SNS기반 유해사이트 판단 및 수집 시스템

장정현*, 나스리디노프 아지즈*
*충북대학교 소프트웨어학과
e-mail : aziz@chungbuk.ac.kr

The system of collecting and judgement of harmful site in SNS

Jeong-Hyun-Chang*, Nasridinov Aziz*
*Dept. of Computer Science, Chung-Buk National University

요 약

소셜 미디어를 이용하는 사용자의 수가 증가함에 따라 소셜 미디어에서 공유되고 있는 유해 정보(불법, 음란)의 심각성의 대두되고 있다. 기존의 단어 DB기반의 유해 사이트 판별 방법은 단어 DB의 갱신 문제점과 유해 정보와 낮은 연관성을 가진 단어가 DB에 저장되는 문제점을 가지고 있었다. 또한 링크 주소를 짧게 해주는 Short URL 서비스를 고려하지 않아 잘못된 웹 문서를 판별 대상으로 삼을 수 있는 문제점이 있다. 본 논문에서 제안하는 유해 사이트 판별 방법은 기 구축한 유해 단어 DB에서 유해 단어를 추출하고, 추출된 단어를 포함하는 소셜 미디어상의 유해 게시물을 조회한다. 유해 단어 DB를 구축하는 방법으로, 유해 게시물 조회시 내용에 포함되는 해시태그를 저장하는 방법을 사용하여 게시물 수집과 동시에 유해 단어 DB를 갱신시킨다. 또한 유해 게시물 내용에 있는 URL 링크의 웹 문서를 문자열로 치환하여, 해당 문자열내의 유해 단어 DB에 있는 유해 단어의 등장 빈도 수를 계산하고 이를 기준치와 비교하여 유해도를 판단한다. Short URL을 사용한 URL 링크인 경우 HTTP 응답 메시지의 헤더 부에 존재하는 실제 목적지 URL 주소를 가져와 유해도 검사를 실시한다.

1. 서론

최근 휴대성과 개인성, 높은 성능을 장점으로 하는 스마트폰의 사용이 일상화가 되면서, 기존의 PC 기반의 소셜미디어을 대체하는 모바일 기반의 소셜미디어가 등장하게 되었다. 모바일 기반의 소셜미디어는 언제 어디서든 자신의 사진이나 글을 친구 또는 지인들에게 공유할 수 있다는 장점으로 인해 많은 사용자가 이용하고 있으며, 계속해서 사용자의 수가 증가하고 있는 추세이다.

하지만 이러한 소셜미디어의 장점과 반대되는 역기능이 사용자 수의 증가에 따라 점차 대두되고 있는데, 바로 소셜미디어를 통한 유해정보(불법, 음란) 유통의 문제점이다. '소셜미디어를 통한 불법·유해정보 유통 실태 및 대응 현황: 국내사례'[1]에서 조사한 결과에 따르면 한국 방송통신심의위원회에서 유해게시물로 지정하여 시정 요구 받은 총 64.446건 중 59,422건이 소셜미디어 출처로 전체 92.2%를 차지한다. 이러한 조사 결과는 소셜미디어상에서 유통되고 있는 유해정보의 심각성을 객관적으로 보여주고 있다. 기존의 유해 사이트 차단 시스템은 방송통신심의위원회가 제공하는 유해사이트 목록 DB를 바탕으로 만들어지고 있으며, '2013-2015 방송통신심의위원회 시정요구 인지방법별, 관계기관별 심의 요청 현황'[2]에서 인지방법별 현황 표를 보면 외부 신고가 77%, 모니터링 23%임을 확인할 수 있으며, 이는 유해사이트 DB 자료의 갱신이 외부 의존적인 것을 알 수 있다.

본 논문에서 유해사이트 DB의 외부 의존적인 문제점을 해결하기 위해, 소셜 미디어에서 공유되고 있는 게시물들 중 유해 정보를 포함한 게시물에 있는 URL 링크를 추출하여 웹 문서를 수집한 후 해당 사이트의 유해도를 판단하기 위해 기 구축된 유해 단어 DB를 이용하여 유해 단어 등장 빈도를 측정하고, 측정치가 설정된 기준치보다 높을 시 유해 사이트로 판단한다.

유해 단어 DB 구축을 위한 단어 수집 방법은 소셜 미디어에서 해당 단어와 연관되는 게시물을 사용자에게 제공하기 위한 해시태그 기술을 이용하여, 유해 정보를 포함한 게시물에서 사용된 해시태그 단어들을 추출하고 이를 유해 단어 DB에 저장하는 방식을 사용하였다. 이러한 방법은 유해 게시물 수집과 동시에 유해 단어 DB가 갱신이 이루어져 유해 단어 DB 최신화를 보장할 수 있다.

실제 사이트의 URL 주소를 숨기기 위해 유해사이트가 Short URL 서비스를 이용한 경우, 실제 목적지 URL 주소를 찾기 위해 Short URL를 제공하는 서버의 HTTP 메시지 헤더를 분석 및 추출하는 방법을 사용하였다. 이로 인해 기존에는 Short URL을 사용한 유해 사이트의 차단의 어려움을 HTTP 메시지 헤더 분석을 통해 해결하였다.

본 논문의 4장 실험 방법 및 결과에서는 제안하는 SNS 기반의 유해 사이트 판단 및 수집 시스템의 성능을 평가하기 위해, 시스템에서 예측한 값과 데이터의 실제 값의 발생 빈도를 나열한 Confusion Matrix를 사용하였다.

2. 관련 연구

기존에 연구된 유해 사이트 판별 기술에 관한 두 가지 연구를 제시하고, 각 연구별 단점과 본 논문에서 제안하는 방법과 상기 두 연구의 차이점을 아래와 같이 서술한다.

‘유해 텍스트 판별 기술’[3]에서는 유해사이트 판별 기법으로 텍스트 기반의 분석을 통한 판별 방법을 제시한다. 해당 논문에서 제시하는 ‘효율적인 유해 텍스트 판별 시스템’은 유해 텍스트 판별기와 학습 기반 유해 텍스트 판별기를 함께 적용한 것으로, 유해 텍스트 판별기를 통해 유해로 오인할 소지가 전혀 없는 무해 텍스트를 우선 걸러낸 후, 나머지 문서들에 대하여 학습기반 유해 텍스트 판별기를 통해 유 무해 여부를 판단한다. 유해 텍스트 판별기를 사용하기 시스템에 적용하기 위해 기계학습을 통해 학습 모델 생성이 판별 이전에 이루어져야 한다는 점과 직접 HTML 웹 문서를 파싱할 시 특수 문자, 불용어를 제거하기 위해 전처리 과정을 가져야 한다는 제약이 있으며, 전처리된 문서와 생성된 학습 모델과의 비교를 위하여 가중치 부여 과정에서의 기준의 모호성과 어플리케이션에 적용할 효과적인 알고리즘이 아직 없어 성능이 우수한지 이론적으로 알 수 없다.

‘문자 기반 유해사이트 판별 기법’[4]에서는 유해사이트판별 기법으로 유해 단어 DB를 사용하여 해당 웹 문서 내의 출현 빈도와 각 단어의 가중치를 바탕으로 유해도를 판단하는 방법을 제시한다. 정보통신 윤리위원회의 인터넷 내용 등급 판정 기준을 참고하여 각 등급에 해당하는 사이트에 사용되는 단어들을 조사해 유해 단어 DB를 구축하였고, 각 단어들의 가중치 값은 각 등급 +1로 설정하였다. 하지만 해당 방법의 경우 유해 단어 DB의 갱신이 주기적으로 이루어지지 않으며, 유해 단어 DB의 구축 방법 또한 직접 입력을 해야 하는 문제점을 가지고 있다.

본 논문에서 제안하는 SNS 기반의 유해사이트 판단 및 수집 시스템은 유해도를 평가하기 위한 기준치 설정을 위해 무작위로 추출한 유해사이트 10곳에서 사용되는 단어와 기 구축된 유해 단어 DB에 있는 유해 단어들을 비교하여 등장 빈도수를 측정하였고, 유해 단어 등장 빈도를 모두 합산한 후 평균을 낸 유해 단어의 평균 등장 빈도수를 유해도 평가 기준치로 설정하였다.

유해 단어 DB의 구축 방법은 초기 유해 단어를 한번 입력하여 해당 단어를 포함하는 SNS 게시물을 조회하고, 조회된 게시물에 사용되는 해시태그를 가져와 유해 단어 DB에 저장하는 방법을 사용한다.

게시물과 연관이 깊은 단어를 사용하는 해시태그의 특성을 이용한 방법으로, 유해한 정보를 가진 게시물에 사용되는 해쉬 태그 용어는 유해한 단어로 가정하고 유해 단어 DB에 저장을 한다. 이러한 방법은 유해 게시물을 수집하는 과정에서 지속적으로 유해 단어 DB가 갱신이 되므로, 유해 단어DB의 최신성을 보장할 수 있으며, 해당 해시태그 용어로 게시물을 조회할 시, 용어와 관련이 깊은 내용을 가진 게시물을 조회할 수 있다. 한 해시태그에 n개 이상의 게시물이 조회될 수 있고, 하나의 게시물에 복수 개의 해시태그가 존재하므로, 다양한 유형의 유해 단어 와 유해 게시물 수집이 가능하다는 장점을 가지고 있다.

3. 제안

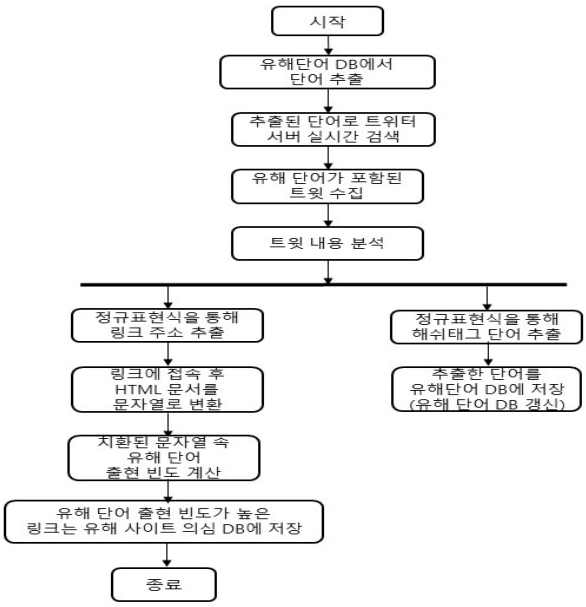
본 논문에서 제안하는 SNS 기반의 유해 사이트 판단 및 수집 시스템의 흐름도와 유해도 판단을 위한 유해 단어 등장 빈도 기준치 설정 방법, Short URL 서비스를 이용한 경우 실제 URL 주소를 추적하는 방법은 다음과 같다.

3.1 유해 사이트 판단 및 수집 흐름도

본 논문에서 제안하는 유해사이트 판단 및 수집 방법의 흐름도 순서도는 (그림.1)과 같다. 제안하는 처리 과정은기 구축한 유해 단어 DB에서 단어를 추출, 해당 단어를 포함하고 있는 소셜 미디어 게시물을 조회하고, 이를 수집한다. 수집한 게시물의 내용을 대상으로 정규표현식을 사용하여 해시태그와 URL 링크를 추출한다. 이때 추출된 해시태그의 용어가 유해단어 DB에 존재하지 않는다면, 유해 단어 DB에 저장을 하고, 이미 존재하는 용어라면 저장을 하지 않는다.

정규표현식을 사용하여 추출한 링크 주소를 대상으로 HTTP 연결을 시도하여, 해당 웹 문서를 읽어오며, 해당 문서의 내용에 대한 조작이 쉽도록 하기 위해 이를 문자열로 치환하는 과정을 진행한다.

치환된 문자열을 대상으로 유해 단어 DB에 존재하는 유해 단어들의 등장 빈도수를 측정을 하고, 이를 유해도 판단 기준치와 비교한다, 기준치보다 높은 유해 단어 출현 빈도수를 가진 웹 문서는 유해 사이트로 판단하며, 웹 문서의 링크 주소와 해당 링크가 게시된 게시물의 정보를 유해 사이트 의심 DB에 저장하여 향후 해당 게시물을 작성한 계정에 대한 제재 및 추가적인 유해 정보 공유 차단을 위한 정보를 제공할 수 있도록 한다.



(그림.1) 유해도 판단 및 단어 DB 갱신 흐름도

3.2 유해 단어 등장 빈도 기준치 설정 방법

본 논문에서 사용하는 유해 단어 등장 빈도 기준 설정 방법은 무작위로 추출한 유해사이트 5곳을 대상으로 기 구축한 유해 단어 DB에 존재하는 단어들 등장 빈도 합의 평균치를 계산하였고, 이를 유해 단어 등장 빈도 기준치로 설정하였다. 해당 기준치를 넘거나 같은 유해 단어 등장 빈도수를 가진 사이트는 유해 사이트로 판단한다. 본 논문에서 실험에 사용된 유해 단어 등장 빈도 기준치는 5로 설정되었다.

3.3 Short URL 서비스를 사용 시 실제 URL 추적 방법

Short URL은 긴 URL 주소를 짧은 형태로 제공하는 서비스이다. Short URL을 사용하여 유해 사이트의 URL 주소 변환한 경우 하나의 유해 사이트를 가리키는 N개의 Short URL 생성이 가능하다. 이로 인해 유해 게시물에 있는 Short URL 링크 주소를 차단한다고 해도 또 다른 Short URL로 다시 유해 사이트를 가리킬 수 있는 문제점이 있다. 이러한 방법으로 유해 사이트들은 URL 주소 기반의 유해 사이트 차단 시스템을 계속해서 피해왔다.

Short URL을 사용한 URL 주소의 경우 Short URL을 서비스하는 서비스 업체명이 도메인에 명시되어있다.(표.1) 실제 Short URL 서비스를 이용하여 만들어낸 URL 항목이다. (표.1)를 토대로 URL 문자열에 특정 서비스 업체명이 포함되어 있다면, 이를 Short URL을 사용한 것으로 본 논문에선 판단하였다.

(표.1) Short URL 서비스 업체별 URL 주소 변환 형태

서비스업체	원래 URL 주소	바뀐 URL 주소
bit.ly	youtube.com	http://bit.ly/ID7AM5
goo.gl	youtube.com	goo.gl/2hIJZl
durl.me	youtube.com	http://durl.me/6jfk8t

본 논문에서는 (표.1)과 같은 Short URL 서비스를 이용한 URL 주소의 실제 목적지 주소를 추출하기 위해 HTTP 응답 메시지의 헤더 내용 캡처 및 파싱 기법을 사용하였다.

Short URL 서비스를 이용한 링크를 추출하고, 해당 링크를 대상으로 HTTP GET 메소드를 이용하여 요청을 보낸다. Short URL 서비스를 제공하는 서버에서의 요청에 대한 응답으로 보내는 메시지 헤더 항목에서 Location 필드 항목에 실제 유해 사이트의 링크가 포함되어있음을 확인할 수 있었다.

따라서 추출되는 URL 링크 문자열에서 Short URL 서비스업체명이 포함되어있다면, 해당 링크를 대상으로 HTTP 응답 메시지 헤더 캡처를 통해 실제 목적지 사이트의 주소를 추출한다.

4. 실험 조건 및 결과

4.1 실험 조건

본 논문에서 제안하고, 실험에 사용된 프로그램은 Java 언어로 개발했으며, Window 운영체제 기반으로 동작한다. 또한, 여러 소셜 미디어중 가장 많은 유해 게시물 시정 요구를 받은 트위터를 대상으로 실험을 진행하였다[1].

트위터에서 제공하는 API를 사용하지 않고 직접 트위터 서버를 대상으로 URL 쿼리 설정을 통해, 실시간으로 트위터에 올라오는 게시물을 대상으로 검색어와 연관되는 게시물들의 내용을 Json 파일 형태로 받도록 설정하였다.

실험에 사용된 검색어는 유해 단어 DB에 저장되어있는 유해 단어를 랜덤으로 추출하여 사용했으며, 실험은 두 차례 진행되었고 실험에 사용된 검색어는 '야동사이트'이며 질의에 따른 추출할 자료의 크기는 100개로 제한했다.

4.2 실험 결과

본 시스템을 구동 후 100개의 유해 정보 포함 게시물을 추출하였으며, 이중에서 실제 유해 사이트 URL을 포함하고 있는 게시물의 수는 78건이며, 22건은 유해하지 않은 게시물이었다.

첫 번째 실험으로, 본 논문에서 사용하는 유해 단어 등장 빈도 기준치를 바탕으로 유해도를 평가한 결과 유해한 게시물로 판단한 게시물의 수는 68건 이며, 유해하지 않다고 판단한 게시물의 수는 32건 이였다.

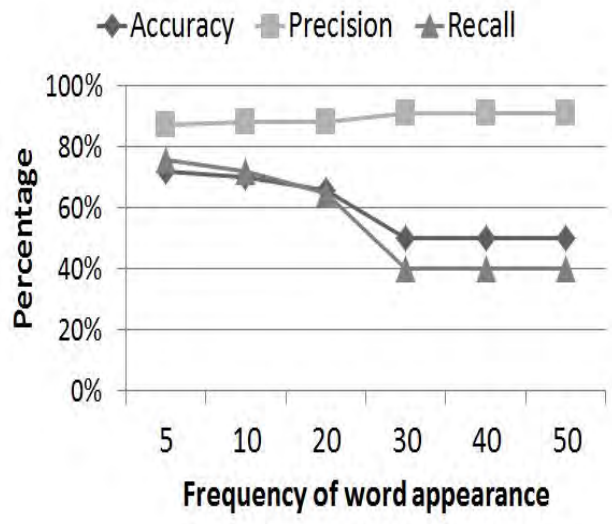
두 번째 실험으로, 본 논문에서 사용하는 유해 단어 등장 빈도 기준치 값의 변화에 따른 실험 결과 수치의 변화를 파악하기 위해 동일한 모집단 데이터를 대상으로 기준치 값을 증가시켜 동일한 실험을 진행하였다.

위 두 실험의 결과를 분석하기 위하여 Confusion Matrix를 사용하였는데, Confusion Matrix는 실제 값과 예측 값을 나열한 표로, 사용된 알고리즘의 성능을 평가하는 지표로 많이 사용 된다. 본 실험에서 Confusion Matrix(혼동 배열)를 이용해, 메트릭 계산식을 통해 얻은 결과 수치는 다음(표.2)과 같다.

(표.2) Confusion Matrix를 사용해 도출한 실험 결과

Metric	Frequency of word appearance					
	5	10	20	30	40	50
Accuracy	0.72	0.7	0.66	0.5	0.5	0.5
Precision	0.87	0.88	0.88	0.91	0.91	0.91
Recall	0.76	0.72	0.65	0.4	0.4	0.4

(그림.2) 기준치 변화에 따른 실험 결과 수치의 변동



5. 실험 결과 분석

본 논문에서 제안한 시스템의 실험 결과를 대상으로 메트릭 계산식을 통해 계산한 결과(표.2)에서 기준치 값이 5인 첫 번째 실험의 수치 결과를 보면, 전체 예측에서 (유해와 무해 무관하게) 올바르게 예측한 비율을 의미하는 Accuracy 값이 0.72로, 소셜 미디어에서 게시되는 게시물의 유해 또는 무해 여부를 효과적으로 판단하고 있음을 알 수 있다.

또한 Y로 예측된 것 중 실제로 Y인 경우의 비율을 의미하는 Precision 값이 0.87로 해당 시스템이 유해하다고 판단한 게시물들의 대다수가 실제로도 유해했음을 알 수 있다.

기준치 변경에 따른 실험 결과 수치 변동 결과(그림.2)를 보면 기준치 값이 높아질수록 Accuracy와 Recall의 수치가 내려가고, Precision의 수치가 높아지는 그래프를 얻었다. 기준치 값은 유해 사이트 내에 등장해야하는 유해 단어의 최소 등장 빈도수이므로, 높은 기준치를 설정하여 판단한 유해 의심 사이트는 많은 유해 단어를 포함하고 있어야 한다. 또한 기준치의 증가에 따라 유해 단어 등장 빈도수가 높아져야 하므로, 이보다 낮은 유해 단어 등장 빈도수를 가진 유해 사이트들은 유해하지 않다고 판단하게 된다. 따라서 높은 기준치 설정을 통하여 판단한 유해 의심 사이트 목록은 유해 여부에 대하여 높은 신뢰성을 보장하지만, 유해 여부에 대한 판단 자체의 신뢰도는 떨어지게 된다는 결과를 얻게 되었다.

또한 기준치 값이 30~50으로 증가하는 구간에서 Accuracy와 Precision, Recall의 수치가 같아지는 현상이 나타났다. 이는 모집단의 크기가 100으로 제한되어 있고, 기준치 값이 30~50으로 설정되어있을 때 유해 사이트로 판단되는 사이트들이 모두 50개 이상의 유해단어를 포함하고 있기 때문에 모두 동일한 값을 가지게 된 것으로 추정된다.

6. 결론

본 논문에서 제안하는 시스템에서 사용하는 유해도 판단 기법은 유해 단어 등장 평균 빈도수에 기반하고 있다. 이로 인해 유해 의심 사이트에 대하여 유해도를 판단할 때 단어의 등장 유무를 파악해야하기 때문에 해당 사이트에 대한 유해도 판단 시간이 길어질 수 밖에 없다. 또한 실험 결과에서 서술한 것 과 같이 기준치보다 낮은 등장 빈도를 가지는 유해사이트는 유해하지 않다고 판단할 수 있다. 이를 보완하기 위해 추후 용어 빈도와 용어 가중치를 활용하는 용어 유해도 판단 모델을 구축하고, 이를 바탕으로 유해도 판단 시간을 단축시켜 제안된 시스템의 성능을 향상시키고자 한다.

ACKNOWLEDGEMENT

"본 논문은 교육부가 지원하고 충북대학교가 수행하는 지역선도대학육성사업의 지원을 받아서 수행되었습니다."

참고문헌

[1] 김경환 "소셜미디어를 통한 불법·유해정보 유통 실태 및 대응현황: 국내사례" 방송통신심의위원회 방송통신 심의동향, 2016

[2] 한국 인터넷 투명성 보고서 "2013-2015 방심위 시정요구 인지방법별, 관계기관별 심의 요청 현황" http://transparency.or.kr/analysis/1345

[3] 김영수, 남택용, 장종수 (2005). "유해 텍스트 판별 기술" 한국통신학회 학술대회 및 강연회, 345-349.

[4] 정규철, 이진관, 박기홍, 이태헌 (2004). "문자 기반 유해사이트 판별 기법" 컴퓨터교육학회 논문지, 83-91