문서의 주제어별 가중치 부여와 단어 군집을 이용한 한국어 문서 자동 분류 시스템

허 준 희[†] · 최 준 혁[††] · 이 정 현[†††] · 김 중 배[††††] · 임 기 욱[†††††]

요 약

새로운 문서를 기존에 존재하는 클래스들에 할당하는 방법을 문서의 자동 분류라고 한다. 문서의 자동 분류는 뉴스 그룹의 기사분류, 웹 문서의 범주화, 전자 메일의 순서화, 사용자의 관심을 학습하여 보다 정확한 정보 검색 결과를 제시하는데 사용될 수 있다. 본 논문에서는 한국어 문서 분류의 정확도를 높이기 위하여 문서내의 모든 단어들에 대한 확률 값을 사용하여, 문서를 분류하는 기존의 방법과 달리 문서의 주제어를 선정하여 주제어로 선정된 단어들에 가중치를 부여하고 그렇지 않은 단어들에 대해서는 제거하거나 낮은 가중치를 부여하는 베이지안 분류자를 사용한다. 문서에서 특징으로 추출된 단어가 적어 문서를 분류하기 위한 만족할 만한 정보를 제공하지 못할 경우에 부족한 문서의 특징을 보충하기 위하여 말뭉치로부터 자동 단어 군집화를 통해 형성된 연관 단어 군집을 사용한다. 이러한 방법을 한국어 문서에 적용한 결과 기존의 베이지안 확률을 사용한 분류법보다 향상된 분류 정확도를 얻을 수 있었다.

An Automatic Classification System of Korean Documents Using Weight for Keywords of Document and Word Cluster

Jun-Hui Hur[†] · Jun-Hyeog Choi[††] · Jung-Hyun Lee[†††]
Joong-Bae Kim[††††] · Kee-Wook Rim[†††††]

ABSTRACT

The automatic document classification is a method that assigns unlabeled documents to the existing classes. The automatic document classification can be applied to a classification of news group articles, a classification of web documents, showing more precise results of Information Retrieval using a learning of users interests. In this paper, we use the weighted Bayesian classifier that weights with keywords of a document to improve the classification accuracy. If the system can't classify a document properly because of the lack of the number of words as the feature of a document, it uses relevance word cluster to supplement the feature of a document. The word clusters are made by the automatic word clustering from the corpus. As the result, the proposed system outperformed existing classification system in the classification accuracy on Korean documents.

키워드 : 문서분류(Document classification), 베이지안 분류자(Bayesian Classifier), 연관단어군집(Relevance word cluster)

1. 서 론

인터넷의 대중적인 보급은 온라인 문서의 양을 기하급수적으로 증가시키고, 이를 만들고 사용하는 사람의 수를 증가하게 만들었다. 문서의 양과 사용자의 수가 증가함에 따라 자동 문서 분류는 방대한 양의 데이터를 정리하는 사람들을 돕기 위해 점점 중요한 도구가 되었다[1].

확률 통계적 문서 분류 알고리즘은 기사를 분류하고 웹 문서를 범주화하며, 전자 메일을 순서화하고 사용자의 관심을 학습하는데 적용되어 왔다.

문서의 분류 및 클러스터링은 유사한 내용을 가진 문서들을 모아 집단화하는 작업이다[2]. 문서의 자동분류는 사전에 분류체계가 만들어져 있는 상태에서 새로운 문서를 적절한 클래스에 배정하는 것이고, 클러스터링은 사전에 분류된 체계없이 문서들 사이의 유사도 계산을 통해 같은 부류의 문서들을 묶어 군집을 형성하는 것이다.

문서의 자동 분류에 대한 연구 역시, 다른 정보검색에 관련된 연구들처럼 오래 전부터 진행되어 왔지만 이는 거의 모두가 영어권의 나라에서 이루어진 것이 대부분이다. 각 나라의 언어적인 특성을 고려해 볼 때 외국의 연구들을 그대로 한국어 문서 자료 집합에 적용하는 것은 문제가 있기 때문에 한국어의 특성에 맞는 자동 분류 시스템에 대한 연

† 정 회 원 : 아이티나라(주) 기술연구소
†† 종신회원 : 김포대학 컴퓨터계열 교수
††† 종신회원 : 인하대학교 전자계산공학과 교수
†††† 정 회 원 : 한국전자통신연구원 컴퓨터소프트웨어기술연구소
††††† 종신회원 : 선문대학교 산업공학과 교수
논문접수 : 2001년 2월 28일, 심사완료 : 2001년 9월 28일

구가 필요하다. 그러나 정보 검색의 효율 향상을 위한 다른 연구들과는 달리 문서 자동 분류에 대한 연구는 아직까지 미흡하다고 할 수 있다.

자동으로 연관성 있는 문서들을 분류하는 작업은 특히 대용량의 문서 자료 집합으로 구성되는 데이터 베이스의 저장과 검색에 효과적이다. 문서의 자동 분류에 의해 만들어진 데이터 집합을 분리하여 저장하고 검색하는 것은 정보 검색 시스템의 평가 기준이라고 할 수 있는 검색의 정확도 측면이나 검색에 소요되는 시간, 저장 공간의 효율성 면에서 훌륭한 결과를 보여줄 수 있다. 그렇지만 문서 분류의 정확성이 떨어진다면 이러한 기대 효과를 얻지 못할 뿐만 아니라 검색의 질을 낮추는 결과를 가져올 수도 있다.

본 논문에서는 각 언어권에서 일반적으로 가장 높은 분류효율을 보이는 베이지안 확률을 사용한 분류자를 기반으로 한국어의 특성에 맞는 문서 자동 분류 시스템을 설계하고 구현한다. 본 논문에서 설계한 시스템은 먼저 말뭉치를 통해 유사 단어 군집화(Clustering)를 수행하여 문서 분류 단계의 지식 베이스(knowledge-base)로 사용한다. 문서 분류에 있어서는 문서 내에 출현하는 모든 단어의 확률을 계산하는 기존의 방식과 달리, 각 문서를 대표할 만한 주제어(keywords)를 선정하여 그 주제어들에 가중치를 부여하는 베이지안 분류자를 사용한다.

2. 가중치가 부여된 베이지안 분류자와 단어 군집을 사용한 한국어 문서 분류

2.1 텍스트 표현(Text Representation)

문서들은 전형적으로 자연 언어 형식으로 쓰여진 텍스트를 나타내는 캐릭터들의 순서열로 저장된다[4]. 정보검색 분야에서 문서를 표현하는 캐릭터 스트링을 변형하는 방법에 대한 많은 연구가 진행되어져 왔다. 이러한 방법들은 문서 분류의 전처리 방법으로 적용될 수 있다. 이러한 방법들은 음성 인식이나 이미지 처리 등의 분야에서 연구되는 특징 선택과 유사하다.

텍스트를 표현하기 위한 많은 통계학적이고 언어적이며 지식 기반 기술을 사용한 연구들이 정보 검색 분야에서 진행되었다[9]. 그러나 언어적 분석이나 지식 기반이 없는 단순한 문서 표현 방법이 다른 방법들과 비교해 비슷한 성능을 보여준다[5]. 따라서 복잡한 전처리 과정을 거치지 않는 단순한 텍스트 표현법을 사용하더라도 시간과 계산의 복잡성 측면에서 이득이 되면서 다른 복잡한 처리를 가지는 방법들과 비슷한 효율을 얻을 수 있기 때문에 대부분의 정보 검색과 문서 분류 시스템에서는 단순한 단어 모델을 사용한다.

가장 단순하면서도 널리 사용되는 문서 표현법은 텍스트를 단어들의 집합으로 간주하는 방법이다. 이러한 방법에서 고려해야 할 부분은 문서들로부터 단어들을 추출하는 방법이다. 따라서 단어들을 추출하기 위해 기존의 n-gram이나 형태소, 어근 추출(word stem)법 등의 방법들을 쉽게 사용할 수 있다.

한국어 문서 처리에서 대표적으로 사용되고 있는 방법은 형태소 분석법이다.

자연언어 처리 기법에서 개발된 형태소 분석 기법은 한국어 문서를 표현하고 이해하는데 효과적이다. 단어 벡터 모델은 형태소 분석의 복잡한 부분인 파싱(parsing)을 통한 의미 분석을 생략하고 명사만을 추출하여 구성할 수 있다. 따라서 본 논문에서는 텍스트로 이루어진 문서를 표현하기 위해 형태소 분석을 통한 명사 추출 과정을 전처리 과정으로 사용하여 각 문서를 단어들의 집합, 즉 단어 벡터 모델로 나타낸다.

2.2 자동 단어 군집화

본 논문에서는 문서의 길이가 짧아 추출된 명사의 수가 기준치 이하일 경우 부족한 특징을 보충하기 위해 자동으로 구축된 단어 군집을 사용한다. 자동으로 단어 군집을 구축하기 위한 기존의 연구들은 단어의 각 의미별로 정확한 군집화를 통하여 정보 검색에 있어 사용자의 질의 확장이나 언어 모델링(language modeling), 시소러스(thesaurus) 구축에 이용하였다[7]. 그러나 본 논문에서는 단어 군집을 부족한 특징을 보충하는데 사용하기 때문에 기존의 연구들이 추구했던 군집화의 정확도, 즉 중의성이 있는 단어에 대해 조금 덜 제약적이다.

단어들 간의 유사성을 측정하기 위하여 본 논문에서는 단어들의 연관성을 정량적으로 나타내는 상호 정보 수식을 사용한다[10]. 크기가 N인 말뭉치에서 단어 x와 y가 출현한 횟수를 각각 $f(x), f(y)$라 하고 x와 y가 한 문장에서 함께 출현한 빈도 수를 $f(x,y)$라고 했을 때 단어 x와 y의 상호정보량은 식 (1)과 같이 정의할 수 있다.

$$I(x,y) = \log \frac{P(x,y)}{P(x)P(y)} \approx \log \frac{N*f(x,y)}{f(x)f(y)} \quad (1)$$

단어들 간의 유사도를 기반으로 단어들을 클러스터링하는 알고리즘으로는 집적적인 클러스터링(agglomerative clustering)[11], k-means 알고리즘[9], EM(Expectation Maximization) 알고리즘[12], 모의 담금질(simulated annealing)[8], k-NN(k-Nearest Neighbor) 알고리즘[2] 등이 있다.

본 논문에서는 상호정보를 단어들의 유사도 측정방법으로 사용하여 집적적인 클러스터링 알고리즘을 통해 단어 군집을 형성한다. 이와같은 방법으로 군집화된 단어들은 문서에서 추출된 특징 단어의 수가 부족할 때 이러한 부족분을 보충하기 위해 사용된다.

2.3 특징 선택 및 사전(Vocabulary)구성

문서를 분류하는데 있어서 문서의 특징이 될 수 있는 단어들을 추출하는 것은 중요하다. 또한 학습 집합에서의 특

징 추출은 사전의 크기를 줄이는데 유용하게 사용할 수 있다. 기존의 일반적인 학습 방법들은 학습 문서에 출현하는 모든 용어를 대상으로 사전을 구축하였기 때문에 분류에 영향을 미치지 않는 단어들에 대한 정보도 유지하게 된다. 이러한 사전 구성은 시스템의 저장 공간의 문제와 수행 속도, 계산의 복잡성 측면에서 비효율적이라고 할 수 있으며 문서 분류에 대한 오 분류의 요인으로 작용할 수 있다.

본 논문에서는 각각의 문서에 대한 특징이 아닌 클래스 특징을 반영하기 위해 클래스 변수와 단어들간의 상호정보를 이용해 특징을 추출하고 사전을 구성하게 되며, 이때 사용하는 식은 식 (2)과 같다[3].

$$I(C;W_t)=p(c,f_t)\log(\frac{p(c,f_t)}{p(c)p(f_t)}) \quad (2)$$

여기서,

$$p(c,f_t)=\frac{frequency\ of\ w_t\ where\ class\ label\ is\ c}{number\ of\ total\ words}$$

$$p(c)=\frac{number\ of\ total\ words\ where\ class\ label\ is\ c}{number\ of\ total\ words}$$

$$p(f_t)=\frac{number\ of\ w_t\ occurence}{number\ of\ total\ words}$$ 이다.

2.4 학습 데이터로부터의 naive Bayes 학습

베이지안 분류자는 이전 데이터로부터 문서의 표현 집합이 되는 단어들의 확률 값을 통해 새로운 문서의 확률 값을 추정하게 된다.

"naive Bayes 가정"에 기반하여 문서를 분류하는 연구들은 크게 두 가지 방법으로 나눌 수 있다. 첫 번째 방법은 문서내의 단어들의 발생과 비 발생만을 고려하여 문서를 분류하는 방법이고, 두 번째 방법은 문서내의 단어의 발생과 비 발생뿐만 아니라 해당 단어의 출현빈도까지 고려하는 방법이다. 첫 번째 방법을 일반적으로 이진 독립 모델(Binary Independence Model)이라 칭하거나 특별히 문서 분류에 있어서 Multi-variate Bernoulli Model이라고도 한다. 두 번째 방법은 일반적으로 다항 모델(multinomial Model)이라 부른다[3, 4].

본 논문에서는 학습 문서들로부터 사전 확률 값을 계산하기 위해 단어의 발생여부만을 사용하는 방법이 아닌 단어의 출현빈도를 고려하는 다항(multinomial) 베이지안 학습법을 사용한다[13]. 학습 알고리즘으로는 형태소 분석을 통해 선택된 명사와 클래스 변수와의 상호정보 계산을 통해 추출된 특징을 이용하여 학습을 수행하는 (알고리즘 1)을 사용하였다.

```
Voc ← words through feature selection
Data ← training data set
for each class variable {
        docs_j ← documents which class variable is v_j
        P(v_j) ← |docs_j| / |Data|   // 각 클래스의 출현확률
        V ← total words in docs_j
        n ← number of words in V
        /* 각 단어에 대한 추정치 계산 */
        for each w_k in Voc{
            n_k ← frequency of w_k in V
            P(w_k|v_j) ← (n_k + 1) / (n + |Voc|)
        }
}
```

(알고리즘 1) 특징이 추출된 사전을 이용한 베이지안 학습 알고리즘

2.5 가중치가 부여된 베이지안 분류자

문서를 대표할 만한 단어 또는 어구를 찾는 작업은 검색의 정확도 측면에 있어 매우 중요하다. 이러한 작업을 자동 색인(indexing)이라고 하는데 자동 색인에 대한 연구는 상당히 오랜 기간 이루어져 왔지만 한국어 문서의 색인의 경우 복합 명사 색인의 문제점 때문에 그다지 만족할 만한 결과를 얻지 못하고 있다.

본 논문에서는 문서 분류에 자동 색인 분야에서 연구된 가중치 부여 식을 분류자에 첨가하여 문서 분류에 사용한다. 베이지안 분류자를 이용한 기존의 방식들은 문서 내에 출현하는 모든 단어들의 확률 값의 곱으로 분류 확률을 계산하기 때문에 문서의 오분류에 영향을 줄 수 있는 단어들, 즉 잡음(noise)들까지 확률 계산에 포함하게 된다. 따라서 기존의 분류 방법에서 분류 효율의 저하를 유도하였다. 이에 본 논문에서는 문서를 대표할 만한 주제어들을 사전에서 추출하여 문서의 주제에 관련이 없는 단어들을 제거하거나 낮은 가중치 값을 부여하여 분류 오류를 줄이고자 하였다.

본 논문에서 제안하는 한국어 문서 자동 분류 방법은 다음과 같다.

먼저, 분류될 문서가 주어지면 형태소 분석을 수행하여 주제어 후보들을 생성한 다음 식 (3)의 가중치 계산을 통하여 주제어를 추출한다[6].

$$W_i = TF \cdot IDF = TF_i(\log_2(n)-\log_2(DF_i)+1) \quad (3)$$

만일 문서의 크기가 작아 문서의 특징을 나타내는 주제어들이 최소 특징 수보다 부족하면, 부족한 특징을 보충하기 위해 말뭉치(corpus)로부터 단어 군집화를 통해 만들어진 유사단어 쌍을 이용하여 부족한 특징을 보충하여 데이터의 희소성을 해소한다.

```
Num_words ← # of words in document ;
w_k ← each word in document ;
Num_class ← # of class variable ;
z ← # of feature ;
```

```
for (k = 1 ; k ≤ Num_words){
        W_k ← Calc_Weight (w_k) ;
        Buf ← Save_Buf (W_k) ;
}
Sort(Buf) ; // TFIDF 값이 큰 순서로 정렬
Select(z) ; // 상위 z개를 주제어로 선택
/* 주제어의 수가 부족하면 단어 군집으로부터 부족한 특징
   을 보충 */
if(feature number < z)
        then(supplement keywords) ;
for (j = 1 ; j ≤ Num_class){
        for (each keyword){
            Prob ← Prob * Calc_W (w_i) * Calc_P (w_i|v_j) ;
        }
        V ← Calc_P (v_j) * Prob ;
        return Max(V) ; // 최대확률 값
}
Assign(document) ; // V값이 최대인 클래스에 할당
```

(알고리즘 2) 주제어별 가중치가 부여된 분류자를 사용한 분류 알고리즘

문서의 특징추출로 각 주제어가 선택되었으면 식 (4)의 각 주제어들의 가중치를 고려하는 베이지안 분류자를 통해 확률 값이 가장 높은 클래스에 문서를 할당하게 된다.

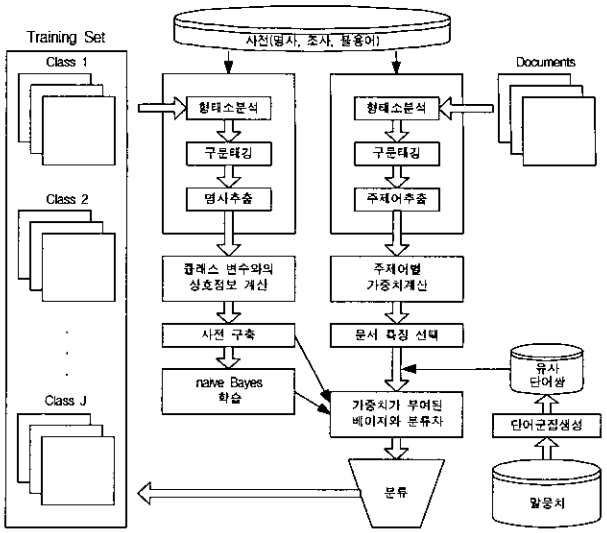
$$v_{NB} = \arg max_{v_j \in V} P(v_j) \prod_{i \in I} W_i P(a_i|v_j) \qquad (4)$$

여기서, $\prod_{i \in I} P(a_i|v_j) = P(w_1|v_j)\ P(w_2|v_j) \cdots P(w_i|v_j)$이고, W_i는 식 (3)와 같이 계산된다. 이때, DF_i는 전체 문서가 아닌 각 클래스의 문서들로부터 얻어진다. (알고리즘 2)는 본 논문에서 설계한 자동분류 알고리즘이다.

3. 전체 시스템 설계

(그림 1)은 본 논문에서 설계한 문서 자동분류 시스템에 대한 구성도이다.

본 논문에서 설계한 시스템은 먼저 지식 베이스를 구축하기 위해 말뭉치로부터 단어 군집화를 수행하여 연관있는 단어의 집합(단어 쌍의 형태)으로 구성한다. 학습 단계에서는 미리 분류된 학습 문서를 대상으로 문서를 문서 내에 출현하는 단어 벡터의 형태로 나타내기 위해 형태소 분석을 수행한 후, 형태소 분석된 학습 문서들로부터 명사로 품사 태깅된 단어들을 추출하여 전체 사전을 구성한다. 각 클래스별로 클래스 변수와의 상호정보를 계산한 후 그 값이 높은 순서대로 정렬된 클래스 사전을 구축한다. 사전이 구축되었으면 레이블 된 학습 데이터의 클래스로부터 각 단어에 대한 베이지안 추정치를 계산하여 저장한다. 분류 단계에서는 새롭게 분류될 문서로부터 주제어를 추출한다. 이렇게 선택된 주제어의 수가 역치(threshold) 이하이면 군집화된 단어 쌍으로부터 얻은 연관 단어들을 문서의 특징으로 분류정보에 추가한다. 최종적으로 각 클래스에 대해 문서의 주제어별 가중치를 계산하고 가중치가 부여된 베이지안 분류자를 통해 계산된 확률 값이 가장 높은 클래스에 문서를 할당한다.



(그림 1) 가중치 부여와 단어 군집을 이용한 문서분류 시스템

4. 실험 및 평가

4.1 실험 환경

실험 환경으로는 Windows 2000 서버를 사용하였으며, 시스템의 구현을 위해 MS Visual C++ 6.0을 사용하였다. 실험 데이터로는 한국어 정보검색 시스템의 성능 평가용 데이터 집합인 KTset95 문서 4,414개 중 1,300개의 문서를 학습 집합으로, 1,000개의 문서를 실험 집합으로 사용하여 실험을 수행하였다. 학습 집합의 클래스는 수작업으로 전산학 각 연구 분야의 14개 클래스로 분류하였다. KTset95 문서 중 정의된 클래스에 해당하지 않는 문서들은 사용하지 않았다. 학습 문서들로부터 78,156개의 용어들이 추출되었고, 이 중에서 중복된 명사들을 제거한 후 클래스 변수들과의 상호정보를 계산하여 총 3,750개의 단어들로 초기 사전을 구성하였다.

4.2 실험 결과

<표 1>은 학습 집합을 말뭉치로 사용하여 연관 단어 군집화를 수행한 결과이다. 군집화 결과, 유사한 단어 또는 문장 내에 공기(co-occurrence)하는 단어들이 하나의 군집을 형성하고 있음을 알 수 있다.

〈표 1〉 연관 단어 군집(일부)

군 집	단 어
1	[컴퓨터, 시스템, 프로그램, 기술, 파일, 데이터, …]
2	[자연언어, 형태소, 한글, 국어, 한국, 음절, 음소, …]
3	[문자, 패턴, 인식, 화상, 얼굴, 필기체, 지문, …]
4	[연구, 논문, 프로젝트, 보고서, 개선, 향상, …]
5	[교환망, 네트워크, 패킷, 프로토콜, 통신, 전송, …]
6	[정보, 검색, 질의, 색인, 유사도, 정확도, 재현율, …]

<id>0537
<title>한국어 형태소 분석기의 정형화
<abstract>한국어와 일반 자연언어 처리에 적합한 범용 형태소 분석기를 구현하였다. 한정된 영역에서 선형 우선 결합하는 형태소의 결합특성을 분석하고, 이진선형 결합관계를 중심으로 한 형태소 분석의 새로운 방법을 제시하였다. 본 논문의 방법은 자동 형태소 분석기의 생성 등 형태소와 자연언어 처리의 정형화에 크게 기여할 것으로 기대된다.

(그림 2) 학습 문서 원문

[한국어와 ((한국어 와) (N PI)) ((한국어 와) (N PCA))]
[일반 ((일반) (N))]
[자연언어 ((자연언어) (N))]
[처리에 ((처리 에) (N PCA))]
[적합한 ((적합 하 ㄴ) (NH AS ED))]
[범용 ((범용) (N))]
[형태소 ((형태소) (N))]
[분석기를 ((분석기 를) (N PCO))]
[구현하였다 ((구현 하 였 다) (NH AS EPF EE))]
.
.
.
[형태소와 ((형태소 와) (N PI)) ((형태소 와) (N PCA))]
[자연언어 ((자연언어) (N))]
[처리의 ((처리 의) (N PCD))]
[정형화에 ((정형 화 에) (N N PCA))]
[크게 ((크 게) (A EC))]
[기여할 ((기 여 하 ㄹ) (V EA VA ED)) ((기 여 하 ㄹ) (V EA AA ED)) ((기여 하 ㄹ) (NH AS ED))]
[것으로 ((것 으로) (ND PCA))]
[기대된다 ((기 대 되 ㄴ 다) (V EE VA EPF EE)) ((기대 되 ㄴ 다) (NH AS EPF EE)) ((기대 어 되 ㄴ 다) (V EC VA EPF EE)) ((기대 되 ㄴ다) (NH AS EE))]

(그림 3) 원문의 형태소 분석 결과

학습 데이터로부터 각 클래스별 사전을 구축하고 학습을 통해 각 단어들의 추정치를 계산하는 방법은 다음과 같다.

(그림 2)는 실험 데이터 중 학습 문서로 사용된 KTset95 문서 원문의 예로 14개의 클래스 중 자연어 처리에 해당하는 문서이다. (그림 3)은 (그림 2)의 원문을 형태소 분석하여 품사 태깅한 결과이고 (그림 4)는 형태소 분석 결과로부터 명사와 그 출현 빈도 수를 계산한 결과이다.

학습 문서들에 대한 형태소 분석 및 명사추출과정이 수행된 후에 시스템은 각 클래스 변수와 그 클래스로 레이블된 문서들로부터 추출된 명사들을 대상으로 식 (2)를 사용하여 상호정보를 계산하고 사전을 구성한다.

한국어 3, 형태소 7, 분석기 3, 정형 2, 일반 1, 자연언어 2, 처리 2, 적합 1, 범용 1, 한정 1, 영역 1, 선형 2, 우선 1, 우선 1, 결합 3, 특성 1, 분석 3, 이진 1, 관계 1, 방법 2, 논문 1, 자동 1, 생성 1, 기여 1

(그림 4) 추출된 명사 및 빈도

<표 2>는 "자연어 처리" 클래스에 해당되는 문서들로부터 추출된 명사들과 클래스 변수와의 상호정보를 계산한 결과 중 일부이다.

<표 2> "자연어 처리"와 단어들간의 상호정보(일부)

단 어	전체 빈도	클래스 빈도	상호정보
한 글	255	167	0.4150
문 자	263	133	0.3300
한국어	79	60	0.1475
형태소	63	59	0.1450
필 기	92	48	0.1175
문 법	41	23	0.5500
텍스트	72	21	0.5000
음 절	21	18	0.4250

각 클래스별로 사전구성이 끝나면 (알고리즘 1)을 수행하여 사전 내의 각 단어들에 대한 추정 값을 구한다. 다음의 (그림 5)와 (그림 6)은 각각 "자연어 처리"와 "네트워크"클래스의 각 단어들의 추정 값의 일부이다.

(그림 5)와 (그림 6)으로부터 각 클래스 변수와 관련이 많은 단어일수록 높은 확률 값을 가지고 있음을 알 수 있고, 이러한 사실은 확률에 의해 문서를 분류하는 방법이 유효하다는 것을 실험적으로 증명해 주고 있다. 각 클래스 사전 내의 단어들에 대해 위와 같이 추정 값들이 구해지면 문서분류에 사용하기 위해 데이터베이스에 저장하게 된다.

문서 분류 단계에서는 분류될 문서에 대해 학습 문서에서처럼 형태소 분석을 통하여 문서를 대표하게 될 단어들을 선정하고 식 (3)을 사용하여 문서의 주제어를 선정한다.

자 연 ; 0.00322
한 글 ; 0.01260
질 의 어 ; 0.00082
처 리 ; 0.00757
지 식 ; 0.00157
추 출 ; 0.00270
자 연 어 ; 0.00120
한 국 어 ; 0.00457
연 구 ; 0.00630
구 문 ; 0.00202
서 술 ; 0.00030
알고리즘 ; 0.00277
부 분 ; 0.00097
분 리 ; 0.00142
영 한 ; 0.00030
구 축 ; 0.00060
갱 신 ; 0.00075
확 장 ; 0.00150
통 한 ; 0.00052
액 세 스 ; 0.00015
.
.
.

(그림 5) 추정치(자연어처리)

분 산 ; 0.00398
회 선 ; 0.00026
교 환 ; 0.00250
설 계 ; 0.00776
지 연 ; 0.00167
네트워크 ; 0.00423
계 층 ; 0.00295
프로토콜 ; 0.01154
통 신 망 ; 0.00545
제 안 ; 0.00532
성 능 ; 0.00731
분 석 ; 0.00635
응답시간 ; 0.00013
최 적 화 ; 0.00032
통 계 적 ; 0.00019
신호처리 ; 0.00026
메카니즘 ; 0.00077
교 환 기 ; 0.00173
링 크 ; 0.00115
드라이버 ; 0.00038
.
.
.

(그림 6) 추정치(네트워크)

(그림 7)은 분류 실험을 위해 사용된 KTset95 문서 중 한 예이고, (그림 8)은 (그림 7)의 원문으로부터 선정된 주제어들의 리스트 중 일부이다.

분류될 문서의 단어 추출과 주제어 선정 작업이 끝나면 각 클래스별로 식 (4)를 통해 그 값을 구하여 가장 높은 값을 갖는 클래스에 해당문서를 배정하게 된다.

<id>3667
<title>지능형 정보 검색에 관한 연구 [Intelligent infomation retrieval environment]
<abstract>본 연구는 검색효율이 우수한 시소러스기반 자동 색인 시스템을 개발하고 등위접속 애매성 해소 기법 연구와 애매성 해소 지식베이스 구축,질의어 재구성 및 학습 연구 그리고 정보검색용 전자사전 구축과 정보검색용 시소러스 구성을 목적으로 한다. 색인과 검색,그리고 자연언어 인터페이스 분야에서는 시스템 사전과 시소러스 구축 분야는 비교적 계획에 맞춰 정보량을 증가시키고 있다. 그러나 사전관리도구의 연구는 미흡하다. 현재 전체적으로 약 55%의 연구 진척율을 보이고 있다. 본 연구가 정보 과학과 전기 전자 분야의 지능형 정보 검색 시스템 연구에 기초자료로 활용되리라 전망한다.

(그림 7) 실험 문서 원문

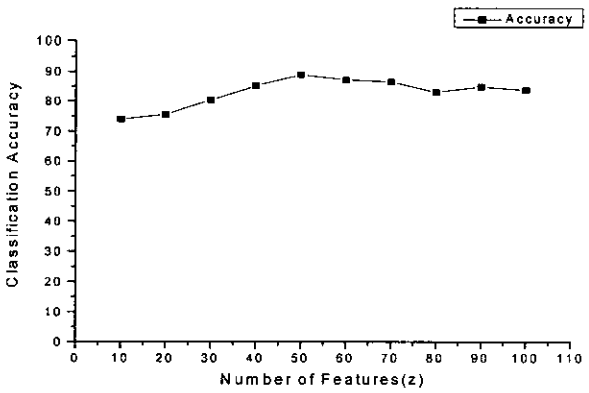
검 색 : 5,36,21.60964
사 전 : 4,23,20.00000
애 매 성 : 2,1,19.08219
색 인 : 3,11,18.19827
해 소 : 2,6,13.90839
과 학 : 2,23,10.00000
구 축 : 4,146,9.28771
연 구 진 : 1,2,8.53916
전 체 : 2,40,8.33985
질 의 어 : 1,4,7.53916
자 연 언 어 : 1,4,7.53916
지식베이스 : 1,7,6.72792
.
.
.

(그림 8) 주제어 후보 리스트

4.3 결과 분석 및 평가

본 논문에서 제안한 문서 분류방법과 기존의 분류법에 대한 비교를 위해 먼저 실험에 의해 문서의 특징 값을 정하였다. (그림 9)는 특징의 개수, 즉 (알고리즘 2)의 z값을 10부터 100까지 증가시키며 각 특징 값에서의 분류 효율을 계산한 결과의 그래프이다. 그래프에서 y축의 분류 효율(classification accuracy)은 정확하게 분류된 문서 수를 전체 문서수로 나누어 백분율로 나타내었다. 이 비교 실험에서 사용된 사전의 크기는 2,500개로 고정하였다.

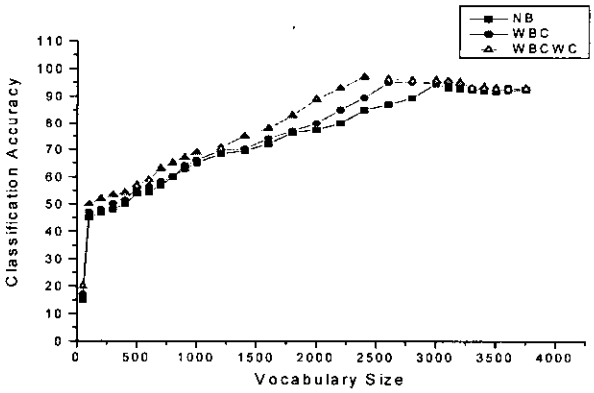
(그림 9)의 그래프 상에서 보면 특징의 개수를 50개로 고정하였을 때의 분류 효율이 가장 우수함을 알 수 있다. 따라서 본 논문에서는 분류될 각 문서에 대한 특징 값을 50으로 지정하여 분류 실험을 수행하였다. 즉, 실험 집합의 각 문서에 대해 특징을 추출하여 높은 가중치 값을 가지는 단어순으로 정렬하여 그 수가 50을 넘으면 상위 50개까지만 문서의 특징을 나타내는 벡터로 구성하고 그렇지 않으면 50개의 특징을 만족하도록 단어 군집으로부터 단어의 수를 보충한다.



(그림 9) 특징 크기에 따른 분류 정확도

(그림 10)은 학습 집합에 대한 분류 성능을 나타내는 그래프이다. 그림에서 NB는 naive Bayes, 즉 기존의 베이지안 확률을 사용한 시스템을 뜻하고 WBC는 본 논문에서 제안한 가중치가 부여된 베이지안 분류자(Weighted Bayesian Classifier)를 통해 분류 실험을 한 결과이다. WBCWC (Weighted Bayesian Classifier with Word Cluster)는 가중치가 부여된 베이지안 분류자와 단어 군집을 통해 부족한 특징을 보충하는 시스템에 대한 결과이다.

그래프로부터 알수 있듯이 전체적으로 가중치 부여와 단어 군집을 사용한 분류방법의 성능이 가장 우수하고, 분류자에 가중치만 부여하더라도 기존의 문서내의 모든 단어의 확률을 고려하는 분류자보다 분류 효율이 향상됨을 알 수 있다. 또한 본 논문에서 제안한 방법이 기존의 방법보다 적은 사전 크기에서 높은 분류 효율을 얻을 수 있었다. 이러한 사실은 학습 집합의 수에 따라 분류 효율에 많은 영향을 받는 확률을 사용하는 분류법에서 중요하다. 따라서 본 논문에서 제안한 방법으로 문서를 분류할 때 기존의 방법보다 적은 학습 집합을 사용하더라도 향상된 분류 정확도를 얻을 수 있다.

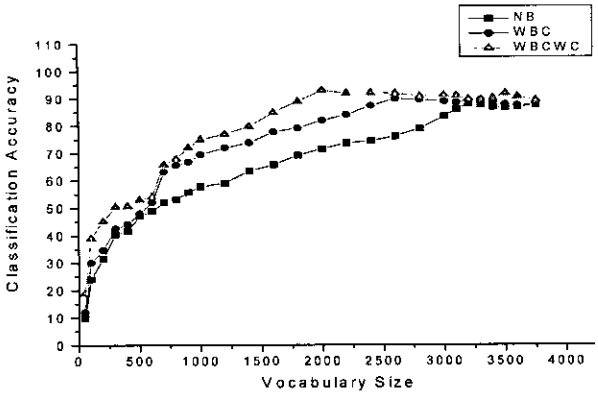


(그림 10) 분류 정확도 곡선(학습 집합)

(그림 11)은 실험 집합에 대한 분류 효율을 보여준다. 이 그래프에서도 위의 학습 집합에 대한 결과와 마찬가지로 본 논문에서 제안한 방법이 기존의 것보다 적은 사전을 가지고도 높은 분류효율을 나타내고 있다.

<표 3>은 분류 정확도가 가장 높았을 때의 사전 크기와 분류 성능 값에 대해 보여주고 있다.

본 논문에서 제안한 시스템을 사용하는 경우 사전의 크기를 고려하지 않고서도 기존의 naive Bayes 분류자를 사용했을 때보다 분류 정확도가 학습 집합에서 2.5%, 실험 집합에서 5.1% 향상되었다.



(그림 11) 분류 정확도 곡선(실험 집합)

〈표 3〉 분류 성능 비교표(1)

		NB	WBC	WBCWC
학습집합	사전 크기(개)	3000	2800	2400
	분류 정확도(%)	94.62	95.20	97.12
실험집합	사전 크기(개)	3200	2600	2000
	분류 정확도(%)	88.00	89.80	93.10

<표 4>는 기존 시스템과 본 논문에서 제안한 시스템의 성능 차이가 현저한 사전 크기에서 각각의 분류 성능을 보여준다. <표 4>로부터 학습 집합에서 사전의 크기가 2,200개일 때 본 논문에서 제안된 시스템의 분류 정확도가 기존의 시스템보다 13% 향상됨을 알 수 있다. 또한 실험 집합에서도 사전의 크기가 2,000개일 때 21.5%의 정확도가 향상됨을 확인할 수 있다.

제안된 시스템의 오분류에 대한 내용을 살펴보면, 문서에 자주 등장하여 높은 가중치를 가지는 단어가 해당하는 문서의 내용과 연관성이 적은 단어이거나, 문서의 내용이 둘 이상의 주제에 대해 다루고 있는 경우가 오분류된 문서의 95% 정도를 차지하고 있다. 기존의 naive Bayes를 이용한 분류실험의 경우 오분류된 문서의 90%이상이 문서의 크기가 50어절 미만인 경우였다. 이는 문서의 특징을 나타내는 단어의 수가 부족하기 때문이다. 본 논문에서는 이러한 50어절 미만의 문서에 대해서도 연관 단어 군집을 이용해 해당되는 클래스로의 분류를 시도하였다.

〈표 4〉 분류 성능 비교표(2)

사전크기 (개)	분류 정확도(%)-학습집합			분류 정확도(%)-실험집합		
	NB	WBC	WBCWC	NB	WBC	WBCWC
1000	65	66	68.9	57.8	69.5	75.1
1200	68.3	69.6	70.6	59	72	77
1400	69.5	70.1	75	63.5	73.9	79.8
1600	72	74	78	65.8	77.8	84.9
1800	76.3	77	83	69.2	79.1	88.9
2000	77.5	80	88.9	**71.6**	**82**	**93.1**
2200	**80**	**85**	**93**	73.8	84.1	92
2400	85	89.5	97.12	74.5	87.3	92.1
2600	87	95	96.3	76.2	89.8	91.9

5. 결 론

기존의 naive Bayes를 사용한 분류는 문서에 출현한 모든 단어에 대해서 추정치를 계산하고 이를 바탕으로 분류를 수행하였기 때문에 문서의 특징들을 정확히 반영하기 어렵고, 많은 잡음들의 영향으로 문서를 오분류하게 된다.

본 논문에서는 분류 정확도를 높이기 위해 클래스 변수와 단어들간의 상호정보 계산을 통해 사전을 구성하고, 분류될 각 문서에서 추출된 주제어들에 가중치를 부여하는 가중치가 부여된 베이지안 분류자를 사용하여 문서를 분류하였다. 또한 상대적으로 길이가 짧은 문서를 기존의 방법에 의해 문서를 분류할 경우, 문서를 대표하는 특징의 수가 적어 오분류하는 경우가 많다. 따라서 본 논문에서는 분류될 문서의 특징을 나타내는 주제어들과의 연관 단어 군집을 통하여 이러한 문제를 해결하고자 시도하였고, 실험을 통하여 이러한 방법이 효과적임을 확인하였다.

실험 결과 사전의 크기를 고려하지 않았을 때, 가중치가 부여된 베이지안 분류자와 단어 군집을 사용한 방법이 기존의 방법보다 학습 집합에서 2.5%, 실험 집합에서 5.1% 향상된 분류정확도를 얻을 수 있었다. 또한 사전의 크기를 고려하였을 때는 학습 집합에서 13%, 실험 집합에서 21.5% 향상된 분류 정확도를 얻었다.

향후 연구 과제로는 본 논문에서 분류한 문서 분류 결과를 사용하여 정보검색의 정확도를 향상시키는 방안에 대해 연구되어야 하고, 문장이 가지고 있는 구문 정보를 이용하여 분류 효율을 높이는 방안에 대한 연구가 필요하다. 또한 충분한 학습 데이터를 확보하는 방법에 대한 연구와 베이지안 분류자를 학습하는 방법에 대한 연구가 진행된다면 더욱 정확하게 문서를 해당 클래스에 분류 할 수 있을 것으로 기대한다.

참 고 문 헌

[1] L. Douglas Baker and Andrew Kachites McCallum, "Distributional Clustering of Words for Text Classification," Proceedings of SIGIR'98, pp.96-103, 1998.

[2] 정영미, *정보검색론*, 구미무역 출판부, 1993.

[3] Andrew McCallum and Kamal Nigram, "A Comparison of Event Models for Naive Bayes Text Classification," AAAI-98 Workshop on Learning for Text Categorization, 1998.

[4] David D. Lewis, "Naive (Bayes) at forty : The Independence Assumption in Information Retrieval," In European Conference on Machine Learning, 1998.

[5] Mehran Sahami, "Learning limited dependence Bayesian classifiers," KDD-96 : Proceedings of the Second International Conference on Knowledge Discover and Data Mining, pp.335-338, AAAI Press, 1996.

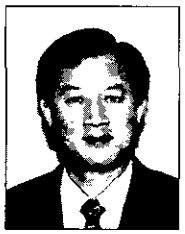
[6] W. Frakes and R. Baeza-Yates, *Information Retrieval*, Prentice Hall, 1992.
[7] 박영자, 사전을 이용한 단어 의미 자동 클러스터링 : 유전자 알고리즘 접근법, 연세대학교 대학원 컴퓨터과학과 박사학위 논문, 1998.
[8] Hang Li and Naoki Abe, "Clustering Words with the MDL Principle," Proceedings of COLING-96, pp.4-9, 1996.
[9] David D. Lewis, Representation and Learning in information retrieval, ph.D.thesis, Dept. of Computer and Information Science, University of Massachusetts, 1992.
[10] 전미선, 박세영, "상호 정보를 이용한 어의 모호성 해소에 관한 연구", 제6회 한글 및 한국어 정보처리학술발표논문집, pp. 369-373, 1994.
[11] G. Salton, "Experiments in Automatic Thesaurus Construction for Information Retrieval," Proceedings of IFIP Congress, pp.43-49, 1971.
[12] A. P. Dempster, N. M. Laird and D. B. Rubin, "Maximum Likelihood from Incomplete data via EM Algorithm," Journal of the Royal Statistical Society, Series B, Vol.39, pp. 1-38, 1977.
[13] Tom M. Mitchell, *Machine Learning*, McGraw-Hill, 1997.

허 준 희
e-mail : jjun2@itnara.net
1998년 인하대학교 전자계산공학과 졸업 (공학사)
2000년 인하대학교 대학원 전자계산공학과 졸업(공학석사)
2000년~현재 아이티나라(주) 기술연구소 연구원
관심분야 : 자연언어처리, 정보검색, 인공지능, 데이터베이스 등

최 준 혁
e-mail : jhchoi@kimpo.ac.kr
1990년 경기대학교 전자계산학과 졸업 (이학사)
1995년 인하대학교 대학원 전자계산공학과 졸업(공학석사)
2000년 인하대학교 대학원 전자계산공학과 졸업(공학박사)
1997년~현재 김포대학 컴퓨터계열(소프트웨어개발전공) 조교수
관심분야 : 정보검색, 데이터마이닝, 신경망, 유전자 알고리즘 등

이 정 현
e-mail : jhlee@inha.ac.kr
1977년 인하대학교 전자공학과 졸업
1980년 인하대학교 대학원 전자공학과 (공학석사)
1988년 인하대학교 대학원 전자공학과 (공학박사)
1979년~1981년 한국전자기술연구소 시스템 연구원
1984년~1989년 경기대학교 전자계산학과 교수
1989년~현재 인하대학교 전자계산공학과 교수
관심분야 : 자연언어처리, HCI, 정보검색, 음성인식, 음성합성, 계산기 구조

김 중 배
e-mail : jjkim@etri.re.kr
1986년 고려대학교 공과대학 산업공학과 졸업(공학사)
1988년 한국과학기술원 산업공학과 졸업 (공학석사)
1998년~현재 한국과학기술원 산업공학과 박사과정
1988년~1991년 대한항공(주) 시스템부
1991년~현재 한국전자통신연구원 컴퓨터소프트웨어기술연구소 모바일응용서버연구팀장
관심분야 : 인터넷 정보검색, 미들웨어, 시스템 소프트웨어 등



임 기 욱
e-mail : rim@omega.sunmoon.ac.kr
1977년 인하대학교 공과대학 전자공학과 졸업
1987년 한양대학교 전자계산학 석사
1994년 인하대학교 전자계산학 박사
1977년~1983년 한국전자기술연구소 선임연구원
1983년~1988년 한국전자통신연구소 시스템소프트웨어 연구실장
1988년~1989년 미 캘리포니아 주립대학(Irvine) 방문연구원
1989년~1997년 한국전자통신연구원 시스템연구부장 주전산기 (타이컴) III, IV 개발 사업책임자
1997년~2000년 정보통신연구진흥원 정보기술전문위원
2000년~현재 선문대학교 산업공학과 교수
관심분야 : 실시간 데이터베이스시스템, 운영체제, 시스템구조 등