Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.2 pp. 167~175

Korean Document Classification using Characteristics of Word Information

Seok Ki Kim[1] · Kyung Soo Han[2] · Jeong Yong Ahn[3]

Abstract

In document classification, target of analysis is not document itself but words appeared in the document. Word information, therefore, is a significant factor in document classification. In this study, we are dealing with the classification of Korean document based on words and feature vectors. First, we present the performance of document classification using nouns and keywords. Second, we compare to the results for the size of feature vectors.

Keywords : 한글 문서 분류, 특성 정보, 특성 벡터, 문서 분류 알고리즘

1. 서론

네트워크 기술의 급격한 발전과 인터넷의 대중화로 인해 전자 문서의 생성이 기하급수적으로 증가함에 따라 문서의 분류(document classification)에 대한 필요성이 널리 인식되고 있으며, 많은 연구들이 진행되고 있다(강현규,박세영, 1998; 한광록 등, 2000; Cunningham *et al.*, 1997; Rijsbergen, 1979). 문서의 분류는 정해진 분류 체계(일반적으로 문서에 출현하는 단어 정보를 이용)하에서 분류하고자 하는 문서들을 가장 적합한 범주(category)에 배정함으로서 문서를 집단화하는 작업으로 정의할 수 있다. 따라서 문서 분류에서 분석의 대상은 문서 자체가 아니라 문서에 포함된 단어이며, 문서에 포함된 단어 정보를 분석 가능한 수치 자료 형태로 표현하는 것이 문서 분류를 비롯한 문서 기반의 분석 기법들이 수행해야 할 첫 번째 절차이다. 즉, 단어 정보를 어떻게 표현하느냐가 문서의 분류 결과에 매우 큰 영향을 줄 수 있다.

1) Graduate School, Department of Computer Science and Statistis, Chonbuk National University, Chonbuk, 561-756, Korea
E-mail : sisyphus@mail.chonbuk.ac.kr
2) Professor, Division of Mathematics and Statistical Informatics, Chonbuk National University, Chonbuk, 561-756, Korea
3) Assistant Professor, Department of Computer Science and Statistics, Seonam University, Chonbuk, 590-711, Korea

영어 문서의 경우 문서 분류 알고리즘에 적절한 단어 정보의 구성은 물론 분류에 영향을 주는 특성 선택(feature selection)과 특성 추출(feature extraction), 단어 군집(word clustering)의 구성 부분에 많은 연구들이 이루어지고 있다. Yang과 Pedersen (1997)은 특성 벡터(feature vector)를 구성하는 특성 선택에 있어서 document frequency, information gain, mutual information, χ^2-test, term strength의 5가지 방법을 k-nearest-neighbor와 linear least square fit에 적용하여 특성 벡터의 크기에 따른 정밀도(precision)를 측정하였다. Joachims(1997)는 유즈넷 뉴스그룹 문서와 Reuters 문서 집합에 대하여 특성 벡터의 크기에 따른 정확률(accuracy)에 대하여 연구하였으며, Lewis(1992)는 Reuters 문서 집합과 MUC-3 문서 집합을 이용한 문서 분류에서 특성 벡터의 차원이 10~15에서 최적의 결과를 보여줄 뿐만 아니라 단어, 단어 군집, 어구(phrase), 어구 군집 중 단어를 이용하여 특성 벡터를 구성하는 것이 높은 정확률과 정밀도를 제공하는 것을 보였다(정확률과 정밀도에 대한 정의는 식 (4.1)과 (4.2) 참조).

한편 국내의 연구들은 영어 문서를 이용하는 형태와 문서 분류 알고리즘을 한글 문서에 적용하는 형태를 취하고 있다. 김진상과 신양규(2000)는 유즈넷 뉴스그룹 문서에 대하여 Naive Bayes를 이용하여 특성 단어의 수에 따른 분류 실험을 실시하였으며, 이윤오와 이정진(2000)은 TREC6 자료의 boolean 변환 결과에 대하여 LAD(logical analysis of data)를 적용한 확률화 정보검색 모형에 대하여 연구하였다. 한글 문서를 다루는 경우에 있어서도 대부분의 연구들은 영어 문서에 대한 연구들을 통하여 얻어진 결과들을 분류를 위한 특성의 선택과 특성 벡터의 구성에 사용하고 있다. 그러나 이러한 접근 방법은 기본적으로 영어 문서에서의 단어와 한글 문서에서의 단어의 특성이 동일하다는 것을 가정하므로 근본적인 문제를 가지고 있다.

이러한 문제의 해결을 위하여 본 연구에서는 문서 분류에 사용하는 특성 정보인 단어와 특성 벡터의 관점에서 한글 문서 분류의 문제를 다루고자 한다. 먼저 명사와 색인어를 특성 정보로 이용하여 분류 알고리즘을 적용할 경우 분류 성능에 미치는 영향을 알아보고자 한다. 또한 특성 벡터의 크기가 분류 성능에 미치는 영향에 대해 살펴본다. 이러한 비교의 결과들은 문서 분류 알고리즘에 적합한 특성 정보의 선택과 특성 벡터의 크기 결정에 유용하게 이용될 수 있으며, 분류 알고리즘의 특징에 영향을 적게 받는 특성의 결정에도 이용할 수 있다. 실험에 사용된 문서 분류 알고리즘은 TFIDF, Naive Bayes와 TFIDF에 단어의 출현 확률이 고려된 PrTFIDF를 사용하였다. 실험 문서 집합은 연구개발정보센터에서 개발한 한글 테스트 컬렉션(HANTEC)의 HKIB94를 이용하였다.

2. 문서 분류 알고리즘

(1) TFIDF

TFIDF(Term Frequency Inverse Document Frequency) 알고리즘은 문서에 출현하는 단어의 빈도(term frequency)와 특정 단어를 포함하는 문서의 역수(inverse document frequency)를 특성으로 사용하여 벡터 형태로 문서를 표현한다. 문서의 분류는 특성 벡터의 학습을 통하여 얻은 범주별 대표 벡터(prototype vector)와 분류 대상 문서의 특성 벡터의 코사인 값을 사용하여 판정하게 된다(Joachims(1997)).

일반적으로 문서는 $\vec{d} = (d_1, d_2, \cdots, d_{|F|})$의 벡터 형태로 표현된다. 여기서 $|F|$는 어떤 문서에 존재하는 전체 단어의 수를 의미한다. 단어 w가 문서에서 나타나는

빈도를 $TF(w, \vec{d})$, 단어 w를 포함하는 문서의 수를 $DF(w)$로, 문서 집합에 포함되는 전체 문서의 수를 $|D|$라 하자. 이 때 $IDF(w)$는 식 (2.1)로, 특성 벡터의 원소 d_i는 식 (2.2)와 같이 표현된다.

$$IDF(w) = \log\left(\frac{|D|}{DF(w)}\right) \tag{2.1}$$

$$d_i = TF(w_i, \vec{d}) \cdot IDF(w_i) \tag{2.2}$$

또한 단어 w가 범주 C의 문서들에서 나타나는 빈도를 $TF(w, C)$라 할 때, 분류 대상 문서 d'에 대한 범주 결정은 식 (2.3)을 이용하여 이루어진다.

$$d'_{TFIDF} = \arg\max_{C \in V} \frac{\sum_{w \in F}\{TF(w, d')IDF(w)\}\{TF(w, C)IDF(w)\}}{\sqrt{\sum_{w' \in F}\{TF(w', C)IDF(w)\}^2}} \tag{2.3}$$

(2) Naive Bayes

Naive Bayes 알고리즘은 문서에 나타나는 각 단어들이 서로 독립이라는 가정 하에, 문서 집합에 대한 단어의 출현 빈도와 문서 내에서의 단어의 출현 빈도를 이용한다. 학습 절차는 먼저 학습 문서 집합에서 범주 c_j에 대한 확률 $\Pr(c_j)$와 범주 c_j에 대한 단어 w_i의 조건부 확률 $\Pr(w_i|c_j)$를 계산한다. 학습을 통하여 계산된 확률 값들은 Bayes 정리를 이용하여 새로운 문서의 범주 결정에 이용된다(Mitchell(1997)). 문서 분류를 위한 Naive Bayes 알고리즘은 <그림 2.1>과 같다.

Learn_Text(*examples*, *V*)

- *voc* ← 문서에 나타나는 모든 상이한 단어
- *V* 내의 모든 범주 c_j에 대하여 $\Pr(c_j)$와 $\Pr(w_i|c_j)$를 계산
  - $docs_j$ ← 범주가 c_j인 *examples*의 집합
  - $\Pr(c_j)$ ← $|docs_j|/|examples|$
  - $text_j$ ← $docs_j$의 모든 요소를 나열하여 만든 문서
  - n ← $text_j$에 나타난 모든 단어의 수
  - *voc*에 있는 모든 단어 w_i에 대하여
    - n_i ← $text_j$에 나타난 단어 w_i의 수
    - $\Pr(w_i|c_j)$ ← $(n_k+1)/(n+|voc|)$

Classify_Text(*doc*)

- *pos* ← *voc*에서 발견된 토큰들을 포함하는 *doc*내의 모든 단어들의 위치 $doc_{NB} = \arg\max_{c_j \in V} \Pr(c_j) \prod_{i \in pos} \Pr(a_i|c_j)$
-
  - $\hat{\Pr}(a_i|c_j)$ ← $(n_c + mp)/(n+m)$
    - n ← 범주가 c_j인 *examples*의 수
    - n_c ← 범주가 c_j이고 특성 벡터 a가 a_i인 *examples*의 수
    - p ← uniform prior
    - m ← p에 대한 가중치

<그림 2.1> 문서 분류를 위한 Naive Bayes 알고리즘

(3) PrTFIDF

PrTFIDF(Probabilistic Classifier Derived from TFIDF) 알고리즘은 단어의 출현 확률을 고려한 TFIDF 알고리즘으로 Joachims(1997)에 의해 제안되었다. 이 알고리즘은 Naive Bayes와 비슷한 정확률을 나타내는 것으로 알려져 있다. PrTFIDF는 식 (2.4), (2.5)와 같이 변형된 $DF(w)$와 $IDF(w)$를 이용한다. $DF'(w)$는 단어 w의 문서 내 출현 횟수 대신 각각의 범주 C에서 출현하는 비율을 사용한다. 또한 $IDF'(w)$는 TFIDF에서의 log 변환 대신에 root 변환을 사용한다.

$$IDF'(w) = \sqrt{\frac{|D|}{DF'(w)}} \tag{2.4}$$

$$DF'(w) = \sum_{C \in V} \frac{TF(w, C)}{\sum_{w' \in F} TF(w', C)} \tag{2.5}$$

새로운 문서 d'에 대한 범주 결정은 식 (2.6)을 사용하며, TFIDF의 범주 결정 식 (2.3)과 동일한 형태를 가진다.

$$d'_{PrTFIDF} = \arg\max_{C \in V} \frac{\sum_{w \in F}\{TF(w, d')IDF'(w)\}\{TF(w, C)IDF'(w)\}}{\sqrt{\sum_{w' \in F}\{TF(w', C)IDF'(w')\}^2}} \tag{2.6}$$

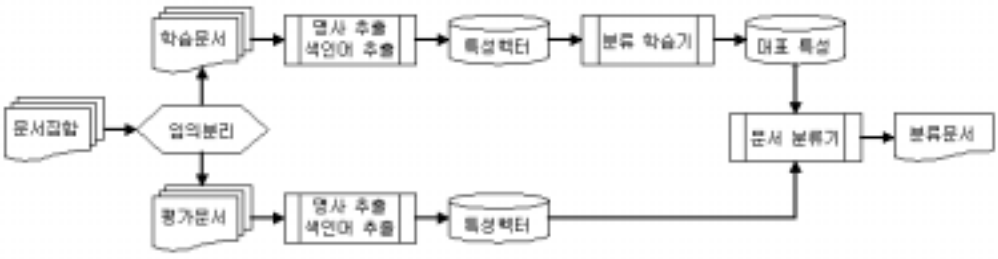
3. 한글 문서 분류

HANTEC 컬렉션은 연구개발정보센터에서 개발한 문서 집합으로 <표 3.1>과 같이 구성되어 있다. 본 연구에서는 HANTEC 컬렉션에서 제공하는 HKIB94를 실험 문서로 이용하였다. HKIB94는 한국일보 1994년 1월에서 7월까지의 기사로 총 19개 분야 22,000건의 기사로 이루어져 있다. 이 중에서 문서의 수가 적은 분야들을 제외하고, 범주가 명확이 구분되는 5개 분야 19,743건의 문서들로 분류 실험을 실시하였다. 실험에 사용된 분야는 정치(2,105건), 경제(9,913건), 사회(3,944건), 국제(3,237건), 스포츠(544건)이다.

<표 3.1> HANTEC 컬렉션의 문서 구성

분야	구성 문서	문서 건수
일반종합	1994년 한국일보 기사	22,000 건
	gov 확장자를 갖는 웹 페이지	9,000 건
	com 확장자를 갖는 웹 페이지	9,000 건
사회과학	1994년 한국경제신문 기사	39,480 건
	한국여성개발원 발행 정기간행물 여성연구 게재 논문	110 건
	경상북도의회 회의록	410 건
과학기술	과학기술처 지원 연구보고서	10,000 건
	해외과학기술 동향	18,000 건
	논문 서지 정보	12,000 건

문서 분류 절차는 <그림 3.1>의 문서 분류를 위한 시스템 구성에 따라 진행된다. 문서 집합은 범주별 대표 특성 학습을 위한 학습 문서 집합과 테스트를 위한 평가 문서 집합으로 임의 분리한다. 이 실험에서 학습 문서 집합은 전체 문서의 66%, 평가 문서 집합은 34%를 임의로 분리하여 구성하였다. 두 번째로 각각의 문서에서 HAM을 이용하여 문서에 포함된 명사와 색인어를 추출하여 분류 알고리즘에 적합한 특성 벡터 생성을 위한 정보 계산을 위해 데이터베이스에 저장하였다.



<그림 3.1> 문서 분류 시스템 구성도

최종적인 문서 분류는 먼저 학습 문서에서 생성된 특성 벡터를 이용하여 TFIDF, Naive Bayes, PrTFIDF 분류 학습을 통하여 범주별 대표 특성 벡터를 생성한다. 다음으로 범주별 대표 특성 벡터와 평가 문서에서 생성된 특성 벡터를 각각의 분류 알고리즘에 적용하여 범주를 판단하는 과정으로 이루어진다.

4. 문서 분류 결과

문서 분류에서 각각의 문서에 대한 특성 벡터는 문서 내에 출현하는 단어 또는 색인어 정보를 이용하여 구성된다(임형근,장덕성(2001), Lewis(1992)). 이때 문서 분류의 정확률의 향상 측면에서는 모든 문서에서 높은 빈도로 나타나는 단어를 제거하고 특정 문서에만 높은 빈도로 나타나는 단어를 채용할 필요가 있다(황도삼 외(1999)). 따라서 본 연구에서는 한글 문서의 분류를 위하여 다음과 같은 4가지 특성 정보를 사용할 때의 정확률과 정밀도를 비교하여 보았다. 정확률과 정밀도는 식 (4.1), (4.2)와 같이 계산되어진다(Junker M., Hoch R., 1999).

- 문서에 출현하는 명사들의 집합
- 문서에 출현하는 색인어들의 집합
- 모든 범주에 공통적으로 나타나는 단어를 제외한 명사들의 집합
- 모든 범주에 공통적으로 나타나는 단어를 제외한 색인어들의 집합

또한 특성 벡터의 크기에 따른 정확률과 정밀도의 차이를 알아보기 위하여 특성 정보의 값을 정렬하여 크기에 따른 문서 분류 실험을 실시하였다. HAM을 이용하여 명

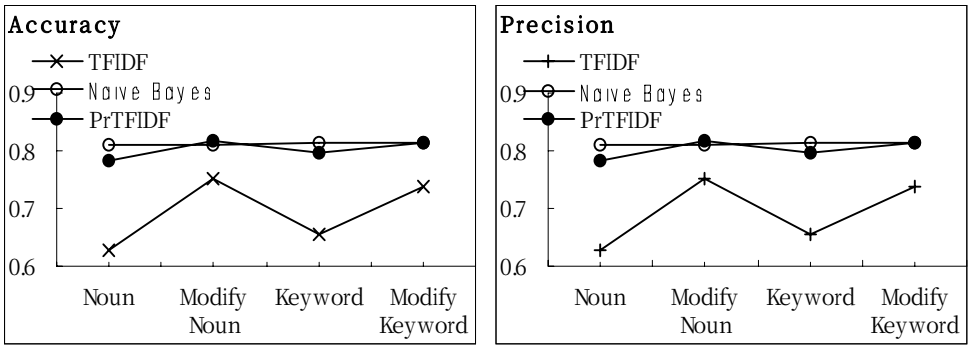
사와 색인어를 추출할 경우 조사, 보조동사, 관사, 전치사와 같은 기능어를 제외하고 추출하는 기능을 가지고 있기 때문에 별도의 기능어 제거 작업은 수행하지 않았다.

$$\text{accuracy} = \frac{\#(\text{documents correctly categorized})}{\#(\text{documents})} \tag{4.1}$$

$$\text{precision} = \frac{\#(\text{documents correctly assigned to categoriy } C)}{\#(\text{documents assigned to category } C)} \tag{4.2}$$

4.1. 특성 정보의 성질에 따른 문서 분류

특성 정보로 위의 4가지 경우를 이용한 한글 문서의 분류 실험 결과는 <그림 4.1>과 같다. 차이는 크지 않지만 PrTFIDF와 Naive Bayes가 TFIDF에 비해 정확률과 정밀도에서 좋은 성능을 가지고 있다. 특히 Naive Bayes의 경우 특성 정보의 성질에 영향을 거의 받지 않음을 알 수 있다. 또한 공통 출현 단어를 제거한 명사를 특성 정보로 이용할 경우 PrTFIDF가 가장 높은 정확률과 정밀도를 제공하는 것을 알 수 있다.



<그림 4.1> 명사, 색인어, 공통 출현 특징을 제거한 경우의 분류 결과

4.2. 특성 벡터의 크기에 따른 문서 분류

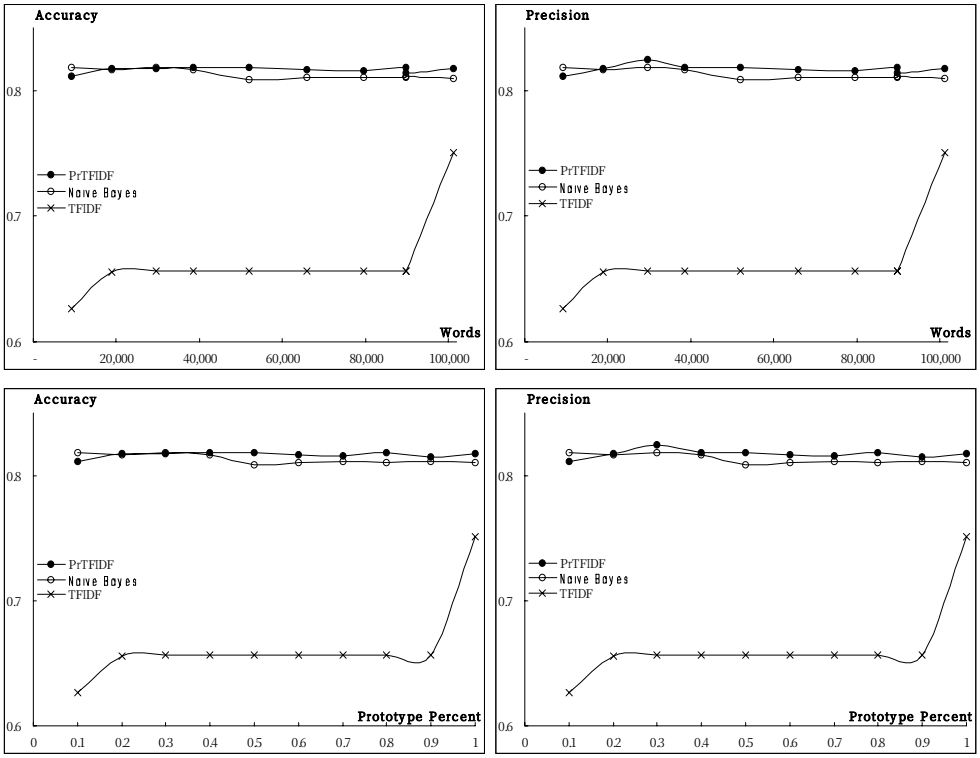
Joachims(1997)의 뉴스그룹 문서와 Routers의 acq 범주 문서의 분류 실험에서 특성 벡터의 크기가 커질수록 정확률이 증가하는 것을 보였다. 반면에 Routers의 wheat 범주의 문서의 경우는 TFIDF를 제외하고 거의 유사한 정확률을 가지는 결과를 나타내고 있다. 또한 김진상, 신양규(2000)의 뉴스 그룹 문서에 대한 분류 실험에서는 특성 단어의 수가 20,000개 정도에서 최소 에러를 가지는 것을 보였다.

한글 문서에서도 위의 연구에서와 같이 특성 벡터의 크기에 따른 분류율의 차이가 존재할 수 있다. 따라서 본 연구에서는 특성 정보의 값을 크기 순으로 정렬하여 벡터의 크기에 따른 분류의 정확률과 정밀도를 측정하였다. 특성 벡터 구성에 사용한 특성은 앞의 실험에서 모든 범주에 나타나는 단어를 제거한 명사를 사용하였다.

실험 결과는 <그림 4.2>와 같이 Naive Bayes와 PrTFIDF는 벡터의 크기에 따른

영향을 거의 받지 않고 근소한 차이를 가지는 것을 알 수 있다. 반면에 TFIDF는 특성 벡터의 크기가 클수록 정확률과 정밀도가 높은 것을 알 수 있다. 또한 PrTFIDF의 경우 특성 정보의 크기가 30%인 단어 크기 30,000 단어 부근에서 높은 정확률과 정밀도를 가진다.

Naive Bayes 분류기의 경우 특성 벡터의 크기가 커질수록 미미한 수준이기는 하지만 분류율이 감소하는 것을 알 수 있다. 반면에 PrTFIDF가 Naive Bayes에 비하여 특성 벡터의 크기에 비교적 영향을 적게 받는 것을 실험을 통하여 알 수 있다.



<그림 4.2> 특성 벡터의 크기에 따른 분류 결과

5. 결론

본 연구에서는 TFIDF, Naive Bayes, PrTFIDF를 이용하여 특성 벡터의 크기에 따른 분류 성능과 특성의 성질에 따른 분류 성능에 대한 실험을 실시하였다. 실험 결과 모든 범주에 나타나는 단어를 제거한 명사를 특성 정보로 사용하는 것이 분류 성능을 높일 수 있는 방법으로 나타났다. 또한 Naive Bayes와 PrTFIDF를 사용하는 것이 작은 특성 벡터 크기에 대하여도 좋은 성능을 가짐을 알 수 있었다. 이러한 결과로 볼 때 한글 문서의 특성에 대한 명확한 이해와 사용은 분류 성능의 향상에 중요한 요소라 할 수 있다.

이 실험에서는 범주의 크기와 각 범주에 포함되는 문서의 수에 대한 변화를 고려하지 않았다. 따라서 향후 범주의 크기에 따른 분류 성능에 대한 연구와 각 범주에 포함되는 문서 수의 차이에 따른 성능에 대한 연구가 이루어져야 할 것이다.

문서 분류에 있어서 어려운 점은 분류 학습을 위한 양질의 문서 데이터 확보이다. 문서라는 것은 인간의 생각을 언어로 표현한 것으로 시간의 흐름에 따라 구현 언어가 변화한다. 또한 모든 분야의 문서 분류에 적합한 특성 정보와 분류기를 개발하는 것은 매우 어려운 과제이다. 따라서 특정 분야의 문서 분류에 적합한 특성 정보를 찾고 분류 알고리즘을 개발하는 것이 바람직할 뿐만 아니라, 시간의 흐름에 따라 변화하는 문서의 내용을 반영하기 위해 자체적으로 학습이 가능한 알고리즘에 대한 연구가 이루어져야 할 것이다.

참고문헌

1. 강현규, 박세영 (1998). 정보 검색, 정보처리, 제5권, 5호, 37-47.
2. 김진상, 신양규 (2000). 베이지안 학습을 이용한 문서의 자동분류, *Journal of the Korean Data & Information Science Society*, Vol. 11, No. 1, 19-30.
3. 이윤오, 이정진 (2000). 논리적 패턴을 이용한 확률화 정보검색 시스템의 연구, 응용통계연구, 제13권, 1호, 1-10.
4. 임형근, 장덕성 (2001). 색인어 연관성을 이용한 의료정보문서 분류에 관한 연구, 정보처리학회논문지 B, 제8-B권, 5호, 469-476.
5. 한광록, 선복근, 한상태, 임기욱 (2000). 인터넷 문서 자동 분류 시스템 개발에 관한 연구, 한국정보처리학회논문지, 제7권, 9호, 2867-2875.
6. 황도삼, 최기선, 김태석 (1999). 자연언어이해, 홍릉과학출판사.
7. Aas K., Eikvil L. (1999). Text Categorisation: A Survey, *Norwegian Computing Center*, Report No. 941.
8. Baker L.D., McCallum A.K. (1998). Distributional Clustering of Words for Text Classification, *Proceedings of the 21st ACM International Conference on Research and Development in Information Retrieval.*
9. Cunningham S.J., Littin J., Witten I.H. (1997). Applications of Machine Learning in Information Retrieval, *Annual Review of Information Science and Technology*, 341-419.
10. HAM, 한국어 형태소 분석기와 한국어 분석 모듈, 국민대학교 자연언어 정보검색 연구실, http://nlp.kookmin.ac.kr.
11. HANTEC, 정보검색 시스템 평가를 위한 한글 테스트 컬렉션, 연구개발정보센터, http://hantec.kisti.re.kr.
12. Joachims T. (1997). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization, *Proceedings of 14th International Conference Machine Learning.*
13. Junker M., Hoch R. (1999). On the Evaluation of Document Analysis Components by Recall, Precision, and Accuracy, *Proceedings of the 5th*

International Conference on Document Analysis and Recognition.
14. Lewis D.D. (1992). Feature Selection and Feature Extraction for Text Categorization, *Proceedings of Speech and Natural Language Workshop*, 212-217.
15. Mitchell T.M. (1997). *Machine Learning*, McGraw-Hill.
16. Rijsbergen C.J. (1979). Information Retrieval, *http://www.dcs.gla.ac.uk/Keith/ Preface.html*.
17. Verbeek J.J. (2000). An Information Theoretic Approach to Finding Word Groups for Text Classification, *Institute for Language, Logic and Computation, University of Amsterdam*, Masters Thesis.
18. Yang Y., Pedersen J.O. (1997). A Comparative Study on Feature Selection in Text Categorization, *Proceedings of the 14th International Conference on Machine Learning*, 412-420.

[2002년 12월 접수, 2003년 4월 채택]