음란 유해사이트에 대한 현황과 신호처리에 기반한 차단 방법의 제안

조 동 욱* 최 병 갑** 김 지 영***
e-mail : ducho@ctech.ac.kr

A Survey of Harmful Internet Sex Siter and Proposal of Blocking Methods Based on Signal Processing

Dong Uk Cho* Byung Kap Choi** Ji Yeong Kim***
* : Chungbuk Provincial Univ. of Science & Technology
** Mokwon University *** Seowon University

요 약

본 논문에서는 인터넷 역기능에 있어 가장 큰 문제로 인식되고 있는 음란사이트에 대한 현황과 이에 대한 지금까지의 정책적, 기술적 대응책에 대한 고찰을 행하고자 한다. 또한 지금까지의 기술적인 방법 등이 목록기반과 단어기반에 기초한 방법이기 때문에 음란사이트 차단이 제대로 되지 않은 경우가 있었다. 이를 위해 본 논문에서는 신호처리에 기반한 음란 유해 사이트에 대한 차단 방법을 제안하고자 한다. 특히 본 논문은 음란 유해 사이트에 대한 현황과 전체 시스템의 개요, 그리고 음향분석에 기초한 음란 유해 사이트 차단 방법에 대해 주로 다루고자 한다.

1. 서론

최근 전 세계를 하나로 묶은 인터넷은 그 쌍방향성, 익명성 등의 순기능에 비해 이제는 그 역기능이 사회적 문제로 대두되고 있다[1]~[3]. 국적 불명의 언어를 사용하는 인터넷 채팅과 문자 메시지, 전자우편 바이러스, 인터넷 게시판에서 나타나는 부적절한 행동과 상대방에 대한 성적 모욕, 욕설의 사용, 상대방 비하 등의 행위 그리고 70만개의 육박하는 전 세계 음란, 도박, 자살 등 유해정보 사이트의 무분별한 배포행위 등이 인터넷 역기능의 주요 사례가 될 수 있다. 이중 국적 불명의 언어를 사용하는 전자우편과 인터넷 채팅은 어휘를 축약하거나 소리나는 대로 적는 방법으로 통신속도를 신속하게 하려는 경제적 동기와 일상용어와 다른 변화를 추구하고 친밀감이나 감정을 전달하고자 하는 표현적 동기, 제도와 규율이라는 경직된 현실 공간에서 벗어나 자유로움과 새로움을 경험하려는 사회 심리적 동기 등이 복합적으로 작용하여 새로운 통신용어들이 출현하는 것으로 이에 대한 분석 및 대책이 연구되고 있는 실정이다[4].

또한 인터넷 게시판에서 행해지는 부적절한 행동은 유해단어를 필터링[5]하거나 비속어 처리 프로그램을 개발[6]하는 기술적 접근 외에 실명제의 도입, 회원삼진아웃제 도입[7]등과 같은 정책적 도입까지 행해져 그 역기능

해소에 주력하고 있다. 그러나 가장 큰 사회적 문제는 음란사이트의 무분별한 배포 행위가 될 것이다. 현재 유해 정보사이트를 언어별로 분류하면 영어 다음으로 한글이 세계 2위를 차지하고 있으며 심지어 초등학생에게도 하루 평균 2개의 음란사이트를 소개하는 메일이 전달되고 있는 실정이다. 이를 위해 정부에서는 e-

메일 주소 추출기에 대한 법적 규제강화, 해외에 서버 컴퓨터를 빌려 한글로 영업하는 불법 음란사이트를 인터넷서비스제공업체(ISP)의 협조를 받아 국제 관문국 단계에서 국내 유입을 차단하는 등의 정책을 시행하고 있다, 또한 '주니어 e-메일' 계정 서비스를 보급하고 음란사이트의 카드결제를 막는 등의 방법으로 인터넷 역기능을 해결하고자 하고 있다.

정책적인 방법외에 기술적으로는 KT등이 월정액제로 음란사이트를 차단하는 서비스를 시행하고 있는데 대부분의 음란사이트 차단은 목록기반과 단어기반 등의 방법을 채택하고 있다. 그러나 현재의 실정으로는 목록기반과 단어기반만으로는 음란사이트 차단이 완벽하게 이루어지지 않아 이에 대한 기술적 방법론의 개발이 시급한 실정이다. 이를 위해 본 논문에서는 신호처리 방법에 기반한 음란 유해사이트 차단 방법에 대해 제안하고자 한다. 특히 본 논문은 전체 시스템 중 음란 유해 사이트에 대한 현황, 전체 시스템의 개요 그리고 음향 분석에 기초한 음란 유해 사이트 차단 방법에 대해 집중적으로 다루고자 한다.

2. 음란사이트의 현황과 기술적 대처에 대한 고찰

2003년 4월 기준으로 우리나라의 성별, 연령별 인터넷 이용률은 아래 <표 1>과 같다[8].

<표 1> 성별, 연령별 인터넷 이용률

연령별	남	여
6~19세	91.4 %	91.4 %
20대	92.3 %	87.3 %
30대	75.3 %	63.2 %
40대	47.5 %	30.8 %
50대이상	14.4 %	5.3 %

위의 표에서 알 수 있듯이 인구대비 청소년, 소녀들의 인터넷 이용은 거의 모두가 이용하고 있다고 할 수 있으며 따라서 이에 대한 음란사이트의 차단은 더욱더 중요한 인터넷 역기능 해소 사안이 될 수밖에 없다. 또한 음란 유해사이트는 70만개에 육박하며 이에 대한 분석을 아래 <표 2>에 나타내었다[9].

<표 2> 음란 유해사이트에 대한 분석

분석 항목	내 용
언어별 분류	영어사용(56만 4천개, 83.6%), 한글사용(6만 4천개, 9.5%) 일어사용(1만 5천개, 2.2%), 독일어(1.3%), 프랑스어(0.6%)
유형별 분류	음란사이트(66만 8천개, 98.9%), 도박사이트(6천 900개, 1.0%) 엽기, 마약, 폭력, 자살사이트(0.1%)
요일별 유해 정보사이트 이용량 분석	월(14.8%), 화(11.5%), 수(9.8%), 목(11.8%), 금(11.9%), 토(20.5%), 일(19.8%)
최근 한글 유해정보 사이트 발현률	하루 평균 268개씩 생겨나고 있으며 이는 같은 기간 전 세계 출현 (하루 평균 58개)의 45%에 해당됨

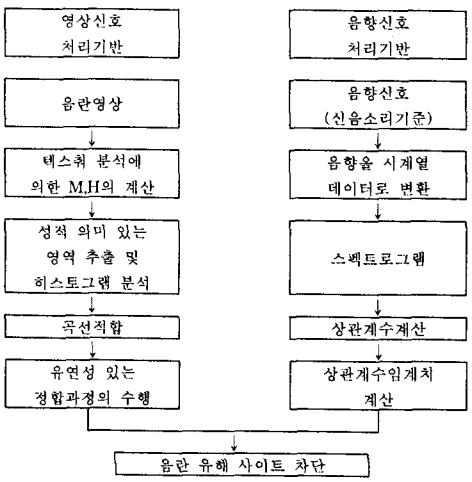
이같이 청소년들에게 막대한 피해를 주는 음란 유해사이트의 차단은 대단히 중요한 과제가 아닐 수 없다. 이는 주로 전자우편을 통해 스팸메일 형태로 전송되기 때문에 이를 사전에 차단하던지 아니면 해당 서버에 대해 요금을 유료화하는 것이 적절한 정책적 방법이 될 수 있다. 그러나 이에 대한 반론도 만만치 않아 결국은 유해사이트를 자동으로 필터링하는 기술적 방법론이 가장 적합한 방법이 될수 밖에 없다. 이를 위해 가장 많이 사용하는 방법이 목록기반의 차단기술이다[10]. 그러나 이는 가정과 같이 하나의 PC를 온 가족이 함께 사용할 때 연령별로 수시로 접속 제한을 하기가 용이치 않고 새로운 내용을 목록 DB에 수시로 업데이트 하는 것이 문제가 된다. 또한 단어를 통해 필터링하는 방법[11]은 음란 서버에서 보내는 단어가 주기적으로 교묘히 변형되고 또한 단어 자체적으로는 음란성을 나타내지 않는 등의 방법으로 단어 필터링을 피해 가는 상황이기 때문에 이를 추적하는 것도 어려운 실정이다. 결국 목록기반과 단어기반 필터링의 문제점을 보완 내지는 해결하기 위해서는 신호처리 기반의 필터링 방법이 강구되어야만 한다. 이를 위해 본 논문에서는 음향과 영상기반의 음란 유해사이트 차단방법을 제안하고자 한다. 특히 본 논문에서는 신호처리 기반 전체 시스템에 대한 개요와 음향 신호처리 기반 방법에 대한 방법에 대한 방법론을 다루고자 한다.

<표 3> 유해사이트 차단 국내 프로그램들의 비교·분석

제품 구분	지키미 2.31	파로스	i-boho	수호천사 2000	컴지기 2.0
업체 이름	인터피아 월드	아이탑	일레아트	플러스 기술	인터정보
배포 형태	유료	유료	유료	유료	유료
운영 체제	Win96, XP	Win98, XP	Win96, 2000, XP	Win98, 2000, XP	Win98, XP
차단 방법	목록, 단어	목록, 단어	목록, 단어	목록, 단어	목록, 등급, 이미지
차단 대상	음란, 폭력, 마약	음란, 폭력, 도박	음란, 폭력, 도박	음란, 폭력, 마약	음란, 폭력, 도박
자동 연결	음란, 폭력, 마약	음란, 폭력, 도박	음란, 폭력, 도박	음란, 폭력, 마약	음란, 폭력, 도박
목록 갱신	음란, 폭력, 마약	음란, 폭력, 도박	음란, 폭력, 도박	음란, 폭력, 마약	음란, 폭력, 도박
추가 삭제	음란, 폭력, 마약	음란, 폭력, 도박	음란, 폭력, 도박	음란, 폭력, 마약	음란, 폭력, 도박
내역 조회	음란, 폭력, 마약	음란, 폭력, 도박	음란, 폭력, 도박	음란, 폭력, 마약	음란, 폭력, 도박
시간 설정	음란, 폭력, 마약	음란, 폭력, 도박	음란, 폭력, 도박	음란, 폭력, 마약	음란, 폭력, 도박

3. 전체 시스템에 대한 개요

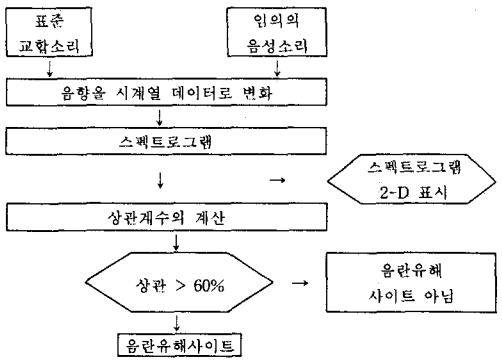
신호처리에 기반한 음란유해사이트 차단 방법에 대한 개요를 아래 (그림1)에 나타내었다.



(그림1) 전체시스템에 대한 개요

4. 음향신호처리에 기반한 음란유해사이트차단

음란 유해 사이트에서 음향신호는 음란사이트임을 알게 해 주는 주요한 요소가 아닐 수 없다. 그리고 그 내용이 한국인이든 일본인이든 아니면 서양인이든 간에 또한 성관계 형태가 일대일 형태이든지 아니면 또는 일대다 형태의 성행위 형태이든간에 성행위시에 나오는 거의 조작된 일정 형태의 신음소리 음향 신호가 존재한다. 다시 말해 성행위 장면시 의도적으로 반드시 교합소리가 나오겠끔 콘텐츠가 제작되어 있다. 본 논문에서는 이를 역으로 이용해 이같은 교합소리에 대한 표준 패턴을 만들어 놓고 이에 대한 상관계수를 계산하여 그 상관계수가 60% 이상이면 음란 유해 사이트로 간주하여 차단하고자 한다. 이 방법은 음향콘텐츠의 표절 감정 [12]방법중 음성을 따로 분리하여 멜로디 부분만을 비교해 보는 PCA(Principal Component Analysis) 또는 ICA(Independent Component Analysis)기법을 생략하여 처리를 행하는 방법이 된다. 아래 (그림2)에 음향신호처리에 기반한 유해 사이트 차단 방법에 대해 나타내었다.



(그림3) 음향에 기반한 음란 유해 사이트 차단 방법

이상을 Matlab코드를 이용하여 설명하면 아래와 같다

(가) Wave read를 통해 wav화일을 숫자 열로 변환

(나) 스펙트로그램을 구한다

specgram(a,nfft,fs,window,numoverlap) 함수를 이용한다.

(다) 스펙트로그램의 결과를 나타낸다.

(라) 상관관계를 계산한다.

r=corr(A,B)이며 이것은 Atls호와 Btls호사이의 상관계수를 구한다는 의미이며 이를 구하는 방법은 하식과 같다.

$$r=\frac{\sum_m\sum_n(Amn-\bar{A})(Bmn-\bar{B})}{\sqrt{(\sum_m\sum_n(Amn-\bar{A})^2)(\sum_m\sum_n(Bmn-\bar{B})^2)}}$$

여기서 $\bar{A}$는 Atls호의 평균값, $\bar{B}$는 B신호의 평균값을 나타낸다.

이때 이 결과가 1에 가까울수록 표준 교합소리에 근접하다는 것을 의미하며, 그 상관 관계 계산 결과가 0.6 즉, 60%이상이면 음란 유해 사이트로 판정하고자 한다.

5. 결론

본 논문에서는 신호처리기반에 기초하여 음란 유해 사이트를 차단하는 방법중 음향 신호 처리에 기반한 방법에 대해 그 방법론을 기술하였다. 현재 음란 유해 사이트는 꼭 메일 형태로 전달이 안되더라도 무료로 접속해서 볼 수 있는 사이트가 다양한 형태로 너무 많이 존재하므로 이를 차단하기 위한 방법론의 개발은 상당히 중요한 사회적 움제로 대두되었다. 특히 현재의 KT등에서 제공하는 방법은 단어기반과 목록기반에 기초한 방법들이기 때문에 이를 벗어난 음란 유해 사이트도 많이 존재한다. 이를 위해 본 논문에서는 신호처리(음향, 영상)에 기반한 방법론을 개발하고자 한 것이다. 향후 음향에 기반한 방법에 대한 실험을 수행하여 전체 시스템에 대한 유용성을 입증하는 것이 앞으로의 과제라 여겨진다.

참 고 문 헌

[1] 추병완, "사이버 윤리의 정립과 방안", 청소년보호 정책토론 자료집, 2001

[2] 한국교육학술정보원, 교육기관 정보화 역기능 방지에 관한 연구, 방문사, 2000

[3] 이재선, 전용회, "인터넷 등급 서비스를 이용한 효과적인 유해사이트 선별 기술에 관한 연구", 한국정보처리학회 추계종합학술대회 논문집, 제9권, 제2호, 2002

[4] 강승식, 김보영, "자모 빈도에 의한 통신 언어의 특성 연구", 한국정보처리학회 추계종합 학술대회 논문집, 제10권, 제1호, 2003

[5] 김치민, 김응곤, "인터넷 게시판에서 정보통신윤리 교육을 위한 유해 단어 필터링 시스템의 설계와 구현", 한국정보처리학회 추계종합학술대회 논문집, 제9권, 제2호, 2002

{6}조아영, "웹 게시판 비속어 처리 프로그램의 설계 및 구현", 한국컴퓨터산업교육학회 논문지, Vol. 2, No. 10, 2001

[7] 중앙일보, 비방하는 글 몸살. 청와대 홈페이지 삼진아웃제 도입, 2003년 7월 5일자 1면

[8] 이상경, "국내 인터넷 이용 행태 및 조사 방법", Telecommunication Review, Vol. 13, No. 3, 2003

[9] 디지털조선, 전세계 유해 사이트 68만개, 2003년 4월 27일

[10] 김재천, "인터넷 유해 사이트 차단 프로그램 분석 및 활용 방안", 홍익대학교, 2001

[11] 장윤정 외 3인, "유해 문자 가중치를 이용한 유해 사이트 차단 방법", 한국정보처리학회 춘계종합학술대회 논문집, 제10권, 제1호, 2003

[12] 이규대, 디지털컨텐츠 감정에 관한 연구, 프로그램심의 조정위원회 최종연구보고서, 2002년 11월