ISSN(Print) : 1976-5304, ISSN(Online) : 2713-931X
https://kdfs.jams.or.kr, DOI : 10.22798/kdfs.2022.16.1.76

불법 웹툰 사이트 정보를 통한 사이트 연관성 분석

최 영 철*, 이 상 진**
고려대학교 정보보안학과 (대학원생)*
고려대학교 정보보호대학원 (교수)**

Analysis of site relevance through illegal webtoon site information

YoungChul Choi*, SangJin Lee**
Dept. of Information Security, Korea University (Graduate Student)*
School of Cybersecurity, Korea University (Professor)**

요 약

최근 웹툰에 대한 수요가 많아져, 인기 웹툰을 불법 복제하여 무단 배포하는 불법저작물 사이트가 지속적으로 늘어나고 있다. 이러한 사이트들은 신고가 되어도, 서버가 외국에 있어 처벌이 어렵고, 사이트 주소만 바꾸어 복제된 웹툰을 재배포하기 때문에 접속 차단 조치만으로는 빠르게 근절하는데 어려움이 있다. 본 연구에서는 불법저작물 사이트에서 얻을 수 있는 데이터와 복제된 웹툰에서 추출한 메타데이터를 이용하여, 불법저작물 사이트간 연관성을 분석했다. 또한 기존에 제시되지 않았던 불법 저작물 사이트 연관성 분석에 필요한 요소를 식별하고 이를 활용하여 효과적인 사이트 차단방법을 제안한다. 연관성 있는 사이트를 구분하는 방법은 해외 공조 등 오랜 시간이 소요되는 수사에서 수사를 착수하기 위한 우선순위를 정하는데 도움을 줄 것이며, 사이트 연관성 분석에서 찾아낸 특징인 불법 저작물 사이트에 게시된 웹툰이 저장되어 있는 서버를 공유하는 특징을 이용해 이미지 저장서버를 차단했을 때의 효과를 보여줄 것이다. 이에 본 연구는 불법 사이트에서 얻을 수 있는 데이터를 통해 사이트 간 연관성을 분석하는 방법과 효과적인 차단 방법에 대해서 소개하고자 한다.

주제어 : 불법 저작물, 웹툰, 메타데이터, 접속차단

ABSTRACT

Recently, the demand for webtoons has increased, and illegal copyright sites that illegally reproduce popular webtoons and distribute them without permission are continuously increasing. Even if these sites are reported, punishment is difficult because the server is abroad, and because only the site address is changed and replicated webtoons are redistributed, it is difficult to quickly eradicate them with access blocking measures alone. In this study, the association between illegal sites was analyzed using data obtained from illegal sites and metadata extracted from copied webtoons. In addition, elements necessary for analyzing the relevance of illegal sites that have not been previously presented are identified and utilized to propose an effective site blocking method. The method of distinguishing related sites will help prioritize investigations in long-term investigations such as international cooperation, and will show the effect of blocking image storage servers by sharing servers posted on illegal sites. Therefore, this study aims to introduce a method of analyzing the association between sites through data obtained from illegal sites and an effective blocking method.

Key Words : Piracy, Webtoon, Metadata, Site Blocking

※ 이 논문은 2022년도 정부(문화체육관광부)의 재원으로 한국저작권보호원의 지원을 받아 수행된 연구임(No 2022. 저작권 특화 디지털포렌식 전문인력 양성사업)
- Received 13 February 2022, Revised 14 February 2022, Accepted 31 March 2022
- 제1저자(First Author) : Youngchul Choi (Email : cyc0703@korea.ac.kr)
- 교신저자(Corresponding Author) : Sangjin Lee (Email : sangjin@korea.ac.kr)

Ⅰ. 서 론

최근 코로나 19의 영향으로 사람들의 외부활동이 제한되면서, 저비용으로 즐길 수 있는 웹소설, 웹툰 등 디지털 콘텐츠에 대한 소비가 늘어나고 있다. 한국저작권보호원의 웹소설 등 저작권 침해 실태조사 및 대응방안 연구[1]에 따르면, 2019년 웹툰 산업의 규모는 플랫폼과 에이전시의 매출을 더해 약 6,400억 원으로 추정되며 네이버, 카카오 등 대형 플랫폼의 글로벌 진출 및 IP사업 확장에 따라 웹툰 산업의 규모는 더 커질 것으로 예상된다. 하지만 규모가 커지는 것과 비례해 저작권 침해로 인한 피해 또한 증가하고 있으며, 피해액은 2019년 기준 합법적인 웹툰 산업 규모의 49.7%인 약 3,189억 원으로 추정된다.

저작권 침해를 막기 위한 방법으로 국내 서버의 경우 저작권법에 의해 '삭제·전송 중단 조치'를 할 수 있지만 해외 서버의 경우 현행 저작권법으로는 대응할 수 없어 정보통신망법에 근거해 방송통신심의위원회(이하 방심위)에서 불법 사이트를 대상으로 망 사업자에게 해당 사이트에 대한 내국인 접속 차단을 요구하고 있다. 2021년 국정감사에서 문화체육관광부로부터 받은 연도별 만화 웹툰 관련 신고 현황 및 불법 웹툰 차단 조치 현황[2]에 따르면 2020년 신고건수는 3,844건이며, 사이트 접속차단 건수는 399건으로 신고대비 최대 10% 정도가 차단되고 있다. 저조한 차단비율을 보이고 있는 이유는 불법 사이트가 생산되는 속도와 도메인이 바뀌는 주기(1~15일)[3]에 비해 부족한 수사 인력과 방심위의 심사기간(4~6일)이 길기 때문이다.

저작물의 불법 유통을 근절하기 위해서는 유포자를 찾는 것이 제일 좋은 방법이다. 하지만 유포자를 찾는 것은 해외 공조 등 많은 시간이 필요하다. 따라서 방심위에서는 저작권자의 피해를 최소화하기 위해 웹사이트 차단을 수행하고 있다. 불법적인 웹사이트의 차단이 관련 사이트의 트래픽을 약 73% 감소시켰고, 합법사이트의 이용률을 6% 증가시켰다는 연구[4]와 불법복제물 이용경로를 접속 차단했을 때 불법 사이트 이용자 55%가 불법 사이트 이용을 포기했다는 보고서[5] 등을 봤을 때, 불법 사이트 차단도 저작물의 불법 유통으로 인한 피해를 줄일 수 있는 충분한 수단이 될 수 있다.

따라서 본 연구에서는 불법 사이트에서 획득할 수 있는 정보를 이용해 여러 불법사이트를 동시에 차단하는 효과를 낼 수 있는 방법과 불법 사이트 수사에 도움을 줄 수 있는 사이트 간 연관성을 분석하는 방법에 대해 기술하고자 한다. 이 논문에서 제안하는 차단 전략을 통해 빠르게 생성되는 불법 사이트에 대한 효과적인 대응과 사이트 간 연관성 분석을 통해 수사의 우선순위를 선정하는데 기여할 수 있을 것이다.

본 논문의 구성은 다음과 같다. 2절에서는 불법 사이트에서 수집될 수 있는 정보에 관한 연구를 소개하고, 3절은 웹툰의 불법 복제 유형에 대해 설명한다. 4절은 고려대학교에서 2022년 1월 동안 링크모음사이트를 크롤링해 추출한 2,976개의 불법 사이트 중 웹툰을 불법으로 유통하는 불법 사이트의 정보와 게시된 이미지의 메타데이터를 활용한 사이트 간 연관성 분석방법과 수집한 사이트의 정보를 이용한 효율적인 불법 사이트 차단방법에 대해 설명하고, 5절에서는 결론을 맺는다.

Ⅱ. 관련 연구

유해 사이트의 정보를 수집하는 연구나 유해 사이트를 탐지하는 연구는 많이 존재하지만, 불법 사이트 간 연관성을 분석하는 연구는 없다. 따라서 사이트 간 연관성을 분석하기 위해 유해 사이트에서 획득할 수 있는 요소가 나타나는 연구를 몇 가지 소개한다.

Kang 등[6]은 실시간 크롤링을 통한 유해 사이트 판별 시스템을 제안하였다. URL(Uniform Resource Locator)을 입력 받아 유해 사이트 DB에 포함되지 않으면, 크롤링을 진행하여 해당 사이트에서 나타나는 단어 집합들을 추출했으며, 학습된 모델에 의해 텍스트기반의 판별을 진행했다. 그러나 형태소 분석을 이용해 단어를 추출하지 않았으며, 유해 사이트의 정보 중 키워드만 사용하여 유해 사이트를 판별하여 정확도가 떨어졌다. 따라서 본 논문에서는 텍스트 외 추가적인 정보도 활용할 것이다.

Choo 등[7]은 웹 크롤링을 통한 유해 사이트 정보 수집에 대한 연구를 수행하였으며, 링크모음 사이트를 크롤링하여, 유해 사이트를 수집하고 수집된 사이트에 대해 URL, 사이트 유형, 운영자 정보, Google Analytics ID, 웹사이트 엔진 정보, IP, 운영상태, SNS(Social Networking Service) 정보, 텔레그램 URL, 트위터 URL, VPN(Virtual Private Network) 사용여부, ASN(Autonomous System Number) 정보 등을 획득했다. 이 논문에서 제안한 웹 크롤러의 유해 사이트 판별율은 95%를 보이고 있지만 링크모음사이트를 기반으로 크롤링하기 때문에 Seed URL에 대한 업데이트를 지속적으로 수행해야 한다는 한계가 있다. 또한 유해 사이트에서 수집될 수 있는 정보의 종류를 보여주었다. 따라서 본 논문에서는 유해 사이트 간 연관성을 보이기 위해 수집된 정보를 통한 연관성 분석에 이를 활용할 것이다.

Kang 등[8]은 유해 사이트에 게시되어 있는 배너광고 추적을 통해서 광고주를 찾아 수익원을 찾아내는 연구를 하였다. 유해 사이트의 배너광고를 계속 추적하여, 광고를 통해 수익을 얻지 않는 최종 광고주를 찾아낼 수 있다는 가정으로 최초 추적을 시작하는 1,172개의 유해 사이트로부터 배너광고를 추적해 최종 광고주 사이트 173개를 찾아내었고, 최종 광고주는 도박, 성인용품, 성인물 등을 통해 수익을 내는 것으로 파악하였다. 최종 광고주의 Google analytics ID를 통해 광고 운영자간 연관성을 분석하기도 하였다. Google analytics ID는 사이트 운영 측면에서 사용자 이용추이를 파악하는데 사용되는 기능이기 때문에, 본 논문의 사이트 간 연관성 분석에 Google analytics ID는 유효한 정보로 사용할 수 있을 것이다.

이 외에도 HTML 태그 순서를 통해 유사도를 측정해 불법 사이트를 탐지하는 기술 연구[9], 도메인 변경 패턴에 기반한 차단 방식을 제안하는 연구[3], 키워드를 통해 웹사이트 간 연결 관계를 이용한 유해 사이트 판별 방법을 제안하는 연구[10] 등 불법 사이트 판별 및 차단에 대한 연구가 활발히 이루어지고 있다.

Ⅲ. 웹툰 불법 복제 유형

3.1 웹툰 불법 복제 사이트

불법 사이트에서는 유료결제를 통해 볼 수 있는 회차에 대한 무료 열람 서비스를 제공하고 있으며, 유료분이 나오는 날을 기점으로 업데이트가 이루어진다. 또한 불법 복제 사이트는 사이트 차단을 회피하기 위해 1일에서 15일 사이의 주기로 도메인 주소를 변경하고 있다. 불법 사이트는 국내법을 회피하기 위해 해외 서버를 기반으로 서비스되고 있어 사이트에 대한 차단만 이루어지고 소송 및 벌금 부과 사례가 없다.

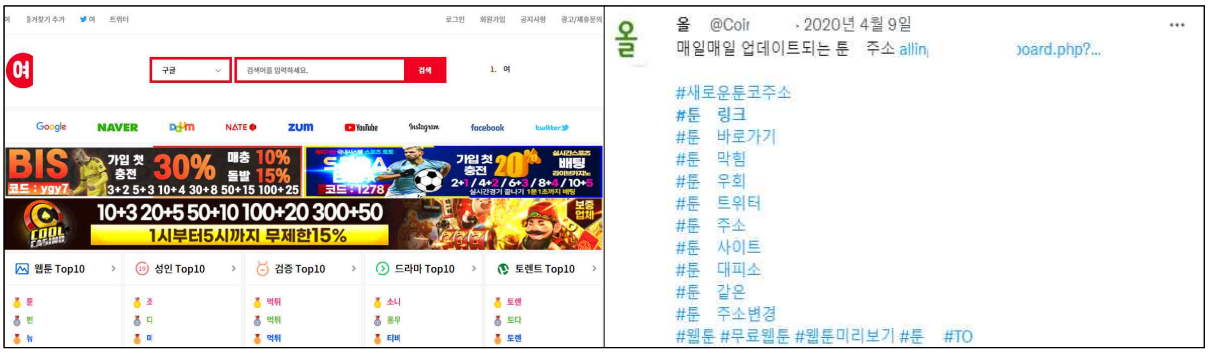
3.2 웹툰 불법 복제 전용 앱

iOS와 Andorid 앱 모두 불법복제 전용 앱을 찾아볼 수 있으며, Android는 apk 파일 형태로 유통된다. 구글의 플레이스토어, 애플의 앱스토어 등 앱을 배포하는 공식 플랫폼에서는 불법으로 웹툰을 배포하는 사이트 이름, 웹툰 명으로 검색했을 때, 불법 웹툰을 배포하는 앱을 찾아볼 수 없다. 하지만 같은 키워드로 검색했을 때 [그림 1]과 같이 apkc****.com, apkp***.com 등 공식 플랫폼이 아닌 사설 플랫폼에서는 불법 웹툰을 배포하는 apk를 확인할 수 있으며, 복제된 웹툰을 배포하는 사이트에서 직접 apk를 배포하기도 한다.

〈Figure 1〉 APK distribution site

3.3 링크 사이트, SNS를 통한 불법 복제 공유

세 번째 유형으로 [그림 2]와 같이 불법 저작물 사이트로 연결하는 링크모음 사이트와 SNS를 통해 사이트가 공유되고 있으며, 과거에는 불법 콘텐츠를 공유하는 것이 불법이 아니었지만 2017년부터 "불법 복제물 링크 행위는 불법 복제물에 대해 실질적으로 접근 가능성을 증대시켜 이용에 제공하는 행위를 용이하게 하므로 저작권자의 전송권 침해 행위를 방조하는 행위"라는 판례[13]가 나오면서, 최근의 판결들은 불법 복제물 링크 행위가 저작권 침해의 방조에 해당한다고 판단하고 있다. 따라서 링크모음사이트와 SNS를 통한 불법 사이트 공유도 불법 저작물 복제 유형 중 하나로 볼 수 있다.



〈Figure 2〉 Link site, SNS distribution

Ⅳ. 불법 웹툰 사이트 연관성 분석 및 차단방법

저작권 침해 행위는 최종적으로 웹 사이트를 통해서 이루어지기 때문에 불법 복제 유형 중 불법 복제 사이트를 중점으로 동일 유형(웹툰)의 사이트에 대한 연관성을 분석하였다. 또한 수집된 사이트 정보를 통한 효과적인 차단방법을 제안한다.

4.1 데이터베이스 내용 분석

링크모음사이트를 중심으로 유해 사이트를 수집하여 저장한 데이터베이스[7]의 칼럼 중 웹툰과 관련된 URL을 수집하기 위해 [표 1]과 같이 3개의 컬럼을 중점으로 확인했으며, 이를 통해 확인할 수 있는 카테고리 별 사이트 수는 [표 2]와 같다.

〈Table 1〉 Database Table Column

Column	Description
main_url	수집된 유해 사이트의 URL
expected_category	예상되는 유해 사이트의 카테고리 “adult”, “gamble”, “link”, “prostitution“, ”sportslive“, ”streaming“, ”torrent“, ”webtoon”
created_at	데이터 수집 일자

〈Table 2〉 The number of URLs stored by expected_category

Expected_category	Count
adult	945
gamble	586
link	1
prostitution	257
sportslive	56
streaming	367
torrent	316
webtoon	448

웹툰으로 구분된 사이트 448개 중 합법 웹툰 사이트, 중복된 사이트, 접근 불가능한 사이트, 웹툰이 아닌 만화를 배포하는 사이트를 제외하고 42개의 웹툰 불법 배포 사이트를 확인했다. 또한 수집된 42개의 웹툰 불법 배포 사이트에서 [그림 3]과 같이 Google Analytics ID(GA), 이미지 저장 서버(Image Server URL), 이미지 저장 하위 경로(Sub URL), 이미지 해시, 이미지 DQT 해시에 대한 정보를 추가로 수집하였다. 이미지 저장서버에 대한 정보는 불법 웹툰 사이트에서 배포되고 있는 웹툰 이미지의 URL에서 획득할 수 있으며, Google Analytics ID는 대부분 사이트 메인 페이지 소스에서 G-XXXXXXXXXX, UA-XXXXXXXXX-X 형태로 획득할 수 있다. 본 논문에서는 크롤러 탐지기술이 적용되지 않은 일부 사이트에서는 각 사이트에 맞춰 제작한 크롤러를 통해서 추가 정보를 획득하고, 그 외 사이트에서는 직접 접근하여 추가 정보를 획득하였다.

site_url	image hash	GA ID	Image Server URL	Sub URL	dqt_hash
https://ww	58.xyz	2b26f5302ef3183e1408055aedf13a83	-	https://w e58.xyz/view	view/	36670c99dd7fd3310227d88cdac8815c
https://nan	n79.com	4531f840ef9bd5d5c4e94449502fb0b4	G-EY0ZT2N93L	https://in nedtoon78.com/toon2/a1/img	toon2/a1/img	00b3186be50bcb5eb28908150b7f5cb4
https://too	g28.com	4531f840ef9bd5d5c4e94449502fb0b4	-	https://c e.africa.com/toon2/a1/img	toon2/a1/img	00b3186be50bcb5eb28908150b7f5cb4
https://cop	202.com	4531f840ef9bd5d5c4e94449502fb0b4	-	https://c e.africa.com/toon2/a1/img	toon2/a1/img	00b3186be50bcb5eb28908150b7f5cb4
https://blac	140.com	6d639f149b0149a7df9e432f239f8360	UA-175188271-1	https://b naimg.com/img/a-upload	img/a-upload/	26baf4e032c43711249f8375d40bdbce
https://agit	m	6d639f149b0149a7df9e432f239f8360	UA-186192900-1	https://b naimg.com/img/a-upload	img/a-upload/	26baf4e032c43711249f8375d40bdbce
https://ww	t15.net	79d7ebf8dec1fd8ca9b74dc3c604ac67	-	https://w 05.net/data/files/editor	data/files/editor/	fae6c8c4297184f408eda79f8fb31d46
https://k1.k	n.com	79d7ebf8dec1fd8ca9b74dc3c604ac67	G-LCHP581VL1	https://w 05.net/data/files/editor	data/files/editor/	fae6c8c4297184f408eda79f8fb31d46
http://wtbo	m	79d7ebf8dec1fd8ca9b74dc3c604ac67	G-F953PGKT7D	https://w 05.net/data/files/editor	data/files/editor/	fae6c8c4297184f408eda79f8fb31d46
https://fli		79d7ebf8dec1fd8ca9b74dc3c604ac67	UA-154925916-4	https://w 05.net/data/files/editor	data/files/editor/	fae6c8c4297184f408eda79f8fb31d46
https://ww	oon17.com	79d7ebf8dec1fd8ca9b74dc3c604ac67	-	https://w nimg.com/data/files/editor	data/files/editor/	fae6c8c4297184f408eda79f8fb31d46
https://sto	om	79d7ebf8dec1fd8ca9b74dc3c604ac67	-	https://w nimg.com/data/files/editor	data/files/editor/	fae6c8c4297184f408eda79f8fb31d46
https://tkor		4c4407ca13044cfb3a510586ff168911	-	https://tk /data/file/wtoon	data/file/wtoon/	fae6c8c4297184f408eda79f8fb31d46
https://yay	3.com	4c4407ca13044cfb3a510586ff168911	-	https://tk /data/file/wtoon	data/file/wtoon/	fae6c8c4297184f408eda79f8fb31d46
https://bed	.com/	4c4407ca13044cfb3a510586ff168911	-	https://tk /data/file/wtoon	data/file/wtoon/	fae6c8c4297184f408eda79f8fb31d46
https://doll	m/	4c4407ca13044cfb3a510586ff168911	-	https://tk /data/file/wtoon	data/file/wtoon/	fae6c8c4297184f408eda79f8fb31d46
https://pillo	com/	4c4407ca13044cfb3a510586ff168911	-	https://tk /data/file/wtoon	data/file/wtoon/	fae6c8c4297184f408eda79f8fb31d46
https://mar	n139.com	4c4407ca13044cfb3a510586ff168911	-	https://tk /data/file/wtoon	data/file/wtoon/	fae6c8c4297184f408eda79f8fb31d46
https://cuc	129.com	4c4407ca13044cfb3a510586ff168911	-	https://tk /data/file/wtoon	data/file/wtoon/	fae6c8c4297184f408eda79f8fb31d46
https://jam	rk	4c4407ca13044cfb3a510586ff168911	-	https://ja ork/data/file/wtoon	data/file/wtoon/	fae6c8c4297184f408eda79f8fb31d46

〈Figure 3〉 Result of collecting additional information on piracy site informations

4.2 추가 수집 정보 분석

추가로 수집된 정보 중 이미지 저장 서버가 동일한 사이트를 확인하면 〔표 3〕과 같다. 불법 복제된 웹툰을 제공하는 사이트는 다르지만 이미지를 저장하고 있는 서버는 동일한 사이트가 다수 존재했으며, 42개 중 26개 사이트가 타 사이트와 중복되는 이미지 저장 서버가 있는 것으로 파악됐다.

〈Table 3〉 Piracy sites by image storage server

이미지 저장 서버 URL	불법 사이트 URL
https://tk*****ks/data/file/wtoon/	https://tk*****ks
	https://yay*****.com
	https://bed*****.com/
	https://do*****.com/
	https://pill*****.com/
	https://man*****.com
	https://cuc*****.com
https://www.yg*****et/data/files/editor/	http://wtb*****.com
	https://k1.ki*****.com
	https://www.*****15.net
	https://fli*****.net
https://zzo*****.com/data/file/download/img	https://frto*****.com
	https://new*****.org
	https://zzol*****.com
https://new*****.org/data/file	https://hoho*****.com/
	https://new*****.com
https://img.tktk*****.com/data/file/webtoonBK2/	https://hdh*****.net/
	https://tk*****.com/
https://i8.*****d.com/	https://fxf*****.com
	https://w*****.com
https://cloudflare.*****.com/toon2/a1/img	https://copy*****.com/
	https://too*****8.com
https://black*****.com/img/a-upload/	https://ag*****.com
	https://bla*****0.com/
https://www.to*****.com/data/files/editor	https://www.s*****7.com
	https://sto*****.com

〈Table 4〉 Piracy Sites by Google Analytics ID

Google Analytics ID	불법 사이트 URL
UA-192*****9-X	https://hob*****.com
	https://buzzt*****6.com
UA-175*****6-X	https://cuc*****1.com
	https://neo*****1.com
	https://real*****80.link
	https://mk*****.link
	https://gobo*****.com

Google Analytics ID는 운영자의 편의를 위해 사용되는 기능으로 동일한 운영자가 웹사이트를 운영하는지 확인하는 지표로 활용될 수 있다. Google Analytics ID를 기준으로 동일한 ID를 사용하는 사이트는 총 2개의 그룹으로 나눌 수 있었으며 〔표 4〕와 같다.

〔표 5〕는 웹 페이지를 생산할 때 동일한 업체 혹은 인물이 만든다면 서버 폴더 구조가 비슷할 것이라는 가정으로 도메인을 제외하고 하위 경로가 같은 불법 사이트를 나타낸 결과이다. 동일한 하위 경로를 가지는 불법 사이트들이 다수 식별되었으며, 이미지 저장서버가 다르고 불법 사이트도 다르지만 하위경로만 같은 사이트가 다수 존재한다는 것은 사이트 생산자가 동일할 가능성을 내포한다는 의미로 해석할 수 있을 것이다.

〈Table 5〉 Piracy sites by sub-path of illegal images

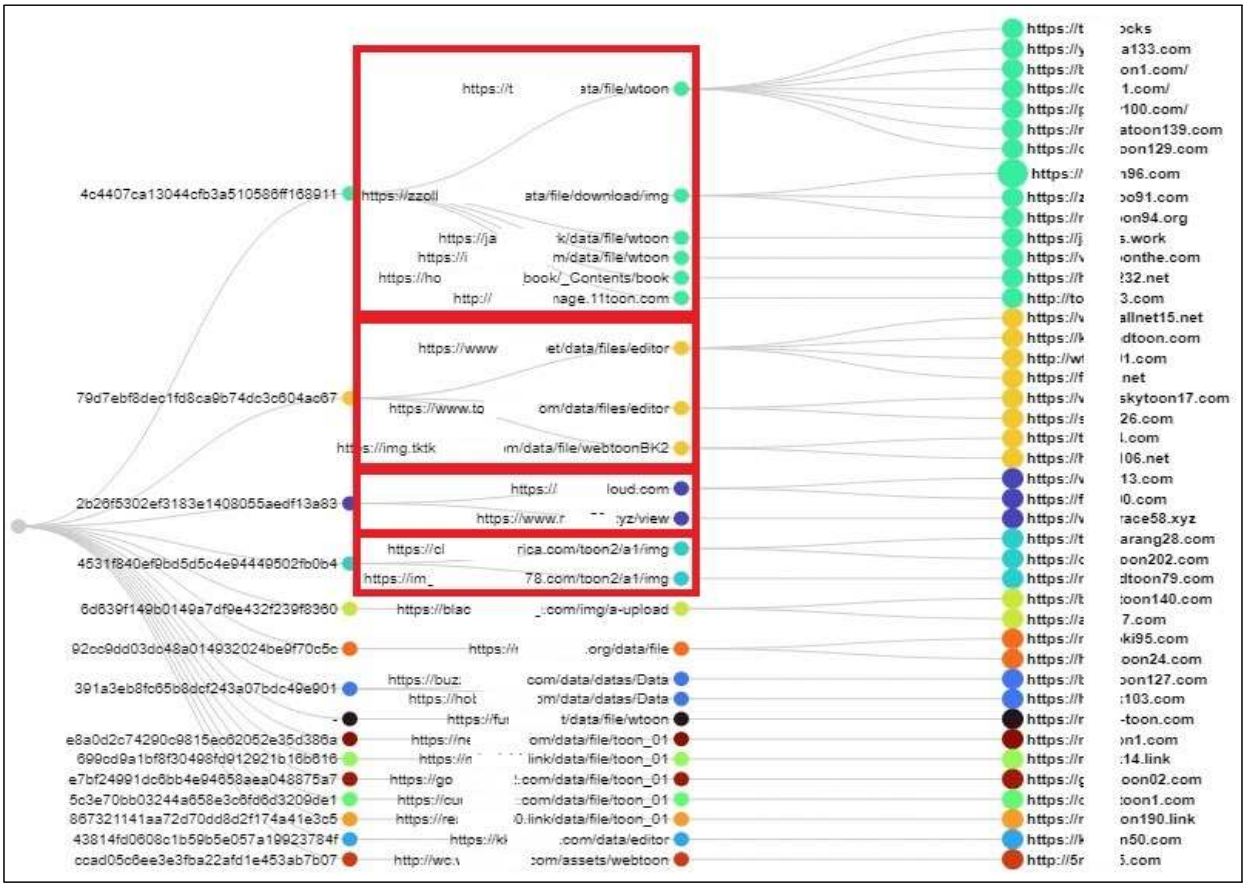
하위 경로	이미지 저장 서버 URL	불법 사이트 URL
/data/file/wtoon/	https://tk*****ks/data/file/wtoon/	https://cuckto*****.com
		https://mangato*****.com
		https://show*****.com
		https://bed*****.com/
		https://do*****.com/
		https://tko*****ks
		https://pill*****.com/
	https://jam*****ork/data/file/wtoon/	https://jam*****ork/
	https://fun*****st/data/file/wtoon/	https://majo*****n.com
	https://i1.to*****.com/data/file/wtoon	https://v34.*****he.com
/data/files/editor/	https://www.y*****et/data/files/editor/	https://k1.k*****on.com/
		http://w*****1.com
		https://www.al*****5.net
		https://f*****.net
	https://www.too*****com/data/files/editor/	https://www.skyt*****.com
		https://st*****.com
/data/file/toon_01/	https://neo*****com/data/file/toon_01/	https://ne*****.com
	https://m*****.link/data/file/toon_01/	https://m*****.link
	https://gob*****02.com/data/file/toon_01/	https://gobot*****.com
	https://cuc*****n1.com/data/file/toon_01/	https://cucci*****.com
	https://rea*****180.link/data/file/toon_01	https://realto*****.link
toon2/a1/img/	https://img3.nam*****n78.com/toon2/a1/img/	https://namedt*****.com/
	https://cloudflare.a*****.com/toon2/a1/img	https://copyto*****.com/
		https://toonsar*****.com
data/datas/Data/	https://buzzto*****.com/data/datas/Data/	https://buzzto*****.com/
	https://hob*****.com/data/datas/Data	https://hob*****.com

4.3 불법 사이트 연관성 분석

이 절에서는 앞서 불법 웹툰 사이트들의 연관성을 분석하는 방법을 제안한다.

첫 번째 방법은 이미지의 해시값을 기준으로 그룹핑하는 것이다. [그림 4]는 동일한 운영자에 의해 운영되는 불법 사이트는 동일한 불법복제 이미지를 사용할 것이라는 가정으로 같은 회차의 웹툰 이미지를 수집하여, 이미지의 해시값이 같은 사이트를 그룹화한 결과이다. 동일한 이미지 저장서버에서 이미지를 가져오는 경우 같은 해시가 나오기 때문에 다른 이미지 저장서버를 사용하지만 동일한 해시 값이 나오는 사이트를 그룹화했다. 해시값을 기준으로 그룹화 했을 때 총 4개의 그룹이 나왔으며, 이미지 저장 서버가 다르지만 동일한 이미지를 사용하여, 불법 복제물을 배포하는 곳이 다수 식별되었다.

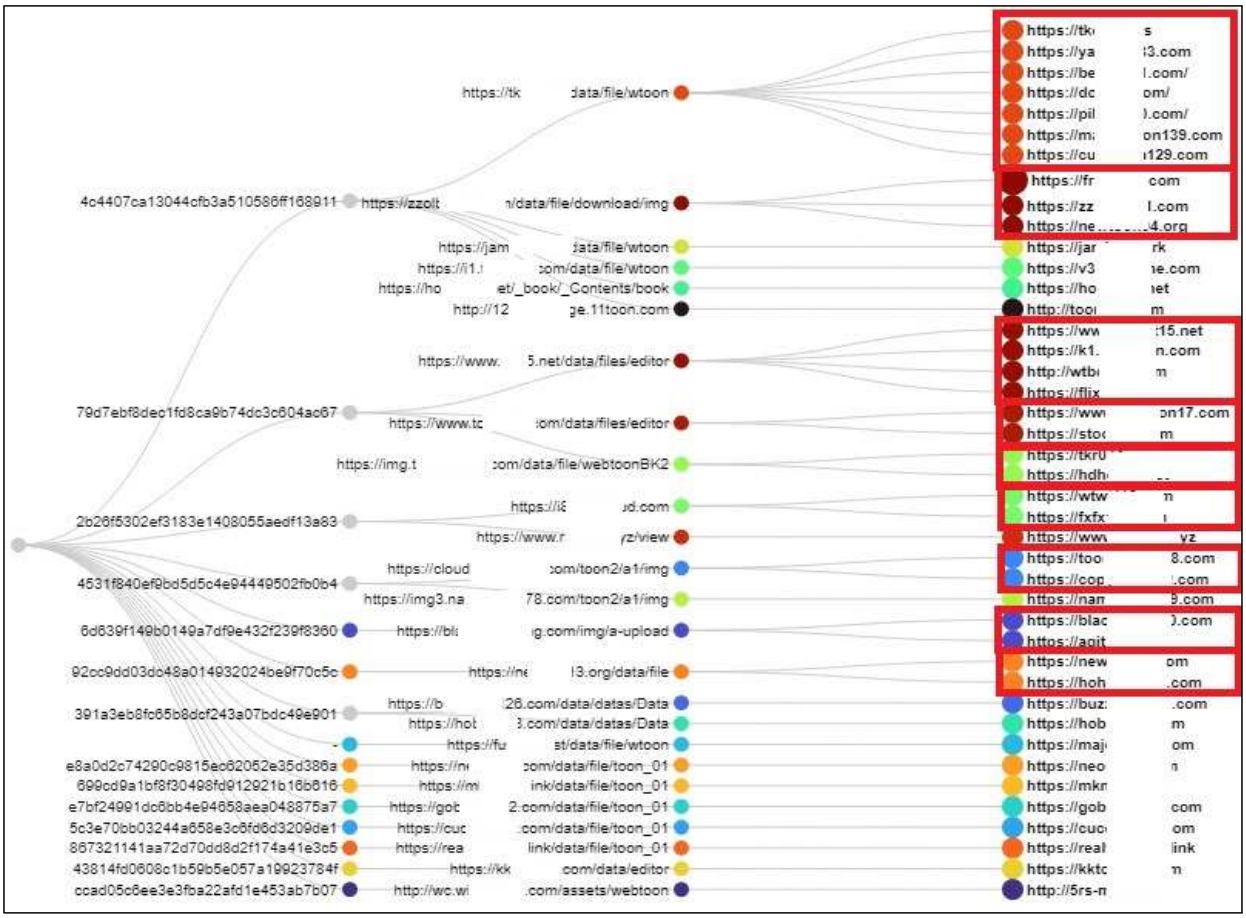
두 번째 방법으로 이미지 저장서버가 동일한 경우를 같은 그룹으로 보고 분석했을 때 결과는 [그림 5]와 같으며 2개 이상의 사이트로 묶이는 그룹은 총 9개의 그룹이 생성되었다.



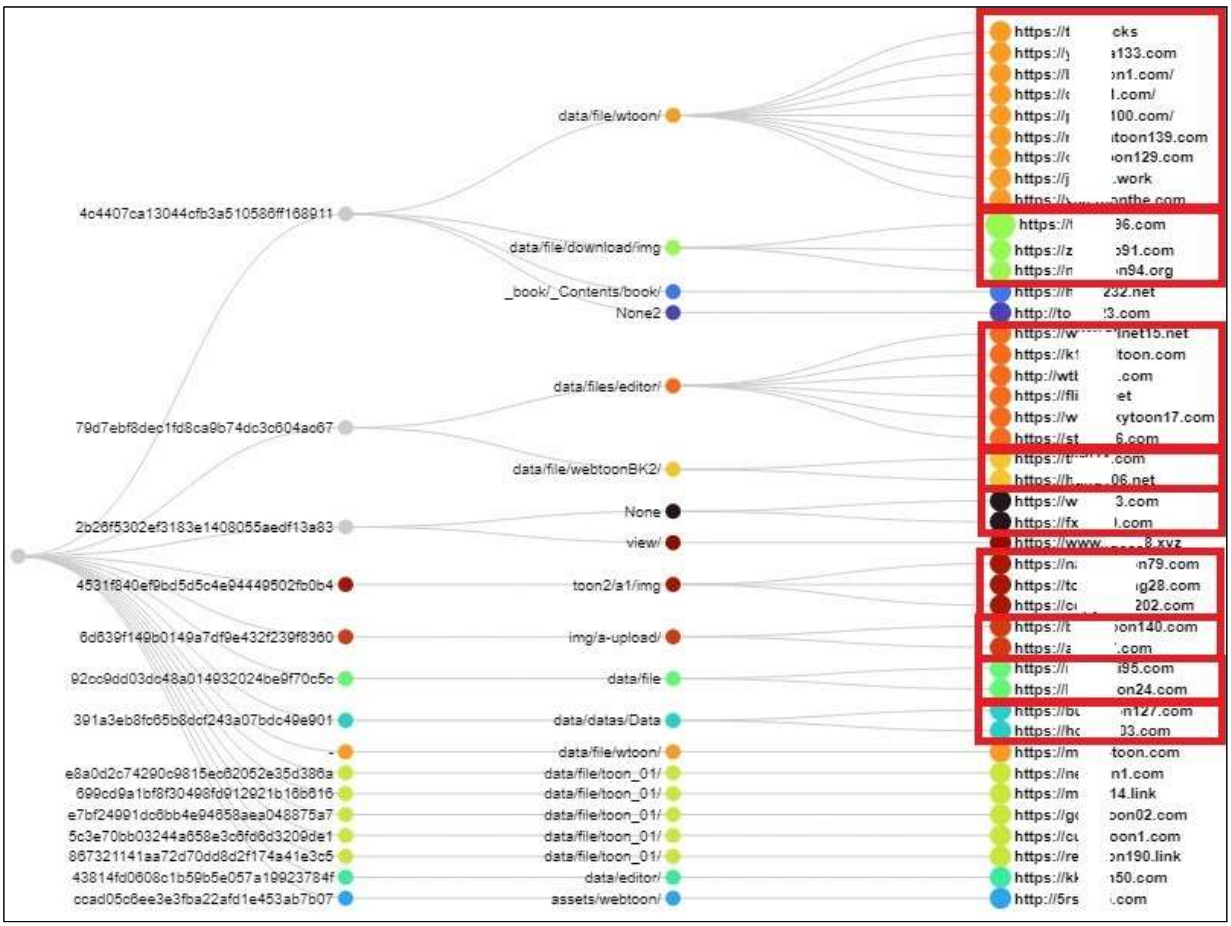
〈Figure 4〉 Result of grouping based on image hash and image storage server

첫 번째 방법인 이미지 해시 값만 같은 경우는 사이트에 대한 정보가 누락되어 너무 많은 사이트가 연관된 것으로 해석되며, 두 번째 방법인 이미지 저장서버를 기준으로 분석한 경우 이미지 저장서버가 같다면 이미지의 해시값도 같기 때문에 첫 번째 기준의 결과에 포함되고 이미지 저장서버가 똑같은 사이트만 그룹핑되기 때문에 연관 있어 보이는 불법 사이트를 포함하지 못하는 지엽적인 결과를 보여준다. 그런데 사이트의 하위 경로는 사이트 생산자의 특성이 있기 때문에 이미지 해시값과 하위 경로가 동일한 경우에 대한 분석을 진행하였다. 이러한 조건하에서 불법 사이트가 두 개 이상 그룹화 되는 경우를 확인해보면, 총 9개의 그룹이 연관성을 보인다. 즉, 총 31개의 불법사이트에서 한 그룹당 최소 2개에서 최대 9개의 사이트가 연관된 것으로 분석되며, 결과는 [그림 6]과 같다.

[그림 6]을 보면 세 번째 방법인 이미지의 해시값과 하위경로를 기준으로 그룹핑한 결과에서 하위경로가 같지만 이미지해시가 다른 불법 사이트들이 다수 식별되었으며, 이를 분석해본 결과 각 이미지의 해상도가 다른 것으로 확인되었다. 이는 이미지의 해시값이라는 기준도 지엽적으로 작용되어 연관성 있어 보이는 사이트를 연관 짓지 못하는 결과를 내는 것으로 보인다. 따라서 이미지의 해시값 대신 이미지 파일에서 추출할 수 있는 데이터인 DQT(Define Quantization Table)의 사용을 제안한다.



〈Figure 5〉 Grouping results based on image storage server



〈Figure 6〉 Result of grouping by image hash, sub-path

DQT는 jpg 이미지에서 획득할 수 있는 메타데이터로 본 논문에서 사용된 불법 사이트 모두 jpg 형태로 웹툰을 배포하고 있어 특징으로 사용할 수 있다. DQT는 jpg 이미지를 생성할 때 압축률에 영향을 미치기 때문에 어플리케이션에 따라 효율적인 값을 사용하며, 동일한 이미지 처리 프로세스를 사용하거나 동일한 어플리케이션을 사용하는 경우 생성된 jpg 파일의 DQT를 통해 구분이 가능하다[11]. 따라서 불법 웹툰 이미지의 "DQT"와 웹서버 구조의 특징인 "하위경로"를 기준으로 사이트간 연관성을 분석해 보았다. 결과는 〔그림 7〕과 같으며, 총 42개의 불법사이트 중 36개의 불법사이트를 10개의 그룹으로 구분할 수 있다.

〔그림 6〕에서 하위경로가 "data/file/toon_01/" 인 사이트의 경우 이미지의 해시값이 모두 달라 같은 그룹으로 묶을 수 없었지만 네 번째 방법을 사용했을 때의 결과인 〔그림 7〕을 보면 위에서 세 번째로 그룹화된 불법 배포 사이트는 모두 같은 DQT를 사용하고 있어 같은 그룹으로 묶인 것을 볼 수 있으며 이 그룹은 동일한 Google Analytics ID를 사용하고 있어, 사이트 간 연관성이 있을 뿐 아니라 동일한 운영자가 운영하고 있음을 추정할 수 있다.

이 결과는 위에서 사이트 연관성을 분석하기 위해 가정한 세 가지 방법인 "이미지의 해시값이 동일한 경우", "이미지 저장서버가 동일한 경우", "이미지 해시값과 하위 경로가 동일한 경우"로 그룹화한 결과에 모두 부합하며, 연관성이 없어 보였던 불법 사이트도 같은 그룹으로 묶어주는 더 나은 결과를 보여주었다.

〔그림 7〕의 결과로 웹툰을 복제하는데 동일한 소프트웨어를 사용하고 웹서버 구조가 같은 웹툰 불법 배포 사이트를 구분할 수 있어, 이 경우 수사 중 유포자 뿐 아니라 불법 복제 이미지를 생성했을 가능성을 고려해 불법 저작물 생산자도 같이 수사하는 것을 고려할 수 있다. 따라서 본 논문에서는 사이트 연관성을 분석하는데 사용하는 특징으로 불법 웹툰 이미지의 DQT와 사이트 하위경로를 제안한다.

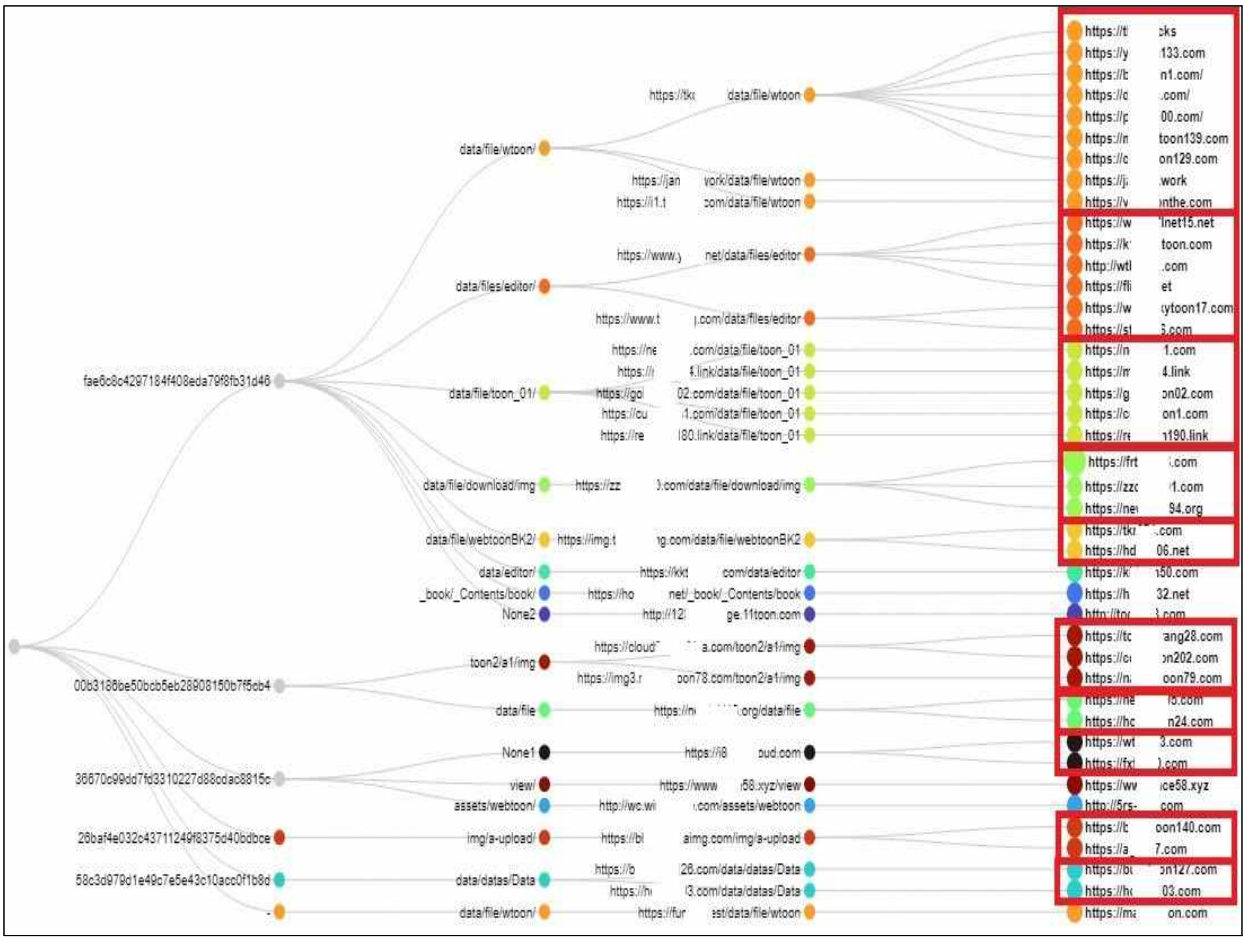
4.4 이미지 저장 서버를 이용한 차단방법

불법 사이트에 대한 추가 정보 수집을 통해 〔그림 8〕과 같이 1개의 이미지의 저장서버가 여러 불법 사이트에서 공유되는 것을 확인할 수 있다. 따라서 불법 사이트를 차단할 때, 신고된 도메인 주소만 차단하는 것이 아닌 해당 이미지가 저장되는 서버도 같이 차단한다면 이미지 저장서버 1개를 차단했을 때 최소 1개에서 최대 7개의 사이트를 차단할 수 있으며, 이 방법을 사용하면 신고되지 않은 다른 불법 사이트도 차단하는 간접적인 효과도 볼 수 있을 것이다.

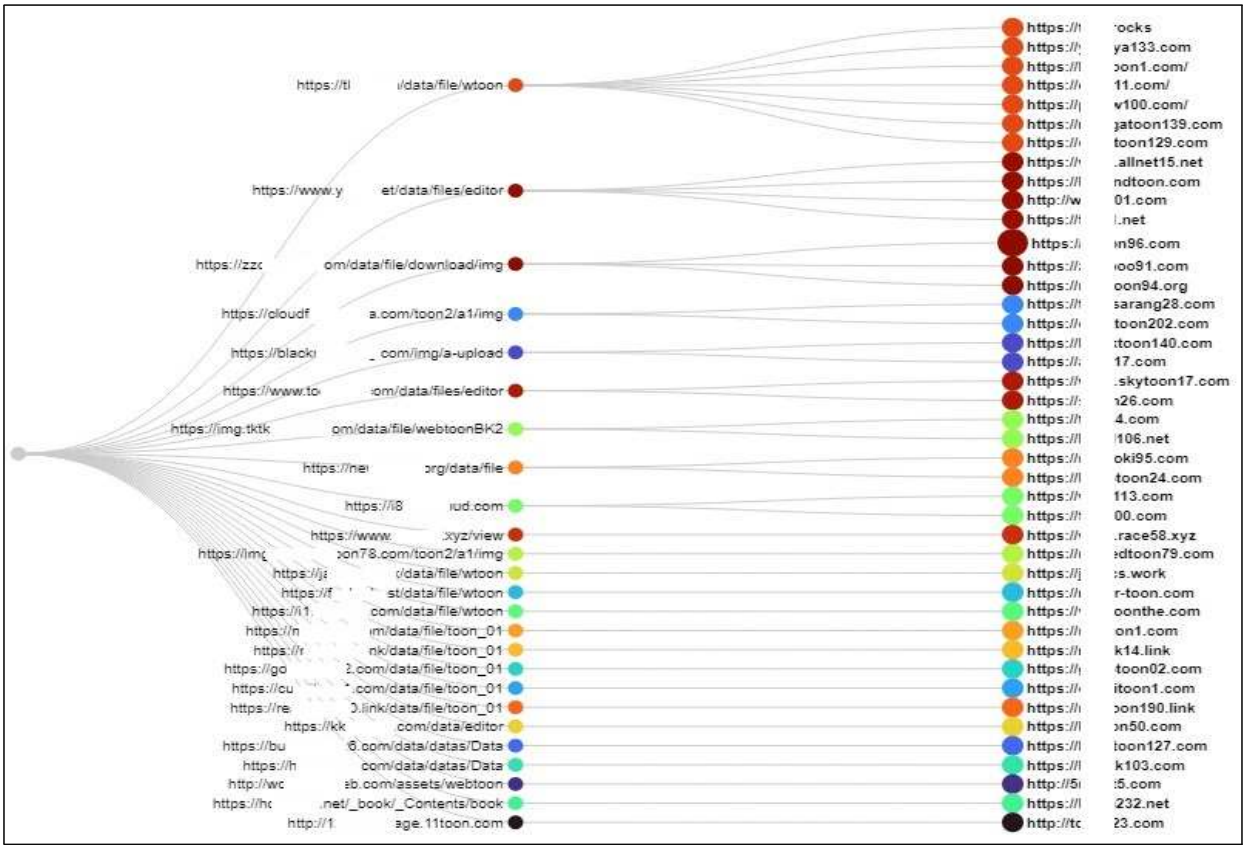
Ⅴ. 결 론

웹툰 불법저작물에 대한 피해를 근절하기 위해서 불법사이트 폐쇄가 필수적이다. 하지만 국제공조수사에 걸리는 시간은 길기 때문에, 현실적으로 모든 불법 사이트에 대해 진행되기 어려워 피해규모가 큰 사이트를 우선으로 수사가 진행된다. 따라서 수사가 진행되기 전 불법 사이트의 규모를 파악해 우선순위를 선정하는 과정이 필수적이다. 2021년 4월 20일 경찰청과 문화체육관광부, 인터폴이 저작권 침해 사이트 합동단속을 실시했으며, 피해가 심각한 웹툰 등을 위주로 총 30개 사이트를 우선 선정하여 국제공조수사를 진행했다[12]. 이처럼 국제공조수사가 필요한 해외 불법사이트 폐쇄는 피해규모가 큰 사이트를 우선 선정해 진행되기 때문에 본 논문에서 제안한 사이트 연관성 분석을 통해 연관성 있는 사이트를 구분할 수 있다면, 그룹화 된 사이트를 바탕으로 불법 사이트의 규모를 파악해 국제공조수사 착수시간을 단축하는데 기여할 수 있을 것이다.

또한 본 논문에서 불법 사이트에 대한 정보를 수집하면서, 불법 웹툰 배포 사이트의 주소는 다르지만 배포되는 웹툰 이미지의 저장서버를 공유하는 사실을 확인했다. 따라서 불법 사이트를 차단할 때, 신고한 해당 URL만 차단하는 것이 아닌 해당 이미지가 저장되는 서버도 같이 차단한다면 기존에 사용하던 도메인 차단보다 최대 7배의 효율로 차단하는 효과를 기대할 수 있을 것이다.



〈Figure 7〉 Result of grouping by image DQT Hash, sub-path, image storage server



〈Figure 8〉 Piracy site group by image storage server

참 고 문 헌 (References)

[1] Korea Creative Content Agency, A Survey on Webtoon Businesses, 2020. KOCCA20-24.

[2] Korea Copyright Protection Agency, A Study on the Status of Webnovel/Webtoon Copyright Infringement and Countermeasures, KPResearch 2021-02

[3] Joongwon Jeong and Sangjin Lee, Blocking method of harmful sites based on domain change pattern, Journal of Digital Forensics 15(3), pp.39-53, 2021.9.

[4] Korea Copyright Protection Agency, A Study on the Countermeasure of Copyright Infringement via International Internet Service, KPResearch 2020-03

[5] Korea Copyright Protection Agency, Annual Report on Copyright Protection, 2021

[6] SeukYoon Kang, JuYoung Cho, GaHyeon Ju and YoungKoo Lee, Harmful Website Detection System Using Real-time Web Crawling, The Korean Institute of Information Scientists and Engineers, pp.1904-1906, 2018.6

[7] Seungyong Choo, Yeseong Hwang and Sangjin Lee. Methods for Collecting Harmful Websites Using Web Crawling. Journal of Digital Forensics, 15(3), pp.127-138, 2021.9.

[8] Hayeon Kang, Youngchul Choi and SangJin Lee, Analysis of advertisers by tracking banner ads on piracy websites, Journal of Digital Forensics 15(3), pp.15-26, 2021.9.

[9] Kiryong Lee and Heejo Lee, An Automated Technique for Illegal Site Detection using the Sequence of HTML Tags, Journal of KIISE 43(10), pp.1173-1178, 2016.10.

[10] Boungjin Kim and Sangjun Lee, Improvement of Methods for Discriminating Harmful Web Sites by using Link Relations between Web Sites and Constructing Whitelist, KIISE Transactions on Computing Practices 25(10), pp.506-510, 2019.10.

[11] Eric Kee, Micah K. Johnson, and Hany Farid, Digital Image Authentication From JPEG Headers, IEEE TRANSACTIONS ON INFORMATION FORENSICS AND SECURITY, VOL. 6, NO. 3, pp.1066-1075, 2011.9.

[12] National Police Agency, Ministry of Culture, Sports and Tourism, Interpol jointly crack down on copyright infringement sites Avilable : https://www.korea.kr/news/pressReleaseView.do?newsId =156454302 2022.03.16. confirmed

[13] 대법원 2021. 9. 9. 선고 2017도19025 전원합의체 판결 〔공2021하,1881〕

저 자 소 개

최 영 철 (YoungChul Choi)
준회원

2016년 2월 : 고려대학교 사이버국방학과 졸업
2020년 9월 ~ 현재 : 고려대학교 정보보안학과 석사과정
관심분야 : 디지털 포렌식, 정보보호

이 상 진 (SangJin Lee)
평생회원

1989년 10월 ~ 1999년 2월 : 한국전자통신연구원 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