이에 완만한 성능의 증감이 있고, *Tr*-2000에서 최대 성능인 41%를 보인다. 그러나 *Tr*-100과 *Tr*-2000의 성능 차이는 1%일 뿐이다. 따라서 이와 같은 성능변화 추이를 고려하지 않고 〈표 5〉에 나타난 '최고성능 문헌집단'만으로는 결론을 내릴 수 없다.

위 내용을 종합하면, 범주화를 수행하기 위한 시스템 효율 측면에서나 적은 수의 문헌집합으로 분류기를 학습시킬 수 있다는 경제적 측면에서 범주별 최적의 문헌 수는 100건으로 보는 것이 역시 타당하다.

〈표 6〉 *Tr*-문헌집단의 범주별 주제범주 부여 정확도

	전체범주	BM0100	BM0200	BM0700	BM0800	BM1000	BM1100	BM1200	ET0204
평균성능	30%*	21%	32%	18%	19%	9%	11%	34%	90%
최고성능 문헌집단	*Tr*-1000	*Tr*-50	*Tr*-1000	*Tr*-1000	*Tr*-100	*Tr*-1500, *Tr*-2000	*Tr*-2000	*Tr*-2000	*Tr*-100

* 위 값은 개별범주 반올림 이전 수치로 계산된 결과임

4.4 개별 범주의 기 부여된 주제범주의 정확성과 범주화 성능 비교

이 실험에서 BM1000 범주의 낮은 성능과 동 범주의 *Tr*-20에서 *Tr*-100까지 정확률이 0%인 사실과 BM1100 범주의 *Tr*-20에서 0%, 그리고 *Tr*-50에서 *Tr*-100 사이에 나타난 급격한 성능저하 등은 결과에 대한 해석을 어렵게 한다.

실험 대상인 8 범주 중 7 범주는 평균적으로 기존 문헌에 보고된 *k*NN분류기를 사용한 것보다 낮은 범주화 성능을 보였다. 이 낮은 성능은 대상 학습문헌집합의 주제범주 정확성과 연관이 있을 것으로 추정되었다. 즉, 학습문헌집합에 기 부여된 주제범주의 낮은 정확성은 적정한 분류기 학습을 달성할 수 없을 것이고 이와 같이 부적정하게 학습된 범주가 위에 나타난 일관성이 결여된 결과를 보이는 것이 아닌가 하는 추정을 하였다.

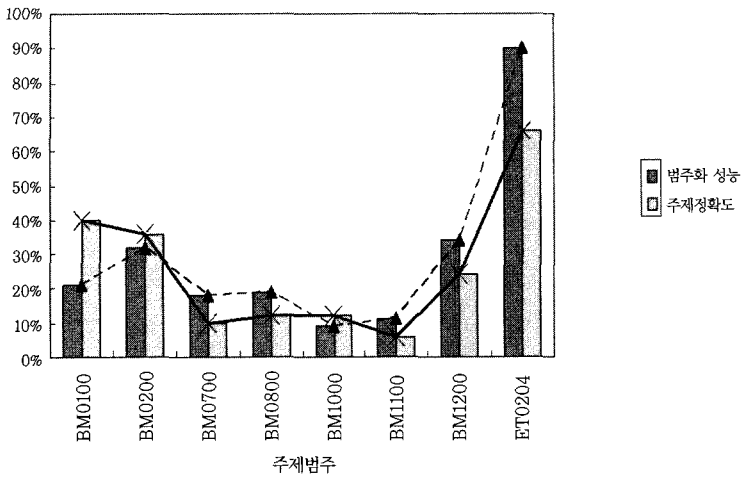
따라서 이 실험에 사용된 문헌집단의 주제범주 부여의 정확성 확인을 위한 추가적인 검증 작업으로 소규모 실험을 하였다. 이 소규모 실험에서 위 8개 범주에서 문헌 50개씩을 무작위로 추출하여 심경, 정영미 (2006)와 동일한 과정을 재현함으로써 전문가의 재분류를 통한 주제범주의 정확성을 확인해 보았다.

재분류 작업 결과, ET0204 한 범주를 제외한 일곱 개 범주의 주제범주 부여 정확도가 50%에 미치지 못하였다 (표 7 참조). 이 결과 중 BM0100 범주의 정확도는 사실은 아래 제시된 것보다 더 낮다고 할 수 있다. 왜냐하면 8개 범주의 재분류 작업 중 부여된 8개 범주의 주제 영역 밖에 속하는 문헌들이 발견되었으며, 특히 이러한 문헌들은 BM0100 범주에 집중되어 있었다. 이들은 주로 면역학, 생리학 등에 관련된 문헌들로, 나머지 7개 범주보다는 BM0100 범주에 가장 가깝다고 판단하여 일단은 옳은 주제범주가 부여된 것으로 간주하였다. 이러한 판단기준이 BM0100 범주의 분류

〈표 7〉 *Tr*-문헌집단의 범주별 주제범주 부여 정확도

주제범주	주제범주 부여 정확도	범주별 범주화 성능
의학 : 내과 (BM0100)	40%	21%
의학 : 외과 (BM0200)	36%	32%
의학 : 피부비뇨기과 (BM0700)	10%	18%
의학 : 재활의학 (BM0800)	12%	19%
의학 : 방사선과 (BM1000)	12%	9%
의학 : 마취과 (BM1100)	6%	11%
의학 : 의학일반 (BM1200)	24%	34%
전자공학 : 고체 디바이스 및 집적회로 (ET0204)	66%	90%
평균 정확도	25%	30%*

* 위 값은 개별범주 반올림 이전 수치로 계산된 결과임



〈그림 4〉 문헌 분류 정확도와 범주화 성능과의 관계

정확도를 실제보다 높였다.

〈표 7〉에 보인 기 부여된 주제범주의 낮은 평균 정확도와 앞 절의 결과를 종합할 때, 대상 문헌집단의 분류 정확도가 이 실험의 전반적으로 낮은 범주화 성능에 영향을 미쳤을 가능성이 크다.

이에 대한 추가적인 확인을 위하여 개별 범주의 분류 정확도와 거시평균 성능과의 관계를 그래프로 나타내면 〈그림 4〉와 같은 상관관계를 보여준다. 앞서 실제보다 높은 분류 정확도를 부여한 BM0100 범주를 제외하면, 분류 정확도와 범주화 성능은 비례하는 것을 볼 수 있다.

세부적으로 앞서 범주화 성능에서 평균치가 낮고 학습문헌 규모 증가에 따른 성능 변화가 매우 불규칙하였던 BM0700, BM1000, BM1100의 세 범주는 역시 분류 정확도가 각각 10%, 12%, 6%로 가장 낮았다. 반면 평균 범주화 성능이 상대적으로 높으며 규칙적인 변화를 보인 BM0200, BM1200, ET0204의 세 범주는 각각 36%, 24%, 66%의 분류 정확도를 보여 앞서 실제보다 높은 분류 정확도를 부여한 BM0100 범주를 제외하면, 가장 높은 분류 정확도를 보였다.

이 결과는 분류 정확도가 낮은 범주의 학습문헌 규모별 범주화 성능 변화가 일반적으로 불규칙하다는 것으로 알 수 있다. 하지만 낮은 분류 정확도가 앞서 언급한 BM1000이나 BM1100범주와 같이 불규칙한 성능 곡선을 보인 원인이었는가는 앞으로 연구가 더 필요하다. 또한 분류 정확도가 높고, 범주화 성능도 높은 ET0204범주의 성능 변화는 위에 결론 내린 최적의 학습문헌 규모인 Tr-100이 가장 적절한 규모임을 확인해 준다.

5. 결론 및 제언

본 연구에서는 실제 서비스 되고 있는 문헌

집단을 대상으로 평균 범주화 성능을 알아보고, 범주별 학습문헌수를 20개에서 2,000개까지 증가시키며 구성한 8개의 문헌집단을 대상으로 범주화 성능을 관찰하여 분류기를 학습시킬 수 있는 최적의 학습문헌 규모를 알아 보았다. 또한 실험 결과 중 일관성이 결여된 내용의 확인을 위하여 소규모 문헌집합을 수작업으로 재색인하여 실험 결과 해석에 대한 신뢰성과 타당성을 부여하였다

본 연구에서 발견한 사실은 다음과 같다.

첫째, 실제 서비스가 되는 JAFO 데이터베이스 중 주로 의학분야의 문헌들을 대상으로 추출한 하위문헌집단에서 평균 범주화 성능은 F_1 값 30%로 동일한 JAFO 데이터베이스에서 추출한 다른 문헌집단에서 얻어진 17% 보다는 높았으나, 선행연구에 보고된 kNN 분류기 범주화 성능에 크게 미치지 못했다.

둘째, 학습문헌집단의 규모를 증가시키며 실험한 분류기 학습에 필요한 최적정 학습문헌수의 거시평균범주화 성능은 Tr-100에서 F_1 값 31%를 보이고, Tr-1000에서 32%를 기록하여 1,000건의 학습문헌집합이 가장 높은 성능을 보였다. 그러나 1%의 성능향상을 위하여 900건의 학습문헌수를 증가시키는 것은 비용대비 효과 면에서 바람직하지 않으며 실험결과 Tr-100 보다 학습문헌수를 증가시켰을 때 성능이 저하하였다가 Tr-1000에서 다시 증가하는 추세를 보여 100건의 학습문헌이 규모 면에서 가장 경제적이며 적정한 범주화 성능을 보이는 것으로 판단되었다. 범주별 학습문헌수 증가에 따른 범주화 성능변화를 관찰한 결과, 평균 범주화 성능 90%를 보인 ET0204의 경우, Tr-100에서 가장 높은 성능을 보인 후 학습문헌수 증가에 따라 성능이 감소하여 위 결론의 신빙성을 높여주고 있다. 또한, 대상 문헌집단을 수작업 재분류를 수행하여 살펴본 결과, 범주별 분석에서 낮은 범주화 성능을 보인 범주들은 일반적으로 낮은 분류 정확도를 보였으며 역으로 높은 범주화 성능을 가진 범주는 역시 높은 분류 정확도를 보였다. 이를 토대로 범주화 성능이 높은 범주의 성능 변화로 확인한 본 실험의 결론에 신빙성을 더하였다.

이 실험은 하나의 분류기 모델을 단일 문헌집단에 적용한 결과이므로 향후 본 실험결과의 일반화를 위하여 다수의 분류기 모델을 여러 문헌집단에 적용하는 것이 필요하다.

참 고 문 헌

김상범, 윤보현, 백대호, 한경수, 임해창. 1999. 문서범주화를 위한 선형분류기와 kNN의 결합모델.『한국인지과학회 춘계학술대회 논문집』, 225-231.

심경, 정영미. 2006. The effect of the quality of pre-assigned subject categories on the text

categorization performance. 『한국정보관리학회지』, 23(2) : 266-285.

이혜원. 2003. 『복합분류기를 이용한 웹 문서 범주화에 관한 실험적 연구』. 석사학위논문, 연세대학교 대학원, 문헌정보학과.

정영미. 2005. 『정보검색연구』. 서울 : 구미무역(주) 출판부.

Apté, Chidnand, Fred Damerau, and Sholom M. Weiss. 1994. "Automated learning of decision rules for text categorization." *ACM Transactions on Information Systems*, 12(3) : 233-251.

Bennett, Paul N. 2003. "Using asymmetric distributions to improve text classifier probability estimates." *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 111-118.

Brank, Janez, Marko Grobelnik, Nataša Milić-Frayling, and Dunja Mladenić. 2002. "Feature selection using linear support vector machines." *Proceedings of the 3rd International Conference on Data Mining Methods and Databases for Engineering.*

Cai, L. and T. Hofmann. 2004. " Hierarchical document categorization with support vector machine." *Proceedings of the ACM 13th Conference on Information and Knowledge Management*, 2004, 182-189.

D'Alessio, Stephen, Keitha Murray, and Robert Schiaffino. 2000. "The effect of using hierarchical classifiers in text categorization." *Proceedings of the 6th Workshop on Very Large Corpora (COLING-ACL '98)*, 66-75.

D'Alessio, Stephen, Keitha Murray, and Robert Schiaffino. 1998. "Category levels in hierarchical text categorization." *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing (EMNLP-3)*, 1-18.

Gao, Sheng, Wen Wu, Chin-Hui Lee, and Tat-Seng Chua. 2003. "A maximal figure-of-merit learning approach to text categorization." *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 174-181.

Jackson, Peter, and Isabelle Moulinier.

2002. *Natural Language Processing for Online Applications : Text Retrieval, Extraction and Categorization.* Amsterdam : John Benjamins Publishing Co.

Joachims, Thorsten. 1999. "Transductive inference for text classification using support vector machines." *Proceedings of ICML-99 : 16th International Conference on Machine Learning*, 200-209.

Joachims, Thorsten. 1998. Text categorization with support vector machines : learning with many relevant features. In *Proceedings of ECML-98, 10th European Conference on Machine Learning* (Chemnitz, Germany, 1998), 137-142.

Kotcz, Aleksander, Vidya Prabakarmurthi, and Jugal Kalita. 2001. "Summarization as feature selection for text categorization." *Proceedings of the 10th International Conference on Information and Knowledge Management*, 365-370.

Lam, Wai, and Chao Yang Ho. 1998. "Using a generalized instance set for automatic text categorization." *Proceedings of SIGIR-98, 21st ACM International Conference on Research and Development in Information Retrieval*, 81-89.

Larkey, L. S., and W. B. Croft. 1996. "Combining classifiers in text categorization." *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 289-297.

Lewis, David D., Robert E. Schapire, James P. Callan, and Ron Papka. 1996. "Training algorithms for linear text classifiers." *Proceedings of SIGIR-96, 19th ACM International Conference on Research and Development in Information Retrieval*, 298-306.

Ruiz, Miguel E., and P. Srinivasan. 2002. "Hierarchical text categorization using neural networks." *Information Retrieval*, 5 : 87-118.

Ruiz, Miguel E., and Padmini Srinivasan. 1999a. "Combining machine learning and hierarchical indexing structures for text categorization." *Proceedings of the 10th ASIS SIG/CR Classification Research Workshop*, 107-124.

Ruiz, Miguel E., and Padmini Srinivasan. 1999b. "Hierarchical neural networks for text categorization." *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval*, 281-282.

Sebastiani, F. 2002. "Machine learning in automated text categori zation." *ACM Computing Surveys*, 34(1) : 1-47.

Shanahan, James. G., and Norbert Roma. 2003. "Boosting support vector machines for text classification through parameter-free threshold relaxation." *Proceedings of the 12th International Conference on Information and Knowledge Management*, 247-254.

Weigend, Andreas S., Erik D. Wiener, and Jan O. Pedersen. 1999. "Exploiting hierarchy in text categories." *Information Retrieval*, 1(3) : 193-216.

Yang, Yiming. 1996. "An evaluation of statistical approaches to MEDLINE indexing." *Proceedings of the AMIA*.

Yang, Yiming. 1994. "Expert network : effective and efficient learning from human decisions in text categorization and retrieval." *Proceedings of SIGIR-94, 17th ACM International Conference on Research and Development in Information Retrieval*, 13-22.

Yang, Yiming, and X. Liu. 1999. A re-examination of text categorization methods. In *Proceedings of SIGIR-99, 22nd ACM International Conference on Research and Development in Information Retrieval* (Berkeley, CA, 1999), 42-49.