적응적 비속어 필터링을 위한 비속어 사용성향 분석

Analysis of Profanity Using Tendency for Adaptive Profanity Filtering

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

ABSTRACT

서열 정렬 값을 이용한 유사도 측정방식의 변형 비속어 필터링은 사용자의 비속어 변형을 통한 필터링 우회를 막을 수 있는 효과적인 방법이다. 그러나 변형에 대한 필터링을 엄격히 적용하기 위해 적용되는 유사도 임계값을 낮게 설정할 경우 비속어와 유사한 여러 일반 단어들이 필터링 되어 사용자의 정상적인 의사소통을 방해하게 된다. 이 문제를 해결하기 위해서는 상황과 사용자에 따라 임계값 적용을 달리하는 적응적 필터링을 수행할 필요가 있다. 본 보고서에서는 이 적응적 필터링을 위하여 사용자의 비속어 사용 성향을 분석하여 등급을 설정하는 알고리즘에 대하여 제안하도록 한다.

KEYWORDS Profanity Filter, Adaptive Filtering, Alignment

1 서 론

우리는 지금까지 실용적인 비속어 필터를 개발하기 위하여 사용자가 입력한 변형 비속어와 데이터베이스 내의 금칙어 간의 서열정렬값 측정을 통한 유사도 분석 방식을 사용하여 필터링하는 기법을 개발하였다[1]. 그리고 근사문자열 검색 방법을 적용하여 실제 시스템에 적용이 가능한 속도를 확보하였다[2]. 그러나 이 유사도를 이용한 측정 방식의 경우 사용자의 정상적인 단어 입력을 변형 비속어로 인식할 위험성을 내포하고 있다.

표 1. 비속어와 유사한 정상단어

비속어	정상단어
개새끼	새끼줄, 소세키,
시발	시발점, 시바, 퍼시발
성기	성기사, 김성기, 성장기

위의 표 1는 비속어와 유사한 일반단어를 정리한 표이다. 이러한 정상단어들의 필터링을 막기 위해서는 유사도의 임계값을 높게 설정하여 일반단어의 필터링을 막을 수 있다. 반면에 유사도의 임계값을 높게 설정하게 되면 그만큼 비속어 필터링의 변형에 대한 대응이 약해지게 된다. 예를 들어 "

시발"과 "퍼시발"을 필터링 하지 않게 하기 위해 임계값을 설정하면 "시ㅅㅂ발" 이라는 변형 욕설을 필터링 할 수 없게 된다.

욕설을 필터링 하는 문제와 정상단어의 필터링은 서로 상충되는 문제이다. 이것을 막기 위해 우리는 사전에 필터링을 하지 않아도 될 단어를 일반 단어 리스트에 저장해서 미리 검증을 거치는 방식을 사용하지만 인터넷 용어는 정확한 맞춤법을 지키는 경우가 드물고 자신들만의 용어를 만들어서 사용하는 경우가 많아서 완벽한 대응은 불가능 하다고 할 수 있다.

비속어의 사용은 사람과 상황에 따라서 다르게 적용된다. 즉 비속어를 자주 사용해 왔던 사람은 앞으로도 비속어를 자주 사용할 가능성이 높은 것이다. 그러므로 우리는 이러한 비속어를 자주 사용하는 사용자들에게 좀더 엄격하게 비속어 필터링을 적용하고 비속어를 자주 사용하지 않은 우량 사용자에게는 비속어 필터링 강도를 낮게 적용하여 커뮤니케이션의 편의를 제공하는 방법을 사용할 수 있다. 이 방법은 비속어 필터링을 비속어 사용에 대한 제재 수단으로서 사용하여 사용자의 비속어 사용을 자제하게 하는 계몽의 수단으로서도 사용할 수 있는 것이다.

이 경우 사용자의 비속어 사용 성향을 분석할 수 있는 모델이 필요하다. 단순히 비속어 사용할 경우 임계값을 낮추고 시간이 지나면 임계값을 다시 높이는 방식을 사용할 수도 있으나 이러한 방식은 다양한 커뮤니케이션 상황에 모두 대응하기 어렵다. 그리고 비속어 사용 검출이 단순히 필터링 시스템을 통해서만 이루어지지 않고 사용자의 신고나 운영자의 모니터링을 통해서도 검출 되기 때문에 여러 상황에 대해서 적용을 달리 하여야 할 것이다. 그리고 비속어로 사용되는 단어의 종류에 따라서도 비속어 사용 성향을 분석에 차이를 둬야 할 것이다. 단순히 장난으로 넘길 수 있는 비속어도 있는 반면에 보는 이의 마음에 깊은 상처를 줄 수 있는 비속어 또한 존재하기 때문이다.

본 보고서에서는 여러가지 상황에 대처할 수 있는 비속어 사용성향 분석 시스템에 대하여 제안하고자 한다[3]. 이 시스템은 사용자의 비속어 사용성향을 단기적, 장기적 관점에서 분석하며 사용되는 비속어의 등급에 따라 차등적인 적용을 하고 또한 이 등급 시스템을 커뮤니케이션의 상황에 따라 적절하게 적용할 수 있는 방법에 대하여 설명하도록 하겠다.

2 사용자 적응적 필터링을 위한 사용자 비속어 사용 성향의 측정 척도

사용자의 비속어 사용 성향 분석에서 가장 기본은 비속어를 많이 사용하는 사용자의 필터링 임계값을 낮춰 많은 단어가 필터링 되게 하고 적게 사용하는 사용자의 임계값을 높여 커뮤니케이션의 편의성을 해치지 않는 일이다. 이것을 위해서는 사용자의 비속어가 검출되는 횟수를 측정하여 그에 맞게 비속어 사용 등급을 높이는 방식을 사용하게 된다.

이때 비속어 사용 횟수에 따른 비속어 사용 등급의 상승폭을 어떻게 설정하는가가 문제가 된다. 너무 빠르게 올릴 경우 비속어를 얼마 사용하지 않아도 등급이 너무 빠르게 상승하여 커뮤니케이션의

편의성을 해치게 되고 너무 천천히 올릴 경우 시스템의 의의가 줄어들게 된다. 이때 고려해야할 사항은 두가지가 있다. 첫째, 사용자간 분쟁상황에 대한 대처이다. 분쟁이 일어날 경우 비속어 사용이 증가하는 것은 필연적이며 평소에 비속어를 사용하지 않는 사용자라 할지라도 분쟁상황에서는 비속어를 남발하게 될 가능성이 있다. 둘째, 지속적으로 비속어를 사용하는 사용자에 대한 대처이다. 비속어를 꾸준히 사용해왔던 사람은 당연히 앞으로도 비속어를 자주 사용할 것이다.

이 두가지 상황은 서로 다른 대처법이 필요하다. 분쟁상황의 경우 일시적으로 비속어의 사용 빈도가 높아지는 상황이다. 그러므로 비속어 사용등급의 변화가 빠르게 일어나야 할 필요가 있다. 반면에 후자의 경우 사용자의 장기적인 언어 사용을 분석하여 적용해야 하므로 사용 등급을 완만하게 변화시켜야 할 필요가 있다. 이러한 두가지 경우를 모두 만족시키기 위하여 우리는 비속어 사용 등급을 이원화하여 적용할 필요가 있다.

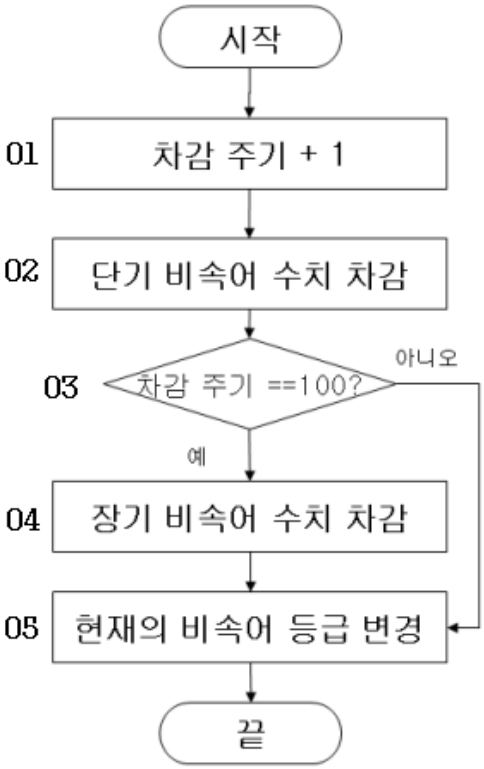

그림 1. 비속어 사용등급의 장기적 수치와 단기적 수치의 차감 알고리즘 - 본 예시에서는 장기적비속어와 단기적 비속어 차감주기가 100배 차이나게 설정하였다. 01 - 설정한 주기에 맞춰 차감주기가 증가하고, 02 - 차감주기 1 증가할때마다 단기비속어 수치는 차감되며, 03 - 차감주기가 100이 되었는지 검사후, 04 - 100에 도달하면 장기 비속어 수치를 차감한다, 05 - 변화된 단기, 장기 비속어 수치를 종합하여 유저의 비속어 등급을 변경한다.

Algorithm 1 비속어 사용등급의 장기적 수치와 단기적 수치의 차감 알고리즘

Input: M_L(장기 비속어 수치), M_S(단기 비속어 수치), T(차감주기), A(비속어 등급 가중치)
Output: R(비속어 등급)

$M_L \leftarrow$ 현재의 장기 비속어 수치
$M_S \leftarrow$ 현재의 단기 비속어 수치
Function Decrease(M_L, M_S)
  $T \leftarrow 0$
  T += 1
  M_S -= 1
  if T == 100 then
    M_L -= 1
  $R \leftarrow (M_L + M_S)/A$
  return R

먼저 단기적인 비속어 사용 등급이다. 사람은 분쟁이 발생하면 비속어를 사용하게 되고 해당 비속어가 필터링되게 되면 변형을 시도하면서 필터링 시스템을 우회하여 비속어를 사용하려고 하는 경향이 있다. 그러므로 비속어가 사용될 경우 단기간에 비속어 사용등급을 높여야 할 필요성이 있다. 그리고 이 등급은 역시 단기간에 감소되어야 할 것이다. 분쟁 상황동안 비속어 사용을 막는 것이 목적이기 때문이다.

장기적인 수치는 사용자의 비속어 사용에 따라 천천히 증가하고 사용하지 않는 기간에 따라 천천히 감소하게 되는 수치이다. 이 수치가 높아질 경우 일반적인 상황에서의 커뮤니케이션에도 문제가 생길 가능성이 높아지므로 불량 사용자에 대한 제재의 의미도 가지게 된다. 이 수치는 필터링으로 인한 비속어 검출 뿐만 아니라 사용자의 신고 혹은 운영자의 모니터링에 의해서도 증가하게 될 것이다. 사용자의 신고나 운영자의 모니터링으로 인한 검출은 단기적인 비속어 사용 수치를 증가시키기에는 적합하지 못하기 때문이다.

사용자의 최종적인 비속어 사용 등급은 이 두가지 등급의 합으로 계산되게 된다. 예를 들어 장기적인 비속어 사용 등급이 낮은 사용자는 한두번의 비속어를 사용하더라도 커뮤니케이션의 문제가 생길 만큼 종합 등급이 높아지지 않을 것이나 장기적인 비속어 사용등급이 높은 사람은 한두번의 비속어 사용으로도 커뮤니케이션에 문제가 생길 것이며 증가한 비속어 사용등급을 낮추는데에도 오랜 시간이 걸리게 될 것이다.

3 비속어의 모욕감 정도에 따른 차등적 수치 증감 적용

앞서 우리는 비속어 사용 검출을 통해 이원화된 비속어 사용등급을 관리하는 방법에 대하여 설명하였다. 비속어 사용이 검출될 경우 비속어 사용 등급이 상승하게 된다. 그러나 비속어는 종류에 따라서 장난으로 넘겨들을 수 있는 비속어가 있는가 하면 보는이로 하여금 깊은 마음의 상처를 줄 수 있는 비속어도 있다. 단순히 비속어가 검출되었다 하더라도 어떤 비속어를 사용하였느냐에 따라 죄의

깊이가 달라지는 것이다.

현실적인 비속어 사용등급 조정을 위해서는 비속어의 정도에 따라서 가중치를 달리 줘야할 필요성이 있다. 이를 위해서는 비속어에 따라서 사용자가 느끼는 불쾌감을 객관적으로 입증할 수 있는 자료가 필요하다. 일반적으로는 사용자의 설문을 통해서 가중치를 정하는 방식을 사용할 수 있을 것이다. 상당한 노동력과 시간이 필요한 작업이지만 가장 객관적이고 검증된 방법이라 할 수 있다. 또다른 방법으로는 사용자의 신고 빈도를 이용하는 것이다. 사용자의 신고를 통해 비속어가 검출되었을 경우 사용자가 충분히 불쾌감을 느낀 단어라고 할 수 있다. 이러한 경우가 빈번히 발생하는 단어의 경우 가중치를 자동적으로 높여줄 필요가 있다고 할 수 있다.

비속어 필터를 통해 자동적으로 검출될 경우 비속어의 유사도 또한 하나의 척도가 될 수 있다. 서열정렬을 통한 유사도 검색을 통해 비속어를 필터링 하는 만큼 유사도가 낮을 수록 잘못 필터링되었을 가능성이 높다고 할 수 있다. 그러므로 검출된 비속어의 유사도에 비례하여 가산 수치를 조절해 줘야할 필요성이 있다. 하나의 비속어 입력에 따른 증가 수치를 계산하는 알고리즘은 다음과 같다.

Algorithm 2 문장 입력에 따른 비속어 사용 수치 증가 알고리즘

Input: S_i(입력문장), w_i(문장의 각 단어), A(가중치 검색 함수), S(유사도 측정 함수), D(비속어 데이터 베이스), R(최종 비속어 사용치 함수)
Output:M(최종 비속어 사용 가산치)

$S_i =< w_{i0}, w_{i1}, ..., w_{in} >$
$D_i =< d_{i0}, d_{i1}, ..., d_{im} >$
Function $S(w_i, D)$
  $s = max(sim(w_i, d_{i0}), sim(w_i, d_{i1})...sim(w_i, d_{im}))$
  return s
$d_i \leftarrow$ 유사도 측정시 최대 값을 보인 금칙어
Function $A(d_i)$
  return d_i 의 가중치
Function $R(s_i)$
  M = 0
  For k = 1 to n
    M += $S(w_{ik}, D) \times A(d_i)$
  return M
end procedure

예를 들어 "0이런 씨0이발년" 이라는 문장을 입력했다고 하자. 여기서 우리는 "0이런" 이라는 단어는 유사도를 보이는 단어가 없으므로 가산되는 비속어 사잉 수치는 0 이다. "씨0이발년" 의 경우 "씨발년" 과 가장 높은 유사도를 보이게 되는데 이때 상대 유사도는 80% 이다. "씨발년" 은 높은 불쾌감을 주는 단어로 일반적인 비속어에 비해 높은 3의 가중치를 주게 된다. 그러므로 이 문장은 최종적으로 2.4의 비속어 수치를 더하게 되는 것이다. 이 방법은 단순히 모든 비속어를 같은 수치로 가산시키는 방법에 비하여 사람들에게 좀더 납득이 가는 비속어 사용 등급 관리 시스템을 제공할 수 있을 것이다.

4 비속어 사용 상황에 따른 비속어 사용수치 증감

사용자의 비속어 사용 등급에 따른 비속어 필터의 차등적 적용은 사용자의 비속어 사용 성향에 맞는 필터링을 제공하여 우량 사용자의 커뮤니케이션 문제를 최소한으로 줄이는데 그 의의가 있다. 그러나 사람에 따라서 적용을 달리하는 것 뿐만 아니라 커뮤니케이션의 상황에 따라서도 차등적인 적용이 필요하다.

예를 들어 사람들은 친한 친구 사이에서 격의 없는 대화를 나눌때에 비속어를 섞어서 대화하는 경우가 자주 있다. 비슷한 경우로 온라인 게임이나 채팅 시스템에서 서로 친구로 등록된 사람들 간에 대화를 할 경우 비속어 필터링을 완화해주는 것은 물론 비속어 사용등급 수치를 증가하는 것도 줄여주거나 무시할 필요성이 있다. 지인과의 대화는 온라인 커뮤니케이션에서 상당부분을 차지하는 영역이므로 해당 부분의 커뮤니케이션을 쾌적하게 해주는 것으로 전체적인 편의성이 증가하는 효과를 볼 수 있다.

반면에 비속어 필터링을 좀더 엄격하게 적용해야 할 부분도 존재한다. 예를 들어 게시판 글의 제목 같은 경우 사용자가 의도하지 않은 상태에서 내용을 접하게 되는 경우가 많고 해당 부분에 비속어가 포함되어 있을 경우 피해가 커질 수 있다. 그러므로 제목 부분에 대한 비속어 필터는 엄격하게 적용하고 비속어 사용으로 인한 사용 수치 증가도 가산해 줄 필요가 있다. 또한 게임 도중의 채팅의 경우 비속어 필터링을 엄격하게 적용해줄 필요성이 있다. 게임을 통한 경쟁의 도중 사람은 흥분할 가능성이 높으며 비속어 사용 빈도 역시 증가할 가능성이 높기 때문이다.

온라인 커뮤니케이션에서는 여러 상황이 있을 수 있으며 각 상황에 맞는 차등적 적용을 통해서 비속어 필터링을 통해 발생할 수 있는 여러 문제점을 완화시키는데 도움을 줄 수 있을 것이다. 이것을 위해서 해당 시스템의 각 상황에 맞는 모니터링을 통하여 상황에 따른 비속어 사용 성향을 파악하는 것이 중요하다.

5 결론

본 보고서는 비속어 필터링의 사용자와 상황에 따른 적응적 적용 방법에 대하여 기술하였다. 비속어 필터링 시스템은 필연적으로 정상단어를 필터링 할 가능성을 내포하고 있으며 이 문제를 해결하기 위해서는 여러 상황을 면밀하게 분석하여 각 상황에 맞는 적응적 필터링을 적용하는 것이 중요하다고 할 수 있다.

우리는 사용자의 비속어 사용등급을 장기적 단기적 수치로 이원화 하여 사용자가 처한 상황에 따라 알맞게 적용할 수 있는 방법을 제안하였으며 더불어 커뮤니케이션의 상황에 따라 차등적 적용을 해야할 필요성에 대해서도 설명하였다. 차후 실제 시스템의 적용과 사용자의 피드백을 통하여 더 구체적인 시스템을 구현할 수 있도록 하겠다.

참고 문헌

1. 한국게임산업진흥원, “게임언어 건전화 지침서 연구,” 2008.
2. Gonzalo Navarro and Edgar Chávez, “A metric index for approximate string matching,” *Theor. Comput. Sci.*, vol. 352, no. 1, pp. 266–279, 2006.
3. Ramachandran A Feamster N and Vempala S, “Filtering spam with behavioral blacklisting,” *In Proceedings of the 14th ACM Conference on Computer and Communications Security (Alexandria, Virginia)*, pp. 342–351, 2001.
4. 조환규 윤태진, “반 전역 정렬을 이용한 온라인 게임 변형 욕설 필터링 시스템,” vol. 9, no. 12, pp. 113–120, 2009.
5. 조환규 윤태진, “한글 자소정렬을 이용한 온라인 욕설 필터링 시스템,” vol. 36, no. No2(C), pp. 194–198, 2009.
6. 조환규 윤태진, “변형 비속어 필터링을 위한 비속어 필터링 시스템 및 방법,” 2009.