온라인게임 채팅에서의 비속어 차단 시스템

이성욱*

A Swearword Filter System for Online Game Chatting

Songwook Lee*

이 논문은 2011년도 충주대학교 교내학술연구비의 지원을 받아 수행한 연구임

요 약

온라인 게임의 활성화로 온라인 게임의 폐해도 증가하고 있는데 온라인 게임의 대표적인 폐해 중 하나인 언어 폭력 문제가 심각한 사회문제를 야기하고 있다. 본 논문은 온라인 게임의 채팅에 나타나는 비속어를 자동으로 차단하는 시스템을 제안한다. 우리는 온라인 게임의 채팅창에 나타나는 문장을 수집하였고 비속어 포함 문장과 정상 문장으로 수동으로 분류하였다. 음절 n-gram과 어휘-품사 쌍을 자질로 사용하며 카이제곱 통계량을 이용하여 자질을 선택한다. 선택된 자질들을 이진 가중치로 표현하여 지지벡터기계(SVM)를 학습한 후, SVM 분류기로 각 문장의 차단 여부를 결정하였다. 실험 결과, 수집된 데이터에 대해 약 90.4%의 F1 정확률을 얻었다.

ABSTRACT

We propose an automatic swearword filter system for online game chatting by using Support Vector Machines (SVM). We collected chatting sentences from online games and tagged them as normal sentences or swearword included sentences. We use n-gram syllables and lexical-part of speech (POS) tags of a word as features and select useful features by chi square statistics. Each selected feature is represented as binary weight and used in training SVM. SVM classifies each chatting sentence as swearword included one or not. In experiment, we acquired overall 90.4% of F1 accuracy.

키워드

온라인게임, 비속어 차단, 지지벡터기계, 카이제곱 통계량

Key word

online game, swearword filter, support vector machine, chi square statistics

* 정회원 : 충주대학교 (교신저자, leesw@cjnu.ac.kr)

접수일자 : 2011. 05. 30
심사완료일자 : 2011. 06. 16

I. 서 론

인터넷의 발달과 더불어 온라인 게임을 이용하는 사용자가 폭발적으로 증가하고 있다. 인기 있는 온라인 게임은 동시접속자수가 40만 명을 넘어서기도 한다. 온라인 게임 산업이 발달하는 가운데 온라인 게임이 우리 사회에 끼치는 문제도 커지고 있다. 온라인 게임 중독에 인한 사회문제가 커지자 여성가족부와 문화부에서는 청소년의 온라인 게임 이용시간을 제한하는 제도를 추진하고 있을 정도이다. 게임 중독과 더불어 온라인 게임의 가장 큰 폐해가 언어 폭력이며 청소년의 73%가 욕설을 사용하며 이중 52%가 온라인 게임에서 폭력적 언어를 경험한다고 한다[1].

네오위즈, 한게임, 엔씨소프트 등의 게임사들은 이러한 비속어를 금칙어로 지정하여 차단하고 있으며 24시간 모니터링을 통해 적발하기도 한다[2, 3]. 그러나 이러한 금칙어를 통한 차단 시스템은 비속어와 동음이의어를 차단하는 문제가 있다. 또한 사용자들은 금칙어를 피하기 위해 자음과 모음을 교묘히 바꾸어 더 다양한 형태의 비속어를 만들어 내는데 이를 처리하기 위해 금칙어를 계속 수동으로 추가해야하는 어려움이 있다.

본 연구의 목적은 온라인 게임 채팅에 발생하는 비속어 문장을 자동으로 차단하는 것이다. 비속어 차단 문제는 내용이 비교적 짧다는 점을 제외하면 스팸 메일 판별 문제와 매우 유사한 문제이다. 스팸 메일 판별과 관련된 연구에서 [4]는 카이제곱 통계량을 이용하여 자질을 선택하였으며 지지벡터기계(Support Vector Machines)를 이용하여 시스템을 학습하였고, 나머지 대부분의 연구는 베이지안 분류기를 기반으로 하고 있으며[5, 6, 7, 8, 9], 그 외, 마코프 랜덤 필드 (Markov Random Field) 모델[10]과 k-Nearest Neighbor(k-NN) 방법[11]을 이용한 연구 등이 있다. 그 외, 유사한 문제로 스팸 블로그 판별 문제가 있으며 해결 방법은 스팸 메일 문제와 유사하다[12, 13].

본 연구에서는 카이제곱 통계량을 이용하여 자질을 선택한 후, 이를 지지벡터기계 학습에 사용하는 방법[4, 14, 15]을 비속어 차단 시스템에 이용하는 것을 제안한다. 비속어 차단 시스템은 채팅에 사용된 문장을 자동으로 비속어 포함 문장과 정상 문장으로 분류하는 이진 분류 시스템이다.

본 논문의 구성은 다음과 같다. 2장에서는 시스템 구조를 살펴보고, 3장에서는 카이제곱 통계량과 지지벡터기계의 학습에 사용되는 자질을 어떻게 구성하고 사용하였는지 설명한다. 4장에서 실험 결과를 보이며 5장에서 결론을 내린다.

II. 시스템 구조

다음 그림 1은 제안하는 시스템의 구조도이다.

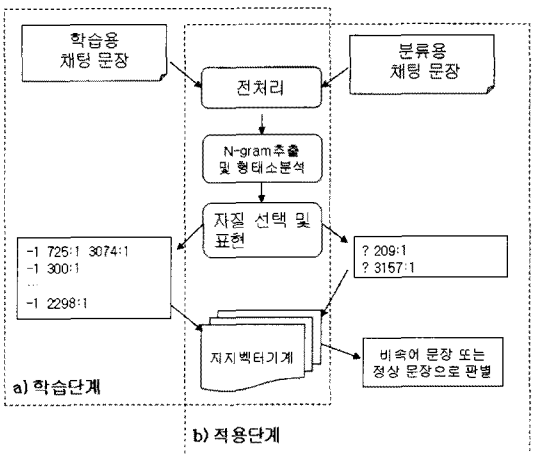

그림 1. 제안 시스템 구조도.
Fig. 1 The system architecture.

제안하는 비속어 차단 시스템은 크게 두 단계로 나뉜다. 먼저 학습단계에서는 학습용 채팅 문장으로부터 지지벡터기계(SVM)의 학습에 사용할 수 있는 자질(feature)을 추출한다.

학습용 채팅 문장은 먼저 문장 부호 등을 제거하는 전처리과정을 거친 후, 음절 n-gram추출과 형태소 분석 단계를 거쳐 음절 n-gram과 어휘/품사 쌍으로 자질을 이룬다. 각 자질은 해당하는 차원의 축을 이루며 각 자질의 가중치가 그 차원의 값이 된다. 학습단계에서 모든 채팅 문장에 대한 벡터 데이터가 만들어지면 SVM을 학습한다.

SVM의 학습이 끝나면 적용단계에 들어간다. 적용단계에서는 학습에 사용되지 않은 분류용 채팅 문장이 입

력된다. 입력된 문장은 학습단계와 유사하게 전처리 단계와 형태소 분석 단계 및 자질 추출 단계를 거쳐 다차원 상의 한 점을 이루는 벡터 데이터가 되고 이를 SVM이 비속어 포함 문장 또는 정상문장으로 판별하게 된다.

Ⅲ. 자질과 자질 선택

기계 학습에서 적절한 자질의 선택은 시스템의 성능에 많은 영향을 끼친다. 채팅 문장의 경우, 한글맞춤법을 무시한 탈락, 축약, 생략 현상과 띄어쓰기 오류가 많이 나타나고 있으며 이는 채팅에 사용되는 문장이 문어보다는 구어에 가깝기 때문이다. 다음 표 1은 채팅 문장의 이러한 특성을 나타낸 것이다.

표 1. 채팅 문장의 특성
Table 1. Characteristics of chatting sentences

특성	예
축약과 생략	설(서울), 겜(게임), 셤(시험), 비됴(비디오), 글쿠나(그렇구나), ㅎㅎ(하하) 등
첨가	그런감, 하세염, 넵 등
발음대로 표기	마니(많이), 조아(좋아),시퍼(싶어),멀져(멀지요) 등
숫자와 알파벳 혼용	20000(이만), 밥5(바보), 바2(바이), 감4, 50쇼(어서 오십시오) 등
약어	정모(정기모임), 강퇴(강제퇴장), 남친(남자친구) 등
의성어 의태어	ㅋㅋ(크크), ㅋㄷㅋㄷ(커득키득), ㄷㄷㄷ(덜덜덜)
이모티콘	-;, ^^, TT, ㅅㅅ 등
신조어	킹왕짱, 듣보잡, 막장, 안습, 고고씽, 지못미, 차도녀, 쩐다 등

표 1에서와 같이 채팅 문장은 한글맞춤법을 무시한 경우와 띄어쓰기를 무시한 경우가 많기 때문에 형태소 분석이 올바르게 이뤄지지 않는다. 따라서 형태소 분석 결과물인 어휘/품사 쌍만을 자질로 사용하기에는 자질의 개수가 충분하지 않아 기계학습의 성능이 떨어지게 된다. 그래서 우리는 다중 언어의 자질 추출 등에 사용되는 n-gram 자질[12]을 추가로 사용하였는데, 1-gram, 2-gram, 3-gram까지 사용하였다. n-gram 자질은 문자 길이 "n"을 윈도우 사이즈로 사용하여 인접한 문자열을 추출한 것이며, 예를 들어 "개무시"라는 단어의 음절 2-gram은 "개무, 무시"가 된다.

본 연구에서 사용된 데이터는 온라인 게임의 채팅창에서 직접 수집하였고, 수동으로 정상 문장과 비속어 포함 문장으로 태깅하여 실험에 이용하였다. 다음 그림 2는 정상인 입력문장 "뭐라해야될까요..?"과 비속어를 포함한 문장 "내말다씹으셔"의 각 단계별 처리 예를 나타낸다.

a) 입력문장
뭐라해야될까요..?
내말다씹으셔

b) 전처리 후
뭐라해야될까요
내말다씹으셔

c) 품사 부착 후
뭐라해야/UNK+되/VX+ㄹ까요/E
내말다/UNK+씹/VV+으셔/E
(UNK:미등록어, VX:보조용언, E:어미, VV:동사)

d) N-gram 자질 및 어휘/품사 자질 추출

1-gram: 뭐, 라, 해, 야, 될, 까, 요
내, 말, 다, 씹, 으, 셔
2-gram: 뭐라, 라해, 해야, 야될, 될까, 까요,
내말, 말다, 다씹, 씹으, 으셔
3-gram: 뭐라해, 라해야, 해야될, 야될까, 될까요
내말다, 말다씹, 다씹으, 씹으셔
어휘/품사: 뭐라해야/UNK, 되/VX, ㄹ까요/E
내말다/UNK, 씹/VV, 으셔/E

그림 2. 품사 부착 전후의 채팅 문장 예
Fig. 2. An example of Part of Speech tagged sentence.

입력된 문장은 먼저 그림 2의 a)와 같이 문장부호 등을 제거하는 전처리기를 거친다. 전처리 된 문장은 이후 그림 2의 c)와 같이 MS Wordbreaker2007 형태소 분석기를 이용하여 자동으로 품사가 부착된다.

입력 문장에 품사가 부착된 후, 그림 2의 d)와 같이 최종적으로 실험에 이용할 n-그램 자질과 어휘/품사 자질을 추출한다.

그림 2의 예와 같이 하나의 입력 문장에서 추출되는

자질의 개수는 문장의 길이에 비해 비교적 많은 편이다. 총 자질의 개수는 수집된 학습데이터에서 발견되는 모든 n-gram과 어휘/품사 쌍이 된다. 이러한 자질들 중에서는 유용한 자질과 불필요한 자질이 섞여있다. 이 중에서 좋은 자질을 선택하기 위해 카이제곱 통계량을 이용하며 카이제곱 통계량을 계산하는 식은 다음 수식(1)과 같다[15].

$$\chi^2(f,s)=\frac{(A+B+C+D)\times(AD-BC)^2}{(A+B)\times(A+C)\times(B+D)\times(C+D)} \quad (1)$$

A는 비속어 문장 s 중에 자질 f를 포함하고 있는 문서의 수이고, B는 범주 s 이외의 문서, 즉 정상 문장 중 속해 있는 문서 중에 자질 f를 포함하고 있는 문서의 수이다. 또한, C는 비속어 문장 s에 속해 있는 문서 중에 자질 f를 포함하지 않는 문서의 수이며, D는 범주 s외의 문서 중에 자질 f를 가지고 있지 않는 문서의 수이다. 자질 f와 범주 s가 완전히 독립적이면 0의 값을 갖는다.

실험 대상인 채팅 문장은 비교적 짧기 때문에 자질의 빈도수를 이용한 TF-IDF 가중치 등을 사용하지 않고 이진 가중치를 각 자질의 가중치로 사용하였다.

SVM은 이진 분류기이므로 우리는 비속어 문장과 정상 문장을 분류하기 위해 하나의 SVM모델만을 학습하면 된다. 비속어 문장인 경우 양(+1)의 자질을, 정상 문장인 경우 음(-1)의 자질을 부여하였다. SVM의 학습을 위해서 각 자질은 벡터의 각 차원을 구성하며 전체 자질의 개수가 다차원 공간의 차원을 결정한다. 한 개의 데이터는 다차원 공간상의 벡터로 표현되는데, 데이터에 존재하는 자질들만 이진 가중치로 표현함으로써 하나의 벡터를 구성하게 된다. 이렇게 만들어진 벡터들을 이용하여 SVM을 학습하는데, 본 연구에서는 LIBSVM[16]을 이용하였고 선형커널을 이용하여 학습하였다.

Ⅳ. 실험 및 결과

실험에는 "던전앤파이터", "리니지", "마구마구", "서든어택", "아이온", "카오스" 등의 온라인 게임에서 자체적으로 수집한 채팅 데이터를 사용하였다. 총 10,520개의 채팅 문장[1]이 수집되었는데 그 중 비속어가 포함된 문장은 585개였다. 비속어 문장의 비율과 정상 문장의 비율이 차이가 크므로 공정한 실험을 위해 정상 문장과 비속어 문장의 비율을 똑같이 맞추어 주기 위해 비속어 문장의 개수만큼 정상 문장을 무작위 추출하였다. 따라서 실험에는 586개의 정상 문장과 585개의 비속어 문장을 사용하였으며, 5:1의 확률로 무작위 추출하여 251개를 평가 데이터로 사용하고 나머지를 학습 데이터로 사용하였다.

다음 표 2는 자질의 종류에 따른 성능을 살펴본 것이다.

표 2. 자질의 종류에 따른 성능
Table 2. Experimental results for types of features

자질의 종류	정확도(%)	비고
어휘/품사	57.4%	모든 자질 사용
음절 n-gram	87.3%	
음절 n-gram, 어휘/품사	88.0%	

표 2에서와 같이 어휘/품사 정보만으로는 좋은 성능을 얻을 수 없었고 음절 n-gram 자질을 사용한 것이 더 좋은 성능을 보였다. 음절 n-gram에 어휘/품사 자질을 추가하여도 큰 성능 향상은 없었는데, 이는 채팅 문장의 특성상 형태소 분석 오류가 많기 때문이다.

다음 표 3은 카이제곱 통계량을 이용하여 자질의 개수를 조정했을 때의 정확도를 나타낸다.

표 3. 자질의 개수에 따른 정확도 비교
Table 3. Experimental Results of # of features

자질의 개수	정확도(%)	비고
9,223	88.0	모든 자질
5,259	86.1	$\chi 2>1$
962	90.0	$\chi 2>2$
454	88.0	$\chi 2>3$

표 3에서와 같이 카이제곱 통계량을 이용하여 자질의 개수를 약 1/10로 줄였을 때 더 좋은 성능을 보였는데 이는 카이제곱 통계량을 이용한 자질 선택 방법이 유용함

1) 사용자가 채팅창에 엔터(Enter) 단위로 입력하며, 한 개 이상의 문장으로 구성됨.

을 보여준다.

다음 표 4는 제안 시스템의 성능을 정확도(accuracy), 정확률(precision), 재현율(recall), F1[2], Hm 오류율, Sm 오류율[17] 등으로 평가한 결과이다.

표 4. 제안 시스템의 성능(%)
Table 4. The performance of the system

정확도	정확률	재현율	F1	Hm	Sm
90.0	90.5	90.2	90.4	0.04	0.16

Hm(%) = 정상 문장을 잘못 분류한 개수/정상 문장의 수*100
Sm(%) = 비속어 문장을 잘못 분류한 개수/비속어 문장의 수*100

Hm 오류율과 Sm 오류율은 스팸메일 분류 등에서 사용하는 평가척도인데 일반적으로 정상문장을 잘못 분류한 Hm 오류율이 Sm 오류율보다 작은 시스템이 사용자에게 더 선호되는 시스템이다. 그 이유는 정상문장을 비속어 문장으로 잘못 분류했을 때 사용자가 입는 피해가 더 크기 때문이다. 제안 시스템은 Hm이 Sm보다 아주 작은 값을 가지는데 이는 바람직한 결과라 할 수 있다.

시스템의 분석 오류 중 데이터 태깅 오류를 제외하면 대부분의 오류들은 '뷴선', '색기' 등과 같이 자-모음을 변형한 비속어 사용으로 인한 오류들이며 이런 오류를 해소하기 위해서는 더 많은 유형의 비속어 채팅 문장의 수집이 필요하며 비속어와 그 변형어의 자-모음열의 유사도 측정 방법에 대한 연구도 뒤따라야할 것이다.

V. 결론 및 향후 과제

본 논문에서는 온라인 게임상의 언어폭력을 막기 위해 채팅 문장의 비속어 유무를 판별하는 시스템을 제안하였다. N-gram 자질과 어휘/품사 자질을 이용하여 SVM을 학습하여 자동으로 비속어 문장을 차단하는 시스템을 제안하였다. 각 자질의 카이제곱 통계량을 이용하여 좋은 자질을 선택함으로써 시스템의 성능을 향상 시켰다. 비속어 문장을 차단하는 다른 연구가 아직 보고 되지 않아 다른 시스템과 직접적으로 비교할 수는 없지만, 제안 시스템은 90.4%의 F1 정확률을 얻어 비교적 좋은 성능을 얻었다. 향후, 시스템의 성능 향상을 위해서는 더 많은 채팅 데이터의 수집이 필요하다. 현 시스템이 수행하는 비속어가 포함된 문장 단위의 차단에서 더 나아가 비속어가 발생한 어절만 차단하는 시스템으로의 발전이 필요하다. 또한 비속어의 순화어로의 변환이나 채팅어의 표준어로의 변환 등에 관한 연구도 필요하다.

2) F1은 정확률과 재현율의 조화평균값

참고문헌

[1] http://www.zdnet.co.kr/news/news_view.asp?artice_id=20110105084601, 2011.01.05.

[2] http://www.edaily.co.kr/news/NewsRead.edy?SCD=DB41&newsid=01922086589626600, 2009.03.24.

[3] http://www.ajnews.co.kr/view.jsp?newsId=20101021000646, 2010.10.21.

[4] 이성욱, "카이제곱 통계량과 지지벡터기계를 이용한 스팸메일 필터", 정보처리학회논문지, 제17-B권, 제3호, pp.249-254, 2010.

[5] V. Keselj, E. Milios, A. Tuttle, S. Wang, and R. Zhang, "TREC 2005 Spam Track: Spam Filtering Using N-gram-based Techniques", *Proceedings of Text REtrieval Conference,* 2005.

[6] 김현준, 정재은, 조근식, "가중치가 부여된 베이지안 분류자를 이용한 스팸 메일 필터링 시스템", 정보과학회논문지, 제31권 8호, 2004, pp.1092-1100.

[7] R. Segal, "IBM SpamGuru on the TREC 2005 Spam Track", *Proceedings of Text REtrieval Conference,* 2005.

[8] A. Brakto and B. Filipic, "Spam Filtering Using Character-Level Markov Models: Experiments for the TREC 2005 Spam Track", *Proceedings of Text REtrieval Conference,* 2005.

[9] L. A. Breyer, "DBACL at the TREC 2005", Proceedings of Text REtrieval Conference, 2005.

[10] F. Assis, W. Yerazunis, C. Siefkes, and S. Chhabra, "CRM114 versus Mr. X: CRM114 Notes for the TREC 2005 Spam Track", *Proceedings of Text REtrieval Conference,* 2005.

[11] W. Cao, A. An, and X. Huang, "York University at TREC 2005: SPAM Track", *Proceedings of Text REtrieval Conference,* 2005.

[12] P. Kolari, A. Java, and T. Finin, "Characterizing the splogosphere", *Proceedings of WWW 2006, 3rd Annual Workshop on the Webloggging Ecosystem: Aggregation, Analysis and Dynamics.* 2006.

[13] 이성욱, "지지벡터기계를 이용한 스팸 블로그(Splog) 판별 시스템", 한국해양정보통신학회 논문지, 제15권, 제1호, pp.163-168, 2011

[14] 은종민, 이성욱, 서정연, "지지벡터기계(Support Vector Machines)를 이용한 한국어 화행분석", 정보처리학회논문지, 제.12-B권, 제3호, pp.365-368, 2005.

[15] Y. Yang and Jan O. Pedersen. "A comparative study on Feature selection in text categorization," *Proceedings of the 14th International conference on Machine Learning,* 1997.

[16] http://www.csie.ntu.edu.tw/~cjlin/libsvm, 2009.

[17] G. V. Cormack and T. R. Lynam. "TREC 2005 spam track overview," *Proceedings of Text REtrieval Conference,* 2005.

저자소개

이성욱(Songwook Lee)

1996년 서강대학교 컴퓨터학과 학사
1998년 서강대학교 컴퓨터학과 석사
2003년 서강대학교 컴퓨터학과 공학박사

2004-2005년 LG전자 기술원 선임연구원
2005-2007년 동서대학교 컴퓨터정보공학부 전임강사
2007년-현재 국립충주대학교 컴퓨터과학과 조교수
※관심분야 : 인터넷응용시스템, 한국어정보처리