Automated Detection Technique for Suspected Copyright Infringement Sites

Hae Seon Jeong[1] and Jin Kwak[2*]
[1] ISAA Lab., Department of AI Convergence Network, Ajou University, Suwon, Republic of Korea
[e-mail: haeseon0363@ajou.ac.kr]
[2] Department of AI Convergence Network, Department of Cyber Security
Ajou University, Suwon, Republic of Korea
[e-mail: security@ajou.ac.kr]
[*]Corresponding author: Jin Kwak

Received November 19, 2020; accepted November 26, 2020; published December 31, 2020

Abstract

With the advances in Information Technology (IT), users can download or stream copyrighted works, such as videos, music, and webtoons, at their convenience. Thus, the frequency of use of copyrighted works has increased. Consequently, the number of unauthorized copies and sharing of copyrighted works has also increased. Monitoring is being conducted on sites suspected of conducting copyright infringement activities to reduce copyright holders' damage due to unauthorized sharing of copyrighted works. However, suspected copyright infringement sites respond by changing their domains or blocking access requests. Although research has been conducted for improving the effectiveness of suspected copyright infringement site detection by defining suspected copyright infringement sites' response techniques as a lifecycle step, there is a paucity of studies on automation techniques for lifecycle detection. This has reduced the accuracy of lifecycle step detection on suspected copyright infringement sites, which change domains and lifecycle steps in a short period of time. Thus, in this paper, an automated detection technique for suspected copyright infringement sites is proposed for efficient detection and response to suspected copyright infringement sites. Using our proposed technique, the response to each lifecycle step can be effectively conducted by automatically detecting the lifecycle step.

Keywords: Automated Analysis, Copyright Detection, Copyright Infringement, Monitoring, Piracy Site

1. Introduction

Owing to the improvement in Information Technology (IT), copyrighted works, such as movies, drama, music, and webtoons, can be conveniently used in a PC or mobile environments, and users can stream or download high-quality content. Given that the content is more convenient to use with the advances in IT, the demand for use and frequency of use for various copyrighted works is increasing [1].

However, the increase in demand for copyrighted works has led to an increase in the distribution of illegal copies as well as the distribution of legal copyrights, thereby resulting in the creation of piracy sites that share illegal copies without the permission of copyright holders [2, 3]. Accordingly, to protect the rights of the copyright holders, content management agencies are conducting copyright protection activities, such as monitoring sites suspected of copyright infringement [4-8]. In Korea, the number of copyright infringement sites blocked by the Korea Communications Standards Commission, as of 2019, has more than tripled since 2017, and the number of blocked sites is expected to increase [9].

However, suspected copyright infringement sites bypass detection by blocking access requests from specific Internet Protocols (IPs) or by creating new domains and changing the site access address [10]. If suspected copyright infringement sites block access to site, then automated access request is impossible. In addition, if suspected copyright infringement sites change their domain, access target cannot be secured, thereby, content management agencies cannot monitor suspected copyright infringement sites.

Therefore, an effective detection technique is required to respond to suspected copyright infringement sites. Hence, studies have been conducted to detect suspected copyright infringement sites by defining the response steps of these sites as a lifecycle model. However, it is difficult to manually detect suspected copyright infringement sites that change domain access addresses and change the lifecycle steps in a short period of time. Therefore, the development of features and processes for automated detection techniques for the lifecycle step is required. Accordingly, in this paper, by analyzing each step of the lifecycle's characteristics and associations, we propose features and processes for automated detection techniques for the lifecycle step of copyright infringement sites. We also demonstrate that the proposed technique applies to actual, suspected copyright infringement sites via the acceptance ratio of automated detection technique for the lifecycle step. Additionally, we improve the effectiveness of suspected copyright infringement site detection by proposing preemptive measures for monitoring and responding to these sites.

The remainder of this paper is organized as follows. In Section 2, we review related research on the characteristics and lifecycle model of piracy sites. In Section 3, we propose features and processes for automated detection techniques of the lifecycle step of copyright infringement sites and preemptive countermeasures for each detected lifecycle step. In Section 4, we analyze the experimental results for the lifecycle step detection, and finally, we present the conclusions in Section 5.

2. Related Work

2.1 Analysis of Piracy Sites

According to the statistics from the Korea Copyright Protection Agency, copyright infringement sites share copyrighted works in the order of video streaming, webtoon posting, webhard, torrents, and portals, and the number of online service providers sharing this work is listed in **Table 1**. In this paper, we select torrent, video streaming, and webtoon publishing piracy sites for research because webhard sites support alliance services and portal sites are operated as a community [9]. Common and type-specific characteristics of piracy sites are as follows:

Table 1. Status of OSP subject to online piracy monitoring in 2018 [9]

Type	Number of Online Service Providers (OSPs)
Video streaming	88
Etc (Webtoon)	87
Webhard	52
Torrent	41
Portal	5
Total	273

❒ Analysis of common characteristics of piracy sites

As piracy sites are operated without formal registration, unlike in **Fig. 1**, the business registration number does not exist.

Fig. 1. Example of business registration number in a legitimate site

Furthermore, most piracy sites generate revenue by posting illegal gambling advertising banners within advertising banners [11]. As shown in **Fig. 2**, gambling-related keywords, such as “Casino,” “Toto,” “Welcome bonus,” and “Deposit,” are commonly used in illegal gambling advertising banners [12].

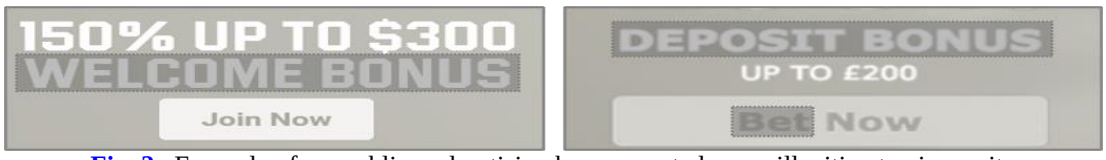


Fig. 2. Example of a gambling advertising banner posted on an illegitimate piracy site

❒ Analysis of characteristics of video streaming sites

In the case of video streaming sites, streaming services are provided on the site by posting video links from other streaming servers. Therefore, it is common for keywords related to other streaming servers, such as “HLSPlay,” “FlashVid,” and “SuperVid,” and keywords related to video streaming types, such as “Link” and “Host,” to exist, as shown in **Fig. 3** [12, 13].

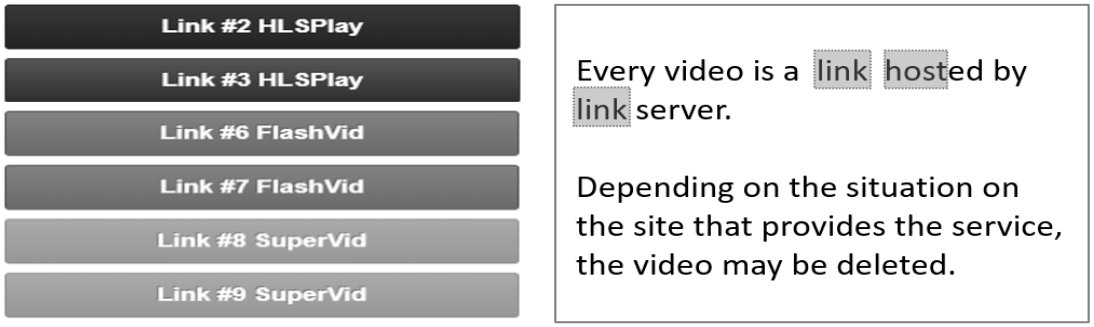


Fig. 3. Example of video streaming type

❒ Analysis of characteristics of torrent sites

In the case of torrent sites, copyrighted works are not provided directly from the server but from users. Therefore, it is common for keywords related to torrent types such, as "Magnet," "Torrent," and "Seed," to exist, as shown in **Fig. 4** [12, 14].

Fig. 4. Example of a torrent type

❒ Analysis of characteristics of webtoon posting sites

In the case of webtoon posting sites, copyrighted works are provided with a watermark containing the site's domain or keywords related to illegal gambling advertisement banners. Therefore, it is common for domain or keywords related to illegal gambling advertisements to exist, as shown in **Fig. 5** [12].

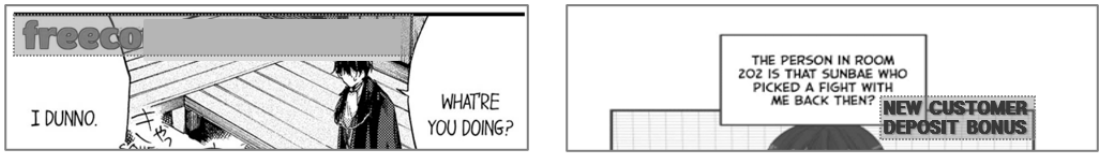


Fig. 5. Example of webtoon types

As piracy sites exhibit common and type-specific features that identify them, it is possible to detect piracy sites by using these characteristics.

2.2 Lifecycle Analysis

2.2.1 Lifecycle Step Definition

The lifecycle model, which is a response step according to the time flow of the piracy sites, can be defined as five steps involving creation, operation, response, change, and closure, as shown in **Fig. 6**. Detailed definitions of each step of the lifecycle are as follows [15]:

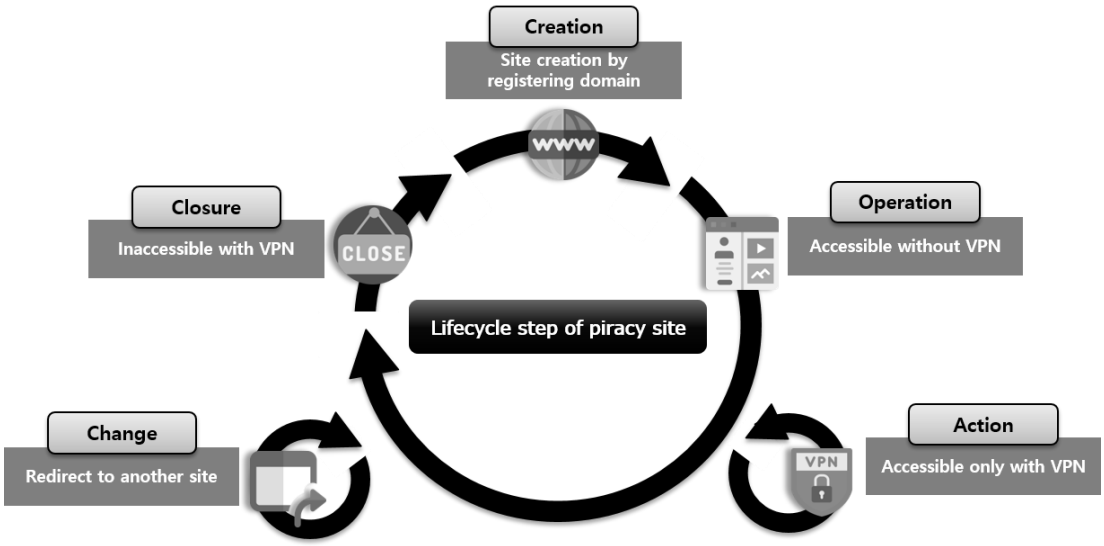


Fig. 6. Lifecycle step of piracy site

- Creation: In this step, the piracy site is first created via domain registration.
- Operation: In this step, copyright infringement sites are operated by sharing copyright infringement contents from operators or users.
- Action: In this step, bypass access is available with the newest domain of a piracy site or normal access is not possible. However, only bypass access is possible with the previous version domain of a piracy site.
- Change: In this step, normal access is available with the newest domain of a piracy site and then redirected to another piracy site or normal access is not possible. However, only bypass access is possible with the previous version domain of a piracy site and then redirected to another piracy site.
- Closure: In this step, normal access is unavailable and the site is no longer operated.

2.2.2 Acceptance Ratio of a Lifecycle Model

According to the research on the development of the lifecycle model for piracy sites by Kim et al., the features (w_{1-6}) corresponding to the characteristics of the piracy sites are weighted. Specifically, w_1 (0.5) feature is added at site creation and w_{2-5} (0.1) features can be added during the operation–action–change steps. Furthermore, w_6 (0.1) feature is added at the point of site closure. Therefore, the maximum sum of weights per cycle is 1.0. By calculating the sum of weights for the corresponding features, an acceptance ratio is measured to verify the application of the lifecycle model to actual piracy sites [15]. Features corresponding to the piracy site characteristics and their corresponding weight values are shown in **Table 2**.

Table 2. Features for an acceptance ratio

Feature	Weight	Description
Site creation	w_1 (0.5)	New site creation with new domain registration
Similar domain	w_2 (0.1)	Generate similar domain
Membership	w_3 (0.1)	Manage membership database
Advertising banner	w_4 (0.1)	Place gambling advertisement in piracy site
Guidance of new domain	w_5 (0.1)	Guide new domain through Social Networking Service (SNS), banner
Site closure	w_6 (0.1)	Close site

The calculation formula for the acceptance ratio refers to the percentage of weight sum ($\sum_{x=1}^{S} w_x$) for each piracy site ($1 \leq i \leq$ N) to the product of the total number of sites (N) and total sum of weights (1.0). Descriptions of the notations used in the calculation formula are shown in **Table 3**.

Table 3. Notations for an acceptance ratio

Notation	Description
AR	Acceptance ratio (%)
N	Total number of suspected infringement sites ($1 \leq N$)
i	Suspected infringement site (i) with the lifecycle model ($1 \leq i \leq N$)
W	Sum of the total weights ($W = 1.0$)
w_x	Weight for corresponding features ($1 \leq x \leq S$)
S	Total number of features ($S = 6$)

$$AR = \frac{\sum_{i=1}^{N} i(\sum_{x=1}^{S} w_x)}{N \times W} \times 100\% \quad (1)$$

However, the features in **Table 2** typically exist at various lifecycle steps. but may also exist for the same lifecycle step. Therefore, the sum of weights may vary for the same lifecycle step and can be the same for different lifecycle steps. For example, sites A and B in **Table 4** are in the same operation step but exhibit different weight sums. Additionally, the lifecycle steps for sites A, C, and D are different with respect to operation, change, and closure, respectively. However, the weight sums are all identical and correspond to 1.2. Therefore, the lifecycle steps cannot be identified by weight sum. This implies that the calculation formula

for the acceptance ratio in expression (1) does not consider whether the actual lifecycle step of each piracy site has been detected, making it difficult to measure the valid acceptance ratio. Therefore, a calculation formula for the acceptance ratio considering the lifecycle step is required.

Table 4. Examples of non-identifiable lifecycle step with respect to weight

Site	Step	Weight						Sum of weights
		w_1 (0.5)	w_2 (0.1)	w_3 (0.1)	w_4 (0.1)	w_5 (0.1)	w_6 (0.1)	
A	Operation	O	X	O	X	X	X	0.6
B	Operation	O	O	O	X	X	X	0.7
C	Change	O	O	X	X	X	X	0.6
D	Closure	O	X	X	X	X	O	0.6

3. Proposed Automated Detection Technique

In this study, we draw features by analyzing the characteristics and associations of lifecycle steps. We propose an automated detection technique for lifecycle steps of suspected copyright infringement sites and present an acceptance ratio calculation formula to verify the performance of detection techniques for the lifecycle steps of suspected copyright infringement sites. Additionally, we improve the effectiveness of the response to each detected lifecycle step by proposing preemptive countermeasures.

3.1 Features for Identifying Lifecycle Step

In this section, we propose features for automated detection techniques that can detect lifecycle steps by analyzing the features and associations. The details of the features are as follows.

❒ Existence of domain registration information

In the creation step, the domain is registered in the WHOIS server ("whois.markmonitor.com" or "whois.nic.ac"), and information, such as date of domain creation and registrar name, is present within the WHOIS server. This feature corresponds to all lifecycle steps. However, the creation step does not exhibit all the characteristics of the other steps. Therefore, the characteristics of the other phases can be examined to identify the creation step. Hence, domain registration information within the WHOIS server can be used as a feature for detecting lifecycle steps.

❒ Normal accessibility

Normal access before Virtual Private Network (VPN) execution is possible during the operation step, and normal access is not possible at the action and closure steps. Although normal access after VPN activation is possible at the action step, it is not possible at the closure step. Therefore, normal accessibility can be used to detect operation, response, and closure lifecycle steps.

❒ Reception of response code, implying closure

A response code indicating site closure is received when sending a connection request to a site in the closure step. Therefore, the reception of response code, implying closure, can be used to detect the closure lifecycle step.

❒ Redirection to another piracy site

During the change step, the suspected copyright infringement site is redirected to another suspected infringement site when accessing the site. As suspected copyright infringement sites create new domains similar to existing domains based on the rules for creating new domains, the reception of non-similar response domains can be checked for redirection to another suspected infringement site. Therefore, redirection to another piracy site can be used to detect the change lifecycle step.

❒ Existence of features for suspected copyright infringement sites

In the operation step, a suspected copyright infringement site shares illegal copies and includes keywords related to illegal gambling advertising and types of piracy within the webpage. In the case of the action step, the characteristics of the suspected copyright infringement sites can be checked after bypass access. In the change step, the characteristics can be checked after redirection. To sum up, the suspected copyright infringement site features are common in the operation, action, and change steps, but do not exist in the creation step. Therefore, the existence of features for suspected copyright infringement sites can be used to detect the creation, operation, action, and change steps.

The notation of features for automated detection of the lifecycle step for suspected copyright infringement sites proposed in this section is shown in **Table 5**, and the details of corresponding features for each lifecycle step are as follows.

Table 5. Features for identifying lifecycle step

Feature	Notation
Existence of domain registration information	f_1
Normal Accessibility	f_2
Reception of response code implying closure	f_3
Redirection to another piracy site	f_4
Existence of features for suspected copyright infringement sites	f_5

- Creation: A domain (f_1) is created. However, it does not possess normal accessibility (f_2), does not receive the response code, implying closure (f_3), is not redirected to another piracy site (f_4), and does not possess features (f_5) of suspected copyright infringement sites.

- Operation: A domain (f_1) is created to enable normal access (f_2) and possesses the features for suspected copyright infringement sites (f_5). However, it does not receive a response code, implying closure (f_3) and does not redirect to another piracy site (f_4).

- Action: A domain (f_1) is created to operate a suspected copyright infringement site and possesses the characteristics of a suspected copyright infringement site (f_5). However, normal access (f_2) is not possible, and it does not receive response code, implying closure (f_3) and does not redirect to another piracy site (f_4).

- Change: A domain (f_1) is created to operate a suspected copyright infringement site and possesses the characteristics of a suspected copyright infringement site (f_5) and redirects to another piracy site (f_4). However, normal access (f_2) is not possible, and it does not receive a response code, implying closure (f_3).

- Closure: Although the domain was created (f_1), normal access is not possible even after activating the VPN (f_2). Alternatively, the response code, implying closure is received (f_3) .

The expressions for lifecycle steps using the notations in **Table 5** are as shown in **Table 6**.

Table 6. Expressions for lifecycle steps

Step	Expression
Creation	$f_1 \wedge (\sim (f_2 \vee f_3 \vee f_4 \vee f_5))$
Operation	$f_1 \wedge f_2 \, (\sim (f_3 \vee f_4)) \wedge f_5$
Action	$f_1 \wedge (\sim (f_2 \vee f_3 \vee f_4)) \wedge f_5$
Change	$f_1 \wedge (\sim f_3) \wedge f_4 \wedge f_5$
Closure	$(f_1 \wedge f_3) \vee (f_1 \wedge (\sim f_2))$

3.2 Automated Detection Process for Lifecycle Steps

The detection of copyright infringement sites will be ineffective if the domain is changed or the domestic IP access is blocked. Therefore, lifecycle step detection should precede the detection of suspected copyright infringement sites. To this end, based on the features in Section 3.1, a process of automated detection for the lifecycle steps of suspected copyright infringement sites is proposed in this paper.

3.2.1 Detection Process for Lifecycle Steps

In this section, a process of automated detection for the lifecycle steps of suspected copyright infringement sites is proposed by subdividing the features in Section 3.1, as shown in **Fig. 7**. The automated detection process is performed in four steps. 1) Check the existence of the domain based on the domain registration information within the WHOIS server of the site being inspected. 2) Analyze whether a valid domain can be accessed via an occurrence of error when requesting access. This generates a list of sites that require VPN bypass because the sites cannot be accessed appropriately and checks if the site can be accessed normally after VPN activation. 3) Analyze the response information to check whether a response code, implying closure has been received and whether it is redirected to another suspected copyright infringement site. 4) Analyze the page source code of the suspected copyright infringement site to confirm the existence of the suspected copyright infringement site features. The detailed process of the automated detection technique for the lifecycle steps is as follows.

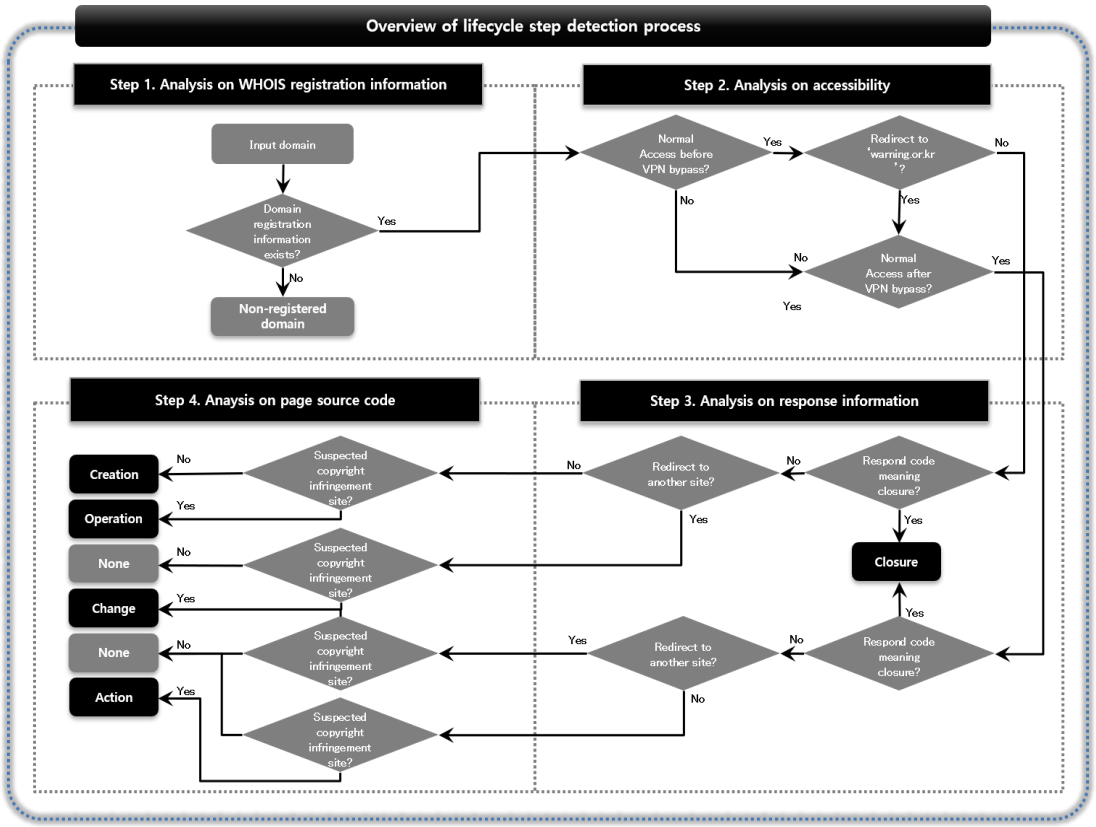


Fig. 7. Automated detection technique for lifecycle step

Step 1. Analysis of existence of domain registration information in WHOIS server

Operating systems (OSs), such as Linux and Windows 10, provide domain registration information within the WHOIS server through the "whois" command, as shown in **Fig. 8** [16].

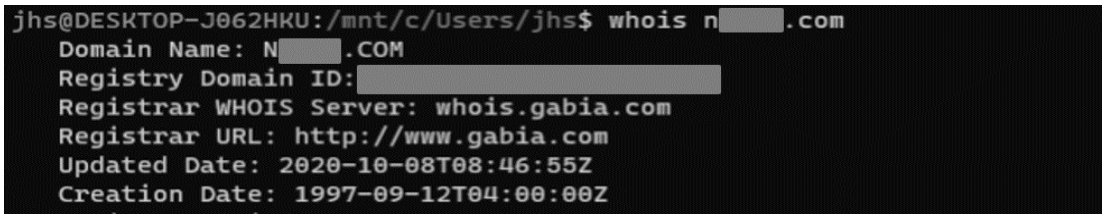


Fig. 8. Example of "whois" command result for registered domain

If the domain is not registered, the WHOIS server sends a response with keywords that imply that the domain is not registered, such as the "No match for domain" keyword in **Fig. 9**.

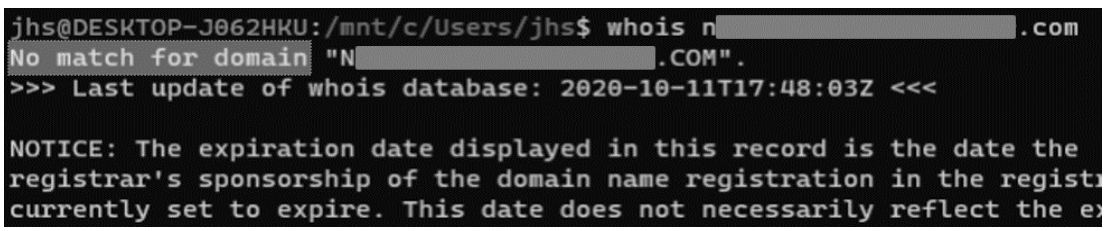


Fig. 9. Example of "whois" command result for non-registered domain

Therefore, the results obtained after executing the "whois" command are analyzed by utilizing Python' s "os" module, and the pseudo-code for this is shown in **Table 7**. If a keyword, which implies that the domain is not registered, is present in the results of the "whois" command response, then the domain is detected as a non-registered domain.

Table 7. Pseudo-code for Step 1

Function for check domain registration
1 def check_WHOIS(domain):
2 command = "whois" + domain
3 process = os.popen(command)
4 result = str(process.read())
5 if "keyword when domain is registered" in result:
6 *# Domain registered in WHOIS*
7 else:
8 *# Domain not registered in WHOIS*

Step 2. Analysis of normal accessibility

The normal accessibility in the case of a valid domain is analyzed. If normal access is not possible, then the Korea Communications Standards Commission redirects access to the "warning.or.kr" domain to block users from accessing the copyright infringement site. If normal access is not possible, then the actual lifecycle step cannot be detected because the response domain, response code, and page source code that are verified during normal access to the site cannot be analyzed. Therefore, before analyzing the actual response information, page source code, etc, are used to analyze the normal accessibility.

To this end, after analyzing normal accessibility in Step 2-1, we analyze suspected copyright infringement sites in which normal access is possible but redirected to "warning.or.kr" in Step 2-2. Thus, we can generate a list of sites requiring VPN bypass access because they cannot normally access and analyze the normal accessibility after VPN activation for the sites in that list. The pseudo-code for Step 2 is shown in **Table 8**, and the details for Step 2 are as follows.

Table 8. Pseudo-code for Step 2

Function for analysis of accessibility
1 def check_accessibility(domain):
2 *# Step 2-1. Normal access without VPN*
3 try:
4 *# Normal access without VPN*
5 res = requests.get(url,Verify=False)
6 *# Step 2-2. Access denied*
7 if res.url == "warning.or.kr":
8 *# Add to VPN Site List*
9 except requests.ConnectionErroras e:
10 *# Add to VPN Site List*
12
13 def Step_2-3(VPN Site List):
14 *# Step 2-3. Normal access with VPN*
15 try:
16 *# Normal access with VPN*
17 res = requests.get(url,Verify = False)
18 except requests.ConnectionErroras e:
19 *# Identify closure step*

Step 2-1. Analysis of normal accessibility before VPN activation

If the suspected copyright infringement site blocks access, normal access is not possible when VPN is not activated and access errors occur. In Python's requests module, "Connection Error" occurs when access is blocked by the site [17]. Therefore, the handling of exceptions in **Table 8** generates a list of VPN bypass access target sites that are not normally accessible.

Step 2-2. Analysis of blocked site by the Korea Communications Standards Commission

If the Korea Communications Standards Commission blocks a site, then normal access is possible. However, it is redirected to the "warning.or.kr" domain where the webpage of **Fig. 10** is printed. Therefore, if the response domain matches "warning.or.kr," then it is determined that the site is not accessible and added to the VPN bypass access target list.

Fig. 10. Website of "warning.or.kr"

Step 2-3. Analysis of normal accessibility after VPN activation

Analyze the availability of normal access after VPN activation for the VPN bypass target site list. In the action step, normal access is possible after VPN activation, and in the closure step, normal access is not possible even if VPN is activated. Therefore, if normal access is not possible after VPN activation, then it is detected as a closure step.

Step 3. Analysis of response information

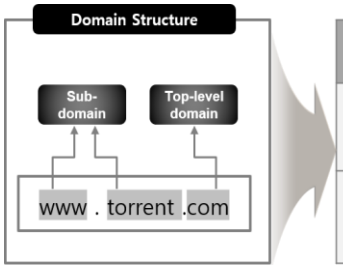
The change and closure steps can be detected via the response domain and response code, etc, received during access requests. Thus, we analyze the response code in Step 3-1 to identify the closure step and analyze redirection to another suspected copyright infringement site based on the response domain and detect the change step in Step 3-2. The details of Step 3 are as follows.

Step 3-1. Analysis of response code

If 404 and 500 response codes, which imply that a server of the domain does not exist [18], are received, then it is detected as a closure step. Furthermore, 403 response code denotes that sites can be operated normally. However, if the response page contains "Forbidden" or "Access denied" keywords, then it is detected as a closure step.

Step 3-2. Analysis of redirection to another suspected copyright infringement site

The change step is redirected to another site when accessing the site, and redirection can be identified by analyzing the response domain. A suspected copyright infringement uses a domain similar to existing domains as a new domain as per rules in **Fig. 11**. Hence, if a non-similar response domain that does not follow the rules in **Fig. 11** are received, then it can be detected as a change step.



# of sub-domain	Rules for generating new domain	Example	
		Current domain	New domain
1	Add new sub-domain	torrent.com	**www.**torrent.com
	Change top-level domain	torrent.**com**	torrent.**site**
	Change/Remove number of keyword domain	torrent.com	torrent**6**.com
2	Remove sub-domain	**www.**torrent.com	torrent.com
	Change top-level domain	www.torrent.**com**	www.torrent.**me**
	Change/Remove number of keyword domain	www.torrent.me	www.torrent**2**.me

Fig. 11. Rules for generating new domains

When components are created based on the separator "." for the domain, they are composed of top-level domains and sub-domains. The top-level domain is the component on the right side, and the sub-domain is the component on the left [19].

Suspected copyright infringement sites create new domains by adding or increasing numbers in the front or rear side of the sub-domains when creating new domains of the same site. If there are two sub-domains, then a new domain is created by adding or changing numbers on the sub-domain on the right side.

For the sub-domain subject to change, the number in the sub-domain is removed and defined as the keyword domain. If the keyword domain of the requesting domain and the response domain do not match, then it is detected as redirected to another site and as the change step. The pseudo-code for Step 3-2 is shown in **Table 9**.

Table 9. Pseudo-code for Step 3-2

Function for check redirection to another site

```
def extract_sub_domain(urls):
    domains = []
    name_keywords = []
    for url in urls:
        # Extract domain
        tmp = url.split("//")[-1]
        domain = tmp.split("/")[0]
        # Delete Top-level domain
        d_split = domain.split(".")
        sub_domains = []
        if len(d_split) == 3:
            # Domain with three element
            if check_top_level(d_split[1]) == 1:
                # One sub-domain
                sub_domains.append(d_split[0])
            else:
                # Two sub-domains
                sub_domains.append(d_split[0])
                sub_domains.append(d_split[1])
        else:
            # Domain with Two element
            sub_domains.append(d_split[0])
```

Step 4. Analysis of page source code

Suspected copyright infringement sites display illegal gambling advertisement banners and

piracy-related keywords on the main page. Furthermore, keywords that can identify piracy types, such as video streaming, torrent, and webtoon posting, exist on the internal page. Thus, suspected copyright infringement sites can be detected via page source code analysis [12].

3.2.2 Detection Process for Suspected Copyright Infringement Sites

The detection process for suspected copyright infringement sites used in page source code analysis within the lifecycle step detection process is shown in **Fig. 12**. The process consists of analyzing common characteristics of suspected copyright infringement sites and then analyzing the existence of keywords related to the types of piracy [12]. The detailed process of automated detection technique for the suspected copyright infringement sites is as follows:

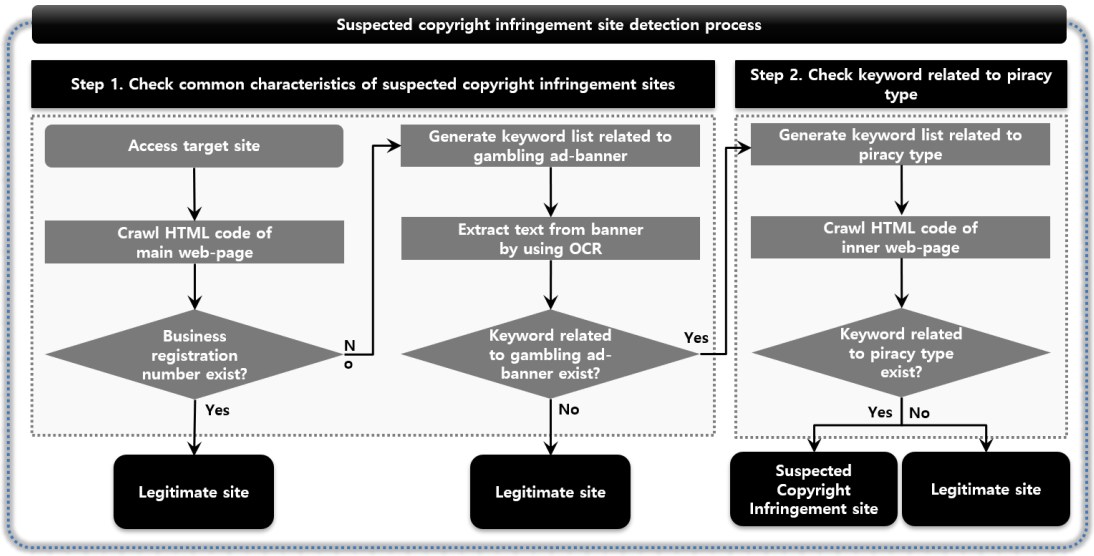


Fig. 12. Copyright infringement site detection process

Step1. Check common characteristics of suspected copyright infringement sites

As suspected copyright infringement sites are not officially registered, an analysis of the existence of a business registration number on the main page is required. If the business registration number does not exist, then it can be estimated that it is a suspected copyright infringement site.

Furthermore, given that such sites generate revenue by posting gambling advertisements, it can be assumed as a suspected copyright infringement sites when the text on the main page extracted through Optical Character Registration (OCR) contains gambling-related keywords such as “Casino,” “Toto,” “Welcome bonus,” and “deposit.”

If the business registration number does not exist and gambling-related keywords are identified, then the following analysis process detects suspected copyright infringement sites.

Step2. Check keywords related to the piracy type

The existence of keywords related to the piracy type should to be analyzed because keywords can identify the type of piracy on the internal page. If the internal page contains keywords in **Table 10**, then it can be detected as a suspected copyright infringement site.

Table 10. List of keywords related to piracy type

Type	Category	Keyword
Video streaming	Streaming server	"JawLoad", "HLSPlay", "FlashVid", "SuperVid"
	Link	"Link", "Host"
Torrent		"Magnet", "Seed", "Torrent"
Webtoon	Domain	The domain of target site
	Gambling advertisement	"Casino", "Toto", "Welcome bonus", "Deposit"

3.2.3 Acceptance Ratio for Automatically Detected Lifecycle Step

The Calculation formula (2) the acceptance ratio is based on the features in Section 3.1 and on the automated detection process in Section 3.2, the percentage of suspected copyright infringement sites ($\sum_{i=1}^{N} m_i$) to the actual lifecycle step detected for the number of sites analyzed (N). The notations and descriptions used in the calculation formula are shown in **Table 11**. Equation (2) considers whether each lifecycle is detected.

Table 11. Notations for acceptance ratio

Notation	Description
AR	Acceptance ratio (%)
N	Total number of suspected infringement site ($1 \leq N$)
i	Suspected infringement site (i) with the lifecycle step ($1 \leq i \leq N$)
m_i	A site that lifecycle step is identified ($1\ if\ matched, 0\ if\ not\ matched$)

$$AR = \frac{\sum_{i=1}^{N} m_i}{N} \times 100\% \tag{2}$$

3.3 Preemptive Countermeasure for Detected Lifecycle Step

Effective detection of suspected copyright infringement sites can be performed if we can preemptively respond to possible threats corresponding to each detected lifecycle step. Preemptive countermeasures with respect to detected lifecycle steps of suspected copyright infringement sites are as follows.

❒ Preemptive countermeasures with respect to the transition to operation when the creation step is detected

After the domain is created, the operation of the suspected copyright infringement site can be initiated. Thus, a preemptive response regarding the transition to the operation step is required. To this end, by continuously sending access requests for the current domain, we can quickly detect the transition to the operation step by analyzing response information and page source code.

❒ Preemptive countermeasure to a new domain when the operation, action, change, and closure steps are detected

In the case of operation, action, change, and closure steps, suspected copyright infringement sites can change access addresses to a newly created domain. Thus, a preemptive response is required because detection cannot be conducted if access addresses are changed to

a new domain. To this end, by creating a list of possible new domains based on the domain creation rules in **Fig. 11**, we can quickly detect the creation of a new domain by continuously checking the existence of WHOIS registration information for that list.

❒ Preemptive countermeasure about the transition to change when the operation and action steps are detected

In the case of operation and action steps, they can be redirected to another suspected copyright infringement site. Thus, a preemptive response is required because detection cannot be performed if the response domain is not secured. A preemptive response is generated by continuously identifying the response domain and quickly detecting the change step.

❒ Preemptive countermeasures about the transition to action when operation and change steps are detected

In the operation and change steps using the newest domain, the suspected copyright infringement site does not block users' access requests. However, if access to a suspected copyright infringement site is blocked, a preemptive response is required because it cannot be detected due to inaccessibility. To this end, the response information is continuously checked, and if the connection is blocked, a preemptive response is considered by bypassing the block through VPN activation. **Table 12** shows the possible threats in each detected step of the lifecycle and the corresponding preemptive countermeasures to them.

Table 12. Preemptive countermeasures for detected lifecycle step

Possible threat	Preemptive countermeasure	Detected lifecycle step				
		Creation	Operation	Action	Change	Closure
Switch to operation step	Analyze characteristics of suspected piracy site by analyzing response information and page source code	O	-	-	-	-
New domain creation	Generate possible new domain list and continuously analyze WHOIS registration	-	O	O	O	O
Switch to change step	Continuously check for matches between the connection request domain and the response domain	-	O	O	-	-
Switch to action step	Continuously check for response information	-	O	-	O	-

4. Experimental Analysis of the Proposed Technique

4.1 Design of Experiments

4.1.1 Experimental Setup

An automated detection technique for the lifecycle step of suspected copyright infringement

site is implemented using Python 3.8 programming language in a Windows 10 Pro 64-bit environment. Additionally, Touch VPN application is executed in a Windows 10 Pro 64-bit environment to analyze sites that require VPN activation by bypassing access blocking. The form of input and output data is in the form of a CSV file. Python's Pandas module is used to read and write data to a CSV file, and the "os" module is used to analyze the results of "whois" command execution. Furthermore, "requests" module is used to transmit access requests and analyze response information.

Table 13. Experimental setup

Component	Specification
OS	Windows 10 Pro 64-bit
Programming language	Python 3.8
VPN	Touch VPN

4.1.2 Dataset

In this paper, experiments are conducted on piracy sites of video streaming, torrent, and webtoon posting. The dataset is composed of a self-built dataset and dataset from the Korea Copyright Protection Agency, which consists of 87 video streaming sites, 75 torrent sites, and 56 webtoon sites, which comprise a total of 218 piracy sites.

If a Cloudflare is applied to a site as a bypass technique for automatic analysis, such as crawling, then the site is excluded from the dataset because the site information that can be verified on normal access is not available with Cloudflare [20]. With the exception for Cloudflare, the piracy site dataset consists of 85 streaming sites, 56 torrent sites, and 54 webtoon sites, which comprise of 195 piracy sites. The number of sites with Cloudflare and piracy sites via lifecycle step is shown in **Table 14**.

Table 14. Actual lifecycle step of the dataset

Step	Actual lifecycle step			
	Video steaming	Torrent	Webtoon	Total
Cloudflare	2	19	2	23
Creation	-	-	-	-
Operation	10	9	9	28
Action	44	18	17	79
Change	6	7	0	13
Closure	25	22	28	75
Total	87	75	56	218

4.2 Results of Experimental Analysis

In the analysis, with the exception of 23 sites with Cloudflare applied, lifecycle steps of 195 copyright infringement sites were detected and their acceptance ratio was measured. Thus, the WHOIS server's domain registration information can be determined in all the experimental datasets.

In the case of video streaming piracy sites, one site was detected as a closure step. However, a normal response code was received and the site was then detected as an operation step. Two sites corresponded to change step. However, they were not redirected to another piracy site

but induced click-through advertising banners. Hence, these two sites were detected as an operation step. In the case of torrent piracy sites, one site corresponded to an operation step. However, the response code, implying closure was received and then detected as a closure step. In the case of piracy sites of webtoon posting, one site corresponded to the operation step. However, the response domain did not comply with domain creation rules and was detected as a change step. Details of the detection results are as shown in **Table 15**.

Table 15. Detection results of lifecycle steps

Step	Actual	Result	Actual	Result	Actual	Result
	Video- streaming		Torrent		Webtoon	
Creation	-	-	-	-	-	-
Operation	10	13	9	8	9	9
Action	44	44	18	18	17	16
Change	6	4	7	7	0	1
Closure	25	24	22	23	28	28
Total	85	85	56	56	54	54

Therefore, the lifecycle steps of 190 sites from 195 piracy sites were detected. Thus, more than 95 % acceptance ratio was measured for all types of piracy sites, and the acceptance ratio of the total dataset was measured at 97.44 %, as shown in **Table 16**. This lifecycle acceptance ratio is more than 20 % higher than that in a previous study [15].

Table 16. Acceptance ratio comparison

	Proposed method				D.H. Kim et al. [15]		
	Video-streaming	Torrent	Webtoon	Total	Torrent	Webtoon	Total
Acceptance ratio	96.47 %	98.21 %	98.15 %	97.44 %	77.02 %	75 %	76.27 %

5. Conclusion

In this paper, we proposed the features and technique for automated detection of the lifecycle step of suspected copyright infringement sites performing video streaming, torrent, and webtoon posting. We improved the effectiveness in detecting suspected copyright infringement sites by proposing processes for detecting suspected copyright infringement sites after automatic detection of lifecycle steps and presenting preemptive countermeasures for each detected lifecycle step. Thus, the acceptance ratio was measured at approximately 97 %, which corresponded to an improvement of approximately 20 % when compared to that of a previous study. Therefore, the automated technique for detecting suspected copyright infringement sites, as proposed in this paper, can improve the accuracy and effectiveness of detection for suspected copyright infringement sites.

Acknowledgement

This research project was supported by Ministry of Culture, Sports and Tourism (MCST), and from Korea Copyright Commission in 2020 (2019-PF-9500).

References

[1] P. Sherrell, "International Comparative Legal Guide to: Copyright 2019," Global Legal Group Ltd., London, United Kingdom, 5th Edition, Oct. 2018. Article (CrossRef Link)
[2] D. Zhang, "Illegal File Sharing & The Film Industry," B.S. thesis, Dept. Economics, California Univ., Berkeley, CA, USA, 2015.
[3] European Union Intellectual Property Office (EUIPO), "Online Copyright Infringement in the European Union: Music, Film and TV (2017-2018), Trends and Drivers," European Union Intellectual Property Office, Alicante, Spain, Nov. 2019. Article (CrossRef Link)
[4] MPA Asia Pacific, "MPA Study on Site Blocking Impact in South Korea," MPA Asia Pacific, Los Angeles, CA, United States, May 2018. Article (CrossRef Link)
[5] Creative Content Australia, "Site Blocking Laws in Australia," Creative Content Australia, Sydney, Australia, Article (CrossRef Link)
[6] High Authority for the dissemination of works and the protection of rights on the internet (Hadopi), "Anti-piracy Strategies Concerning Cultural and Sports Content -2019 International Survey," High Authority for the dissemination of works and the protection of rights on the internet (Hadopi), Paris, France, May 2020. Article (CrossRef Link)
[7] Ofcom, "Site Blocking" to reduce online copyright infringement," Ofcom, London, United Kingdom, May 2011. Article (CrossRef Link)
[8] K. E. Noonan, "2019 Review of Notorious Markets for Counterfeiting and Piracy," Office of the United States Trade Representative Executive Office of the President, Washington DC, United States, May 2020. Article (CrossRef Link)
[9] Korea Copyright Protection Agency, "2019 annual report on copyright protection," Korea Copyright Protection Agency, Seoul, Korea, Aug. 2019. Article (CrossRef Link)
[10] Incopro, "Site blocking efficacy in Portugal -September 2015 to October 2016," Incopro, London, United Kingdom, Feb. 2020. Article (CrossRef Link)
[11] Office for Harmonization in the internal market, "Digital advertising on suspected infringing websites," Office for Harmonization in the internal market, Provincia de Alicante, Spain, Jan. 2016. Article (CrossRef Link)
[12] S. K. Choi and J. Kwak, "Feature analysis and detection techniques for piracy sites," *KSII Transactions on Internet and Information Systems*, vol. 14, no. 5, pp. 2204-2220, May. 2020. Article (CrossRef Link)
[13] S. Oh, S. Wallsten, and N. Lovin, "Do pirated video streams crowd out non-pirated video streams? evidence from online activity," Technology Policy Institute, Washington DC, United States, Jan. 2020. Article (CrossRef Link)
[14] A. K. Sharma and N. Sharma, "Bit torrent (peer to peer network): Antipiracy and Anonymity," *International Journal of Science and Research*, vol. 4, no. 7, pp. 253-256, July 2015. Article (CrossRef Link)
[15] D. H. Kim, H.S. Jeong, and J. Kwak, "Development of lifecycle model for copyright infringement site," *Journal of The Korea Institute of Information Security & Cryptology*, vol. 30, No. 1, pp. 101-121, Feb. 2020. Article (CrossRef Link)
[16] E. Paz, "Cyber-Attacks discovery via analysis of DNS," M.S. thesis, Dept. Computer Science Division, The Open Univ., Ranana, Israel, 2020.
[17] K. Reitz, "Requests Documentation," K. Reitz, Release 2.25.0, Nov. 2020. Article (CrossRef Link)

[18] DELL, "REST API Programmer's Guide," DELL, Version 5.x, Austin, Texas, June 2019.
Article (CrossRef Link)
[19] Huawei Cloud, "Domain Name Service," *Service Overview*, no. 6, Feb. 2020.
Article (CrossRef Link)
[20] Cloudflare, "Securing Applications in the Cloud," Cloudflare, Austin, Texas, 2019.
Article (CrossRef Link)

Hae Seon Jeong is a postgraduate student in master's course at Dept. of AI Convergence Network in Ajou University, Republic of Korea. She received the B.S. degree from Ajou University, Republic of Korea. Her research interests include Copyright protection, IoT security, Vulnerability & Malware analysis.

Jin Kwak is a professor at Dept. of AI Convergence Network and Dept. of Cyber Security in Ajou University, Republic of Korea. He received the Ph.D. degree from SKKU, Republic of Korea. His research interests include Copyright protection, Cryptographic protocols, Applied security mechanisms for cloud and big data system and so on.