한국산학기술학회논문지
Vol. 11, No. 12 pp. 5089-5096, 2010

Okapi BM25 단어 가중치법 적용을 통한 문서 범주화의 성능 향상

이용훈[1], 이상범[1*]
[1]단국대학교 전자계산학과

A Research on Enhancement of Text Categorization Performance by using Okapi BM25 Word Weight Method

YongHun Lee[1] and SangBum Lee[1*]
[1]**Dept. of Computer science, Dankook University**

요 약 문서 범주화는 정보검색 시스템의 중요한 기능중의 하나로 문서들을 어떤 기준에 의해 그룹화를 하는 것을 말한다. 범주화의 일반적인 방법은 대상 문서에서 중요한 단어들을 추출하고 가중치를 부여한 후에 분류 알고리즘에 따라 문서를 분류한다. 따라서 성능과 정확성은 분류 알고리즘에 의해 결정됨으로 알고리즘의 효율성이 중요하다. 본 논문에서는 단어 가중치 계산 방법을 개선하여 문서분류 성능을 향상시키는 것을 소개하였다. Okapi BM25 단어 가중치법은 일반적인 정보검색분야에서 사용되어 검색 결과에 좋은 결과를 보여주고 있다. 이를 적용하여 문서 범주화에서도 좋은 성능을 보이는지를 실험하였다. 비교한 단어 가중치법에는 가장 일반적인 TF-IDF법와 문서분류에 최적화된 가중치법 TF-ICF법, 그리고 문서요약에서 많이 사용되는 TF-ISF법을 이용하여 4가지 가중치법에 따라 결과를 측정하였다. 실험에 사용한 문서로는 Reuter-21578 문서를 사용하였으며 분류기 알고리즘으로는 Support Vector Machine(SVM)와 K-Nearest Neighbor(KNN)알고리즘을 사용하여 실험하였다. 사용된 가중치법 중 Okapi BM25 법이 가장 좋은 성능을 보였다.

Abstract Text categorization is one of important features in information searching system which classifies documents according to some criteria. The general method of categorization performs the classification of the target documents by eliciting important index words and providing the weight on them. Therefore, the effectiveness of algorithm is so important since performance and correctness of text categorization totally depends on such algorithm. In this paper, an enhanced method for text categorization by improving word weighting technique is introduced. A method called Okapi BM25 has been proved its effectiveness from some information retrieval engines. We applied Okapi BM25 and showed its good performance in the categorization. Various other words weights methods are compared: TF-IDF, TF-ICF and TF-ISF. The target documents used for this experiment is Reuter-21578, and SVM and KNN algorithms are used. Finally, modified Okapi BM25 shows the most excellent performance.

Key Words : Text Categorization, Document Classification, TF-IDF, TF-ICF, TF-ISF, Okapi BM25, SVM, Reuter-21578

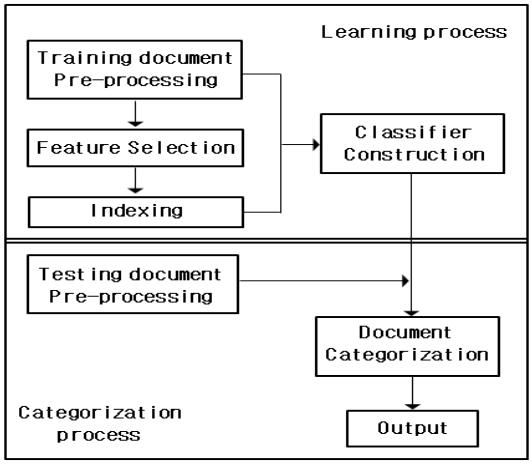
1. 서론

현대 사회에서는 정보의 양이 급증함에 따라 정보자원을 관리하고 효율적으로 다루는 것이 점점 더 중요시 되고 있다. 조직화되고 관리되어 있지 않은 대용량 문서에서 원하는 정보를 획득하기 위해서는 일일이 문서를 비

이 연구는 2010학년도 단국대학교 대학연구비 지원으로 연구되었음.
*교신저자 : 이상범(sblee@dankook.ac.kr)
접수일 10년 10월 21일 수정일 (1차 10년 11월 19일, 2차 10년 12월 09일) 게재확정일 10년 12월 17일

교해 보면서 원하는 정보인지를 확인하는 과정이 필요할 것이며 이는 많은 시간과 비용이 필요하다. 이러한 문제를 해결하기 위한 하나의 방법으로써 문서 범주화 기법을 사용한다. 문서 범주화는 정보검색의 주요 분야로써 문서의 내용에 기반하여 미리 정의되어 있는 범주로 문서를 분류한다. 분류된 문서를 통하여 원하는 정보를 보다 신속하고 정확하게 검색해 내기 위한 것이다. 정보의 효과적인 탐색과 이용을 위한 출발점이 된다. 초기의 문서 범주화 기술은 문서의 범주를 수작업으로 분류 하였으나 급증하고 있는 정보의 양을 해결하기 위해 문서 범주화의 자동화 분야가 활발하게 연구되고 있다[1].

자동화 문서 범주화는 이미 범주화 되어 있는 문서에서 자질을 추출하여 각 범주의 특성을 파악한 후 새로운 문서에 대해서 어떠한 범주로 할당하는 것이 가장 적합한가를 결정한다. 그림 1은 자동화 문서 범주화시스템에 대한 전체 구성도를 보여주고 있다.



[그림 1] 문서범주화 시스템의 전체 구성도

일반적으로 자동화 문서 범주화는 학습 문서로부터 자질을 먼저 추출하고 추출된 자질에 따라 범주화에 대한 정보가 분류기 알고리즘에 의해 학습된다. 학습된 정보를 토대로 테스트 문서에 대한 범주가 결정된다. 분류 데이터로 부터 정확한 자질을 추출하여 표현하는 것과 표현된 자질을 이용하여 적합한 범주로 분류하는 문제는 기계학습 분야의 기본 개념이다. 문서에서 표현된 자질을 기계학습에서 연구된 많은 분류기 알고리즘 중 어떠한 알고리즘을 적용하느냐에 따라 그 성능이 달라지며 또한 이에 대한 연구가 진행되고 있다[2].

본 논문에서는 문서에서 자질의 가중치에 따른 문서 분류의 성능에 관하여 논의한다. Okapi BM25 단어 가중치법은 정보검색분야에서 사용되어 지는 가중치 법으로써 검색 결과에 좋은 성능을 보이는 방법이다. 이 가중치법을 문서 범주화에 적용해 보고 더 나은 성능을 보이는지를 실험하였다. 비교한 단어 가중치법에는 가장 일반적인 TF-IDF법와 문서분류에 최적화된 가중치법 TF-ICF법, 문서요약에서 많이 사용되는 TF-ISF법을 이용하여 4가지 가중치법에 따라 결과를 측정하였다. 논문의 구성은 다음과 같다. 2장 관련 연구에서는 문서 범주화에 대한 관련 연구를 기술하며 3장에서는 관련 연구에서 논의된 기술을 가지고 실험 환경을 구축한다. 4장 실험 결과에서는 3장에서 제시한 실험 방법에 대한 결과를 기술하며 마지막으로 5장에서는 결론과 향후 연구에 대해서 논한다.

2. 관련연구

2.1 자질선택

자질선택이란 문서 분류 과정에서 자질의 개수를 축소하여 분류과정에서 발생하는 계산량을 줄이기 위한 방법이다. 학습문서에서만 추출된 단어의 수는 수십에서 수만까지 추출되어지며 많은 자질의 수는 분류과정에서 많은 시간비용을 요하게 된다. 자질선택을 통하여 단어 중에 중요한 내용어를 선택하여 자질의 개수를 줄이는 동시에 문서분류의 성능 저하 없이 분류 할 수 있는 방법이 자질선택의 목적이다. 문서범주화에서 사용하는 유명한 자질선택 방법은 카이제곱 통계량과 정보 획득량이 있다[3].

카이제곱 통계량(x2 statistics)은 용어 t와 범주 c와의 의존성을 측정하는 방법으로 식(1)은 다음과 같다.

$$x^2(t,c) = \frac{N \times (AD - CB)^2}{(A+C) \times (B+D) \times (A+B) \times (C+D)} \quad (1)$$

A : 범주 c에 속해 있는 문서 중에 용어 t를 포함하고 있는 문서 수

B : 범주 c에 속하지 않은 문서 중 용어 t를 포함하고 있는 문서 수

C : 범주 c에 속해 있는 문서 중 용어 t를 포함하고 있지 않은 문서 수

D : 범주 c에 속하지 않은 문서 중 용어 t를 포함하고 있지 않은 문서 수

N : 학습에 사용된 전체 문서 수

카이제곱 통계량은 용어 t와 범주 c가 완전히 독립적이면 0의 값을 갖는다. 통계량 계산 후 전체 용어에서 정해진 개수만큼 선택하는 방법으로는 식(2)(3)의 두 식 중

하나를 이용하여 최종 선택하게 된다.

$$x^2_{avg}(t) = \sum_{i=1}^{m} \Pr(c_i) x^2(t, c_i) \quad (2)$$

$$x^2_{avg}(t) = \max_{i=1}^{m} x^2(t, c_i) \quad (3)$$

2.2 단어 가중치법

2.2.1 TF-IDF

정보검색과 문서범주화 분야에서 문서로 부터 중요한 단어를 추출하여 문서를 수치화하는 방법들 중 가장 단순하면서 많이 사용하는 가중치법은 TF-IDF 법이다[4]. 각 단어의 가중치를 다음 식(4)로 표현된다.

$$w_{ij} = tf_{i,j} \times idf_i \quad (4)$$

가중치는 단어출현빈도(Term Frequency, tf)와 역문헌빈도(Inverse Document Frequency, idf)의 곱으로 표현되며 단어출현빈도식는 다음 식(5)과 같다.

$$tf_{i,j} = \frac{freq_{i,j}}{\max_l freq_{l,j}} \quad (5)$$

단어출현빈도는 특정 한 문서(i)에서 나타난 단어 중 가장 많은 빈도수(l)와 특정 단어의 빈도수(j)로 나눈 값이다. 이 계산식은 문서 내에서 많은 빈도수를 가진 단어가 문서를 표현하는데 더 적합하다고 가중치 값을 높이는 방법이다. 다음 식(6)은 역문헌빈도의 식이다.

$$idf_i = \log\left(\frac{N}{n_i}\right) \quad (6)$$

역문헌빈도는 특정 단어를 포함하는 문서의 개수(n_i)를 전체 문서의 개수(N)값으로 나눈 값에 로그 값을 취한 것이다. 단어의 출현빈도는 문서의 내부적 영향 값을 나타는 것이라면 역문헌빈도는 문서의 외부적 영향 값을 고려한 값이다.

2.2.2 TF-ICF

Inverted Category Frequency(ICF, 역범주 빈도치)값은 역범주 빈도 가중치 법이라 하며 문서범주화에 특화된 가중치 법이다. 역범주 빈도 가중치법은 문서 간의 분리도가 높은 자질에 더 높은 가중치를 부여하는 방법이다. 즉 소수의 범주에 많이 나온 자질에 대해서는 더 높은 가중치를 주고, 여러 범주에 고르게 나오는 자질에 대해서는 낮은 가중치를 주는 방법이다[5]. 가중치 값은 단어빈도수(TF)와 역범주 빈도수(ICF)를 곱한 값으로 표현되며 ICF값은 아래의 식과 같다.

$$icf_i = \log(M) - \log(CF_i) + 1 \quad (7)$$

여기서 M은 전체 범주의 개수이며 CF_i 값은 색인어 w_i를 포함하는 범주의 개수이다.

2.2.3 TF-ISF

문서요약에서 가장 흔하게 사용하는 단어 가중치법은 $tf-isf$이다. 이 가중치법은 $tf-idf$와 유사하지만 $tf-idf$은 전체 문서에서 하나의 문서를 계산하기 위한 방법이라면 $tf-isf$은 하나의 문서 안에서 문장을 단위로 가중치를 계산하는 방법이다. 그러므로 $tf(w_i, s_j)$가 되며 여기서 w는 단어 가중치이며 s는 하나의 문장이 된다. $isf(w_i)$는 다음과 같다.

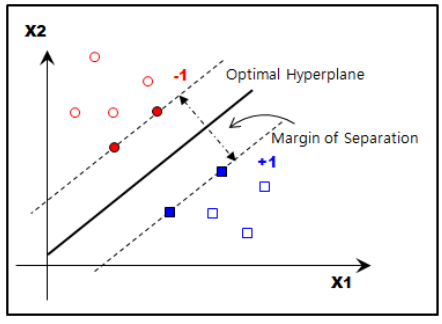
$$isf(w_i) = \log\left(\frac{|S|}{SF(w_i)}\right) \quad (8)$$

여기서 |S|는 하나의 문서 안에서 문장의 개수 이며 SF(wi)는 특정한 단어 wi 가 출현한 개수이다[6].

2.3 분류기 알고리즘

2.3.1 Support Vector Machine(SVM)

SVM은 통계학자 Vapnik에 의해 개발된 분류 기법으로 N차원을 구성하고 있는 x_i 와 각 클래스에 대한 인덱스를 가지고 있는데 y_i를 대상으로 학습을 통해 얻어진 함수 f를 추정하여 x를 {+1, -1} 중 하나로 분류한다. 이때 함수 f는 두 클래스 +1, -1를 선형 분리하는 경계면을 찾게 되는데 이 경계 면적을 최대화 하는 문제를 SVM의 철학이라고 할 수 있다[7]. 그림 2는 간단한 선형 SVM의 예이다.



[그림 2] 선형 SVM 예

그림 2와 같이 두 클래스를 나누는 경계면(Hyperplane)이 존재 하며 그 경계면을 결정하는 데이터들을 Support Vector라 한다. 식 (9)은 SVM의 최종식이다.

$$Maxmimize: Q(\alpha) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} \alpha_i \alpha_j y_i y_j (X_i^T \cdot X_j)$$

$$subejctto: \begin{cases} \sum_{i=1}^{N} a_i y_i = 0 \\ 0 \le \alpha_i \le C, \quad i = 1,2,..., N \end{cases} \tag{9}$$

파라미터 C는 오분류와 성능 간의 상충관계(trade-off)를 나타내는 비용(Cost) 변수이다. 몇몇 데이터는 올바른 클래스 영역에 포함되는 것이 아닌 다른 클래스 영역에 포함될 수 있도있으며 이러한 문제점을 해결하기 위하여 C값을 사용한다[7]. C의 값이 크면 오분류 오차가 적어지며, C의 값이 작다면 최소거리가 최대화되어 풀이의 복잡도는 낮아게 된다.

SVM은 선형 분리가 불가능한 경우 고차원의 특징 공간으로 이동시킴으로써 비선형 분류 문제를 선형 모델로 구현 가능하게 한다. 이러한 함수들을 커널 함수라 하며 다항식 커널, 가우시안 RBF 함수 등이 있다.

다항식 커널 함수 : $K(\mathrm{x}, \mathrm{x}') = (\mathrm{x}\,\mathrm{x}' + 1)^d$

가우시안 RBF :

$$K(\mathrm{x}, \mathrm{x}') = \exp(-1/\delta^2(\mathrm{x} - \mathrm{x}')^2) \tag{10}$$

여기서 d는 다항식 커널의 차수이고, δ^2 는 가우시안 RBF 커널의 대역폭이다.

2.3.2 k-Nearest Neighbor(KNN)

KNN은 패턴인식분야에서 가장 잘 알려진 분류기로써 문서 범주화 연구가 시작된 초기부터 많이 사용되어지는 알고리즘이며 좋은 성능을 가진 분류기중 하나이다[8]. KNN은 비교적 단순한 알고리즘으로써 학습문서 중에서 테스트 문서와 유사도가 가장 높은 k개의 문서를 추출하고 추출된 문서를 비교하여 특정 범주로 분류하는 방법이다.

$$sim(D_x, D_j) \frac{\sum t_{xk} \times \sum t_{jk}}{\sqrt{\sum (t_{xk})^2} \times \sqrt{\sum (t_{jk})^2}} \tag{11}$$

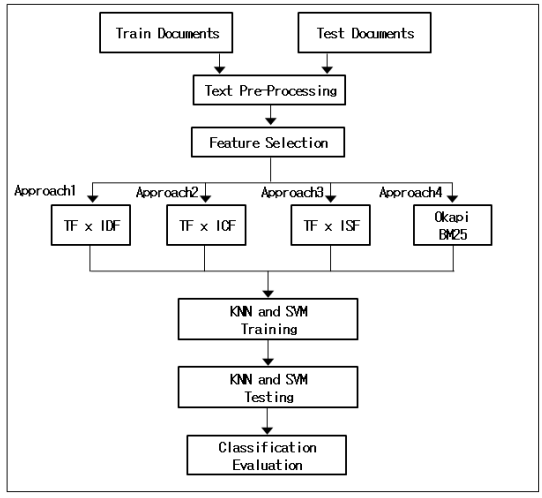
위 공식은 코사인 유사도법을 이용한 KNN 공식이다. D_j는 학습 문서이고 D_x가 입력문서 일 때의 유사도를 측정하는 방법이다. 여기서 t_{xk}는 D_x에 출현한 용어 k의 가중치이며 t_{jk}는 D_j에 출현한 용어 k의 가중치이다.

KNN에서 k값은 위의 공식에서 특정 D_x 테스트 문서에서 가장 유사한 D_j의 개수를 k 만큼 선택하게 된다. k개 만큼 선택되어진 유사한 문서 D_j 중에서 가장 많이 나타난 범주로 D_x의 문서가 분류되어진다.

KNN알고리즘은 SVM알고리즘과 함께 문서분류에서 가장 많이 사용되는 분류기알고리즘이며 SVM처럼 문서를 하나의 공간에 표시하면서 범주의 영역을 나누는 복잡한 방법이 아니므로 비교적 간단하게 구현 가능한 알고리즘이다. 그러나 KNN은 각 입력문서에 대해서 모든 문서를 비교해야 하기 때문에 수행속도가 느리다는 단점을 가지고 있다.

3. 범주화 실험

본 논문의 범주화 실험 과정은 그림 3과 같다.



[그림 3] 범주화 실험 아키텍처

3.1 실험 범위

본 범주화 실험에서 논하고자 하는 부분은 그림 3의 4가지의 방법(Approach)들이다. 범주화의 일반적인 성능 향상은 [9]에서 논하는 문서집합의 특성에 따른 범주화 성능에 관한 논문, 또는 [10]에서 연구한 신경망 알고리즘의 개선에 따른 성능 개선이며, 또한 [3]에서의 자질축소 방법에 따른 범주화 성능 개선을 연구한다.

본 논문은 [5]에서 연구한 단어 가중치법에 따른 범주화의 성능을 평가하는 방법과 유사하다 정보검색에서 질의어 가중치로 사용한 Okapi BM25법은 검색 결과에 좋은 성능을 보였으며 이 가중치법이 문서를 표현하는데 더 적합한 방법이라면 범주화 실험에서도 좋은 성능을 보일 것이라고 판단하여 실험해 보고자 하였다.

3.2 실험 문서 선택

실험 문서로 사용한 Reuter-21578 말뭉치는 1987년 Reuter 뉴스 통신에서 사용했던 문서를 모아 놓은 것으로 범주화 실험에서 많이 사용되어지는 말뭉치이다[11]. Reuter-21578문서 중 가장 많은 문서를 포함하는 10개의 클래스를 선택했다. 표 1은 선택한 클래스 종류와 문서들의 개수를 나타내고 있다.

각 클래스당 문서 수는 학습문서 100개, 테스트 문서 50개로 총 150개의 문서이며 500개의 클래스가 없는 문서를 추가하였다. 이는 잡음 데이터를 주기 위해서 추가하였다. 총 2000개의 문서에 대해서 실험하며 이중 500개가 테스트 문서로 사용되었다.

[표 1] 실험에 사용한 문서 정보

클래스이름	학습문서수	테스트문서수	합 계
acq	100	50	150
crude	100	50	150
earn	100	50	150
grain-wheat	100	50	150
interest	100	50	150
money-fx	100	50	150
money-supply	100	50	150
shop	100	50	150
sugar	100	50	150
trade	100	50	150
NO CLASS	-	-	500
총 합	1000	500	2000

3.2 문서 전처리

본 과정은 문서에서 단어들을 추출하고 불용어 제거 및 스테밍 처리를 한다. 불용어 제거란 영어 단어 a, the, that 과 같이 문서를 표현하는데 특별한 의미를 지니지 않으면서 많은 문서에서 공통으로 나타나는 단어를 말한다[12]. 스테밍 처리는 단어의 어원을 복구하는 작업이다. 영어의 경우 'computing', 'computed', 'computational' 또는 'computer'와 같이 하나의 어원에서 여러 가지 어미에 따라 품사가 달라지는 단어들이 존재한다. 이러한 단어들의 어원을 복구하는 작업을 스테밍 처리라 하며 본 실험에서는 Porter Stemmer 알고리즘이 있다[13]. 표 3은 불용어 목록표이다.

[표 2] 불용어 목록표

a, an, and, are, as, at, be, but, by, for, if, in, into, is, it, no, not, of, on, or, s, such, t, that, the, their, then, there, these, they, this, to, was, were, will, with

문서에서 불용어 목록표에 해당하는 단어들을 제거하며 제거 후에는 스테밍 처리를 한 후 최종적으로 단어들을 저장한다. 본 실험에서 전처리 과정 후 자질로 선정되는 단어의 수는 9331개로 추출되었다.

3.3 자질선택

본 실험에서 자질선택 방법은 카이제곱 정보량을 사용하여 실험하였다. 계산된 정보량 후 실험에서 사용될 단어들을 고르는 방법으로는 식(12)을 사용 하였다.

$$x^2_{\max}(t) = \max_{i=1}^{m} x^2(t, c_i) \tag{12}$$

자질선택의 개수는 250개 단위로 증가 시키며 결과 값을 측정하였다.

3.4 단어 가중치법

본 논문에서 제안하는 정보검색분야에서 사용된 좋은 단어 가중치법을 문서 범주화에서도 적용하여 범주화 결과에 좋은 성능을 보이는지를 위한 방법으로 Okapi BM25법을 문서 범주화에서 적용해 보았다.

포아송 모델-2의 확률적 모델을 바탕으로 하는 Okapi BM25법은 정보검색분야에서 검색의 결과인 성능을 높이기 위한 방법으로 검색 시스템 사용하는 Ranking function 중의 하나로 질의에 대하여 일치하는 문서들을 순위화 하는 방법이다[14]. 각 문서에서의 단어 빈도수를 이용하여 확률적 모델을 적용하는 계산법을 이용하며 수식은 식(13)과 같다.

$$score(D,Q) = \sum_{i=1}^{n} idf(q_i) \frac{f(q_i, D)(k_1 + 1)}{f(q_i, D) + k_1 (1 - b + b \frac{|D|}{avg\,dl})} \tag{13}$$

단어 q1, ... ,qn 를 포함하고 있는 질의 문서 Q에 대한 문서 D의 식이다. 여기서 f(qi , D)는 문서 D에서 단어 qi 가 등장한 빈도를 나타내며 | D |는 문서 D의 단어 개수를 의미하며 avg dl 은 비교 대상 문서 군의 평균 단어 개수를 나타낸다. k1과 b는 자유 파라미터이다.

또한 Okapi BM25에서는 역문헌빈도수에도 변형된 방법을 사용한다. 식(14)는 Okapi BM25에서의 역문헌빈도수 식이다.

$$idf(q_i) = \log \frac{N - n(q_i) + 0.5}{n(q_i) + 0.5} \quad (14)$$

Okapi BM25와 같이 문서에서 중요한 단어를 찾는 법을 단어출현빈도와 역문헌빈도를 적절한 혼합한 방법과 파라미터를 설정함으로써 새로운 단어 가중치법이 적용되었다.

자질축소 과정이 끝나면 단어들을 가중치를 부여하는 과정을 거치게 된다. 본 실험에서는 TF-IDF, TF-ICF, TF-ISF 그리고 Okapi BM25 가중치법을 4가지를 사용하였다. 여기서 Okapi BM25 가중치의 파라미터 값으로 k 파라미터에 2를 설정하였으며 b 파라미터에는 0.75를 설정하였다. 다음 식(15)는 본 실험에서 사용된 Okapi BM25 가중치법의 공식이다.

$$w_{ij} = idf_{ij} \times \frac{tf_{ij} \times 3}{tf_{ij} + 2 + (1 - 0.75 + 0.75 \times \frac{dl}{avg\,dl})} \quad (15)$$

자질선택의 개수에 따른 4가지 가중치법에 대한 실험결과를 측정하였다.

3.5 분류기 알고리즘를 이용한 문서 학습 및 테스트

분류기를 이용한 분류 학습과 테스트는 KNN알고리즘과 SVM알고리즘을 이용하여 실험하였다. KNN알고리즘에서 k 값은 30으로 실험 결과를 측정하였으며 유사도 측정 방법으로는 코사인유사도를 사용하였다.

SVM 분류기는 타이완국립대학에서 개발한 LIBSVM 2.91 버전을 사용하였다[15]. 실험에서 사용한 파라미터로는 C값을 10으로 설정했으며 커널함수로는 가우시안 RBF 함수를 사용하였다. 먼저 학습문서 1000개의 대해서 학습을 시킨 후 테스터문서 500개 대해서 결과를 측정하였다.

3.6 분류 평가

정보검색과 문서 범주화 분야에서 실험에 대한 평가 방법으로는 정확률과 재현률 그리고 이들 둘을 결합한 방법으로 F-측정률 값이 있다[16].

[표 3] 부류에 대한 분류결과 경우의 개수

부류		분류기의 분류	
		속함	속하지 않음
실제분류	속함	a	b
	속하지 않음	c	d

표 3에 따라 정확율, 재현율 그리고 F-측정률을 표시하면 다음과 같다.

$$\text{정확율} = \frac{a}{a+c} \qquad \text{재현율} = \frac{a}{a+b}$$

$$F\text{-측정률} = \frac{2(\text{정확률} \times \text{재현율})}{\text{정확률} + \text{재현율}} \quad (16)$$

본 실험에서는 F-측정률을 가지고 실험 결과를 표현하였다.

4. 실험결과

4.1 분류 결과

Reuter-21578문서에서 10개의 클래스를 선정 후 학습문서 1000개와 테스트 문서 500개를 구성하였다. 자질선택 방법으로는 카이제곱 정보량을 사용하였으며 4가지 가중치법을 사용하였다. 이후 KNN알고리즘과 SVM알고리즘을 이용하여 문서를 분류한 결과는 표 4와 표 5과 같다.

[표 4] SVM 알고리즘를 이용한 F-측정률

자질개수	가중치법			
	TF-IDF	TF-ICF	TF-ISF	Okapi
250	0.79	0.81	0.54	0.82
500	0.77	0.81	0.58	0.84
750	0.79	0.81	0.57	0.84
1000	0.79	0.81	0.56	0.84
1250	0.79	0.81	0.56	0.84
1500	0.79	0.80	0.56	0.85
1750	0.80	0.81	0.57	0.84
2000	0.80	0.81	0.55	0.85

표 4에서는 SVM알고리즘을 이용한 자질선택에 따른 F-측정률값이다. 이 표에서 알 수 있듯이 Okapi 가중치 값이 가장 좋은 성능을 보였으며 그 다음으로 TF-ICF순으로 이어진다. TF-ISF값은 0.5대의 낮은 분류률을 보였다.

다음으로 KNN알고리즘을 이용한 자질선택에 따른 F-측정률 분류 결과의 표이다.

[표 5] KNN 알고리즘를 이용한 F-측정률

자질개수	가중치법			
	TF-IDF	TF-ICF	TF-ISF	Okapi
250	0.89	0.88	0.82	0.90
500	0.88	0.87	0.83	0.90
750	0.90	0.88	0.80	0.88
1000	0.89	0.89	0.82	0.89
1250	0.88	0.89	0.82	0.89
1500	0.88	0.88	0.82	0.88
1750	0.87	0.89	0.80	0.89
2000	0.87	0.89	0.80	0.88

표 5은 비교적 분류 결과가 비슷하게 나왔으나 근소한 차이로 Okapi BM25가중치 값이 가장 좋은 성능을 보이는 보였다. KNN알고리즘을 이용하고 자질선택의 개수가 250, 500개 일 때 Okapi BM25의 가중치법이 0.90을 나타내며 실험 과정에서 가장 좋은 분류 결과를 보였다.

전체적인 실험 결과의 성능은 Okapi BM25, TF-ICF, TF-IDF, TF-ISF순으로 이어진다. 가장 성능이 낮게 나온 TF-ISF 가중치법은 외부영향이 아닌 한문서의 내부적영향만을 가중치법으로 사용하기에 문서 분류를 위한 가중치법으로 적합하지 않으며 가장 낮은 성능 결과가 나타났다.

4.2 실험 결과에 따른 비교

본 실험에서 분류 범주에 더 높은 가중치를 부여하는 법을 소개한 [5]논문의 TF-ICF 가중치법 보다. 정보검색에서 질의어에 가중치를 부여하는 Okapi BM25 가중치법이 더 좋은 성능을 보이는 것으로 나타났다.

TF-ICF법은 TF-IDF법를 변형한 방법으로 역문헌빈도수를 역범주빈도치로 계산하여 범주화 성능을 개선한 것이라면 본 연구는 포아송 모델-2를 기반으로 한 확률모델을 적용한 Okapi BM25 방법이 범주화에 적용한 것이며 기존 연구보다 좋은 성능을 나타냈다.

5. 결론 및 향후연구

본 논문에서는 문서 범주화의 성능을 높이는 방법에 대해서 논하였다. 대부분의 문서 범주화의 연구는 자질축소 방법과 분류기 알고리즘의 특성에 따른 분류기 선택 문제로 연구되고 있다. 본 논문에서는 범주화에 있어서 가장 근본적인 문제라고 할 수 있는 문서의 자질을 어떻게 더 잘 표현하는가에 대한 문제에 대하여 연구하였다. Okapi BM25 단어 가중치법은 정보검색분야에서 사용되며 검색 결과에 좋은 결과를 보이는 방법이며 문서 범주화에서도 더 나은 성능을 보이는지를 실험하였다. 비교한 단어 가중치법에는 가장 일반적인 TF-IDF법와 문서분류에 최적화된 가중치법 TF-ICF법, 문서요약에서 많이 사용되는 TF-ISF법을 이용하여 4가지 가중치법에 따라 결과를 측정하였으며 Okapi BM25 법이 가장 좋은 성능을 보였다.

앞으로의 향후 과제로는 4가지 단어 가중치법을 이용하여 많은 다른 분류기에 적용해 보는 실험을 통해서 실험 결과의 정당성을 높여야 할 것이며 또한 많은 데이터 문서를 사용해서 증명하는 것도 필요할 것이다.

문서 범주화의 기술은 정보의 양이 급증하는 현 시대에서 꼭 필요한 기술이다. 문서에서 특징이 되는 자질을 추출하고 추출한 자질을 축소하며 알맞은 분류기를 선택하는 문서 범주화의 연구는 앞으로도 많은 부분에서 발전이 필요하다. 지금까지 많은 방법과 알고리즘이 나왔지만 문서라는 정해진 틀이 아닌 자유롭게 작성 가능한 문서정보에서 그 정보를 분류하고 검색하는 문제는 계속될 것이다.

참고문헌

[1] Sebastiani. “Machine learning in automated text categorization.” Techinical report, Consigilo Nazionale delle Rieche, Italy. 1999.

[2] T.Mitchell. “Machine Learning.” MCGraw Hill, NY, US, 1996.

[3] Yang, Y. and J. O. Pderson. “A comparative study on feature selection in text categorization." Proceedings of the 14th International Conference on Machine Learning. 1997.

[4] Gerard Salton and Michael J. McGill. “Introduction to Modern Information Retrieval.” McGraw-Hill Book Company, New York, 1983.

[5] 조광제, 김준태. “역카테고리 빈도에 의한 계층적 분류체계에서의 문서의 자동분류." 한국정보과학회 봄 학술발표논문집(B), 507-510. 1997.

[6] Larocca Neto, Joel. “A Text Mining Tool for Document Clustering and Text Summarization.", Proceedings of The Text Mining Tool for Document Clustering and Text Summarization Fourth International Conference on The Practical Application of Knowledge Discovery and Data Mining, 41-56.

Manchester, UK. Apr, 2000.
[7] Osuna, E., Freund R., and Girosi, F. “Training support vector machines: An application to face detection", Proceedings of Compuer Vision and Pattern Recognition, pp. 130-136, 1997.
[8] Dasarathy, Belur V. “Nearest Neighbor(NN) Norms: NN Pattern Classificatioin Techniques." McGraw-Hill Computer Science Series, CA: IEEE Computer Society Press. 1991.
[9] 리청화, “BPNN의 효율적인 개선방법 및 개념에 기초한 문서분류 시스템 응용” 전북대학교 대학원 박사논문. 2009
[10] 정은경, ”문서 범주화 효율성 제고를 위한 정보원 평가에 관한 연구 ”, 한국정보관리학회, 제24권, 제4호, pp. 305-321, 12월, 2007.
[11] David D. Lewis. “Distribution 1.0 readme file (v1.2) for reuters-21578”, AT&T Labs - Research, 1997.
[12] GSalton, “Automatic Information Organization and Retrieval.” New York:McGraw-Hill, 1968.
[13] M. F. Porter. “An algorithm for suffix stripping.” Program, Vol. 14 no.3 130-137. 1980.
[14] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford. (1994) “Okapi at TREC-3”. In Proceedings of the Third Text REtrieval Conference (TREC 1994).
[15] Chin-Chung Chang and Chih-Jen Lin, LIBSVM: a library for SVM, URL : http://www.csie.ntu.edu.tw/~cjlin/libsvm
[16] D.D.Lewis, “Evaluating text categorization”, in Proceedings of the Speech and Natural Language Workshop, 1991.

이 용 훈(Yong-hoon Lee) **[준회원]**

- 2009년 2월 : 단국대학교 컴퓨터과학과 (공학사)
- 2009년 3월 ~ 현재 : 단국대학교 전자계산학과 석사과정

<관심분야>
정보 검색, 데이터 마이닝, 기계학습

이 상 범(Sang-Bum Lee) **[정회원]**

- 1989년 12월 : 루우지애나주립대 (전산학석사)
- 1992년 8월 : 루우지애나주립대 (전산학박사)
- 1992년 9월 ~ 1993년 10월 : 전자통신연구원 선임연구원
- 1993년 10월 ~ 현재 : 단국대학교 교수

<관심분야>
소프트웨어공학, 정보검색, 모바일컴퓨팅