준지도 학습 기반의 자동 문서 범주화
(Automatic Text Categorization based on Semi-Supervised Learning)

고 영 중 † 서 정 연 ††
(Youngjoong Ko) (Jungyun Seo)

요 약 자동 문서 범주화란 문서의 내용에 기반하여 미리 정의되어 있는 범주에 문서를 자동으로 할당하는 작업이다. 자동 문서 범주화에 관한 기존의 연구들은 지도 학습 기반으로서, 보통 수작업에 의해 범주가 할당된 대량의 학습 문서를 이용하여 범주화 작업을 학습한다. 그러나, 이러한 방법의 문제점은 대량의 학습 문서를 구축하기가 어렵다는 것이다. 즉, 학습 문서 생성을 위해 문서를 수집하는 것은 쉬우나, 수집된 문서에 범주를 할당하는 것은 매우 어렵고 시간이 많이 소요되는 작업이라는 것이다.

본 논문에서는 이러한 문제점을 해결하기 위해서, 준지도 학습 기반의 자동 문서 범주화 기법을 제안한다. 제안된 기법은 범주가 할당되지 않은 말뭉치와 각 범주의 핵심어만을 사용한다. 각 범주의 핵심어로부터 문맥간의 유사도 측정 기법을 이용한 부스트래핑(bootstrapping) 기법을 통하여 범주가 할당된 학습 문서를 자동으로 생성하고, 이를 이용하여 학습하고 문서 범주화 작업을 수행한다. 제안된 기법은 학습 문서 생성 작업과 대량의 학습 문서 없이 적은 비용으로 문서 범주화를 수행하고자 하는 영역에서 유용하게 사용될 수 있을 것이다.

키워드 : 문서 범주화, 준지도 학습, 부스트래핑 기법

Abstract The goal of text categorization is to classify documents into a certain number of pre-defined categories. The previous studies in this area have used a large number of labeled training documents for supervised learning. One problem is that it is difficult to create the labeled training documents. While it is easy to collect the unlabeled documents, it is not so easy to manually categorize them for creating training documents.

In this paper, we propose a new text categorization method based on semi-supervised learning. The proposed method uses only unlabeled documents and keywords of each category, and it automatically constructs training data from them. Then a text classifier learns with them and classifies text documents. The proposed method shows a similar degree of performance, compared with the traditional supervised learning methods. Therefore, this method can be used in the areas where low-cost text categorization is needed. It can also be used for creating labeled training documents.

Key words : Text Categorization, Semi-Supervised Learning, Bootstrapping Techniques

1. 서 론

인터넷이 폭 넓게 보급되어 온라인(on-line)상에서 얻을 수 있는 텍스트(text) 정보의 양이 급증함에 따라 효율적인 정보 관리 및 검색이 요구 되고 있으며, 이를 위한 기법으로 자동 문서 범주화(automatic text categorization)가 중요하게 사용되고 있다. 자동 문서 범주화는 미리 정의된 범주(category)에 문서를 자동으로 할당하는 기법과 관련된 연구분야로서, 대량의 문서의 효율적인 관리 및 검색을 가능하게 하는 동시에 방대한 양의 수작업을 감소시키는 데 그 목적이 있다.

자동 문서 범주화에 관한 기존의 연구에서는 보통 수

· 이 논문은 2008 학년도 동아대학교 학술연구비 지원에 의하여 연구되었음

† 종신회원 : 동아대학교 컴퓨터공학과 교수
yjko@dau.ac.kr
†† 종신회원 : 서강대학교 컴퓨터학과/바이오 융합기술 협동과정 교수
seojy@sogang.ac.kr

논문접수 : 2007년 7월 13일
심사완료 : 2008년 3월 31일

정보과학회논문지: 소프트웨어 및 응용 제35권 제5호(2008.5)

작업에 의해 범주가 할당된 대량의 학습 문서를 사용해서 학습하고, 범주화 작업을 수행한다. 이러한 지도 학습 기반의 문서 분류기에 사용되는 기계 학습 알고리즘(algorithm)으로는 베이지안 확률 모델(Bayesian probabilistic approach)[1,2], 결정트리(decision tree)[3], 지지 벡터 기계(SVM : Supprot Vector Machine)[4,5], 최근린법(k-nearest neighbors classifier)[6], 선형 모델(linear model)[7], 신경망(neural networks)[8] 등이 있다.

그러나, 이러한 기존의 지도 학습 기반의 문서 범주화 과정은 대량의 학습 문서가 필요하다는 어려움이 있다. 특히, 자동 문서 범주화의 영역이 신문 기사(news article), 전자 도서관(digital library), 전자 우편(email), 뉴스 그룹(news group) 등 적용영역이 넓어지고 다양해짐에 따라, 각 영역에 맞추어 대량의 학습 문서를 생성한다는 것은 많은 작업 인원과 많은 시간을 필요로 하는 어려운 작업이다.

본 논문에서는 학습 문서를 생성하기 위한 작업 없이 각 범주의 핵심어(keyword)의 입력만으로, 범주가 할당되지 않은 학습 문서를 사용하는 준지도 학습(semi-supervised learning) 기반으로 한 새로운 문서 범주화 기법을 제안한다. 제안된 기법은 핵심어로부터 범주가 할당된 학습문서를 부스트래핑(bootstrapping) 기법을 사용하여 자동으로 생성한다. 이러한 부스트래핑 기법의 출발점이 핵심어이기 때문에, 의미의 단위를 문서에서 문맥(context) 단위로 낮추어 고려한다. 먼저, 수집된 문서를 문맥 단위로 나눈 후에 사용자에 의해 입력된 각 범주의 핵심어와 문맥 간 유사도 측정(similarity measure) 기법을 사용하여 각 문맥들의 범주화를 수행한다. 여기서 범주별로 모아진 문맥들을 학습 데이타로 이용하여 1차적으로 문서 분류기를 학습할 수 있고, 학습된 문서 분류기를 이용하여 문서들을 분류함으로써 문서들에 자동으로 범주를 할당한다. 이 과정을 통하여 핵심어로부터 지도 학습을 위한 범주가 할당된 문서들을 학습 문서로서 최종적으로 생성하게 되고, 이들을 이용하여 지도 학습 방식으로 문서 분류기를 학습하여 최종적인 문서 분류기를 획득하게 된다. 제안된 기법은 수집된 문서로부터 대량의 학습 문서를 생성하기 위한 좋은 기초 자료를 제공할 수 있으며, 대량의 학습 문서 없이 적은 비용으로 문서 범주화를 수행하고자 하는 응용 영역에 유용하게 사용될 수 있을 것이다. 본 논문에서 사용한 자질 추출의 방법은 카이 제곱 통계량[9]을 사용하였으며 문서 분류기로는 단순 베이지안 문서 분류기(Naive Bayes text classifier)를 사용하였다.

본 논문의 구성은 다음과 같다. 먼저 2장에서는 관련 연구들을 소개하고, 3장에서 본 논문에서 제안하는 자동 문서 분류 시스템 모델에 대하여 자세히 설명하고, 4장에서 실험 및 평가를 하며, 5장에서는 결론 및 향후 연구를 기술한다.

2. 관련 연구

대량의 학습 문서를 필요로 하는 지도 학습 기반의 문서 분류기의 문제점을 해결하기 위해서 Nigam은 범주가 할당되어진 소량의 학습 문서(labelled documents)와 범주가 할당되지 않은 대량의 문서 집합(unlabelled documents)을 이용해서 문서 범주화를 수행하는 새로운 기법을 제안하였다[10]. Nigam은 [10]에서 베이지안 확률 모델과 EM(Expectation Maximization) 알고리즘을 사용하여 범주가 할당된 300개의 학습 문서와 범주가 할당되지 않은 10,000개의 문서를 사용하여 학습한 결과 1,000개의 범주가 할당된 학습 문서를 사용하여 학습한 결과와 거의 동일한 성능을 나타냄을 보였다.

또한, Languillon은 범주가 할당되어진 학습문서와 할당되지 않은 학습문서를 결합해서 문서 범주화를 수행하는 또 다른 방법을 제시하였는데, 소량의 범주가 할당된 학습문서를 이용하여 군집화 알고리즘의 성능을 향상시킴으로써 문서범주화를 수행하고자 하였다[11].

McCallum은 [12]에서 범주화 대상 영역인 컴퓨터분야의 개발자들로부터의 영역지식이 반영된 범주체계 정보와 EM 알고리즘을 사용해서 문서 범주화 수행하는 새로운 방식의 기법을 제안하고, 그 결과를 컴퓨터 분야의 검색엔진을 개발하는데 사용하였다. 그러나, 이들 연구에서도 일정량의 범주가 할당된 학습 문서나 문서 범주화 대상 영역에 대한 전문가로부터의 정확한 사전 지식이 필요하다는 문제가 여전히 존재한다.

한국어 문서에 대한 비지도 학습 기반의 문서 범주화의 연구는 고영중에 의해 수행되었으며 문장단위의 부스트래핑 기법을 사용하여 한국어 문서에 대한 성능을 평가하였다[13]. 본 연구에서는 [13]의 연구에서 발생한 핵심어 추출의 어려움과 문장 단위의 부스트래핑의 한계를 극복하고자 했다.

본 논문에서 사용한 것과 같이 범주가 할당된 학습 문서를 사용하지 않고 의미의 개념을 확장시키는 부스트래핑 기법은 어휘 의미 중의성 해소(word sense disambiguation) 기법 등에서 사용되었다. Yarowsky는 적은 수의 종자 단어(seed word)와 의미 집합을 사용하여 부스트래핑을 함으로써 단어 중의성을 해소하는 알고리즘을 제시하였다[14].

3. 준지도 학습을 기반으로 한 문서 범주화 시스템

본 논문에서 제안하는 시스템은 그림 1과 같이 크게 전처리 과정, 학습문맥 집합 생성 과정, 그리고, 문서분류기 학습 및, 학습된 문서분류기를 통해 분류된 범주가

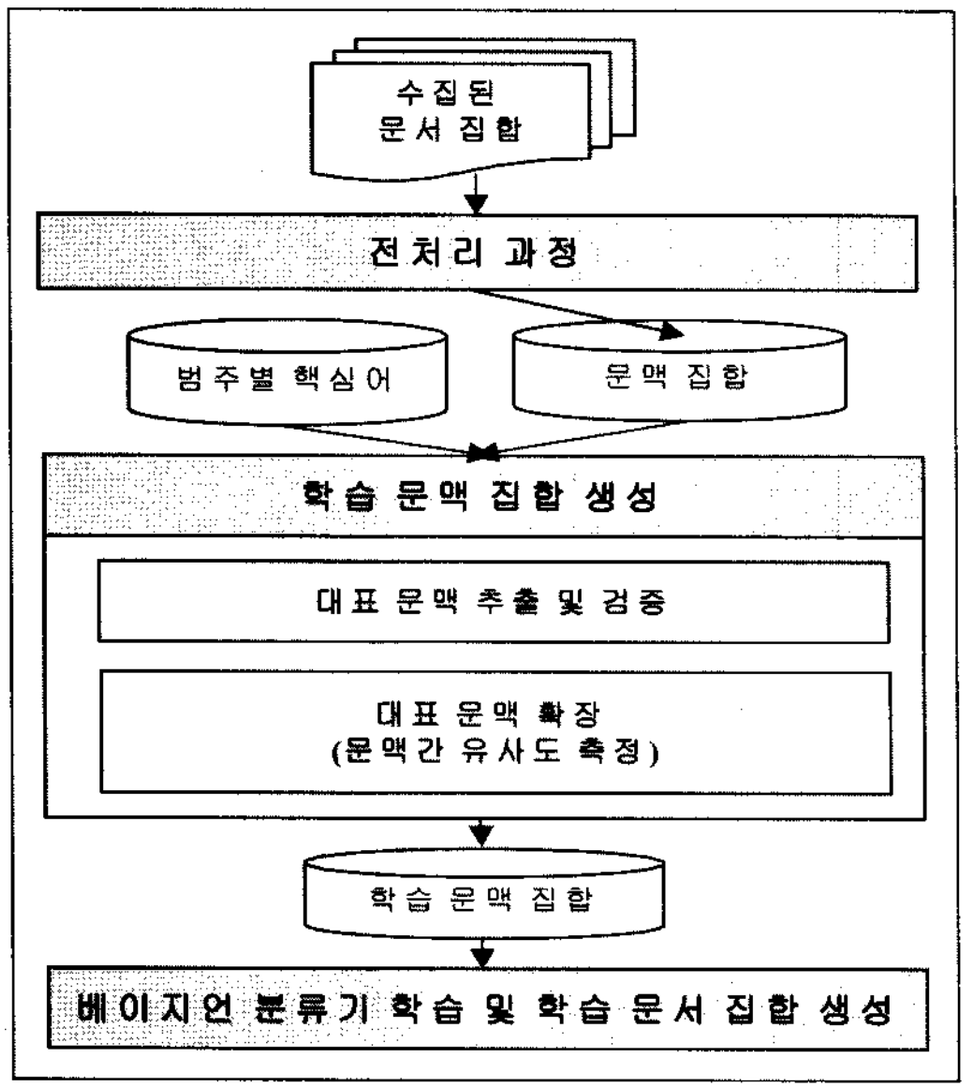

그림 1 전체 시스템 구성도

할당된 학습 문서 집합을 생성하는 과정으로 나누어진다. 제안된 방법은 각 범주별 학습 문맥 집합을 생성하고, 이를 통해 학습하여 문서 범주화를 위한 범주가 할당된 학습 문서를 자동으로 생성한다.

3.1 전처리 과정

전처리 과정은 크게 내용어(content word)를 추출하는 단계와, 추출된 내용어를 사용하여 문서의 내용을 문맥 단위로 나누는 단계로 나눌 수 있다.

문맥의 내용이나 특징을 잘 반영하는 단어를 내용어라고 한다. 이러한 내용어를 추출하기 위해서는 먼저 형태소 분석기를 사용하여, 각 형태소 별로 나누고 품사를 결정한다. 본 논문에서는 결정되어진 품사 중에 명사와 동사만을 추출하여 내용어로 사용하였다. 또한 불용어를 제거하기 위해 불용어 사전(stopword list)을 정의하고 내용어 추출 시 불용어에 해당하는 용어들을 제거하였다.

문맥이란 특정 단어나 혹은, 구문을 둘러싸고 있는 문서의 일부분으로써, 어떤 단어나 구문의 문맥은 그들의 의미를 결정해주는데 유용하게 사용될 수 있다. 본 논문에서는 각 범주의 핵심어로부터 부스트래핑 작업이 시작되기 때문에, 기존의 문서 범주화의 의미 단위인 문서 단위 보다는 문맥단위를 부스트래핑의 기본 단위로 사용하고 있다. 본 논문에서는 인접한 60개의 내용어를 슬라이딩 윈도우(sliding window) 기법[15]으로 추출하고 이를 하나의 문맥으로 사용한다.

3.2 핵심어 추출

학습 문서 생성을 위한 부스트래핑 작업은 각 범주의 핵심어로부터 시작되기 때문에 핵심어의 선택은 본 시스템의 성능에 매우 중요한 요소이다. 물론, 핵심어를 사람이 직접 생성하여 사용하는 것이 가장 좋은 방법이겠지만, 각 범주의 핵심어를 추출한다는 것은 쉽지 않은 작업이다. 특히, 사용자가 잘 알지 못하는 영역에서 핵심어를 추출한다는 것은 매우 어려운 일임에 틀림없다. 핵심어를 손쉽게 추출하기 위하여 각 범주의 주제어(subject word)와 공기 정보(co-occurrence information)를 사용하여 후보 핵심어를 자동으로 추출하여 제시하고, 사용자는 이들 중에 유용한 단어들만을 선택하여 핵심어로서 사용한다. 서론에서의 정의에 따르면 문서 범주화는 미리 정의된 범주 정보를 바탕으로 하는 작업이기 때문에, 각 범주의 주제어를 선택하는 것은 그리 어렵지 않은 일이다. 그러므로, 각 범주의 주제어와의 공기 정보를 다음과 같은 코사인(cosine) 유사도 계산식을 사용하여 주제어와 의미상으로 가장 근접한 단어들을 추출할 수 있다.

$$sim(T,W)=\frac{\sum_{i=1}^{n} t_i \times w_i}{\sqrt{\sum_{i=1}^{n} t_i^2 \times \sum_{i=1}^{n} w_i^2}} \quad (1)$$

식 (1)에서 T는 주제어를, W는 유사도를 계산하기 위한 단어를 나타내며 t_i와 w_i는 i번째 문서에서 두 단어의 출현을 이진값(0 혹은 1)으로 나타낸 것이다. 즉, 각 단어는 학습 문서 집합에서 주제어와 같이 출현한 문서의 수가 많을수록 그 주제어와 의미상으로 근접하다. 따라서, 가장 높은 코사인 유사도를 갖는 범주로 각 단어를 할당한다. 하지만, 핵심어는 주제어와의 유사도 뿐만 아니라 범주간의 의미를 잘 분별할 수 있어야 한다. 그러므로, 다음의 식으로 각 범주에 할당된 단어에 중요도 값을 부여하고 정렬한다.

$$Score(W,c_{\max}) = sim(T_{\max},W) + (sim(T_{\max},W) - sim(T_{secondmax},W)) \quad (2)$$

식 (2)에서 T_{max}는 가장 높은 유사도 값을 갖는 주제어를, $T_{secondmax}$는 두 번째로 높은 유사도 값을 갖는 주제어를 의미한다. 이 식에 의해서 한 주제어와 높은 유사도를 갖고 그 외의 주제어와는 낮은 유사도를 갖는 단어가 높은 순위를 갖게 된다. 이러한 과정을 통해 10개의 후보 핵심어를 사용자에게 제시하고 사용자는 이 중에서 유용한 핵심어만을 추출하여, 각 범주의 핵심어를 추출하게 된다. 다음 표 1은 *Reuters* 문서 집합에서의 범주별 핵심어이다. 본 논문에서는 공정한 평가를 위하여 각 범주의 주제어를 선정할 때 각 범주의 제목을 그대로 이용하였다. 다만, *Reuters* 문서 집합의 경우에 몇 개의 범주에 줄임말로 표현된 것이 존재하기 때문에 Reuters 문서 집합의 *ReadMe* 파일에 기술되어 있는

표 1 Reuters 문서 집합의 범주별 핵심어

범주	주제어	핵심어
acq	acquisition, merger	inc, shares, shareholders
corn	corn	bushel, soybeans, bushels, soybean
crude	crude oil	bpd, barrels, barrel, petroleum, opec, gasoline
earn	earnings	quarter, revenues, income, profits
grain	grain	usda, harvest, maize,buenos, aires, argentine
interest	interest rate	lending, bonds, inflation, banks, bundesbank
money-fx	foreign exchange	currency, intervention, dollar, dollars, auction, yen
ship	shipping	missiles, chinese-made, silkworm, tehran, missile, iran, ships, vessel
trade	trade	tariffs, surplus
wheat	wheat	tones, winter, barley, flour

범주 설명을 참조하여 주제어를 선정하였다.

3.3 학습 문맥 집합 생성

본 절에서는 선택된 핵심어와 문맥 간 유사도 측정 기법을 사용하여 단순 베이지안 문서 분류기를 학습하기 위한 학습 문맥 집합을 생성하는 부스트래핑 기법을 기술한다. 먼저, 핵심어를 사용하여 각 범주의 핵심어를 직접 포함하고 있는 문맥을 그 범주의 특성을 가장 잘 내포하고 있는 문맥으로 고려하고, 전처리 단계에서 생성되어진 수집된 문맥의 내용어 중에 미리 정의된 핵심어를 직접 포함하고 있는 문맥을 각 범주의 대표 문맥으로 추출한다.

추출된 대표 문맥들만으로 각 범주의 학습문맥 집합이 되기에는 양이 부족하므로, 대표 문맥으로 추출되지 못한 미 분류 문맥들을 추출된 대표 문맥과의 유사도 측정을 통해 각 범주에 할당시키고, 대표 문맥 집합과 할당된 범주별 문맥 집합을 합하여 최종적인 범주별 학습 문맥 집합을 생성한다.

3.3.1 범주별 대표 문맥 추출

본 논문에서 정의하는 각 범주별 대표 문맥이란 미리 정해진 범주의 핵심어를 직접 내용어로 가지고 있는 문맥이다. 이들을 추출하기 위해서 전처리 과정에서 추출되어진 각 문맥의 내용어 중에 각 범주의 핵심어를 직접 포함하고 있는 문맥을 추출하는데, 이때 두 가지 이상의 범주에 해당하는 내용어를 가진 문맥은 대표 문맥의 중의성을 해결하기 위하여 제외시킨다. 본 논문에서는 이렇게 추출되어진 대표 문맥들을 각 범주의 특성을 가장 잘 나타내는 문맥으로 고려한다. 그러나, 실제로는 어떤 범주의 핵심어를 포함하고 있는 문맥이라 할지라도 그 범주의 특성을 잘 나타내지 못하는 문맥들이 있다. 추출된 문맥들을 각 범주의 특성을 잘 나타내는 순서로 순위화 하기 위해 다음과 같이 문맥의 가중치(weight)를 계산하고 순위화 한다.

단계 1. 추출된 대표 문맥들의 단어에 가중치를 부여하기 위해서 다음과 같은 용어빈도(TF : Term Frequency)와 역범주 빈도(ICF : Inverse Category Frequency)를 사용한다[16].

① 각 범주 대표 문맥 집합에서의 용어 t_i의 용어 빈도(TF)를 다음 식으로 계산한다.

$$TF_{ij} = j\text{번째 범주에서의 용어 } t_i\text{의 출현빈도} \quad (3)$$

② 기존의 정보 검색 분야에서는 일반적으로 역문헌 빈도(IDF : Inverse Document Frequency)를 사용하나 제안된 기법에서는 문맥 단위로 프로세스(process)가 진행됨에 따라 문서 출현 빈도를 계산할 수 없으며, [16]에서 범주 간의 분리도가 높은 단어에 높은 가중치를 주는 역범주 빈도를 정의하고 효율성을 입증하였으므로, 제안된 기법에서는 역범주 빈도를 사용한다. t_i를 포함하는 범주의 개수는 CF_i이고 총 범주의 개수를 M이라고 할 때 역범주 빈도는 다음 식과 같다.

$$ICF_i = \log(M) - \log(CF_i) \quad (4)$$

③ 위에서 계산된 용어 빈도(TF_{ij})와 역 범주 빈도(ICF_i)를 이용해서 용어 t_i의 j번째 범주에서의 가중치 w_{ij}는 다음과 같이 계산된다.

$$w_{ij} = TF_{ij} \times ICF_i \quad (5)$$

단계 2. 단계 1에서 계산된 용어 가중치를 이용해서 다음과 같이 j번째 범주에서의 문맥 ($X=\{w_{1j}, w_{2j}, \cdots, w_{Nj}\}$)에 가중치 W를 부여한다. (N은 문맥의 용어의 총수)

$$W = \frac{w_{1j} + w_{2j} + \ldots + w_{Nj}}{N} \quad (6)$$

단계 3. 단계 2에서 계산된 문맥 가중치를 해당 범주에서의 문맥의 대표성, 즉 얼마나 해당 범주의 특성을 잘 내포하는지의 척도로 보고 범주에 속해 있는 문맥을 가중치가 큰 순서로 순위화하여 문맥 유사도 계산에 사용될 대표 문맥들을 결정한다.

3.3.2 대표 문맥 집합의 확장

추출된 대표 문맥 집합은 각 범주의 특성을 잘 나타내는 문맥들로 구성되어 있지만 문서 범주화의 학습 데이타로 사용하기 위해서는 그 양이 아직 적기 때문에 이를 보충하기 위해서 대표 문맥으로 추출되지 못한 문

맥들을 각 범주의 대표 문맥들과의 유사도 측정을 통해 측정된 유사도가 가장 높은 범주에 할당시킴으로써 학습 문맥 집합을 확장한다.

(1) 문맥 간 유사도 계산

유사한 단어는 유사한 문맥에 위치하는 경향이 있으므로 이를 이용해 문맥 정보를 반영하여 문맥 간 유사도를 측정한다[17,18]. 본 논문에서는 [17]에서 어휘 의미 중의성 해소 분야(Word Sense Disambiguation)에 적용되어 좋은 성능을 보인 문맥 간 유사도 측정 기법을 개량하여 사용한다. 사용된 문맥 간 유사도 측정 기법에서는 단어와 문맥은 상호 보충적인 역할을 수행한다. 문맥은 유사한 단어들을 많이 포함할수록 유사한 문맥이고 단어는 유사한 문맥에서 많이 사용될수록 유사한 단어이다. 이 정의는 순환적이며 이를 반영하기 위해 그림 2와 같이 두 개의 행렬(matrix)를 이용하여 반복 계산한다.

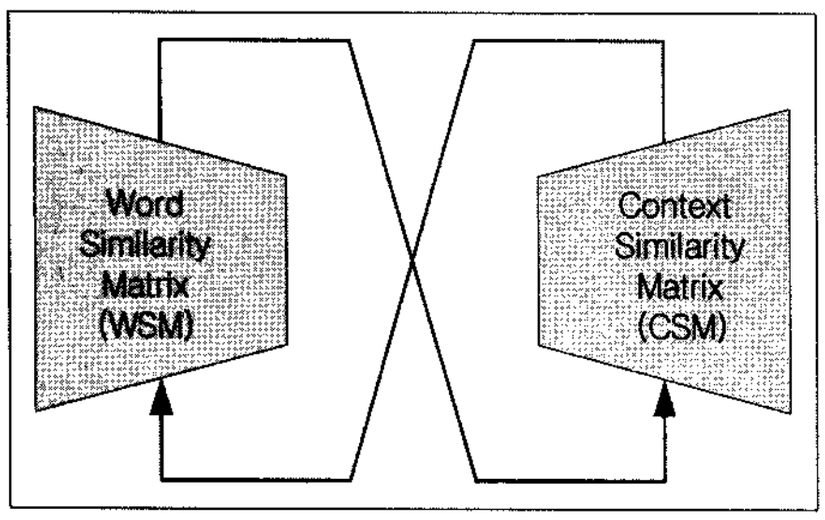

그림 2 문맥 유사도 측정의 반복 계산

그림 2의 WSM(word similarity matrix)의 행(row)과 열(column)은 범주별로 대표 문맥과 미 분류 문맥들에 포함되어 있는 모든 내용어들로 구성되며, 행렬의 각 요소(cell)는 단어 사이의 문맥적 유사도를 나타내는 0에서 1사이의 값을 가진다. *CSM*(context similarity matrix)은 행에 미 분류 문맥들이 위치하고 열에는 각 범주의 대표 문맥들을 위치함으로써 이들 문맥간의 유사도 값을 나타내게 된다. 본 논문에서는 각 범주의 대표 문맥과의 유사도 측정을 위한 입력 문맥(미 분류 문맥)의 수를 수행 속도, 메모리 할당 등을 고려하여 200개로 제한하며, 대표 문맥의 수도 3.3.1절에서 계산된 대표 문맥의 순위별로 상위 200개로 제한하여 각 범주마다 *WSM*과 *CSM*을 생성한다. 이들 단어간, 문맥간 유사도를 측정하기 위해서는 먼저 *WSM*을 단위 행렬(identity matrix)로 초기화한다. 즉, 각 단어는 같은 단어의 유사도 값은 1로 하고 다른 단어와의 유사도 값은 0으로 한다. 그리고, 유사도 값의 변화가 충분히 작아질 때까지 다음과 같은 과정 1), 2)를 반복 수행한다.

1) WSM_n을 사용해서 CSM_n을 갱신(update)한다.

2) CSM_n을 사용해서 WSM_n을 갱신한다.

(2) 친밀도 계산식

단어와 문맥의 유사도 측정 계산을 단순화하기 위해서 단어와 문맥 사이의 관계를 정의하는데 이를 친밀도(affinity)라 한다. 단어(W)는 모든 문맥과 친밀도를 가지는데 이는 단어(W)와 문맥을 구성하고 있는 단어들과의 문맥적 관계를 나타낸다. 비슷한 방법으로 문맥(X)도 모든 단어와 친밀도를 가지는데 이는 문맥(X)와 단어(W)를 포함하고 있는 문맥들과의 유사도를 반영한다.

친밀도 계산식은 식 (7)과 같다[17]. 여기서 단어(W)가 문맥(X)에 속해 있을 때 $W \in X$로 표시한다.

$$aff_n(W,X) = \max_{W_i \in X} sim_n(W, W_i)$$
$$aff_n(X,W) = \max_{W \in X_j} sim_n(X, X_j) \quad (7)$$

n은 반복 횟수를 나타내며 유사도 값은 WSM_n과 CSM_n에 의해 정의된다. 이 식에 의해 모든 단어는 모든 문맥과 친밀도에 의해 표현될 수 있고, 문맥은 포함하고 있는 단어들의 친밀도를 나타내는 벡터(vector)들로 표현된다.

(3) 유사도 계산식

단어 W_1과 W_2의 유사도는 W_1과 W_2를 포함하고 있는 문맥들의 평균 친밀도에 의해 정의되고, 문맥 X_1와 X_2의 유사도는 X_1와 X_2를 구성하는 단어들의 가중치 평균 친밀도로 정의된다. 유사도 계산식은 다음과 같다.

$$sim_{n+1}(X_1, X_2) = \sum_{W \in X_1} weight(W, X_1) \cdot aff_n(W, X_2) \quad (8)$$

$$\text{if } W_1 = W_2$$

$$sim_{n+1}(W_1, W_2) = 1$$

$$else \quad (9)$$

$$sim_{n+1}(W_1, W_2) = \sum_{W_1 \in X} weight(X, W_1) \cdot aff_n(X, W_2)$$

여기서 식 (8)의 가중치는 각 단어의 출현빈도, 품사 등을 고려하여 계산되었으며, 식 (9)의 가중치는 합이 1이 되도록 단어 W_1를 포함하고 있는 문맥 수의 역수를 사용하였다. 이 식에서 계산된 유사도 값은 WSM_n과 CSM_n의 각 요소에 대응되는 값이다.

(4) 미 분류 문맥의 범주 할당

한 번에 입력되는 200개의 미 분류 문맥들은 각 범주별로 *WSM*과 *CSM*을 생성하고 문맥 간 유사도 측정에 의해 미 분류 문맥별로 각 범주의 대표 문맥들과의 유사도 값을 계산한다. 그리고, 다음과 같은 식으로 표현되는 평균값으로 각 미 분류 문맥의 범주별 유사도 값을 할당하고, 그 값이 가장 큰 범주로 미 분류 문맥을 분류한다.

$$\underset{c_i \in C}{sim}(X, c_i) = \frac{1}{n} \sum_{j=0}^{n} \underset{S_j \in R_c}{sim}(X, S_j) \quad (10)$$

여기서 X는 미 분류 문맥이고 $C=\{c_1,c_2,\cdots c_m\}$는 범주 집합(category set)이며 $R_{c_i}=\{S_1,S_2,\cdots,S_n\}$는 범주 c_i의 대표 문맥 집합이다. 즉, 총 범주의 수가 10개라면 각각의 200개의 문맥이 10개의 범주별 유사도 값을 가지게 된다. 문맥 중에는 어느 범주의 특징도 잘 포함하고 있지 않는, 즉, 어느 범주에도 속하지 않는 문맥들도 많은데 이를 제거하여 학습에 참여하지 않게 하기 위해서 다음과 같이 임계값을 설정하고, 식 (10)에 의해 계산되는 미 분류 문맥의 범주별 유사도 값이 임계값 이상이 될 경우에만 해당 문맥에 범주를 할당한다.

• 임계값(Threshold Value) 설정

총 범주의 수가 10개라고 했을 때, 입력 미 분류 문맥 200개가 10개의 각 범주별 유사도 값을 가지므로, 총 2,000개의 범주별 유사도 값이 존재한다. 이들의 분포를 정규분포(normal distribution)로 보고, 각 미 분류 문맥의 범주별 유사도 값이 일정 임계값(상위 %)안에 해당하는 문맥에만 범주를 할당한다. 여기서 임계값의 수치는 실험값으로서 결정되며 그 결과는 또한 4.3.1절에서 실험하고 평가하였다. 본 논문에서 사용된 임계값은 범주별 유사도의 값이 정규 분포를 따르므로 다음과 같은 식으로 계산된다.

$$\max\left\{\underset{c_i\in C}{sim}(X,c_i)\right\}\geq \mu+\theta\sigma \tag{11}$$

여기서, X는 미 분류 문맥이고, μ는 2,000개의 유사도 값의 평균값이며, Θ는 표준 정규 분포표에서 임계값(상위 %)에 해당하는 값에 대응되는 수치이다. 또한, σ는 표준편차 값이다.

3.4 자질 추출 및 문서 범주화

3.4.1 자질 추출

Yang은 [9]에서 여러 가지의 자질 추출 방법을 사용하여 실험을 한 결과 카이 제곱통계량과 정보 획득량을 사용하는 것이 가장 효과적임을 보이고 있다. 본 논문에서는 이를 바탕으로 비교적 구현이 쉽고 고빈도 단어에 친화적인 카이제곱 통계량을 사용하여 자질을 추출한다. 카이제곱 통계량에서 용어(t)와 범주(c)와의 공기 문맥 혹은 문서 빈도는 표 2와 같이 나타낸다. 기존의 카이 제곱 통계량 기법에서는 공기 문서 빈도를 사용하나, 본 논문에서 제안되는 기법에서 학습 문맥을 사용하는 1차 문서 분류기 학습에서는 문서 단위가 존재하지 않으므로 공기 문맥 빈도를 사용한다.

$N(=A+B+C+D)$을 학습 데이타에서의 총 문맥 혹은 문서수라고 할 때 용어(t)의 해당범주(c)에서의 카이 제곱 통계량 값은 다음 식과 같다[9,19].

$$\chi^2(t,c)=\frac{N\times(AD-CB)^2}{(A+C)\times(B+D)\times(A+B)\times(C+D)} \tag{12}$$

위와 같은 식으로 계산되어진 각 범주(c)에서의 용어(t)의 자질 값을 전체 학습 데이타에서의 자질 값으로 변환하기 위하여 다음 식을 사용한다.

$$\chi^2(t)=\max_{i=1}^{m}\left\{\chi^2(t,c_i)\right\} \tag{13}$$

즉, 용어(t)의 전체 학습 데이타에서의 자질값은 각 범주(c_i)에서의 자질값의 최대값으로 할당한다. 식 (13)에 의해서 계산되어진 자질 값을 순위화하여 자질 값이 높은 순서로 자질을 추출한다.

3.4.2 문서 범주화

본 논문에서 사용한 문서 분류기는 베이지안 확률 모델을 사용한다[2,10]. 베이지안 확률 모델은 주어진 입력 문서의 각 범주에 할당될 확률을 구하기 위해서 문장에 속해 있는 용어들과 범주와의 결합 확률값(joint probability)을 사용하는 방법이다. 즉, 문서 d는 $Pr(c|d)$가 최대가 되는 범주 c에 할당되게 된다. 이는 다음과 같은 식으로 나타낸다.

$$\Pr(c|d)=\Pr(c)\times\frac{\Pr(d|c)}{\Pr(d)} \tag{14}$$

여기서 d는 문서를 나타내고, c는 범주를 나타낸다. 이 식은 구하고자 하는 확률($\Pr(c|d)$)을 베이지안 규칙(Bayes' rule)을 사용하여 바꾼 것이다. 이 식을 계산 가능하게 하기 위하여 문서 d를 문서 안에 포함되어 있는 용어(t_i)의 집합($d=\{t_1,\ldots,t_i\cdots\}$)으로 표현하고 용어간에 독립 발생을 가정하면 식 (15)를 유도할 수 있다.

$$\Pr(c|d)=\Pr(c)\times\frac{\prod_{i=1}^{T}\Pr(t_i|c)^{N(t_i,d)}}{\Pr(d)} \tag{15}$$

식 (15)에서 계산되는 확률 값을 문서가 각 범주에 할당될 확률로 보고 가장 확률이 높은 범주로 문서를 할당한다. 식 (15)에서 $N(t_i,d)$는 문서 d에서의 용어 t_i가 출현하는 빈도(TF)를 의미하고, T는 전체 문서 집합 내의 용어의 수를 나타낸다.

식 (15)을 통해 계산되는 확률 계산값이 용어의 수가 많아져서 컴퓨터의 부동소수점 연산의 불확실성에 의해 0이나 혹은 1에 치우친 극단 값이 나오는 문제가 있는데, 이를 완화하기 위하여 Kullback-Leiber Divergence를 사용하여 식 (16)과 같이 변환하여 사용한다[20,21].

$$\begin{aligned}&\Pr(c)\prod_{i=1}^{T}\Pr(t_i|c)^{N(t_i,d)}\\&\propto\frac{\log\Pr(c)}{n}+\sum_{i=1}^{T}\Pr(t_i|d)\log\left(\frac{\Pr(t_i|c)}{\Pr(t_i|d)}\right)\end{aligned} \tag{16}$$

표 2 용어와 범주간의 공기 문맥/문서 빈도수

범주 \ 용어	C 할당됨	C 할당안됨
t 있음	A	B
t 없음	C	D

n은 문서 d에 출현하는 모든 용어의 수이고 $Pr(c)$는 전체 학습 집합에서의 해당 범주가 나타날 확률을 의미하며, $Pr(t_i|c)$는 해당 범주에서 용어 t_i가 나타날 확률을, 그리고, $Pr(t_i|d)$는 대상 문서에서 용어 t_i가 나타날 확률을 의미한다. 각각의 확률식은 다음과 같이 계산된다[20].

$$\Pr(t_i|c) = \frac{N(t_i,c)+1}{\sum_{j=1}^{T_c} N(t_i,c)+1\times T_c} \tag{17}$$

$$\Pr(t_i|d) = \begin{cases} \dfrac{N(t_i,d)+1}{\sum_{j=1}^{T_c} N(t_i,d)+1\times T_d} & \text{if } N(t_i,d)\neq 0 \\ 0 & \text{if } N(t_i,d)=0 \end{cases}$$

여기서 $N(t_i,c)$는 범주 c에서의 용어 t_i가 출현한 빈도수이며 T_c는 범주 c의 용어의 총수이다. 식 (17)은 Laplace smoothing이라 불리우는 식으로서 스무딩(smoothing)기법으로 사용되어진다. 학습 문맥 집합을 이용한 학습에서는 각 범주의 학습 문맥 집합으로부터 빈도수가 계산되며, 제안된 기법에 의해 자동으로 생성된 학습문서를 이용한 학습에서는 각 범주의 문서 집합에서 빈도수가 계산된다.

4. 실험 및 결과

4.1 실험 데이타

실험에서 사용한 테스트 문서 집합은 문서 범주화 영역에서 주로 사용되는 대표적인 두가지를 사용한다. 첫 번째 문서 집합은 뉴스 그룹(UseNet discussion group)의 문서들을 모아 놓은 테스트 문서 집합(Newsgroups)[10,22]으로써, 20개의 범주에 총 20,000개의 문서들로 구성되어 있다. 하지만, 본 논문에서는 이들 범주를 모두 다 사용하지 않고 16개의 범주(16,000개 문서)만을 사용한다. 제외된 4개의 범주 중 3개의 범주는 범주의 내용이 기타에 해당하는 범주이고, 다른 하나의 범주는 'hardware'라는 주제어가 중복되어서 제외하였다. 뉴스 그룹 문서 집합은 학습 문서와 테스트 문서의 구분이 없으므로, 공정한 평가를 위해서 five-fold cross validation 기법으로 평가하였다. 즉, 전체의 20%를 테스트 문서로 하고 나머지를 학습문서로 사용하여, 총 다섯 개의 학습 문서와 테스트 문서의 집합을 만들어 각각 실험하고, 실험 결과의 평균값으로 성능을 평가하는 기법이다. 불용어 사전을 사용하였으며, 스테밍(stemming)은 사용하지 않았다.

두 번째 문서 집합은 *Reuters 21578*으로써 총 12,902개의 신문기사와 90개의 범주로 구성되어 있다[21,23]. 본 논문에서는 문서를 가장 많이 보유하고 있는 10개의 범주를 대상으로 실험한다. 학습 문서와 테스트 문서의 구분을 위해서는 'ModApte' 분류 기준을 따랐으며, 불용어 사전은 사용하였으나 스테밍은 사용하지 않았다.

4.2 성능 평가 방법

뉴스 그룹 문서 집합의 평가 방법으로는 정보 검색 분야에서 일반적으로 사용되는 정확율(precision)과 재현율(recall)을 사용하였으며, 정확율과 재현율을 하나의 값으로 표현해주기 위해서 다음 식 (17)과 같이 F_1-*measure*를 사용하였다.

$$F_1(r,p) = \frac{2rp}{r+p} \tag{18}$$

식 (18)에서 r은 재현률에 해당하고 p는 정확률에 해당한다.

하지만, *Reuters* 문서 집합은 한 문서가 여러 개의 범주에 할당 될 수 있기 때문에 각 범주별로 이진 분류기(binary classifier)를 만들어서 각 범주별로 정확율과 재현율이 같아지는 지점에서의 값인 손익 분기점(break-even point)로 평가하였다[23].

모든 범주의 성능을 통합하여 평가하기 위한 기법으로는 문서 범주화 기법의 성능 평가에 주로 사용되는 마이크로평균(micro-averaging)기법을 사용한다[23].

4.3 실험 결과

제안된 방법의 학습 문서는 지도 학습 기반(supervised learning based)의 문서 범주화를 위해 생성된, 범주가 할당된 학습 문서 집합을 범주가 할당되지 않은 것으로 가정하여 사용한다. 이들을 범주 구분 없이 사용하여 학습 문맥 집합을 추출하고, 본 기법을 통해 각 문서에 범주를 할당하고, 그들을 학습하여 범주화를 수행한다.

4.3.1 미 분류 문맥의 범주 할당에서의 임계값에 따른 실험

3.3.2절의 (4)에서 식 (10)의 미 분류 문맥의 범주 할당의 임계값을 상위 10%, 상위 15%, 상위 20%, 상위 25%, 상위 30%로 나누고 각각 실험하여 성능을 비교하였다. 실험은 뉴스 그룹 문서 집합의 학습 문서 집합에서, 다시 20%를 검증집합(validation set)으로 추출하여 평가하였다.

그림 3에서 문맥 할당의 임계값은 상위 15%를 사용하는 것이 좋은 성능을 나타내고 있으므로, 이후의 실험에서는 문맥 할당의 임계값은 상위 15%를 사용하여 실험하였다.

4.3.2 범주가 할당된 학습 문서를 사용한 지도 학습 기반 시스템과의 성능 비교

본 논문에서는 제안된 기법을 정확히 평가하기 위하여 기존의 지도 학습 기반의 문서 범주화 시스템을 구현하고 같은 자질 추출 기법(χ^2 statistics)과 같은 단순

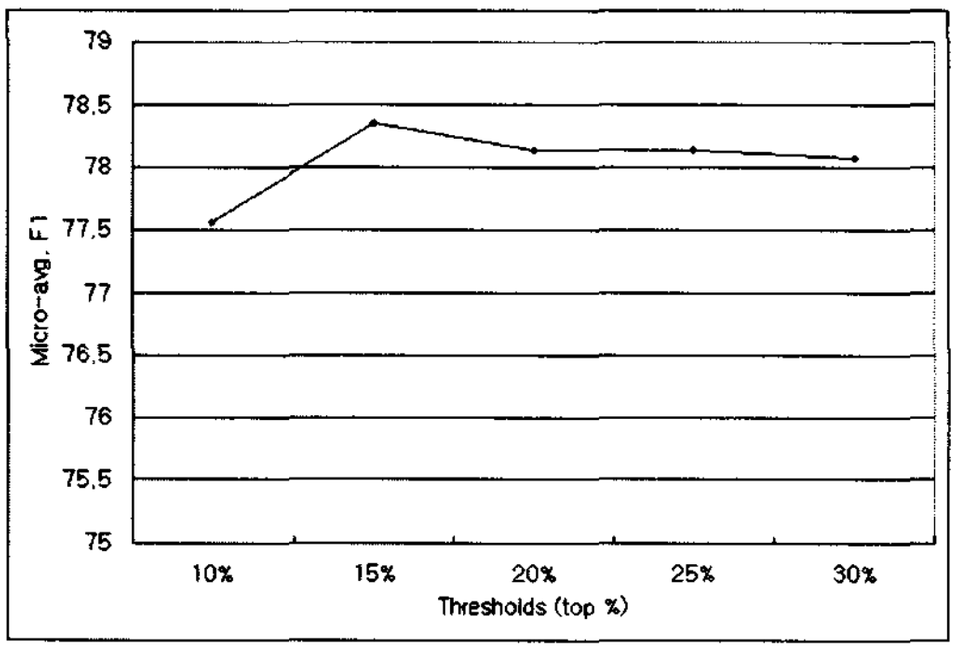

그림 3 문맥 할당 임계값에 따른 성능 비교

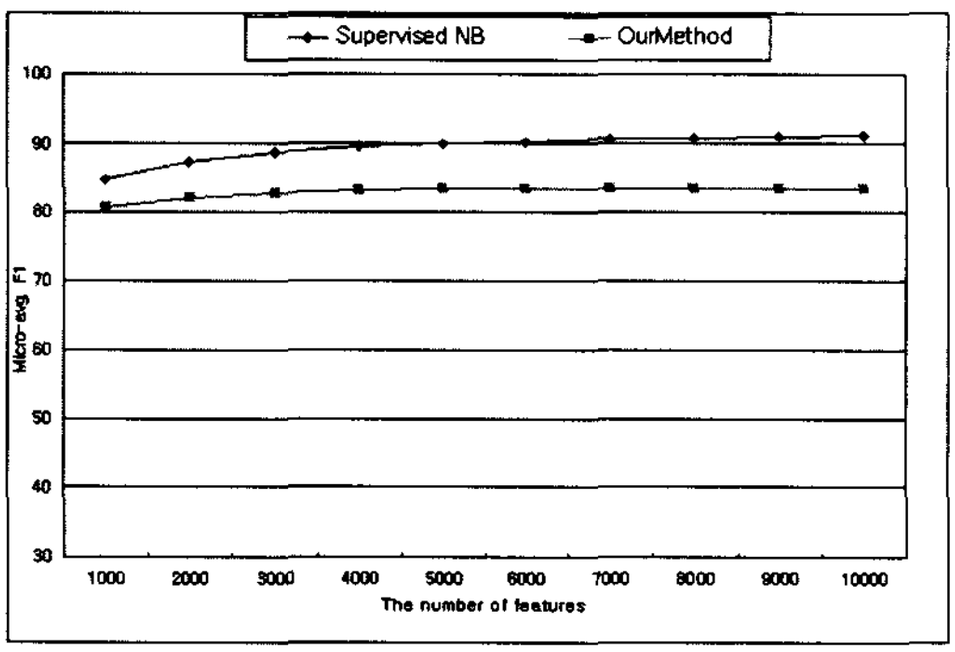

그림 4 뉴스 그룹 문서 집합에서의 지도 학습 기반의 기법과 제안된 기법과의 자질 수에 따른 성능 비교

베이지안 문서 분류기(Naive Bayes Classifier)를 사용하여 실험하고 성능을 비교하였다.

그림 4는 자질 수에 따른 각 기법에서의 단순 베이지안 분류기의 성능 변화를 보이고 있고 표 3은 범주별 성능비교를 보이고 있다. 제안된 기법은 83.46%의 성능을 보이고 있고, 지도 학습 기반의 성능은 91.2%를 보임으로써 7.74%의 성능 차이를 보이고 있다.

다음 표 4는 *Reuters* 문서 집합에서의 각 범주별 손익 분기점과 micro-averaging 손익 분기점을 보이고 있다. 제안된 기법은 88.7%의 성능을 보이는 반면, 지도

표 4 Reuters 문서 집합에서의 성능비교

범주	OurMethod	Supervised NB
acq	93.74	96.24
corn	64.28	66.07
crude	85.71	89.41
earn	96.22	97.42
grain	69.79	92.61
interest	78.62	77.09
money-fx	75.41	78.21
ship	83.14	85.39
trade	78.81	81.35
wheat	64.78	67.6
micro-avg.	88.7	91.64

학습 기반의 성능은 91.64%를 보이고 있다. 따라서, *Reuters* 문서 집합에서는 2.94%라는 근소한 차이만을 보이고 있다.

두 개의 문서 집합에서 제안된 기법을 사용한 문서 분류기는 범주가 할당된 학습 문서 없이, 각 범주의 핵심어들만을 사용해서 80%가 넘는 높은 성능을 보이고 있으며 지도 학습 기반의 분류기와도 적은 성능 차이를 보이고 있다. 특히, *Reuters* 문서 집합에서는 2.94%라는 거의 근접한 성능을 보이고 있다. 이 결과는 본 논문에서 제안된 기법을 사용한다면 대량의 학습 문서를 생성하는 작업 없이, 적은 시간과 적은 인력을 들이고도 충분히 문서 범주화를 수행할 수 있음을 보이고 있다.

4.3.3 제안된 기법의 성능을 얻기 위해 필요한 학습 문서의 양 추정

본 논문에서는 제안된 기법을 사용했을 때 얻을 수 있는 가장 큰 장점은 범주의 할당 작업을 수행하지 않고도 문서 범주화 작업을 할 수 있다는 점이다. 따라서 제안된 기법의 성능을 좀 더 분석하기 위해서, 지도학습 기반의 분류기가 제안된 기법의 성능을 내고자 했을 때 어느 정도의 범주가 할당된 학습 문서가 필요한지를 실험으로 추정하여 보았다. 그림 5에서 보는 바와 같이 학습 문서의 수를 10개로 시작하여 7,193개까지 단계적으로 늘려가며 성능을 비교해 봤을 때 제안된 기법이 얻

표 3 Newsgroups 문서 집합에서의 범주별 성능 비교

범주	OurMethod	Supervised NB	범주	OurMethod	Supervised NB
atheism	75.29	90.02	hockey	96.44	96.44
graphics	72.18	80	cryptography	81.03	93.23
mac	80.5	83.96	electronics	65.58	82.29
window.x	82.45	86.17	medicine	82.71	91.42
forsale	79.68	77.57	space	90.27	90
autos	85.37	90.5	christian	87.01	93.53
motorcycle	93.9	92.8	gun	88.4	91.82
baseball	96.44	95.97	mideast	87.7	92.14
			micro-avg.	83.46	91.2

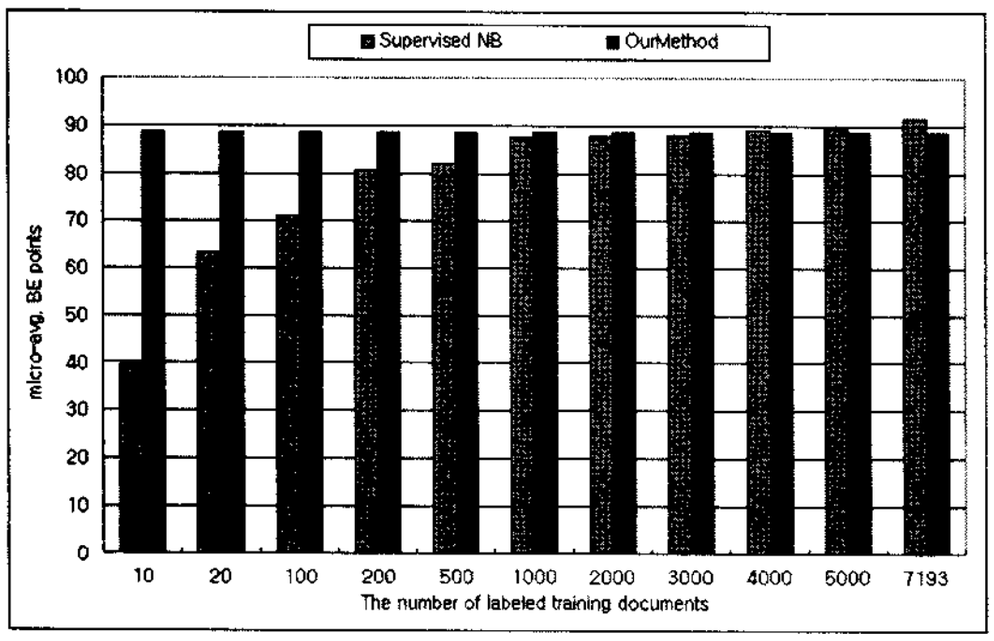

그림 5 지도학습 기법의 학습 문서의 개수에 따른 성능과 제안된 기법의 성능 비교

은 88.7%를 얻기 위해서는 약 4,000개의 학습 문서가 필요했다. 즉, 4,000개의 학습 문서를 범주화 작업을 수작업으로 수행하지 않고도 비슷한 성능을 얻을 수 있다는 결론을 얻을 수 있었다. 그러므로, 제안된 기법을 사용하면 범주화 작업의 학습 문서 획득의 어려움을 완화할 수 있을 것이다.

5. 결론 및 향후 과제

본 논문에서는 기존의 지도 학습 기반의 문서 범주화 기법과는 달리 수작업에 의한 대량의 학습 문서 생성 작업 없이, 각 범주의 핵심어의 입력만으로 문서를 자동으로 분류해내는 준지도 학습 기반의 새로운 기법을 제안한다. 그리고, 지도 학습 기반의 문서 범주화 기법과의 실험 결과를 살펴보면, 제안된 방법은 지도 기반의 문서 범주화 시스템과 근소한 성능 차이를 보이고 있다.

온라인상으로 얻을 수 있는 텍스트 문서의 양이 많아짐에 따라 학습 문서 생성을 위해 문서를 수집하는 것은 점점 쉬워지고 있으나, 각 영역에 맞는 대량의 학습 문서를 생성하는 것은 대단히 어려운 작업이다. 제안된 기법을 사용한다면 대량의 학습 문서 없이 적은 비용으로 문서 범주화를 수행하고자 하는 응용 영역에 유용하게 사용될 수 있을 것이다. 또한, 높은 성능을 요구하는 문서 분류 작업에서는 제안된 기법을 통해 손쉽게 범주가 할당된 학습 문서를 생성할 수 있을 것이다.

향후 과제로는, 먼저 범주별 핵심어를 사용하여 대표 문맥을 추출하는 과정에서 내용어의 의미 중의성(word sense ambiguity)문제가 발생하여 잘못된 대표 문맥이 추출되는 경우가 있었다. 이를 해결하기 위해 의미 중의성을 해결하기 위한 시스템을 개발하고 활용한다면 좀 더 좋은 성능을 보일 수 있을 것이다. 또한, 핵심어로부터의 부스트래핑 방법의 개선이 필요할 것으로 생각된다.

참 고 문 헌

[1] D. D. Lewis. "Naive (bayes) at forty: The independence assumption in information retrieval," *European Conference on Machine Learning*, 1998.

[2] A. McCallum and K. Nigram, "A comparison of Event Models for Naive Bayes Text Classification," *AAAI '98 workshop on Learning for Text Categorization*, 1998.

[3] D. D. Lewis and M. Ringuette, "A comparison of Two Learning Algorithms for Text categorization," *Proceedings of the 3rd Annual Symposium on Document Analysis and Information Retrieval*, 1994.

[4] C. Cortes and V. Vapnik. "Support vector networks," *Machine Learning*, 20:273-297, 1995.

[5] T. Joachims. "Text Categorization with Support Vector Machines: Learning with Many Relevant Features," *European Conference on Machine Learning (ECML)*, 1998.

[6] Y. Yang. "Expert netword: Effective and efficient learning from human decisions in text categorizatin and retrieval," *17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'94)*, pp. 13-22, 1994.

[7] D. D. Lewis, R. E. Schapire, J. P. Callan and R. Papka, "Training Algorithms for Linear Text Classifiers," *Proceedings of the 19th International Conference on Research and Development in Information Retrieval (SIGIR'96)*, pp. 289-297, 1996.

[8] E. Wiener, J. O. Pedersen, and A. S. Weigend. "A neural network approach to topic spotting," *Proceedings of the Fourth Annual Symposium on Document Analysis and Information Retrieval (SDAIR'95)*, 1995.

[9] Y. Yang and J. O. Pederson, "A Comparative study on feature selection in text categorization," *Proceedings of the 14th International Conference on Machine Learning*, 1997.

[10] K. Nigam, A. McCallum, S. Thrun, T. Mitchell, "Learning to Classify Text from Labeled and Unlabeled Documents," *Proceedings of 15th National Conference on Artificial Intelligence (AAAI-98)*, 1998.

[11] C. Languillon, Partially Supervised Text Categorization: Combining Labeled and Unlabeled Documents Using an EM-like Scheme, *Proceedings of the 11th Conference on Machine Learning, (ECML 2000)*, Vol.1810, LNCS, Springer Verlag, pp. 229-237, 2000.

[12] A. McCallum, K. Nigam, J. Rennie, and K. Seymore, Automatic the Construction of Internet Portals with Machine Learning, *Information Retrieval*, Vol.3, No.2, pp. 127-163, 2000.

[13] 고영중, *비지도 학습을 기반으로 한 자동 문서 범주화*, 서강대 석사학위 논문, 1999.
[14] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," *Proceeding of the 33rd Annual Meeting of the Association for Computational Linguistics*, pp. 189-196.
[15] Y. Maarek, D. Berry, and G. Kaiser, "An Information Retrieval Approach for Automatically Construction Software Libraires," *IEEE Transaction On Software Engineering*, Vol.17, No,8, pp. 800-813, August 1991.
[16] 조광제, 김준태. "역카테고리 빈도에 의한 계층적 분류 체계에서의 문서의 자동분류", 한국 정보과학회 봄 학술발표논문집(B), pp. 507-510, 1997.
[17] Y. Karov and S. Edelman, "Similarity-based Word Sense Disambiguation," *Computational Linguistics*, Vol.24, No.1, pp. 41-60, March 1998.
[18] S. Park, H. Kim, Y. Ko, and J. Seo, "Implementation of an efficient requirements analysis supporting system using similarity measure techniques," *Information and Software Technology, Elseviser*, Vol.42, No.6, pp. 429-438, 15 April, 2000.
[19] 김상범, 윤보현, 백대호, 한경수, 임해창, "문서 범주화를 위한 선형 분류기와 kNN의 결합 모델", *한국 인지과학회 춘계 학술대회 논문집*, pp. 255-231, 1999.
[20] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, and S. Slattery, "Learning to Extract Symbolic Knowledge from the World Wide Web," *Proceedings of the International Workshop on AAAI'98*, 1998.
[21] 오효정, 임정묵, 이만호, 맹성현, "점진적으로 계산되는 분류정보와 링크정보를 이용한 하이퍼텍스트 문서 분류 모델", *한글 및 한국어 정보처리 학술 대회 논문집*, pp. 89-96. 1999.
[22] Y. Ko, J. Park, and J. Seo, "Automatic Text Categorization using the Importance of Sentences," *Proceedings of the 19th International Conference on Computational Lin- guistics (COLING'2002)*, pp. 474-480, 2002.
[23] Y. Yang, "An Evaluation of statistical approaches to text categorization," *Information Retrieval Journal*, May, 1999.

고 영 중
정보과학회논문지 : 소프트웨어 및 응용
제 35 권 제 1 호 참조

서 정 연
정보과학회논문지 : 소프트웨어 및 응용
제 35 권 제 1 호 참조