문서분류에서 가상문서기법을 이용한 성능 향상

이 경 순[†] · 안 동 언[††]

요 약

본 논문에서는 문서분류의 학습단계에 *가상적합문서기법*을 적용하여 성능을 향상시킬 수 있는 방법을 제안한다. 어떤 범주에 대해 적합하다고 판단된 두 개의 적합문서를 결합해서 생성된 문서 또한 적합문서가 된다는 관찰을 통해서, 문서분류기가 학습할 수 있는 새로운 정보를 추가함으로써 분류기의 학습을 돕는다. 제안하는 방법은 학습문서집합에 있는 적합문서들의 쌍을 조합해서 단순히 변환함으로써 가상의 문서를 생성한다. 이 방법에 의해서 생성된 가상 문서는 두 개의 적합문서에 같이 발생하는 어휘들에 대해서는 높은 가중치를 갖고, 문서 내의 어휘 공간이 확장되는 특성을 갖는다. 대량의 문서를 포함하는 TREC-11 필터링 태스크 참여에서 제안한 방법은 제공되는 학습문서를 이용한 기본 성능에 비해 71%의 성능 향상을 보였다. 또한 문서분류 연구에서 일반적으로 비교를 위해 이용하는 실험집합인 Reuters-21578에서 학습집합의 적합문서 개수가 100개 이하인 범주에 대해서 기본 학습문서를 이용한 분류기에 비해 11%의 성능향상을 보였다. 가상문서를 계속 추가해 나가면서 성능의 변화를 분석한 결과, 가상문서의 추가는 문서분류기의 학습능력을 도와 성능이 꾸준히 향상되고 있음을 보였다.

Performance Improvement by a Virtual Documents Technique in Text Categorization

Kyung-Soon Lee[†] · Dong Un An[††]

ABSTRACT

This paper proposes a *virtual relevant document* technique in the learning phase for text categorization. The method uses a simple transformation of relevant documents, i.e. making virtual documents by combining document pairs in the training set. The virtual document produced by this method has the enriched term vector space, with greater weights for the terms that co-occur in two relevant documents. The experimental results showed a significant improvement over the baseline, which proves the usefulness of the proposed method: 71% improvement on TREC-11 filtering test collection and 11% improvement on Reuters-21578 test set for the topics with less than 100 relevant documents in the micro average F1. The result analysis indicates that the addition of virtual relevant documents contributes to the steady improvement of the performance.

키워드 : 가상적합문서(Virtual Relevant Document), 담화단위(Discourse Unit), 지지벡터기계(Support Vector Machine), 문서분류(Text Categorization)

1. 서 론

본 논문에서는 문서분류에서 학습문서집합에 사전 지식인 적합문서들을 조합해서 가상의 적합문서를 생성해서 통합하는 가상적합문서(virtual relevant document) 기법을 제안한다. 연구의 동기는 학습문서집합의 크기가 작을 경우 문서분류시스템이 적절한 경계를 추정하기 어렵다는 관찰에 기반해서, 기존의 지식을 이용해서 새로운 정보를 생성, 추가함으로써 문서분류기가 학습할 수 있는 정보를 보다 풍부하게 제공하는 것이다.

문서분류(text categorization)는 이미 정의된 범주들 중에 새로운 문서가 속하는 범주를 결정하는 작업이고, 정보 필터링(information filtering or batch filtering)은 사용자의 관심을 나타내는 사용자 프로파일에 대해서 새로운 문서가 사용자의 관심에 부합하는 문서인지 아닌지를 거르는 작업이다. 라우팅(routing)은 질의에 대해 적합한 문서가 학습자료로 주어지고, 그것을 이용해서 새로운 질의를 생성해서 검색을 하는 것이다. 이들은 모두 학습자료를 이용해서 적절한 결정의 경계를 학습하는 지도 학습(supervised learning) 방법이고, 사용자의 질의, 프로파일, 범주를 하나의 주제로 본다면 같은 작업이라고 볼 수 있다. 이러한 연구에서는 학습 방법과 학습집합에 포함된 예제들이 시스템의 성능에 영향을 미친다. 학습집합에 포함된 문서들과 관련된 연구는 다음과 같다.

문서분류, 라우팅과 정보검색에서는 학습문서집합에 대해서 샘플링을 하여 다양한 기계학습에 적용함으로써 성능을

※ 본 연구는 한국과학재단 목적기초연구(R01-2003-000-11588-0) 지원으로 수행되었음.
† 정 회 원 : 전북대학교 전자정보공학부 교수
†† 종신회원 : 전북대학교 전자정보공학부 교수
논문접수 : 2003년 6월 30일, 심사완료 : 2004년 5월 25일

향상시키려는 연구가 많았다. Allen[1]의 연구에서는 라우팅 질의의 학습 단계에서 적합문서의 개수와 부적합문서의 개수를 동일하게 만들기 위해서 부적합문서의 일부를 제거하였다. 이는 Rocchio학습[9]에서 긍정적인 증거와 부정적인 증거의 균형을 맞추기 위한 것이다. Singhal[13]의 연구에서는 질의에 대해 어느 정도 관련이 있는 질의 영역(query zone)에 속하는 부적합문서들을 선택하여 라우팅 질의의 학습에서 이용하였다. Kwok[5]의 연구에서는 유전자 알고리즘에 기반한 피드백 질의를 생성하기 위해서 가장 효과적인 적합문서들의 일부를 선택하여 학습에 이용하였다. 이와 같은 샘플링 기법은 기본적으로 학습문서집합의 크기를 축소시키기 때문에, 이용가능한 학습문서의 크기가 작은 경우 학습이 어렵게 되어 성능 저하를 초래한다.

영상 인식, 영상 분류, 문자 인식 등의 기계학습 응용 연구에서는 샘플링보다는 사전 지식(prior knowledge)을 학습예제에 통합하는 연구가 많았다. 사전 지식은 학습예제 이외 이용 가능한 정보를 나타내는 것으로, 학습예제로부터 새로운 예제로 일반화시키는 것을 가능하게 한다[2]. 예를 들어, 영상 인식 시스템은 입력 영상을 이동, 회전, 스케일링 등으로 조금 변형시킴으로써 새로운 예제를 생성해서 이용한다. 음성 인식 시스템은 시간 왜곡(time distortion)이나 피치 변형(pitch shift)에 의해서 새로운 예제를 생성한다. 3차원 물체 인식 연구에서, Poggio[7]는 2차원 관점으로부터 새로운 관점을 생성하기 위해 적절한 변형을 적용하였다. 필기체의 숫자 인식 연구에서, Decoste[2]는 학습예제들을 네 방향에서 하나의 픽셀을 변형함으로써 가상의 새로운 예제를 생성하였다. 학습예제에 대한 어떠한 변형은 성능을 더 저하시킬 수 있기 때문에, 이러한 방법에서 고려할 문제는 학습예제에 어떤 변형을 통해 생성된 사전 지식을 통합하는 방법이 안전하게 적용될 수 있는지 하는 것이다.

문서분류에서의 어휘 공간에 대한 관찰에 의해 다음과 같은 것을 알 수 있다.

- 어휘 공간은 학습문서에서 나타난 어휘들로 제한된다.
- 새로운 문서의 자질(feature)도 학습문서에 나타난 어휘 공간으로 제약된다.
- 각 문서를 표현하는 문서의 어휘 공간은 그 문서에 나타난 어휘들로 제한된다.

이와 같이 학습문서의 제약 범위 안에서 학습을 위해 적절한 어휘 공간을 생성하는 방법의 개발이 필요하다. 이를 위해서 본 연구에서는, 문서분류에서 문서를 담화 단위(discourse unit)로 보고, 범주에 속하는 문서들을 같은 주제에 속하는 문서들의 그룹으로 보기보다는 일관된 담화(topical discourse)의 범주라고 가정한다. 이러한 관점은 학습문서집합에 속하는 문서들을 기존의 문서 단위(document unit)로 제한시키지 않고, 자유로이 학습문서들을 조작하는 것이 가능하게 해준다.

이러한 관찰과 가정을 바탕으로 하여, 본 연구에서 제안하는 가상적합문서기법은 학습문서집합의 적합문서에 속하는 문서들에 대해서, 두 개의 문서를 조합하여 하나의 문서의 단위로 확장을 한다. 새로 생성되는 가상문서의 어휘 공간은 각 하나의 문서에 포함된 어휘 공간에서 두 문서에 포함된 어휘의 공간으로 확장된다. 또한 가상문서의 어휘에 대한 가중치 계산에서는, 두 문서에 공통으로 나타나는 어휘는 담화 주제와 연관성이 높을 것이라는 가정으로, 그 어휘의 가중치들을 곱함으로써 높은 가중치를 부여한다. 생성된 가상문서는 하나의 문서 어휘 공간에서 두 개의 문서 어휘 공간으로 확장되고, 두 문서에 같이 나타나는 어휘는 다른 어휘들에 비해 높은 가중치를 가지게 되어, 문서분류기의 학습시 새로운 정보를 제공할 수 있을 것이다.

본 연구에서 제안하는 방법을 지지벡터기계에 기반한 분류기에서 실험을 하였다.

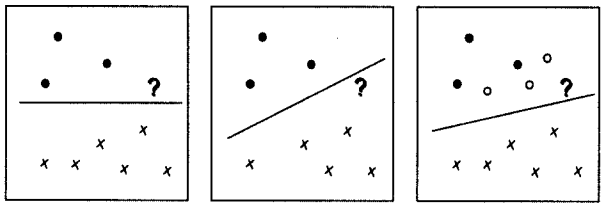
2장에서는 문서 공간에서의 가상 예제에 대해 설명한다. 3장에서는 기존의 학습문서집합에 가상적합문서를 통합하는 방법을 제안하고, 4장에서 실험과 결과에 대한 분석을 보이고, 5장에서 본 연구의 결론을 맺는다.

2. 문서분류에서 가상 문서의 기대 효과

하나의 문서를 그 문서가 나타내는 주제의 담화 공간에서의 하나의 객체 또는 단위라고 볼 수 있다. 뿐만 아니라, 하나의 문장도 그 주제를 설명하는 단위로 볼 수도 있다. 극단적으로 본다면, 어떤 주제와 관련이 있다고 판단된 문서들의 집합도 각 문서의 경계를 지워버리고 하나로 만들면 그 주제의 담화 공간에서 하나의 객체/단위라고 볼 수 있게 된다.

이와 같이, 담화 공간에서 어떤 주제에 대해 관련성을 유지한다는 조건하에서 새로운 단위는 자유롭게 생성이 가능하다. 직관적으로, 어떤 주제를 설명하는 두 개 또는 그 이상의 문서들을 연결함으로써 가상 문서를 생성할 수 있다. 이렇게 생성된 가상 문서는 그 주제에 대한 일관성을 유지할 뿐만 아니라 각각의 실제 문서에서 높은 중요도를 갖지 않지만, 주제와 관련이 있는 어휘들을 이용함으로써 그 주제의 표현을 더 명확하게 할 수도 있다.

문서분류기에서 학습문서집합을 이용하여 가상 문서를 생성하여 기존의 학습문서집합에 통합함으로써 기대하는 효과는 (그림 1)과 같다[12]. 주어진 학습문서집합에 의해 학습된 결정 경계 (그림 1)(b)에서는 '?'로 표시된 새로운 문서에 대해서 잘못된 분류를 하게 된다. 가상 문서가 주제를 적절히 표현한다면, (그림 1)(c)에서와 같이 보다 나은 결정 경계를 학습하는데 도움이 될 것이다.



●: 적합 문서 x: 부적합 문서 ○: 가상 적합 문서 ?: 테스트 문서

(그림 1) 결정 경계에서 가상 문서들에 대한 기대 효과 (a) 실세계에서의 결정 경계 (b) 주어진 학습문서집합에 대한 학습을 통해서 생성된 결정 경계 (c) 학습문서집합에 가상적합문서를 통합한 상태에서 학습된 결정 경계

3. 학습문서집합에 가상적합문서의 통합

3.1 가상적합문서 생성 방법

가상적합문서(VRD)는 학습문서집합에 있는 실제 적합문서를 임의로 결합하여 생성한다. 2개의 적합문서에서부터 n개의 적합문서를 결합할수 있으나, 본 연구에서는 두개의 적합문서를 결합하여 하나의 가상문서를 생성한다. 각 문서는 어휘와 어휘의 가중치로 된 벡터로 표현이 되어있는데, 새로 생성되는 가상문서에서의 어휘의 가중치는 두 실제 문서 벡터에 있는 어휘의 두 가중치를 곱한 값으로 취한다. 이는 두 문서에 같이 나타나는 어휘는 범주의 주제와 높은 연관이 있을 것이라는 가정을 바탕으로 한 것이다.

문서분류의 각 범주 C_k에 포함된 학습문서집합 중 적합문서 $d_1, d_2, \cdots, d_n$에 대해서

- [문서 표현] 각 문서는 어휘의 가중치 벡터 $d_i = <w_1, w_2, \cdots, w_n>$로 표현한다.
  - 어휘의 가중치는 정규화된 logTF*IDF로 계산한다.
- [가상문서 생성] 적합문서 두개에 대해서 하나의 가상 문서를 생성한다.

```
for i = d1 to dn {
   for j = di+1 to dn
        GenerateVirtualDoc(vrdij, i, j)
}
```

- **GenerateVirtualDoc** (vrd_{ij}, d_i, d_j) :
  문서벡터 $d_i = <w_{i1}, w_{i2}, \cdots, w_{in}>$와 $d_j = <w_{j1}, w_{j2}, \cdots, w_{jn}>$를 이용해서, vrd_{ij} 벡터를 생성한다.

$$vrd_{ij} = <w_{ij1}, w_{ij2}, \cdots, w_{ijn}>$$

  - 새로 생성되는 가상문서 vrd_{ij}의 각 어휘의 가중치 w_{ijk}는 다음과 같다 :

$$w_{ijk} = w_{ik} \cdot w_{jk} \quad (1)$$

  - 만약 w_{ik}와 w_{jk} 두 값 중 하나가 0이면, 0 대신 디폴트 값을 할당한다.
  - 가상 문서 벡터 vrd_{ij}의 어휘의 가중치는 코사인 정규화를 한다.

(그림 2) 학습문서의 적합문서벡터를 이용한 가상적합문서벡터 생성 방법

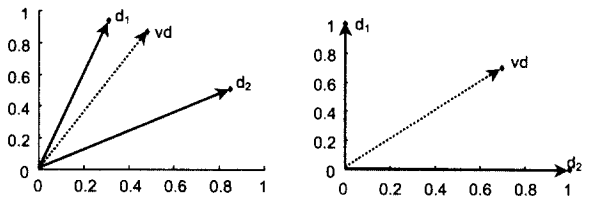
가상적합문서 생성에서의 가설은 두 벡터에 있는 모든 어휘를 포함하여 어휘 공간을 확장하고, 두 벡터에 함께 포함하는 어휘와 그렇지 않은 어휘를 구분하여 가중치에 변별력을 부여함으로써 문서분류에서의 결정 경계를 학습하는데 있어서 새로운 정보를 제공할 수 있다는 것이다.

가상적합문서를 생성하는 알고리즘은 (그림 2)와 같다.

개념적으로 2개 이상의 문서를 이용해서 하나의 가상문서를 생성할 수 있다. n개의 적합문서가 있다면, n(n-1)/2개의 가상 문서가 생성된다. 새로운 학습문서가 추가된다면, 기존에 생성된 가상문서에는 변화가 없고, 새로운 적합문서와 기존의 적합문서들을 이용하여 새로운 가상적합문서들을 생성하여 추가하게 된다.

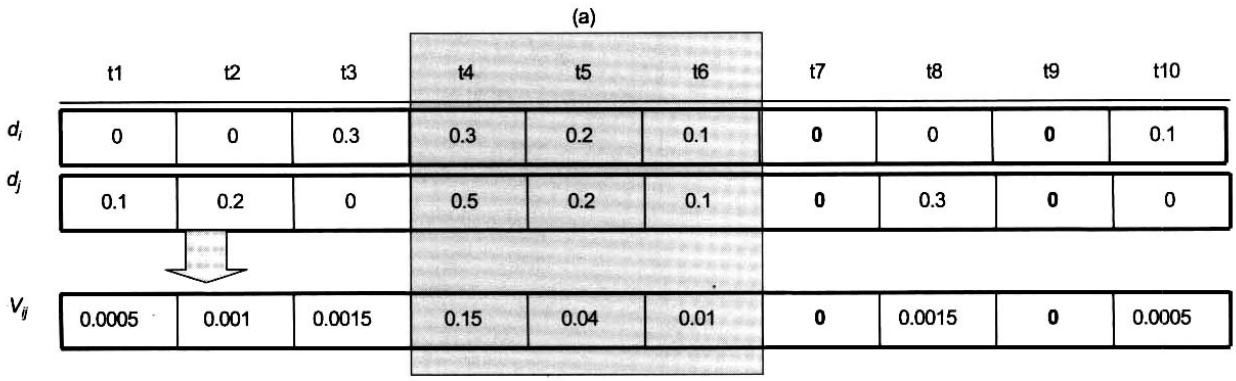
가상문서의 가중치 계산에서 만약 어휘 k가 di와 dj 중에서 하나의 벡터에만 나타난다면, 다른 하나의 벡터에서는 그 가중치가 0이 되므로((그림 3)(b)), 가상 문서에서 그 어휘는 드러나지 않게 된다. 가상 문서에서 그 어휘에 대한 정보를 유지하기 위해서, 가중치를 0대신 디폴트 값 0.005을 할당하였다. 두 벡터 모두 나타나지 않는 어휘는 가상문서에도 나타나지 않는 어휘가 된다. (그림 3)은 어휘가 두개로 된 2차원 문서벡터에서 생성되는 가상문서의 예를 나타내고, (그림 4)는 10개의 어휘로 구성된 문서에서 가상문서의 생성예를 보여준다.

어떠한 주제를 설명하기 위해서는 다양한 어휘들이 사용될 수 있고, 두 문서는 서로 다른 용어로 그 주제를 설명할 수 있기 때문에, 본 연구에서는 가상적합문서의 어휘 공간에 대해 두 문서에서 공유하는 어휘만을 취하지 않고, 두 문서에 포함하는 모든 어휘들을 포함하도록 하여, 어휘 공간을 확장시켰다. 2차원 어휘 공간에서 예를 들면, 벡터 d_1이 하나의 단어 t1을, 벡터 d_2가 하나의 단어 t2를 가질 때, 가상 벡터는 (그림 3)(b)에 나타난 것처럼 모든 단어를 갖게된다. 따라서 생성된 가상 문서는 어휘공간이 확장되고, 두 문서에서 공유하는 어휘는 높은 가중치를 갖게 된다.



(그림 3) 2차원 공간에서 실제 두 문서 d_1과 d_2에 대해 생성된 가상문서 vd

만약 두 문서를 단순히 연결해서 VRD를 생성한다면, VRD의 가중치는 어휘빈도수(tf)와 역문서빈도수(idf)에 의한 tfidf 가중치 기법으로 계산되었다면, 단순히 두 문서를 합한 것이 된다. 반면에, 곱셈 연산은 두 벡터에서 공유하는 어휘에 대해서 그렇지 않은 어휘에 비해 높은 변별력을 부여해 준다.



(그림 4) 두 적합문서 d_i와 d_j에서 생성된 v_{ij} 벡터 어휘의 가중치. (a)부분에 있는 어휘들은 두 벡터에 같이 포함하고 있는 어휘이다. 두 적합문서에 모두 나타나고 가중치가 높은 어휘는 범주와 관련이 높은 어휘일 가능성이 높다. 가상벡터에서 (a)부분의 어휘는 다른 어휘들에 비해 상대적으로 높은 가중치를 갖게된다(이 예에서의 가중치는 코사인 정규화를 거치기 전 상태임). 가상문서 벡터에서는 한쪽 벡터에 나타나지 않는 어휘(t1, t2, t3 등)라 하더라도 가상문서에는 그 정보를 유지하기 위해 가중치가 부여된다.

3.2 가상 문서와 지지 벡터

문서분류에서 가상적합문서의 효과를 살펴보기위해 지지벡터기계(Support Vector Machine)[14]를 이용하였다. SVM에서 지지벡터(Support Vector) 집합은 긍정적인 학습 자료와 부정적인 학습자료의 경계 결정을 위한 주요한 부분 집합으로, SVM이 학습집합에 대해서 학습을 통해서 선택한다.

Schölkopf[11]와 Vapnik[14]의 연구에 의해 SV 집합은 주어진 분류 문제를 풀기 위해서 필수적인 모든 정보를 포함한다는 것이 관찰되었다. 필기체의 숫자 인식에 대한 연구인 DeCoste[2]에서는 SV 집합에 대해서 가상 예제를 생성하는 것이 전체학습예제에 대해서 가상 예제를 생성하는 것과 비슷한 성능을 낸다는 것을 보였다.

이와 같은 연구들에서의 관찰에 따라, 본 연구에서는 모든 적합문서집합에 대해서 VRD를 생성하는 대신, SV 집합에 대해서 VRD를 생성하였다. 다음은 각 범주에 대해서 VRD를 생성하고, 주어진 학습집합에 통합해서 결정 경계를 학습하는 과정이다.

① 주어진 학습문서집합에 대해서 SVM을 학습시켜서, SV 집합을 얻는다.

② 생성된 SV 집합에 대해서 VRD를 생성한다.

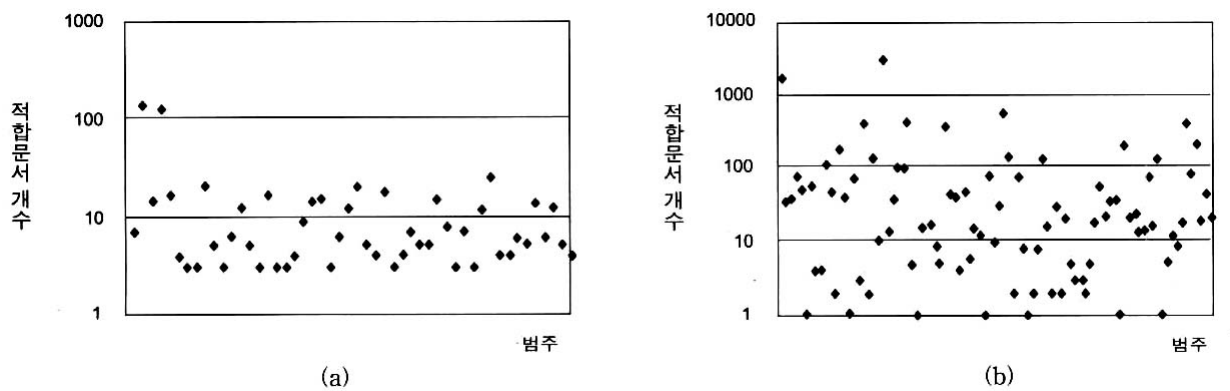
③ ①에서 생성된 SV 집합과 ②에서 생성된 VRD 집합을 통합한 학습문서집합에 대해서 SVM을 학습시킨다. 학습에 의해 생성된 SV 집합이 테스트 문서에 대한 분류를 결정하는 결정 경계로 이용된다.

SVM 학습은 두 번 이루어지고 있는데, 한번은 주어진 학습문서집합에 대해서 초기의 SV 집합을 얻기 위한 것이고, 또 한번은 SV 집합과 VRD 집합을 통합한 학습집합에 대해서 분류를 위한 결정 경계를 찾기 위한 것이다.

4. 실험 및 평가

4.1 실험 환경

본 연구에서 제안한 가상적합문서 기법의 효율을 검증하기 위해 TREC-11 필터링 테스트컬렉션과 Reuters-21578 테스트컬렉션에 대해서 평가를 하였다.



(그림 5) 적합문서 개수의 분포 (a) TREC-11 필터링 테스트컬렉션의 50개 주제에 대한 적합문서 수의 분포 (b) Reuters-21578 테스트컬렉션의 90개 범주에 대한 적합문서 수의 분포

TREC-11 일괄 필터링(batch filtering) 태스크[8]에서는 로이터코퍼스 1[10]을 이용하였다. 문서집합은 83,650개의 학습문서집합과 723,141개의 테스트문서집합으로 나누어져 있다. 주제(또는 범주)는 2가지 유형으로, 일반적인 TREC 질의 형태로 된 것(assessor topic)이 50개이고, 로이터의 범주들에 대해서 임의로 만든 것이 50개이다. 로이터 범주에 대해 임의로 생성된 질의는 TREC-11의 대부분의 참여자들의 결과에서 성능이 매우 낮은 문제점이 있어서[8], 본 연구에서는 TREC 질의 형태의 50개에 대해서 평가를 하였다. (그림 5)(a)는 50개의 범주들이 갖는 적합문서 개수의 분포를 보여준다. 전체 50개의 범주중에서 48개의 범주가 전체 83,650개의 학습집합문서 중에서 25개 이하의 문서를 적합문서로 갖고 있다. 작은 개수의 적합문서를 이용해서 분류를 위한 학습을 하기는 매우 어렵다.

Reuters-21578 집합은 문서분류의 많은 연구에서 이용해오고 있는 것이다[15, 4]. 본 연구에서는 Reuters-21578 집합에 대해서 학습문서집합과 테스트문서집합으로 분리한 것들 중에서 Lewis[6]에 의해 만들어진 ModApte 분리 집합을 이용하였는데, 7770개의 학습문서집합과 3019개의 테스트문서집합으로 구성되어 있다. 범주는 1개 이상의 학습문서와 테스트문서를 포함하는 것으로, 90개이다. (그림 5)(b)는 90개의 범주 각각이 갖는 적합문서의 분포를 보여준다. 30개의 범주가 10개 이하의 적합문서를 갖고 있고, 74개의 범주가 100개 이하의 적합문서를 포함하고 있다.

어휘에 대해 불용어(stop word)를 제거하고, 어근처리(stemming)를 해서, TREC-11 필터링 테스트컬렉션에 대해서는 125,846개의 어휘가, Reuters-21578 집합에 대해서는 16,422개의 어휘가 추출되었다. Reuters-21578에 대한 실험에서는 카이제곱에 의한 자질 추출을 거쳐서, 그 값이 10 이하인 어휘에 대해서는 두 벡터에 같이 공유된다 하더라도, 가상 문서의 가중치 계산에서 최소값을 갖도록 하여, 일반적으로 자주 발생하는 어휘에 대한 영향을 줄였다. 문서를 표현하기 위해서, 각 어휘는 logTF*IDF에 의한 가중치 계산을 하고, 문서내에서 정규화를 하였다. 문서분류 실험을 위해서 SVM^{light}시스템[3]을 이용하였는데, RBF 커널에 기반한 학습을 하였고, 학습을 위한 다른 파라미터는 디폴트값을 그대로 이용하였다.

비교 평가를 위해서, 각 범주에 대해 주어진 학습문서집합에 대한 SVM의 기본 성능과 제안한 가상적합문서 기법에 의한 SVM의 성능을 비교하였다:

- baseTR : 주어진 학습집합에 대한 SVM의 성능
- VRDsv : 본 연구에서 제안한 방법. 주어진 학습집합에서 학습된 SV 집합에 대해서 생성된 가상적합문서 집합과 SV 집합을 통합한 새 학습집합에 대한 SVM의 성능

문서분류의 성능을 비교를 위해, TREC-11 필터링태스크[8]에서 정의한 T11SU, T11F 평가척도와 문서분류에서 주로 이용하는 마이크로 평균 F1을 평가척도를 이용하였다.

① T11SU(scaled linear utility) : 선형 유틸리티 평가 척도는 검색된 적합문서에 대해서는 2배를 가산점으로 부여하고, 검색된 부적합문서에 대해서는 1배를 벌점으로 부여한다. MeanT11SU는 전체 범주에 대한 T11SU를 평균한 것이다.

$$T11SU = \frac{\max(T11NU, MinNU) - MinNU}{1 - MinNU} \quad (2)$$

여기서, T11NU = T11U/MaxU이고, T11U = $2R^+ - N^+$이다. MaxU = 2×(전체 적합문서 개수), MinNU = -0.5이다. R^+는 검색된 적합문서의 개수, N^+는 검색된 부적합문서의 개수이고, R^-는 검색되지 않은 적합문서의 개수를 나타낸다.

② T11F(F-beta) : F-beta는 재현율과 정확률에 대한 함수로, 재현율과 정확률에 각각에 가중치 파라미터를 부여해서 성능을 계산한다. 여기서는 베타값을 0.5로 하였다. MeanT11F는 전체 범주에 대한 T11F의 평균이다.

$$T11F = \frac{1.25 \times R^+}{0.25 \times R^- + N^+ + 1.25 \times R^+} \quad (3)$$

여기서, $R^+ + N^+$ 값이 0이 아닐 때 위와 같이 계산한다.

③ F1 : F1값은 재현율과 정확률의 값을 평균한 것으로, 마이크로 평균 F1은 모든 범주를 하나로 보고, 계산한 값이다.

$$F1 = \frac{2 \times P \times R}{P + R} \quad (4)$$

4.2 실험 결과

4.2.1 TREC-11 필터링 테스트컬렉션에 대한 결과 비교

<표 1>은 TREC-11의 필터링 테스트컬렉션에 대한 성능을 나타내고 있다. <표 2>는 각 범주에 대한 학습 과정에서의 통계 정보를 나타내고 있다. 적합 SV 집합 중 VRD로부터 추출된 것이 많이 포함되어 있다. SV 집합은 문서분류기가 새로운 문서에 대해서 분류를 결정하는 결정 경계에 직접 이용되는 정보인데, 분류 결정의 학습 결과 VRD가 결정 경계에 새로운 정보로 이용되고 있음을 보여주는 것이다.

〈표 1〉 TREC-11 필터링 테스트컬렉션에 대한 성능 비교

평가 척도	baseTR	VRDsv	성능 변화
MeanT11SU	0.359	0.376	4.74%
MeanT11F	0.090	0.190	111.11%
MicroAvgF1	0.181	0.310	71.27%

〈표 2〉 SVM 학습에 이용된 정보의 통계

	baseTR	VRDsv
학습문서집합의 평균 개수	861.48	328.24
적합문서의 평균 개수	12.78	216.08
VRD의 평균 개수		(204.92)
SV 집합의 평균 문서 개수	123.32	119.44
VRD에서 선택된 것		(15.32)
SV집합의 적합문서 중에서 선택된 것		(10.64)
SV집합의 부적합문서 중에서 선택된 것		(93.48)

baseTR에서 학습된 SV집합의 문서 크기는 123.32이고, VRDsv에서 학습된 SV집합의 문서 크기는 119.44로 비슷하다는 것을 알 수 있다. 이것은 가상적합문서 기법이 SVM의 결정 경계를 학습하는데 새로운 정보를 제공했음을 나타내는 것이다. 하지만, TREC-11 필터링 테스트컬렉션에서의 기본 SVM 성능 그 자체가 매우 낮다.

4.2.2 Reuters-21578 집합에 대한 결과 비교

Reuters-21578 집합에 대한 실험에서는 학습문서집합에 포함된 적합문서 개수에 따라 비교를 하였다. 100개이하의 적합문서(1 < rd < 100)를 포함하는 67개 범주에 대한 평가, 300개이하의 적합문서(1 < rd < 300)를 포함하는 76개 범주에 대한 평가, 500개이하의 적합문서(1 < rd < 500)를 포함하는 81개 범주에 대해서 평가를 하였다.

전체 범주 90개 중에서 7개 범주는 적합문서를 각 1개만 포함하고 있다. 제안한 가상적합문서 기법은 두 개 이상의 문서에 대해서 가상 문서를 생성할 수 있으므로, 본 실험에서는 제외시켰다. 또한 2개의 범주는 500개 이상의 적합문서를 포함하고 있어, 너무 많은 적합문서가 생성되고 실험집합 전체의 크기를 볼 때 충분한 학습이 되었다고 보므로 본 실험에서 제외시켰다. 참고로, 다른 연구와의 비교를 위해서 전체 90개 범주에 대한 성능을 제시하면, baseTR의 성능은 0.8495이고, VRDsv 성능은 0.8634이다.

<표 3>은 적합문서의 개수에 따른 범주들에 대해서 성능을 보여준다. 제안한 방법이 마이크로 평균 F1 평가 척도에서 67개의 범주에서는 10.9%, 76개 범주에서는 7.4%, 81개 범주에 대해서는 5.9%의 성능 향상을 보였다.

가상 문서의 가중치 계산에서 곱셈 연산을 하지 않고, 덧셈 연산을 이용해서 가상 문서를 생성한 경우에 대한 평가를 하였는데, baseTR의 성능 0.538과 비슷한 성능인 0.542를 보였다(67개 범주에 대해서 마이크로 평균 F1에 의한 성능 평가). 이는 가상 문서의 생성에서 덧셈 연산에 의한 가중치 계산은 학습에서 새로운 정보를 제공하지 못했음을 나타내고, 곱셈 연산에 의한 가중치 계산을 통해서 두 벡터에서 공유하는 어휘에 대해 높은 가중치를 부여함으로써, 학습에서 새로운 정보를 제공했음을 나타낸다. 또한, 가상문서를 주어진 전체학습집합에 대해서 생성하여 비교해 보았는데, SV 집합에서 생성한 경우는 0.596이고, 전체 학습집합에서 생성한 경우는 0.588을 나타냈다. 이는 SV 집합에 대해서 가상 문서를 생성하는 방법이 안전함을 알 수 있다.

〈표 3〉 Reuters-21578 집합에서 적합문서 개수에 따른 범주의 성능 비교(rd는 적합문서임)

평가 척도	67 범주(1 < rd < 100)			76 범주(1 < rd < 300)			81 범주(1 < rd < 500)		
	baseTR	VRDsv	chg%	baseTR	VRDsv	chg%	baseTR	VRDsv	chg%
MeanT11SU	0.507	0.543	7.14%	0.538	0.575	6.97%	0.553	0.591	6.87%
MeanT11F	0.460	0.503	9.31%	0.506	0.550	8.62%	0.527	0.566	7.42%
MicroAvgF1	0.538	0.596	10.92%	0.634	0.681	7.39%	0.707	0.748	5.90%

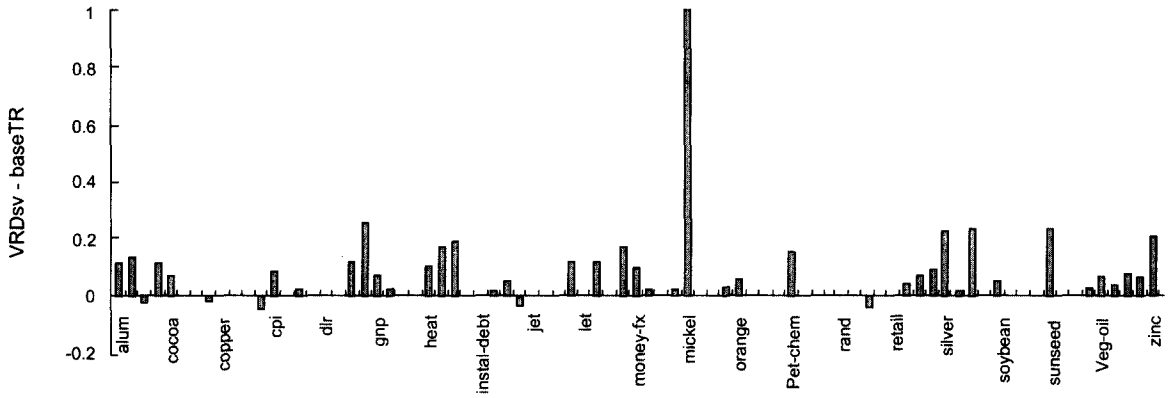
4.3 결과 분석

(그림 6)은 Reuters-21578 집합에서 81개 범주에 대해 baseTR의 성능에 대한 VRDsv의 성능의 변화를 살펴보았다(VRDsv의 F1값 - baseTR의 F1값). 대부분의 범주에 대해서 제안한 방법에 의해 성능이 향상되었음을 알 수 있다.

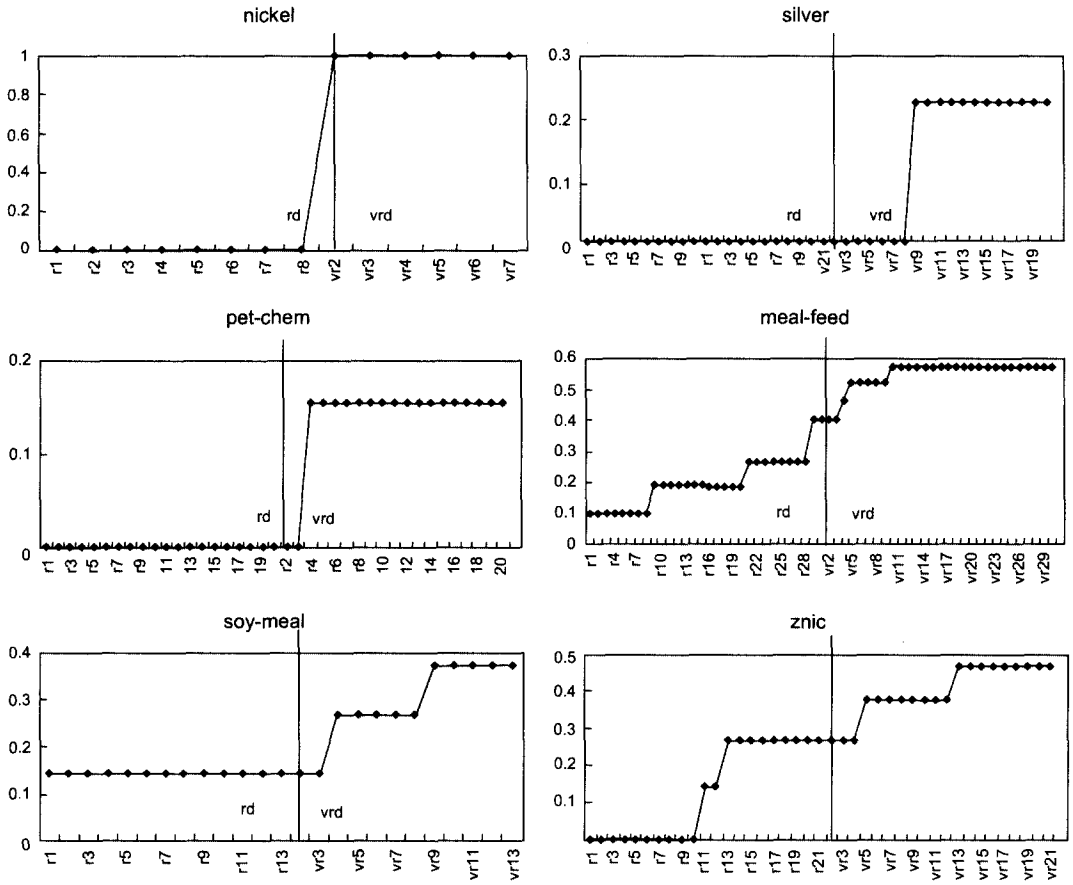
가상문서는 문서분류의 경계 결정을 위해서 이용되는 것으로, 문서분류기의 학습에서 가상적합문서의 효과를 자세히 분석하기 위해서, 각 범주에 대해서 적합문서의 개수를 점차적으로 증가시키면서 성능을 측정하고, 적합문서를 점진적으로 증가시키면서 생성된 가상적합문서에 대해서 성능을 측정했다. 적합문서의 개수를 증가시키면, 그에 따라 생성되는 가상문서의 수도 증가된다.

(그림 7)에서 x-축의 r1, r2, rN은 적합문서의 개수를 1개, 2개, N개에 대해서 각각 학습을 했을 때의 성능이고, vr2, vr3, vrN은 적합문서 2개, 3개, N개에 대해서 각각 생성된 가상문서집합과 모든 적합문서집합을 통합해서 학습했을 때의 성능이다.

학습에서 적합문서 또는 가상적합문서의 개수를 점점 증가시킬수록 성능이 향상되고 있음을 보인다. 특히, VRD는 학습문서집합에 적합문서의 개수가 작은 경우(예 : nickel, silver, pet-chem, soy-meal, zinc)에 대해서 상당히 높게 기여하고 있다. 가상적합문서를 추가할수록 성능이 꾸준히 증가되고 있는 것을 볼 수 있다.



(그림 6) Reuters-21578의 81개 범주에서의 baseTR에 대한 VRDsv의 성능 변화(F1)



(그림 7) 적합문서(rd)의 개수와 가상적합문서(vrd)의 개수를 점차적으로 증가시키면서 학습하고 테스트한 경우 성능 변화(F1 측정)

5. 결 론

본 연구에서 제안한 문서분류의 학습 단계에서의 가상적합문서를 주어진 학습문서와 통합하는 가상적합문서 기법은 좋은 효과를 보였다. 가상적합문서의 추가는 본 연구에서 이용한 분류 방법인 SVM의 범주 결정을 위한 정보인 지지벡터 집합을 변화시켰다. 제안한 방법은 TREC-11 필터링 테스트컬렉션에서 마이크로 평균 F1 측정에서 71%의 성능 향상을 보였고, Reuters-21578 집합의 67개의 범주에 대해서 11%의 성능 향상을 보였다. 이러한 성능 향상은 단순한 방법으로 생성된 가상적합문서가 시스템의 성능을 꾸준히 향상시키는데 기여하고 있다는 점에서 아주 중요하다. 결과 분석을 통해서, 학습집합의 제한 범위내에서 학습을 돕기 위해, 어떤 주제에 대한 담화 공간에서 기존의 문서의 단위에 한정되지 않고, 어휘 공간이 적절하게 조작될 수 있음을 알 수 있다.

향후 연구를 통해서, 문서분류를 위한 다른 학습 방법에 대해서도 가상적합문서 기법에 의해서 생성된 가상적합문서들을 적용하여 학습하는 방법을 찾는 연구가 필요하다.

참 고 문 헌

[1] Allan, J., Ballesteros, L., Callan, J., Croft, W. and Lu, Z., Recent experiments with INQUERY. In Proc. of the Fourth Text REtrieval Conference (TREC-4), 1996.

[2] DeCoste, D. and Schölkopf, B., Training invariant support vector machines. Machine Learning, 46(1), pp.161-190. 2002.

[3] Joachims, T., Making large-scale support vector machine learning practical. In Advances in Kernel Methods : Support Vector Machines (Schölkopf et al., 1999), MIT Press, 1999.

[4] Kawatani, T., Topic Difference Factor Extraction between Two Document Sets and its Application to Text Categorization. In International ACM-SIGIR Conference on Research and Development in Information Retrieval. 2002.

[5] Kwok, K. and Grunfeld, L., TREC-5 English and Chinese retrieval experiments using PIRCS. In the Proc. of the Fifth Text REtrieval Conference, 1997.

[6] Lewis, D., Reuters-21578 text categorization test collection distribution 1.0., http://www.daviddlewis.com/, 1999.

[7] Poggio, T. and Vetter, T., Recognition and structure from one 2D model view : observations on prototypes, object classes and symmetries. A.I. Memo No. 1347, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1992.

[8] Robertson, S. and Soboroff, I., The TREC 2002 Filtering Track Report. In Proc. of the Eleventh Text Retrieval Conference, 2002.

[9] Rocchio. J., Relevance feedback information retrieval. In Gerard Salton (ed.), The Smart retrieval system| experiments in automatic document processing, Prentice Hall, 1971.

[10] Rose, T. G., Stevenson, M. and Whitehead, M., The Reuters Corpus Volume 1 - from Yesterday's News to Tomorrow's Language Resources. In Proc. of the Third International Conference on Language Resources and Evaluation, 2002, http://about.reuters.com/researchandstandards/corpus.

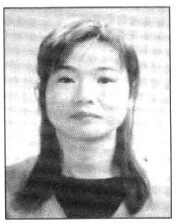
[11] Schölkopf, B., Burges, C. and Vapnik, V., Extracting support data for a given task. In Proc. of the First International Conference on Knowledge Discovery & Data Mining, Menlo Park, AAAI Press, 1995.

[12] Schölkopf, B. Support Vector Learning. R., Oldenbourg Verlag, Munchen. Doktorarbeit, TU Berlin, http://www.kernel-machines.org, 1997.

[13] Singhal, A., Mitra, M. and Buckley, C., Learning routing queries in a query zone. In Proc. of the 20th ACM SIGIR Conference on Research and Development in Information Retrieval, pp.21-29, 1997.

[14] Vapnick, V., The Nature of Statistical Learning Theory, Springer-Verlag, New York, 1995.

[15] Yang, Y. and Liu, X., A re-examination of text categorization methods. In Proc. of the 22nd ACM SIGIR Conference on Research and Development in Information Retrieval, 1999.



이 경 순

e-mail : selfsolee@chonbuk.ac.kr

1994년 계명대학교 컴퓨터공학과 학사

1997년 한국과학기술원 전자전산학 석사

2001년 한국과학기술원 전자전산학 박사

2001년~2003년 일본 국립정보학연구소 (National Institute of Informatics) 연구원

2004년~현재 전북대학교 전자정보공학부 전임강사

관심분야 : 정보검색, 지식 마이닝, 자연언어처리

안 동 언

e-mail : duan@chonbuk.ac.kr

1981년 한양대학교 전자공학과(공학사)

1987년 KAIST 전산학과(공학석사)

1995년 KAIST 전산학과(공학박사)

2001년~2002년 전북대학교 정보검색시스템 연구센터 센터장

1995년~현재 전북대학교 전자정보공학부 부교수

관심분야 : 정보검색, 한국어정보처리, 문서분류, 문서요약