SVM 기반 기술정보 문서분류를 위한 특징 선택 기법

강윤희*
*천안대학교 정보통신학부
e-mail:yhkang@infocom.cheonan.ac.kr

Feature Selection for Document Classifier for IT documents based on SVM

Yun-Hee Kang*
*Division of Information & Communication, Cheonan University

요약

인터넷상의 정보의 급증에 따라 필요한 정보를 발견하고 관련된 정보를 조직화하기가 더욱 어려워지고 있으며 정보 접근의 부하를 줄이기 위한 효율적인 문서 분류의 중요성 및 필요성이 증가하고 있다. 본 논문에서는 디렉토리 내의 학습 문서 집합을 기반으로 구성된 디렉토리 내의 대표 용어 집합으로 구성된 모델을 학습 및 분류하기 위해 SVM을 사용한다. 문서분류를 위해 정보통신 웹 디렉토리 내의 문서로부터 추출된 용어 집합을 기반으로 학습을 수행한 후 문서 분류를 수행한다. 또한 TFiDF를 기반으로 특징을 표현하기 위해 벡터공간 모델을 사용하였고 이를 기반으로 성능 평가를 수행한다.

1. 서론

최근 인터넷 급속한 성장과 보급에 따라 전자우편과 웹을 통해 제공되어지는 정보의 양은 기하급수적으로 증가하고 있다. 그러나 웹을 통해 접근 가능한 정보에서 실제 사용자에게 필요한 정보는 극히 일부분이며 이러한 현상을 정보과부하(information overload) 라고 한다. 정보과부하 문제의 해결을 위해서는 인터넷 상의 유용한 정보를 접근하기 위해 필요한 정보를 분류하는 작업이 필수적이다[1,5,6].

텍스트 분류에 대한 연구는 신경망과 통계적 접근으로 이루어지고 있다. E.D Wiener 는 텍스트 분류에 역전파(back propagation)를 적용하였으며 샘플 텍스트로부터 추출되어진 벡터의 차원은 최소 200 이다[9]. Lewis와 Schapire 는 학습가능 선형 모델인 perceptron과 EM 알고리즘을 텍스트 분류에 적용하였다[10]. Joachims는 SVM(Support Vector Machine)의 응용으로 차원의 문제를 완화하기 위해 텍스트 분류에 적용하였다[6]. Yang과 Liu는 k-NN, SVM, 역전파와 LLSF(Linear Least Square Fit)의 통계적 기준을 사용하는 텍스트 분류의 성능을 비교하였다. Yang은 K-NN, Linear Least Square Fit와 WORD를 성능 측면에서 분석하였다[5].

본 논문에서는 감독자 기반 문서분류 기법인 SVM을 이용한 기술문서의 분류를 위해 확장된 특징 추출 기법을 사용한다. 이를 위해 정보 통신 분야 11개 대분류 디렉토리를 문서의 카테고리로 간주하며 기술 문서의 분류 작업을 적용한다.

본 논문의 2장에서는 관련연구를 기술하며 3 장에서는 특징 선택기법을 4장에서는 문서 분류기의 설계 및 구현을 기술하고 실험 결과 및 평과를 보인다. 5장에서는 결론 및 향후 연구에 대해 기술한다.

2. 관련 연구

문서 분류를 위한 특징 추출 기법으로 감독 학습 기법에 대한 색인어 추출 기법에는 DF(Document Frequency), IG(Information Gain), MI(Mutual Information), χ^2(chi Square), TS(Term Strength), LSI(Latent Semantic Indexing) 등이 있다[5]. DF 기법은 단어가 출현한 문서의 절대 빈도수만을 고려하는 기법으로 단순하고, 분류 성능도 비교적 우수하지만 출현 빈도만으로 문서의 카테고리를 비교할

수 없는 경우도 많이 발생할 수 있다. IG 기법은 정보 검색에서 주로 사용되는 색인어 추출 기법으로 카테고리 별 단어의 평균 빈도수를 고려하는 방법이다. χ^2 기법은 우연성 테이블을 이용하여 단어와 카테고리의 독립성을 고려하는 기법으로 텍스트 분류에서 비교적 높은 정확도를 나타낸다. 그러나 저빈도 단어들에 대해서는 고려하기 어려운 단점을 가지고 있다. MI 기법은 텍스트 분류에서 많이 사용되는 색인어 추출 기법으로 단어와 카테고리 간의 독립성을 고려하여 카테고리별 대표 단어를 추출한다. TS 기법은 코사인 계수와 같은 식에 의한 유사도 계산을 통해 문서들을 사전에 클러스터링 한 후, 유사한 문서 쌍 내에서 출현 확률이 높은 단어만을 대표 색인어로 추출하는 기법이다.

본 연구에서는 SVM의 입력을 위해 [5]의 실험을 기반으로 DF 기법을 사용하여 특징을 선택한다. DF는 문서내의 용어의 출현 빈도를 기반으로 문서내의 유일한 단어에 대한 발생빈도를 계산한다.

3. 문서 분류를 위한 제안된 특징 선택 기법

본 절에서는 문서내의 주요한 키워드인 특징의 선택 방법을 기술한다. 특징 선택은 문서 분류의 학습을 위한 전처리 과정으로 문서 학습을 위한 입력 벡터 생성을 수행한다.

3.1 벡터공간 모델

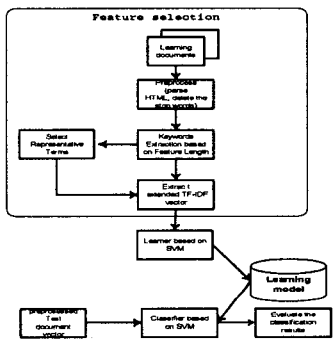
벡터공간 모델에서 질의와 각 문서는 용어 공간 내의 벡터로서 표현한다. 벡터 $w_{ij} \geq 0$를 (ki, dj)쌍의 가중치라고 하며, 이 가중치는 문서의 의미적 내용을 설명하기 위한 색인어의 중요도를 정량화 한다. 시스템 내의 색인어 수를 t라 하고 ki를 색인어라 하면 모든 색인어 집합 K={k_1, k_2,... ,k_t}이다. 문서 dj에서의 색인어 k_i의 가중치는 $w_{ij} \geq 0$이고 따라서, 문서 내에 한번도 출현하지 않은 색인어의 가중치는 0 이 된다. 문서 $\overrightarrow{d_j}$ 는 색인어 벡터 $\overrightarrow{d_j}$ = <w_{1j}, w_{2j},..., w_{tj}>로 표현된다. 문서 내의 용어에 대한 가중치 벡터($\overrightarrow{d_j}$)의 계산은 용어 빈도(Term Frequency)와 역문서 빈도수(Inverse Document Frequency)로서 정의한다[2].

3.2 특징 선택 기법

본 논문에서는 분류기 구성을 위해 웹으로부터 추출된 데이터들을 이용하여 전처리(Preprocessing) 과정을 수행한다. 전처리 과정에서는 문서로부터 추출된 용어 중 불필요한 용어 제거 및 빈도수에 따라 학습을 위한 입력 특징벡터를 구성한다.

학습기는 특징 값이 결정된 학습 데이터(training data)와 목적값(desired value)의 쌍을 이용하여 각 카테고리의 학습데이터로 SVM을 사용하여 학습한 후 새로운 데이터들을 분류하기 위한 모델을 생성한다. 분류기는 생성된 학습 모델을 기반으로 반입된 새로운 문서들을 분류한다.

본 논문에서는 전체 대상문서에서 특징선택(Feature Selection) 과정을 거친 후 만들어진 입력 특징 벡터를 각 카테고리의 SVM에 기반 학습기에 입력하여 모델을 구성한다. 또한 분류기를 사용하여 테스트 문서가 임계값 이상의 값을 출력하는 카테고리는 이 문서를 포함하는 것으로 결정한다. (그림 1)은 문서 분류 흐름을 보인 것이다.



(그림 1) 문서 분류 흐름도

(그림 1)의 문서 흐름에서 학습을 위한 문서는 HTML 태그 제거 및 불용어 처리의 전처리 후 빈도수에 따른 선택 과정을 통해 확장된 TF-IDF 벡터를 구성한다. TF-IDF 벡터는 SVM 학습기의 입력으로 제공되며, 학습기는 분류를 위한 모델를 생성한다. 분류를 위한 문서는 학습과 동일한 전처리 과정을 통해 입력 벡터를 생성한 후 구성된 학습 모델을 기반으로 분류가 이루어진다.

4. 실험 및 평가

4.1 실험 환경

본 실험에서는 불필요한 특징을 제거하기 위한 특징 선택 기준으로 "불용어 및 길이가 2이하인 키워드를 제거한 후 전체 키워드에서 2000개를 선택한 후 가중치를 구하고 없는 경우 0의 값을 갖는다"을 사용한다.

본 실험을 위해 SVM은 SVM Litght[7]를 사용

하여 11개의 분류기를 구성한다. 개별 SVM은 학습 문서를 사용하여 문서 분류를 위한 모델을 구성한다. 모델 구성을 위한 학습은 긍정 문서 집합과 부정 문서 집합 모두를 사용하여 수행하며 다수의 클래스의 문서 분류에 적용할 수 있도록 모델링 한다. 본 실험에서는 1) 신규 카테고리 추가시의 문서 분류의 성능 비교, 2) 특징 벡터의 크기 구성에 따른 문서 분류의 성능 비교의 2가지 평가 기준에 따라 실험 시나리오를 작성한 후 실험을 수행한 후 결과를 평가한다.

4.2 입력 벡터 구성

본 실험에서는 문서 카테고리내의 출현 용어와 클래스간의 연관성을 고려하여 <규칙 1>에 따라 클래스 대표용어를 추출하였다.

규칙 1: 클래스 대표용어 추출을 위해 class document frequency/total document frequency 의 값이 임계값(0.25)를 넘는 키워드를 모아 클래스 별 대표용어 리스트를 구성한다.

문서 학습을 위해 트레이닝 데이터를 수동으로 추출하여 이용하였으며 <규칙 2>을 사용하여 학습 벡터를 구성한다. 학습 벡터의 차원 학습에서의 벡터 크기의 연관성 검토가 필요하며 500, 1000, 2000으로 설정하였다. 수집된 문서에 대한 색인 구축은 문서 내에 발생된 용어의 빈도 수와 카테고리 대표용어를 기반으로 하여 가중치를 설정한다. 이를 위해 기존의 TF*IDF 수식을 규칙 2를 적용하여 개선하였다.

규칙 2:

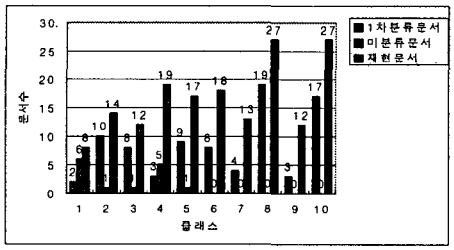
1. 발생 빈도가 높은 키워드부터 정렬하여 t 차원 벡터를 구성한다.
2. 각 학습 문서로부터 학습 벡터의 키워드들의 빈도수를 구해서 tf * idf 값으로 벡터를 구성한다.
3. 키워드가 해당하는 카테고리 키워드리스트에 있으면 df * tf 값에 1보다 큰 값(1.2)을 곱하고 다른 클래스의 키워드 리스트에 있으면 1보다 작은 값(0.8)을 곱한다.

학습을 위한 문서 벡터는 테스트 데이터를 이용하여 학습벡터 구성 단계에서 구한 벡터값과 테스트 데이터에서 얻은 빈도수를 곱하여 학습 벡터를 구성한다. 본 실험에서는 SVM 학습기의 입력 문서벡터 집합 구성을 위해 <규칙 3>를 적용하였다.

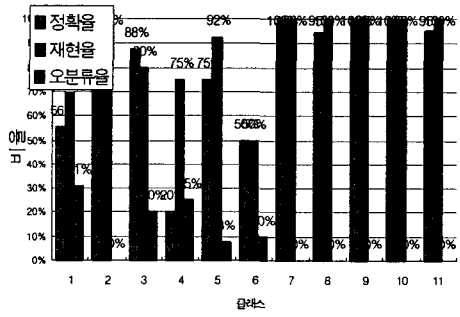
규칙 3: 학습 문서에 대해 특징벡터를 만들어 각 카테고리 별로 두개의 파일 정보 집합인 feature.pos와 feature.neg로 구성한다. 구성된 각 클래스의 긍정 학습 특징 벡터와 부정 특징 벡터를 통합하여 개별 클래스에 대한 종합 특징 벡터를 클래스 수만큼 구성한다.

4.3 실험 결과 및 평가

본 실험은 신규 카테고리 추가에 대한 문서 분류의 성능 평가와 특징 벡터의 크기에 따른 성능 비교를 수행하였다. 1차 평가는 전체 테스트 문서의 수는 126개이고 특징벡터의 크기는 500으로 한정한 후 수행하였다. (그림 2)와 같이 실험 결과는 8,9,10번 카테고리의 경우 100%의 정확율을 얻었으며 낮은 분류 정확율 갖는 경우는 1, 4번 카테고리이다. 특정 카테고리에 대한 분류 정확율이 70 % 미만인 것들이 10개중 4개이다. 특히, 유선통신, 인터넷의 1,4 클래스의 경우 클래스 연관도가 높은 이유로 낮은 성능을 보였다. 또한 학습문서에서의 정제에 의해 3개의 클래스의 경우 재현율이 90% 이상을 보인다.



(그림 2) 정보 통신 문서 분류 성능 비교

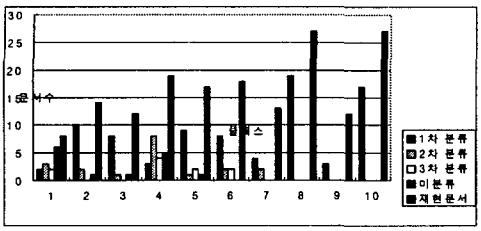


(그림 3) 문서 분류 성능 비교

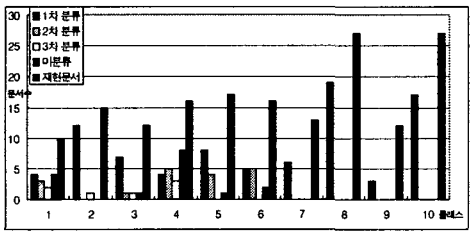
(그림 3)는 전체 테스트 문서의 수는 303개이고 기존의 10개의 클래스와 무관한 신규의 클래스를 추가하였을 때 효과적인 문서 분류가 이루어지는 것을 보인다. 실험 결과 85.7%의 정확율과 78.8%의 높은 재현율 및 0.845의 F0.5 값을 얻었다.

(그림 4)와 (그림 5)은 126개의 전체 테스트 문서

에 대해 특징벡터 크기에 따른 성능을 보인 것으로 (그림 5)이 특징 벡터의 크기가 2000 증가됨에 따라 전체적으로 정확율이 증가되었으며, 1번 클래스는 29%에서 44%로 20%에서 33%으로 정확율이 높아짐을 보였다.



(그림 4) 특징 벡터의 크기(500)



(그림 5) 특징 벡터의 크기(2000)

5. 결론

본 논문에서는 기술정보 분야의 웹 문서의 텍스트 자동 분류 시스템의 설계 및 구현을 기술하였다. 학습 문서 벡터는 웹 카테고리 내의 문서로부터 추출된 용어 및 관련 문서를 기반으로 구성하였으며 학습 문서 구성 후 SVM 학습기를 통해 모델을 구성하였고 문서 분류를 수행한다.

본 실험의 문서는 정보통신 분야 디렉터리 서비스 시스템인 itfind로부터 수집된 문서를 대상으로 하였으며 2가지 시나리오에 따라 수행하여 각 시나리오 별로 재현율/정확율 및 미분류율을 성능 요소로 계산하였다. 본 실험을 통해 학습 벡터 구성과정에서 신규 카테고리 추가에 의해 다른 카테고리의 문서 분류에 미치는 영향을 평가하여 SVM을 기반으로 한 문서 분류 기법이 우수함을 보였다.

향후 클래스간의 유관성에 따른 분류성능의 영향에 대한 정량적 분석을 통해 SVM의 분류 결과 조정을 수행할 예정이며 전체 시스템에 대한 자동 학습 기능을 위해 사용자 피드백을 통해 문서 분류 시스템의 성능 향상을 수행할 예정이다.

참고문헌

[1] Tak W.Yan, Hector Garcia-Molina, "Sift - A Tool for Wide-Area Information Dissemination," In Proceedings of the 1995 USENIX Technical Conference, pp. 177-186, 1995.

[2] Salton, G. Automatic Text Processing: The Transformation, Analysis and Retrieval of Information by Computer, Addison-Wesley, 1989

[3] Chapelle, O., Haffner, P. and Vapnik, V., "SVM for histogram-based image classification," IEEE Trans. on Neural Networks, 10(5), pp.1055-1065,1999.

[4] T. Doszkocs, J. Reggia, and X. Lin. "Connectionist models and information retrieval," Annual Review of Information Science & Technology 25:209-260, 1990.

[5] Yang Y., J.O. Pdedersen, "A Comparative Study on Feature Selection in Text Categorization," Proc. Of the 14th Internatinal Conference on Machine Learning ICML-97, pp.412-429, 1997.

[6] Joachims, T., "Text categorization with support vector machines: Learning with many relevant features," Proc. European Conference on Machine Learning (ECML), pp. 137-142,1998

[7] Joachims, T., SVMLight, http://ais.gmd.de/~thorsten/svm_light, 1998.

[8] E. Wiener, J. O. Pedersen and A. S. Weigend, A neural network approach to topic spotting, Proc. SDAIR '95, pp. 317-332, Las Vegas, NV, 1995.

[9] D. Lewis, W. A. Gale, A sequential algorithm for training text classifiers, Proc. SIGIR '94, pp. 3-12. Dublin, Ireland. 1994.

[10] D. Lewis, R. Schapire, J. Callan, and R. Papka, "Training Algorithms for Linear Text Classifiers," Proceedings of ACM SIGIR, pp.298-306, 1996.