# 문헌간 유사도를 이용한 자동분류에서 미분류 문헌의 활용에 관한 연구
# Utilizing Unlabeled Documents in Automatic Classification with Inter-document Similarities

김판준(Pan-Jun Kim)*, 이재윤(Jae-Yun Lee)**


## 초록

문헌간 유사도를 자질로 사용하는 분류기에서 미분류 문헌을 학습에 활용하여 분류 성능을 높이는 방안을 모색해보았다. 자동분류를 위해서 다량의 학습문헌을 수작업으로 확보하는 것은 많은 비용이 들기 때문에 미분류 문헌의 활용은 실용적인 면에서 중요하다. 미분류 문헌을 활용하는 준지도학습 알고리즘은 대부분 수작업으로 분류된 문헌을 학습데이터로 삼아서 미분류 문헌을 분류하는 첫 번째 단계와, 수작업으로 분류된 문헌과 자동으로 분류된 문헌을 모두 학습 데이터로 삼아서 분류기를 학습시키는 두 번째 단계로 구성된다. 이 논문에서는 문헌간 유사도 자질을 적용하는 상황을 고려하여 두 가지 준지도학습 알고리즘을 검토하였다. 이중에서 1단계 준지도학습 방식은 미분류 문헌을 문헌유사도 자질 생성에만 활용하므로 간단하며, 2단계 준지도학습 방식은 미분류 문헌을 문헌유사도 자질 생성과 함께 학습 예제로도 활용하는 알고리즘이다. 지지벡터기계와 나이브베이즈 분류기를 이용한 실험 결과, 두 가지 준지도학습 방식 모두 미분류 문헌을 활용하지 않는 지도학습 방식보다 높은 성능을 보이는 것으로 나타났다. 특히 실행효율을 고려한다면 제안된 1단계 준지도학습 방식이 미분류 문헌을 활용하여 분류 성능을 높일 수 있는 좋은 방안이라는 결론을 얻었다

## ABSTRACT

This paper studies the problem of classifying documents with labeled and unlabeled learning data, especially with regards to using document similarity features. The problem of using unlabeled data is practically important because in many information systems obtaining training labels is expensive, while large quantities of unlabeled documents are readily available. There are two steps in general semi-supervised learning algorithm. First, it trains a classifier using the available labeled documents, and classifies the unlabeled documents. Then, it trains a new classifier using all the training documents which were labeled either manually or automatically. We suggested two types of semi-supervised learning algorithm with regards to using document similarity features. The one is one step semi-supervised learning which is using unlabeled documents only to generate document similarity features. And the other is two step semi-supervised learning which is using unlabeled documents as learning examples as well as similarity features. Experimental results, obtained using support vector machines and naive Bayes classifier, show that we can get improved performance with small labeled and large unlabeled documents then the performance of supervised learning which uses labeled-only data. When considering the efficiency of a classifier system, the one step semi-supervised learning algorithm which is suggested in this study could be a good solution for improving classification performance with unlabeled documents.

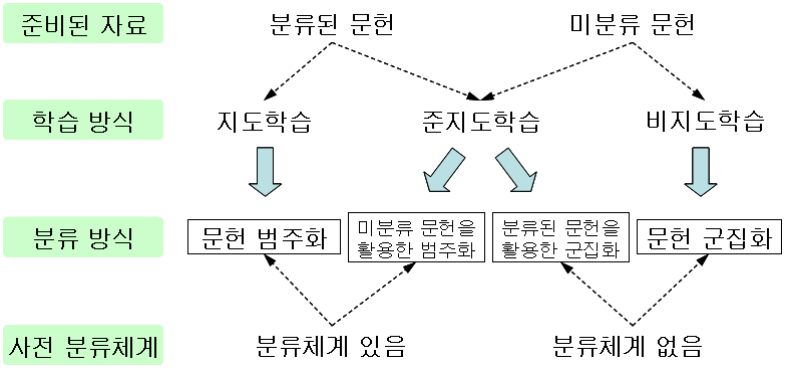

키워드: 문헌자동분류, 문헌 범주화, 준지도학습, 미분류문헌, 문헌유사도, SVM 분류기, 나이브베이즈 분류기, automatic classification, text categorization, semi-supervised learning, unlabeled documents, document similarities, SVM classifier, naive Bayes classifier

* 연세대학교 문헌정보학과 시간강사(dpbluesea@hanmail.net)
** 경기대학교 문헌정보학전공 조교수(memexlee@kgu.ac.kr)

## 1. 서론

1960년대에 시작된 문헌의 자동분류에 관한 연구는 1990년대 초반에 기계학습 기법이 도입된 이후 크게 활성화되었다. 전통적으로 문헌 자동분류는 사전 분류체계의 존재 여부에 따라 크게 두 가지 유형으로 구분하고 있다(정영미 2005; Dattola 1969). 사전 분류체계가 있는 경우에는 여러 기계학습 알고리즘을 통해 각 문헌을 가장 적절한 주제범주에 할당하게 되며 이 유형을 문헌 범주화라고 부른다(Lewis & Gale 1994; Vapnik 1995; McCallum & Nigam 1998), 반면에 사전 분류체계가 없이 유사한 문헌끼리 군집을 형성하는 방식은 문헌 군집화라고 부른다(Jain & Dubes 1988; Wagstaff et al. 2001).

최근에는 지도학습의 경우에도 기계학습 알고리즘을 위한 학습집합의 구성에서 미분류 문헌의 활용 여부에 따라 다시 두 가지 유형으로 구분하고 있다. 즉, 사전에 범주가 부여된 분류된 문헌(labeled documents)만을 학습집합으로 사용하는 완전지도학습(complete supervised learning)과 범주가 부여되어 있지 않은 미분류 문헌(unlabeled documents)을 학습에 포함시키는 준지도학습(semi-supervised learning)이 그것이다(Liu et al. 2003). 또한, 대표적인 비지도학습 기법이라 할 수 있는 클러스터링에서도 소수의 분류된 문헌을 활용하는 준지도학습 클러스터링(semi-supervised clustering)에 관한 연구가 다수 이루어지고 있다(Basu, Banerjee, & Mooney 2002; Basu, Bilenko, & Mooney 2004; Cohn, Caruana, & McCallum, 2003). 그러므로, 최근 연구되는 문헌 자동분류는 사전 분류체계의 유무와 미분류 문헌의 활용이라는 두 가지 측면을 고려하여 <그림 1>과 같이 구분할 수 있다.




<그림 1> 기계학습 방식과 문헌분류 방식의 구분 및 대응 관계

대규모 문헌집단의 모든 문헌을 특정 시점에 일괄적으로 수작업으로 분류하는 것은 많은 시간과 노력이 필요하다. 대부분의 경우에 자동분류 시스템을 도입하려는 이유는 수작업분류가 어렵기 때문이므로 자동분류 시스템 도입 시점에는 분류되지 않은 문헌이 상당수 존재하기 마련이다. 따라서, 수작업으로 분류된 문헌이 많지 않은 상황에서 나머지 대부분의 미분류 문헌들을 학습에 활용하기 위해 제안된 준지도학습은 실제적인 대안으로서 상당한 가능성을 보여주고 있다.

이러한 준지도학습 관련 연구는 문헌을 표현하는 자질로서 단어를 사용하고 있으며, 흔히 분류된 문헌만을 활용하는 1단계와 미분류 문헌을 추가로 활용하는 2단계로 구성된다. 1단계에서는 소규모의 분류된 문헌들만으로 학습된 분류기를 사용하여 남아있는 미분류 문헌들

을 자동 분류한 다음, 2단계로 1단계에서 분류된 전체 문헌들과 범주정보를 학습집합으로 사용하여 새로운 분류기를 생성한다. 결과적으로 검증집합에 대한 분류는 2단계에서 생성된 새로운 분류기를 사용하여 이루어진다.

이 연구에서는 문헌 범주화에서 미분류 문헌을 이용하되, 기존 연구와 달리 단어자질을 사용하지 않고, 최근 제안된 문헌간 유사도 기반 자질 표현 방식(이재윤 2005)을 이용하는 새로운 접근을 시도한다. 문헌간 유사도 기반 자질 표현 방식은 전통적인 단어 기반 자질 표현 방식보다 뛰어난 분류 성능을 나타내므로, 이에 대해서 미분류 문헌을 활용하여 성능을 더 높일 수 있는 방안을 모색하는 것이 이 연구의 목표이다.

## 2. 미분류 문헌을 활용한 범주화

### 2.1 범주화를 위한 준지도학습의 유형과 알고리즘

1990년 후반부터 소규모의 분류된 문헌과 나머지 대부분의 미분류 문헌을 활용하는 준지도학습 방법을 사용한 문헌 범주화에 관한 연구가 여러 학자에 의해 활발히 수행되어 왔으며, 그 결과 미분류 문헌이 분류기의 학습에서 상당한 기여를 할 수 있다는 사실이 판명되었다(Basu, Banerjee, & Mooney 2002; Bennett & Demiriz 1998; Blum & Mitchell 1998; Bockhorst & Craven 2002; Ghani 2002; Goldman & Zhou 2000; Hofmann 1999; Joahcims 1999; Muslea, Minton, & Knoblock 2002; Nigam & Ghani 2000; Nigam et al. 2000; Zhang 2000).

최근 연구되고 있는 미분류 문헌을 활용한 범주화(준지도학습)는 분류된 문헌 내 부정예제(negative examples)의 사용여부에 따라 다음과 같이 크게 두 가지 유형으로 구분할 수 있다.

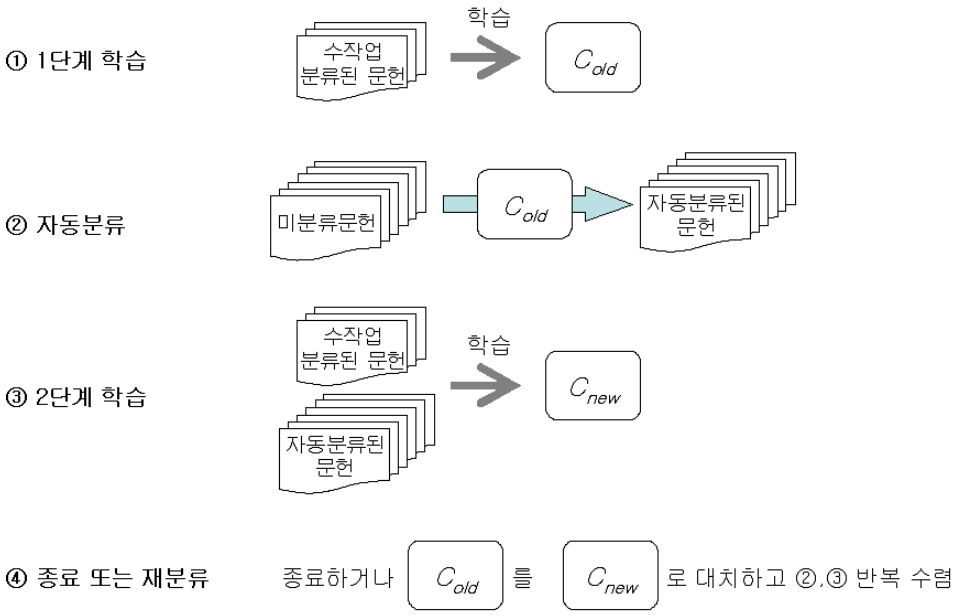
첫째, 소규모의 분류된 문헌과 대규모의 미분류 문헌 사용, 즉 분류된 문헌 중에서 긍정예제와 부정예제를 모두 사용하는 경우

둘째, 소규모의 긍정예제와 대규모의 미분류 문헌을 사용, 즉 분류된 문헌 중에서 긍정예제만을 사용하는 경우

소규모의 분류된 문헌과 대규모의 미분류 문헌을 활용한 준지도학습 알고리즘은 대부분 2단계로 이루어지는 학습에 기반하고 있다. 먼저, 소규모의 분류된 문헌으로 학습하여 생성된 분류기($c_{old}$)로 나머지 미분류 문헌을 분류하는 선행 분류단계를 거치고, 그 결과를 반영하여 학습한 새로운 분류기($c_{new}$)로 입력문헌을 분류한다. 이러한 2단계 알고리즘은 일반적으로 다음과 같은 절차에 따라 수행된다.

① 소규모의 분류된 문헌들만으로 학습한 분류기 $c_{old}$ 생성
② 나머지 미분류 문헌들을 분류하는데 분류기 $c_{old}$를 사용
③ 분류기 $c_{old}$로 자동으로 분류된 문헌을 학습집합에 추가하여 새로운 분류기 $c_{new}$ 생성
④ 그대로 학습을 마치거나 $c_{old}$를 $c_{new}$로 대치하여 ②와 ③ 단계를 반복하여 수렴시킴

준지도학습을 위한 2단계 알고리즘의 수행 절차를 그림으로 나타내면 <그림 2>와 같다.




<그림 2> 준지도학습을 위한 2단계 학습 알고리즘

## 2.2 최근 관련 연구

앞에서 살펴본 준지도학습을 위한 2단계 알고리즘에서는 처음의 선행 분류 단계에서 미분류 문헌에 자동으로 범주를 부여한 다음, 이 범주 정보를 이후의 2단계 분류에서는 수작업으로 부여한 범주와 마찬가지로 학습에 활용한다. 따라서 미분류 문헌을 분류하는 선행 분류 단계의 성능이 준지도학습에 의한 분류 성능에 큰 영향을 미친다. 준지도학습의 성능 향상을 위해서 최근 검토된 개선 방안은 동시-학습(co-training) 알고리즘, 부정예제의 배제, 1단계와 2단계에 사용되는 분류기의 최적 조합 모색 등이다.

Ghani(2002)는 분류된 문헌과 미분류 문헌을 조합하는 접근법으로 먼저, 복수클래스 문제를 복수의 이진 분류 문제로 분리하고 개별 이진 분류 문제를 학습하기 위한 동시-학습 기법을 사용하였다. 그는 학습을 위한 분류기로는 나이브베이즈를 사용하고 각 이진 분류의 학습에서 동시-학습 기법을 적용하는 자신의 접근법이 특히 많은 수의 범주를 포함하는 텍스트 분류에 유용하고 다른 준지도학습 기법보다 성능이 우수하다고 주장하였다. 그러나 Park과 Zhang(2004)은 준지도학습과 동시-학습 알고리즘을 함께 사용한 연구 결과에서 동시-학습 과정에서 미분류 문헌에 잘못된 범주가 할당될 수 있어 분류기들에 부정적인 영향을 줄 수도 있다고 하였다.

학습과정에서 미분류 문헌을 사용한 SVM(transductive SVM)의 성능을 평가한 Silva와 Ribeiro(2004)는 특히 재현율 측면에서의 성능 향상을 보고하였는데, 이는 대부분의 실제 분류 환경에서 아주 작은 수의 긍정예제들만이 존재하는 경우에 미분류 문헌의 활용이 아주 중요할 수 있음을 증명하는 것이다. 이처럼 소규모의 분류된 문헌과 대규모의 미분류 문헌을 활용한 준지도학습 관련 연구들은 학습과정에서 미분류 문헌의 활용이 분류 성능을 크게 향상시킬 수 있음을 보여주었다.

준지도학습 방법 중에서도 소규모의 긍정문헌과 대규모의 미분류 문헌을 사용한 문헌의

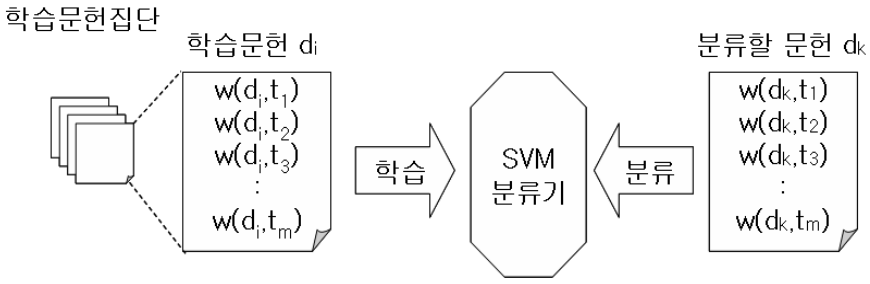
자동분류는 부정예제를 사용하지 않는다는 특징을 갖는데(Yu, Zhai, & Jiawei 2003), 이러한 Probably Approximately Correct(PAC) 학습에 대한 이론적 연구는 Denis(1998)에 의해 처음 소개되었다. Blum과 Mitchell(1998)은 서지 데이터의 검색에서 긍정예제들을 두 가지 측면에서 추출하였는데, 첫 번째 학습집합은 표제, 저자, 초록, 편자로 구성하였고 두 번째 학습집합은 논문의 전문으로 구성하였다. 여기서 두 가지 측면에서 추출한 자질집합 각각에 대한 기본 분류기들을 점진적으로 구축하여 이들의 결과를 조합하는 것을 동시-학습 알고리즘이라 하였다. 또한 Denis 등(2002)은 긍정예제와 미분류 문헌들을 사용한 나이브 베이즈 알고리즘(PNB)을 제안하였는데, 사용가능한 긍정예제의 수가 작을 경우에 동시-학습 알고리즘이 분류 성능을 상당히 개선할 수 있다고 보고하였다. 한편 이 방법의 단점은 이용자에게 긍정 범주의 확률을 입력할 것을 요구하므로 실제적인 환경에서 이용자에게 제공하기 어려울 뿐만 아니라 학습과정에서 미분류 데이터를 무시하고 긍정 데이터로만 학습할 가능성이 있다는 것이다.

이상에서 살펴본 미분류 문헌을 사용한 준지도학습 관련 연구들은 부정예제의 사용 여부에 상관없이 거의 모두가 2단계로 분류기를 학습하는 알고리즘을 기초로 하고 있다. 결국 사람이 분류한 문헌 이외에 기계가 자동분류한 문헌도 학습 데이터로 활용하므로 일부 잘못된 판정 결과가 학습 정보로 포함되는 것을 피할 수 없다. 따라서 분류 성능은 미분류 문헌도 활용하는 준지도학습 방식이 분류된 문헌만 활용하는 지도학습 방식보다 좋지만, 모든 미분류 문헌을 수작업 분류한 후 학습에 활용하는 경우의 지도학습방식보다는 분류성능이 낮을 수밖에 없다. 또한 기존의 모든 연구들은 분류기에 입력되는 문헌 표현 방식으로 전통적인 단어 자질 표현 방식을 채택하고 있으므로, 문헌 유사도 표현 방식에 준지도학습 방식인 2단계 학습 알고리즘을 그대로 적용하는 것이 바람직한가는 검증할 필요가 있다.

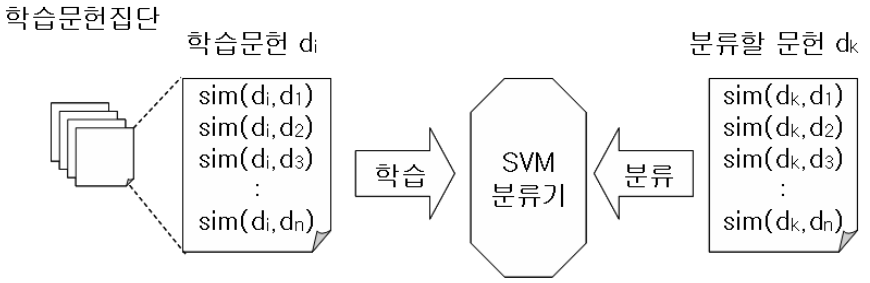
## 3. 실험 설계

### 3.1 문헌간 유사도를 이용한 분류기에서의 준지도학습

문헌간 유사도를 이용한 분류기에서 문헌을 표현하는 자질집합은 <그림 3>과 같이 단어 가중치가 아닌 문헌간 유사도이다. 각 문헌을 표현하는 벡터의 요소들이 학습집합에 포함된 각 문헌과 그 문헌과의 유사도로 구성되므로 학습문헌이 n개라면 각 문헌벡터의 크기도 n 차원이 된다.



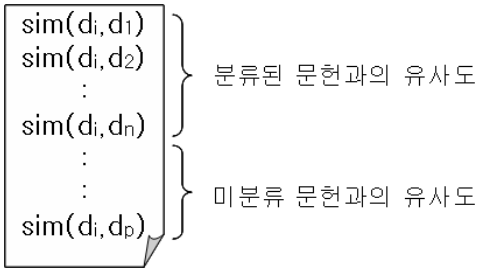

(a) 자동분류에서 색인어 자질을 이용한 문헌 표현




(b) 자동분류에서 문헌유사도 자질을 이용한 문헌 표현

<그림 3> 문헌 벡터의 색인어 자질 표현과 문헌유사도 자질 표현의 비교 (이재윤 2005a)

이와 같은 문헌유사도 자질 표현 방식에서는 문헌간 유사도를 산출하기 위해서 학습문헌과 유사도를 구하는 대상 문헌이 반드시 분류된 문헌일 필요가 없다. 만약 자동분류를 실행하기 이전에 준비된 문헌이 p개이고 이중에서 n개는 분류가 되어있다면 <그림 4>와 같이 n개의 분류된 학습문헌을 p개의 문헌과의 유사도로 나타낼 수 있는 것이다.



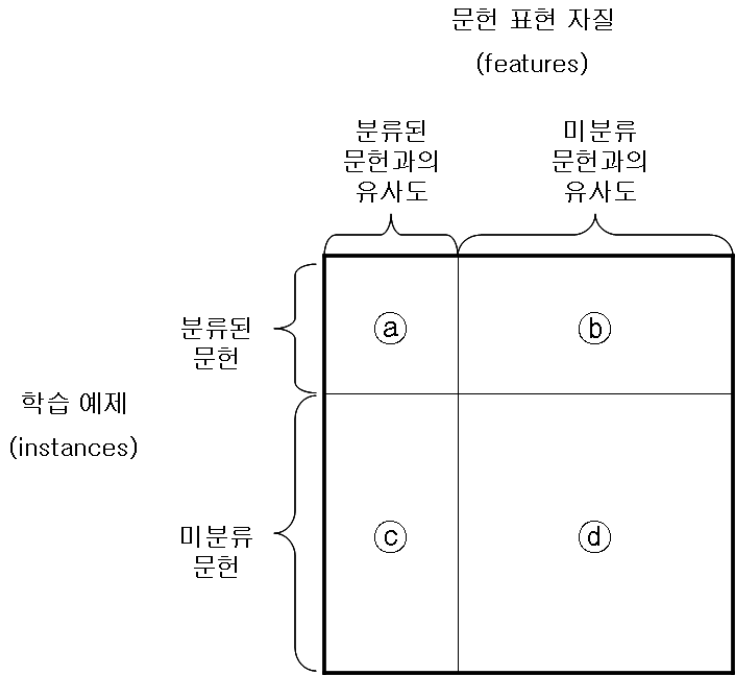

<그림 4> 미분류 문헌을 포함한 문헌유사도 표현 방식

미분류 문헌을 활용한 문헌의 자동분류에서 문헌간 유사도 자질에 기반한 분류기를 사용하는 경우에는, 학습집합에 대한 미분류 문헌의 포함에서 단어 자질에 기반한 분류기와는 다른 고려사항이 필요하다. 왜냐하면, 문헌간 유사도를 이용한 분류기에서는 학습집합에서 각 문헌을 표현하는데 사용된 자질들 자체가 특정문헌과 다른 문헌간의 유사도이므로, 미분류 문헌을 자질집합에만 추가하여도 해당 문헌과의 유사도를 반영하여 학습하는 효과를 갖기 때문이다. 따라서, 문헌간 유사도를 이용한 분류기를 사용하는 경우에는 다음과 같이 두 가지 측면으로 미분류 문헌을 학습에 활용할 수 있다.

첫째, 미분류 문헌을 자질값인 유사도 산출에만 사용함.
둘째, 미분류 문헌을 자질값인 유사도 산출에 사용하면서 동시에 학습예제로도 사용함.

문헌유사도 행렬을 표현한 <그림 5>를 참고하여 설명하자면, 분류된 문헌만 학습에 활용하는 전통적인 지도학습 방식에서는 행렬의 ⓐ부분만 사용하는 반면에, 미분류 문헌까지 포함하여 학습예제로 삼는 전통적인 2단계 알고리즘의 준지도학습 방식에서는 행렬 전체(ⓐ+ⓑ+ⓒ+ⓓ)를 사용한다. 반면에 행렬의 ⓐ+ⓑ부분만 사용하는 것은, 수작업 분류된 문헌만을 예제로 학습하되 미분류 문헌과의 유사도까지 자질로 삼는 1단계 알고리즘의 준지도학습 방식이라고 할 수 있다.

이 논문에서는 문헌간 유사도를 이용한 자질표현 방식에서 미분류 문헌을 자질집합에만 포함하는 경우를 1단계 준지도학습 방식이라 하고, 미분류 문헌을 자질집합과 예제로서 포함하는 경우를 2단계 준지도학습 방식으로 표현하기로 한다. 2단계 준지도학습 방식은 기존의 준지도학습 방식을 문헌유사도 기반 자질표현 방식에 적용한 것이며, 1단계 준지도학습 방식은 이와 달리 미분류 문헌의 자동분류 절차를 생략하므로 더 간단한 방식이다.




<그림 5> 문헌유사도 기반 자질표현 방식에서의 준지도학습 개념을 설명하는 문헌간 유사도 행렬

### 3.2 실험 문헌집단과 소프트웨어

이 연구에서 사용한 실험문헌집단은 TREND-2287 분류실험집단으로서 정보검색용 실험집단인 HANTEC v.2.0(김지영 외 2000)에서 분류정보가 포함된 해외과학기술문헌속보 문헌 2,287건을 추출한 것이다. 이중에서 학습문헌은 1997년 4/4분기에 등록된 1,178건이며, 검증문헌은 1998년 1/4분기에 등록된 1,109건으로 구성되어 있다. 각 문헌은 8가지 과학기술 주제로 구분되어 있다. 각 문헌의 색인어는 제목과 본문으로부터 자동 추출한 다음 학습문헌집단에서 문헌빈도가 2 이하인 단어를 제거하고 실험에 사용하였다.

자질 추출과 유사도 산출을 위한 프로그램은 Visual FoxPro로 구현하였고, 자동분류 실험을 위한 분류 시스템으로는 공개용 기계학습 실험 패키지인 WEKA 3.4 버전(Witten & Frank 2005)를 사용하였다. 분류 성능의 평가 척도로는 가장 많이 사용되는 마이크로 평균 $F_1$ 척도를 채택하였다. 실험 문헌에 범주가 하나씩 할당된 경우에 마이크로 평균 $F_1$ 척도는 마이크로 평균 정확률과 같으며 이를 계산하는 공식은 다음과 같다.

$$\text{마이크로 평균 } F_1 = \frac{\text{검증문헌이 적합 범주에 할당된 횟수}}{\text{검증문헌의 총 할당 횟수}}$$

### 3.3 실험 구성

문헌의 자동분류에서 미분류 문헌의 활용에 따른 성능을 크게 두 가지 측면에서 살펴보았다. 우선 1차 실험에서는 성능 비교를 위해서 분류된 문헌만을 활용하는 지도학습 방식의 성능을 학습집합의 규모를 달리하면서 살펴보았다. 2차 실험에서는 미분류 문헌을 문헌유사도 자질을 산출하는 용도로만 학습집합에 포함하는 1단계 준지도학습 방식의 성능을 살펴보았다. 마지막으로 3차 실험에서는 2단계 준지도학습 방식의 성능을 실험한 후 앞서 수행한 지도학습 방식 및 1단계 준지도학습 방식의 성능과 비교해보았다.

문헌간 유사도 자질 표현 방식에서는 학습예제로 사용되는 문헌과 유사도를 구하기 위해 비교되는 문헌이 일치하지 않을 경우에는 학습문헌별로 유사도 값의 분포 차이가 극심해지며 이는 분류 성능 저하의 원인이 된다. 따라서 이를 보정하기 위해 유사도 값을 보정해주는 굽이변환 방법(이재윤 2005a)을 사용하였다. 굽이변환 방식은 문헌과 문헌간의 코사인 유사도 값을 다음 공식으로 변환해준다.

$$y = \frac{x + ax}{x + a}$$

이 공식에서 $x$는 원래의 문헌간 코사인 유사도이고 $a$는 매개변수이며 $y$는 변환된 유사도이다. 이때 매개변수 $a$는 선행연구에서 좋은 성능을 보인 0.5로 설정하였다.

자동분류 알고리즘으로는 여러 분류 알고리즘 중에서 성능이 가장 좋은 것으로 알려진 지지벡터기계(이하 SVM으로 약칭)와 알고리즘이 간단하면서 실행속도가 빨라서 실용성이 높은 나이브베이즈 분류기(이하 NB로 약칭)를 사용하였다.

## 4. 실험 결과 및 분석

### 4.1 지도학습 실험

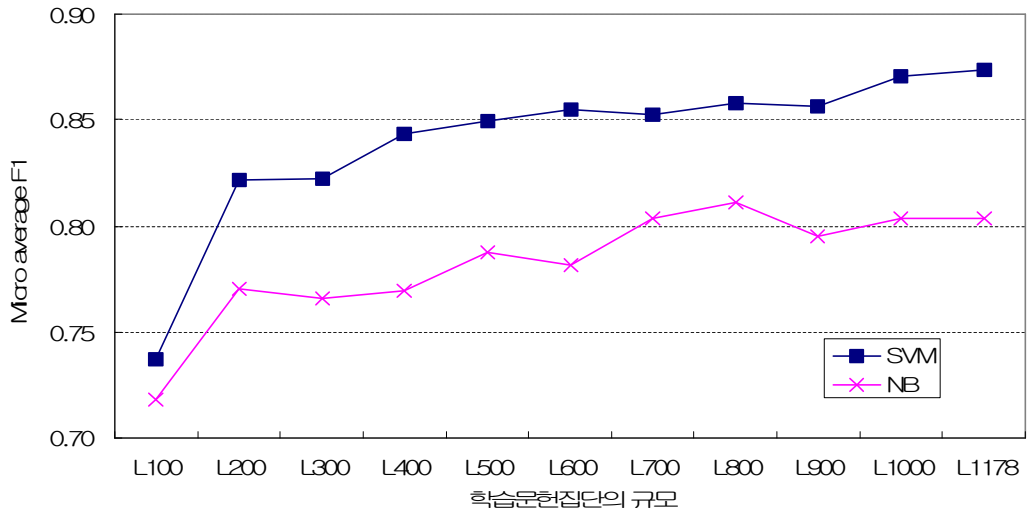
일단 준지도학습 방식의 성능을 평가하는 기준을 삼기 위해서 지도학습 방식의 성능을 파악하였다.

<표 1>과 <그림 6>은 실험문헌 집단에 대하여 2개 분류기(SVM, NB)에서 분류된 문헌만을 사용한 결과로서, 이후 미분류 문헌의 활용에 따른 분류 성능의 상한선으로 제시한 것이다. 이론적으로는 학습문헌집단의 모든 문헌이 수작업으로 분류된 경우가 가장 성능이 좋을 것이기 때문이다. 여기서 L은 분류된 문헌을 뜻하며 그 뒤의 숫자는 학습에 사용된 문헌의 수를 나타낸다. 따라서 L200이면 분류된 문헌 200개를 학습에 사용했다는 뜻이다.

<표 1> 지도학습 방식에서 분류된 학습문헌의 규모에 따른 성능 (마이크로 평균 $F_1$)

| 분류기 | L100 | L200 | L300 | L400 | L500 | L600 | L700 | L800 | L900 | L1000 | L1178 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SVM | 0.7367 | 0.8215 | 0.8224 | 0.8431 | 0.8494 | 0.8548 | 0.8521 | 0.8575 | 0.8566 | 0.8702 | 0.8738 |
| NB | 0.7178 | 0.7701 | 0.7656 | 0.7692 | 0.7872 | 0.7818 | 0.8034 | 0.8106 | 0.7953 | 0.8034 | 0.8034 |

* L 뒤의 숫자는 분류된 학습문헌의 수




<그림 6> 문헌유사도 기반 자질 표현을 사용한 지도학습 방식에서 학습문헌의 수에 따른 성능 비교

SVM과 NB 두 분류기 모두 학습문헌집단의 규모가 커질수록 문헌유사도 기반 자질 표현을 적용한 지도학습 방식의 분류성능이 대체적으로 향상되는 경향을 보였다. 다만 SVM 분류기의 경우에는 준비된 문헌이 모두 수작업으로 분류되어 학습에 사용된 경우(L1178)의 성능이 가장 좋은 반면에, NB 분류기는 800개의 문헌만 학습에 사용한 경우(L800)의 성능이 가장 좋게 나타났다. 이는 상대적으로 SVM 분류기의 성능이 더 안정적임을 뜻한다. 또한 학습문헌집단의 규모가 어떤 경우라도 항상 SVM 분류기가 NB 분류기보다 더 좋은 성능을 보였다. 이와 같은 결과는 SVM 분류기의 성능이 NB 분류기의 성능보다 좋으며, 학습

문헌집단의 규모가 클수록 분류성능이 더 좋다는 일반적인 경향과 일치한다.

<그림 6>을 보면 학습문헌의 수가 100개에서 200개로 증가할 때 성능이 가장 뚜렷하게 향상되는 것으로 나타났다. 학습문헌이 100개에서 200개로 증가할 때의 성능 향상값은 SVM 분류기가 0.085, NB 분류기가 0.052인데 반해서, 200개로부터 1178개로 학습문헌을 대폭 증가시키더라도 성능이 향상되는 폭은 대략 2/3에 불과한 0.052(SVM 분류기)와 0.033(NB 분류기)에 그친다. 이를 달리 생각해보면 분류된 문헌이 100여개로 최소한일 경우에 미분류 문헌을 사용하여 성능을 개선시킬 여지가 매우 크다고 할 수 있다.

## 4.2 1단계 준지도학습 실험

두 번째 실험에서는 미분류 문헌을 문헌유사도 자질을 산출할 경우에만 활용하고 학습예제로는 포함하지 않는 1단계 준지도학습 방식의 성능을 살펴보았다. 이 방식은 전통적인 단어자질 표현 방식에서는 불가능하고 문헌유사도 자질을 사용하는 경우에만 가능한 방식이다.

미분류 문헌의 활용에 따른 효과를 알아보기 위해 다음의 두 가지 측면에서 1차 실험을 수행하였다. 먼저, 분류된 문헌과 미분류 문헌을 조합하여 학습집합을 구성한 다음, 문헌간 유사도를 이용한 분류기의 성능을 알아보았다. 다음으로, 소규모의 분류된 문헌만이 확보된 경우(L100)를 가정하여, 소규모의 분류된 문헌에 미분류 문헌을 일정하게 추가하면서 학습집합을 구성한 경우의 분류 성능을 알아보았다. 1단계 준지도학습 실험에서는 학습집합에 포함되는 미분류 문헌은 자질로만 사용되고 예제로는 사용되지 않으므로 미분류 문헌의 범주 정보를 획득하기 위한 선행 분류 과정이 필요하지 않다.

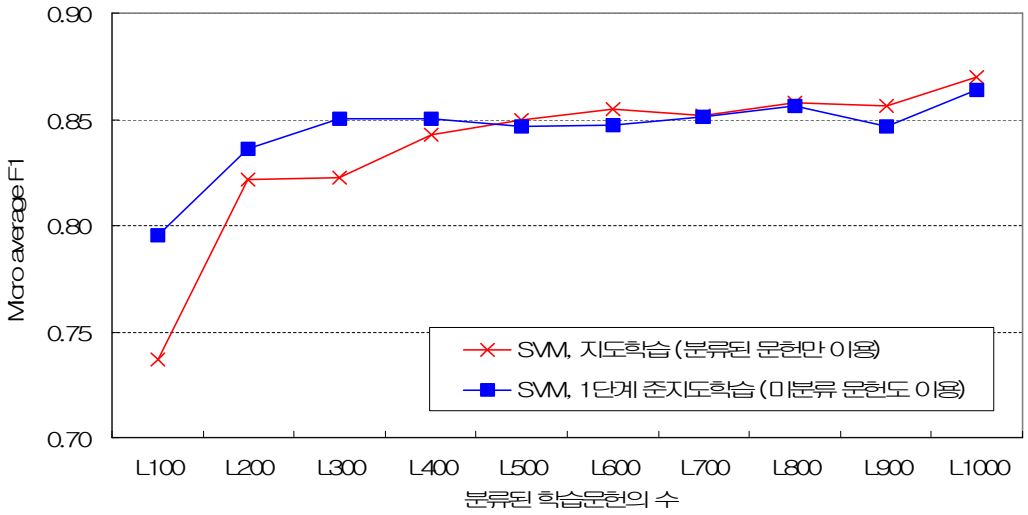
<표 2>는 분류된 문헌(L: Labeled documents)과 미분류 문헌(U: Unlabeled documents)의 조합에 따른 각 분류기의 분류 성능이다. 여기서 L100U1078은 분류된 문헌은 100개에 불과하고 미분류 문헌이 1078개로 훨씬 많은 경우이고, 반대로 L1000U178은 분류된 문헌이 1000개로 훨씬 많고 미분류 문헌은 178개에 불과한 경우이다. 이는 전체 학습문헌의 수는 1178개로 일정하지만 분류된 문헌의 비중이 증가함에 따라 달라지는 성능을 살펴보려는 것이다. 또한 1단계 준지도학습 방식의 성능을 지도학습 방식의 성능와 비교한 결과를 분류기별로 <그림 7>과 <그림 8>에 제시하였다.

<표 2> 1단계 준지도학습 방식에서 분류된 학습문헌의 비중에 따른 성능 (마이크로 평균 $F_1$)

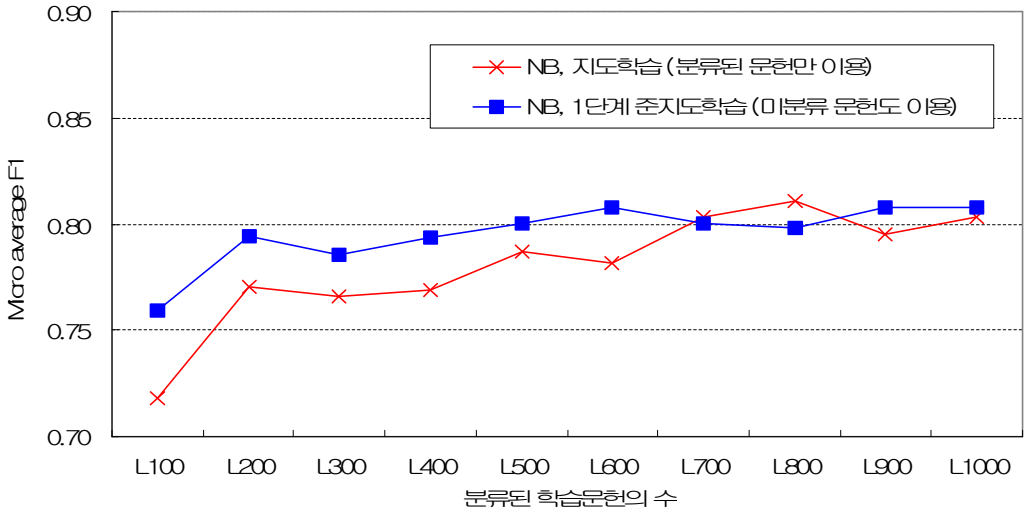
| 분류기 | L100 U1078 | L200 U978 | L300 U878 | L400 U778 | L500 U678 | L600 U578 | L700 U478 | L800 U378 | L900 U278 | L1000 U178 | L1178 U0 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SVM | 0.7953 | 0.8359 | 0.8503 | 0.8503 | 0.8467 | 0.8476 | 0.8512 | 0.8566 | 0.8467 | 0.8638 | 0.8738 |
| NB | 0.7592 | 0.7944 | 0.7854 | 0.7935 | 0.8007 | 0.8079 | 0.8007 | 0.7980 | 0.8079 | 0.8079 | 0.8034 |

* L 뒤의 숫자는 분류된 학습문헌의 수

* U 뒤의 숫자는 문헌유사도 자질값 산출에 사용한 미분류 문헌의 수




<그림 7> SVM 분류기에서 분류된 문헌만을 사용한 경우와 미분류 문헌을 포함하여 학습한 경우의 성능 비교




<그림 8> NB 분류기에서 분류된 문헌만을 사용한 경우와 미분류 문헌을 포함하여 학습한 경우의 성능 비교

1단계 준지도학습에서 분류된 문헌과 미분류 문헌의 비중을 달리한 실험의 결과는 다음과 같다.

첫째, 분류기 별로는 역시 SVM 분류기가 NB 분류기보다 항상 더 좋은 성능을 보였다.

둘째, 지도학습 방식과 비교해서 1단계 준지도학습 방식의 성능이 더 좋은 경우는 분류된 문헌의 비중이 낮고 미분류 문헌의 비중이 높은 경우였다. SVM 분류기에서는 1178개 학습 문헌 중에서 분류된 문헌의 비중이 대략 1/3 이상이 되었을 때(L400 이상) 지도학습과 1단계 준지도학습의 성능이 비슷하게 나타났다. NB 분류기에서는 분류된 문헌의 비중이 대략 2/3 이상이 되었을 때(L700 이상) 두 방식의 성능이 비슷해졌다. 이는 분류된 문헌이 어느

정도 규모 이상인 경우에는 소수의 미분류 문헌을 학습에 추가하는 것이 성능 향상에 그다지 도움이 되지 않음을 뜻한다.

셋째, 지도학습 방식과 1단계 준지도학습 방식의 성능 차이는 SVM과 NB 분류기 모두 분류된 문헌의 수가 100개로 비중이 가장 적은 경우에 최대로 나타났다(SVM/L100U1078: 0.059, NB/L100U1078: 0.042). 이는 문헌유사도가 아닌 단어 자질을 이용한 선행 연구의 결과와 일치하는 것이다(Nigam & Ghani 2000; Nigam et al. 2000; Muslea, Minton, & Knoblock 2002; Liu et al. 2003).

넷째, 분류된 문헌의 비중이 커짐에 따른 1단계 준지도학습 방식의 성능 향상 정도를 살펴보면 SVM 분류기에서는 분류된 문헌이 300개 이상인 경우부터, NB 분류기에서는 200개 이상인 경우부터 미미하게 나타났다. 이는 미분류 문헌을 문헌유사도 자질로 활용하는 1단계 준지도학습 방식을 사용한다면, 준비된 문헌을 모두 수작업으로 분류하지 않고 20%-30% 내외의 문헌만 분류해서 학습하더라도 비슷한 성능을 얻을 수 있음을 뜻한다.

다섯째, NB는 분류된 문헌이 600개일 때 최고 성능을 나타내었지만(L600U578: 0.8079), 전체적으로 SVM의 성능보다는 낮은 수준에 있을 뿐만 아니라 분류된 문헌만을 사용한 결과와 크게 차이가 없는 성능을 보여주어 미분류 문헌의 활용에 따른 효과가 크지 않음을 알 수 있다. 이렇게 볼 때 SVM은 최소한의 분류된 문헌만이 존재할 경우에 분류되지 않은 나머지 미분류 문헌을 활용하는 목적에 가장 적합한 분류기라고 할 수 있다. 문헌유사도가 아닌 단어 자질 표현에 기반하여 미분류 문헌을 학습에 활용하는 기존의 2단계 준지도학습 방식 연구에서도 이 연구와 마찬가지로 분류된 문헌이 매우 적은 경우에는 SVM 분류기가 미분류 문헌을 활용하기에 가장 적합한 분류기라고 결론지은 바 있다(Joachims 1999; Silva & Ribeiro 2004; Yu et al. 2003).

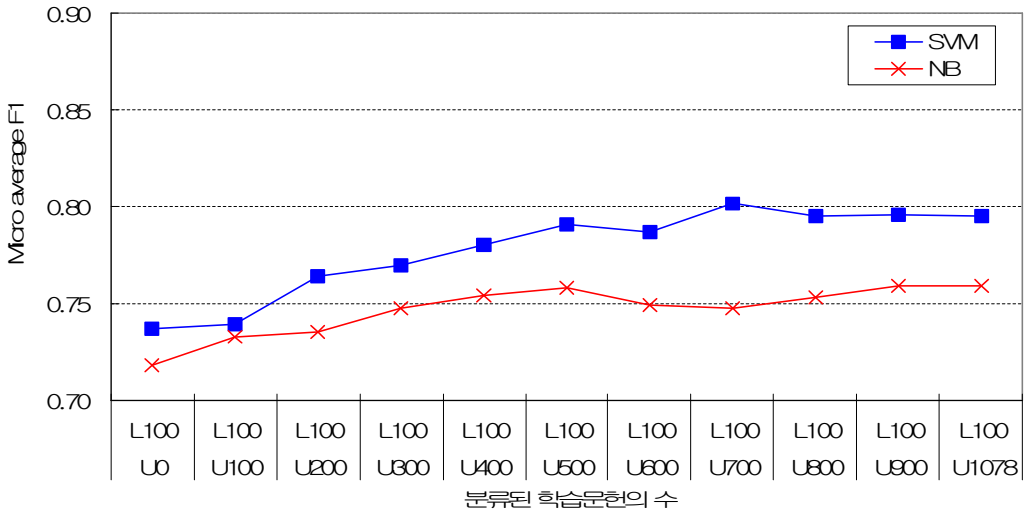
미분류 문헌의 활용 효과는 분류된 문헌이 100개로 매우 적을 때 가장 효과적임을 확인하였으므로, 다음으로는 분류된 문헌 100개에 미분류 문헌을 100개씩 추가하면서 1단계 준지도학습 방식의 성능을 살펴보았다. 이 실험의 목적은 최소한의 분류된 문헌이 존재하는 상황에서, 미분류 문헌을 추가하는 것이 지속적으로 성능 향상 효과를 얻는지 여부를 알아보고자 하는 것이다.

<표 3>과 <그림 9>는 분류된 문헌이 100개로 매우 적은 상황(L100)에서 미분류 문헌을 100개씩 지속적으로 추가하여(L100U100부터 L100U1078까지) 학습한 분류기의 성능이다. 미분류 문헌은 900개까지는 100개씩 추가하고 마지막은 나머지 178개를 모두 추가한 결과를 제시하였다.

<표 3> 1단계 준지도학습 방식에서 미분류 문헌의 추가 규모에 따른 성능 비교 (마이크로 평균 $F_1$)

| | L100 U0 | L100 U100 | L100 U200 | L100 U300 | L100 U400 | L100 U500 | L100 U600 | L100 U700 | L100 U800 | L100 U900 | L100 U1078 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| SVM | 0.7367 - | 0.7394 (0.37%) | 0.7638 (3.67%) | 0.7701 (4.53%) | 0.7800 (5.88%) | 0.7908 (7.34%) | 0.7872 (6.85%) | 0.8016 (8.81%) | 0.7953 (7.96%) | 0.7962 (8.08%) | 0.7953 (7.96%) |
| NB | 0.7178 - | 0.7331 (2.14%) | 0.7349 (2.39%) | 0.7475 (4.15%) | 0.7538 (5.03%) | 0.7583 (5.65%) | 0.7493 (4.40%) | 0.7475 (4.15%) | 0.7529 (4.90%) | 0.7592 (5.78%) | 0.7592 (5.78%) |

* L 뒤의 숫자는 분류된 학습문헌의 수
* U 뒤의 숫자는 문헌유사도 자질값 산출에 사용한 미분류 문헌의 수
* 괄호 안은 분류된 100개의 문헌만으로 지도학습한 경우(L100U0)의 성능 대비 향상율




<그림 9> 미분류 문헌의 추가 규모에 따른 1단계 준지도학습 방식의 성능 비교

실험 결과를 보면 두 분류기 모두 미분류 문헌을 추가할수록 대체적으로 성능이 향상되었으나 어느 정도 규모 이상이면 성능 향상 효과는 미미하게 나타났다. SVM 분류기는 미분류 문헌을 700개 추가한 경우(L100U700)에 가장 성능이 좋았고 이보다 더 많이 추가하면 오히려 약간 낮은 성능을 보였다. NB 분류기는 미분류 문헌을 최대한 많이 추가한 경우(L100U1078)가 가장 성능이 좋긴 하지만 500개 추가한 경우의 성능과는 차이가 미미하게 나타났다.

이런 결과에 근거해서 판단해보면 문헌유사도 자질 표현을 사용하는 1단계 준지도학습 방식에서 분류된 문헌이 100개 정도로 매우 적은 경우에는, 미분류 문헌을 다섯 배 내지 여섯 배 정도만 추가하는 것이 효율적이며 그 이상 추가하는 것은 성능 향상 효과를 크게 기대하기 어렵다고 할 수 있다.

## 4.3 2단계 준지도학습 실험

세 번째 실험은 소규모의 분류된 문헌과 대규모의 미분류 문헌을 활용한 선행연구들에서 주로 적용된 2단계 알고리즘에 기초한 것이다. 미분류 문헌 활용과 관련한 대부분의 기존 연구는 단어 자질에 기반하여 분류기를 학습하기 때문에, 미분류 문헌을 지도학습 방식으로 자동분류한 후 이를 학습예제로 사용하는 2단계 준지도학습 방식을 채택하고 있다. 그런데 문헌간 유사도에 기반한 분류기는 미분류 문헌에 범주를 자동할당할 경우에도 미분류 문헌을 자질로만 사용하는 1단계 알고리즘과 자질 및 예제로 동시에 사용하는 2단계 알고리즘의 두 가지 방식이 가능하다. 여기서는 이러한 두 가지 경우의 성능을 모두 살펴보았다.

2단계 준지도학습을 위해서는 선행 단계로 미분류 문헌을 자동으로 분류하는 작업이 필요하다. 여기서는 분류된 학습문헌을 100개로 가정했으므로 나머지 미분류 학습문헌 1078개는 범주를 자동으로 할당한 후 학습문헌으로 사용해야 한다. 이때 미분류 문헌 1078개를 자동분류하는 선행 분류 단계에서 다음과 같은 두 가지 방법이 가능하다.

- 지도학습 방식으로 미분류 문헌의 분류정보 획득: 이미 수작업 분류가 되어있는 100개만을 학습예제로 삼아서 지도학습 방식으로 미분류 문헌을 분류한다. 이는 전통적인 2단계 준지도학습에서의 선행 분류 방식과 같다.
- 1단계 준지도학습 방식으로 미분류 문헌의 분류정보 획득: 이미 수작업 분류가 되어있는 100개 문헌을 학습예제로 삼고 문헌유사도 자질로는 미분류 문헌까지 포함한 1178개 문헌과의 유사도를 사용하는 1단계 준지도학습 방식으로 미분류 문헌의 분류정보를 얻는다. 이는 문헌유사도 기반 자질 표현 방식을 적용했기 때문에 가능한 방식이다.

앞 절에서 1178개 학습문헌 전체와의 유사도를 자질로 사용한 1단계 준지도학습 방식의 성능이 지도학습 방식의 성능에 비해서 SVM은 7.96%, NB는 5.78% 향상된 성능을 보인 바 있다. 이는 검증문헌집단을 대상으로 얻은 성능이지만, 학습문헌집단에 속한 미분류 문헌을 분류하는 선행 분류에서도 지도학습 방식보다 1단계 준지도학습 방식이 더 좋을 것으로 기대할 수 있다. 실제로 미분류 문헌 1078개를 두 가지 방식으로 선행 분류하여 분류정보를 획득하였을 때의 성능을 구해보면 <표 4>와 같다.

<표 4> 2단계 준지도학습을 위한 선행 분류에서 미분류 문헌의 분류 성능 비교

| 분류기 | 미분류 문헌의 분류 방식 | |
|---|---|---|
| | 지도학습 (L100) | 1단계 준지도학습 (L100U1078) |
| SVM | 0.7282 | 0.7709 |
| NB | 0.5612 | 0.7393 |

* L 뒤의 숫자는 분류된 학습문헌의 수
* U 뒤의 숫자는 문헌유사도 자질값 산출에 사용한 미분류 문헌의 수

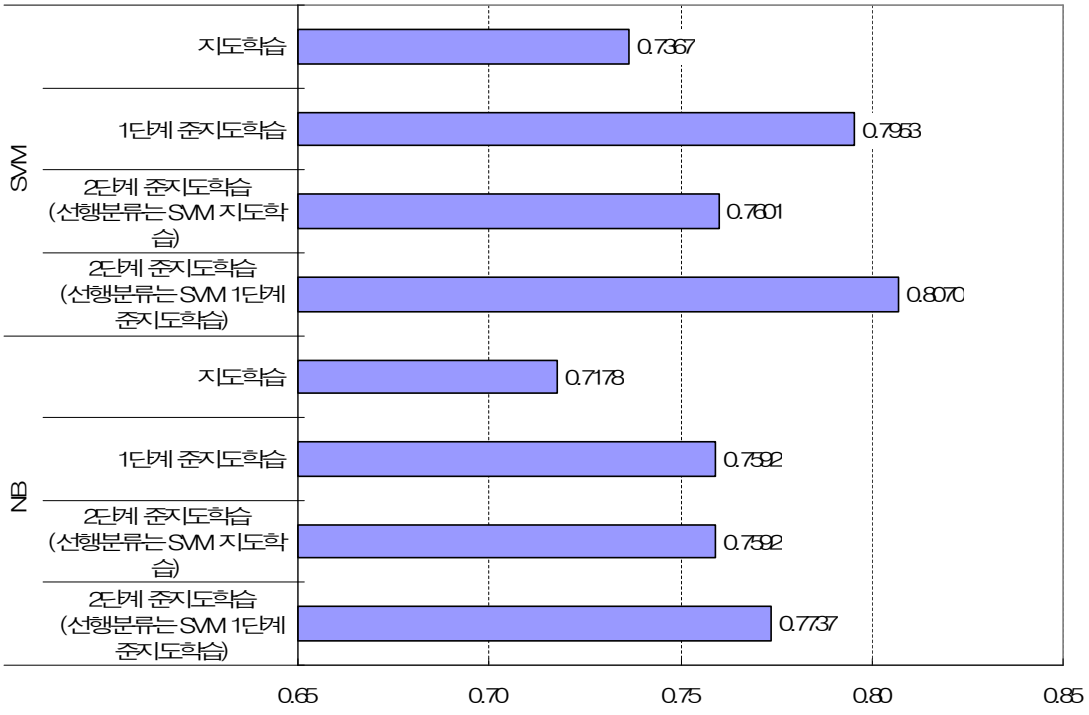
이 두 가지 선행 분류 방식을 미분류 문헌 1078개에 적용한 결과를 살펴보면 <표 4>와 같다. 이 표를 보면 1단계 준지도학습으로 선행 분류를 수행하면 SVM 분류기나 NB 분류기 모두 미분류 문헌의 분류정보를 훨씬 더 정확하게 얻을 수 있었다. 특히 NB 분류기는 지도학습 방식보다 1단계 준지도학습 방식으로 미분류 문헌을 분류한 결과가 월등하게 좋은 결과를 보였다. 그렇지만 지도학습이나 1단계 준지도학습 방식 모두에서 NB보다는 SVM 분류기의 성능이 더 좋았다. 2단계 준지도학습에서는 선행 분류에서 얻어진 미분류 문헌의 범주 정보가 정확할수록 이를 활용하여 학습하는 2단계 분류의 성능도 좋을 것이므로, 선행 분류를 기존과 다른 1단계 준지도학습 방식으로 SVM 분류기를 사용하여 수행하였을 때 더 좋은 성능이 기대된다.

최종적으로 전체 학습문헌 1178개 중에서 분류된 학습문헌이 100개로 매우 적을 경우의 성능을 학습방식과 분류기별로 구분하여 실험한 결과는 <표 5> 및 <그림 10>과 같다. 결과를 보면 SVM과 NB 분류기 모두 미분류 문헌을 사용하지 않은 지도학습 방식(<표 5>의 ⓐ)의 성능이 가장 낮으며, 선행 분류를 SVM 분류기로 1단계 준지도학습 방식으로 수행한 2단계 준지도학습 방식(<표 5>의 ⓓ)의 성능이 가장 좋았다. 전통적인 2단계 알고리즘처럼 미분류 문헌을 지도학습으로 분류한 후 학습집단에 포함시킨 2단계 준지도학습 방식(<표 5>의 ⓒ)의 성능은 선행 분류 절차가 없는 1단계 준지도학습 방식(<표 5>의 ⓑ)에 비해서 SVM 분류기에서는 더 낮았고 NB 분류기에서는 동일하게 나타났다. 문헌유사도 자질 표현 방식에서만 가능한 1단계 준지도학습 방식(<표 5>의 ⓑ)의 성능은 복잡한 과정을 거치는 2

단계 준지도학습 방식(<표 5>의 ⓓ)의 성능보다는 약간 낮게 나타났으나 SVM 분류기에서의 차이는 미미했다. 따라서 비용대 효과면에서 본다면 비교적 간단한 1단계 준지도학습 방식이 실용적이라고 판단된다.

<표 5> 분류된 문헌이 100개일 때 지도학습과 준지도학습의 성능 비교

| 분류기 | 학습방식 | 분류 성능 | 지도학습 대비 성능 향상율 |
|---|---|---|---|
| SVM | ⓐ 지도학습 | 0.7367 | - |
| | ⓑ 1단계 준지도학습 | 0.7953 | 8.0% |
| | ⓒ 2단계 준지도학습 (선행 분류는 SVM 지도학습) | 0.7601 | 3.2% |
| | ⓓ 2단계 준지도학습 (선행 분류는 SVM 1단계 준지도학습) | 0.8070 | 9.5% |
| NB | ⓐ 지도학습 | 0.7178 | - |
| | ⓑ 1단계 준지도학습 | 0.7592 | 5.5% |
| | ⓒ 2단계 준지도학습 (선행 분류는 SVM 지도학습) | 0.7592 | 5.5% |
| | ⓓ 2단계 준지도학습 (선행 분류는 SVM 1단계 준지도학습) | 0.7737 | 7.4% |




<그림 10> 분류된 문헌이 100개일 때 지도학습과 준지도학습의 성능 비교

결론적으로 문헌유사도 자질 표현의 장점을 살리는 1단계 준지도학습 방식이 단독으로 사용되거나 2단계 준지도학습 방식의 선행 분류로 사용할 경우에 모두 좋은 성능을 보였다.

따라서 단어 자질이 아닌 문헌유사도 자질 표현방식에서는 미분류 문헌을 활용할 때, 실행 효율을 최우선으로 고려한다면 지도학습 방식에 의존하는 전통적인 2단계 준지도학습 방식보다는 1단계 준지도학습을 채택하는 것이 바람직하다. 또한 최고 성능을 얻기 위해서는 선행 분류에서 1단계 준지도학습으로 미분류문헌에 범주를 할당해서 활용하는 2단계 준지도학습 방식을 채택해야 할 것이다.

## 5. 결론

문헌 자동분류를 도입하는 주된 이유는 수작업 분류에 드는 비용이 너무 크기 때문인데, 자동분류를 하기 위해서는 오히려 대량의 문헌을 사전에 수작업으로 분류해놓아야 한다는 것이 시스템 도입의 걸림돌이 되기도 한다. 이런 경우에 미분류 문헌을 이용한 자동분류는 좋은 대안이 될 수 있다. 이 연구에서는 최근 좋은 분류 성능을 보인 것으로 보고된 문헌간 유사도를 자질로 사용한 문헌 자동분류에서 미분류 문헌을 활용한 성능 향상을 모색해보았다.

1차 실험에서는 학습집단의 규모에 따른 지도학습 방식의 분류 성능을 살펴보았다. 분류 성능은 SVM 분류기가 NB 분류기보다 항상 좋은 것으로 나타났으며, 두 분류기 모두 학습문헌집단의 규모가 커질수록 문헌유사도 기반 자질 표현을 적용한 지도학습 방식의 분류성능이 대체적으로 향상되는 경향을 보였다. 특히 학습문헌의 수가 100개에서 200개로 증가할 때 성능이 가장 뚜렷하게 향상되는 것으로 나타났으므로, 분류된 문헌이 100개 정도로 매우 소규모일 때 미분류 문헌을 사용하여 성능을 개선시킬 여지가 크다는 것을 확인하였다.

2차 실험에서는 미분류 문헌을 문헌유사도 자질을 산출할 경우에만 활용하고 학습예제로는 포함하지 않는 1단계 준지도학습 방식의 성능을 살펴보았다. 이 방식은 전통적인 단어자질 표현 방식에서는 불가능하고 문헌유사도 자질을 사용하는 경우에만 가능한 방식이다. 분류된 문헌이 100개에 불과하다고 설정하고 나머지 미분류 문헌을 100개씩 추가해보면 SVM 분류기는 700개 정도, NB 분류기는 500개 정도 추가한 이후에는 더 추가하더라도 성능 향상 효과가 거의 없었다.

3차 실험에서는 2단계 준지도학습 방식을 미분류 문헌의 분류 방식에 따라 다시 두 가지로 구분하여 실험한 다음 지도학습 및 1단계 준지도학습 방식의 성능과 비교해보았다. 실험결과 문헌유사도 자질 표현 방식에서만 가능한 1단계 준지도학습 방식의 성능은 SVM 분류기를 사용한 지도학습의 성능보다 8.0% 향상되는 것으로 나타났다. 한편 지도학습과 준지도학습을 단계적으로 적용하는 전통적인 2단계 준지도학습 방식은 SVM 분류기를 사용한 지도학습의 성능보다 3.2% 향상되는데 그쳤다. 그렇지만 2단계 준지도학습을 수행하되 미분류 문헌의 선행분류를 지도학습이 아닌 1단계 준지도학습으로 분류한 경우에는 SVM 분류기를 사용한 지도학습의 성능보다 9.5%의 향상된 성능을 얻어서 가장 좋은 결과가 나타났다.

이상의 실험 결과를 바탕으로 다음과 같은 결론을 내릴 수 있다.

첫째, 문헌유사도 자질 표현 방식을 채택한 자동분류 시스템을 도입할 때, 대규모의 분류된 학습문헌을 사전에 확보하기가 어렵다면 미분류 문헌을 활용하여 분류 성능을 향상시킬 수 있다.

둘째, 문헌유사도 자질 표현 방식을 채택한 자동분류 시스템에서 분류된 문헌이 100개

내외로 소규모라면 1단계 준지도학습 방식에서 미분류 문헌을 다섯 배 내지 여섯 배 정도만 추가하는 것이 비용대 효과면에서 유리하다.

셋째, 문헌유사도 자질 표현 방식을 채택한 자동분류 시스템에서 미분류 문헌을 활용할 때, 비용대 효과면을 고려한다면 비교적 간단하면서 높은 성능을 보이는 1단계 준지도학습 방식을 채택하고, 분류 성능을 최우선으로 고려한다면 1단계 준지도학습으로 선행 분류를 수행하는 2단계 준지도학습을 채택해야 한다.

결론적으로 미분류문헌을 활용하는 것이 문헌유사도 자질 표현 방식을 채택한 자동분류 시스템을 도입할 때 효율성을 높일 수 있는 대안이라고 판단된다. 다만 이 연구에서 이용한 실험집단이 하나였으므로 실험 결과를 일반화할 수 있도록 추후 다양한 문헌집단에 대해서 문헌 유사도 기반의 준지도학습 방법을 적용해볼 필요가 있다. 또한 분류 성능 향상을 위해 개별 분류기가 아닌 다양한 분류기를 조합하는 방안도 모색해보고자 한다.

## <참고문헌>


김지영, 장동현, 맹성현, 이석훈, 서정현, 김현. 2000. 한국어 테스트 컬렉션 HANTEC의 확장 및 보완.『제12회 한글 및 한국어 정보처리 학술대회 논문집』, 210-215.

김판준. 2006. 기계학습을 통한 디스크립터 자동부여에 관한 연구.『정보관리학회지』, 23(1): 279-299.

이재윤. 2005a. 문헌간 유사도를 이용한 SVM 분류기의 문헌분류성능 향상에 관한 연구.『정보관리학회지』, 22(3): 261-287.

_____. 2005b. 자질 선정 기준과 가중치 할당 방식간의 관계를 고려한 문서 자동분류의 개선에 관한 연구.『문헌정보학회지』, 39(2): 123-146.

정영미. 2005. 정보검색연구. 서울: 구미무역(주) 출판부.

Basu, S., A. Banerjee, and R. Mooney. 2002. “Semi-supervised clustering by seeding.” *Proceedings of the Nineteenth International Conference on Machine Learning (ICML-02)*, 19-26.

Basu, S., M. Bilenko, and R. J. Mooney. 2004. “A probabilistic framework for semi-supervised clustering.” *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 59-68.

Bennett, K. P., and A. Demiriz. 1998. “A semi supervised support vector machines.” *Proceedings of the 1998 Conference on Advances in Neural Information Processing Systems II*, 368-374.

Blum, A., and T. Mitchell. 1998. “Combining labeled and unlabeled data with co-training.” *Proceedings of the Eleventh Annual Conference on Computational Learning Theory (COLT-98)*, 92-100.

Bockhorst, J., and M. Craven. 2002. “Exploiting relations among concepts to acquire weakly labeled training data.” *Proceedings of the 19th International Conference on Machine Learning (ICML-02)*, 43-50.

Cohn, D., R. Caruana, and A. McCallum. 2003. *Semi-supervised clustering with user feedback.* Technical Report TR2003-1892, Cornell University. [cited 2006. 11.

9]. <http://citeseer.ist.psu.edu/cohn03semisupervised.html>.

Dattola, R. T. 1969. "A fast algorithm for automatic classification." *Journal of Library Automation*, 2(1): 31-48.

Denis, F. 1998. "PAC learning from positive statistical queries." *Proceedings of the 9th International Conference on Algorithmic Learning Theory (ALT-98)*, 112-126.

Denis, F., R. Gilleron, and M. Tommasi. 2002. "Text classification from positive and unlabeled examples." *Proceedings of the 9th International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems (IPMU-02)*. [cited 2006. 10. 30]. <http://www.cmi.univ-mrs.fr/~fdenis/ipmu02.ps>

Ghani, R. 2002. "Combining labeled and unlabeled data for multiclass text categorization." *Proceedings of the 19th International Conference on Machine Learning (ICML-02)*, 187-194.

Goldman, S., and Y. Zhou. 2000. "Enhancing supervised learning with unlabeled data." *Proceedings of the 17th International Conference on Machine Learning (ICML-00)*, 327-334.

Jain, A. K., and R. C. Dubes. 1988. *Algorithms for Clustering Data*. Englewood Cliffs, NJ: Prentice-Hall.

Joachims, T. 1999. "Transductive inference for text classification using Support Vector Machines." *Proceedings of 16th International Conference on Machine Learning (ICML-99)*, 200-209.

Lewis, D. D., and W. A. Gale. 1994. "A sequential algorithm for training text classifiers." *Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 3-12.

Liu, B., Y. Dai,, X. Li, W. S. Lee, and P. S. Yu. 2003. "Building text classifiers using positive and unlabeled examples." *Proceedings of the Third IEEE International Conference on Data Mining (ICDM-03)*, 179-188.

McCallum, A., and K. Nigam. 1998. "Employing EM and pool-based active learning with keywords, EM and shrinkage." *Proceedings of 16th International Conference on Machine Learning (ICML-98)*, 359-367.

Muslea, I., S. Minton, and C. Knoblock. 2002. "Active + semi-supervised learning = robust multi-view learning." *Proceedings of the Nineteenth International Conference on Machine Learning (ICMA-02)*, 435-442.

Nigam, K., and R. Ghani. 2000. "Analyzing the effectiveness and applicability of co-training." *Ninth International Conference on Information and Knowledge Management (CIKM-00)*, 86-93.

Nigam, K., A. McCallum, S. Thrun, and T. Mitchell. 2000. "Text classification from labeled and unlabeled documents using EM." *Machine Learning*, 39(2/3): 103-134.

Park, Seong-Bae, and Byong-Tak Zhang. 2004. "Co-trained support vector

machines for large scale unstructured document classification using unlabeled data and syntactic information." *Information Processing & Management*, 40(3): 421-439.

Silva, C., and B. Ribeiro. 2004. "Labeled and unlabeled data in text categorization." *Proceedings of the IEEE International Joint Conference on Neural Networks (IJCNN-04)*, 2971-2976.

Wagstaff, K., C. Cardie, S. Rogers, and S. Schroedl. 2001. "Constrained k-means clustering with background knowledge." *Proceedings of 18th International Conference on Machine Learning (ICML-01)*, 577-584.

Witten, I. H., and E. Frank. 2005. *Data Mining: Practical Machine Learning Tools and Techniques*. 2nd ed. San Francisco: Morgan Kaufmann.

Yu, Hwanjo, ChengXiang Zhai, and Jiawei Han. 2003. "Text classification from positive and unlabeled documents." *Proceedings of the Twelfth International Conference on Information and Knowledge Management (CIKM-03)*, 232-239.

Zhang, T. 2000. "The value of unlabeled data for classification problems." *Proceedings of 17th International Conference on Machine Learning (ICML-00)*. [cited 2006. 10. 21.]. <http://citeseer.ist.psu.edu/549184.html>.