한글 비속어 필터링 시스템의 문제점과 그 해결법

Problems of Korean Vulgar Words Filtering System and its solution

윤태진
Yoon Taijin
부산대학교 컴퓨터공학과
ytj@pusan.ac.kr

ABSTRACT

비속어 필터링 시스템을 개발하다보면 여러가지 문제점에 부딪히게 된다. 장기간의 언어생활을 통해 비속어를 구분하는 능력을 얻은 인간과는 달리 컴퓨터는 한정된 자원과 정보를 통해서 비속어를 구별해내야 하기 때문이다. 특히 타 시스템에서 부가적인 모듈로써 돌아가야하는 비속어 필터링 시스템의 특성상 시스템에 걸리는 부하를 최소한으로 줄여줘야 할 필요가 있다. 더군다나 비속어 필터링 시스템은 필터링 성능보다도 정상적인 대화를 방해하지 않는 것이 중요하기에 성능과 편의성 사이에서 타협점을 찾을 필요가 있다. 그래서 본 보고서에서는 속도, 정확성, 편의성의 측면에서 비속어 필터링 시스템이 해결해야 할 문제점과 그 해결책에 대해서 서술하고자 한다.

KEYWORDS vulgar word, language processing, HCI

1 서 론

온라인 게임, 인터넷 게시판 등의 온라인 커뮤니케이션 등이 활발해지고 있는 요즘 온라인 상의 언어 폭력은 심각한 문제라 할 수 있다. 익명성으로 인한 책임의식의 부족이 비속어 사용에 대한 죄의식과 제약을 줄여주기 때문이다. 반면에 비속어 필터링에 관한 연구는 매우 취약하여 고성능의 비속어 필터링 시스템의 개발은 시급한 과제라 할 수 있다.

한글 비속어 필터링 시스템을 개발하면서 단순한 단어 필터링 시스템으로는 해결하기 어려운 여러가지 문제점을 발견하게 되었다. 그 중 가장 큰 문제라 할 수 있는 것은 정상단어의 필터링 문제이다. 실제 온라인 게시판이나 게임의 채팅 시스템에서 필터링 시스템에 대한 사용자의 불만은 비속어 필터링 성능 보다는 정상단어 필터링으로 인한 의사소통의 불편함이다. 사용자가 비속어 필터링 시스템에 대한 삭제를 요구하는 경우는 그리 드문 일이 아니다.

다음으로 어려운 문제는 복합 비속어로 인한 parsing문제이다. 한글은 단어의 결합을 통해서 새로운 단어를 만드는 경우가 많기때문에 복합 단어에 대해서 형태소 단위로 분리해줄 수 있는 언어 처리 시스템이 필요하다. 여러 문장이 입력되는 게시판이나 채팅에서 단순 빈칸 단위로 단어를 분리해서는 올바른 비속어 필터링을 수행할 수 없다. 이러한 빈칸은 비속어를 숨기기 위한 트릭으로도 사용되기 때문이다.

다음은 속도 문제이다. 비속어 필터링 시스템은 실시간으로 동작하는 온라인 게임을 위한 채팅 시스템이나 수많은 글이 등록되는 인터넷 BBS 등에 사용 되는데 수많은 텍스트를 서버에서 실시간으로

처리하려면 시스템에 너무 과도한 부하를 걸어서는 안된다. 비속어 필터링 시스템으로 인해 본시스템의 성능이 저하된다면 본말전도이기 때문이다. 이것을 위해서는 효과적인 데이터 구조와 비속어 검증 시스템이 필요하다.

그리고 사람의 행동성향에 대한 접근 방법이다. Email spam filtering system의 경우 black listing과 gray listing 등 email을 보낸 사람의 신용도에 따라서 email의 필터링 강도를 다르게 하는 방법을 사용한다. 평소의 유저의 성향을 분석해서 언어 폭력적인 수준을 평가하여 비속어 필터링 수준을 강화하는 등의 방법을 이용하여 비속어 사용에 대한 경각심을 불러일으키는 방법이 있을 수 있을 것이다.

본 보고서에서는 위에 언급된 비속어 필터링 시스템의 주요 문제점을 분석하고 해결방안을 모색하여 앞으로의 연구방향을 제시해보고자 한다.

2 정상 단어의 필터링 문제

온라인 게임을 하다 보면 채팅 시스템에 비속어 필터링 시스템 사용하는 경우를 많이 볼 수 있다. 그러나 너무 민감하게 적용된 경우 정상적인 단어를 필터링하여 게임에서 중요한 정상적인 의사소통을 방해하게 되는 경우가 있다.

표 1. 비속어 필터링 시스템으로 인한 정상 단어 필터링 예시

정상 문장	필터링 된 단어
상자위에 적이 있다.	자위
이번 섹터에서는 보스만 치세요.	섹
텔레비젼만 보지 말고 공부 좀 하세요.	보지
남자성기사입니다.	남자성기

표 2는 정상적인 대화가 비속어 필터링 시스템에 의해 차단된 예시를 보여주고 있다. 현재 개발중인 변형 비속어에 대응할 수 있는 비속어 필터링 시스템의 경우 이러한 정상단어로 인한 필터링 문제가 더 심각하다. 유사도를 통해 비속어 여부를 판정하게 되는 시스템의 특성상 비속어와 유사한 정상 단어의 경우도 필터링해버리는 문제점이 있기 때문이다.

그림 1는 개발된 비속어 필터링 시스템의 sensitivity와 specificity 그래프이다. sensitivity는 전체 비속어 중에서 필터링 된 비속어의 비율이고 specificity는 전체 필터링 된 단어 중에서 필터링 된 비속어의 비율이다. 이 두 수치는 80%에서 교차되는데 이 수치는 sensitivity에는 적합한 수치이지만 specificity에는 적합하다고 보기 어렵다.

전체 필터링 되는 단어 중에서 20%나 되는 정상 단어가 섞여 있다는 것은 심각한 문제이다. 단순하게 계산해도 문장에 단어가 5개 정도 사용된다면 이 중에서 한 단어 정도는 필터링 된다는 것이다. 실제로는 더 낮은 비율이겠지만 그렇다 하더라도 사용자에게 불편함을 주기에는 충분한 수치일 것이다.

이러한 정상 단어 필터링 문제를 해결하기 위한 해결책은 두가지 정도가 고려되고 있다. 첫째, 비속어 필터링 시스템의 성능을 향상시켜서 높은 threshold 값을 사용하더라도 sensitivity를 유지할 수

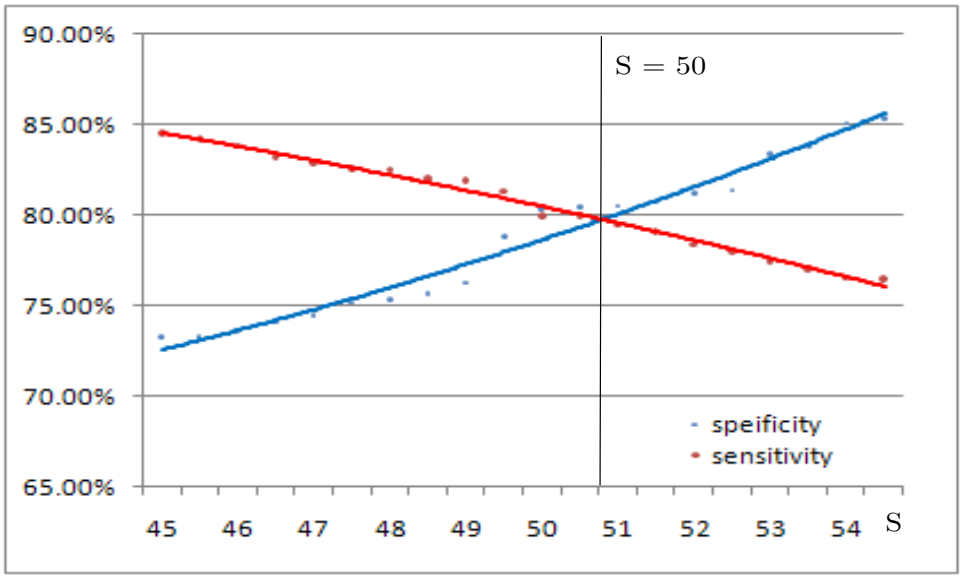

그림 1. 개발 중인 비속어 필터링 시스템의 필터링 threshold 수치에 따른 sensitivity와 specificity의 그래프

있도록 하는 것이다. 비속어 필터링 시스템의 성능을 높이기 위해서는 알고리즘의 개선과 더 많은 비속어 데이터를 수집하여 분류하는 것이다. 이것은 비속어 필터링을 연구해 내가면서 자연스럽게 해결되는 문제이다.

두번째 방법은 정상 단어를 미리 검출하는 것이다. 정상 단어를 미리 단어 사전에 저장 시켜 둔 뒤 입력된 단어를 미리 이 정상 단어와 비교한 뒤에 정상 단어 사전에서 발견되지 않은 단어를 비속어 사전에서 검색하는 것이다. 정상 단어인 만큼 변형형에 대해서 크게 신경쓸 필요가 없으므로 Hash 등의 데이터 구조를 이용해 검색한다면 시스템에 크게 부하를 걸지 않고 specificity를 향상 시킬 수 있을 것이다.

3 복합 비속어 문제

한글은 단어와 단어를 조합해서 새로운 단어를 만들 수 있으며 이것은 비속어에도 그대로 적용된다. 가장 흔하게 쓰이는 "개새끼"라는 욕도 "개"라는 욕과 "새끼"라는 단어의 합성이다.

표 2. 복합 비속어의 예시

복합비속어	원래 단어	복합비속어	원래 단어
개새끼	개 + 새끼	씹새끼	씹 + 새끼
씨팔년	씨팔 + 년	갈보년	갈보 + 년
개씹좆	개 + 씹 + 좆	개자지	개 + 자지
개양아치	개 + 양아치	개자식	개 + 자식

표 2은 한국게임산업 진흥원의 게임 언어 건전화 지침서 연구[1]에서 발췌한 복합비속어 모음이다. 복합 비속어의 경우 비속어 + 비속어의 경우나 비속어 + 일반단어와 같은 비속어를 이용해 복합 비속어를 만든 경우에는 크게 문제가 생기지 않는다. 정상 단어 검증을 통해서 일반단어 부분이 떨어져 나간다 하더라도 남은 비속어 부분을 통해서 필터링 되기 때문이다. 그러나 정상단어 + 정상단어의 조합으로 생성되는 비속어의 경우 문제가 있다. "개자식"의 경우 "개", "자식" 양쪽 모두 따로 쓰일때는 정상단어로 사용되는 단어이다. 그러나 "개자식"으로 합쳐지면서 욕으로 사용되는 단어이다. 이것을 일반단어 검증을 통해 검증하면 각각 "개", "자식"으로 정상단어로 분류되어 비속어 필터링 시스템을 벗어나는 경우가 생겨버린다.

이것을 위한 해결책으로는 정상단어 + 정상단어로 형성되는 비속어를 정상단어 검증 전에 한번 필터링하는 방법이 있을 수 있다. 즉 비속어 검증 단계를 3단계로 나눠서 복합비속어, 일반단어, 변형 비속어의 순으로 필터링을 진행하는 것이다. 당연히 상위 단계가 하위 단계보다 우선시 된다. 복합 비속어 검증의 경우도 Hash를 사용하여 시스템에 가해지는 부하를 최소한으로 줄일 수 있다.

4 속도와 데이터 구조

본시스템은 입력단어와 비속어 간의 유사도를 파악하기 위해서 semi-global alignment score를 측정하는 방법을 사용한다. 단어가 완전히 일치하지 않아도 유사도를 측정하여 비속어인지 아닌지 판별할 수 있으므로 변형 비속어에 대한 대응책이 될 수 있으나 연산량이 많아서 시스템에 가하는 부하가 크므로 비교횟수를 최소화 해줄 필요가 있다.

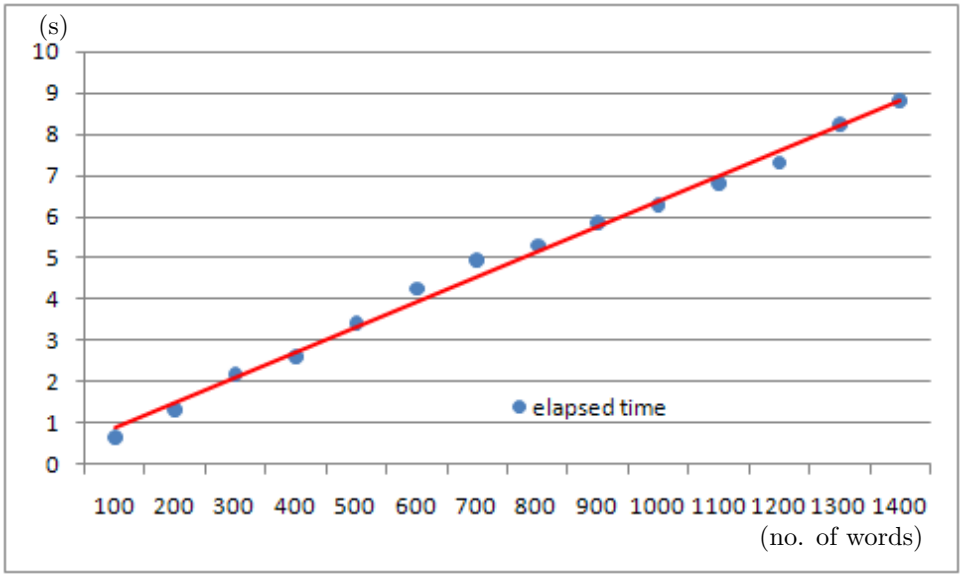

그림 2. 개발 중인 비속어 필터링 시스템의 입력단어 수에 따른 수행시간 그래프

그림 2은 현재 개발중인 비속어 필터링 시스템의 입력단어 수에 따른 수행시간 그래프이다. 100개의 단어를 처리하는데 0.7초 정도의 시간이 필요한데 일반적으로 채팅에는 한 문장에 10개 내외의 단어가 사용되기 때문에 개인 클라이언트에서 처리하기에는 크게 문제가 없다. 그러나 서버에서 채팅

내용에 대한 비속어 필터링을 수행할 경우 수천, 수만의 단어가 오가게 되는데 1400 단어를 처리하는데 9 초가량의 시간이 걸리는 현 시스템으로는 무리가 있다.

비속어 비교 문제에서 가장 큰 문제는 유사도가 sequencial 하지 않다는 것이다. "씹새끼"의 변형인 "띱때끼"의 경우 비속어 목록을 binary search 로 검색할 경우 해당 비속어를 찾을 수가 없다. 그렇다고 모든 비속어와 유사도를 비교하면 수천, 수만 단어와 비교가 이루어져야 하므로 지나치게 시간이 오래 걸리게 된다.

이러한 문제점을 해결하기 위해서 가장 적합한 자료구조는 Tree 구조라고 할 수 있다. 순서가 없더라도 유사한 비속어 끼리 무리를 지을 수 있기 때문에 각 무리를 나누어서 tree 의 노드로 사용할 수 있다. tree 구조를 사용하게 될 경우 비교 횟수도 비약적으로 줄여 줄 수 있으므로 시스템의 성능도 크게 향상될 수 있을 것이다. 그러나 이 경우 상위 노드로 선택할 대표 단어를 선정하는 방법과 노드의 단어 갯수, 노드의 분할 방법 등 생각해 봐야 할 문제가 많이 있다.

다른 방법으로 2 차원 공간에 단어를 배열하는 것이다. 선형적인 구조에서 벗어나 데이터를 배치할 수 있으므로 배치 문제도 해결될 수 있고 R-tree 와 같은 기존의 데이터 구조를 응용하여 사용할 수 있기 때문에 확실한 성능을 보장할 수 있다. 그러나 이경우 x, y 좌표를 어떠한 기준으로 배치할 것인가가 문제가 된다. 단순히 초성과 중성, 모음과 자음으로 나누는 방법이 있을 수 있으나 어느것도 최적화 된 방법이라는 근거가 부족하다.

5 사용자의 성향 분석을 통한 비속어 필터링

spam filtering 기법에는 email 주소와 서버의 신뢰도 측정을 통한 방법으로 blacklisting 과 graylisting 기법이 있다. 본 비속어 필터링 시스템의 장점은 threshold 값 조정을 통해서 얼마든지 비속어 필터링 레벨을 조절할 수 있다는 점이다. 이것을 이용한다면 사용자의 사용실태를 통한 성향 분석을 통해서 사용자의 비속어 필터링 레벨을 조정할 수 있을 것이다.

먼저 사용자의 언어 생활 평판 시스템이다. 사용자가 비속어를 사용해서 필터링 시스템에 적발이 되거나 다른 사용자의 고발을 통해서 해당 사용자의 비속어 필터링 레벨을 높일 수 있는 것이다. 필터링 레벨이 높아진다면 정상단어의 필터링 확률도 높아지므로 의사소통에 지장을 받게 된다. 단순히 시스템적인 필터링만이 아니라 사용자에게 피드백을 줌으로써 비속어 사용에 대한 경각심을 높일 수 있어서 자체 정화 작용에 도움을 줄 수 있다.

비속어가 필터링 시스템에 감지 된 직후의 대화에 대한 일시적으로 비속어 필터링 레벨을 높이는 방법이 있을 수 있다. 일반적으로 사용자가 비속어를 사용하고 비속어 필터링 시스템에 의해 방해를 받았을 경우 필터링 시스템을 피하기 위해서 다른 비속어나 비속어를 변형하여 입력하는 경우가 일반적이다. 이때 비속어 필터링 레벨을 일시적으로 높이는 방법을 통해서 이러한 우회방법을 저지할 수 있다.

비속어 필터링 레벨을 사용자가 속한 집단에 적용하는 방법도 있을 수 있다. 온라인 게임의 경우 사용자끼리 친목을 목적으로 집단을 형성하는 경우가 많고 시스템적으로도 지원하고 있다. 집단에는 유사한 성향을 가진 유저가 모이기 마련이다. 그러므로 비속어를 자주 사용하는 사람들은 서로 같은 집단을 형성할 가능성이 높다. 그러므로 해당 집단에도 언어생활에 대한 평판 시스템을 적용하여 구

성원이 비속어를 사용한다면 해당 집단에도 비속어 필터링 레벨을 높이는 제재를 가할 수 있다. 이 경우 구성원이 사용한 비속어가 해당 집단에 영향을 미치게 되므로 비속어 사용에 대한 책임의식을 더욱 높일 수 있다.

반대로 신뢰도가 높은 사용자에 대해서는 필터링 레벨을 낮추는 방법도 사용할 수 있다. 서로 친구로서 등록되어 있거나 같은 집단에 속하는 경우 서로간의 비속어 필터링 레벨을 낮추는 방법으로 원활한 의사소통을 수행할 수 있다. 친근한 관계에서는 서로 인사로 비속어를 사용하거나 빠른 의사소통을 위해 은어를 사용하는 경우가 많으므로 사용자의 편의성 향상에도 도움을 줄 수 있다.

더 나아가서 인터넷 실명제 등이 본격화 된다면 하나의 사이트나 게임에 국한 되지 않고 각 시스템간의 연계를 통하여 해당 사용자의 종합적인 인터넷 비속어 사용실태를 파악하여 온라인 언어 생활에 대한 등급을 메기는 방법이 있을 것이다. 이러한 종합적이고 체계적인 관리를 통하여 사용자가 내뱉은 말에 책임을 지게 함으로써 익명성으로 인한 언어 폭력의 문제를 해결할 수 있을 것이다.

6 결론

본 보고서에서는 개발 중인 비속어 필터링 시스템의 문제와 그 해결방안에 대해서 제안하였다. 한글의 특성을 이용한 변형 비속어에 대해서 효과적으로 대처하기 위한 방법으로 다음과 같은 문제점과 방안이 검토되었다.

1. 정상단어의 필터링 문제 : 본 시스템은 변형 비속어에 대응하기 위해 등록된 비속어와 입력된 단어간의 유사도를 이용해서 비속어 여부를 판단한다. 이 경우 기존 비속어 필터링 시스템의 정상단어 필터링 문제가 더욱 두드러지게 된다. 이것을 방지하기 위하여 비속어 필터링 시스템의 정확도를 높이는 방법과 정상단어 사전을 Hash로 등록하여 비속어 필터링 전에 정상단어를 미리 검증하는 방법을 제시하였다.
2. 복합 비속어 문제 : 한글은 단어와 단어의 조합을 통해서 새로운 단어를 만들어내는 경우가 많다. 비속어 또한 단어간의 조합을 통하여 새로운 비속어를 만들어 내는 경우가 많다. 이러한 복합 비속어 문제를 해결하기 위해서는 고성능의 parsing 알고리즘이 필요하고 더 불어서 정상단어 검증을 통해서 정상단어 + 정상단어의 조합으로 만들어진 복합 비속어를 정상단어 검증 보다 먼저 필터링 해 줄 필요가 있다. 이것 역시 Hash를 사용하면 시스템에 가하는 부하를 줄여줄 수 있다.
3. 속도와 데이터 구조 : semi-global alignment score를 측정하는 방법은 단어와 단어가 완전히 일치 하지 않더라도 그 유사도를 알아내서 비속어를 검증할 수 있다는 장점이 있다. 그러나 요구하는 연산량이 많으므로 횟수를 최소화 해줄 필요가 있다. 변형 비속어를 검증하기 위해서는 여러 단어와 유사도를 측정해 봐야 하는데 이 측정 횟수를 최소로 줄이기 위해서는 Tree구조 혹은 2차원 Map에 단어를 배열하는 방법 등이 있을 수 있으나 정확한 알고리즘은 차후 좀더 연구가 필요할 것이다.
4. 사용자 성향 분석을 통한 비속어 필터링 : 단순한 알고리즘적인 방법으로는 한글 비속어 사용을 원천 봉쇄하는데 한계가 있다. 사람이 만드는 시스템인 만큼 필터링을 벗어날 수 있는 허

점이 존재하기 때문이다. 그렇기 때문에 spam 필터링 시스템에서 사용되는 blacklisting 기법과 graylisting 기법 같은 사용자의 신뢰도를 이용하는 방법이 있을 수 있다. 사용자나 사용자가 속한 집단의 비속어 사용빈도나 사용 성향 등을 파악하여 해당 집단의 언어 생활 평판을 부여하는 방법이다. 비속어 사용을 남발할 수록 비속어 필터링 레벨을 높여서 온라인내 커뮤니 케이션이 어려워지므로 비속어 사용행위에 대한 책임을 물을 수 있어 비속어 사용에 대한 경각심을 높여 줄 수 있을 것이다.

온라인 커뮤니케이션에서 비속어 필터링 문제는 점점 부각되고 있으나 학술적인 연구 사례는 매우 적다. 활발한 연구가 이루어지고 있는 spam 필터링 과는 달리 실시간 처리가 이루어져야 한다는 문제점과 판단의 대상이 email에 비해서 비교적 짧은 문장이기 때문에 단어 단위의 판단이 이루어져야 한다는 어려움 때문일 것이다. 이러한 문제점을 해결하기 위해서는 시스템적인 해결 방법과 더불어서 사용자와 피드백을 주고 받아서 신종 비속어 및 복합 비속어 등에 대해서도 빠르게 대응할 수 있고 사용자에게 비속어 사용에 대한 경각심을 불러 일으킬 수 있는 장치의 개발이 무엇보다도 중요할 것이다.

참고 문헌

1. 한국게임산업진흥원, “게임언어 건전화 지침서 연구,” 2008.