SVM을 이용한 온라인게임 비속어 필터링 시스템

박교현° 이지형
성균관대학교, 전자전기컴퓨터공학과
megagame@skku.edu, jhlee@ece.skku.ac.kr

Developing a Vulgarity Filtering System for Online Games using SVM

Kyo-hyeon Park°, Jee-hyong Lee
Dept. of Electrical and Computer Engineering, Sungkyunkwan University

요 약

최근 온라인 게임 산업이 커짐에 따라 이를 즐기는 유저도 급증하고 있다. 온라인 게임에서는 일반적으로 유저들이 서로를 구분하기 위해 사용하는 사용자 이름과 상호간 의사소통을 하기 위한 채팅을 지원한다. 유저의 수가 증가함에 따라 대화의 양은 더욱 더 많아지고, 선정성, 폭력성을 띄는 언어의 문제로 이어지고 있다. 이는 특히 18세 이하도 이용가능한 게임을 만드는 경우 더욱 중요하다. 하지만 대부분의 게임들이 금지어 리스트에 따른 단어 매칭방식의 비속어 필터링만을 제공하고 있다. 이러한 방법은 금지어로 지정된 단어를 포함한 정상적인 채팅도 막을 뿐만 아니라 일부 음절을 다른 기호로 바꾸어 표기한 비속어는 걸러내지 못한다. 변형된 단어들을 충분히 처리하지 못한다면 비속어 필터링 시스템은 단지 무력하고 쓸모없는 존재가 될 뿐이다. 본 논문에서는 SVM을 이용하여 학습이 가능한 비속어 필터링 시스템을 제안하고자 한다. SVM을 이용하면 사용자 편의성을 해치지 않고서도 보다 많은 종류의 비속어들을 효과적으로 걸러낼 수 있다.

1. 서 론

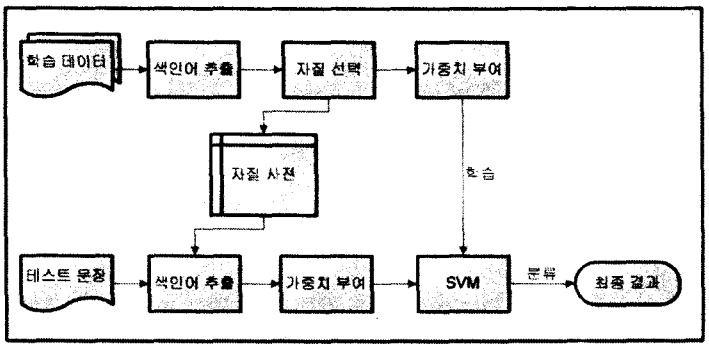
2004년 기준, 온라인게임 시장 규모는 34억불이며 전년 대비 57.9%의 성장을 이루었다. 2004년과 비교하여 2007년에는 2.4배가 성장한 80억불의 시장 규모를 이룰 것으로 예상된다[1]. 그러나 이에 따른 여러 가지 문제들도 점차 부각되기 시작하였다.

온라인 게임에서는 유저들이 서로를 구분하기 위해 사용하는 사용자 이름과 상호간 의사소통을 하기 위한 채팅을 지원한다. 온라인 게임 시장 규모가 커지고 이를 즐기는 유저의 수가 증가함에 따라 대화의 양은 더욱 더 많아지고, 이는 선정성, 폭력성을 띄는 언어의 문제로 이어지고 있다. 특히 18세 이하도 이용가능한 게임을 만드는 경우 더욱 중요하다[2].

현재 서비스 중인 대부분의 온라인 게임들은 금지어 리스트에 따른 단어 매칭방식의 비속어 필터링만을 제공하고 있으며 필터링하지 못하는 비속어의 경우 유저 신고 제도를 통하여 적발 및 처벌하는 방식을 채택하고 있다. 그러나 단어 매칭 방식은 금지어로 지정된 단어를 포함한 정상적인 대화도 막기도 한다. 이를 막기 위해 금지어를 추가하면 할수록 일상적 대화도 제약받을 확률이 높으며 비교 검사를 위해 더 많은 CPU 자원을 소모하게 된다.

단어 매칭 방식은 일부 음절을 다른 기호로 바꾼 비속어는 걸러내지 못하는 문제도 있다. 신고 제도를 이용하여 처벌하더라도 언어폭력에 상처받은 유저의 마음까지는 보상받지 못한다. 정상적인 대화를 방해하거나 변형된 비속어 단어들을 충분히 걸러내지 못한다면 비속어 필터링 시스템은 불편한 애물단지에 지나지 않는다. 심지어 최근엔 유저가 필터링 시스템의 작동 여부를 결정할 수 있도록 하여 비속어를 보게 되더라도 대화가 방해받지 않는 것과 불편하더라도 비속어를 최대한 보지 않도록 하는 것을 선택하도록 하는 게임마저도 나오고 있다.

이러한 문제를 해결하기 위해 본 논문에서는 Support Vector Machine(SVM)을 사용한 감독자 학습(supervised learning)방식의 비속어 필터링 시스템을 제안하고자 한다. SVM은 문서 분류 문제에 있어서 매우 효율적인 방법이다. 이를 통해 게임 상에서 사용되는 비속어 문장과 정상적인 문장을 구분하는 방법을 제시한다.



<그림 1> 비속어 필터링 시스템의 처리 과정

2. 관련 연구

본 논문에서 비속어 분류기로 사용할 Support Vector Machine(SVM)은 기계 학습(machine learning)의 한 가지 방법이다. +1과 -1의 두 클래스로 분류한 입력 데이터들을 감독자 학습(supervised learning) 방법을 통해 각 클래스에 포함된 패턴들을 분리하는 절단 평면인 hyperplane을 학습한다. hyperplane을 결정하는 입력 패턴들을 support vector라 하며, 이 패턴을 두 개의 집단으로 분리할 수 있을 때, 이 hyperplane은 면으로부터 support vector까지의 거리를 최대화하고 모든 support vector는 hyperplane으로부터 동일한 최소 거리를 가지게 된다[3].

SVM은 현재 알려져 있는 많은 기법 중에서 가장 인식 성능이 뛰어난 학습 모델중 하나이다. 특히 분류 문제에 있어서 일반화 기능이 높기 때문에 현재에는 생물정보학, 문자인식, 필기인식, 얼굴 및 물체인식 등 다양한 분야에서 성공적으로 적용되고 있다. SVM은 명백한 이론적 근거에 기반하므로 결과 해석이 용이하고, 실제 응용에 있어서 인공신경망 수준의 높은 성과를 내면서도 한계점으로 지적되었던 과대적합, 국소최적화 등을 완화하는 장점을 가진다. 또한 적은 학습 자료만으로도 신속하게 분별 학습을 수행할 수 있다.

3. 비속어 필터링 시스템

본 논문에서 제안하는 비속어 필터링 시스템은 <그림 1>과 같은 과정을 통해 대화를 정상과 비속어 두 분류로 나눈다. 감독자 학습(supervised learning)을 기반으로 하고 있기에 먼저 주어진 학습 데이터를 통해 SVM 분류기를 학습시킨 후, 학습과정에서 얻어진 자질을 기반으로 학습된 SVM을 통해 비속어 분류를 수행한다. 각 단계에 따른 세부 처리는 다음과 같다.

3.1 색인어 추출

색인어 추출 과정은 수집된 데이터를 변환하여 해당 데이터의 내용이나 특징을 잘 반영하는 자질(feature)을 생성하는 전처리 단계이다. 본 논문에서 분류하고자 하는 채팅 데이터에는 주로 길이가 매우 짧고 항상 정확한 맞춤법이 지켜지지는 않는다는 특징이 있다. 특히 부정확한 띄어쓰기 외에도 음절을 유사한 모양이나 발음의 문자로 변형하여 사용하는 경우도 자주 볼 수 있다. 이는 게임의 주 이용자층이 통신어체를 자주 사용하는 세대일수록 두드러진다. 따라서 형태소 분석이나 구문 분석의 방식으로는 분류를 위한 핵심어를 추출하기 힘들다. 이에 본 논문에서는 *n*-gram을 이용한 4단계의 처리 과정을 거쳐 자질을 생성하는 방식을 채택하였으며 각각의 단계는 다음과 같다.

- 단계 1. 데이터 내의 어절 분리
  빈칸, 마침표, 쉼표, 따옴표 등을 구분자로 하여 각 어절을 분리한다.
- 단계 2. 각 어절에서 비색인 분절을 절단
  조사, 어미, 접미사 등이 결합된 다양한 형태의 음절인 비색인 분절을 제거한다.
- 단계 3. 불용어 제거
  대화를 분류하는데 있어 특징으로 쓰기에 무의미한 어절들을 제거한다. 예를 들면 '내년' 같이 비속어와 정상 문장을 분리할 능력이 없는 단어들을 제거한다.
- 단계 4. 각 분절들을 *n*-gram들로 분할
  복합명사를 인접한 n개의 음절로 나누어 사용하는 n-gram으로 분할한다. 예를 들면 '정보과학'이란 어절을 2-gram은 '정보', '보과', '과학' 으로, 3-gram

은 '정보과', '보과학'으로 분리한다.

n-gram 방식은 다음과 같은 장점을 가진다. 첫째로 어절 단위로 분리할 때의 절단 오류로 인한 파급 효과를 완화한다. 둘째로 복합 명사의 띄어쓰기 문제를 완화한다. 셋째는 철자 오류나 일관성 없는 외래어 표기 문제를 적절히 극복할 수 있다.[4]

3.2 자질 선택

자질 선택은 전처리 단계를 통해 생성된 학습 문서에 나타나는 자질들 중에서 분류 학습에 유용하게 사용될 만한 자질만을 선택하는 단계이다. 본 논문에서는 카이제곱 통계량(Chi-square statistic)을 사용하여 비속어 분류에 효과적인 자질을 선택하도록 하였으며 그 공식은 다음과 같다[5].

$$x^2(t,c)=\frac{N(AD-CB)^2}{(A+C)(B+D)(A+B)(C+D)}$$

A: 범주 c에 속해있는 문장 중 단어 t 를 포함하고 있는 문장의 수
B: 범주 c에 속하지 않는 문장 중 단어 t 를 포함하고 있는 문서의 수
C: 범주 c에 속해 있는 문장 중 단어 t를 포함하고 있지 않은 문서의 수
D: 범주 c에 속하지 않는 문장 중 단어 t를 포함하고 있지 않은 문장의 수
N: 학습 데이터의 양

본 논문에서는 카이제곱 통계량이 일정치 이상인 단어만을 자질로 선택하도록 하였다. 이렇게 학습과정에서 선택된 자질은 자질 사전에 저장되고, 테스트 문장을 분류하고자 할 때 색인화 과정에서 자질 사전을 참조하여 등록된 단어들로만 색인화한다.

3.3 가중치 부여

가중치 부여 과정은 문장을 표현할 선택된 자질에 가중치를 부여하는 단계이다. 본 논문에서는 가중치를 결정하기 위해 TF-IDF 기법을 사용하였다. 각 자질의 가중치는 해당 문장에서 각 자질의 빈도(TF)와 역문헌빈도(IDF)의 곱으로 나타내며 그 식은 다음과 같다.

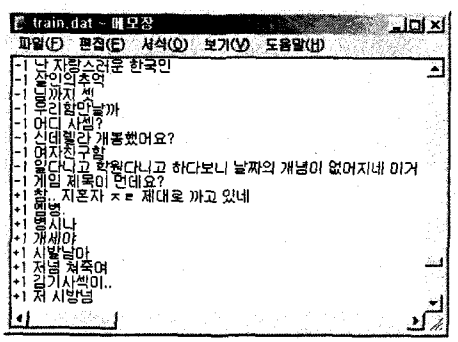
$$w_{ij}=tf_{ij}\log\left(\frac{N}{df_j}\right)$$

TF(term frequency)는 한 단어가 해당 문장 내에 등장하는 빈도수를 나타내며 IDF(inverse document frequency)는 N개의 문장 집합 중에서 해당 단어가 포함된 문장의 수의 역수이다. 어떠한 단어가 한 문장에 여러번에 걸쳐 나타난다면 그 단어의 중요도는 높을 것이고, 여러 문장에 걸쳐 모두 나타난다면 그 단어가 가지는 문장 분류 능력은 떨어진다고 볼 수 있다. 불필요한 실시간 연산을 줄이기 위해 IDF값은 학습과정의 자질 선택 단계에서 선택된 자질과 함께 저장된다.

4. 실험 및 결과

4.1 실험 방법

SVM의 학습 및 테스트에 쓰일 데이터는 게임 환경과 유사하게 대화체의 채팅이 오가며 역시 비속어 문제가 존재하는 온라인 채팅사이트의 대화 기록에서 수집하였다. 대부분의 채팅사이트들이 대화 기록을 보존하는 기능을 제공하고 있어 게임 환경보다는 데이터를 얻기 수월한 면이 있다. 먼저 채팅 기록을 정상과 비속어 두 분류로 나누고 정상적 문장은 -1(negative), 비속어 문장은 +1(positive) 값으로 설정하였다.



<그림 2> 수집 및 분류된 데이터

앞서 수집 및 분류한 데이터를 가지고 학습과 테스트의 두 그룹으로 나누어 <그림 1> 과정을 수행한다. 여기서 사용된 데이터셋의 구성은 다음과 같다.

<표 1> 데이터셋의 구성

	학습 데이터	테스트 데이터
정상	100	100
비속어	100	100
총합	200	200

본 실험에서는 색인어 추출에 2-gram 방식을 사용하였으며, SVM 분류기로는 Joachims의 SVMlight[6]를 사용하였다.

4.2 실험 결과

제안한 시스템의 분류 성능을 평가하기 위해 실험결과를 다음의 수치들로 나타내었다.

$$Precision = \frac{TP}{TP+FP} \times 100$$

$$Recall = \frac{TN}{TN+FN} \times 100$$

TP: True Positive FP: False Negative
TN: True Negative FN: False Negative

위의 공식에서 Precision은 총 비속어 대화 데이터 중에 맞게 분류된 데이터(true positive)의 비율이고, Recall은 정상 대화 데이터 중에 맞게 분류된 데이터(true negative)의 비율이다. Precision이 높을수록 비속어의 필터링 성능은 높아지며 Recall이 높을수록 잘못된 필터링의 수가 줄어들어 사용자의 편의성이 증대된다.

SVM을 이용한 테스트 데이터의 분류 수행으로 다음과 같은 결과를 얻을 수 있었다.

<표 2> SVM 분류 결과

Precision	86%
Recall	92%

5. 결 론

본 논문에서는 온라인 게임을 위한 학습 가능한 비속어 필터링 시스템을 제안하였다. SVM을 이용하면 기존의 단어 매칭 방법보다 사용자의 편의성을 해치지 않으면서도 보다 많은 종류의 욕설과 상스러운 말들을 걸러낼 수 있다. 특히 금지어 리스트를 관리하지 않고서도 비속어 대화들을 수집하여 학습시키는 것만으로 비속어의 분류가 가능하다.

그러나 이러한 방식의 필터링 시스템도 역시 한계는 있다. 그 중 한 가지는 학습 데이터의 구성에 따라 분류 성능이 크게 차이가 난다는 것이다. 같은 의미의 비속어라고 하여도 음절을 바꿔 표현하기 때문에 이를 충분히 학습하지 못한다면 제대로 분류되어지지 않는다. 또한 문장 단위로 자르거나 은유적인 표현의 욕설 혹은 음란성 대화는 본 시스템으로도 여전히 막기 힘들다. 시스템이 미처 막지 못하는 이러한 부분은 유저 신고 기능을 통하여 보조를 받아야 할 것이다.

향후, 분류 결과를 더욱 향상시킬 수 있는 자질 선택과 색인 알고리즘에 대하여 연구가 필요하다. 또한 좀더 정확한 실험결과를 위해 보다 다양한 유형의 채팅 데이터를 수집할 필요가 있으며, 비속어와 정상 대화를 구분하기에 적합한 학습용 데이터가 더욱 필요하다.

감사의 글 : 본 연구는 21세기 프론티어 연구개발 사업의 일환으로 추진되고 있는 정보통신부의 유비쿼터스컴퓨팅 및 네트워크원천기반기술 개발사업의 지원을 받았습니다.

참 고 문 헌

[1] 한국게임산업개발원, 2005년 게임백서, 4부, 1장, p552, 2005

[2] Shekhar Dhupelia, "Designing a Vulgarity Filtering System", Game Programming Gems 5, Charles River Media, 2005

[3] 서정우, "Support Vector Machine을 사용한 스팸메일 탐지 방안", 한국정보과학회, 2003

[4] 이준호, "한글 문서의 효과적인 검색을 위한 n-Gram 기반의 색인 방법", 정보관리 학회지, 1996

[5] 고영중, "문서 관리를 위한 자동 문서 범주화에 대한 이론 및 기법", 정보관리 연구논문지, 제33권, 2호, pp.16-32, 2002

[6] SVMlight, http://svmlight.joachims.org/