반 전역 정렬을 이용한 온라인 게임 변형 욕설 필터링 시스템

The Online Game Coined Profanity Filtering System by using Semi-Global Alignment

윤태진, 조환규
부산대학교 공과대학 정보컴퓨터공학부

Taijin Yoon(ytj@pusan.ac.kr), Hwan-Gue Cho(hgcho@pusan.ac.kr)

요약

온라인 게임에서의 언어폭력 문제는 매우 심각하지만 그에 대한 효과적인 정책이나 기술적인 방법은 부족한 상황이다. 온라인 게임 서비스 업체에서는 금칙어 리스트를 작성하여 Swear Filter를 이용한 고정된 형식의 문자열 검색 방식을 통해 문제를 해결하려고 하고 있으나 사용자들은 다양한 방법으로 욕설을 조합 또는 변형시켜 기존의 필터링을 회피하고 있다. 특히 한글은 욕설의 변형이 매우 쉬운 특성을 가지고 있다. 본 논문에는 한글에 기초한 변형 욕설을 효율적으로 탐색하여 걸러내는 알고리즘을 제시한다. 이 알고리즘의 주된 특징은 변형 욕설의 표준형 변환과 자소단위의 반 전체 정렬(semi-global alignment), 이다. 실험 결과 저자들이 다양한 인터넷 게임 환경에서 직접 수집한 다종의 욕설 단어들에 대하여 약 90%의 우수한 필터링 성능을 보였다.

■ 중심어 : | 욕설 | 한글 처리 | 정보 검색 | Alignment |

Abstract

Currently the verbal abuse in text message over on-line game is so serious. However we do not have any effective policy or technical tools yet. Till now in order to cope with this problem, the online game service providers have accumulated a set of forbidden words and applied this list on the textual word used in on-line game, which is called 'Swear filter'. But young on-line game players easily avoid this filtering method by coining another words which is not kept in the list. Especially Korean is very easy to make new variations of a vulgar word. In this paper, we propose one smart filtering algorithm to identify newly coined profanities. Important features of our method include the canonical form transformation of coined profanities, semi-global alignment between in the level of consonant and vowel units. For experiment, we have collected more than 1000 newly coined vulgar words in on-line gaming sites and tested these word against our methods. where our system have successfully filtered more than 90% of those newly coined vulgar words.

■ keyword : | Profanity | Korean Language Processing | Information Retrieval | Alignment |

I. 서론

정보 통신 기술의 발달로 인터넷을 통한 커뮤니케이션은 우리 일상에 중요한 한 부분으로 자리 잡았다. 그러나 온라인의 익명성을 이용한 언어폭력은 날이 갈수록 심해져서 사용자에게 심리적 상처를 입히고, 인터넷

* 본 연구는 한국학술진흥재단 연구과제(과제번호 : 20090291000)로 수행되었습니다.
접수번호 : #090717-002
접수일자 : 2009년 07월 17일
심사완료일 : 2009년 10월 30일
교신저자 : 조환규, e-mail : hgcho@pusan.ac.kr

환경을 어지럽힐 뿐만 아니라 궁극적으로는 올바른 국어사용을 저해하고 있다[1]. 하지만 그에 대한 대책은 미비하여 사용자의 신고에 의한 사후 처리 방식을 취하고 있어 언어폭력을 사전에 방지할 수 있는 시스템의 마련은 시급히 해결해야할 문제이다.

온라인상의 언어폭력을 가장 확실하게 예방하는 방법은 욕설을 원천적으로 사용하지 못 하도록 입력 자체를 막는 것이라 할 수 있다. 이미 온라인 게임이나 인터넷 BBS 등에서는 이러한 욕설 필터링 시스템을 도입해서 사용하는 경우가 많으나 실제로 욕설 사용을 막아내는 경우는 미미하다. 그 원인은 변형 욕설의 사용에 있다.

일반적인 욕설을 금칙어 리스트에 넣어서 필터링 하더라도 그 욕설을 약간의 변형을 통해서 의미는 유지하면서 필터링 시스템을 통과할 수 있기 때문이다. 특히 한글은 여러 자소 단위의 하위문자를 조합하여 하나의 글자를 만드는 형태이기 때문에 이러한 변형이 매우 자유롭게 이루어져서 문제가 더욱 심각하다고 할 수 있다. 예를 들어 “자식”, “자씩”, “쨔식”, “자쉭”, “개짜식”의 경우, 일반인들은 모두 같은 의미로 욕설로 받아들일 수 있으나 컴퓨터는 전혀 다른 문자열의 다른 단어로 인식하게 된다.

변형 욕설을 필터링 하지 못하는 문제보다 더 심각한 문제가 욕설 필터가 일반단어를 필터링 하는 문제이다. 일반 단어가 의도하지 않게 필터링 되는 문제를 Scunthorpe Problem 이라고 부른다. 그 이유는 Scunthorpe라는 마을이 여성의 신체의 일부를 의미하는 cunt라는 단어 때문에 인터넷에서 필터링된 것이 그 어원이다. 이와 비슷한 문제는 한글 필터링에서도 자주 발생한다[6]. 예를 들어 “자위”라는 단어를 금칙어에 추가해서 필터링하는 경우 “상자위에”라는 단어에서 “자위”를 필터링 하여 “상##에” 라는 의미 불명의 단어를 출력하게 된다. 이 문제는 욕설 필터링의 민감도를 높이는데 어려움을 주게 된다. 욕설 필터링 시스템이 정상적인 대화를 방해하여 사용자에게 불편함을 주게 된다면 서비스 제공자의 입장에서는 욕설 필터링 시스템을 도입하는 것을 꺼리게 되기 때문이다. 이 문제는 이미 인터넷 게임에서의 언어사용에 관한 연구에서 많은 연구자들이 지적하고 있는 문제로서 어간과 어미의 결합이 다양한 한글의 경우에 유독 심각한 문제이다[1].

상기의 두 가지 문제는 서로 상충되는 문제로 단순히 금칙어 리스트를 확충하는 방식으로는 근본적인 해결이 이루어지지 않는다. 많은 금칙어가 등록될 수 록 필터링 되는 일반단어는 많아지게 되고 반면에 사용자는 간단한 변형만으로 필터링을 무력화하여 욕설을 자유롭게 사용할 수 있기 때문이다. 본 논문에서는 단어 단위의 필터링을 중심으로 이 두 가지 문제를 해결하기 위하여 욕설 표준형 변환 기법을 이용한 반 전역 정렬 기법과 정상 단어 검증을 기초로 하는 다단계 필터링 방식을 제안하고자 한다.

II. 욕설 필터링에 관련된 이전 연구

욕설 필터링 시스템은 이미 여러 분야에서 실제 사용되고 있다. 그러나 서론에서 언급한 문제를 완전히 해결하지는 못하여서 성능이 높다고는 할 수 없다. 한국게임산업진흥원에서는 2009년 봄에 “게임언어 건전화 지침서 연구”를 통해서 대표형 2,308항, 총 8,508개의 금칙어 리스트를 작성하여 배포하였으나 위의 두 가지 문제를 이유로 네티즌들의 반응은 부정적이었다[1]. 이 연구의 의미는 실제 사용되는 인터넷 욕설을 총망라했다는 점에서 큰 의미가 있으나 인터넷 욕설의 진화속도를 볼 때 새로운 욕설은 매우 빠르게 생산되고 있어 위와 같은 종합적인 욕설탐색 작업이 그 속도를 따라잡기에는 여러 어려움이 있다.

일반적으로 온라인 게임 채팅 시스템에 사용되는 욕설 필터링 시스템은 Swear Filter[8]가 쓰인다. 입력된 문장을 검색해서 부적합한 단어가 검색되면 입력을 거부하거나 해당 단어를 다른 단어로 변형 시켜서 출력하는 방법이다. 간단하고 효율적인 방법이나 변형 욕설을 필터링 하지 못하고 Scunthorpe Problem에서 자유롭지 못하여 실용성은 그렇게 높다고 하기 어렵고 오히려 사용자들의 불만을 사고 있다. 국내 주요 온라인 게임포탈에서도 이 방법을 응용한 자체적인 욕설 필터링 기법을 사용하고 있으나 만족할 만한 성능을 보이고 있지 않다[7]. Shekhar Dhupelia가 언급했듯이 변형 욕설에

대해서는 매우 취약한 방식이기 때문이다[9].

Spam Filtering 분야에서는 문서에 포함된 단어를 이용한 문서 분류 기법을 사용하여 Spam을 걸러낸다. 주로 Naive Bayes, SVM, K-nearest neighbor 등의 기법이 사용되는데 Machine Learning 방법을 통해서 Spam에 주로 사용되는 단어를 학습시킨 후 그 데이터를 이용하여 sSpam을 걸러내는 방식을 사용한다[2]. 한국에서는 "SVM을 이용한 온라인게임 비속어 필터링 시스템"에서 이러한 방식을 사용한 비속어 필터링 시스템을 제안하였다[3]. 그 외에도 e-mail 주소나 서버의 신뢰도를 이용한 spam filter 방식도 연구되고 있다[5].

"String Alignment는 문자서열 간의 유사성을 비교하는데 아주 유용한 방법으로 특히 DNA 서열의 전체 유사성, 국부 유사성을 찾는데 사용되고 있다. 우리는 이 방법을 원용하여 한글단어를 자소단위로 풀어 나열하고 그것의 주어진 단어의 DNA 서열과 같이 간주하여 그 자소단위의 스트링을 alignment하여 그들 간의 상호 유사성을 찾아내고자 한다.

자연어 처리기법을 이용한 비속어 필터링 방식도 생각해 볼만하다. 비유적인 내용을 사용한 문장의 저속한 의미를 형태소 분석 등을 통하여 파악하는데 도움을 줄 수 있다. 그러나 온라인 용어의 특성상 문법과 맞춤법이 지켜지지 않고 특히 욕설을 사용할 때 단어 그 자체를 극단적으로 변형시켜서 사용하는 경우가 많기 때문에 일반적인 자연어 처리 기법으로는 이러한 변형 단어와 문법에 어려움을 겪게 된다. (주)아이모션에서 출원한 특허 "음절결합 정보를 이용한 음란/비속어 차단시스템"에서 자연어 처리기를 기반으로 한 비속어 처리기를 제안하였으나 이 역시 중심음절 그 자체가 변형되어 버리면 올바른 동작을 보장할 수 없게 된다[10].

Ⅲ. 변형 욕설 필터링

1. 표준화를 통한 매칭 기법

한글의 경우 조합 욕설은 수많은 변형 형태를 지닐 수 있고 그러한 변형된 형태로 사용될 경우 단순한 문자열 대조로는 검출해내기 어렵다. 가능한 변형 형태를 모두 데이터베이스화 하는 방법이 있을 수 있겠으나 어디까지나 부가적인 모듈이 되는 욕설 필터링 기능이 지나치게 시스템이 큰 부하를 가하게 되면 시스템의 성능을 저하시켜 사용자에게 불편을 끼치게 된다. 그리고 단순히 데이터베이스를 확충하는 방식으로는 Scunthorpe Problem을 해결할 수 없다. 그러나 다양한 변형형태를 지니고 있다고는 하나 변형된 형태를 살펴보면 어느 정도 일관된 규칙을 찾을 수 있고 이 규칙을 이용한다면 간단한 알고리즘을 도입하여 변형 욕설 문제에 대한 해결책을 제시할 수 있다.

표 1. 한글 발음을 이용한 변형된 욕설의 다양한 예

기본 단어	변형 단어
개새끼	개새기, 개새뀌, 개색기, 개색휘, 개색히, 개세이, 개쉐리, 개쉐이, 개색
개놈	개늄, 개눔, 개늠, 개넘, 개너므, 게 넘, 게넘, 게놈, 게늄, 게놈, 게놈, 궤놈
병신	뱅쉰, 뱅쉰, 벵신, 벵틴, 병시인, 퓡신, 퓡 쉰, 븅신, 븅쉰, 비웅신
씨팔	쉬발, 쉬밸, 쉬뱅, 쉬벌, 쉬벧, 쉬볼, 쉬불, 쉬빌, 쉬파, 쉬팍, 쉬팔, 쉬팕, 쉬팡
불알	붕알, 부랄, 부럴, 브랄, 브라알, 블알, 뽕알, 뽕알

먼저 가장 일반적으로 사용되는 변형형태는 비슷한 발음을 이용하여 유사한 발음의 단어로 변화시키는 것이다. 예를 들어 욕으로 많이 사용되는 "개"라는 단어의 경우 "걔", "궤", "캐", "컈" 등의 발음으로 변형되어 사용되는 경우가 많다. 자음의 경우 된소리, 센소리 등으로 변화시키는 경우가 많고 모음의 경우 유사한 발음군으로 변화시켜 사용하게 된다. 이렇게 변화로 주로 사용되는 자소들 끼리 하나의 무리로 통합하여 대표되는 발음을 선정, 해당 자소로 표준화시켜서 그 표준형태를 검색을 위한 색인으로 사용한다면 발음을 이용한 변형형태를 사용하더라도 원래 형태의 욕설을 찾아 낼 수 있다.

[표 2]는 본 시스템에서 사용되는 표준형 변환 규칙을 나타낸 표이다. 자음의 경우 된소리, 센소리를 기본 발음으로 통합하였으며 모음의 경우 비슷한 발음이라 생각되는 것들을 대표되는 하나로 통합하였다. 이 표준형 변환만으로도 인터넷에서 사용되는 욕설 변형형의

상당수가 검색 가능해져서 필터링이 가능하게 되는 것을 알 수 있다.

표 2. 욕설 표준형 변환을 위한 변환규칙

초 성			
원래 문자	표준화 문자	원래 문자	표준화 문자
ㄱ, ㄲ, ㅋ	ㄱ	ㅅ, ㅆ	ㅅ
ㄷ, ㄸ, ㅌ	ㄷ	ㅈ, ㅉ, ㅊ	ㅈ
ㅂ, ㅃ, ㅍ	ㅂ		
중 성			
원래 문자	표준화 문자	원래 문자	표준화 문자
ㅏ, ㅑ	ㅏ	ㅗ, ㅛ	ㅗ
ㅐ, ㅒ, ㅔ, ㅖ, ㅙ, ㅚ, ㅞ, ㅞ	ㅐ	ㅜ, ㅠ	ㅜ
ㅓ, ㅕ	ㅓ	ㅟ, ㅡ, ㅣ	ㅣ
종 성			
원래 문자	표준화 문자	원래 문자	표준화 문자
ㄱ, ㄲ, ㅋ	ㄱ	ㅂ, ㅍ	ㅂ
ㄷ, ㅌ, ㅅ, ㅈ, ㅉ, ㅊ	ㄷ		

다른 변형 방법으로는 욕설의 글자사이에 무의미한 빈칸이나 기호 등을 포함시켜 필터링을 피하는 방법이 있다. “멍청이”를 “멍 청 이”나 “멍,청,이”로 쓰게 된다면 사람은 같은 단어라는 것을 판단할 수 있으나 컴퓨터의 경우 다른 단어로 받아들여 필터링을 피해갈 수 있는 것이다. 일반적으로 많이 쓰이는 Swear Filter의 경우 빈칸과 Non-alphanumeric 문자를 제거하고 필터링하는 방법을 사용하나 이모티콘을 이용한 욕설이나 특수문자를 이용하는 변형형도 있기 때문에 주의를 요한다. 그래서 본 시스템에서는 특수문자와 빈칸을 제거한 문자열과 제거하지 않은 문자열 두 가지 모두 검사하는 방식을 사용한다.

본 시스템에서는 무의미한 문자를 추가하는 방식의 변형에 대해서 좀 더 탄력적으로 대처하기 위하여 자음, 모음을 분리한 형태로 검사를 수행한다. 예를 들어 “깍쟁이”를 “깍쟿잉”으로 변형시켜 입력할 경우 Semi-global Alignment 수치가 0으로 측정되게 된다. 그러나 “ㄲㅏㄱㅈㅐㅇㅣ”와 “ㄲㅏㅋㅈㅐㅎㅇㅣㅇ”의 경우 받침만 변경되어 mismatch를 만들어낼 뿐이므로 충분한 유사도가 측정될 수 있다. 욕설의 표준화 작업에도 이러한 자모 분리를 사용하는 방법이 유리하다.

표 3. 외래어와 특수문자를 이용한 변형 예

기본 단어	변형 단어
씨팔	cval, 씨팔, 씨발, ㅆ1바, 씨발
개새끼	dog새끼, dog새
니미	ㄴ1ㅁ1, ㄴ1미
뒈진다	D질래
게이	gay, g@y
니에미	ㄴ1ㅇㅐ미, ㄴ1에미, 니ㅐ미
망할년	亡할년, 望할년
미친	ㅁ1친, 미친, ㅁ| 친

외래어와 특수문자를 이용한 욕설 변환의 경우 변환 방법이 매우 다양하고 사용되는 문자가 광범위 하여 매우 까다로운 문제라 할 수 있다. 주로 사용되는 것은 영어와 shift-number 문자를 이용하는 방식으로 손쉽게 사용할 수 있어 사용 빈도가 높다. 한자나 외국 문자를 이용하는 방법은 사용이 번거롭기는 하나 매크로 등에 미리 저장해두고 사용하는 방식도 있어 무시할 수 없는 문제이다.

표 4. 외래어 및 특수문자를 이용한 변환규칙

자 음			
원래 문자	표준화 문자	원래 문자	표준화 문자
g, k, c	ㄱ	b, v, p, f	ㅂ
n, L	ㄴ	s, A	ㅅ
d, t, E	ㄷ	o, w	ㅇ
l, r	ㄹ	j, g, z	ㅈ
m, 口	ㅁ	h	ㅎ
모 음			
원래 문자	표준화 문자	원래 문자	표준화 문자
a	ㅏ	u	ㅜ
@, H	ㅐ	l, i, y, !, ㅣ	ㅣ

특수문자의 경우 무의미하게 욕설 사이에 추가된 것인지 변형 욕설을 위해 사용된 것인지 판단이 어렵다. 그러므로 신뢰성을 높이기 위해서는 특수문자를 제거하고 필터링 한 것과 포함하고 필터링하는 두 번에 걸친 작업이 필요하다.

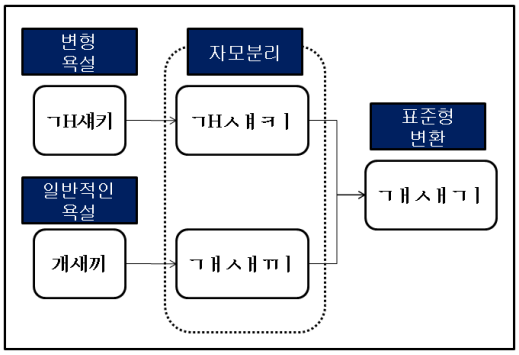

그림 1. 표준형(canonical form) 변환

[그림 1]은 욕설 표준형 변환 알고리즘을 다이어그램으로 표현한 것이다. 변형 욕설인 "ㄱH새키"가 일반 욕설인 "개새끼"와 동일한 형태로 변환되는 것을 알 수 있다. 이 변환 알고리즘을 이용하면 다양한 변형형태들을 간단한 형태로 통일할 수 있어 검색 알고리즘의 효율성을 증대할 수 있다. 이것은 단어를 발음을 기준으로 정렬하는 phonetic algorithm과 연관하여 유사한 단어 검색의 효율성을 증대 시킬 수 있을 것이다[4].

표 5. 변형 욕설의 표준형 변환 예

변형욕설		욕설의 표준형
1	ㄱHㅅHㄲㅣ	ㄱㅐㅅㅐㄱㅣ
	깨쉐끼	ㄱㅐㅅㅐㄱㅣ
2	미친년	ㅁㅣㅈㅣㄴㅕㄴ
	미췬년	ㅁㅣㅈㅣㄴㅕㄴ
	ㅁ1친년	ㅁㅣㅈㅣㄴㅕㄴ
3	씨팔	ㅅㅣㅂㅏㄹ
	쉬팔	ㅅㅣㅂㅏㄹ
	쉬빨	ㅅㅣㅂㅏㄹ
4	썅년	ㅅㅏㅇㄴㅕㄴ
	쐉ㄴㅕㄴ	ㅅㅏㅇㄴㅕㄴ
5	자지	ㅈㅏㅈㅣ
	자ji	ㅈㅏㅈㅣ

[표 5]는 여러 욕설의 표준형 변환 예를 보여주고 있다. 주로 쓰이는 다양한 변형 욕설들이나 모두 같은 형태의 표준형으로 변환되는 것을 볼 수 있다.

2. Semi-Global Alignment를 이용한 유사도 측정

표준형 변환 기법을 통해서 우리는 입력 단어와 가장 유사한 욕설을 찾아내었다. 그리고 입력 변형 욕설이 우리의 데이터베이스에 있는 욕설인지를 판별하기 위해서 semi-global alignment를 이용한 패턴 매칭 방법을 통해 단어 간의 유사도를 분석하여 욕설 여부를 판정한다.

semi-global alignment는 global alignment와는 달리 입력 단어의 전체와 비교 대상 단어의 일부를 사용해서 alignment를 수행한다. 그러므로 semi-global alignment는 유사한 욕설을 찾는데 우수한 성능을 보여준다고 할 수 있다. 본 시스템에서 추가한 핵심아이디어는 동일문자 간에만 matching이 이루어지는 것이 아니라 표준형 변환 알고리즘에서 사용된 유사한 발음과 형태의 문자 간에도 match값을 줘서 변형 욕설과 일반 욕설 간의 matching이 이루어지게 하였다.

표 6. 변형 자소의 matching value

기본 자소	매칭 자소	점수
ㄱ	ㄲ, ㅋ	0.8
	g, k, c	0.6
	〉	0.4
ㄴ	n	0.6
	L	0.5
ㄷ	ㄸ, ㅌ	0.8
	d, t	0.6
ㄹ	l, r	0.6

[표 6]은 본 시스템에서 사용되는 변형 자소에 대한 matching value의 일부이다. 동일한 자소를 1.0으로 두고 'ㄱ'과 'ㅋ'는 0.8의 매칭값을 가진다. "만득이"를 자모 분리 시켰을 때 "ㅁㅏㄴㄷㅡㄱㅇㅣ"의 8개의 자소로 최대 8.0의 값을 가지는데 "만득이"를 alignment할 경우 "ㅁㅏㄴㄷㅡㅋㅇㅣ"는 7.8의 매칭값을 가지게 되므로 상대적 유사도는 7.8/8.0 = 97.5%로 계산된다. 본 시스템은 sensitivity와 specificity가 일치하는 75%를 threshold로 설정하였으므로 "만득이"는 금칙어로 정상적으로 판정되게 되는 것이다.

표 7. 욕설의 Semi-Global Alignment 예

	욕설	Alignment 결과
Source	ㅅㅣ입팔	ㅅ ㅣ ㅇ ㅣ ㅂ ㅍ ㅏ ㄹ
Target	씨입팔	ㅆ ㅣ ㅇ ㅣ ㅂ ㅍ ㅏ ㄹ
Score	7.8	0.8 1.0 1.0 1.0 1.0 1.0 1.0 1.0
Source	시이방	ㅅ ㅣ ㅇ ㅣ ㅂ ㅏ ㅇ
Target	쒸이방	ㅆ ㅟ ㅇ ㅣ ㅂ ㅏ ㅇ
Score	6.4	0.8 0.6 1.0 1.0 1.0 1.0 1.0
Source	ㅅㅣ입팔	ㅅ _ _ ㅣ ㅇ ㅣ ㅂ ㅍ ㅏ ㄹ
Target	씨이이이팔	ㅅ ㅣ ㅇ ㅣ ㅇ ㅣ _ ㅂ ㅏ ㄹ
Score	3.6	1.0-1.0-1.0 1.0 1.0 1.0-1.0 0.6 1.0 1.0

[표 7]은 여러 욕설들의 semi-global alignment 결과 값을 보여주고 있다. 몇몇 글자를 변형 시켜서 변형 형태를 만들더라도 충분히 높은 유사도가 측정되는 것을 볼 수 있다.

Ⅳ. 다단계 필터링을 통한 정상 단어 처리

앞서 설명한 Scunthorpe Problem은 욕설 필터링 시스템의 발전을 저해하는 아주 중요한 문제이다. 욕설을 원천 봉쇄하기 위해 강도 높은 필터링을 수행하게 된다면 필터링 되는 정상 단어의 빈도도 높아지기 때문이다.

변형 욕설의 경우 이 문제가 더 민감하게 작용하게 된다. 일반 욕설보다 변형 욕설이 일반 단어와 유사한 형태일 가능성이 높기 때문이다. 특히 semi-global alignment를 이용하여 유사도 측정을 통해 변형 욕설을 검색하는 본 시스템의 특성상 욕설과 유사한 일반 단어는 높은 측정치를 보여 필터링 될 가능성이 높다. 인터넷 용어는 은어와 신조어가 많이 사용되고 외래어 등의 유입도 빠르기 때문에 이러한 단어들이 필터링 되는 경우 또한 매우 빈번하게 일어날 것이다.

이 문제의 해결책은 정상 단어를 모아놓은 사전을 이용하여 미리 정상적인 단어를 검증하고 나머지 부분에 대해서만 욕설 필터링을 수행하는 것이다. 예를 들어 "시발점"이라는 단어를 입력 했을 때 "씨발"이 금칙어 리스트에 포함되어 있다면 "씨발"과 "시발"은 같은 표준형인 "ㅅㅣㅂㅏㄹ"로 변환되고 유사도도 높기 때문에 욕설 필터링에 의해 걸러지게 된다. 그러나 "시발점"이라는 단어를 미리 정상 단어 사전에 추가하고 단어를 미리 검증해 둘 경우 정상적으로 단어를 입력 할 수 있다.

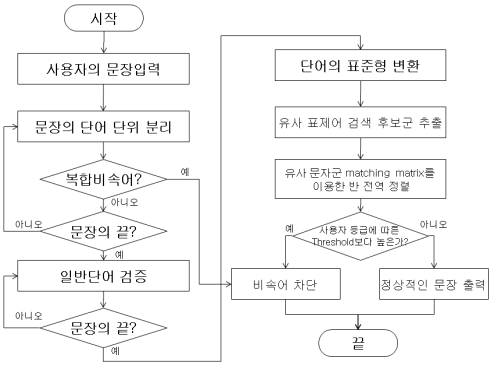

그림 2. 다단계 욕설 필터링 시스템

반면에 정상 단어 검증은 욕설 필터링과 같은 부작용을 일으키게 된다. 욕설 필터가 정상 단어를 욕설로 필터링하는 경우처럼 정상 단어 필터가 욕설을 정상 단어로 필터링하는 경우이다. "말뼈다귀"의 경우 "말"라는 단어와 "뼈다귀"라는 단어의 복합어이다. 이 단어를 그냥 일반단어 검증을 하고 욕설 필터링을 한다면 "말뼈다귀"는 정상 단어 두개로 인식되어 욕설 필터링을 빠져나가게 된다.

이 부작용을 막기 위한 해결책은 정상 단어 검증 방법과 유사하다. 정상 단어 검증 전에 욕설 필터링을 한 번 더 거치는 것이다. 이 욕설 필터링은 변형 욕설 필터링과는 달리 명백히 욕설이라 할 수 있는 것만 필터링 하여 Scunthorpe problem이 일어나지 않도록 하여야 한다.

Ⅴ. 변형 욕설 필터링 실험

우리는 본 시스템의 성능을 측정하기 위하여 무작위로 수집된 일반 욕설 1,672개의 필터링 실험과 무작위로 추출된 정상 단어 1,505개에 대한 검증 실험을 수행하였다. 실험에 사용된 금칙어 리스트는 한국게임산업진흥원에서 배포한 "게임언어 건전화 지침서 연구"에 독자적으로 수집한 비속어를 추가하여 9,265개의 금칙

어를 이용하여 실험하였다.

$$sensitivity = \frac{T_p}{T_p + F_n} \quad specificity = \frac{T_n}{T_n + F_p} \qquad (1)$$

위 공식은 성능의 기준으로 사용되는 sensitivity와 specificity의 공식이다. Tp(True Positive)는 필터링 된 단어들 중에서 욕설이 맞는 경우에 해당하고 Fn(False Negative)은 필터링 되지 않은 단어 중에서 욕설 수를 의미한다. Tn(True Negative)은 필터링 되지 않은 단어 중 정상 단어를 의미하고 Fp(False Positive)는 필터링 된 단어 중에서 정상 단어이다.

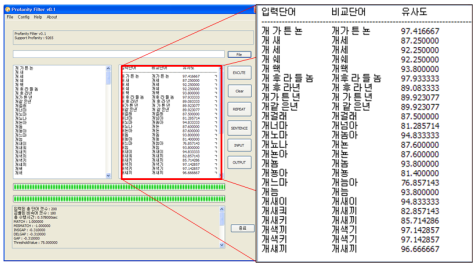

그림 3. 변형 욕설 필터링 시스템 실행 화면

위 [그림 3]은 실험에 사용된 구현 프로그램이다. 미리 저장된 9,265개의 금칙어와 입력단어 간의 상대유사도를 반 전역 정렬 값을 계산 파악할 수 있도록 만들어져 있다.

표 8. 실험에 사용된 입력단어와 검출된 결과

	입력 단어 수	검출 단어수
정상 단어	1,505	18
일반 욕설	1,672	1,648

(1)의 공식과 실험 결과로 계산되는 sensitivity와 specificity는 각각 98.5%와 98.8%이다. specificity가 아주 높게 측정되어 Scunthorpe problem이 상당 부분 해결된 것을 알 수 있다. 일반 욕설에 대한 sensitivity 또한 threshold값 75%에서 거의 유사한 수치인 98.5%를 기록하여 일반적인 비속어 필터로서의 성능은 만족할 만한 수준인 것을 알 수 있다.

표 9. 변형 욕설 필터링 비율 비교 실험 결과(200단어)

	검출된 단어 수	sensitivity
제안된 시스템	180	90.0%
게임포탈(N사)	60	30.0%
게임포탈(H사)	23	11.5%
게임포탈(P사)	20	10.0%

[표 9]는 본 시스템과 유명 게임포탈 N사, H사와 P사의 금칙어 필터링 시스템 간의 변형 욕설 필터링 비율에 관한 비교 실험 결과이다. 입력된 단어는 일반 욕설을 의미가 통하는 정도로 변형 시킨 욕설들을 사용하였다. N사가 30% 정도, 다른 두 게임 포탈의 필터링 시스템이 각각 11.5%와 10%로 매우 낮은 성능을 보인 것에 반해 본 시스템에서는 threshold 75%를 적용하였을 때 90% 정도의 필터링 비율을 보여 단어 단위의 변형 욕설 필터링 문제에 대해 매우 우수한 성능을 보이는 것을 알 수 있다.

Ⅵ. 결론 및 향후 연구 과제

본 논문은 욕설 필터링에서 발생하는 두 가지 주요 문제점인 변형 욕설의 필터링 문제와 정상 단어가 같이 필터링 되는 문제에 관한 해결 방법에 대해서 설명하였다. 제안된 해결 방법은 다음과 같다.

1. 변형 욕설 필터링 : 단어를 유사한 발음과 형태의 문자로 바꾸거나 특수 문자, 빈칸 등을 욕설 사이에 삽입하는 방법으로 욕설 필터링을 피해가는 문제이다. 여러 문자를 대표 되는 문자 몇 가지로 통일하여 단순화 시키는 방법으로 표준형을 만들어 유사한 발음의 단어를 검색하는 색인으로 사용하였다. 해당 단어가 변형 욕설인지 확실히 검증하기 위한 방법으로는 semi-global alignment를 이용한 유사도 값 측정방법을 사용하여 정확도를 높였다.
2. 정상 단어 필터링 문제 : 유사도를 이용하여 변형 욕설을 필터링하는 본 시스템은 정상 단어 필터링 문제에 더욱 민감하다고 할 수 있다. 이러한 정상 단어 필터링을 피하기 위해서는 사전에 정상 단어를 검증하여 정상 단어에 속하지 않는 영역만 욕설

을 검증하는 방식을 사용하였다. 그리고 정상 단어 검증이 가져다주는 부작용 즉 욕설의 substring이 정상 단어일 경우 욕설을 정상 단어로 간주하는 문제점을 막기 위해서 사전에 복합 욕설 및 명백히 욕설인 단어들을 필터링하는 방법으로 3단계에 걸친 필터링 시스템을 적용하였다.

본 시스템은 변형 욕설 문제에 대해 우수한 성능을 보이며 특히 정상 단어 필터링 문제를 상당부분 해결하여 실용성이 높다고 할 수 있다. 해결해야 할 과제로는 semi-global alignment를 통해 욕설 여부를 판정하기 때문에 많은 계산이 요구되어 서버 단위로 처리하고자 할 때 어려움이 따를 수 있다. 욕설 사전의 데이터 구조와 alignment방법의 개선을 통하여 속도의 향상이 필요하다고 생각된다.

참 고 문 헌

[1] 한국게임산업진흥원, *게임언어 건전화 지침서 연구*, 2008.

[2] C. Lai "An empirical study of three machine learning methods for spam filtering," Know.-Based Syst, Vol.20, No.3, pp.249-254, 2007.

[3] 박교현, 이지형, "SVM을 이용한 온라인게임 비속어 필터링 시스템", 2006년도 한국정보과학회 가을 학술발표논문집, 제33권, 제2(B)호, pp.260-263, 2006.

[4] G. Kondrak, "Identifying cognates by phonetic and semantic similarity," In Second Meeting of the North American Chapter of the Association For Computational Linguistics on Language Technologies 2001, pp.1-8, 2001.

[5] A. Ramachandran, N. Feamster, and S. Vempala, "Filtering spam with behavioral blacklisting," In Proceedings of the 14th ACM Conference on Computer and Communications Security, pp.342-351, 2007.

[6] http://en.wikipedia.org/wiki/Scunthorpe_Problem

[7] http://www.khgames.co.kr/week_01/main_content.htm?mCode=1615871101284119&sCode=381101173848313&code=laboratory&idx=8

[8] http://en.wikipedia.org/wiki/Swear_filter

[9] D. Shekhar, "Designing a vulgarity filtering system," in Game Programming Gems 5. 2005, Charles River Media, 2005.

[10] (주)아이모션, "*음절결합 정보를 이용한 음란/비속어 차단시스템*", 특2001-0067853, 2001.

저 자 소 개

윤 태 진(Taijin Yoon) 준회원

- 2009년 2월 : 부산대학교 정보컴퓨터공학부(공학사)
- 2009년 3월 ~ 현재 : 부산대학교 정보 컴퓨터 공학부 석사과정

<관심분야> : 한글 언어 처리, 정보 검색, 이미지 프로세싱

조 환 규(Hwan-Gue Cho) 정회원

- 1984년 2월 : 서울대학교 계산통계학과(석사)
- 1986년 2월 : KAIST 대학원 전산학과(공학석사)
- 1990년 2월 : KAIST 대학원 전산학과(공학박사)
- 1990년 3월 ~ 현재 : 부산대학교 공과대학 정보 컴퓨터공학부 교수(한국정보올림피아드 운영위원)

<관심분야> : 알고리즘, 응용 그래프 이론, 생물정보학