웹 게시판에서 비속어사용실태와 문제 해결 방안의 제시

Usage Analysis of Swearing Words on Web Board and Proposal of Problems Resolution Method

조동욱
충북과학대학 정보통신과학과

중심어 : 웹 게시판, 비속어

요 약

최근 인터넷상의 웹 자유게시판에 쓰여지는 글들이 비속어를 많이 사용함으로써 인터넷 역기능의 대표적인 문제로 대두되고 있다. 이를 위해 본 연구에서는 웹 게시판에서 사용되는 비속어에 대한 실태 파악을 통해 비속어가 쓰여진 단어나 문장을 기술적으로 차단하는 방법론을 개발하고자 한다. 이는 크게 세 단계로 나누어 개발이 진행된다. 첫째가 비속어 사용 실태 및 이를 차단키 위한 알고리즘의 개발, 둘째가 비속어의 정도 차이를 파악키 위한 구체적이고 방대한 설문조사의 수행, 셋째가 이를 프로그램하여 실제 웹 상에서 비속어가 어느 정도 효과적으로 차단이 가능한지에 대한 시스템 구현 등으로 나누어진다. 본 논문은 이 같은 전체 시스템 중 첫째 단계인 웹 게시판에서의 비속어 사용 실태와 이를 해결키 위한 알고리즘 개발 부분에 대해 다루고자 한다.

Dong-Uk Cho (ducho@ctech.ac.kr)
Dept. of Information & Communications Engineering,
Chungbuk Provincial University of Science & Technology

Keyword : Web Board, Swearing Words

Abstract

Recently, usage of swearing words on web board is the most typical Internet negative-functions. For this, technical method is proposed for blocking swearing words or sentences by analyzing swearing words usage types and behaviors. This system consists of 3 steps. Firstly, a survey , analysis of swearing words on web board and algorithm proposal for blocking these words must be studied. Secondly, sufficient and concrete opinion researches about every generations for measuring swearing degree must be accomplished. Finally, implementation on web board by programming will be done. This paper, in the first, deals with usage analysis of swearing and algorithm developement for solving these problems.

I. 서론

최근 들어 익명성이라는 이유로 표현의 자유를 누린다는 것, 인터넷의 쌍방향성을 이용한 다양한 의견의 교환등과 같은 인터넷 웹 자유게시판의 순기능과 더불어 웹 자유게시판에 폭언, 비방, 근거 없는 흑색 선전 등과 같은 역기능이 사회적 문제로 대두되고 있다. 청와대에서는 이 같은 문제 때문에 청와대 홈페이지에 올라오는 글들에 대해 3회 이상 근거 없는 비방이나 욕설의 글이 올라오면 삼진 아웃제[1]를 도입하여 그 글을 삭제하고 있으며 법원에서도 사이버 명예훼손과 관련하여 실형을 선고하는등 처벌을 강화하고 있지만 이는 역으로 인터넷의 특성인 표현의 자유와 쌍방향성을 법으로 그리고 강제로 규제하는 것이기 때문에 네티즌으로 하여금 강한 반발을 사고있는 것이 현 실정이다[2],[3]. 그리고 이에 따라 현재 인터넷에 실명제로 글을 올리는 것이 적절하다는 의견도 많이 개진되고 있으며, 실명제로 하더라도 진지한 토론을 하는 것이 얼마든지 가능하기 때문에 익명성을 통한 표현의 자유라는 것은 적합지 않다는 주장이 제기되고 있는 것도 사실이다[4]. 역으로 익명제를 주장하는 사람들은 익명제를 통해 욕설과 비방등 역기능이 존재하는 것은 사실이지만 마음껏 자신의 의견을 피력할 수 있다는 장점과 더불어 역기능을 해결키 위한 기술적 방법이 개발되기 때문에 이를 어떠한 사회적 제도나 방법으로 규제하기 보다는 인터넷 발전 과정의 한 수순으로 보는 것이 적합하다는 의견이다. 본 논문에서는 근래 웹 게시판에 쓰여지는 글들에 대한 형사 처벌이 강화되고 있는 실정[5]에 맞추어 이에 대한 적절한 대응책을 마련하고자 한다. 이를 위해 우선적으로 가정과 학교에서 어떤 방안이 마련되어야 하는것에 대해 제안하고자 한다. 아울

접수번호 : #031029-005
접수일자 : 2003년 10월 29일, 심사완료일 : 2003년 12월 11일

*교신저자 : 조동욱, e-mail : ducho@ctech.ac.kr

러 최종 목표는 원하는 사람들은 누구나 비속어 처리 프로그램만 가지고 있다면 웹 게시판의 역기능을 해결할 수 있도록 하기 위한 기술적 방법론의 개발이다. 이를 위해 크게 삼단계의 과정을 걸쳐 개발하고자 한다. 첫째가 사이버공간의 특성과 장·단점을 분석하고 비속어를 사용하는 사이트의 종류, 그 사용 이유와 비속어 종류 및 사용에 따른 문제점 실태를 조사하고자 한다. 이를 통해 비속어를 차단하는 알고리즘을 제안한다. 둘째가 제안한 알고리즘의 유용성과 효율성을 뒷받침하기 위한 구체적이고 방대한 양의 비속어 파악과 이에 따른 설문조사의 수행 그리고 마지막으로 이를 프로그래밍하여 구현함으로서 실제의 차단율을 계산하여 개발한 시스템의 효용성을 입증하는 것이다. 본 논문은 전체 시스템중 첫 번째 단계로 웹 게시판에 대한 종합적 분석, 학교와 가정에서의 대처 방안에 대한 제시 그리고 기술적 차단법에 대한 알고리즘 제시와 이에 따른 설문조사의 예에 대해 다루고자 한다.

II. 웹 게시판의 특성

인터넷 웹 게시판이 가지는 특성은 아래와 같다.

□ 익명성

가명(ID)으로 상징되는 '접속인' 들간의 만남에서 비롯되는 것이다. 접속인들간의 만남은 사회적 관습과 자신의 정체성이라는 맥락에서 벗어날 수 있다는 점에서 발생한다. 이는 정보의 다양성을 무한히 증식시키고 개인들의 자발성을 촉진시키면서 다양한 커뮤니티를 형성할 계기가 되기도 하는 장점이 있지만 폭력이나 사기행각을 벌이는 등의 문제를 야기할 수 있는 단점도 존재한다.

□ 개방성

개방적이지 않은 통신망은 이미 사이버공간이라고 할 수 없다. 인터넷의 상업화가 가속화되어 정보공유와 쌍방향의 자유로운 접속이 제한될 경우 인터넷 자체의 성장이 중단될 수 있다.

□ 자율성

사이버공간의 가장 강력한 도덕적 원리로서, 오프라인에서의 일방향적인 소통방식과 통제기구 및 규법으로부터 벗어나 자율적 개인들이 스스로 표현의 자유를 만끽할 수 있는 원리이다.

III. 웹 게시판 제공

통상 홈페이지등을 제작할시 웹 게시판의 기능은 웹 게시판을 제공해 주는 곳에서 다운 받아 사용하는 것이 일반적인 방법이다. 아래에 대표적으로 웹 게시판을 제공해 주는 곳에 대해 나타내었으며 어떠한 형태로 비속어등을 차단하는지에 대해 나타내었다.

□ com.ne.kr

이 곳에서 제공하는 게시판은 무료이다. 단, 유해단어 설정과 차단을 해주는 설치가 없다.

□ zero board

제로보드에서 사용되는 게시판은 주로 유료사이트로 리눅스 시스템에서 사용할 수 있는 게시판 중 가장 많이 사용하는 것 중 하나가 바로 제로보드이다.

제로보드 게시판을 사용하는 웹 사이트에 대하여 어떻게 생각하는가에 대한 설문조사에 의하면 사용하기에 편리하면 아무런 상관없다고 68.3%가 응답하였다. 하지만 제로보드는 주로 유료사이트로 무료계정을 받기가 힘들다. 아울러 제로보드는 비속어를 사용 시 무조건 차단되는 문제가 존재한다. 예로서 "귀여운 내 새끼"라는 문장을 입력해도 충분히 허용이 가능한 문장임에도 불구하고 무조건 차단이 되는 문제가 존재한다.

□ myboard

이 곳은 운영자가 직접 유해 단어를 입력해서 차단하는 방법을 가지고 있다.

□ http://www.superboard.com/

이 곳은 관리자사칭, 불량 단어, 불량 IP등 세가지 타입으로 차단하고 있다.

□ http://www.cgiserver.net/

이 곳은 금지 단어라고 하는 항목이 메뉴에 있으며 이곳에 원하는 금지 단어를 쓰면 자동으로 차단해 주고 있다.

□ http://ttcgi.com/(티티보드)

이 곳은 등급에 따른 각종 권한 설정, 관리자 공지기능, 욕설 방지 기능, DB 백업, 게시판 언어 설정 등 다양한 설정 기능을 제공하는 곳이다.

□ http://www.cgiworld.net/

음란어, 광고 차단 기능, 도배 방지 기능, 특정단어 등록 거부 기능 등을 사용하고 있다.

IV. 비속어 사용 종류, 사용 이유 및 문제점 실태

1. 비속어를 사용하는 사이트의 종류

□ http://boa_xxx.wo.to/(보아 안티 사이트)

특별한 제재가 없는 것 같다. 비속어가 있는데로 다 써진다. 대표적인 것으로 씨발, 병신, 주둥이, 년등의 단어가 써진다. 이보다 훨씬 심한 욕설과 비속어를 써도 차단이 않된다.

□ http://redjin.onair.co.kr/(레드존)

이 곳은 3류가수, 표절의혹, 립싱크 가수의 안티를 다룬 사이트이다. 여기는 게시판을 여러 용도로 나누고 거기에 맞는 글을 올리도록 되어있다. 본 사이트의 운영자는 자신이 운영하는 게시판을 따로 만들어서 오직 운영자만 글을 올릴 수 있도록 해 놓고 이용자는 볼수만 있도록 해놓았다.

□ http://krmusics.tripod.com

이 사이트도 위의 레드존과 비슷한 사이트이다.

그러나 최근에는 많이 이용되고 있지 않는 실정이다.

2. 비속어의 사용 이유

비속어를 사용하는 이유는 다음과 같이 분석되어 진다 [6],[7].

첫째가 심리적 안정성이다. 이는 강압적 사고가 우위인 우리 사회에서 비속어를 사용하는 것이 자신의 위치를 더욱 확대하고 자신을 방어할 수 있다는 심리적 안정성을 준다는 이유에서 사용되어 진다.

둘째가 심리적 통일성과 유대감 확인이다. 통상 어떤 그룹에 속하기 위해서는 그들이 사용하는 언어를 공유해야만 한다. 즉, 타인과 함께 하나의 그룹에 속하고자 하는 욕구가 비속어 사용으로 나타난다고 할 수 있다.

셋째가 화를 표출하기 위해 사용 되어 진다. 비속어는 사회적으로 나쁘다고 인식되고 저속하다고 정해진 말이지만 이를 사용하면서 한편으로는 화를 나타내고 이를 통해 카타르시스를 느끼게 되는데, 이와 같은 아이러니한 경험으로 인해 비속어를 사용한다.

넷째가 유머러스한 말을 구사하기 위해 사용한다. 즉, 정상적인 표준어 사용이 진부하게 느껴져서 만족을 느끼지 못할 경우 사용하게 된다는 것이다..

다섯째가 습관적으로 사용하게 된다는 것이다. 비속어 사용시 주변에서 특별한 제재가 없었을 경우, 비속어 사용에 대한 문제점을 인식하지 못한 상태에서 자연스럽게 하나의 용어로 정착이 되게 된다.

여섯째가 예의바른 표현이나 권위에 반항하고 싶은 심리에서 사용한다.

일곱 번째 이유는 사실적 표현으로 구체성을 강하게 표현하고자 할 때 사용한다.

마지막으로 신기한 표현을 통해 상대방을 놀라게 하고자 하는 의도에서 사용되게 된다.

3. 비속어의 종류

대표적인 비속어에 대해 논하고자 한다.

3.1. 비어

- 주둥아리, 대가리, 이 놈, 저 놈, 죽어라, 싸가지, 꺼져, 닥쳐, 씹창, 새끼, 지랄, 미친, 개자식, 미친놈, 병신, 씨발
- 신체 : 대가리(대갈통) , 주둥이(아가리)
- 호칭 : 아버지-애비(아비, 꼰대, 꼰상), 어머니-에미(어미), 뱃사람(뱃놈), 아이년, 요년, 아이놈 등
- 일반적인 것들 : 먹는다(처먹는다), 죽는다(뒈진다), 닫아라(닥쳐라)등
- 범생(모범생), 왕따(왕따돌림), 십빠빠 울랄(진짜 못생겼다), 야리다(째려보다)

3.2. 속어

- 호박(못 생긴 얼굴)
- 골때린다. (어처구니없다.)
- 뽕간다 (기막히다.)
- 끝내주다 (최고다)
- 호박씨까다 (겉과 달리 몰래 다른 짓을 하다.)
- 캡 (최고)
- 삥땅(부분적인 횡령행위)
- 공갈(거짓말)
- 사꾸라(야바위)

등과 같은 비속어가 웹 게시판에서 많이 사용되어 지고 있는 실정이다

4. 속어 사용 문제점 실태[8]

학생 50명을 대상으로 한 인터넷 언어폭력에 관한 실태조사에 따르면 언어 폭력중 심한 욕설과 인격 모독(44%)이 가장 많은 비중을 차지한 것으로 조사 되었다. 심지어 26%는 성폭력까지 당한 경험이 있는 것으로 나타났다. 이런 언어폭력은 주로 채팅(68%)에서 이루어졌으며 '자신도 언어폭력을 해보았느냐?'는 질문에 48%가 그런 경험이 있다고 응답해 실제 사이버상에서는 과반수가 넘는 넷티즌이 언어폭력을 하고 있는 것으로 나타났다. 또한 요즘 '인터넷상에서 철자를 무시하고 이상한 국적 불명의 용어를 만들어 사용하는 것에 대해 어떻게 생각하는가?'라는 질문에는 '그렇게 중요한 문제가 아니다'(40%), '그냥 자연스러운 현상'(38%)으로 응답해, 무려 78%가 한글의 엉터리 표기와 맞춤법을 지키지 않는 것을 대

수롭지 않게 받아들이고 있는 것으로 조사됐다. 그러나 진작 중요한 것은 웹 자유 게시판에서 무분별하게 일어지고 있는 욕설과 비방이며 이에 대한 대응이 가장 중요한 일로 여겨진다.

5. 속어 사용의 문제점

비속어 사용의 문제점은 다음과 같이 요약되어 질수 있다.
첫째, 원만한 인간관계에 악영향을 미친다.
'비속어(卑俗語)'란 어떤 대상을 아주 얕잡아 보고 경멸하는 태도로 사용하는 용어이므로 비속어를 사용할 경우, 대화 당사자들 간의 기분이 상할 수 있다.
둘째, 정서적인 면에 있어 문제가 발생할 수 있다. '비속어(卑俗語)'는 주로 발음이 과격하거나 표현이 거친 말이 대부분이므로 올바른 정서 함양에 문제를 가져올 수 있다.
셋째, 언어체계의 혼란을 가져올 수 있다.
넷째, 사회의 질적 가치 저하를 가져올 수 있다.
다섯째, 어린이들에게 악영향을 가져올 수 있다.

V. 웹 게시판 문제점을 해결하기 위한 학교와 가정에서의 대응방안

웹 게시판의 문제점을 해결하기 위해 학교와 가정에서 어떻게 대응하는 것이 효과적인가에 대해 다루고자 한다.

1. 학교에서의 대응방안

우선 학교에서 전산 교육을 수행 시 운영체제, 정보 검색등과 같은 지식교육, 프로그래밍 및 소프트웨어 활용등과 같은 기술 교육과 더불어 정보통신에 대한 가치관 교육이 이루어져야 하리라 여겨진다. 아래 표 1에 주요 가치관 교육의 내용에 대해 나타내었다. 특히 이는 2002년만 해도 6만 68건에 해당하는 사이버 범죄가 발생하여 전년도 보다 무려 80%가 증가하는 양상을 보이고 있기 때문에 더욱 주요한 교과 과정으로 운영이 되어야 하리라 여겨진다. 또한 범죄를 일으키는 연령층이 컴퓨터를 가장 잘 사용하는 10대가 가장 높기 때문에 더욱이 중요한 요소가 되리라 여겨진다. 특히 웹 자유 게시판에서의 욕설과 비방은 "네티켓 준수"과목으로 다루어 져야 하리라 여겨지며 아무리 기술이 발전해도 기본적인 윤리 규범은 바뀌지 않음을 주요 내용으로 해야 하리라 여겨진다.

표 1. 정보 통신 가치관 교육의 예

이수 영역	학습 요소
컴퓨터 범죄와 보안	해커, 컴퓨터 보안의 중요성과 대비 방법
바이러스	바이러스의 종류, 바이러스 예방책
컴퓨터 중독과 네티켓	인터넷의 역기능, 네티켓 준수, 디지털 지적 재산권

2. 가정에서의 교육

가정에서는 인터넷 역기능과 관련된 문제를 해결할 수 있는 툴들을 설치하여야 하며 이에 대한 충분한 상황을 자녀들로부터 동의를 얻어내야 한다. 그러나 현재는 예로써 스팸메일 차단을 예로 들면 IT종사자들조차도 14%밖에 사용하지 않는 실정 [9이므로 무엇 보다 학부모들의 적극적인 대응 움직임이 필요한 실정이다.

VI. 기술적 해결방안의 제안

1. 기존 방법에 대한 고찰과 제안하는 방법과의 비교

웹 게시판에 쓰여지는 글중 일반인이 용납하기 어려운 글들에 대해 단어를 삭제하거나 심한 경우 해당 문장을 삭제해야 한다. 이를 기술적으로 해결하기 위한 방법들이 나와 있다. 대표적인 방법들이 조아영 방법[10], 김응곤 방법[11], 제로보드, 조동욱 방법[12] 등이다. 이중 조아영 방법은 비속어에 대한 패턴 분석으로 비속어를 차단하였으며 이때 비속어 등급을 1, 2, 3등급으로 나누어 이를 수행하였다. 그러나 이 방법은 비속어에 대한 등급을 부여한 것이 작위적이라, 실제 허용 가능한 비속어도 필터링 될 수 있는 문제가 존재한다. 이에 비해 김응곤 방법은 유해 단어 사전을 제작하고 이를 통해 유해 단어를 필터링하는 방법을 제안하였다. 또한 부적절한 글쓰기형태를 판별하는 기준을 마련하였으며 이를 통해 비방이나 욕설을 행하고자 하는 작성자의 의도를 파악하고자 하였다. 가장 중요한 것은 개인 기록 조회를 통해 누가 기록 DB를 기록함으로서 불순 의도가 있는 글 작성자를 필터링할 수 있도록 하였다. 그러나 이도 ID를 수시로 변경하여 글을 올리면 필터링이 않되며 아울러 유해 단어에 대한 분명한 기준이 없는 실정이다. 끝으로 조동욱의 방법은 퍼지 의사 모델을 이용해 0~1사이의 값으로 불쾌정도나 비방정도를 계산하

였다. 이러한 방법은 작성자가 직접 입력하거나 제삼자가 입력해줘야 한다는 문제가 존재한다. 이에 비해 제안하고자 하는 방법은 비속어를 무조건 차단하는 것이 아니라 여러 사람들의 의견을 세대별로 반영하여 허용할 수 있는 비속어는 허용하고 허용이 않되는 비속어에 대해 차단하는 방법을 채택하고자 한다. 사실 웹상에 비속어나 비방에 대한 것은 어느 정도가 욕설이고 비방인가에 대한 일반인들이 의견이 가장 중요한 요소이고 본 논문에서는 이를 반영하여 비속어를 차단하고자 한다. 아래 표 2에 기존 방법과 제안하고자 하는 방법과의 기본 알고리즘에 대한 기술 그리고 장·단점을 비교, 분석하였다.

표 2. 기존 방법과의 비교, 고찰

주된 개발자	주된 내용	장 점	단 점
조아영 방법	패턴분석으로 색인어 추출,비속어리스트와 패턴 매칭	비속어 단어 등급을 정의	비속어 단어 등급에 대한 기준이 모호함
김응곤 방법	유해단어 사전 분류, 부적절한 글쓰기 판별	개인기록조회를 누가 DB에 기록	ID변경시 조회 어려고, 유해단어 사전 분류 애매함
제로보드	비속어 DB매칭	필터링 확률 높음	허용 가능한 단어까지 차단
조동욱 방법	퍼지의사 모델	비방정도 계산 가능	문장 평가치 입력
제안한 방법	각 세대별로 비속어에 대한 평가 결과를 활용	여러 사람의 의견이 반영된 합리적 차단 방법 구현	각 계층의 의견을 수렴하는 것의 어려움이 존재

2. 제안한 방법

무조건적으로 비속어를 차단하는 기존 방법의 문제를 해결하기 위해 본 논문에서는 세대별로 대표적인 비속어에 대한 샘플을 가지고 설문조사를 행하여 기술적으로 웹 게시판의 정화 기능을 부여하는 방법을 제안하고자 한다.

제안한 방법은 크게 단어 필터링과 문장 필터링으로 나누어진다. 우선적으로 유해 단어, 속어, 비어등에 대해 설문조사를 행하여 아래 표 3과 같은 가중치를 부여한다. 설문조사를 행하는 것은 일반인들이 느끼는 비속어나 욕설에 대해 그 평가 기준을 삼고자 하는 것이며 이는 퍼지 이론을 예로 든다면 퍼지 멤버쉽함수[3],[4]의 경우에 해당된다. 퍼지 멤버쉽함수는 데카르트 사상이 가지고 있는 이산적이고 디지털적인 이진 논리를 사람의 사고방식과 가깝도록 처리하여 정보의 손실을 줄이고자 하는 접근 방식이며 본 논문에서도 무조건적인 비속어 차단이 아닌 사람들의 일반적인 의견이 반영된 다시 말해 모델링 기법으로는 퍼지 이론과 개념에 기초한 비속어 차단 방법을 제시하기 위해 각 세대별로 비속어의 의견이 반영된 설문조사 결과를 근거로 한 차단 방법을 제안하고자 한다.

우선 정규화되고 선형적인 평가치를 계산하기 위해 아래 표 3과 같은 비속어에 대한 가중치를 부여한다. 이때 최종 평가치는 식 (1)과 같이 가중합(weighted sum)을 이용하여 행하며 이 값은 0과 1사이의 값을 가진다. 아울러 이 값이 1에 가까울수록 단어 필터링의 가능성은 커지게 된다.

표 3. 유해 단어, 속어, 비어등에 대한 가중치

속어	절대안됨 (1.0)	안됨 (0.75)	중간 (0.5)	됨 (0.25)	문제안됨 (0.0)

다시 말해 표 3과 같은 가중치를 바탕으로 세대별로 설문조사를 행하여 각 단어의 삭제 가중치를 부여하면 하식과 같은 평가치가 정의 가능하게 된다.

$$\text{평가치(E1)} = \sum_{i=1}^{n} \sum_{i=1}^{5} W_i f_i \qquad (1)$$

여기서 W_i는 가중치, f_i는 빈도 수를 뜻하며 각각 0과 1사이의 값으로 정해진다. 즉, 가중치는 표 3과 같이 그리고 빈도 수는 어느 항목에 얼마나 표했는가를 백분율로 나타낸 값이 된다.

이제 세대별로 비속어나 욕설에 대해 받아들이는 정도가 다르므로 세대에 대한 반영을 아래 표 4와 같이 정하여 최종적인 즉, 세대별 의견이 반영된 여과치를 정의한다. 여기서 반영률을 10대와 20대는 50%로, 30대는 80%의견을 반영하고 40대 이상은 100% 의견을 반영하였다. 이는 설문 조사시 세대별 반영률을 어느 정도로 했으면 좋겠는가에 대해 그 결과를 반영한 것이기도 하지만 이러한 결과가 나오게 된 배경은 나이 든 세대일수록 비속어와 욕설에 대해 민감한 반응을 보이지만 나이 든 사람들이 정보통신 윤리 교육에는 적합한

의견을 가지고 있다는 연속적인 윤리 개념을 반영한 결과라 할 수 있다. 다시 말해 정보통신 윤리는 정보사회와 기존의 산업 사회와는 다른 "새로운 윤리"와 그리고 기술이 발전하더라도 기본적인 윤리 규범은 바뀌지 않는다는 "연속적인 윤리"로 나뉘어 진다. 이때 나이가 어릴수록 그 반영률을 높게 정의하면 "새로운 윤리"를 강조하는 것이 되고, 나이 든 세대의 의견을 많이 반영하면 "연속적인 윤리"를 높게 반영하는 것이 된다. 본 논문에서는 연속적인 윤리를 기준으로 하였으며 이는 설문조사 결과에서도 뒷받침이 되어 아직은 우리 사회가 연속적 윤리를 기본으로 하는 사회임을 확인할 수 있었다. 이를 기초로 아래 표 4와 같은 세대별 반영률이 정의되게 된 것이다.

표 4. 세대별 반영율

세대	반영율
10대, 20대	0.5
30대	0.8
40대이상	1.0

최종적으로 비속어에 대한 여과치는 하식과 같이 정의한다.

$$\text{여과치(F1)} = \sum\sum Gi/2.3 \qquad (2)$$

여기서 Gi는 세대별 반영율과 평가치의 곱을 뜻한다. 또한 식 (2)에서 2.3은 정규화를 위해 세대별 반영률의 가중치를 모두 더한 값이다. 그리고 비속어가 한 문장에서 여러 번 쓰여져서 전체 문장을 삭제해야 하는 경우에 대한 문장 삭제는 아래 표 5와 같이 구한다. 문장삭제란 똑 같은 비속어를 사용했더라도 한 문장 내에 그것이 몇 번 쓰여졌는가에 따라 단어 삭제와 문장 삭제로 나뉘어 진다. 즉, 비속어가 한 문장 내에 많이 쓰여졌다면 이는 단어 삭제보다는 아예 문장 삭제를 하는 것이 적합할 것이라는 결론에 도달하여 문장 삭제 모듈로 들어가게 된다. 물론 이때도 쓰여진 비속어의 욕설정도에 따라 그 결과 값이 달라지며 유해단어, 비속어, 욕설 단어의 개수에 따라 그 값이 달라져야 한다. 따라서 표 5에서는 유해단어의 수에 따라 일정한 식으로 증가해야 하므로 이를 구현하기 위해 선형함수식을 적용해 0.5, 0.75, 1.0의 가중치를 적용하여 문장 삭제에 적용하였다. 이때 4개 이상은 같은 가중치를 부여하였다.

표 5. 문장 삭제

단어수	문장삭제식
유해 단어 2개	$0.5*(\sum_{i=0}^{2} WFTi/2)$
유해 단어 3개	$0.75*(\sum_{i=0}^{3} WFTi/3)$
유해 단어 4개이상	$1.0*(\sum_{i=0}^{4} WFTi/4)$

이제 불규칙 패턴에 대한 처리도 행해야 하는데 이는 숫자로 시작하거나 여러 가지 패턴이 합성된 경우에 대한 예가 된다. 아래 표 6에 불규칙 비속어 패턴에 대한 예를 나타내었다.

표 6. 불규칙 비속어 패턴의 예

불규칙 패턴	예
합성어	개씹새끼
비속어중간에 특수문자	지~랄하네
영문으로 시작	C8년
음절 띄어쓰기	개 씨 끼 야
숫자로 시작	18년
음소별 띄어쓰기	ㅅ ㅅ ㅣ ㅂ

VII. 비속어처리를 위한 설문조사의 예 및 적용

본 논문에서 제안한 방법에 대해 실험하기 위해 우선적으로 임의로 선정한 20개의 비속어를 대상으로 10대와 20대, 30대, 40대 세대별로 나누어 설문조사를 행하였다. 설문조사는 본 대학 재직 교직원과 학생들을 대상으로 시행하였다. 이중 10대는 청주 지역 거주 학생들을 대상으로 본 대학에 재학중인 학생들의 동생들이 다니고 있는 고교생을 대상으로 하였다. 이는 본 대학 학생들이 동생을 통해 쉽게 설문조사 결과를 얻을 수 있다는 편리함에 기초하였다. 표 7은 10대와 20대의 설문조사 결과이고 이 표에서 알 수 있듯이 10대와 20대는 설문에 응한 인원수를 기재하지 않았다. 이는 500여 명이 넘는 인원에 대해 설문조사한 것이기 때문에 특별히 설문조사 인원수를 기재하지 않아도 객관성을 유지할 수 있기 때문에 설문조사 인원을 표기하지 않았다. 그러나 표 8의 30대는 14명 그리고 표 9의 40대 이상은 20명에 대해서 설문조사를 행했는데 이는 본 대학 직원들을 대상으로 설문조사를

행해 조사 인원이 충분치 않기 때문에 이를 나타내어 실상을 알려주는 것이 객관성을 유지할 것으로 판단되어 설문조사 인원을 나타낸 것이다. 이때 30대와 40대는 비록 인원수는 작지만 본 대학 교직원을 대상으로 한 이유는 교직원은 학력이 고졸부터 대학원졸업까지 다양하게 존재하고 그리고 남녀간의 비율이 적절히 50%, 50%가 되기 때문이다. 이는 남녀간의 의견차를 확인하기 위함의 의도도 있었는바 설문조사 결과로는 세대간의 의견차는 확연히 존재하지만 남녀간의 의견차는 존재하지 않았다. 다시 말해 성교육이나 성관련 설문일 경우는 남녀간에 의견차가 존재하지만 욕설과 비속어의 경우는 의견차가 거의 없었다. 다만 아래 표 7, 표 8, 표 9에서 알 수 있듯이 똑같은 비속어라도 세대가 높을수록 거부감이 컸고 낮은 연령층일수록 비속어에 대한 허용 범위가 큼을 확인할 수 있었다. 또한 비속어에 대한 설문조사항을 20개로 한정시켜 행한 것은 항문수가 많아지면 나이가 들수록 이를 귀찮게 여겨 진실되게 응답하지 않기 때문이며 차후로는 20개씩 약 15개의 설문지를 만들어 다양한 비속어와 욕설에 대한 세대별 의견을 반영하고자 한다. 이도 충분한 인원 확보와 다양한 욕설 파악이 선행되어야 하므로 약 6개월 정도의 조사 기간이 소요될 것으로 여겨진다. 다음에 표 10에 가중치를 내포한 세대별 평가치를, 그리고 표 11에 세대별 가중치를 포함한 최종 평가치를 나타내었다. 또한 그림 1은 제안하고자 하는 알고리즘에 대한 전체 흐름도를 나타내었다. 본 방법은 제로 보드처럼 무조건적으로 비속어로 분류된 단어에 대해 삭제를 행하는 것이 아니라 세대별로 비속어에 대한 의견을 반영한 여과 방식이기 때문에 효과적으로 비속어를 차단할 수 있는 시스템이 되리라 여겨진다. 현재는 20개의 비속어에 대해 본 대학에 재직중 이거나 재학중인 학생들에 대해서만 설문조사를 수행하였지만 전체 시스템 구축을 위해서는 2단계 작업인 다양한 욕설 파악과 충분한 세대별 설문조사 인원 확보를 통한 객관성 유지에 대한 연구가 수행되어야 하리라 여겨진다. 따라서 그림 1에서의 임계치인 TH1과 TH2도 2단계 작업이 완료되어야 그 값을 얼마로 하는 것이 적절한지에 대한 선정이 가능할 것으로 여겨지며 제3단계 작업인 구현 과정을 통해 최종 차단율이 계산 가능하게 될 것으로 여겨진다. 본 연구는 전체시스템중 제1단계 작업인 웹 자유게시판의 현황과 비속어 사용 실태 그리고 이를 차단키 위한 알고리즘의 제시가 주된 내용이고 따라서 후속 연구가 조속히 수행되어 최종적인 차단율 계산까지 결과가 나와 그 유용성과 효율성을 입증하기 위한 연구가 지속적으로 행해져야 하리라 여겨진다.

표 7. 10대, 20대의 비속어 설문조사 결과

속어	절대 안됨	안 됨	보통	됨	문제 안됨
좆나	15.8%	39.5%	29%	10.5%	5.2%
씨발	15.8%	39.5%	29%	10.5%	5.2%
열라	5.2%	15.8%	44.7%	18.4%	15.8%
개새끼	21.1%	39.5%	21.1%	15.8%	2.6%
아가리	18.4%	36.8%	26.3%	13.2%	5.2%
지랄한다	15.8%	7.9%	47.4%	15.8%	13.2%
재수없다	5.2%	7.9%	50%	26.3%	10.5%
뒈진다	10.5%	23.7%	42.1%	18.4%	5.2%
병신	7.9%	34.2%	44.7%	7.9%	5.2%
아리다	10.5%	23.7%	44.7%	10.5%	10.5%
대갈통	13.2%	39.5%	23.7%	15.8%	7.9%
씹창	44.7%	34.2%	15.8%	2.6%	2.6%
싸가지	7.9%	10.5%	34.2%	26.3%	21.1%
끝내주다	2.6%	7.9%	34.2%	26.3%	29%
미친놈(년)	26.3%	21.1%	31.6%	13.2%	7.9%
애자	15.8%	44.7%	23.7%	10.5%	5.2%
띨박	26.3%	18.4%	29%	21.1%	5.2%
빠구리	52.6%	34.2%	7.9%	2.6%	2.6%
염병	23.7%	26.3%	34.2%	13.2%	2.6%
씨다발	26.3%	42.1%	18.4%	7.9%	5.2%

표 8. 30대의 비속어 설문조사 결과

속어	절대 안됨	안 됨	보통	됨	문제 안됨
좆나	10 (71.4%)	3 (21.4%)	1 (7.1%)	0%	0%
씨발	10 (71.4%)	3 (21.4%)	1 (7.1%)	0%	0%
열라	4 (28.6%)	7(50%)	2 (14.3%)	1(7.1%)	0%
개새끼	6 (42.9%)	7(50%)	0%	1(7.1%)	0%
아가리	8 (57.1%)	5 (35.7%)	0%	1(7.1%)	0%
지랄한다	3 (21.4%)	7(50%)	3 (21.4%)	1(7.1%)	0%
재수없다	4 (28.6%)	2 (14.3%)	7(50%)	1(7.1%)	0%
뒈진다	6 (42.9%)	6 (42.9%)	1(7.1%)	1(7.1%)	0%
병신	6 (42.9%)	6 (42.9%)	1(7.1%)	1(7.1%)	0%
야리다	5 (35.7%)	7 (50%)	1(7.1%)	1(7.1%)	0%
대갈통	5 (35.7%)	6 (42.9%)	1(7.1%)	2 (15.8%)	0%
씹창	10 (71.4%)	4 (28.6%)	0%	0%	0%
싸가지	3 (21.4%)	6 (42.9%)	3 (21.4%)	2 (14.3%)	0%
끝내주다	1(7.1%)	3 (14.3%)	7(50%)	3 (21.4%)	1(7.1 %)
미친놈 (년)	5 (35.7%)	6 (42.9%)	2 (14.3%)	1(7.1%)	0%
애자	7(50%)	4 (28.6%)	3 (21.4%)	0%	0%
띨박	6(42.9 %)	7(50%)	1(7.1%)	0%	0%
빠구리	10 (71.4%)	4 (28.6%)	0%	0%	0%
염병	4 (28.6%)	7(50%)	1(7.1%)	1(7.1%)	1(7.1 %)
씨다발	7(50%)	6 (42.9%)	0%	0%	1(7.1 %)

표 9. 40대의 비속어 설문조사 결과

속어	절대 안됨	안 됨	보통	됨	문제 안됨
좆나	16 (80%)	4(20%)	0%	0%	0%
씨발	17 (85%)	3(14%)	0%	0%	0%
열라	11 (55%)	4(20%)	4(20%)	1(5%)	0%
개새끼	17 (85%)	3(15%)	0%	0%	0%
아가리	16 (80%)	3(15%)	1(5%)	0%	0%
지랄한다	16 (80%)	3(15%)	1(5%)	0%	0%
재수없다	12 (60%)	4 (20%)	3(15%)	1(5%)	0%
뒈진다	16 (80%)	3(15%)	1(5%)	0%	0%
병신	16 (80%)	3(15%)	1(5%)	0%	0%
야리다	14 (70%)	4(20%)	1(5%)	1(5%)	0%
대갈통	16 (80%)	3(15%)	1(5%)	0%	0%
씹창	17 (85%)	3(15%)	0%	0%	0%
싸가지	11 (55%)	5(25%)	3(15%)	1(5%)	0%
끝내주다	9(45%)	4(20%)	5(25%)	2(10%)	0%
미친놈 (년)	12 (60%)	3(15%)	4(20%)	1(5%)	0%
애자	14 (70%)	5 (25%)	0%	1(5%)	0%
띨박	14 (70%)	4(20%)	1(5%)	1(5%)	0%
빠구리	17 (85%)	3(15%)	0%	0%	0%
염병	17 (85%)	3(15%)	0(0%)	0(0%)	0%
씨다발	17 (85%)	3(15%)	0%	0%	0%

표 10. 가중치를 포함한 비속어 평가치

속어	10대,20대	30대	40대
좆나	0.63	0.91	0.950
씨발	0.63	0.91	0.955
열라	0.44	0.75	0.813
개새끼	0.65	0.82	0.963
아가리	0.62	0.86	0.938
지랄한다	0.38	0.71	0.938
재수없다	0.43	0.66	0.838
뒈진다	0.56	0.80	0.938
병신	0.58	0.80	0.938
아리다	0.53	0.79	0.888
대갈통	0.59	0.75	0.938
씹창	0.79	0.93	0.963
싸가지	0.39	0.68	0.825
끝내준다	0.32	0.48	0.750
미친놈(년)	0.61	0.77	0.825
애자	0.66	0.82	0.900
띨박	0.60	0.84	0.888
빠구리	0.83	0.93	0.962
염병	0.55	0.71	0.963
씨다발	0.69	0.82	0.963

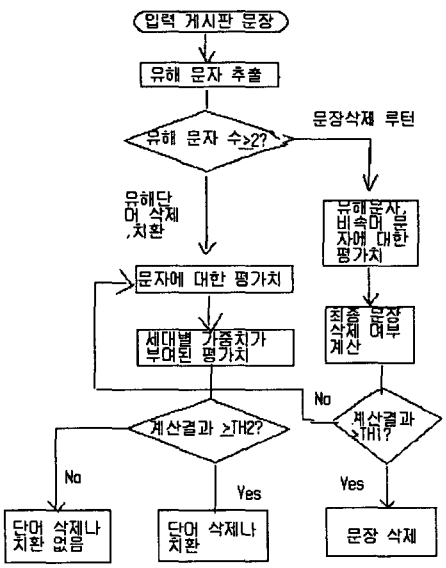

그림 1. 전체 시스템의 개요도

표 11. 세대별 가중치를 포함한 비속어 평가치

속어	평가치
좆나	0.867
씨발	0.867
열라	0.578
개새끼	0.844
아가리	0.843
지랄한다	0.738
재수없다	0.684
뒈진다	0.809
병신	0.809
아리다	0.813
대갈통	0.773
씹창	0.798
싸가지	0.913
끝내준다	0.739
미친놈(년)	0.554
애자	0.757
띨박	0.785
빠구리	0.805
염병	0.784
씨다발	0.853

VIII. 결론

본 논문에서는 웹 자유게시판에서 비속어를 기술적으로 차단하는 전체시스템의 개발중 첫 단계인 웹 자유게시판에 쓰여지는 글들에 대한 체계적인 분석과 비속어 사용 실태 그리고 이를 해결하기 위한 알고리즘을 제안하였다. 이를 통해 세대별 의견도와 여러 사람들의 의견이 반영된 유해단어, 비속어 처리 알고리즘을 제안하고자 하였다. 차후 문장중간의 비속어, 합성어, 그림 문자등 다양한 유해단어나 비속어에 대한 파악과 처리 알고리즘의 개발이 필요하다. 그리고 다양한 욕설과 비속어에 대한 구체적이고 충분한 양의 확보와 파악 및 충분한 설문조사 인원의 확보가 필요하다. 또한 이를 구현하기위한 최종 차단율을 유도한 후 개발하고자 하는 전체 시스템에 대한 유용성과 효율성을 입증하기 위한 연구가 지속적으로 행해져야 하리라 여겨진다.

참 고 문 헌

[1] 중앙일보, 비방하는 글 몸살, 청와대 홈페이지 삼진 아웃 제 도입, 2003년 7월 5일자 1면.
[2] http://www.aks.ac.kr/event/cyber_cult/html/4_15.htm
[3] http://cafe.daum.net/dksxlhomepage
[4] 충청일보, 인터넷 실명제찬반논란, 2003년 10월 15일자, 디지털라이프.
[5] 충청일보, 사이버 범죄 증가, 2003년 10월 24일자, 사회면
[6] http://kr.ks.yahoo.com 지식검색
[7] http://cafe.daum.net/dkuuri 자료실
[8] 중앙일보, 한글날 특집 기사, 2003년 10월 9일.
[9] 조동욱외, "스팸메일에 대한 현황과 분석", 한국정보과학회 충청지부 학술대회논문집, 2003년 12월 12일.
[10] 조아영, "웹자유게시판 비속어 처리 프로그램의 설계 및 구현", 한국컴퓨터산업교육학회 논문지, Vol. 2, No. 10, 2001년.
[11] 김응곤, "인터넷게시판에서 정보 통신 윤리 교육을 위한 유해 단어 필터링 시스템의 설계와 구현", 한국정보처리학회 추계종합학술대회 논문집, Vol. 9, No. 2, 2002년
[12] 조동욱, 신승수, "인터넷 역기능을 해결키 위한 기술적 방법론에 대한 검토", 한국콘텐츠학회 논문지, Vol. 2, No. 4, 2002년.
[13] 오길록, 이광형, 퍼지이론 및 응용, 홍릉출판사, 1991년.
[14] Klir & Folger, Fuzzy Sets, Uncertainty and Information, Prentice Hall, 1988.

조 동 욱(Dong-Uk Cho) 정회원

1983년 2월 : 한양대 공대 전자공학과 (공학사)
1985년 9월 : 한양대 전자공학과 (공학석사)
1989년 2월 : 한양대 전자통신공학과 (공학박사)
1991년 3월 ~ 2000년 2월 : 서원대학교 정보통신공학과 부교수
2000년 3월 ~ 현재 : 충북과학대학 정보통신과학과 교수
<관심분야> : 유해콘텐츠의 기술적 차단, 문화콘텐츠 보호, 지적 재산권 보호, 영상콘텐츠공학, 영상생체인증