논문 2023-3-1 http://dx.doi.org/10.29056/jsav.2023.09.01

불법 저작권 침해 사이트 특정을 위한 특징 정보 수집 방안 연구

최은석*, 김영모**, 박명찬*†

Research on Methods of Feature Information Gathering for Identifying Illegal Copyright Infringement Sites

Eun-Seok Choi*, Young-Mo Kim**, Myung-Chan Park*†

요 약

최근 불법 저작권 침해 사이트는 수사 단속이 어려운 Cloudflare와 같은 클라우드 서비스 이용하고 있으며 도메인을 주기적으로 변경하는 방법으로 도메인 차단을 우회하고 있다. 기존 도메인 차단기술은 변경된 불법 사이트가 발견되어도 동일한 절차에 따라 심의하는데, 불법 사이트는 도메인 변경 주기를 단축시켜 차단을 우회하고 있다. 본 논문에서는 불법 사이트를 대상으로 사이트 로고, 메뉴, 카테고리, 콘텐츠, 광고 배너 등 사이트 내 공개된 정보를 기반으로 대상 사이트를 특정할 수 있는 특징 정보를 정의하고 수집하기 위한 방안을 제시한다. 실험 결과, 도메인 변경 전·후 사이트 특징 정보 값이 동일하게 나타났으며, 이를 통해 불법 사이트 차단 시 불필요한 검증 및 심사 기간을 단축하여 효과적인 단속과 조치에 기여할 것으로 기대한다.

Abstract

Recently, illegal copyright infringement websites have been utilizing cloud services such as Cloudflare, which makes investigation and enforcement challenging. They evade domain blocking by regularly changing domains. Existing domain blocking technologies follow the same process even when a modified illegal site is discovered and these illegal sites bypass blocking by reducing the domain change interval. In this paper, we propose an approach to define and collect feature information about target sites, such as site logos, menus, categories, content, and advertisement banners, based on publicly available information within the sites that can be utilized to identify target sites engaged in illegal activities. Through experiments, it was found that the feature information values of sites before and after domain change were the same. Accordingly, it is expected to contribute to establishing a system that can take immediate action by reducing the unnecessary verification and screening period in blocking illegal copyright infringement sites.

한글키워드 : 불법 저작권 침해 사이트, 특징 정보, 불법 사이트 특정, 저작권 침해, 웹툰, 스트리밍

keywords : Piracy Sites, Feature Information Illegal Site Identification, Copyright Violation, Webtoon, Streaming

* 에이치엠컴퍼니㈜
** 숭실대학교 컴퓨터학과
† 교신저자:박명찬
(email: myungchan.park@hmcom.co.kr)

접수일자: 2023.08.28. 심사완료: 2023.09.01.
게재확정: 2023.09.20.

1. 서 론

한국콘텐츠진흥원의 ‘2022년 웹툰 사업체 실태조사’[1]에 따르면 웹툰 산업의 전체 규모 추정 결과, 2020년 웹툰 산업 매출 규모 1조 538억 대비 약 48% 늘어난 수치로서, 한국 웹툰 산업이 지속적으로 성장하고 있음을 보여준다. 불법적인 콘텐츠 유통의 일반적 유형은 저작권을 가진 콘텐츠를 저작권자로부터 허가 없이 복제하거나 수정 및 변환하여 유통하는 행위를 말하는데[2,3], “2021 저작권 백서”[4]에 따른 유통 매체별 온라인상 불법 복제물 시정권고 현황을 보면 온라인 불법 복제물을 상대로 시정 권고를 내린 건수가 2020년 대비 30,160건이 늘었다. 웹툰뿐만 아니라 저작권의 보호를 받는 동영상을 불법 스트리밍 서비스하는 사이트들 또한 폭증하고 있다. 정보통신 기술의 발전으로 디지털 콘텐츠 소비가 급격히 증가하고 있으며 이에 따른 불법적인 복제 콘텐츠 역시 증가하여 여러 문제를 야기하고 있다[3].

현재 해외 불법 저작권 사이트에 대해서는 국내 유입을 차단하는 도메인 차단 방식을 적용하고 있다. 그러나 차단된 도메인의 주소를 변경하는 방식으로 차단을 우회하고 있으며, 현재 대부분의 불법 저작권 사이트는 Cloudflare와 같은 해외 클라우드 기반 서버를 사용하여 CDN (Content Delivery Network), 프록시 등을 구축하여 IP 주소 특정 및 추적을 어렵게 함으로써 국내 수사망을 회피하고 있다. 2021년 문화체육관광부에서 발표한 자료에 따르면 2020년 디지털 콘텐츠에 관련된 불법 저작물 사이트 신고 건수는 약 4,000건이며 사이트 접속 차단 건수는 약 400건으로 약 10%의 저조한 접속 차단율을 보인다[5].

실제로 불법 저작권 침해 사이트를 효과적으로 근절하는 가장 좋은 방법은 불법 웹서비스의 IP 주소를 특정하여 불법 콘텐츠 유포자를 검거하는 것으로 CDN 서비스로 숨겨진 불법 사이트의 실제 IP 주소를 추적[6]하는 것과 같은 방안들이 있으나 실제 적용에는 한계가 있으며 현행 제도하에서 시도할 수 있는 일반적인 방법들에서 해결 방안을 모색해야 한다. 이에 최근에는 저작권 콘텐츠의 불법 유통을 판단하기 위한 시그니처 정보 생성 방법[7]이 제안되는 등 불법 사이트 근절을 위한 노력은 다각도로 이루어지고 있다.

일반적인 온라인상 불법 정보에 대해 방송통신심의위원회(이하 방심위)의 심사 기일은 평균 10.8일이 소요[8]되는 등 기술적 및 정책적 한계로 인해 현재로서는 어려운 상황이다. 도메인의 경우 변경 간격이 점점 짧아지고 있는 것뿐 아니라, 표 1과 같이 변경 방식 또한 도메인 내의 숫자를 단순히 증가하는 방식에서 TLD(Top-Level Domain) 변경, 전체 주소 변경 등 기존의 패턴에서 벗어나는 경우가 발생하고 있다.

불법 사이트의 도메인 변경 패턴의 경우 Jeong 등에 의해 상세하게 연구[9]된 바가 있으며 표 1과 같은 방식으로 도메인이 변경된 후의 사이트를 비교했을 때 변경 전 도메인과 99.28%로 일치하는 것으로 나타났다. 도메인의 변경 유형에서 가장 많은 비중을 차지하는 번호 증가 방식의 특징은 상호명 부분의 변화 없이 오직 숫자만 증가하는 것이며 이는 확실한 특징 정보임을 알 수 있다.

표 1. 도메인 변경 방식
Table 1. Examples of Domain Change

번호 증가	변경 전	https://*****65.net
	변경 후	https://*****66.net
TLD 변경	변경 전	https://*****.org
	변경 후	https://*****.info
전체 주소	변경 전	https://********.*1.com
	변경 후	https://1********.com

따라서 짧아지는 도메인 변경 주기와 증가하는 변경 방식에 대응하기 위해 지속적인 관찰과 추적을 통해 불법 사이트를 감시해야 하며 지속적으로 도메인을 변경하는 불법 저작권 사이트에 대해 기존 원본 사이트를 특정하거나 원본 사이트와 변경된 사이트간의 동일성을 판단하기 위해 필요한 요소 및 특징 정보를 확인하고, 이를 수집하기 위한 기술 연구가 필요하다.

본 논문에서는 불법 저작권 사이트의 특징 정보를 파악하여 이를 분석 및 분류하고 정보를 자동으로 수집하여 효과적인 단속과 불법 저작권 행태를 근절할 수 있는 방안을 제안하고자 한다.

본 연구에 사용된 분석 대상 사이트의 경우 링크 모음 사이트를 크롤링하여 확보하고 이들 중에서 해외 저작물에 대한 불법 서비스와 서비스 중지 및 폐쇄된 사이트를 제외한 유효한 불법 사이트를 대상으로 테스트를 진행한다.

논문의 구성은 다음과 같다. 2장에서는 다수의 불법 웹툰 사이트, 불법 스트리밍 사이트의 유형을 분류하고 특징 정보를 선별한다. 3장에서는 분류·분석 결과를 바탕으로 각 특징 정보를 상세 정의하며 4장에서 제안한 내용에 대한 실험 결과와 5장을 마지막으로 본 논문의 결론으로 마무리한다.

2. 불법 사이트 레이아웃 및 특징 정보 정의

2.1 불법 저작권 침해 사이트 레이아웃 현황

불법 저작권 침해 사이트의 특징 정보를 수집하기 위해 총 61개의 불법 웹툰 및 스트리밍 사이트 구조를 분석 및 분류했다. 대부분의 불법 사이트의 구조는 그림 1과 같다. 일반적으로 메뉴, 광고, 콘텐츠로 크게 구분되며 메뉴에는 대상 사이트 고유의 로고가 존재한다. 광고의 경우 눈에 잘 띄는 곳에 배치하며 사이트 구조에 따라 다양한 형태를 가진다. 콘텐츠는 크게 최신, 인기, 장르로 분류되고, 하위 요일별로 구분하여 게시한다. 이는 웹툰의 경우 게시물 등록 일정이 정해져 있어 해당 요일을 기준으로 구분하는 경우가 일반적이다.

그림 1. 불법 저작권 침해 사이트 구조
Fig. 1. Piracy Website's Layout

2.2 대상 사이트 정보 수집 항목

불법 저작권 침해 사이트에서 수집할 수 있는 항목은 표 2의 항목에 따라 메뉴(로고, 메뉴, 검색창), 광고 배너, 콘텐츠, 카테고리, 로그인, 게시판, 도메인으로 구분할 수 있으며 각 정보는 사이트 레이아웃에 따라 달라질 수 있다. 대상 웹페이지에서 해당 항목이 존재할 경우 해당 항목의 레이아웃을 기준으로 분류, 비슷한 유형끼리 묶어 하나의 유형으로 분류, 분류된 유형을 대상으로 수집 대상을 정의, 마지막으로 수집 가능 유무 및 중요도를 분석하여 사이트 특정 가능 유무를 최종 판단한다.

대상 사이트 수집 항목을 바탕으로 61개의 불법 저작권 침해 사이트의 레이아웃 유형을 분석하여 콘텐츠형, 광고배너형, 카테고리형, 게시판형으로 분류했다.

표 2. 웹페이지 유형 분류 대상
Table 2. Criteria and Attributes for Classifying Webpage Layout Types

항목	예시	
메뉴	로고	
	메뉴바	
	검색창	
광고 배너		
콘텐츠		
카테고리		
로그인		
게시판		
도메인	https://b*n247.com	

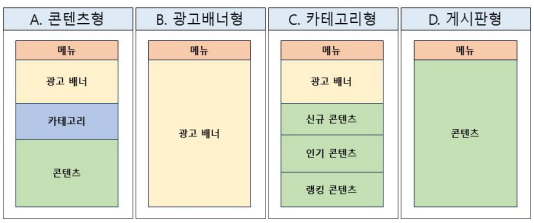

그림 2. 침해 사이트 레이아웃 유형
Fig. 2. Webpage Layout Types

콘텐츠형은 가장 일반적인 유형으로 메뉴 > 광고 > 카테고리 > 콘텐츠로 구성되며, 광고배너형은 메인 화면에 콘텐츠보다 광고를 중심으로 구성하는 것이 특징이다. 카테고리형은 콘텐츠형과 유사하나 신규, 인기, 랭킹 등의 카테고리를 이용하여 클라이언트의 편의성 및 배너 등을 통해 인기 콘텐츠를 쉽게 접할 수 있도록 한다. 마지막으로 게시판형은 광고 없이 콘텐츠만 제공하는 형태로 일반적인 불법 사이트와는 다소 차이를 보인다.

표 3. 레이아웃 유형
Table 3. Proportions of Webpage Layout Types

유형	사이트	비율
콘텐츠형	32	52%
광고배너형	9	15%
카테고리형	16	26%
게시판형	4	7%
합계	61	100%

총 61개의 사이트 유형을 분류한 결과, 표 3의 결과와 같이 콘텐츠형은 32개 사이트에서 발견되어 전체의 절반이 넘는 약 52%를 차지하였고, 광고배너형 9개(15%), 카테고리형 16개(26%), 게시판형 4개(7%)로 분석되었다.

콘텐츠형은 32개로 가장 많은 비율(52%)을 차지하며, 일반적이고 직관적인 형태를 가진다. 최상단에 메뉴 > 광고 배너 > 카테고리 > 콘텐츠 등 일자로 구성되고, 상단 메뉴에는 사이트 로고(이름)와 서비스 메뉴 등이 위치하며, 그 아래에는 광고 배너가 존재한다. 하단에는 카테고리(최신, 인기 등)와 최신 콘텐츠로 구성된다. 웹툰 서비스의 경우 요일별로 신규 업데이트가 이루어지기 때문에 요일별 카테고리가 존재하며, 스트리밍 서비스는 최신 등록 순으로 나열된다.

광고배너형은 9개로 15%를 차지하며 메인 화

면이 광고 배너 위주로 구성된 것이 특징이다. 광고배너형은 최상단에 메뉴 > 광고 배너로 구성되며, 메뉴를 통해서 콘텐츠로 접근할 수 있다. 이동 후 실제 콘텐츠가 제공되는 페이지의 레이아웃은 콘텐츠형 구조와 유사하다.

카테고리형은 16개로 사이트 26%를 차지하며 콘텐츠형 다음으로 많이 분포되어 있다. 콘텐츠 유형과 비슷한 구조이나, 콘텐츠 부분에서 최신, 인기(랭킹) 등 추천 콘텐츠 위주로 상위 5-10개 콘텐츠를 보여주는 것이 특징이다. 메뉴를 통해 웹툰 서비스에 접속하면 콘텐츠형으로 서비스를 제공한다. 특히 웹툰이나 스트리밍 서비스 등 복합 서비스를 제공하는 사이트에서 많이 사용되는 유형이며 클라이언트가 쉽게 선택할 수 있도록 배너 형태로 제공하는 것이 특징이다.

게시판형은 전체 대상 61개 사이트 중 4개로 7%를 차지하며 광고가 없는 것이 특징이다. 대부분 텍스트 위주로 작성되어 있으며 카테고리별 최근 등록된 게시물 순으로 보여준다. 일반적인 불법 사이트의 경우 광고 수입을 통해 수익을 창출하는 반면, 게시판형의 경우 특별한 수익원이 확인되지 않는다.

2.3 불법 저작권 침해 사이트 특징 정보 정의

전체 대상 사이트를 표 2의 수집 항목들과 그 항목 각각의 HTML 코드를 분석했다. 그 결과, 게시판의 경우 4개의 게시판형 사이트에서만 발견되어 효과적인 특징 정보로 수집하기엔 적은 개체수이며 로그인, 검색창의 경우 역시 개체수가 많지 않을 뿐 아니라 속성 사이트마다의 고유한 HTML 특징이 잘 드러나지 않는다. 메뉴, 콘텐츠, 카테고리 역시 웹툰, 스트리밍 서비스 특성상 어떤 사이트든 유사한 카테고리 및 콘텐츠를 제공하기 때문에 제외한다.

반면, 로고는 모든 사이트가 고유한 형태로 가지고 있다. 광고 배너 역시 게시판형 사이트를 제외한 모든 사이트들에 존재하며 각 사이트마다의 고유한 가입코드가 광고 이미지에 표시되어 있거나 HTML 소스에 적시되어 있었다. 따라서 각 사이트의 로고 이미지와 광고 배너를 주요 특징 정보로 선정하고, 도메인과 함께 분석하였다.

3. 불법 저작권 침해 사이트 특징 정보 상세 정의

3.1 로고

표 4는 사이트별 로고 이미지를 수집한 결과로서 이미지가 아닌 텍스트를 로고로 사용하는 7개 사이트를 제외하면 54개 사이트에서 고유한 이미지를 가지고 있었다. 해당 사이트가 단속 회피를 위한 도메인 변경 방법을 사용하여도 로고 이미지는 그대로 유지되는 것을 확인했다.

표 4. 로고 분석
Table 4. Logo Analysis

형태	로고 이미지 파일 링크
image	/thema/Miso-Simple/logo-top.png
image	/img/new_logo.gif
image	/img/k**logo.gif
text	text
image	/d********e_media/sites/2/2019/10/d******_logo.png
image	/files/attach/imaes/135/bdb2c8e058f8507792c60412dbfbd5d1.png
image	/thema/Miso-Simple/logo-top.png
.....	.....

그림 3. 로고 위치 예시
Fig. 3. Logo Location Example

그림 3과 같이 대부분 HTML에서 로고 이미지 위치를 확인할 수 있었으나, 일부 사이트의 경우 style에 적용하는 등 일반적인 방법으로 추출하기 어려웠고 style을 호출해야 확인이 가능했다. 로고 이미지는 주로 로고 파일명에 “logo”라는 텍스트 키워드가 포함되는 것이 특징이다.

3.2 광고 배너

3.2.1 광고 배너 유형

분석 대상 사이트에서 발견된 광고 배너를 바탕으로 광고사 비율을 조사한 결과, ‘1BET1’이 총 51개 사이트 중에 가장 많이 발견되었다. 표 5는 가장 많이 발견된 빈도를 기반으로 상위 4개의 광고를 선별한 것이며, ‘1BET1,’ ‘winner,’ ‘WIN’ 순으로 확인되었다.

표 5. 광고 노출 순위 상위 4개
Table 5. Top 4 Most Frequently Encountered Advertisements

광고	배너 이미지
• 1BET1 (51개) • 회원가입 시 가입코드 입력	
• winner (47개) • 회원가입 시 가입코드 입력	
• WIN (42개) • 회원가입 시 가입코드 입력	
• suncity (17개) • 회원가입 시 가입코드 입력	

불법 사이트의 배너광고에 관해 진행된 연구는 광고 배너를 추적하여 광고주를 분석하는 연구[10]와, 많은 배너 광고가 존재하는 특징을 이용하여 유해사이트 여부 판별을 시도한 연구[11]가 있지만 광고의 가입코드를 이용하여 대상 사이트를 특정하는 연구는 없었다.

광고 배너의 HTML 소스 코드를 확인하면 그림 4와 같이, 각 사이트마다 'regcode'와 같은 키워드가 함께 사용되며, 이 키워드와 연결된 고유한 가입코드 값이 링크에 포함되어 있는 광고들이 존재한다. 그림 4에 나와 있는 해당 사이트의 예시에서는 3330이라는 특정 가입코드를 사용하며, 광고 배너를 클릭하고 사이트에 접속하여 가입할 경우 해당 가입코드가 자동으로 적용된다. 따라서 이러한 경우 고유 가입코드가 링크상에 표기되어 있어 특징 정보로서의 활용이 용이하다. 또한, 'suncity'는 이러한 광고 중에서 가장 많이 발견되었으므로, 분석 대상 광고에 포함시켜 함께 분석하였다.

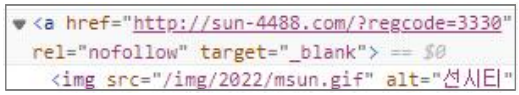

그림 4. suncity 가입코드
Fig. 4. Suncity Regcode

3.2.2 광고사 가입코드

‘1BET1’, ‘winner’, ‘WIN’은 전부 광고 배너 이미지에 코드 및 가입코드가 있기 때문에 각 사이트를 구별하고 특정함에 있어서 특징 정보로 사용하기 용이했다. 표 6은 61개의 분석 대상 사이트에서 식별된 광고사 가입코드를 정리한 표이다. 선별된 광고사는 표 5에 해당되는 네 가지 광고사이다.

표 6. 광고 배너 코드 분석
Table 6. Advertisement Banner Code Analysis

No.	1BET1	winner	WIN	suncity
1	TTC	2580	1111	5555
2	-			
3	-			
4	9466	9915	3882	8006
5	4545	9954	-	-
6	2055	7033	9055	-
7	4000	-		
…	…	…	…	…

4. 실험 및 결과

4.1 실험 방법

본 논문에서 제안하는 불법 저작권 침해 사이트 특정을 위한 특징 정보 수집 방안으로의 실험 과정은 그림 5와 같다. 링크 모음 사이트 등록되어 있는 불법 저작권 침해 웹툰 및 스트리밍 사이트를 크롤링한 후 각 사이트들의 특징 정보를 수집하여 데이터베이스화한다. 저장된 사이트에 도메인 변화를 주기적으로 추적하여 도메인이 변경됐을 시의 변경된 사이트의 특징 정보를 수집한다. 이를 데이터베이스에 저장된 특징 정보와의 비교 및 검증하여 변경 전·후 사이트의 동일성을 판단한다.

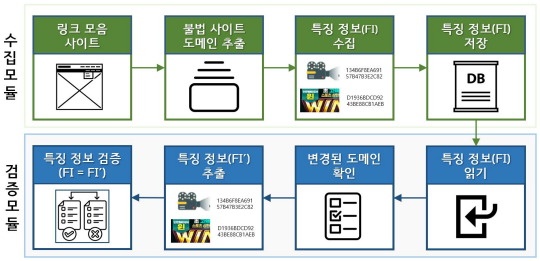

그림 5. 특징 정보 수집 및 검증 과정
Fig. 5. Process of Collecting and Verifying Feature Information

4.2 실험 및 결과

4.2.1 특징 정보 수집

불법 저작권 침해 사이트에서 파생된 1,044개의 URL을 대상으로 특징 정보 수집 및 식별 테스트를 수행했다. 도메인의 경우 연구를 진행하는 도중에도 지속적으로 변경되기 때문에 특징 정보 수집 테스트를 시작하는 2023년 7월 둘째 주를 기준으로 최신 도메인 조회 및 수집한 후 표 7과 같이 식별 테스트를 진행했다. 테스트는 표 내의 사이트들의 로고 이미지, 광고 배너, 도메인을 수집한 시점 이후로 도메인이 변경되었을 때의 특징 정보를 재수집하여 진행했다.

표 7. 특징 정보 비교
Table 7. Feature information comparison

도메인 주소	변경 도메인	로고	광고
https://b*n248.com	248 → 249	-	O
https://a*t312.com	312 → 313	O	O
https://f*e265.com	265 → 266	O	O
https://h*u319.net	319 → 320	O	O
https://w*f278.com	278 → 279	O	O
https://t*r272.com	272 → 273	O	O
https://y*n108.com	108 → 109	O	O
http://a*l89.com	89 → 90	O	O
https://s*n101.com	101 → 102	O	O
https://t*t.org	org → cc	O	O
·····	·····	·····	·····

1차적으로 수집된 도메인은 각각의 사이트로부터 로고 이미지 URL과 로고 이미지 파일을 수집한다. 로고 이미지 파일의 경우는 MD5 해시값을 추가 수집하여 로고 이미지의 무결성을 확보한다. 또한 해당 광고가 존재할 시 '1BET1', 'winner', 'WIN'의 광고 배너상의 가입코드와 'suncity'의 'regcode' 값을 수집한다. 도메인의 경우 숫자 부분에서의 변경만 이루어진 경우 변경된 숫자를 기록하고 로고 이미지의 해시값을 비교하여 동일할

경우 “O”, 주요 네 가지의 광고 배너의 가입코드를 비교하였을 때 모두 동일한 경우 “O” 표시로 작성한다.

4.2.2 특징 정보 식별 테스트

도메인 변경 전·후 사이트의 로고, 광고 배너, 변경된 도메인 패턴을 특징 정보로서 수집하여 식별 테스트를 진행하였다. 그 결과 수집된 정보와 변경 후 수집한 정보가 100% 동일하다는 결과를 도출할 수 있었다. 단, 광고의 경우 교체 주기가 불특정하여 다수의 광고 중 일부만 동일하여도 동일한 사이트로 간주하여 진행하였다.

로고의 경우 모든 사이트에 해당되는 고유한 특징 정보로 이미지 형태가 대부분이다. 로고가 텍스트 형태인 사이트를 제외하고 모든 사이트에서 이미지 파일로 조회 및 다운로드가 가능하다. 로고는 또한 도메인 주소가 변경되어도 동일한 위치를 유지하고 있어 변경 전·후의 사이트가 같은 운영자에 의해 서비스됨을 증명하는 데에 이용 가능하며 파일의 해시값을 비교해본 결과 도메인 변경 전·후의 해시값이 동일하게 나타났다. 특히 텍스트 형태의 로고의 경우 텍스트를 기반으로 해시값을 계산하여 적용하였다.

광고 배너의 경우 사이트별 광고를 분석한 결과, 광고에는 사이트를 식별할 수 있는 가입코드가 부여되어 있으며 이 가입코드는 도메인이 변경된 이후에도 동일한 값이 유지된다. 이를 활용하여 사이트를 특정할 수 있다. 다만, 광고 배너의 경우 배너의 위치 혹은 광고사가 계약 만료 등의 이유로 변경되는 경우가 있는데, 한 사이트의 경우 도메인이 변경되면서 대상 광고 중 하나가 광고 게시가 종료되어 총 3개 중에 2개의 가입코드만 일치하는 것을 확인하였다.

한편, 일부 사이트에서는 동일한 가입코드를 가지는 경우도 발견되었다. 대부분의 불법 사이트의 주 수익원이 가입코드라는 점에서 동일한 가입코드를 사용한다는 것은 동일 운영자이거나 혹은 동일한 광고 관리자인 것으로 추정할 수 있다. 이 경우 각 사이트의 레이아웃 및 이미지 파일 출처 등도 동일한 구성을 보였다.

도메인의 경우 보통 도메인 내 숫자가 하나씩 증가하는 방식으로 도메인 변경이 이루어지는 것이 일반적인 반면 몇몇 그렇지 않은 변경 패턴도 발견되었다. 도메인 내의 숫자가 둘 증가하는 경우, 도메인 변경 후에 변경 전 사이트도 중복으로 운영하는 경우, TLD를 변경하는 경우가 있다.

도메인 내의 숫자가 둘 증가한 경우는 그 사이 숫자의 도메인을 확인해 본 결과, 해당 사이트와 전혀 무관한 광고 사이트로 리다이렉션(redirection)되는 것을 확인했다. 이는 숫자가 하나씩 증가하는 불법 사이트의 패턴을 파악한 제3자의 전략적 광고 행위로 보인다. 실제 이런 경우에 “t*a65.com”을 접속 시 “http://www1.t*a65.com”과 같이 https가 아닌 http를 사용하는 것과 www 다음에 www1과 같이 숫자가 위치하는 것이 특징이다. 이러한 제3자 광고 사이트의 경우 내용은 사이트마다 상이하지만 사이트의 레이아웃 및 생김새는 대체로 그림 6과 같다.

그림 6. 불법 사이트와 무관한 제3자 광고 사이트
Fig. 6. A third-party advertisement site

4.2.3 실험 결과

본 논문에서는 61개 불법 저작권 침해 사이트를 대상으로 약 150일간 1,044개의 URL을 기반으로 수집할 수 있는 정보를 정의하고 각 정보별 식

별 가능성을 분석하였다.

분석 결과, 로고 이미지 파일과 광고 가입코드는 각 사이트 도메인과 결합하였을 때 식별 가능성이 높았다. 특히, 로고 이미지의 경우 모든 사이트에서 고유하고, 변경된 후에도 로고 이미지를 유지하고 있어 사이트 특정에 적합하였다.

또한 광고의 경우 각 사이트별 광고사로부터 부여받은 가입코드가 존재하고 있어 사이트 특정이 용이하였다. 단, 운영자가 동일하거나 광고 관리자가 동일한 다수의 사이트에서 동일한 가입코드가 발견되어 가입코드만으로 사이트를 식별하는 것보다 도메인 정보, 로고 이미지 해시값과 함께 적용함으로써 식별력을 높일 수 있었다.

5. 결 론

본 논문은 해외 불법 저작권 사이트에 대해서 수사 단속이 어려운 Cloudflare와 같은 클라우드 서비스를 통해 도메인 차단 방식을 우회하는 불법 저작권 침해 사이트를 효과적으로 근절하기 위해 기존의 방법보다 빠르게 불법 사이트를 판단할 수 있는 사이트 특징 정보를 제안하였다. 61개의 불법 저작권 침해 사이트를 대상으로 약 150일간 1,044개의 URL을 기반으로 수집할 수 있는 정보를 정의했다. 각 정보별 식별 가능성을 분석하여 로고 이미지 파일과 광고 가입코드가 각 사이트 도메인과 결합하였을 때 식별 가능성이 높은 정보로 정의하고 이를 실험하였다.

실험 결과, 도출된 특징 정보인 사이트 로고, 광고 배너와 도메인, 가입코드를 기반으로 불법 저작권 침해 사이트를 특정할 수 있다는 것을 확인하였다. 특히 로고 이미지의 경우 변경 전·후 사이트 분석에 주요한 특징 정보로 활용할 수 있었다. 특히 광고 가입코드의 경우 운영자가 동일하거나 광고 관리자가 동일한 경우에 있어서는 가입코드만으로 사이트를 식별하는 것보다 도메인 정보, 로고 이미지 해시값과 함께 적용함으로써 식별력을 높일 수 있었다.

향후 본 연구를 기반으로 기존 차단 시스템에 필요한 정보를 제공하여 보완한다면 불법사이트의 도메인 변경 등이 이루어졌어도 사이트를 빠르게 특정할 수 있고 불필요한 검증 및 심사 기간을 줄여 즉시 조치할 수 있는 체계를 갖출 수 있을 것으로 기대한다.

본 연구는 문화체육관광부 및 한국콘텐츠진흥원의 2023년도 문화기술 연구개발 사업으로 수행되었음 (과제명 : 클라우드 기반 원격지 저장장치 증거 수집 및 저작권 침해 데이터 DB 개발, 과제번호 : R2022020109, 기여율: 100%)

참 고 문 헌

[1] Korea Creative Content Agency. 2022 Fact-finding Survey on Illegal Distribution of Comics - Webtoons: Korea Creative Content Agency (2022). ISBN : 979-11-6677-099-9

[2] K. J. Park. A Study on Effects of Relative Benefits and Costs of Piracy of Digital Contents on Attitudes and Behaviors of Illegal Duplication. Journal of the Korea contents association, 15(7), 489-499. (2015). DOI : 10.5392/JKCA.2015.15.07.489

[3] M. S. Choi. Efficient video matching method for illegal video detection. Journal of Digital Convergence, 20(1), 179-184. (2015). DOI : 10.14400/JKCA.2015.15.07.489

[4] Korea Creative Content Agency. 2021 Korea Copyright White Paper. : Korea Creative Content Agency (2022). ISSN : 2234-392X

[5] Korea Copyright Protection Agency. (2021).

2021 Copyright Protection Annual Report. https://www.mcst.go.kr/kor/s_policy/dept/deptView.jsp?pSeq=1526&pDataCD=0417000000&pType

[6] Y. S. Hwang, J. H. Hwang, & S. J. Lee. IP address tracking techniques for illegal sites using Cyber Threat Intelligence search services. Journal of Digital Forensics, 16(2), 116-125.(2022). DOI : 10.22798/KDFS.2022.16.2.116

[7] I. J. Yoo, J. C, Lee, B. C. Park, S. Y. Kim & Y. M. Kim. A Method for Generating Signature Information to Determine Illegal Distribution of Cloud-based Webtoon Contents. (2022). Journal of Software Assessment and Valuation, 18(2), 77-85. DOI : 10.29056/JSAV.2022.12.08

[8] J. E. Jo & J. E. Choi. Operational Status and Improvement Tasks of Digital Sexual Crimes Response Policy. : National Assembly Research Service. (2018). ISSN : 2586-5668

[9] J. W. Jeong & S. J. Lee. Blocking method of harmful sites based on domain change pattern. Journal of Digital Forensics, 15(3), 39-53. (2021). DOI : 10.22798/KDFS.2021.15.3.39

[10] H. Y. Kang, Y. C. Choi, & S. J. Lee. Analysis of advertisers by tracking banner ads on piracy websites. Journal of Digital Forensics, 15(3), 15-26. (2021). DOI : 10.22798/KDFS.2021.15.3.15

[11] S. H. Park, S. M. You, D. H. Song, & K. J. Lee. A Study on Harmful Site Discrimination Methods using Site Banner Advertisements. (2023). Korean Institute of Communication and Information Sciences/Proceedings of Symposium of the Korean Institute of communications and Information Sciences. https://www.dbpia.co.kr/journal/articleDetail?nodeId=NODE11227654

저 자 소 개

최은석(Eun-Seok Choi)

2023.2 한림대학교 정보법과학전공 졸업
2023.2-현재 : 에이치엠컴퍼니㈜ 연구원
<주관심분야> 저작권 보호 및 이용 활성화

김영모(Young-Mo Kim)

2003.2 대전대학교 컴퓨터공학과 졸업
2005.2 대전대학교 컴퓨터공학과 석사
2011.2 대전대학교 컴퓨터공학과 박사
2012-현재 : 숭실대학교 교수
<주관심분야> 저작권 보호 및 이용 활성화

박명찬(Myung-Chan Park)

2005.8 대전대학교 컴퓨터공학과 박사
2012.11-현재 : 에이치엠컴퍼니㈜ 이사
<주관심분야> 저작권 보호 및 이용 활성화