ISSN(Print) : 1976-5304, ISSN(Online) : 2713-931X
https://kdfs.jams.or.kr, DOI : 10.22798/kdfs.2022.16.1.54

HTML 및 URL 특징을 이용한 유해사이트 수집 시스템

장 준 영*, 임 경 대*, 이 상 진**
고려대학교 정보보안학과 (대학원생)*
고려대학교 정보보호대학원 (교수)**

An Harmful site collection system using Characteristic of HTML and URL

JunYoung Jang*, Kyungdai Lim*, Sangjin Lee**
Dept. of Information Security, Korea University (Graduate Student)*
School of Cybersecurity, Korea University (Professor)**

요 약

정부는 저작권법에 의해 유해사이트를 차단하고 있으나, 유해사이트는 이를 우회하여 지속적으로 증가하고 있다. 유해사이트 이용 근절을 위해선 유해사이트들을 지속적으로 수집하여 모니터링하고 해당 사이트를 차단하거나 그로부터 얻을 수 있는 추가적인 불법 행위를 식별하는 등의 노력이 필요하다. 본 논문은 지속적인 유해사이트 수집을 위해 유해사이트의 여부를 판별할 수 있는 방법과 해당 방법을 이용해 설계한 유해사이트 수집 시스템을 제시한다. 유해사이트 판별을 위해 총 60개 유해사이트의 html 코드를 수집하고 분석하여 4가지의 특징을 식별했고, 해당 특징을 토대로 유해사이트 수집 시스템을 설계하여 유해사이트를 수집하였다. 수집 결과 98.79%의 정탐율로 743개의 유해사이트를 수집할 수 있었다. 따라서 제안한 유해사이트 수집 시스템이 유해사이트 근절에 도움을 줄 것으로 기대한다.

주제어: 디지털 저작권 포렌식, 유해사이트 탐지, 크롤링

ABSTRACT

Based on copyright law, goverment has been blocked harmful site. Nevertheless, harmful sites bypass them and continue to increase. For exterminate using harmful site, we need efforts, such as collect them continuously and identifying additional illegal activities that may result from it. For collect harmful site continuously, This paper presents method for determine whether or not a site is harmful and a system for collecting harmful sites designed using the method. we compiled 60 harmful site‘s HTML code sample and analyze them. As a result, We found four characteristics of harmful sites and designed harmful site collection system with founded characteristics. As a result of collecting harmful sites, we were able to collect 743 harmful sites with 98.79% probability of accuracy. Therefore, it is proposed harmful site collection system will be able to help exterminating harmful sites.

Key Words : digital copyright forensic, harmful site detection, crawling

※ 이 논문은 2022년도 정부(문화체육관광부)의 재원으로 한국저작권보호원의 지원을 받아 수행된 연구임(No 2022. 저작권 특화 디지털포렌식 전문인력 양성사업)
- Received 11 January 2022, Revised 12 January 2022, Accepted 31 March 2022
- 제1저자(First Author) : Junyoung Jang (Email: wkdwns420@korea.ac.kr)
- 교신저자(Corresponding Author) : Sangjin Lee (Email : sangjin@korea.ac.kr)

Ⅰ. 서 론

우리나라는 저작권법을 제정하여 저작물에 대한 불법 복제 및 전송을 차단하고, 저작권을 침해한 경우에 대한 수사를 진행할 수 있는 제도를 마련하였다. 또한 국내에서 유통되던 불법 저작물에 대한 단속을 피하여, 해외 서버에 개설한 유해사이트가 저작권을 침해하면 한국저작권보호원과 방송통신심의위원회의 심의를 거쳐 해당 사이트 및 게시물에 대한 접속을 차단하는 제도가 존재한다. 그뿐만 아니라, 범정부차원에서도 2017, 2018년도 주요업무계획을 통해 저작권 정책 방향을 언급하였고, 이를 통해 문화체육관광부에서는 링크 사이트에 대한 불법 유통 대응 방안을 제시하였다[1].

그럼에도 불구하고, 2017년 불법 복제물 유통 실태조사[2]에 따르면 만 13세 이상 69세 이하의 사람들 중 40% 이상이 '불법 복제물을 이용한 경험이 있다'고 응답하였을 정도로 대중들의 불법 복제물에 대한 접근이 쉽고 빈번하다는 것을 알 수 있으며, 저작물 시장의 침해 금액 또한 2조 5,600억원대로, 2016년 대비 7.6%가 증가하는 등 지속적으로 저작물 침해 규모가 커지고 있다. 2022년에도 여전히 인터넷상에는 다양한 유해사이트가 존재하고, 이러한 유해사이트는 저작권법에 의해 차단되더라도 금세 신규 사이트로 재 오픈하여 저작권자들의 정당한 권익과 저작권 산업을 침해하고 있다. 따라서 이러한 유해사이트들을 수집하여 모니터링하고 해당 사이트를 차단하거나 그로부터 얻을 수 있는 추가적인 불법 행위를 식별하는 등의 노력이 필요하다.

이에 본 논문에서는 유해사이트의 특징을 분석하여 정상 사이트와 토렌트, 웹툰, 드라마, 영화, 성인동영상 등을 유포하는 유해사이트를 구분할 수 있는 기술을 제안한다. 실험을 위하여 유해사이트의 html 코드 100여개와 합법 사이트의 html 코드 60여개를 수집하여 유해사이트의 특징을 분석하고, 해당 특징을 이용하여 390분 동안 사이트를 수집하여 분석한 결과 743개의 유해사이트를 98.79%의 정탐율로 수집할 수 있었다. 논문의 구성은 다음과 같다. 2장에서 관련 연구 및 본 논문이 해당 분야에서 기여 할 수 있는 아이디어를 제시한다. 3장에서는 기존 수집된 유해사이트를 대상으로 분석한 결과 식별된 유해사이트의 일반적인 특징을 서술한다. 4장에서는 3장에서 서술한 유해사이트 특징을 이용해 설계한 유해사이트 수집 시스템을 설계한다. 5장에서는 유해사이트 판별 기준을 정하기 위한 실험과 설계한 수집 시스템으로 유해사이트를 수집한 결과를 통해 제시한 유해사이트 수집 시스템의 성능을 측정하고, 일부 오탐의 원인을 식별한다. 마지막으로 5장에서 결론으로 마무리한다.

Ⅱ. 관련 연구

웹 크롤링이란 조직적, 자동화된 방법으로 월드 와이드 웹을 탐색하는 일련의 방법을 의미한다. Manning 등[3]에 의하면 크롤러는 유용한 페이지를 가져올 수 있어야 하고(Quality), 자원을 효율적으로 이용할 수 있어야 한다(Performance and effciency). Linxuan[4] 또한 Quality와 Performance and effciency는 크롤러에 있어 반드시 필요한 속성이라고 하였다.

크롤러는 크롤러에서 아직 수집하지 못한 URL을 가져오는 URL Frontier, 지정된 페이지 중 하나를 가져오는 DNS Resolution, http을 통해 웹페이지 내용을 가져오는 Fetcher, 가져온 웹에서 텍스트와 링크세트를 추출하는 Parser, 검색된 URL 중 중복되거나 이전에 찾았던 링크들을 걸러내는 Eliminator로 구성되어 있다고 Manning 등[3]이 주장하였고, Linxuan[4]는 Seed URL를 제공하는 scheduler, 웹페이지의 정보를 받아오는 downloader, 특정 전략에 따라 필요 정보를 추출하고 추가 URL을 식별하는 information extractor, 중복 URL을 수집하지 않게 하기 위한 URL Queue로 구성되어 있다고 하였다. 즉, 크롤러는 기본적으로 페이지에서 유용한 정보만을 가져오는 기능과, 중복 URL을 제거하는 기능, 페이지를 다운로드 하는 기능으로 구성되어 있다.

Kiryong Lee 등[5]은 google 검색을 통해 유해사이트 후보를 가져오고, 해당 후보들 중에서 웹 페이지에 존재하는 태그의 순서와 배치를 분석하여 신규 유해사이트를 탐지하였다. 그들은 최장공통부분수열(LCS) 알고리즘을 통해 기존에 차단된 유해사이트의 html 태그 순서와 신규 식별한 사이트의 html 태그 순서의 유사도를 측정하여 재 오픈한 유해사이트를 탐지하였다. Angelo 등[6]은 웹 사이트와 기존 피싱사이트의 레이아웃 구조 간 유사성을 이용하여 피싱사이트를 탐지하였으며, Maurer 와 Herzner[7]는 웹사이트와 기존 피싱사이트의 픽셀 데이터 간 유사성을 통해 피싱사이트를 탐지하였다. 이러한 방법은 기존에 차단된 사이트의 소스코드를 재사용하여 오픈한 유해사이트는 탐지할 수 있지만, 알려지지 않은 신규 사이트를 추가로 식별할 수 있는 보장이 없다는 한계점이 있다. 또한 웹페이지의 모든 내용을 각각의 사이트와 비교하는 방법은 유해사이트 탐지의 정확도를 높일 수 있는 반면, 많은 연산이 필요하다는 단점이 있다.

Roopak와 Tony[8]는 구글 검색으로 피싱사이트 후보를 가져왔으며 해당 후보들의 HTML 코드 간 코사

인 유사도를 측정하여 피싱 사이트를 탐지했다. 이러한 방법은 유해사이트 후보 중 실제 유해사이트의 비중이 높지 않기 때문에 높은 정확도의 분석 기술이 필요하지만, 검색 대비 유해사이트 탐지율은 적다는 한계점이 존재한다.

Joonho Sa와 Sangjin Lee[9]는 피싱사이트를 탐지함에 있어서 Http referer 등 피싱사이트가 가지고 있는 간략한 특징을 감지하여 패턴분석이나 코드 유사도 측정을 이용한 피싱사이트 탐지 기법보다 범용성과 효율을 높힐 수 있는 방법을 제시했다.

따라서 기존 유해사이트 탐지 방법과 달리 식별된 유해사이트를 대상으로 크롤링하고, 유해사이트가 가지고 있는 특징을 식별한 뒤, 유해사이트가 가지고 있는 특징을 통해 관련사이트들을 수집하면 많은 연산을 필요로 하지 않으므로 높은 Perfomance and effciency와 Quallity를 보장할 수 있고, 기존 유해사이트와의 유사성이 아닌 일반적인 유해사이트의 특징을 이용하므로 알려지지 않은 신규 유해 사이트 역시 식별할 수 있다. 본 논문에서는 해당 아이디어를 통해 비교적 빠른 속도로 신규 사이트를 포함한 유해 사이트를 수집하는 방법을 제안한다.

Ⅲ. 유해사이트 특징

다른 사이트들과 구별되는 유해사이트만의 특징을 비교하기 위해 고려대학교에서 연구 중인 링크모음사이트에서 추출한 유해사이트 데이터베이스를 분석하였다[10]. 문서 전체를 수집하여 비교하는 것은 정확도를 큰 폭으로 높일 수 있지만, 비교하는 데이터가 많을수록 크롤러의 속도는 낮아지기 때문에 유해사이트에서 식별되는 다수의 링크를 대상으로 판별하기 위해선 최대한 간단하고 짧은 데이터를 통해 특징을 구분해야 한다. 유해사이트를 조사한 결과 다음과 같은 특징을 확인할 수 있었다.

3.1 도메인에 붙는 시퀀스 번호

유해사이트는 사이트가 적발되는 것을 방지하기 위해 주기적으로 도메인을 바꾸며, 사용자들이 바뀐 도메인을 쉽게 찾고 도메인 생성을 편하게 하기 위해 메인 도메인의 마지막에 번호를 붙이고, 해당 번호를 증가시키는 방법으로 도메인을 바꾼다. 해당 시퀀스 번호는 일반적인 사이트에서는 발견되는 경우가 매우 적으나, 유해사이트에서는 자주 드러나는 특징이다.

3.2 HTML 메타 데이터

HTML에서는 여러 태그로 데이터를 표시하는데, 이 중 head 태그 안에 표현되는 자식 태그에는 사이트의 메타 정보가 표시되어있는 meta 태그가 존재한다. 이 태그 안에는 사이트의 키워드 등에 대한 정보가 포함되어있고, 유해사이트는 인터넷 검색하는 사용자에게 사이트를 노출하기 위해 [그림 1]과 같이 해당 meta 태그 안에 저작물 관련 키워드를 삽입한다.

```
<meta name="robots" content="index, follow" />
<meta name="apple-mobile-web-app-title" content="마나팡 - 무료만화, 유료만화, 성인만화, BL만화, 인기만화, 망가, 동인지, 일본만화 미리보기" />
<meta name="title" content="마나팡 - 무료만화, 유료만화, 성인만화, BL만화, 인기만화, 망가, 동인지, 일본만화 미리보기" />
<meta name="subject" content="마나팡 - 무료만화, 유료만화, 성인만화, BL만화, 인기만화, 망가, 동인지, 일본만화 미리보기" />
<meta name="description" content="만화 미리보기 No.1 마나팡, 마나팡에서 최신 만화를 찾아보세요!" " />
<meta name="keywords" content="마나팡, 만화마나팡, 마나팡만화, 만화순위, 만화미리보기, 무료만화, 네이버만화, 다음만화, 만화사이트, BL만화, 레진코믹스,
<meta property="og:title" content="마나팡 - 무료만화, 유료만화, 성인만화, BL만화, 인기만화, 망가, 동인지, 일본만화 미리보기" />
<meta property="og:type" content="article" />
<meta property="og:url" content="https://manapang9.com" />
<meta property="og:image" content="https://manapang9.com/web_cdn/img/logo/ogimg_manapang.png" />
<meta property="og:description" content="만화 미리보기 No.1 마나팡, 마나팡에서 최신 만화를 찾아보세요!" " />
<meta property="og:site_name" content="마나팡 - 무료만화, 유료만화, 성인만화, BL만화, 인기만화, 망가, 동인지, 일본만화 미리보기" />
<meta name="twitter:card" content="summary" />
<meta name="twitter:title" content="마나팡 - 무료만화, 유료만화, 성인만화, BL만화, 인기만화, 망가, 동인지, 일본만화 미리보기" />
<meta name="twitter:url" content="https://manapang9.com" />
<meta name="twitter:image" content="https://manapang9.com/web_cdn/img/logo/ogimg_manapang.png" />
<meta name="twitter:description" content="만화 미리보기 No.1 마나팡, 마나팡에서 최신 만화를 찾아보세요!" " />
```

〈Figure 1〉 Keywords in HTML meta data

3.3 하이퍼링크를 포함한 이미지

유해사이트는 〔그림 2〕와 같이 수익을 얻기 위해 배너 광고를 삽입한다. 이러한 배너는 google 광고 배너 등이 아닌, 하이퍼 링크와 하위 이미지 태그를 포함한 a 태그를 사용한다.

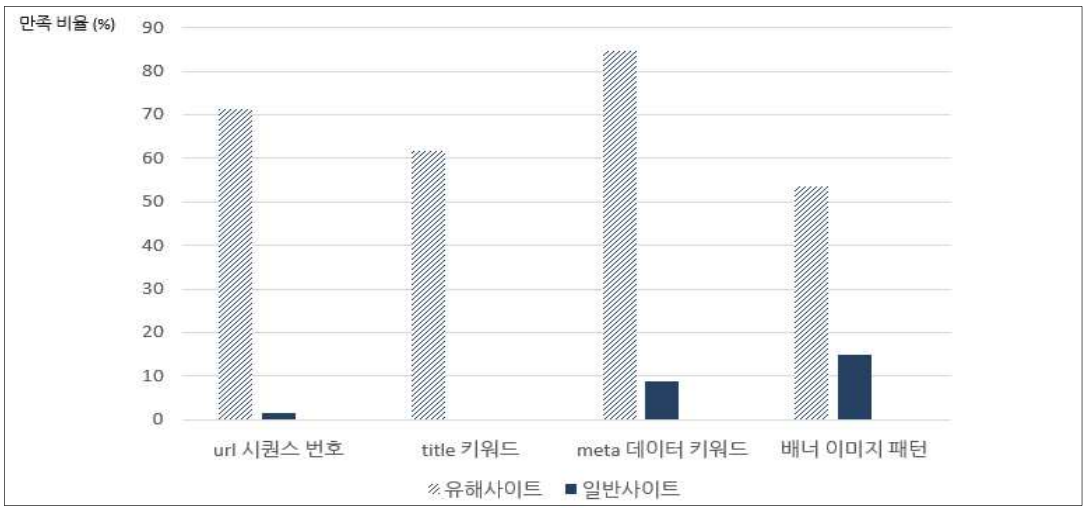
〈Figure 2〉 'img' tags under 'a' tag which include hyperlink

3.4 타이틀에 사용되는 키워드

HTML은 브라우저 상에서 사이트의 제목을 표시하기 위해 title 태그를 사용하며, 이러한 title 태그에는 사이트에 대한 특성이 나타난다. 따라서 유해사이트의 title 태그에는 높은 확률로 해당 사이트와 관련된 키워드가 포함된다.

위와 같은 특징들은 대부분의 유해 사이트에서 공통적으로 식별되는 특징이다. 조사한 112 종의 유해사이트 중 도메인에 시퀀스가 붙는 유해사이트는 90개 이상 존재했으며, 일반 사이트에도 HTML 메타데이터와 타이틀에 관련 분야 키워드가 존재하는 경우가 있었으나, 유해사이트에서 사용하는 복합적인 키워드를 포함하고 있지 않아 일반사이트와 유해사이트 간에 차이를 확인할 수 있었다. 이러한 복합적인 키워드를 포함하고 있는 유해사이트는 각각 70개 이상 존재하였다. 수집한 모든 유해사이트는 타 사이트의 하이퍼링크를 포함한 이미지를 가지고 있었고, 이는 대부분 배너 광고를 위해 사용되었다는 것을 확인했다. 일반 사이트에도 타 사이트의 하이퍼링크를 포함한 이미지 태그가 있었으나, 그 수가 적었고, 이러한 일반 사이트는 유해사이트보다 적은 수의 이미지 태그가 있는 것을 확인했다.

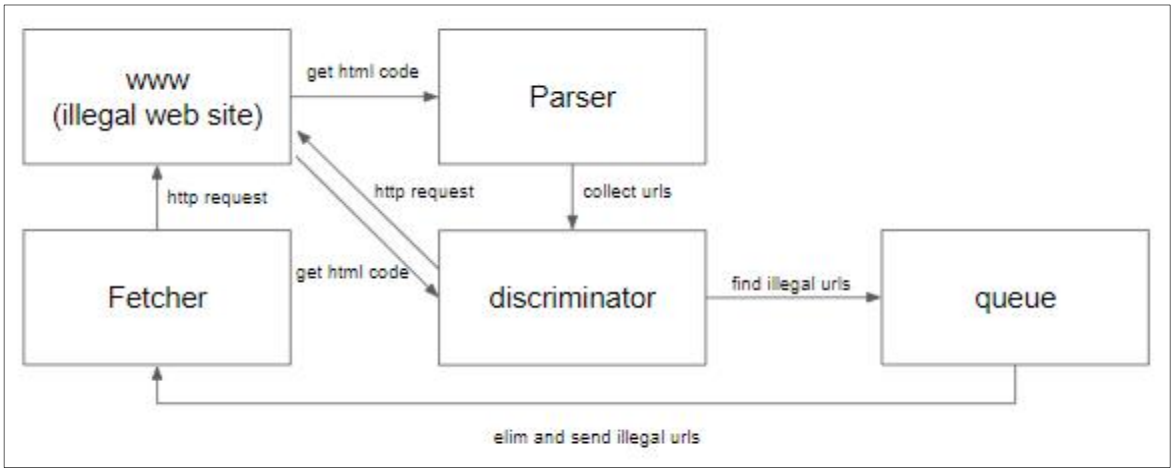
해당 특징들이 유해사이트에서만 식별되는 특징인지 확인하기 위해서 위 유해사이트 112종과 similar web에서 집계한 국내 트래픽 상위 50종 및 일반 웹툰 사이트 및 스트리밍 사이트 17종을 합친 67개의 일반 사이트와 비교하여 해당 특징들이 각각의 사이트에서 식별되는 정도를 파악한 결과 〔그림 3〕과 같이 각 특징이 발현되는 정도가 일반 사이트와 유해사이트 간에 확실한 차이가 있는 것을 확인할 수 있다.



〈Figure 3〉 Characteristic Satisfaction Rate by General and Harmful Sites

Ⅳ. 유해사이트 수집 시스템 설계

〔그림 4〕는 유해사이트의 수집 시스템 구조를 나타낸다. 수집시스템은 총 5가지 역할로 구성된다. 먼저, Fetcher는 인터넷상에서 식별한 유해사이트의 html 코드를 수집하고, Parser는 수집한 웹사이트와 중복되지 않는 URL 목록을 선별한다. 그 후 선별된 URL에 해당하는 사이트에서 discriminator가 html 코드를 가져와 실제 유해사이트인지 판별하고, 해당 사이트들의 URL을 queue에게 넘겨준다. queue는 수집된 URL들을 저장 및 관리하고, 다음 크롤링 대상 URL을 식별한다.



〈Figure 4〉 Structure of Harmful site colloction system

4.1 화이트리스트 비교

사이트에 http 요청을 통해 전달받은 html 코드에서 URL을 분해한 후, 유해사이트를 판별하기 전 불필요한 연산 및 오탐을 줄이기 위해 화이트리스트에 등록된 도메인을 필터링한다.

1) 유해사이트에서 URL 변경 공지 및 광고의 목적으로 쓰이는 SNS 도메인 (예, twitter.com)
2) 오픈소스 및 리소스를 활용하기 위해 사용하는 도메인 (예, bootstrap.com)
3) 기존 합법 저작물 관련 사이트의 도메인 (예, lezhin.com)

1)과 2)에 해당하는 도메인에 대해선 유해사이트의 여부를 판별할 필요가 없고, 3)에 해당하는 도메인에 대해선 3절의 title 및 meta 태그 데이터에 유해사이트와 중복되는 키워드가 존재해 오탐 가능성이 존재하므로 해당 도메인을 화이트리스트에 등록한 후 URL이 화이트리스트에 존재하는 도메인을 포함할 경우 유해사이트 판별 과정을 거치지 않도록 한다.

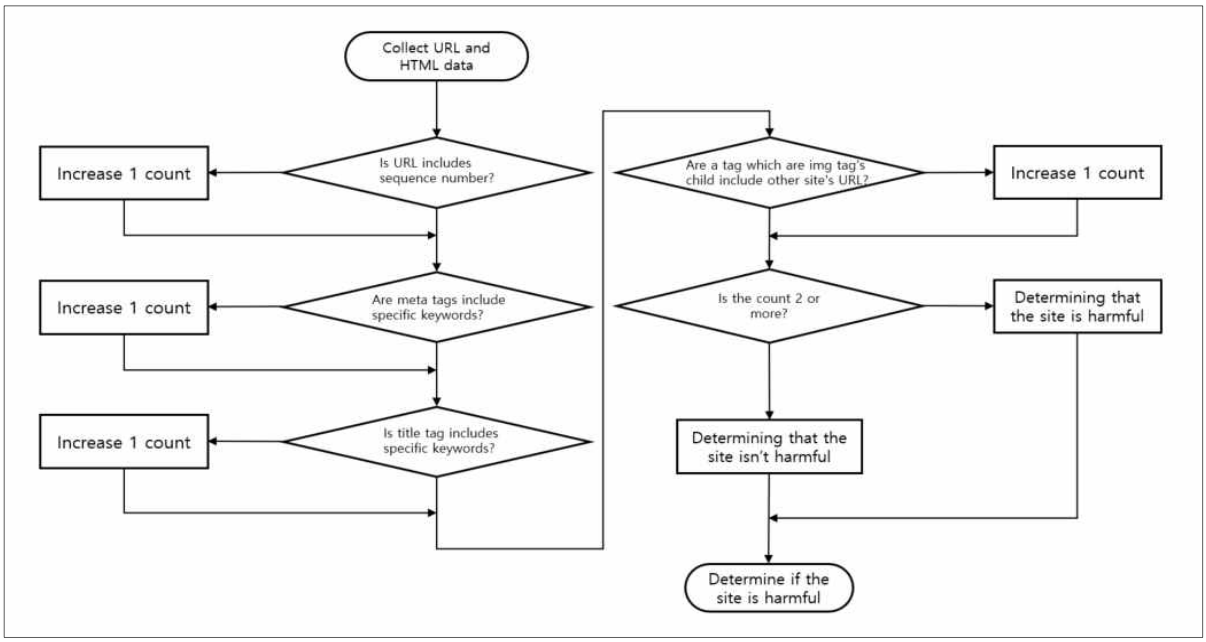
4.2 유해사이트 판별

탐지시스템은 신규 발견된 URL에서 다운로드 받은 html 코드 중, 3절에서 설명한 특징을 탐지하여 유해사이트를 식별한다.

1) URL에서 시퀀스 번호를 식별하는 경우, URL 도메인의 마지막에 숫자가 존재하는 경우 유해사이트의 특징을 만족한다고 판단한다.
2) 타이틀 명과 html 메타데이터에서 키워드를 식별하는 경우, 기존 수집된 유해사이트 샘플들에서 존재하는 키워드를 각각 수집한 뒤, 일반적인 사이트에서 범용적으로 사용하는 키워드를 제거하여 키워드 리스트를 생성하였다. 그 후 신규 식별된 사이트의 title과 meta 태그의 데이터에 생성한 리스트 내에 존재하는 키워드가 특정 개수 이상 존재할 시 해당 사이트는 유해사이트의 특징을 만족한다고 판단한다.
3) 배너 광고의 경우 먼저 img 태그를 식별하고 해당 태그의 부모태그인 a 태그의 하이퍼링크가 해당 사이

트가 아닌 다른 사이트의 URL이 포함되어있는 경우 이를 카운트 하고, 해당 링크의 개수가 특정 개수를 넘는 경우 유해사이트의 특징을 만족한다고 판단한다.

4) 위의 특징들을 특정 개수 이상 만족된다고 판단했을 때 해당 사이트를 유해사이트라고 판단한다.



〈Figure 5〉 Harmful site identification concept

유해사이트의 특징을 만족하는 기준은 수집된 유해사이트 샘플과 정상 사이트 샘플을 대상으로 실험을 진행하여 정한다. 실험 결과는 5.1에서 서술한다.

4.3 크롤링 큐

식별된 유해사이트 URL 중 중복된 URL을 제외하고 다음 크롤링 대상 URL을 정하기 위해 크롤링 큐를 사용한다. 크롤링 큐는 트리(tree) 구조를 사용하여 URL을 관리하며, 특정 URL에서 다른 유해사이트 URL을 탐지하였을 시, 해당 URL을 하위 노드로 하여 트리에 추가한다. 이는 순차적으로 유해사이트를 탐색하여 탐색되지 않는 유해사이트를 최소화 하고, 사이트들 간의 관계를 쉽게 파악하기 위함이다.

중복된 유해사이트 URL을 식별하기 위해 preorder로 해당 트리를 탐색하며, 사이트 중 크롤링 우선순위는 depth가 낮은 순위로 한다.

4.4 중복 수집된 URL 제거

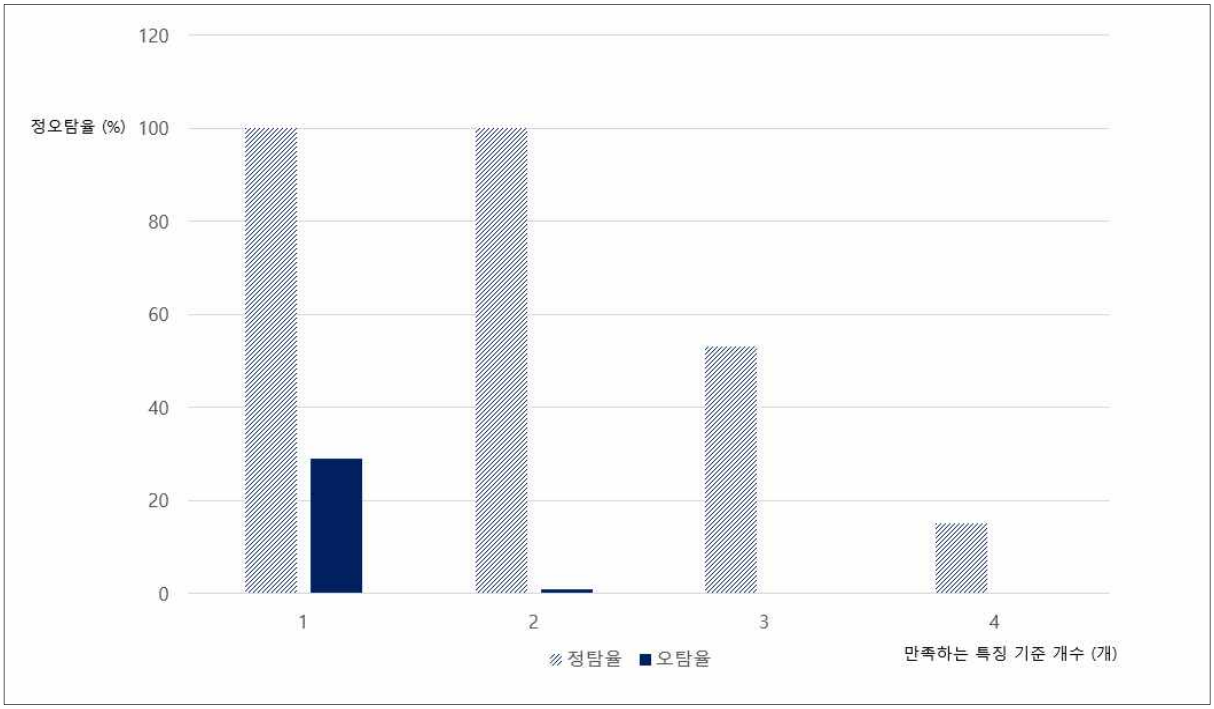
유해사이트를 크롤링 하는 중 유해사이트로 판별이 나지 않지만 대다수의 유해사이트가 해당 URL을 포함하는 경우 해당 URL에 반복적으로 접속함으로써 불필요한 연산이 발생할 수 있다. 해당 문제를 해결하기 위해 크롤링 중 유해사이트로 판별나지 않는 경우의 URL은 별도로 수집하여 관리한다. 그 후 다른 사이트에서 확보한 URL에 대하여 유해사이트 식별을 하기 전에 해당 수집된 URL 목록 중에 있는지 검사하여 해당한다면 유해사이트 식별을 하지 않게 함으로써 불필요한 연산 및 접속으로 인한 시간적 낭비를 줄인다.

Ⅴ. 실험 결과 및 분석

5.1 유해사이트 판별 기준

4.1에서 서술한 유해사이트 판별 시 유해사이트 판별 기준이 되는 특징 만족 개수를 정하기 위해 특징 만족 개수 기준 별 유해사이트 판별 정 오탐율을 실험했다. 실험 결과 〔그림 6〕과 같이 3장에서 서술한 특징들을 2번 이상 만족할 때 유해사이트 판별의 오탐율이 가장 낮고 정탐율이 가장 높은 것을 확인했다. 특징을 하나라도 만족하였을 때 유해사이트라고 판정하는 경우 정상사이트를 유해사이트로 판단 경우가 많았고, 반대로 특징

을 3개 이상 만족해야 유해사이트라고 판정하는 경우 유해사이트를 정상사이트라고 판단하는 경우가 많았다.



〈Figure 6〉 Harmful site's false positive rate by characteristic

5.2 유해사이트 수집 결과

4절에서 제안한 시스템을 검증하기 위해 2021년 5월 9일 16시부터 불법 웹툰 관련 저작물을 게시하는 사이트에서 크롤링을 통해 더 이상의 추가 URL이 수집되지 않을 때까지 유해사이트 URL을 수집하였다. 크롤링을 시작하는 시드 사이트는 다른 유해 사이트의 링크를 다수 보유하고 있는 유해사이트로 선정하였다.

실험결과는 〔표 1〕과 같다. 하나의 사이트에서 시작하여 더 이상 새로운 URL이 수집되지 않을 때 까지 수집한 결과 중복되지 않는 URL을 모두 탐색하는 데 390분이 소요되었고, 해당 시간동안 총 1244개의 URL이 수집되었다. 식별된 사이트들 중 일부는 2개 이상의 URL을 사용하고 있었는데, 중복되는 사이트를 제거한 결과 수집된 사이트는 모두 743개로, 평균적으로 하나의 사이트가 1.67개 이상의 URL을 사용하는 것을 확인할 수 있다. 해당 사이트들을 육안으로 확인한 결과 유해사이트가 아닌 사이트는 총 9개 식별되어 유해사이트 판별의 오탐율이 1.21%로 나타났다. 오탐으로 판정된 사이트는 유해사이트에서 영상, 사진 등을 인용할 때 쓰는 사이트이거나 합법적으로 저작물을 유통하는 사이트인 것으로 확인되었다. 해당 사이트들의 오탐 원인은 모두 유해사이트에서 쓰는 타이틀이나 메타 데이터에 사용하는 키워드들이 동일하기 때문인 것으로 확인됐다. 해당 키워드들은 웹툰, 드라마 등 불법과 관련되지 않은 내용이 포함된 키워드이므로 해당 키워드를 탐색 대상에서 제거하고 무료 웹툰 등 유해사이트에서 사용하는 조합된 단어들의 키워드를 탐색하는 것으로 오탐율을 낮출 수 있다.

〈Table 1〉 Crawling result

Crawled time	Number of detected harmful URLs	Number of Detected harmful sites	Number of normal sites	false positive rate
390 minutes	1244	743	9	1.21%

또한 〔표 2〕와 같이 식별된 각 사이트가 몇 개의 특징을 만족하였는지를 확인하여 유해사이트 탐지 만족 기준에 따른 정탐율을 환산하여 5.1에서의 실험 결과와 비교하였을 때, 3개를 만족했을 때와 4개를 만족했을 때의 정탐율이 거의 일치하는 것을 확인 할 수 있었다. 2개를 만족했을 때의 결과는 해당 기준으로 유해사이트를 수집한 데이터를 대상으로 통계를 냈기 때문에 1개를 만족하는 사이트와 정탐율을 비교할 수 없지만, 1개를 만족

하는 사이트의 경우 높은 비율로 오탐을 내기 때문에 수집시스템이 유해사이트 판별 기준을 특징을 2개 이상 만족하는 경우로 정한 것은 유의미한 판단이라 할 수 있다.

〈Table 2〉 Crawling result

Number of satisfied features	2	3	4
Number of Detected harmful sites	334	305	104
Calculated Accuracy	100%	55%	14%

유해사이트 수집 중 중복 수집된 URL 목록을 확인해 본 결과 유해사이트임에도 수집되지 않은 사이트가 7개 존재했다. 해당 사이트가 유해사이트로 판별되지 않은 이유를 분석해 본 결과 해당 사이트들은 직접 사용자에게 웹사이트 내용을 제공하기 전에 Capcha, 로그인 등의 인증을 사용하거나, 사용자가 웹 사이트의 주소에 2차적으로 접속하게 함으로써 한 번의 요청으로 인해 받아온 HTML 데이터에는 실제 사이트의 특징이 드러나지 않는 것으로 확인됐다.

```
<br><br>
<center><a href="/?OK=OK"><img src="/img/bam_logo.gif" border="0"><br>
<font size="3"><b>밤토끼 시즌2 보안연결</font></b>
<br>(클릭 하세요!!)</a><br><br>
<a href="/" target="_parent"><b>[연결 안될 경우 새로고침 하기]</b></a><br>(클릭 하세요!!)</a></center>
<br>
<center><iframe src="https://cartoon.inde.biz/uchat.html" style="display:inline-block; width:330px; height:500px;" frameborder=0></iframe></center>
```

〈Figure 7〉 HTML code of harmful websites that do not directly serve websites

Ⅵ. 결론

우리는 유해사이트의 HTML 코드를 분석하여 유해사이트를 판별할 수 있는 특징을 추출하고, 이를 바탕으로 유해사이트의 URL을 수집하는 크롤러를 제작하였다. HTML 코드 분석을 통해 식별한 유해사이트의 특징은 총 4가지였으며, 크롤러를 통한 실험 결과 해당 특징을 가지고 있는 사이트들은 대부분 유해사이트임을 알 수 있었다. 따라서, 우리가 제시한 4가지의 특징을 이용한다면 신규 개설 또는 재오픈한 유해사이트를 쉽게 식별할 수 있을 것으로 예상한다. 유해사이트의 특징을 추가적으로 더 식별 할 수 있다면 수집 시스템의 오탐율을 낮추고 유해사이트 수집 양을 더욱 늘릴 수 있을 것으로 예상된다. 이러한 기법으로 유해사이트를 지속적으로 수집하여 유해사이트 차단을 하고, 수집된 유해사이트 분석을 통해 수집시스템 개선 및 추가적인 불법 행위를 차단한다면 유해사이트 근절에 많은 도움이 될 수 있을 것이다.

참 고 문 헌 (References)

[1] Ministry of Culture, Sports and Tourism, 2017 Workplans. Available: https://www.mcst.go.kr/kor/s_policy/plan2021/plan2021.jsp?pTab=02#. 2017.01.06. confirmed

[2] Korea Copyright Protection Agency, Annual Report on Copyright Protection. Available: https://www.kcopa.or.kr/lay1/bbs/S1T11C283/F/25/view.do?article_seq=380. 2018.05 confirmed

[3] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, "Introduction to Information Retrieval," Cambridge University Press. pp.443-459. 2008.

[4] Linxuan Y., Yeli L., Qingtao Z., Yanxiong S., Yuning B. and Wei H., "Summary of web crawler technology research," Journel of Physics: Conference Series, Volume 1449, 012036, 2020.

[5] Kiryong Lee and Heejo Lee, "An Automated Technique for Illegal Site Detection using the Sequence of HTML Tags," Journel of KIISE, Volume 43 Issue 10, pp. 1173-1178, 2016

[6] A. P. E. Rosiello, E. Kirda, and C. Kruegel, "A Layout-Similarity-Based Approach for Detecting Phishing Pages," SecureComm, pp. 454-463, 2007.

[7] ME. maurer, D. Herzner, "Using visual website similarity for phishing detection and reporting," CHI'12 Extended Abstracts on Human factors inComputing systems, pp. 1625-1630, 2012.

[8] S. Roopak, T. Thomas, "A Novel Phishing Page Detection Mechanism Using HTML Source Code Comparison and Cosine Similarity," Advances in Computing and Communications (ICACC), pp. 167-170, 2014.

[9] Joon ho Sa and Sangjin Lee, "Real-time Phishing Site Detection Method," Journel of the Korea Instittute of Information Security and Cryptology, Volume 22 Issue 4, pp. 819-825, 2012.

[10] Seungyoung Choo, Yeseong Hwang, Sanjin Lee, "Methods for Collecting Harmful Websites
Using Web Crawling", Journel of Digital Forensics, Volume 15 Issue 3, pp. 127-138. 2021.

저 자 소 개

장 준 영 (JunYoung Jang)
준회원

2019년 2월 : 고려대학교 사이버국방학과 졸업
2020년 9월 ~ 현재 : 고려대학교 정보보안학과 석사과정
관심분야 : 디지털 포렌식, 정보보호, 리버싱 등

임 경 대 (Kyungdai Lim)
준회원

2019년 2월 : 고려대학교 사이버국방학과 졸업
2020년 9월 ~ 현재 : 고려대학교 정보보안학과 석사과정
관심분야 : 디지털 포렌식, 저작권 보호 등

이 상 진 (Sangjin Lee)
평생회원

1989년 10월 ~ 1999년 2월 : 한국 전자통신연구원 선임연구원
1999년 3월 ~ 2001년 8월 : 고려대학교 자연과학대학 조교수
2001년 9월 ~ 현재 : 고려대학교 정보보호대학원 교수
2017년 3월 ~ 현재 : 고려대학교 정보보호대학원 원장
관심분야 : 대칭키 암호, 정보은닉 이론, 디지털 포렌식