DOI: 10.3745/KIPSTB.2008.15-B.5.457

점진적 특징 가중치 기법을 이용한 나이브 베이즈 문서분류기의 성능 개선

김 한 준† · 장 재 영††

요 약

실제 운용 환경에서 자동문서분류시스템의 성공을 위해서 충분하지 못한 학습문서의 문제와 특징 공간들에 대한 사전지식이 없는 상황을 해결하는 것이 관건이다. 이런 맥락에서 많은 자동문서분류 시스템의 구축을 위해 나이브 베이즈 문서분류 알고리즘을 사용한다. 이는 기존 학습된 분류모델과 특징 공간을 점진적으로 갱신함으로써 분류모델을 향상시키는 것이 매우 용이하기 때문이다. 본 논문에서는 특징 가중치를 이용하여 문서분류기의 성능을 향상시키는 기법을 제안한다. 기본 아이디어는 문서분류 모델의 인자로서 특징들의 분포뿐만 아니라 각 특징들의 중요도를 반영하는 것이다. 속성 선택을 미리 수행하여 학습모델을 만드는 것이 아니라, 속성 중요도를 나이브 베이즈 학습 모델에 포함시킴으로써 보다 정확한 모델을 생성할 수 있다. 또한 동적 환경에서 점진적인 특징 가중치 부여를 위해 기존의 특징 갱신 기법을 확장한 알고리즘도 제안한다. 본 논문에서 제안된 기법을 평가하기 위해서 Reuters-21578과 20Newsgroup 문서집합 이용한 실험을 실시하여, 제안된 기법이 전통적인 나이브 베이즈 분류기의 성능을 크게 향상시킴을 증명한다.

키워드 : 문서분류, 나이브 베이즈 분류기, 특징 가중치, 특징 선택, χ^2-통계량

Improving Naïve Bayes Text Classifiers with Incremental Feature Weighting

Han-joon Kim† · Jae-young Chang††

ABSTRACT

In the real-world operational environment, most of text classification systems have the problems of insufficient training documents and no prior knowledge of feature space. In this regard, Naïve Bayes is known to be an appropriate algorithm of operational text classification since the classification model can be evolved easily by incrementally updating its pre-learned classification model and feature space. This paper proposes the improving technique of Naïve Bayes classifier through feature weighting strategy. The basic idea is that parameter estimation of Naïve Bayes considers the degree of feature importance as well as feature distribution. We can develop a more accurate classification model by incorporating feature weights into Naive Bayes learning algorithm, not performing a learning process with a reduced feature set. In addition, we have extended a conventional feature update algorithm for incremental feature weighting in a dynamic operational environment. To evaluate the proposed method, we perform the experiments using the various document collections, and show that the traditional Naïve Bayes classifier can be significantly improved by the proposed technique.

Keywords : Text classification, Naïve Bayes classifier, feature weighting, feature selection, χ^2-statistics

1. 서 론

최근 들어 블로그(blog), 전자 도서관(digital library), 뉴스 등과 같은 인터넷 환경에서의 온라인 문서들이 꾸준히 증가함에 따라 자동문서분류(automated text classification)에 대한 관심이 학계뿐만 아니라 산업계에도 점차 확대되고 있다. 자동문서분류란 학습문서 집합을 미리 확보하고 있지 않은 상태에서 지속적으로 유입되는 문서만을 사용해서 그 문서들을 자동으로 분류하는 것을 의미한다. 이 기술은 인터넷 환경에서 다양한 응용분야를 갖는데, 웹페이지의 계층적 분류, 스팸 메일의 분류 또는 사용자의 기호에 따른 뉴스기사의 분류 등을 그 예로 들 수 있다. 또한 최근에는 인터넷 게시판 등에서 특정 제품에 대한 소비자의 상품평과 같은 주관적 의견들에 대한 여론을 판단하는 분야에도 자동문서분류 기법을 적용하려는 시도가 이루어지고 있다[1][2].

최근의 문서분류 기법은 주로 기계학습(machine learning) 기술을 사용한다. 이 방식에서는 카테고리(category)의 특징을 설명하는 분류모델(classification model)이 미리 준비되

※ 본 연구는 2008년도 한성대학교 교내연구비 지원과제이며, 또한 지식경제부 및 정보통신연구진흥원의 대학 IT연구센터 육성·지원사업 (IITA-2008-C1090-0801-0031)의 연구결과로 수행되었음.
† 정 회 원 : 서울시립대학교 전자전기컴퓨터공학부 조교수
†† 종신회원 : 한성대학교 컴퓨터공학과 부교수
논문접수 : 2008년 3월 19일
수 정 일 : 2008년 5월 17일
심사완료 : 2008년 5월 18일

어야 하는데, 학습문서집합으로부터 각 카테고리의 특징을 학습하는 귀납적 프로세스(inductive process)를 거쳐 자동으로 문서를 분류할 수 있는 모델을 생성하게 된다. 일반적인 문서분류를 위한 기계학습 방법으로는 나이브 베이즈(Naïve Bayes)[3], K-nearest neighbor[4], Support Vector Machine(SVM)[5] 등이 있다.

일반적인 기계학습에 기반을 둔 문서분류 알고리즘에서는 분류성능에 영향을 미치는 다음의 문제들을 해결해야한다. 우선 분류모델을 수립하기 위해서는 양질의 학습문서를 적정 개수 이상 확보해야한다. 그러나 현실적으로 만족스러운 양과 질을 모두 갖춘 학습문서를 확보하는 것은 쉽지 않은 문제이다. 오히려 실제 환경에서는 완전한 형태의 학습문서가 한순간에 확보되기 보다는, 데이타 스트림(data stream)과 같이 연속적인 형태로 학습문서들이 제공된다. 따라서 이러한 환경에서는 새로운 학습문서들이 제공될 때마다 현재의 분류모델을 지속적으로 갱신하는 방법이 필요하다. 위에서 나열한 기계학습 방법 중에서 나이브 베이즈 알고리즘은 이러한 지속적 분류모델의 갱신이 가능한 기법이다. 또한 이 알고리즘은 모델 인자(parameter)의 구성이 간단하면서도 복잡한 모델 구성을 가지는 다른 기법들과 견주어 성능이 뒤지지 않아 많은 문서분류 시스템에서 채택되고 있다.

두 번째는 각 카테고리를 표현하기 위한 특징(feature)을 선택하는 문제이다. 특징 선택(feature selection)은 '차원의 저주(curse of dimensionality)'라고 불리는 문제를 해결하기 위한 방법으로 제시되었다. 차원의 저주란 문서의 수가 증가할수록 특징이라고 불리는 단어(word) - 또는 용어(term) - 의 수가 지수적으로 증가하는 문제를 말하며, 특징 선택이란 이러한 특징들 중에서 카테고리를 가장 잘 표현할 수 있는 일부 특징을 추출하는 과정을 말한다. 일반적으로 문서들은 단어들의 다중집합(multi-set 또는 bag)으로 부터 추출된 특징들로 표현되며, 모든 문서에 대한 특징들로 구성된 특징 공간(feature space)은 결국 수많은 단어들이 모인 어휘(vocabulary)를 구성하게 된다. 따라서 문서분류의 성능을 향상시키기 위해서는 특징 선택을 통해서 특징 공간을 줄이고 왜곡된 특징들을 삭제하는 과정이 필수적이다. 이를 위해 [6]에서는 χ^2-통계량(Chi-Square statistics), mutual information, information gain, document frequency와 같이 특징을 선택하기 위한 다양한 기법들을 비교 분석한 결과를 제시하였다, 반면에 [5]에서는 분류모델을 수립하기 위해서는 문서상에 존재하는 모든 특징들이 사용되어야한다는 주장을 하기도 했으나 이는 성능 문제로 인해 큰 주목을 받지는 못하고 있다.

특징 선택(feature selection)의 문제는 학습문서가 시간에 따라 연속적으로 제공되는 실제 환경에서는 매우 비효율적이라는 점이다. 이러한 환경에서의 특징 공간은 미리 주어진 사전 정보도 존재하지 않으며 시간에 따라 수시로 변화한다. 따라서 새로운 학습문서가 현재의 문서집합에 새롭게 유입될 때마다 문서분류기는 기존 구성된 분류모델을 재구성하지 않고, 새롭게 생성된 특징 집합을 이용하여 분류모델을 점진적으로 개선해나가는 것이 바람직하다. 예를 들어 나이브 베이즈 문서분류기에서도 새로운 학습문서의 유입에 따라 단어들에 대한 통계량을 매번 재계산하여 모델 인자를 수정하게 된다. 이때 기존의 방법에서는 각 카테고리에 대한 단순한 특징 선택만으로 모델을 설정하였으나, 각 특징에 대한 해당 카테고리에서 차지하는 비중에 따라 가중치를 부여하여 모델을 설정할 경우 더욱 높은 분류성능을 기대할 수 있다. 다시 말해서, 학습과정의 결과로 도출되는 분류모델이 이미 확정된 특징 집합으로 결정되는 것이 아니라, 학습 알고리즘의 원리를 활용하여 각 특징들의 비중을 학습모델의 인자에 반영함으로써 보다 개선된 분류모델을 생성될 수 있는 것이다.

본 논문은 기존의 특징 선택 과정을 나이브 베이즈 학습 알고리즘에 결합하는 방안에 대한 연구이며, 이는 특징 가중치를 나이브 베이즈 모델의 인자로 활용했다는데 그 의의가 있다. 특징 가중치란 각 카테고리를 가장 잘 표현하는 대표적인 특징들에 가중치를 부여하는 값을 의미한다. 가중치가 부여된 특징들은 기존의 분류모델과 결합하여 문서분류기의 성능을 크게 높일 수 있다. 본 논문에서는 우선 학습문서가 점진적으로 추가되는 환경에서 χ^2-통계량에 기반을 둔 특징 가중치 기법을 활용하여 분류모델을 동적으로 진화시키기 위한 알고리즘을 제안한다. 또한 특징 가중치를 부여하기 위해서 [7]에서 제안한 특징 갱신(feature update) 기법을 확장한 알고리즘도 제시한다. 제안된 알고리즘의 우수성을 평가하기위해서 Reuters-21578과 20Newsgroups 문서 집합을 이용한 실험을 실시하였다. 실험은 다양한 인자의 변화에 따라 특징 가중치를 적용한 알고리즘과 그렇지 않은 기존의 알고리즘에 대한 문서분류의 정확도를 비교하여 본 논문에서 제안한 알고리즘의 우수성을 증명하였다.

본 논문의 구성은 다음과 같다. 우선 2장에서는 문서분류에 가장 많이 이용되는 나이브 베이즈 분류모델에 대해 살펴보고, 3장에서는 특징 가중치를 적용하여 나이브 베이즈 분류모델의 인자들을 점진적으로 진화시키는 알고리즘을 제안한다. 4장에서는 실험 결과를 제시하고 마지막으로 5장에서는 결론과 향후 연구 과제를 제시한다.

2. 나이브 베이즈 문서분류 기법

나이브 베이즈 학습기법은 단순하면서도 정확한 추정능력을 발휘한다고 알려져 있어 많은 문서분류 프로젝트에서 채택되어 왔다[3]. 또한 특징 공간이 동적으로 변화하는 환경에서 나이브 베이즈 학습기법은 SVM이나 결정 트리(decision tree)와 같은 복잡한 기계학습 기법에 비해 매우 적합한 분류모델이 되고 있다. 따라서 나이브 베이즈 분류기의 성능을 향상하기 위한 연구는 자동문서분류 분야에서 매우 중요한 의미를 갖는다고 할 수 있다.

나이브 베이즈 문서분류기에서는 소속될 카테고리가 미리 정해진 학습문서 집합으로 부터 분류모델에 대한 인자들을 추정하게 된다. 추정된 분류 모델 $\hat{\theta}_{NB}$는 다음과 같은 두개

의 인자로 구성된다.

$$\hat{\theta}_{NB} = \{\hat{\theta}_{w|c}, \hat{\theta}_{c}\} \quad (1)$$

여기서 $\hat{\theta}_{w|c}$는 카테고리 c에 속하는 문서집합에서 임의로 추출된 단어가 w일 확률값을 나타내고, $\hat{\theta}_{c}$는 전체문서집합에서 임의 추출한 문서가 카테고리 c에 속할 사전확률값(prior probability)을 나타낸다. 각 인자들은 MAP 가설(Maximum A Posteriori hypothesis)에 따라 가장 큰 사후확률값(posterior probability)으로 추정된다.[1)]

나이브 베이즈 알고리즘에 의한 문서분류는 주어진 문서를 분류하기 위해 베이즈(Bayes) 정리에 의해 아래 식과 같이 주어진 문서에 대한 카테고리의 사후확률값을 추정함으로써 이루어진다.

$$Pr(c_j|d_i) = \frac{Pr(c_j)Pr(d_i|c_j)}{Pr(d_i)} \quad (2)$$

여기서 $Pr(c_j)$는 전체문서집합에서 임의로 추출한 문서가 카테고리 c_j에 속할 사전확률값, $Pr(d_i|c_j)$는 카테고리 c_j에 속하는 문서집합에서 임의로 추출된 문서가 d_i일 확률값, $Pr(d_i)$는 전체문서집합에서 임의 추출한 문서가 d_i일 확률값을 의미한다. 이 식을 이용하여 문서 d_i는 전체 카테고리 집합 C중에서 사후확률값으로 $argmax_{c_j \in C} Pr(c_j|d_i)$인 카테고리 c_j로 할당된다.[2)] 여기서 문서 d_i의 각 단어들은 한 번이상 나타날 수 있으며 출현하는 단어들의 빈도도 중요한 요소가 되므로 d_i는 단어들의 다중 집합인 $(w_{i1}, w_{i2}, ..., w_{i|d_i|})$로 표현된다. 또한 나이브 베이즈 분류기는 문서에 출현하는 단어들은 서로 간에 영향력이 없이 독립적이며, 단어가 출현하는 문서내의 위치와는 관계없다는 단순한 가정에 기반을 두고 있다. 이러한 가정 하에 분류함수는 다음과 같이 표현될 수 있다.

$$\begin{aligned} \Phi_{\hat{\theta}_{NB}} &= argmax_{c_j \in C} Pr(c_j|d_i) \\ &= argmax_{c_j \in C} Pr(c_j) \cdot \prod_{k=1}^{|d_i|} Pr(w_{ik}|c_j) \end{aligned} \quad (3)$$

이 함수를 생성하기 위해서 우선 $Pr(c_j)$의 계산은 카테고리가 정해진 학습문서 집합 D^t 중에서 c_j에 존재하는 문서들의 수를 카운트함으로써 쉽게 얻을 수 있다. 즉, $Pr(c_j|d_i) \in \{0,1\}$라고 할 때 $Pr(c_j)$는 다음의 수식으로 계산된다.

$$Pr(c_j) = \hat{\theta}_{c_j} = \frac{\sum_{i=1}^{|D^t|} Pr(c_j|d_i)}{|D^t|} \quad (4)$$

이제 $Pr(w_{ik}|c_j)$의 값을 추정하는 문제만 남게 되는데, $TF(w_{ik}, c_j)$를 카테고리 c_j에서 단어 w_{ik}가 출현하는 빈도수라 하고 V를 전체 학습문서의 어휘 집합이라 할 때, $Pr(w_{ik}|c_j)$의 MAP 추정값은 $\frac{TF(w_{ik}, c_j)}{\Sigma_{w_{ik} \in V} TF(w_{ik}, c_j)}$와 같다. 그런데 이 식은 특정 카테고리에 존재하는 않는 단어에 대해서는 0값을 가지게 되어 전체 식의 값을 0이 되게 할 수 있다. 전체 어휘들 중에서 특정 카테고리에 존재하는 않는 단어가 다수 존재할 수 있기 때문에 이를 보정해 주어야 한다. 이를 위해 일반적으로 Laplace smoothing[3]이 사용되는데, 이는 모든 단어의 사전 출현횟수가 모두 같음을 의미한다. Laplace smoothing을 위의 식에 적용하면 다음의 수식을 얻는다.

$$Pr(w_{ik}, c_j) = \hat{\theta}_{w_{ik}|c_j} = \frac{1 + TF(w_{ik}, c_j)}{|V| + \Sigma_{w_{ik} \in V} TF(w_{ik}, c_j)} \quad (5)$$

수식 (4)와 (5)에서 보는 바와 같이 나이브 베이즈 분류기에서의 학습과정에서는 TF에서 수집된 것 이외에는 더 이상의 다른 통계정보를 필요로 하지 않을 뿐만 아니라, 다른 기계학습 방법과는 달리 추가적인 복잡한 과정도 요구되지 않는다.

지금까지 설명한 나이브 베이즈 학습기법은 문서분류 시스템의 관점에서 다음과 같은 장점을 갖는다. 첫째, 나이브 베이즈 분류기는 주어진 학습문서 집합을 한 번의 스캔만으로 인자들을 결정할 수 있으므로 학습과정의 속도가 다른 기법에 비해 매우 빠르다. 둘째, 주어진 카테고리 집합에 대한 모델 인자를 점진적으로 개선하는 것이 매우 간단하다. 이는 재학습이 매우 간단하게 수행될 수 있음을 의미하는데 수식 (4)와 (5)에서 보는 바와 같이 학습문서가 추가될 경우 새롭게 $\hat{\theta}_{w|c}$와 $\hat{\theta}_{c}$을 추정하기만 하면 된다. 이러한 특징은 문서의 집합이 동적으로 진화하는 경우에 매우 유리하다. 이러한 점을 고려하여 [7]에서는 나이브 베이즈 학습 모델을 기반으로 점진적 특징 갱신 알고리즘을 제안하였으며, 본 논문에서는 학습문서가 동적으로 추가되는 환경에서 특징 가중치에 필요한 특징들의 순위결정을 위해 이 알고리즘의 확장된 형태를 제안한다. 마지막으로 나이브 베이즈 문서분류기는 문서상에 나타나는 특징들에 대한 중요한 정보를 비교적 쉽게 수용할 수 있다. 예를 들어 학습 과정에서 새로운 뉴스나 기고문의 타이틀에 출현하는 용어들은 적당한 비중값을 곱하여 그 용어의 중요도를 반영할 수 있으며, HTML 형식의 반구조적(semi-structured) 전자문서에 포함되어 있는 태그(tag) 정보의 중요도도 같은 방식으로 처리할 수 있다. 본 논문에서는 이 같은 장점을 최대로 활용하기 위해 각 문서에 나타나는 특징들에 대해서 카테고리간의 상

1) MAP 추정은 주어진 학습문서 D에 대해서 가능한 모델 Θ중 최대의 확률을 갖는 모델 θ_{MAP}을 추정하는 것이다.
즉, $\theta_{MAP} \equiv argmax_{\theta \in \Theta} Pr(\theta|D) = argmax_{\theta \in \Theta} \frac{Pr(D|\theta) \cdot Pr(\theta)}{Pr(D)} = argmax_{\theta \in \Theta} Pr(D|\theta) \cdot Pr(\theta)$

2) $argmax_{x \in X} F(x)$는 $F(x)$값이 최대가 되는 x를 의미한다.

대적 중요도에 따라 인위적인 가중치를 부여함으로써 문서 분류기의 성능을 향상시키는 방법을 고안하였으며, 3장에서 이 방법에 대해 구체적으로 논한다.

3. 특징 가중치를 이용한 문서분류 기법

3.1 χ^2-통계량을 이용한 특징 가중치

카테고리의 성격을 규정짓는 방법에 있어서 단순히 주요 단어들의 빈도수만을 카운트하여 결정하기 보다는 특징들에 대한 가중치를 부여하는 많은 방법들이 제안되어 왔다. 예를 들어 위치에 따른 가중치 기법에서는 각 용어들이 문서 내에 위치에 곳에 따라 가중치가 다르게 부여된다[8]. 구체적인 예로 타이틀이나, 머리글, 혹은 문서 앞부분에 위치한 문장들에 포함된 용어들에는 더 많은 가중치를 부여할 수 있다. [9]에서는 타이틀이나 타이틀에 들어있는 단어 중에서 최소한 3개 이상을 포함한 문장으로 부터 주요 특징들을 추출한다. 그러나 이와 같은 휴리스틱(heuristic)들은 매우 직관적이고 다소 임의적인 방법이므로 일반화하여 적용하기에는 부족한 면이 있다.

문서분류기의 성능을 높이기 위해서는 각 카테고리의 주제를 명확히 담고 있는 주요한 특징들로 표현되어야한다. 따라서 각 카테고리를 위한 학습문서를 명확히 차별화 할 수 있는 주요 특징들에 보다 높은 가중치를 부여하는 것이 요구된다. 물론 뉴스 기사와 같이 주요 특징들이 문서의 타이틀에 포함되는 경우도 많다. 그러나 이메일이나 뉴스그룹(newsgroup)등에서는 그렇지 못한 경우가 자주 발생한다. 예를 들어 20Newsgroups에 있는 "*I have a question*"과 같은 타이틀은 문서의 주제를 결정하는데 아무런 도움을 주지 못한다.

χ^2-통계량에 기반을 둔 특징 선택

본 논문에서 제안하는 특징 가중치 기법은 χ^2-통계량에 기반을 둔다. χ^2-통계량은 문서분류에 대한 연구에서 일정 개수의 최적의 특징을 추출하는 데 폭넓게 응용되고 있다[6]. χ^2-통계량에서는 모든 특징에 대해 문서의 주제를 표현하는 정도를 평가하여 가장 적합한 특징들을 선택하게 된다. 주어진 단어 w와 카테고리 c에 대해서 χ^2-통계량 $\chi^2(c,w)$는 w와 c의 관련성 정도를 평가하는 것으로, 이 값이 작으면 서로 독립적이라는 것을 의미하며 반대로 크면 상호 관련성이 크다는 것을 의미한다[6][10]. 단어 w와 카테고리 c에 대한 2원 분할표(2-way contingency table)를 구성한 후 χ^2-통계량 $\chi^2(c,w)$는 다음과 같이 계산된다.

$$\frac{N\times(DF(w,c)\times DF(\overline{w},\overline{c})-DF(w,\overline{c})\times DF(\overline{w},c))^2}{(DF(w,c)+DF(\overline{w},c))\times(DF(w,c)+DF(w,\overline{c}))\times(DF(w,\overline{c})+DF(\overline{w},\overline{c}))\times(DF(\overline{w},c)+DF(\overline{w},\overline{c}))} \quad (6)$$

여기서 $DF(w,c)$는 w가 포함되는 문서 중에 카테고리 c에 해당에는 문서의 빈도수를 나타내고, $DF(\overline{w},c)$은 카테고리 c에 포함된 문서 중에 w를 포함하지 않는 문서의 빈도수를 나타낸다. 역으로, $DF(w,\overline{c})$는 c에 포함되지 않으면서 w를 포함한 문서의 빈도수를 나타내며, 마지막으로 $DF(\overline{w},\overline{c})$는 c에도 포함되지 않고 w도 갖고 있지 않는 문서의 수를 나타낸다. N은 총 문서의 수를 나타낸다. 이 식에 의해서 c와 w가 서로 독립적이면 $\chi^2(c,w)$는 0의 값을 갖게 되며, 반대로 w가 카테고리 c의 주제를 반영하는 단어이면 $\chi^2(c,w)$ 값은 증가될 것이다. 따라서 각 단어들의 χ^2 값은 해당 카테고리에서 각 단어들의 주제표현 정도를 나타내는 수단으로 사용되며, 각 카테고리 c에 대해서 χ^2 값이 가장 높은 단어 집합을 c를 대표하는 특징들로 선택하게 된다.

특징 가중치

χ^2 값을 이용한 특징 선택은 각 카테고리의 성격을 규정짓는 특징을 선별하는 좋은 방법을 알려져 있으며 많은 문서분류기에서 채택되어왔다. 특히 각 카테고리의 특징을 더욱 부각시키기 위해 선택된 특징들 중에 더욱 중요한 특징들에 가중치를 부여하는 방식도 사용되고 있다. 그러나 지금까지의 가중치 부여 기법은 단순히 주요 특징들에 두 배 또는 세 배의 일정량의 가중치를 부여하는 단순한 휴리스틱을 사용하였다. 그러나 이러한 임의적인 방법으로는 만족스러운 분류성능을 기대하기는 힘들다. 그 이유는 가중치 부여에 있어서 해당 특징이 주어진 카테고리에서 차지하는 비중만을 고려하였고 그 특징이 다른 카테고리들에 미치는 영향까지는 고려하지 않아 카테고리간의 차별성이 뚜렷하지 못한 경우가 많기 때문이다. 따라서 각 특징에 대해 해당 카테고리뿐만 아니라 그 특징이 다른 카테고리에 미치는 영향까지 고려해서 가중치를 결정하는 것이 각 카테고리간의 성격을 명확히 구분 짓는데 더욱 효율적이다.

앞서 언급한 바와 같이 나이브 베이즈 문서분류기는 각 카테고리 c_i에 분포하는 단어 w의 분포(또는 확률) $Pr(w|c_i)$을 추정함으로써 이루어진다. 따라서 단어 w가 특정 클래스 c_i의 성격을 규정짓는 주제단어라면 $Pr(w|c_i)$을 다른 카테고리 c_j의 $Pr(w|c_j)$보다 항상 큰 값을 갖는 것이 가장 이상적이라 할 수 있다. 따라서 $Pr(w|c_i)$이 $Pr(w|c_j)$보다 항상 큰 값을 갖도록 가중치를 둠으로써, 다른 카테고리와는 확연히 구별되는 차별화된 카테고리의 주제를 표현할 수 있게 된다. 본 논문에서는 이와 같이 기존의 나이브 베이즈 문서분류기가 갖는 단순성 및 효율성을 유지하면서 문서분류기의 성능을 향상시킬 수 있는 방안을 제안한다.

3.2 특징 가중치를 이용한 나이브 베이즈 분류 모델의 개선 방안

단순한 형태의 특징 선택 기법에 의해서 해당 카테고리에서의 특징들로 선택된 단어들은 각각의 중요도와 관계없이 동일하게 취급된다. 따라서 앞서 설명한 바와 같이 각 특징의 중요도에 따라 가중치를 부여하게 되면 나이브 베이즈 분류기 모델의 인자들을 보다 정확하게 추정할 수 있게 된다.

그러나 기본적인 나이브 베이즈 문서분류기에는 특징들의 중요도를 고려하는 인자가 없이 각 특징들의 TF 정보로 확률을 계산하게 된다. 따라서 각 특징의 TF 값을 조정함으로써 가중치를 부여할 수 있게 된다.

(그림 1)은 본 논문에서 제안하는 특징 가중치를 적용한 나이브 베이즈 분류기의 학습 알고리즘을 보여준다. 이 알고리즘은 초기의 나이브 베이즈 분류모델과 각 카테고리의 어휘를 이용하여, 특징 가중치 기법을 적용한 개선된 분류모델을 생성하게 된다. 이 알고리즘은 우선 각 카테고리 c_i에 포함된 문서로 부터 추출된 어휘 집합 $V(c_i)$로부터 상위의 χ^2 값을 갖는 특징들의 집합인 $V_{\chi^2}(c_i)$를 구성한다. 다음 단계에서는 앞에서 선택된 각 특징이 c_i가 아닌 다른 카테고리에 존재하는 문서들에도 발견되는지 확인한다. 만약 카테고리 c_i에서 선택된 단어 w가 다른 카테고리에서도 나타난다면, 카테고리 c_i에 대한 단어 w의 확률 $Pr(w|c_i)$을 모든 다른 카테고리에 대한 그것보다 높게 설정한다. 즉, $Pr(w|c_i) \gg \forall_{c_j \neq c_i} Pr(w|c_j)$이 만족되어야 한다. 이는 $Pr(w_x|c_i)$ 값을 인위적으로 높여줄 필요가 있음을 의미한다. 알고리즘 작성을 위해 이 식을 다시 작성하면 다음과 같다.

$$Pr(w_x|c_i) > max_{c_j \neq c_i} Pr(w_x|c_j) + \delta \qquad (7)$$

여기서, δ는 가중치를 위한 확률편차(PDW: Probability Difference for Weighting)라고 부르며 가중치의 정도를 결정하는 인자로, 보통 이 값은 실험을 통하여 분류성능을 높이는 최적의 값으로 결정된다. 식(7)에서 $Pr(w_x|c_i)$을 조정하는 것은, 결국 관련 TF값을 조정하는 문제가 된다. $TF(c_i, w_x)$의 증가량을 α라고 한다면, 식(5)와 식(7)을 이용하여 그 α값을 계산해보자. 식(5)에서 $Pr(w_x|c_i)$ 값은 $\frac{1+TF(c_i,w_x)}{|V|+\sum_{w_x \in V} TF(c_i,w_x)}$

이므로, 조정된 $Pr(w_x|c_i)$ 값은 $\frac{1+\{TF(c_i,w_x)+\alpha\}}{|V|+\{SumOfTF(c_i)+\alpha\}}$ 이다(여기서, $SumOfTF(c_i)$는 $\sum_{w_x \in V} TF(c_i,w_x)$를 의미함). 조정된 $Pr(w_x|c_i)$ 값을 식(7)에 대입하면 α값이 다음과 같이 계산된다.

$$\alpha > \frac{(\max_{c_j \neq c_i} P(w_x|c_j)+\delta)*(SumOfTF(c_i)+|V|) - TF(w_x,c_i) - 1}{(1-\max_{c_j \neq c_i} P(w_x|c_j)-\delta)} \qquad (8)$$

a는 정수이어야 하기 때문에 우변보다 큰 정수를 취하며, 그 값을 $TF(w_x,c_i)$과 $SumOfTF(c_i)$에 더해주면 된다. (Lines 9~14 참조).

지금까지는 $V_{\chi^2}(c_i)$의 단어가 다른 카테고리 c_j의 $V(c_j)$에는 존재하지만 중요 단어로는 선택되지 않아 $V_{\chi^2}(c_j)$에는 존재하지 않는 경우의 가중치 부여 기법을 제시하였다. 그러나 그 이외에 경우에 대한 처리 방법도 고려되어야 하는데, 첫째는 카테고리의 주제단어 집합 $V_{\chi^2}(c_i)$의 특정 단어가 다른 카테고리 c_j의 $V_{\chi^2}(c_j)$에도 공통적으로 포함되는 경우이며, 둘째는 $V_{\chi^2}(c_i)$의 특정 단어가 다른 카테고리에 전혀 존재하지 않는 경우이다.

우선 가장 이상적인 형태는 각 카테고리의 주제 단어들이 서로 겹치지 않는 것이다. 하지만 첫째 경우와 같이 선택된 특징이 겹칠 경우에는 어떠한 가중치도 부과하지 않는다. 그 이유는 같은 특징을 갖는 제3의 문서에 대해서 이 특징만으로는 분류될 카테고리를 결정하기 어려울 뿐만 아니라, 일방적으로 한 카테고리에만 가중치를 부과하면 자칫 잘못된 모델이 생성될 가능성이 높기 때문이다. 두 번째 경우는 카테고리 c_i의 주제 단어 w가 다른 카테고리의 문서들에는 나타나지 않는 경우로 이때는 카테고리의 차별성을 높이기 위해서 $Pr(w|c_i)$을 충분히 높은 값으로 설정할 필요가 있다. 따라서 (그림 1)의 Line 16에서 보는 바와 같이 $TF(w,c_i)$를 해당 카테고리의 최대 TF값으로 설정한다.

Algorithm: Feature_Weighting

입력: 기존 나이브 베이즈 모델 $\widehat{\theta_{NB}} = \{\widehat{\theta_{w|c}}, \widehat{\theta_c}\}$
어휘 V
카테고리 c의 어휘 $V(c)$
카테고리 집합 C

출력: 개선된 나이브 베이즈 모델

```
    BEGIN
1     각각의 단어 w와 카테고리 c에 대해서 χ²(c, w)를 계산
2     For each c_i ∈ C {
3           V_χ²(c_i) ← 집합V에서 가장 높은 χ²값을 갖는 M개의 단어를 선택
4     }
5     For each c_i ∈ C {
6           SumOfTF(c_i) ← Σ_{w∈V} TF(c_i, w)
7           For each w_x ∈ V_χ²(c_i) { /* 각각의 주제 단어에 대해서*/
8           if (w_x ∈ V_χ²(c_j) (i ≠ j) ) continue
9           if (w_x ∈ V(c_j) (i ≠ j)) then {
10            MaxPr(c_i) = max_{c_j ≠ c_i} Pr(w_x|c_j)
11            α = ⌈ ((MaxPr(c_i)+δ)*(SumofTF(c_i)+|V|) − TF(w_x,c_i) − 1) / (1 − MaxPr(c_i) − δ) ⌉
12            TF(w_x, c_i) += α
13            SumOfTF(c_i) += α
14          }
15          else
16            TF(w_x, c_i) ← max_{w∈V_χ²(c_i)} TF(w, c_i) + 1
17        } /* The end of for each */
18    } /* The end of for each */
    END
```

(그림 1) 특징가중치를 이용한 나이브 베이즈 분류모델 개선알고리즘

3.3 점진적 특징 가중치를 위한 특징 갱신 기법

(그림 1)의 알고리즘은 기존의 나이브 베이즈 분류모델과 어휘 집합에 대해서 모델을 개선하는 과정을 보여줬다. 앞서 언급한 바와 같이 대부분의 자동문서분류 시스템에서는 점진적으로 새로운 학습문서들이 추가되는 환경을 가정한다. 따라서 새롭게 문서들이 추가될 경우 기존의 어휘에 없는 새로운 단어들이 추가하기도 하고 기존 카테고리들의 단어들에 대한 통계량도 변하게 된다. 결국 이러한 갱신된 정보는 (그림 1)에서 χ^2-통계량을 재계산하는 입력인자로 사용되며 최종적으로 특징 가중치를 이용한 점진적인 분류모델의 개선이 가능해진다.

(그림 2)는 새로운 학습문서가 추가될 때 각 카테고리에 대한 통계량을 갱신하는 알고리즘을 보여준다. 이 알고리즘에서 보는 바와 같이 카테고리 c에 새로운 학습문서 d_{new}가 추가될 때, 우선적으로 d_{new}의 각 단어 w가 기존의 어휘에 없는 새로운 단어라면 이러한 단어들을 어휘 V에 추가한다. 다음으로 이러한 단어에 대해서 수식 (6)의 각 DF 값들을 초기화하는데, 우선 $DF(w,c)$와 $DF(\overline{w},c)$는 알고리즘 후반부의 계산을 위해 0으로 초기화 된다. 그리고 $DF(w,\overline{c})$는 c이외의 카테고리에서 w가 존재하는 문서의 수를 의미하는데, w는 기존에 존재하지 않았던 새로운 단어이므로 이 값은 0으로 설정된다. 마지막으로 $DF(\overline{w},\overline{c})$는 c이외의 카테고리에서 w가 존재하지 않는 문서의 수를 나타내므로 같은 이유로 이 값은 $|D(\overline{c})|$가 된다. 여기서 $|D(\overline{c})|$는 c를 제외한 카테고리에 존재하는 총 문서의 수를 의미한다.

다음 단계로 V의 각 단어 w에 대해서 TF 통계량과 수식 (6)의 각 DF 값 중에서 카테고리 c와 관련된 통계량을 재계산한다. 나이브 베이즈 문서분류기에서는 단순히 모델인자인 $Pr(w|c)$을 추정하면 되므로, 각 카테고리에 대한 TF 통계량만을 수정함으로써 모델인자의 개선이 가능하게 된다. 따라서 $TF(w,d_{new})$와 $TF(w,c)$을 각각 문서 d_{new}와 카테고리 c에 나타나는 단어 w의 출현 빈도수라 할 때, (그림 2)의 line 10의 수식과 같이 새로운 문서 d_{new}에 존재하는 각 단어에 대해서 $TF(w,c)$의 값을 $TF(w,d_{new})$만큼 증가시킨다.

다음으로 특징들의 우선순위를 정하기 위해서 χ^2-통계량의 계산을 위한 인자들을 수정해야하는데, (그림 2)의 Line 11 과 Line 14에서 보는 바와 같이 수식 (6)의 $DF(w,c)$, $DF(\overline{w},c)$ 값들을 수정하게 된다. 즉, V의 각 단어 w가 d_{new}에 나타나면 $DF(w,c)$을 1 증가시키고, 그렇지 않으면 $DF(\overline{w},c)$을 1만큼 증가시킨다. 하지만 새로운 문서 d_{new}는 카테고리 c에 추가된 것으로 나머지 카테고리들에는 변화가 없으므로 $DF(w,\overline{c})$와 $DF(\overline{w},\overline{c})$는 수정이 필요없다.

이와 같이 (그림 2)의 특징갱신 알고리즘은 학습문서의 추가에 따라 DF와 TF를 점진적으로 관리한다. 정리하면 초기 학습문서 집합에 대한 나이브 베이즈 모델은 (그림 1)의 특징 가중치 알고리즘을 이용하여 설정하고, 이후에 추가되는 학습문서에 대해서는 (그림 2)의 특징갱신 알고리즘

Algorithm: Feature_Update

입력: 새 문서 $<d_{new},c>$
어휘 V
카테고리 c의 어휘 $V(c)$
카테고리 집합 C

출력: 갱신된 TF와 DF 통계량

```
    BEGIN
1   For each w∈d_new {
2       if (w∉V) then {
3           V ← V∪w
4           DF(w,c) ← 0          DF(w̄,c) ← 0
5           DF(w,c̄) ← 0          DF(w̄,c̄) ← |D(c̄)|
6       }
7   }
8   For each w∈V {
9       if (w∈d_new) then {
10          TF(w,c) ← TF(w,c) + TF(w,d_new)
11          DF(w,c) ← DF(w,c) + 1
12      }
13      else
14          DF(w̄,c) ← DF(w̄,c) + 1
15  }
    END
```

(그림 2) 특징 갱신 알고리즘

을 이용하여 DF와 TF를 점진적으로 관리하게 된다. 이렇게 수정된 DF와 TF는 다시 (그림 1)의 특징 가중치 알고리즘을 반복적으로 적용하여 나이브 베이즈 분류모델을 점진적으로 개선하게 된다. 이와 같이 각 단어와 카테고리에 대한 통계량만으로 분류모델의 갱신이 이루어지므로 기존 이미 학습된 문서를 재학습할 필요가 없는 장점을 갖게 된다.

4. 성능 분석

4.1 실험 환경

본 논문에서는 제안한 알고리즘의 성능을 평가하기 위해서 Reuters-21578과 20Newsgroups을 이용한 실험을 실시하였다. Reuters-21578과 20Newsgroups는 문서분류의 성능의 평가하기 위해 일반적으로 많이 사용되는 문서집합이다.[3] Reuters-21578는 135개의 카테고리로 묶여진 21578개의 뉴스 기사들로 구성되어 있으며, 20Newsgroups는 20개의 카테고리를 갖는 19,997개의 유즈넷(Usenet) 기사로 구성되어 있다. 이 두 문서집합은 전혀 다른 성격을 갖는다. Reuters-21578는 상대적으로 작은 크기의 문서(평균 861 bytes)들로 구성되어 있으며 카테고리 간에 문서들의 분포가 균형 잡히

3) http://kdd.ics.uci.edu/summary.data.application.html

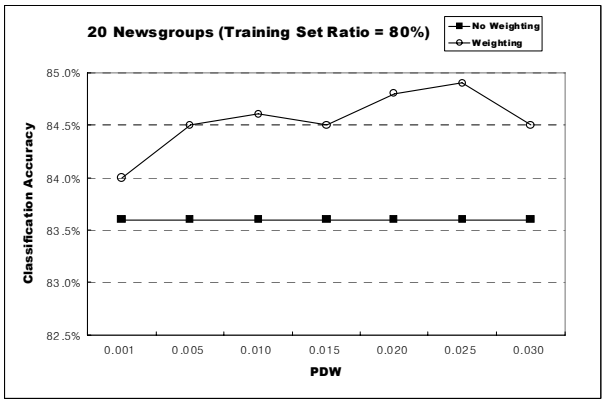
지 못하고 특정 카테고리들에 집중되는 경향을 보인다. 이러한 이유로 이 문서집합은 문서분류의 성능문석에 적합하지 못하다는 평가를 받기도 한다. 따라서 본 논문에서는 이러한 불균형 문제를 해소하기 위해서 20개 이하의 문서들로 구성된 카테고리는 실험대상에서 제외하여 53개의 카테고리에 총 9,133개의 문서에 대해서 실험을 실시하였다. 반면에 20Newsgroups는 상대적으로 큰 용량의 문서(평균 1,892 bytes)로 구성되어 있으며, 카테고리 간에 문서들이 비교적 균형 있게 분포되어 있다. 특별한 언급이 없을 경우 각 문서집합에 대해서 기본적으로 80%는 학습을 위해 사용하였고 나머지 20%는 분류성능을 평가하는데 사용하였다. 분류기의 성능은 각 문서가 자신에게 속할 카테고리에 얼마나 정확하게 분류되는가를 기준으로 평가하였으며, 본 논문에서 제안한 특징 가중치 기법이 기존 나이브 베이즈 문서분류기의 성능을 얼마만큼 향상시키는가에 중점을 두고 진행하였다.

4.2 실험 결과

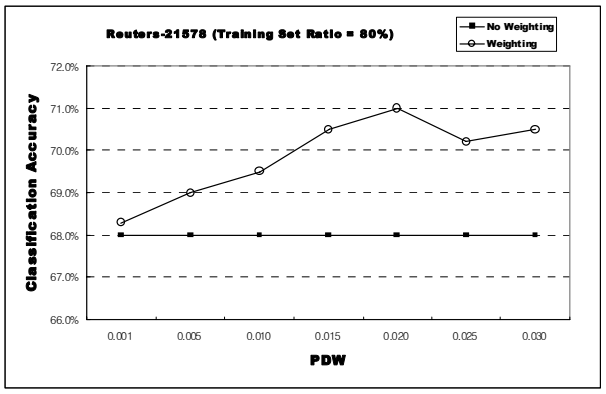
(그림 3)과 (그림 4)는 각각 Reuters-21578과 20Newsgroups 문서집합에서 PDW의 변화에 따른 분류 정확도의 변화를 보여주고 있다. 실험은 10회를 반복하여 실시하였으며 각 그래프의 데이타 값은 반복된 실험 결과에 대한 평균값을 보여준다. 우선 Reuters-21578을 이용한 실험 결과를 보면, 문서수가 매우 적은 카테고리를 실험대상에서 제외했음에도 불구하고 기존의 전통적인 문서분류기에서는 문서들의 불균형된 분포로 인해 분류 성능이 매우 낮은 것을 볼 수 있다. 하지만 본 논문에서 제안한 방법은 PDW의 값에 거의 비례하여 분류 정확도가 높아지고 있다. 특히 PDW가 0.02에서 0.03의 구간에서는 PDW의 값에 따른 변화가 있지만 비교적 좋은 성능을 보여주고 있다. PDW가 0에 가까우면 특징 가중치가 거의 부여되지 않으므로 기존 방법에 비해 큰 성능 향상을 기대하기 어렵고, 반대로 너무 클 경우 일부 특징만으로 전체 분류 모델의 성능이 좌지우지될 수 있어 오히려 역효과가 날 가능성이 높다. 또한 PDW가 너무 크면 가중치 부과 과정에 지나치게 많은 시간이 소요되는 것도 고려해야한다.

다음으로 (그림 4)의 20Newsgroups에 대한 실험 결과를 보면 전반적인 분류 성능이 Reuters-21578보다 좋은 것을 알 수 있다. 그 이유는 앞서 설명한 바와 같이 학습문서들이 비교적 균형 있게 분포되어 있기 때문이다. 단, 특징 가중치를 부여하여 얻을 수 있는 성능 향상의 정도는 Reuters-21578에 비해 그 효과가 미미하지만 PDW에 변화에 큰 영향 없이 전반적으로 고르게 향상되는 것을 알 수 있다.

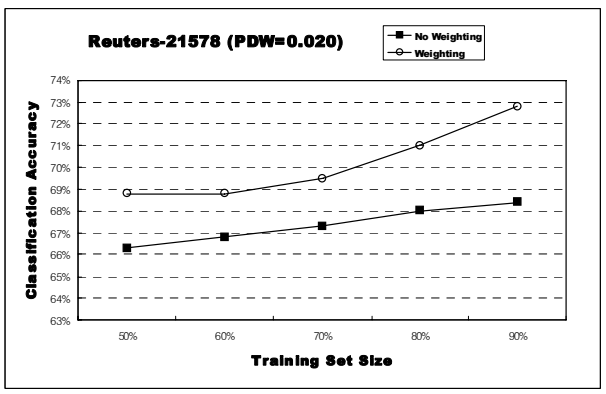
다음으로 (그림 5)와 (그림 6)은 학습문서 집합의 비율에 따라 기존의 문서분류기와 특징 가중치를 부여한 분류기의 성능을 보여준다. 이 실험에서는 PDW의 값을 0.02로 고정하였는데 그 이유는 두 문서 집합 모두에서 이 값에서 가장 우수한 분류 성능을 보였기 때문이다. 이 그림에서 보는 바와 같이 Reuters-21578와 20Newsgroups에 모두에서 학습문서의 비율이 높아질수록 문서분류기의 성능이 대체로 일정하게 향상되는 것을 알 수 있다. 또한 특징 가중치를 적용한 방법이 기존의 방법에 비한 성능향상도 비교적 균일한 정도로 증가되는 것을 알 수 있다. 다만 Reuters-21578의 경우 학습문서 집합의 비율이 높아질수록 특징 가중치를 적용한 방법의 성능 향상 정도가 20Newsgroups에 비해 더 두드러지게 나타나고 있다. 이러한 결과는 학습문서의 비율과 함께 특징 집합의 규모가 커지면서 특징 집합의 분포를 더 정확하게 예측할 수 있기 때문이다. 이는 실제 온라인 환경의 문서집합이 Reuters-21578 문서집합과 유사한 패턴을 보인다는 점에서 고무적인 결과라고 할 수 있다.



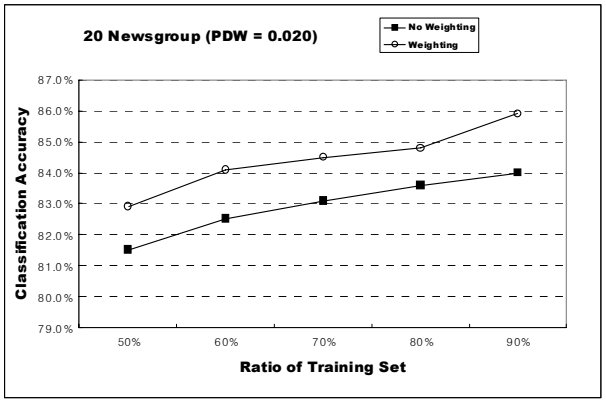
(그림 4) PDW의 변화에 대한 분류 정확도 (20Newsgroups)



(그림 3) PDW의 변화에 대한 분류 정확도 (Reuters-21578)



(그림 5) 학습문서 비율에 따른 분류 정확도 (Reuters-21578)



(그림 6) 학습문서 비율에 따른 분류 정확도 (20Newsgroups))

5. 결 론

본 논문에서는 점진적인 특징 가중치 기법을 이용하여 전통적인 나이브 베이즈 문서분류기의 성능을 향상하기 위해 방법을 제안하였다. 기존의 전통적인 나이브 베이즈 문서분류 방식은 단어들의 분포만을 고려하여 분류모델의 인자를 추정하며 특징들의 중요성 정도는 고려하지 않았다. 본 논문에서는 χ^2-통계량을 이용하여 각 카테고리의 주제가 되는 특징들을 선별한 후, 이 특징들에 대해 다른 카테고리와는 분명한 차별을 가질 수 있도록 가중치를 부과하는 기법을 고안하였다. Reuters-21578, 20Newsgroups 문서집합에 대한 실험을 통해, 제안 기법이 상당한 효과가 있음을 보였으며, 특히 학습문서 비율이 증가함에 따라 그 효과가 커짐을 알 수 있었다. 향후, 이러한 성능 향상에 대한 이론적 근거를 연구할 것이며, 또한 나이브 베이즈 문서분류기의 특징 공간이 동적으로 변하는 환경에서의 특징을 점진적으로 갱신하는 알고리즘을 개발할 계획이다.

참 고 문 헌

[1] M. Hu and B. Liu, "Mining and Summarizing Customer Reviews," Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'04), pp.168-177, 2004.

[2] N. Jindal and B. Liu, "Identifying Comparative Sentences in Text Documents," Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06), pp.244-251, 2006.

[3] T.M. Mitchell, "Bayesian Learning," Machine Learning, McGraw-Hill, pp.154-200, 1997.

[4] E.H. Han, G. Karypis G and V. Kumar, "Text Categorization Using Weight Adjusted k-Nearest Neighbor Classification," Proceedings of The fifth Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD '91), pp.53-65, 1991.

[5] T. Joachims, "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," Proceedings of the 10th European Conference on Machine Learning (ECML'98), pp.137-142, 1998.

[6] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), pp.412-420, 1997.

[7] I. Katakis, G. Tsoumakas and I. Vlahavas, "Dynamic Feature Space and Incremental Feature Selection for the Classification of Textual Data Streams," Proceedings of ECML/PKDD-2006 International Workshop on Knowledge Discovery from Data Streams, pp.107-116, 2006.

[8] K.J. Mock, "Hybrid hill-climbing and Knowledge-based techniques for Intelligent News Filtering," Proceedings of the National Conference on Artificial Intelligence (AAAI'96), pp.48-53, 1996.

[9] A. Kolcz, V. Prabakarmurthi and J. Kalita, "Summarization as Feature Selection for Text Categorization," Proceedings of the 10th International Conference on Information and Knowledge Management (CIKM'01), pp.365-370, 2001.

[10] F. Sebastiani, "Machine learning in automated text categorization," ACM Computing Surveys, Vol.34, No.1, pp.1-47, 2002.

김 한 준

e-mail : khj@uos.ac.kr

1994년 서울대학교 계산통계학과 졸업(학사)

1996년 서울대학교 전산과학과 대학원 졸업(이학석사)

2002년 서울대학교 컴퓨터공학부 대학원 졸업(공학박사)

2002년 2월~2002년 12월 서울대학교 공과대학 Post-Doc

2002년 12월~현 재 서울시립대학교 전자전기컴퓨터공학부 조교수

관심분야 : 데이터베이스, 데이터마이닝, 정보검색

장 재 영

e-mail : jychang@hansung.ac.kr

1992년 서울대학교 계산통계학과 졸업(학사)

1994년 서울대학교 계산통계학과 대학원 졸업(이학석사)

1999년 서울대학교 계산통계학과 대학원 졸업(이학박사)

2000년 3월~현 재 한성대학교 컴퓨터공학과 부교수

관심분야 : 데이터베이스, 데이터마이닝