사이트 내 배너광고를 이용한 유해사이트 판별 기법에 대한 연구

박시현, 유성민, 송동호, 이광재*
*상명대학교
{201821235, 201821241, 201821406}@sangmyung.kr, *begleam@smu.ac.kr

A Study on Harmful Site Discrimination Methods using Site Banner Advertisements

Si-Hyeon Park, Seong-Min You, Dong-Ho Song, Kwangjae Lee*
*Sangmyung Univ.

요 약

본 논문에서는 사이트 내 배너광고 분석을 통한 유해사이트 판별 기법을 제안한다. 유해사이트가 배너 광고를 주 수입원으로 다수 게재하며 운영한다는 점에 초점을 맞춰 유해 여부를 판별하였다. 먼저 웹 크롤링을 통해 2,557개의 배너광고 이미지들을 수집하고 배너광고 특징 및 유사성을 분석할 수 있는 6종류의 수집 데이터를 제작하였다. 그리고 OCR을 통해 배너광고 속 문구를 추출하고 공통으로 발견되는 핵심 키워드를 추출하였다. 마지막으로 Average Hash를 통해 특정 다수 사이트 내 유사한 배너광고 또한 찾아내었다. 실험은 웹사이트를 입력하면 유해사이트의 특이점 및 키워드를 도출하는 방식으로 진행하였고 그 결과 유해사이트가 판별됨을 확인할 수 있었다. 본 논문에서 제안한 배너광고와 관련된 판별 방법 및 수집 데이터를 활용하면 청소년 도박, 웹툰 불법 유통, 음란물 등 사회에 손해를 끼치는 유해사이트 근절에 도움을 줄 수 있을 것으로 기대한다.

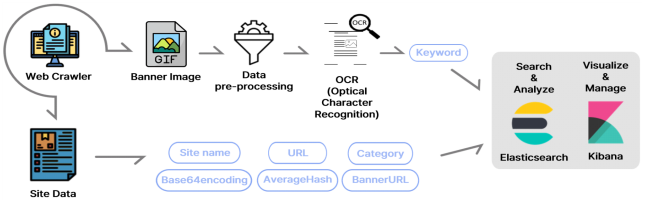
Ⅰ. 서론

유해사이트란 음란물, 도박, 마약 거래, 자살, 저작권 위반 등의 정보가 들어간 사이트를 말한다[1]. 이런 유해사이트에 의해 발생하는 피해는 웹툰 불법 유통이 약 6조에 달하고, 온라인 불법 의약품 판매가 2022년 한 해 동안 약 2만 건이 적발되는 등 사회적인 문제로 대두되고 있다[2], [3]. 그리고 유해사이트에 접속한 청소년이 불법 도박 사이트와 성매매업소 사이트로 연결되는 배너광고에 무차별 노출되고 있으며 청소년 불법 도박으로 이어지고 있다[3]. 현재 웹사이트 신고와 같이 하나씩 대응해서 유해사이트를 접속 차단하고 있지만, 다양한 기법으로 회피하는 유해사이트에 전부 대응하는 것은 현실적으로 불가능하다[4]. 따라서 웹사이트의 유해 여부를 자동으로 판단하는 방법이 필요하다.

본 논문에서는 유해사이트 속 배너광고를 이용하여 유해사이트를 판별하는 방법을 제안한다. 해당 시스템을 활용한다면 주 수입원으로 두고 있는 배너광고를 색출될 수 있는 치명적인 약점으로 작용할 수 있으며 배너광고를 통해 이동하는 또 다른 유해사이트도 색출하여 음란물, 불법 의약품 거래, 불법 도박 사이트 등 연쇄적인 색출이 가능하다는 장점이 있다.

Ⅱ. 본론

본 논문에서 분석한 유해사이트는 다음과 같은 공통적인 특징을 갖는다. 첫째, 사이트는 불법 배너광고를 통해 수익을 창출하기 때문에 다수의 배너광고가 존재한다. 둘째, 불특정 다수의 사이트에서 동일/유사한 배너광고를 반복 사용한다. 셋째, HTML 코드 속 <a> 태그 안에 배너광고의 이미지 파일과 누르면 이동되는 유해사이트 URL 링크가 존재한다. 넷째, 사정기관의 단속을 회피하기 위해 주기적인 URL 변경 패턴을 보인다[4]. 본 논문에서는 배너광고 특징을 분석하여 7개의 수집 데이터를 선정하였다. 그림 1은 본 논문에서 제안하는 시스템의 개념도이다. 웹 크롤링을 통해 6종류의 수집 데이터와 배너광고 이미지를 수집한다. 배너광고 이미지를 전처리, OCR 과정을 통해 배너광고 속 키워드를 추출한다. 수집된 7종류의 데이터는 ELK stack을 이용해 데이터베이스를 구축하고 관리하였다.



[그림 1] 제안하는 시스템의 개념도

1. 웹크롤링을 통한 배너광고 및 관련 데이터 수집

앞서 언급한 대로 유해사이트의 특징을 분석하여 사이트명, URL, 범주, Base64 encoding, Average Hash, 배너 URL, 키워드, 총 7종류의 데이터를 선정하여 수집하였다. 사이트명과 URL은 배너광고를 포함한 사이트 이름과 URL이다. 해당 정보를 통해 같은 도메인에서 URL을 일부 변경하는 사이트를 판별할 수 있다. 범주는 유해사이트에서 제공하는 주요 서비스를 기준으로 도박, 토렌트, 포르노, 스트리밍, 웹툰으로 분류하였다. Base64 encoding는 배너광고 이미지 바이너리 파일이며 Average Hash는 유사 이미지를 색출하는 방법으로 이미지 픽셀의 해시값이다. 키워드는 배너광고에서 OCR을 통해 발견되는 20개의 단어이며 불법 배너광고 게재 여부를 알 수 있으며 해당 데이터로 유해사이트를 판별한다.

2. OCR을 통해 배너광고 속 문구를 추출

본 연구에서는 이미지 속 단어들을 키워드로 규정하고 공통으로 등장하는 키워드를 중심으로 20개를 선정하였다. OCR과 빈도 분석을 이용하여

핵심 키워드를 추출했다. 핵심 키워드로는 '첫충', '매충', '카지노', '코드', '돌발', '미니게임' '룰렛' 등이 있다. 대부분의 배너 광고는 움직이는 GIF 이미지이다. GIF 이미지는 여러 개의 정적 이미지들로 연결되어 이루어져 각각 담고 있는 정보가 조금씩 다르다. 그렇기에 배너광고 속에 있는 모든 텍스트 정보를 얻기 위해서 프레임 분할을 수행했다. 그리고 이미지 크기 및 회색조(Grayscale)로 변환하여 문자 인식률을 높였다.

[그림 2] 배너광고 속 문구를 추출하는 OCR 전처리 과정

현재 가장 잘 알려진 오픈 소스 OCR 엔진에는 Tesseract OCR, Easy OCR이 있다. 본 논문에서는 수집한 배너광고 중 임의로 선정하여 각 엔진들의 인식률 비교를 진행했다. 다양한 비교를 위해 EasyOCR 엔진에 배너광고에 특화된 12,500개 학습데이터를 학습시킨 사용자 학습 EasyOCR도 비교 대상으로 추가하였다. 식 (1)은 이미지 속에 있는 모든 문자에서 모델이 정확하게 인식한 문자의 비율을 정의한다.

$$Accuracy = \frac{Number\ of\ Correct\ Predict\ Characters}{Number\ of\ Characters \in the\ Ground\ Truth} \times 100\% \quad (1)$$

표 1에서 보는 바와 같이 Tesseract OCR은 63% 인식률을 보였고 EasyOCR은 84%를 보였다. 하지만 사용자 학습 EasyOCR의 인식률은 측정할 수 없었다. 이 엔진은 훈련과정에서의 인식률은 높으나 실제 배너광고에 적용하여 OCR을 수행하면 낮은 인식률을 보이며 대부분 오인식이 되는 현상이 나왔다. 이러한 이유는 배너광고마다 폰트와 크기가 다른 점과 학습 데이터셋의 부족으로 오히려 성능이 떨어지기 때문이다. 따라서 본 작품에서는 기본 EasyOCR을 사용했다.

[표 1] OCR 정확도 비교

	Tesseract OCR	기본 EasyOCR	사용자 학습 EasyOCR
Accuracy	63%	84%	-

3. AverageHash를 통한 유사한 배너광고 찾기

다수의 유해사이트에서 동일/유사한 이미지를 사용한다는 점에 초점을 두어 유사 이미지를 판별할 수 있는 Average Hash 알고리즘을 채택하였다[5]. Average Hash는 이미지에 고유한 fingerprint를 부여하기 위해 다양한 방법으로 압축한 이미지를 해시 함수를 이용하여 대표할 수 있는 하나의 값으로 변환하는 방법이다. 본 논문에서는 추출한 이미지와 데이터베이스에 저장된 이미지의 Average Hash의 차를 이용하여 유사도를 구하였다. 이 값은 0에 가까울수록 동일/유사한 이미지로 판별한다.

Ⅲ. 실험 결과

유해사이트의 판별 시연은 선정한 수집 데이터 중 사이트명, Average Hash, 키워드를 활용하였다. 먼저 사이트명으로는 유해사이트로 사용된 사이트 이름인지 확인하여 판별 가능했다. 다음으로 데이터베이스 내의 Average Hash 값과 현재 Average Hash 값의 차이를 계산하여 0 또는 일정한 수 이하이면 유해사이트 배너광고가 게재된 것을 간주했다. 마지막으로 이미지에서 유해사이트 불법 배너광고에서 주로 사용하는 특정 키워드가 발견되는 경우 유해사이트로 판별했다. 그림 3은 정상사이트인 네이버 쇼핑과 유해사이트인 xx닷컴을 판별 시스템에 적용한 결과이다.

[그림 3] 유해사이트 판별 시연

Ⅳ. 결론

본 논문은 유해사이트의 특징 분석을 위해 유해사이트의 2,557개의 배너광고 이미지 및 관련 데이터를 수집하였다. 추가로 OCR을 이용해 키워드를 추출하고 데이터베이스에 삽입하였다. 수집된 데이터베이스와 검사 대상의 배너광고 속 문구를 핵심 키워드와 비교하였고, 배너광고 이미지의 Average Hash 또한 비교하여 유해사이트 판별이 가능함을 확인하였다.

본 논문에서 제안하는 시스템을 활용하면 주 수입원인 배너광고를 치명적인 약점으로 바꿀 수 있으며 또한 배너광고를 통해 또 다른 유해사이트로 이어지므로 연쇄적으로 유해사이트를 색출할 수 있다는 장점이 있다. 해당 시스템을 발전시켜 활용한다면 유해사이트로 인하여 발생하는 청소년 불법 도박, 웹툰 불법 유통, 불법 의약품 거래, 음란물 등 사회적인 피해를 근절하는데 기여할 수 있을 것으로 기대한다.

참 고 문 헌

[1] J. Shin, "Methods for discriminating harmful web sites using link relations between web sites," Ph.D. dissertation, Soongsil Univ., Seoul, 2014.

[2] H.-I. Park. "The surge of illegal distribution of K Webtoon... trampled up industrial growth" sedaily.com. https://www.sedaily.com/NewsView/22U7VJRF1W (accessed Jan. 4, 2023).

[3] H.-H. Kim. "73% of illegal webtoon banners are illegal gambling sites", Aju Economy. https://www.ajunews.com/view/20221019164958143 (accessed Jan. 4, 2023).

[4] J.-W. Jeong and S.-J. Lee, "Blocking method of harmful sites based on domain change pattern", *J. Digit. Forensics*, vol. 15, no. 3, pp. 39-53, 2021.

[5] E. Taskesen, "Detection of Duplicate Images Using Image Hash Functions," Towards Data Science, Jan. 29, 2022. [Online]. Available: https://towardsdatascience.com/detection-of-duplicate-images-using-image-hash-functions-4d9c53f04a75