SVM을 이용한 악성 댓글 판별 시스템의 설계 및 구현

김묘실[0], 강승식
국민대학교 교육대학원 전자계산교육 전공[0], 국민대학교 컴퓨터공학부
forjudy@naver.com[0], sskang@kookmin.ac.kr

A Design and Implementation of Malicious Web Log Identification System by Using SVM

Myo-Sil Kim[0], Seung-Shik Kang
Graduate School of Education[0], School of Computer Science, Kookmin University

요 약

댓글은 온라인 상에서 자신의 의견을 달고 다른 사람의 의견을 공유함으로써 필요한 정보를 쉽고 빠르게 얻을 수 있다. 본 논문에서는 익명성을 이용해서 특정인을 근거 없이 비방하거나 명예를 훼손하는 악성 댓글을 판단하는 시스템을 구현한다. 자질의 추출 방법을 여러 가지로 실험하여 동사, 형용사 등을 추가했을 때 자질의 출현빈도를 이용한 가중치를 계산하고, 용어 벡터로 표현된 입력 문서를 이진 분류기(Binary Classifier)인 SVM^{light}을 이용하여 악성 댓글인지를 판단하는 시스템을 구현하고 그 성능을 평가한다.

1. 서론

정보화 시대의 대량으로 생겨나는 문서들에 대해서 익명으로 자신의 의견을 달 수 있는 댓글은 상호작용적 의견 개진이라는 취지로 시작되었다. 그러나, 포털 사이트 뉴스는 악성 댓글로 인해 심각한 명예훼손 및 개인 정보 침해의 위험에 노출되어 있다.

일반적으로 문서 범주화에 관한 연구는 효과적인 범주화 모델의 계산 방법과 학습 자질의 추출 방법이라는 두 가지 문제를 중심으로 발전되어 왔다. 문서 범주화 모델 중 기계 학습을 이용한 신경망(Neural Network), SVM (Support Vector Machine) 등 다양한 방법을 이용하여 구성하고 있고, 각각의 방법은 학습 정도에 따라 높은 정확도를 가진 분류기로 구성되어 활용되고 있다.

최근에는 학습에 대한 빠른 처리 및 대용량 데이터처리를 위한 성능이 높다고 평가되는 SVM을 활용한 방법이 많이 사용되고 있다.[12]

문서의 자질 추출 방법은 학습 문서에서 형태소 분석기를 이용하여 해당 자질에 대한 가중치를 계산하는 것이 일반적이다. 문서를 표현하기 위해서는 자질에 대한 출현빈도(Term Frequency: TF)를 이용하여 하나의 문서를 표현하는 방법과 역 문헌빈도(Inverse Document Frequency: IDF)를 같이 이용하여 가중치(Weighting)를 표현하는 방법으로 구분한다.

본 논문에서는 TF-IDF 가중치를 이용한 벡터 수치화 방법을 선택하였고, 두 가지 범주로 문서를 분류하는 SVM^{light}을 사용하였다.

본 논문의 구성은 다음과 같다. 2장에서는 문서처리와 SVM^{light}에 대한 관련 연구에 대해 살펴보고, 3장에서는 악성 댓글을 판단하는 시스템 구조와 자질 선택 방법에 대한 내용을 설명한다. 그리고 4장에서는 실험 및 평가를 하고, 5장에서는 결론 및 향후 연구과제에 대해서 기술한다.

2. 관련 연구

인터넷 문서 특히 웹로그(블로그)의 주관적이고 감정적인 표현들을 SVM 같은 기계 학습 알고리즘에 의해 학습하고 새로운 문서를 자동적으로 긍정 또는

[1] 본 연구는 첨단정보기술 연구센터를 통하여 과학재단의 지원을 받았음.

부정으로 분류하는 연구들이 활발히 진행되고 있다.[7] 영화 비평, 상품 사용기, 고객 피드백과 같은 특정 영역별로 다르게 사용되는 감정적 표현들을 인지하고 그것을 범주화 한다.[1,3,4,8,9,10]

본 장에서는 본 논문에서 제안한 댓글 문서의 악성 여부를 판단하는 분류 시스템을 구현하기 위한 문서 처리 방법과 SVM^{light}에 대해 설명한다.[2]

2.1. 문서처리 및 자질 추출

문서처리는 학습(Learning) 과정과 분류 과정에서 분류기가 인식할 수 있는 데이터로 변환하기 위한 처리 과정이다. 이러한 과정은 기계 학습의 처리가 가능하도록 문서의 내용이나 특징을 잘 반영하는 자질을 추출하여 수치화를 한다. 이는 벡터 모델에서 사용이 가능한 벡터로서 표현이 가능하고, 벡터는 자질의 중요도에 따라서 가중치 부여하는 방법으로 표현한다.

문서처리는 <그림 1>과 같은 단계로 처리한다.

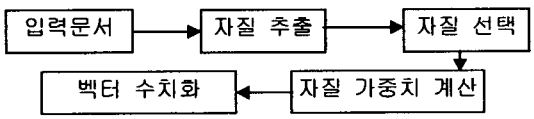

그림 1. 문서 처리 단계

2.2. SVM^{light}

SVM은 2개의 범주로 분류하는 이진 분류기이다. SVM은 기계학습 알고리즘으로써 분류 프로그램에 응용되어 높은 성능을 보여주고 있다. SVM^{light}는 C언어로 Vapnik의 SVM 알고리즘을 구현한 분류기 프로그램이다. 본 논문에서는 SVM^{light}을 이용하여 분류기를 수행하였으며 분류에서 입력되는 데이터는 SVM^{light}에서 호환하는 형식에 따른다. SVM^{light}는 svm_learn 프로세스와 svm_classify 프로세스로 구성되어 있다. 프로세스의 데이터 흐름은 <그림 2>와 같다.

학습 데이터는 긍정적 학습 데이터와 부정적 학습 데이터를 구분하여 SVM 학습기에서 학습이 가능하도록 파일 형태로 생성하여 처리한다. 범주 별로 학습 모델을 별도로 생성하며 분류 서비스에서 제공하는 처리 방법에 따라 학습 모델을 적용한다.

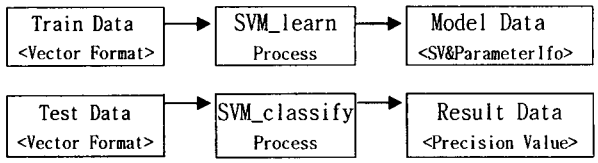

그림 2. SVM^{light} 데이터 구성도

3. 악성 댓글을 판단하는 시스템의 설계 및 구현

3.1. 시스템의 구조

악성 댓글을 판단하는 시스템은 크게 두 가지 부분으로 나눌 수 있다. 범주화 된 학습 문서와 입력 문서에 대해 자질들을 선별하고 가중치를 계산하는 문서 처리 부분과 SVM^{light}에서 문서 처리된 데이터를 학습하고 입력 문서를 악성 또는 Non악성으로 분류하는 과정이다. 학습하기 위해 수집된 문서와 분류하기 위해 입력 받은 문서는 공용으로 같은 문서 처리 모듈을 이용한다.

자질 추출은 형태소 분석기(KTL)를 이용하여 자질을 추출한다.[11] 자질 선택은 형태소 분석기로부터 얻은 품사 정보를 이용하여, 명사, 동사, 형용사 등을 조합하여 선택한다. 선택된 자질들은 오름차순으로 자질 정보(Word List)를 구성한다. 가중치 계산은 문서 분류 분야에서 보편적으로 사용되는 TF-IDF 방식을 적용하였다.

3.2. 자질 선택

형태소 분석기에 의해 추출된 자질들은 각각 품사 정보를 가지고 있다. 각 자질이 가지고 있는 품사 정보를 이용하여 자질 데이터를 구성한다. 이렇게 구성된 자질 데이터를 실험하고 그 성능을 비교함으로써 특정 품사와 악성 표현과의 관계를 분석한다.

표 1. 자질 선택 데이터 구성

구분	내용
K_0	명사만 추출
K_1	명사, 형용사, 동사 추출
K_4	모든 품사의 통합 추출
K_S0	어절 자체와 명사만 추출
K_S1	어절 자체와 명사, 형용사, 동사 추출
K_S4	어절 자체와 모든 품사의 통합 추출

3.3. TF-IDF 가중치

TF-IDF 가중치는 문서의 자질에 가중치를 부여하여 문서를 표현하는 방법이다. 문서에서 나타나는 자질의 빈도수 TF와 역문헌빈도 IDF의 곱으로 표현하며 아래 수식과 같다.

$$a_{ik} = f_{ik} \times \log\left(\frac{N}{n_k}\right)$$

where

$_{ik}$: I문서내 k단어 빈도수

$_{k}$: 전체문서 중 k단어가 출현한 문서 수

본 논문에서 사용한 TF-IDF 가중치 계산은 문서 내 자질이 나타난 빈도수와 역 문헌빈도수만을 고려하여 수식에서 적용하였다. <그림 4>의 가중치 계산 알고리즘에서 구조체 배열 ReplyData는 각 댓글에 대한 배열이다. ReplyData는 각 댓글에 대해 자질의 개수(ndata), 자질(word), 자질 고유 번호(word_index), 문장 빈도(word_df), 자질 빈도(word_tf), 가중치(word_score) 등의 정보를 가지고 있는 구조체이다.

```
Algorithm word_weighting(ReplyData, max_reply)
{
int cnt_reply=0, j=0;
char *filename = "word_cnt";/*자질 정보 사전(WordList)*/
load_worddata(filename);

/*각 댓글에 대해 자질들이 저장되어 있는 파일*/
fpin = fopen("replyData","r");
fpout = fopen("train","w");/*학습파일생성*/

/*catagory : 악플내지 Non악플 구분자*/
/*n : 각 댓글이 포함하고 있는 자질의 갯수*/
for(int  i=0; i<max_reply-1;i++) {

  for (j=0; j<ReplyData[i].ndata;j++) {
    ReplyData[i].word_index[j] =
        bsearch_string(ReplyData[i].word[j], WordData.data,
0, WordData.ndata)+1;

    /*자질(word)에 대한 WordList의 DF을 얻는다.
    DF는 WordList생성시 자질과 함께 저장되었다.*/
    ReplyData[i].word_df[j]=
      atoi(WordData.data[ReplyData[i].word_index[j]]+
      strlen(WordData.data[ReplyData[i].word_index[j]])+1);

    ReplyData[i].word_score[j]=1*log10(max_reply/word_df[j]);
}//end of for
/*자질들을 오름차순 정렬 및 중복 자질 제거하면서 TF 계산 */
sortWord();
```

```
for(j=0; j<ReplyData[i].ndata; j++) {
  fprintf(fpout, " %d:%3.10f",
    ReplyData[i].word_index[j],
    ReplyData[i].word_score[j] *
    ReplyData[i].word_tf[j]);
}//end of for
}//end of for
}//end of Algorithm
```

그림 4. 가중치 계산 알고리즘

TF-IDF 가중치 계산이 처리된 데이터는 <자질:값> 형태로 표현되어 학습이 가능하도록 파일 형태로 생성하여 처리한다. SVM^{light}을 이용하기 위해서는 <그림 5>와 같이 벡터 데이터를 표현하는 표현법을 따른다.

```
<line> .=. <target> <feature>:<value>
<feature>:<value> ... <feature>:<value> # <info>
<target> .=. +1 | -1 | 0 | <float>
<feature> .=. <integer> | "qid"
<value> .=. <float>
<info> .=. <string>
```

그림 5. 벡터 데이터 표현법

본 논문에서는 한 line이 댓글 문서 1개와 동일한 의미를 가진다. 또한 앞서 자질 선택 후 자질 정보 사전(WordList)의 index가 feature에 해당하며, 자질별 계산된 TF-IDF가 value에 해당된다. target의 경우 문서 분류의 결과값인 True/False의 개념과 일치시켜 Positive의 경우 +1, Negative의 경우 -1로 구성이 가능하다. <info>의 경우 설명 정보이므로 생략도 가능하다. 본 논문에서는 Positive는 Non악성 댓글을 의미하며, Negative는 악성 댓글을 의미한다.

4. 실험 및 성능 평가

본 논문은 <표 1>과 같이 6가지의 경우에 대한 악성 댓글을 판단하는 시스템의 성능을 평가하였다.

4.1. 실험 데이터

트레이닝 및 테스트에 사용된 데이터들은 실제 포탈사이트의 뉴스에 대한 댓글들을 수집하여 구성하였다. 데이터의 구성은 <표 2>와 같다.

실험을 하기 위해 총 2,200개의 댓글을 사용하였고, 이에 학습 문서로 1,980개, 테스트 문서로 220개를

이용하였다.

표 2. 데이터 셋의 구성

	Training Data	Test Data
악성 댓글	990	110
Non악성 댓글	990	110
합 계	1,980	220

학습의 정확도를 높이기 위해 악성 댓글과 Non악성 댓글의 비율을 같게 하였다. 문서 분류의 성능을 평가하기 위한 기준은 정확률, 재현률 또는 F_1-measure 측정식이 사용된다. 아래 수식에서 P는 정확률, R은 재현율이다. SVM^{light} 분류기에서 반환하는 정확률과 재현율을 그대로 사용하였다.

$$P(presicion)=\frac{\text{Non악플로 분류된 실제 Non악플 수}}{\text{Non악플로 분류된 수}}$$

$$R(recall) = \frac{\text{Non악플로 분류된 실제 Non악플 수}}{\text{전체 Non악플 수}}$$

$$F_1\text{-measure} = \frac{2PR}{P+R}$$

4.2. 실험 결과

본 논문에서는 자질 추출 후 자질 선택 방법을 6가지로 구성하여 실험하였다. <표 3>은 6가지의 자질 선택 방법에 따라 문서 처리 과정을 마친 2,200개의 댓글 문서를 10-fold cross validation 방식으로 실험한 평균값이다. Accuracy은 전체 댓글 수 중 분류기가 맞게 분류한 개수를 의미한다.

F_1-measure 측정 결과에서 자질 선택시 명사만(K_0) 선택한 것은 같은 조건의 다른 품사를(K_1, K_4) 포함하는 것보다 1.1%정도 성능이 높게 나왔고, 공백을 기준으로 얻은 단어, 어절 자체를(K_S0) 포함했을 때는 어절 자체를 포함하지 않은 데이터보다(K_0) 1.5% 향상된 성능을 얻을 수 있었다.

표 3. 자질 선택 방법에 따른 성능 비교

구분	accuracy	precision	recall	F_1-measure
K_0	0.6610	0.6331	0.7840	0.7005
K_1	0.6580	0.6377	0.7460	0.6876
K_4	0.6575	0.6399	0.7400	0.6863
K_S0	0.6805	0.6552	0.7870	0.7151
K_S1	0.6690	0.6452	0.7710	0.7025
K_S4	0.6590	0.6415	0.7410	0.6877

어절 자체를 포함한 명사(K_S0) 실험이 성능면에서 향상된 결과를 얻을 수 있었던 이유는 댓글에는 변칙적으로 변형된 단어들을 네티즌들이 많이 사용하고 있고, 그런 단어들의 빈도수가 높기 때문이다.

SVM^{light}의 성능이 높게 나오려면 해당 범주를 잘 표현하는 문장이 텍스트 문서 내에 다수 포함하고 있어야 한다. 그러나, 댓글의 특성상 대부분이 100자 이내의 한 줄 문장이며, 해당 범주를 잘 표현하는 자질이 한 문장에 한번내지 두번 등장하기 때문에 높은 성능을 내는데 한계가 있었다.

5. 결론

본 논문에서는 포탈사이트의 뉴스에 대한 악성 댓글을 판단하는 시스템을 제안하고 구현하였다. 제안한 시스템의 성능은 문서 처리 과정의 자질 선택시 어절 자체를 포함한 명사 데이터의 실험이 다른 품사 실험보다 우수한 성능을 보였다.

향후 과제로는 문서에 자주 사용되는 단어 중에는 문서 범주화에 영향을 미치지 않은 것이 있다. 그러나, 이 단어는 높은 빈도에 의해 가중치 계산에 큰 영향을 준다. 이런 불용어를 제거하는 방법이 연구되어야 할 것이다.

현재 수집된 금칙어 사전을 이용해서 분류기가 잘못 판단한 댓글을 다시 한번 필터링 하거나, "극도로 싫어한다.", "쌤통이다.", "얼마나 잘되나" 와 같은 악성 댓글에 나타나는 독특한 표현들에 대한 연구를 적용한 최적의 시스템을 제시할 수 있을 것이다.

참고문헌

[1] Sara Owsley, Sanjay C. Sood, and Kristian J. Hammond, "Domain Specific Affective Classification of Documents", The AAAI Spring Symposia on Computational Approaches to Analysing Weblogs, pp.181-183, 2006.

[2] http://svmlight.joachims.org, T. Joachims, Cornell University, 2002.

[3] Hong Qu, Andrea La Pietra, Sarah Poon, "Automated Blog Classification: Challenges and Pitfalls", The AAAI Spring Symposia on Computational Approaches to Analysing

Weblogs, pp.184-186, 2006.

[4] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up? Sentiment Classification using Machine Learning Techniques", In Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp.79-86, 2002.

[5] Soo-Min Kim and Edward Hovy, "Determining the sentiment of opinions", In COLING-2004, pp. 1367-1373, 2004.

[6] N.Hiroshima, S. Yamada, O. Furuse and R. Kataoka,"Searching for Sentences Expression Opinions by using Declaratively Subjective Clues", In Proceedings of the Workshop on Sentiment and Subjectivity in Text, c2006 Association for Computational Linguistics, pp.39–46, 2006.

[7] P.D. Turney and M.L. Littman, "Unsupervised Learning of Semantic Orientation from a Hundred-billion-word Corpus", National Research Council, Institute for Information Technology, (No. ERB-1094, NRC #44929), 2002.

[8] Michael Gamon. "Sentiment Classification on Customer Feedback Data: noisy data, large feature vectors, and the role of linguistic analysis", In Proceedings the 20th, International Conference on Computational Linguistics, pp.841–847, 2004.

[9] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan, "A Sentimental Education: Sentiment Analysis using Subjectivity Summarization Based on Minimum Cuts". ACL 2004, pp.271–278, 2004.

[10] P.D. Turney and M.L. Littman. "Measuring Praise and Criticism: Inference of Semantic Orientation from Association" ACM Transactions on Information Systems (TOIS). Vol. 21, No. 4. October 2003. pp.315-346, 2003.

[11] 강승식, 한국어 형태소 분석 및 정보 검색, 홍릉과학출판사, 2002.

[12] Joachims, T. "Text categorization with Support Vector Machines: Learning with Many Relevant Features. In Machine Learning", ECML-98, Tenth European Conference on Machine Learning, pp.137-142, 1998.