문서분류에서 SVD(Singular Value Decompotion)기법에 기초한 효율적인 특징 선택방법 연구

리청화, 변동률, 박순철,
전북대학교 컴퓨터공학과
e-mail: bdrsea@chonbuk.ac.kr

An Efficient Selection Method for Document Classification Based On Singular Value Decompostion

Cheng-hua Li, Dong Ryul Byun, Soon Cheol Park
Dept of Computer Engineering, Chonbuk National University

요 약

본 논문에서는 문서분류를 위하여 SVD(Singular Value Decomposition)을 이용한 효율적인 특징 선택 방법을 제안한다. 분류기 알고리즘은 문서를 효과적으로 분류할 수 있지만 분류기에 입력되는 특징공간이 너무 크다는 단점이 있다. SVD를 이용하면 입력 데이터의 차원을 줄여줄 수 있으며 문서와 문서 사이의 관계성을 내포하는 벡터공간을 만들 수 있다. 따라서 SVD를 이용하면 문서분류의 시간과 효율을 동시에 증가시킬 수 있다. 본 논문에서는 실험을 통하여 SVD을 이용한 문서분류 시스템이 입력데이터에 대한 차원을 감소시키면서 훌륭한 분류 결과를 얻을 수 있음을 보여준다.

1. 서론

수만 건의 문서를 정확히 검색하여 사용자가 필요로 하는 문서를 제공할 필요성이 커지고 있다. 문서분류는 정보를 체계화하여 정보를 찾으려는 사용자에게 보다 정확한 문서를 검색할 수 있도록 한다. 문서 자동 분류는 문서의 특성을 분석하여 문서의 범주를 정한다. 분석을 위하여 문서의 특징을 잘 추출하여 범주와의 유사도를 측정하여 문서를 분류하게 된다[1].

본 논문에서는 분류하려는 문서를 알맞은 범주에 지정하기 위하여 이미 분류가 이루어진 문서들을 통하여 학습을 수행한다. 학습이 완료되면 분류기는 문서의 특징을 분석하여 해당하는 범주에 문서를 분류하게 된다. 학습과 분류을 위하여 본 논문에서는 벡터 공간 모델을 사용하였다. 문서의 특징을 벡터 공간으로 표현하고 분류 알고리즘을 통하여 학습을 수행한다. 학습이 완료되면 새로운 문서는 벡터로 입력되어 분류가 이루어진다. 문서 특징 벡터는 고차원의 벡터가 된다. 고차원의 벡터는 분류성능에 비하여 수행속도가 느리다는 단점을 가진다.

본 논문에서는 SVD(Singular Value Decomposition)을 이용하여 고차원의 벡터를 저차원의 벡터로 분해하여 분류기의 수행시간을 빠르게 하였다. SVD로 분해된 벡터는 문서의 의미는 가지고 있게 된다.

본 논문의 구성은 다음과 같다. 2장에서 SVD와 분류 알고리즘에 대하여 설명한다. 3장에서 실험결과를 보여주고, 4장에서 결론 및 향후 연구 방향에 대하여 설명한다.

2. SVD와 신경망 알고리즘

LSI(Latent Semantic Indexing)에서 사용된 SVD는 벡터 공간 모델에서 사용하여 색인어와 문서로 이루어진 고차원의 벡터를 저차원의 공간에 사상시키는 것이 가능하다. 이렇게 줄어든 공간은 문서의 의미 정보를 정확하게 가지고 있다. 매우 효율적으로 줄여준다[2].

본 논문에서 사용된 문서의 특징으로 TF 값을 사용하였다. TF(Term Frequency)는 문서내의 단어의 빈도수를 나타낸다. TF값을 이용하여 문서내의 단어의 가중치를 계산하여 문서의 특징을 가지는 문서 벡터를 생성한다. k번째 문서 벡터 d는 n개의 단어 가중치를 가진다.

$$d_k = \langle w_{k,1}, \ldots w_{k,i}, \ldots w_{k,n} \rangle \qquad (2.1)$$

식 (2.1)로 생성된 문서벡터는 k×n개의 행렬형태가 된다. 원본 벡터를 이용하여 분류기를 실행했을 때 가장 좋은 결과를 보이지만 분류 속도는 현저히 떨어지게 된다. SVD를 이용하면 원본벡터의 의미는 유지하면서 행렬의 크기는 줄일 수 있어 분류성능과 속도 모두를 만족할 수 있다. 원본 행렬을 A는 식 (2.2)에 의하여 $U\sum V^T$ 로 분해할 수 있다[2,3,4].

$$A = U\sum V^T \qquad (2.2)$$

U는 단어 간의 관계를 나타내는 벡터이며 V는 문서간의 관계를 나타내는 벡터가 된다. 본 논문에서는 문서벡터에 단어 간의 의미정보를 가지고 있는 U벡터를 곱하여 입

력 가중치 벡터 $\hat{d}$를 k개의 차원으로 줄여서 사용한다.

$$\hat{d} = d^T U_k \qquad (2.3)$$

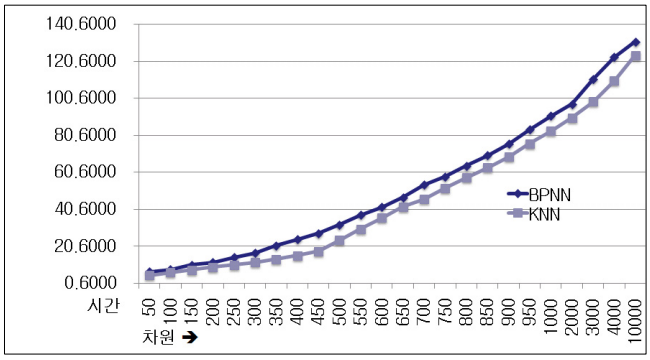
원본벡터보다 $\hat{d}$ 벡터를 이용하면 보다 적은 차원의 입력값을 이용하여 같은 성능의 분류기를 만들 수 있다. $\hat{d}$를 만드는 데는 감소된 차원 k값을 선정하는 방법이 중요하다. k값을 선정하는 방법은 Kruskal이 사용한 SSTRESS 측정법을 사용하여 선정하였다[6].

문서 분류를 위하여 KNN과 신경망 알고리즘을 사용하였다. 신경망 알고리즘은 인간의 신경조직을 응용한 알고리즘이다. 신경조직의 기본 구성요소는 뉴런의 생물학적 특성을 수학적으로 표현한 것이 신경망 처리소자인 퍼셉트론이다. 신경망은 다수의 퍼셉트론으로 구성되어진다. 하나의 퍼셉트론은 입력되는 값에 가중치를 부여하고, 활성함수를 이용하여 출력값을 조절한다. 분류에 대한 해답을 찾을 때까지 학습을 수행한다. 해답을 찾아가는 과정에서 가중치와 활성함수를 조정하여 분류 시스템을 완성한다. 신경망 알고리즘과 비교를 위하여 본 논문에서는 문서분류에 사용되는 대표적인 알고리즘인 KNN을 이용한다. KNN 알고리즘은 분류할 문서 X와 범주별 표본문서 Y사이의 거리를 통하여 문서를 분류한다[5].

3. 실험결과

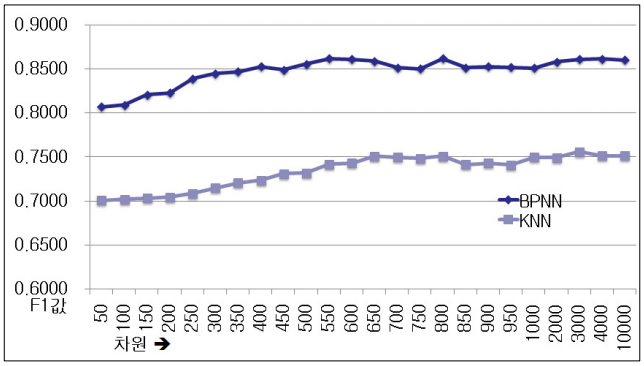
분류시스템의 학습과 분류실험을 위하여 사용된 데이터 셋은 '한국일보-20000'(HKIB-20000)이다. 한국일보-20000 실험문서집합은 한국일보-40075 집합의 기사 중 20,000건을 별도로 추출하여 분류체계를 보다 현실적으로 수정하였으며, 3단계 분류체계의 모든 노드에 기사를 할당하여 구축한 계층적 분류체계의 문서범주화용 실험문서집합이다.

본 논문에서는 10개의 범주를 선택한다. 학습문서와 실험문서의 개수를 달리하여 10차례 이상의 실험을 통하여 평균 F1 값을 측정하였다.



(그림 1) 문서를 분류하는데 필요한 시간

그림 1는 500개의 문서를 차원에 따라서 분류하는데 필요한 시간이다. 500개의 문서는 총 10704개의 단어로 이루어져 있다. 400개의 차원과 1000개의 차원은 약 6배의 차이가 나는 것을 볼 수 있다.



(그림 2) 문서 분류기의 분류 성능

그림 2는 500개의 문서를 통하여 학습을 수행하고 500개의 문서를 분류기를 통하여 분류를 했을 때 KNN과 BPNN 알고리즘의 성능 결과를 보여준다. 400개의 차원과 100개의 차원의 분류 성능이 거의 유사하다는 것을 볼 수 있다.

4. 결론

문서분류기에서 사용되는 벡터 공간 모델은 고차원의 공간을 사용하게 되어 문서가 많아질수록 시간이 기하급수적으로 증가한다. SVD를 이용하면 문서의 의미는 유지하면서 벡터의 차원을 줄일 수 있으므로 그림 1과 그림 2에서 보여지는 바와 같이 짧은 시간에 비슷한 성능의 분류를 할 수 있게 된다. 향후 연구방향은 벡터 공간 모델을 사용하는 분류 알고리즘은 다양하다. SVD를 여러 가지 알고리즘에 결과를 비교해볼 필요성이 있다.

참고문헌

[1] Thorsten Joachims, Text categorization with vector machines : Learning with many relevant features, In International Conference on Machine Learning(ICML), 1998

[2] Richardo Baeza-Yates, Berthier Ribeiro-Neto, Modern Information Retrieval. Addison Wesley, 1999.

[3] David A. Grossman, Ophir Frieder, Information Retrieval Algorithms And Heuristics. Kluwer Academic Publishers, 1998.

[4] Robert R. Korfhage, Information Storage And Retrieval. John Wiley & Sons, Inc, 1997.

[5] 오일석, 패턴인식. 교보문고, 2008.

[6] Johnson, Richard Arnold, Wichern, Dean W. Applied multivariate statistical analysis, A Prentice-Hall, 1998