Intelligent Piracy Site Detection Technique with High Accuracy

Eui-Jin Kim[1] and Jin Kwak[2*]
[1] ISAA Lab., Department of Computer Engineering
Ajou University, Suwon, Republic of Korea
[e-mail: dmlwls0403@ajou.ac.kr]
[2] Department of Cyber Security, Department of AI Convergence Network
Ajou University, Suwon, Republic of Korea
[e-mail: security@ajou.ac.kr]
*Corresponding author: Jin Kwak

Received November 20, 2020; accepted November 26, 2020;
published January 31, 2021

Abstract

Recently, with the diversification of media services and the development of smart devices, users have more opportunities to use digital content, such as movies, dramas, and music; consequently, the size of the copyright market expands simultaneously. However, there are piracy sites that generate revenue by illegal use of copyrighted works. This has led to losses for copyright holders, and the scale of copyrighted works infringed due to the ever-increasing number of piracy sites has increased. To prevent this, government agencies respond to copyright infringement by monitoring piracy sites using online monitoring and countermeasure strategies for infringement. However, the detection and blocking process consumes a significant amount of time when compared to the rate of generating new piracy sites. Hence, online monitoring is less effective. Additionally, given that piracy sites are sophisticated and refined in the same way as legitimate sites, it is necessary to accurately distinguish and block a site that is involved in copyright infringement. Therefore, in this study, we analyze features of piracy sites and based on this analysis, we propose an intelligent detection technique for piracy sites that automatically classifies and detects whether a site is involved in infringement.

Keywords: Copyright Infringement, Copyright Protection, Digital Content, Government Monitoring, Piracy Site

http://doi.org/10.3837/tiis.2021.01.016 **ISSN : 1976-7277**

1. Introduction

Recently, with the improvement in communication speed and popularization of smartphones, a culture that streams music, movies, and videos in a short time has emerged as an environment irrespective of time and space; consequently, resulting in the diversification of online-based media services [1]. Subsequently, users have more opportunities to use digital content, and as the reach of the copyrighted works increases and expands into overseas markets, the size of the copyright market has also increased accordingly [2].

However, as the copyright market size increased, piracy sites that illegally distributed copyrighted works appeared. Piracy sites have many problems, mainly involving illegal copying and propagation of copyrighted works without the copyright holder's permission. This leads to significant losses for the copyright holder [3]. Piracy sites generate revenue by illegally distributing content, such as movies, dramas, TV programs, and webtoons, and generate revenue for site operation by posting illegal advertisement banners on their site. Distributing malicious codes in such advertisement banners can lead to various types of damage, such as leakage of personal information [4].

To prevent this piracy, several public agencies use online surveillance to monitor piracy sites and diversify their piracy response strategies such as billing blocks and domain blocks. Additionally, copyright infringement is prevented through enforcement activities by imposing a fine for online service providers who violate copyright or by closing piracy sites [5].

Although efforts to block piracy sites has continued in the aforementioned manner, the rate of generation of new piracy sites is faster than the speed of detecting piracy sites, and it is difficult to prevent copyright infringement using this technique. Additionally, piracy sites are developing in the same way as legitimate sites, and it is difficult to determine whether they are involved in copyright infringement [3]. Therefore, in this study, we propose an intelligent piracy site detection technique with high accuracy that can respond to piracy site's rapid creation speed and accurately determine whether the site is involved in copyright infringement. In the remaining sections of this paper is structured as follows: In Section 2, we analyze the features of piracy sites and legitimate sites as related research. Furthermore, we analyze the features of the illegal advertisement banner existing in the piracy site and analyze the F1 score to verify the accuracy of the proposed technique. In Section 3, we propose an intelligent detection technique that considers the infringement of a piracy site, and in Section 4 we describe the experimental setup in which the intelligent detection technique is tested. Additionally, the dataset derived from the link collection site, and the results of applying the dataset to the intelligent detection technique are described. Finally, we conclude with Section 5.

2. Related Work

2.1 Features of Piracy Sites and Legitimate Sites

We analyze the features of piracy sites that illegally distribute copyright works and legitimate sites that legally distribute copyright works.

2.1.1 Features of Piracy Sites

Types of piracy sites include torrent sites, video streaming sites, webtoon sites, and link collection sites, and we analyzed the features of each piracy site. **Table 1** presents the types of copyright works infringed by piracy sites, and **Fig. 1** represents the features of the piracy site.

Table 1. Type of content infringement exhibited by piracy sites

	Piracy Sites			
	Torrent	Video Streaming	Webtoon	Link Collection
Broadcasting Films	O	O	X	O
Movie	O	O	X	O
Webtoon	X	X	O	O

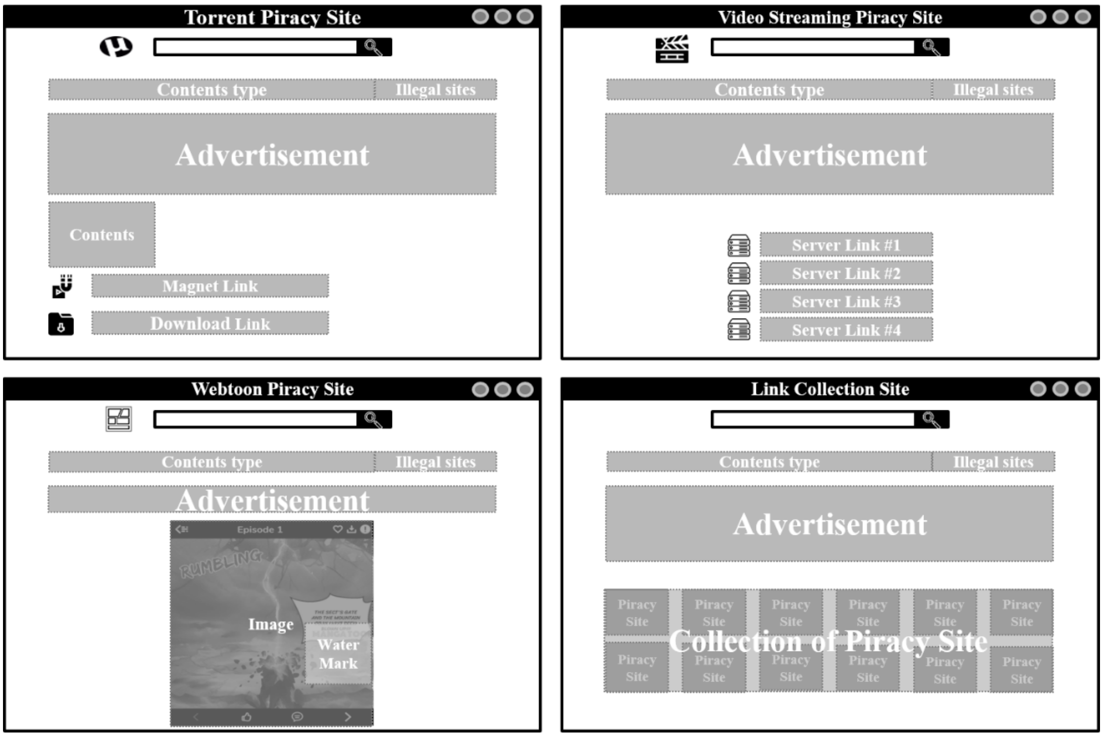


Fig. 1. Example of piracy sites

In general, a torrent site shows the type of content, such as a broadcast or movie, on the top menu, and a link to an illegal site and an advertisement banner. Furthermore, the magnet link and download link are provided for the users to download the content for free [3].

As torrent site, an video streaming site shows the type of content on the top menu, a link to the illegal site, and an advertisement banner. Additionally, external links can use the corresponding content, and it is generally configured to provide a link to another server to stream video [3].

Webtoon is a compound word of web and cartoon and refers to cartoons serialized on the web. Webtoon sites include various webtoon genres, links to illegal sites, and advertisement banners on the top menu. Furthermore, to prevent data leakage to other piracy sites, watermarks may be inserted into webtoon image files [3].

The link collection site is a site that lists piracy sites and provides them to clients. As a feature of a link collection site, an illegal advertisement banner is usually posted inside the site, and many piracy sites exist in the form of a banner in the frame inside the website. The link collection site provides torrents, video streaming, webtoons, and illegal sites, such as adult-rated and gambling sites, and posts links to sites reflecting domains that change periodically due to the short life cycle of the piracy site.

2.1.2 Features of Legitimate Sites

As legitimate site features, there are legitimate video streaming sites that provide broadcasting and movies and legitimate webtoon sites that provide webtoons. As a feature of a legitimate site, there is a business registration number, which implies that the site provides legitimate content services. This business registration number is the simplest way to determine whether the site is legitimate. **Table 2** lists the types of works distributed on legitimate sites, and **Fig. 2** shows the features of a legitimate site.

Table 2. Types of content featured on legitimate sites

	Legitimate Sites	
	Video Streaming	Webtoon
Broadcasting Film	O	X
Movie	O	X
Webtoon	X	O

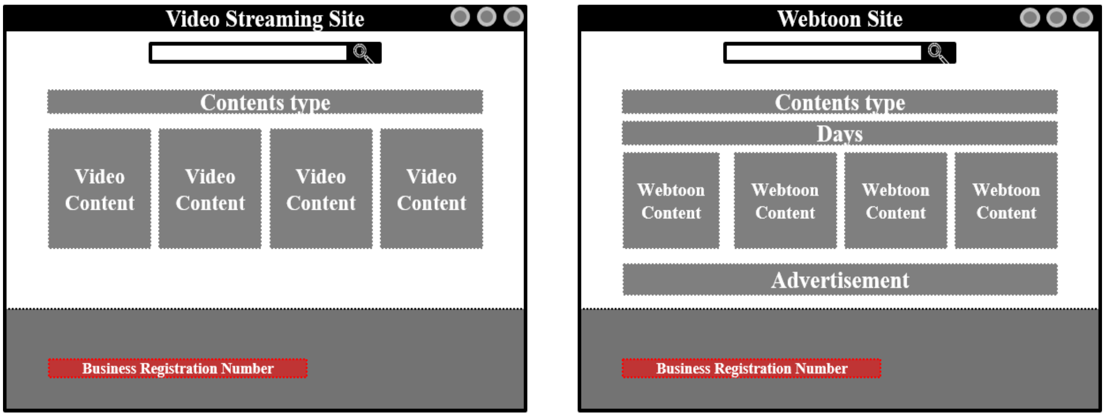


Fig. 2. Example of a legitimate site

As a feature of a legitimate video streaming site, content types are listed on the menu at the top of the website, and there are contents in the form of images. A business registration number is provided by the business operator at the bottom of the website to indicate legal content services.

The types of works are listed at the top of the website in a legitimate webtoon site, and copyright works are shown in the form of images. Furthermore, a business registration number is available at the bottom of the website. **Table 3** presents a comparison of piracy sites and legitimate sites.

Table 3. Comparison between a piracy site and legitimate site

	Piracy Site			**Legitimate Site**	
	Torrent	Video Streaming	Webtoon	Video Streaming	Webtoon
Feature	■ Ad Banners on the website ■ Contents type at the top of the website ■ Link to an illegal site(gambling, porn, etc.) within the website			■ Business registration number at the bottom of the website ■ Contents type at the top of the website	
	■ Magnet Link ■ Download Link	■ External link for streaming video to another server	■ Watermark in middle of the webtoon image		

2.2 Advertisement Banner in Piracy Sites

The piracy site distributes the work free of charge without permission from the copyright holder. However, it has a feature of posting advertisements for the site's operation. Advertising forms within piracy sites exist as AD ads and Ad banners on the main page. Advertisement banners are posted within specific frames existing within the piracy site, and such Ads are most widely used within piracy sites. Additionally, to accurately determine the in-depth information of an Ad banner, and to determine the piracy site—it is necessary to analyze the text area and image area within the Ad banner. However, there are features, such as those shown in **Fig. 3**, that make it difficult to extract text because of the boundaries between the text area and the image area within the Ad banner are ambiguous [6, 7].

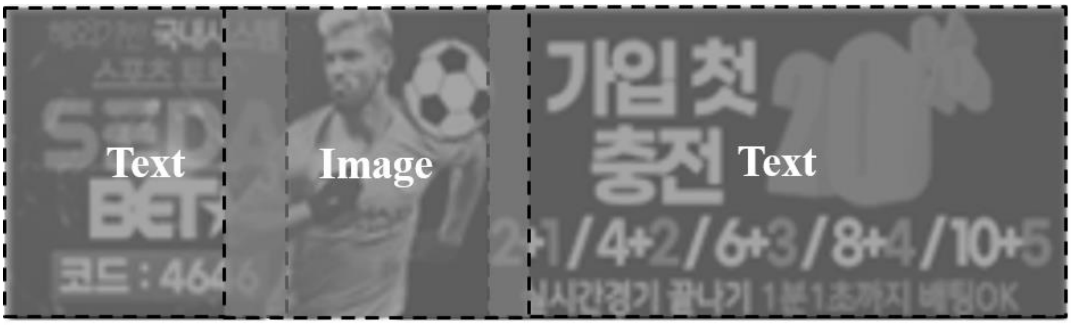


Fig. 3. Example of an advertisement banner in a piracy site

2.3 F1 Score

To calculate the F1 score, a 2 × 2 confusion matrix is used as a method for accuracy verification. As per **Table 4**, the confusion matrix is divided into true positive (TP), true negative (TN), false positive (FP), and false negative (FN) [8].

Specifically, TP denotes the case of predicting true that is true in reality, and FP denotes the case of predicting true that is false in reality. Furthermore, FN denotes the case of predicting false that is true in reality, and TN denotes the case of predicting false that is false in reality.

Table 4. Format of the confusion matrix

		True Condition	
		True	False
Predicted Condition	True	TP	FP
	False	FN	TN

The precision and recall to derive the F1 score can be derived using the confusion matrix in **Table 5**, and precision denotes the ratio of cases correctly predicted as true to those predicted as true. Hence, precision can be derived from (1). Additionally, recall is the ratio of cased predicted to be true to those that are true, and the recall can be derived through (2). The F1 score can be calculated using the derived precision and recall, and the F1 score can be derived as (3) through the harmonic average of (1) and (2) [8].

$$Precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall = \frac{TP}{TP+FN} \tag{2}$$

$$F1\ score = 2\ \times \frac{Precision\ \times Recall}{Precision\ + Recall} \tag{3}$$

By using the F1 score derived using precision and recall, the accuracy of the proposed intelligent piracy site detection technique can be verified.

3. Intelligent piracy site detection technique

In this section, the intelligent piracy site detection technique is described. Furthermore, **Fig. 4** illustrates the proposed technique.

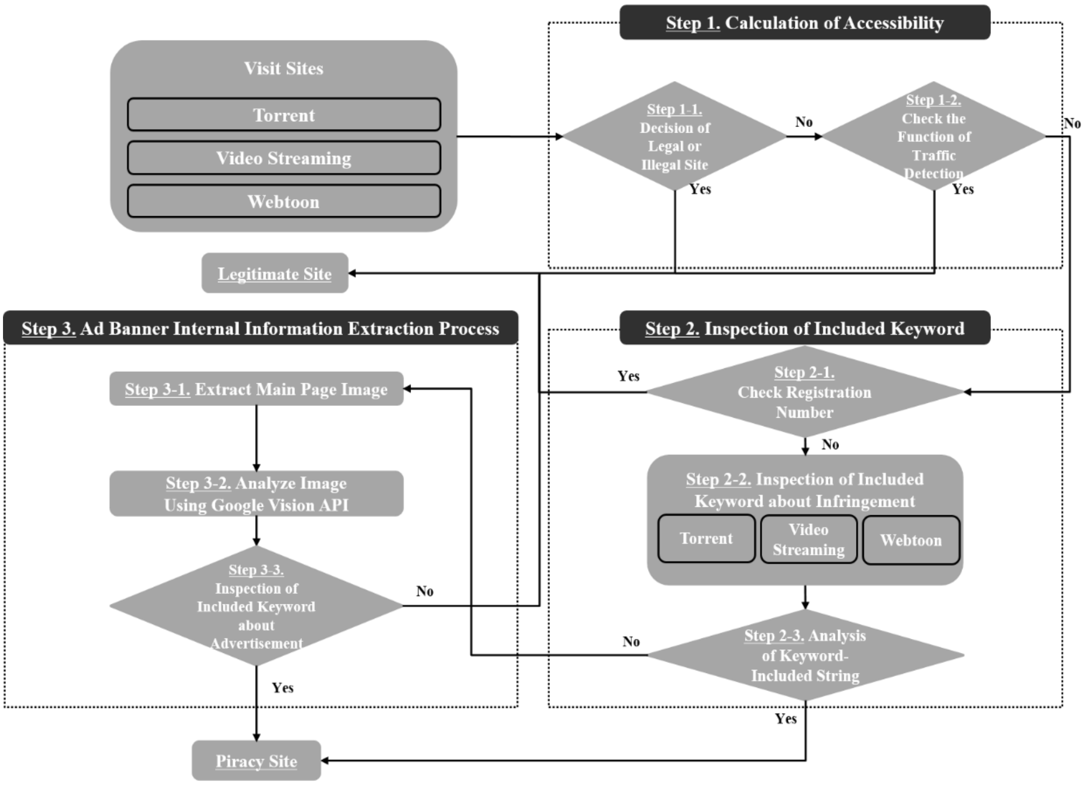
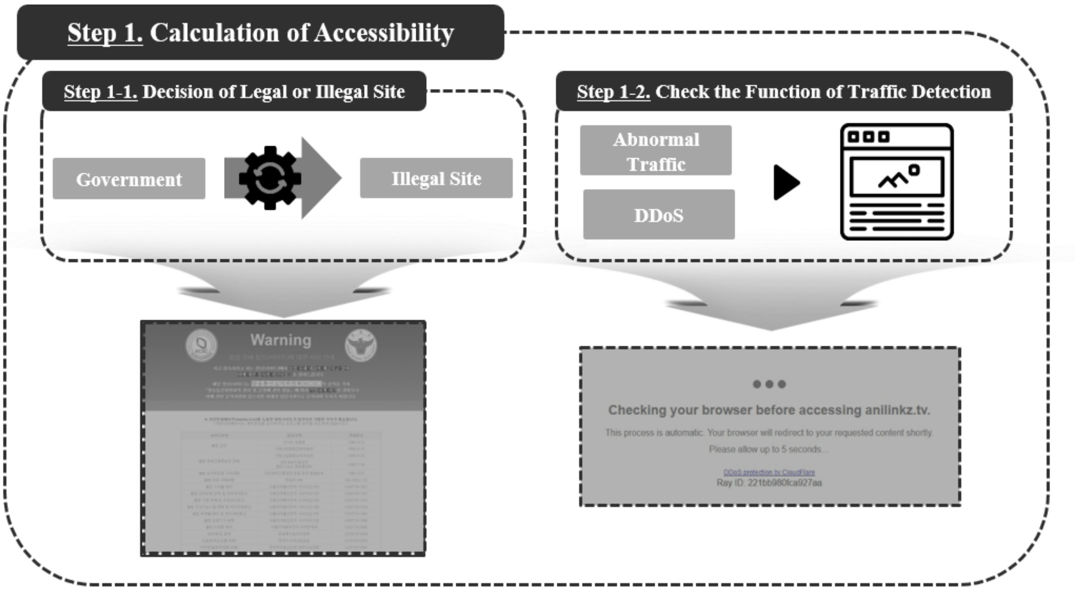

Step 1. Calculation of Accessibility
Visit Sites
Torrent
Video Streaming
Webtoon
Step 1-1. Decision of Legal or Illegal Site
No
Step 1-2. Check the Function of Traffic Detection
No
Yes
Yes
Legitimate Site
Step 3. Ad Banner Internal Information Extraction Process
Step 3-1. Extract Main Page Image
Step 3-2. Analyze Image Using Google Vision API
Step 3-3. Inspection of Included Keyword about Advertisement
No
Yes
Step 2. Inspection of Included Keyword
Yes
Step 2-1. Check Registration Number
No
Step 2-2. Inspection of Included Keyword about Infringement
Torrent
Video Streaming
Webtoon
No
Step 2-3. Analysis of Keyword-Included String
Yes
Piracy Site

Fig. 4. Overview of intelligent piracy site detection technique

Step 1. Calculation of Accessibility

This process determines whether the input site can be accessed normally and extracted from the page source. Furthermore, **Fig. 5** illustrates the process for Step 1, and **Table 5** presents the pseudo code for Step 1.


Step 1. Calculation of Accessibility
Step 1-1. Decision of Legal or Illegal Site
Government
Illegal Site
Step 1-2. Check the Function of Traffic Detection
Abnormal Traffic
DDoS
Warning

Checking your browser before accessing anilinkz.tv.

This process is automatic. Your browser will redirect to your requested content shortly.

Please allow up to 5 seconds...

DDoS protection by CloudFlare
Ray ID: 221bb980fca927aa

Fig. 5. Calculation for accessibility

Table 5. Pseudo code for calculation of accessibility

```
Function for calculation of accessibility:
str_html ← Crawled site using Selenium
warning_site ← Sites previously defined as illegal or harmful sites
traffic_tech ← Traffic detection page url
IF there is warning_site url that matches str_html,
   IF there is a traffic detection url in str_html that is matched to traffic_tech
   ELSE, RETURN -1
ELSE, RETURN -1
```

Step 1-1. Decision of Legitimate or Piracy Site

First, the input site is accessed, and it is determined in advance whether it is an illegal or harmful site by the government and agencies. If the input site is redirected to an illegal or harmful site, further analysis of the site is not required. If the input site is previously defined as an illegal or harmful site, the new site is analyzed. Otherwise, the analysis process moves on to Step 1-2.

Step 1-2. Check the Function of Traffic Detection

In browsing the input site, it is determined whether a technology capable for detecting and preventing abnormal traffic, such as 'DDoS CloudFlare', is used. If abnormal traffic detection technology is present, then analysis is impossible because the page cannot be accessed. If the site contains a URL that detects abnormal traffic, then the new site is analyzed. Otherwise, the analysis process moves on to Step 2-1.

Step 2. Inspection of Included Keywords

This process identifies keywords that exist in the piracy site page source, and if they exist, then they are judged as infringement. However, if they do not exist, then they are not judged as infringements. Furthermore, **Fig. 6** illustrates the process for Step 2, and **Table 6** presents the pseudo code for Step 2.

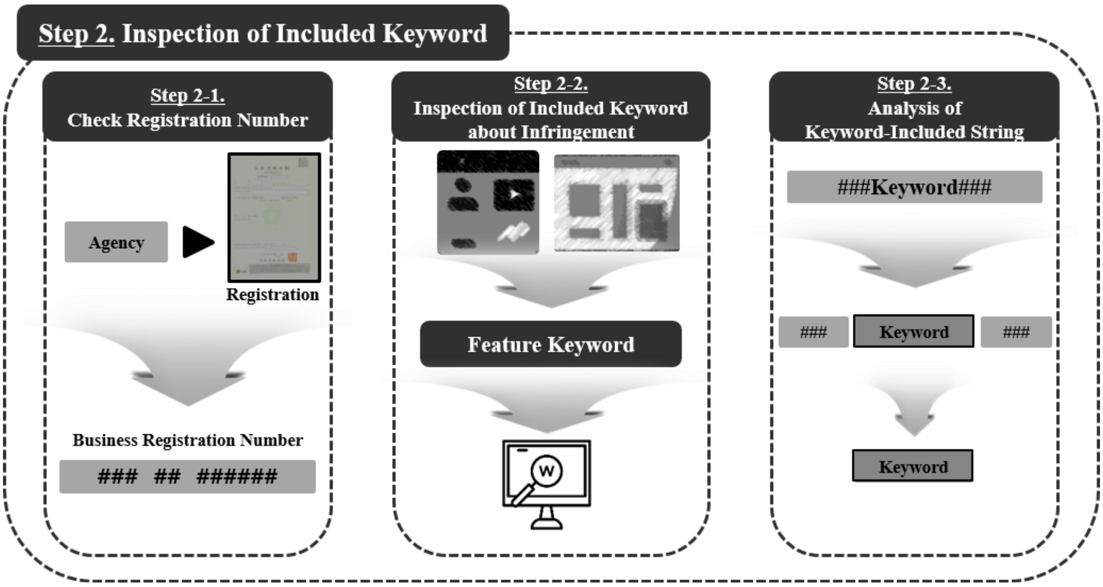

Step 2. Inspection of Included Keyword
Step 2-1.
Check Registration Number
Agency
Registration
Business Registration Number
######
Step 2-2.
Inspection of Included Keyword about Infringement
Feature Keyword
Step 2-3.
Analysis of Keyword-Included String
###Keyword###
###
Keyword
###
Keyword

Fig. 6. Inspection of included keywords

Table 6. Pseudo code for inspection of included keywords

```
Function for inspection of included keyword:
str_html ← Crawling Sites Received from Step 1
business_license_number ← Regular expression for determining business registration number
Keyword_list ← Keyword list used in piracy site
compoundKeyword ← Array to store after separating compound words
IF there is a business_license_number matching str_html, RETURN -1
FOR keyword in Keyword_list:
  IF Extracting strings with keywords if they match through regular expressions
  Separating the string containing keyword and saving it in compoundKeyword
  IF there is a keyword in the found keyword, RETURN 0
RETURN -1
```

Step 2-1. Check Registration Number

By following Step 1, the proposed detection technique can determine whether the business registration number exists from the extracted page source. A business registration number exists on a legitimate site, and the existence of the business registration number is analyzed using features that are not present on the piracy site. If a business registration number exists, then the new site is analyzed. Otherwise, the analysis process moves on to Step 2-2.

Step 2-2. Insepection of Included Keyword about Infringement

It is determined whether there is a keyword for each piracy site in the input page source. The keywords for each piracy site used in this process are derived from each piracy site's features. The list of keywords on the piracy site are shown in **Table 7**.

Table 7. List of keywords on piracy sites

Category	List of keywords
Torrent	'torrent', 'magnet', 'seed' etc.
Video Streaming	'HDVid', 'Streamango', 'HLSPlay', 'FlashVid' etc.
Webtoon	'BL', 'GL' etc.

Step 2-3. Analysis of Keyword-Included String

If a keyword exists in a page source, then a string containing that keyword is imported. The reason for importing a string that contains the keywords is to analyze the compound word. The compound word containing the keyword is judged to be an infringement keyword when determining whether the keyword exists. In separating compound words, the 'segment' module of Python is used. After separating the compound words, the existence of keywords is rechecked. If there is an infringement keyword, then the site is judged as a suspected infringement site. Otherwise, the analysis process moves on to Step 3-1.

Step 3. Ad Banner Internal Information Extraction Process

This process extracts the information inside the advertisement banner that exists on the piracy site to check for infringement. Furthermore, **Fig. 7** illustrates the process for Step 3, and **Table 8** presents the pseudo code for Step 3.

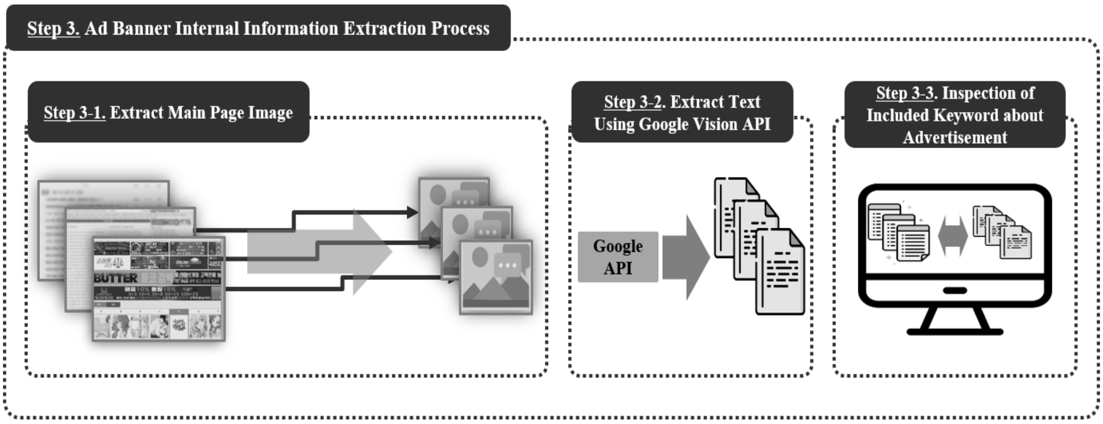

Step 3. Ad Banner Internal Information Extraction Process
Step 3-1. Extract Main Page Image
Step 3-2. Extract Text Using Google Vision API
Step 3-3. Inspection of Included Keyword about Advertisement
Google API

Fig. 7. Ad banner internal information extraction process

Table 8. Pseudo code for Ad banner internal information extraction process

```
Function for Ad banner internal information extraction process:
Ad_keyword ← Ad Banner Internal Keywords
img ← Image of web page extracted using selenium
img_text ← Text within the extracted img using Google Vision API
FOR keyword in Ad_keyword:
    IF a regular expression indicates that the keyword matches img_text, RETURN 0
RETURN -1
```

Step 3-1. Extract Main Page Image

First, the input site is accessed to extract the main page image. And, the main page image is saved in the clipboard using the Python 'Selenium' module. The saved image is used to extract the text on the main page.

Step 3-2. Extract Text Using Google Vision API

In this step, the Google Vision API is used to extract text from the extracted main page image. The Google Vision API is trained with millions of images using deep learning algorithms and convolution neural networks and has various functions [9, 10]. Among them, the optical character recognition (OCR) function can extract text within the main page image. This is used to extract the internal text of the advertisement banner that cannot be detected in the page source analysis.

Step 3-3. Inspection of Included Keyword about Advertisement

The extracted text is checked to verify if keywords are used in illegal advertisement banners and determine the existence of illegal advertisement banners for the input sites. If an ad banner keyword exists, then it is judged that illegal advertising exists on the site and that the site is suspected of infringement. Otherwise, the new site is analyzed. **Table 9** presents the keywords that exist in illegal advertising banners.

Table 9. List of keywords used in advertisement banners

Category	**List of Keyword**
Advertisement	'JACKPOT', 'CASINO', 'TOTO', 'PROTO' etc.

4. Experimental Analysis

4.1 Experimental Setup

The experiment on intelligent piracy site detection technique, proposed in this study, was performed on Windows 10 Pro 64-bit environment, and Python 3.8.5 was used as the programming language. Additionally, the Touch VPN application was used in the experimental environment's operating system to bypass the access block of the piracy site. The input data used in the experiment was in the form of a.csv file, and the output data was also in the form of a .csv file. **Table 10** presents the environment in which the experiment was conducted.

Table 10. Experimental setup

Component	Specification
Operating System	Windows 10 Pro 64bit
Programing Language	Python 3.8.5
VPN	Touch VPN

4.2 Experiments for Data Set

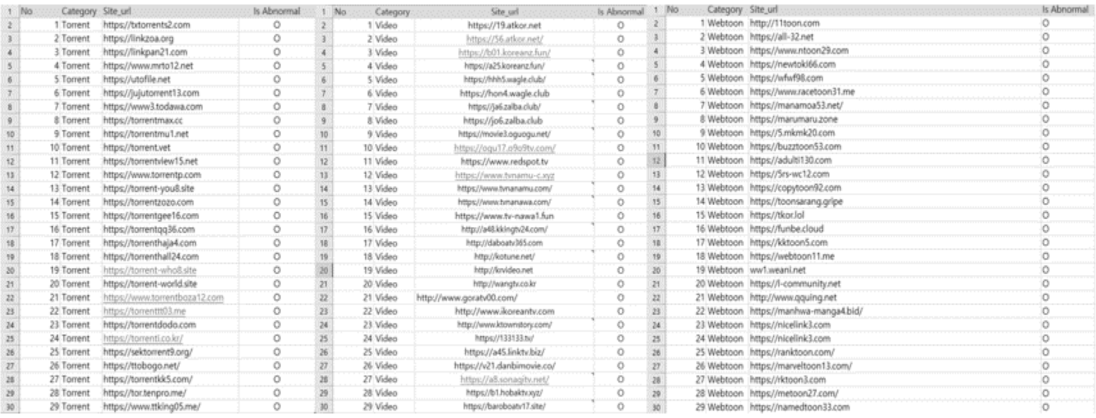
To verify the accuracy of the proposed technique, we proceeded with a 1:1 ratio of piracy sites to the legitimate sites.

The piracy site dataset was created by extracting the URL of the piracy site banner by using the feature that the piracy site exists in the form of a banner in the link collection site. During the extraction process, the Selenium Python module and web driver were used to crawl, and the data set was extracted by saving the piracy site as a.csv file. The link to the piracy site listed after the 'href' tag in the link collection site was extracted and saved as a csv file. **Table 11** presents the pseudo code that was used to extract the piracy site from the link collection site.

Table 11. Pseudo code for extracting piracy site in the link collection site

```
Function for extracting piracy site in link collection site:
str_html ← Link collection site url using Selenium
piracy_url ← Piracy site within link collection site url
data.csv ← csv file where piracy site is stored
IF there is a regular expression href tag that matches str_html,
  Save piracy_url in data.csv
RETURN 0
```

Piracy sites in the extracted data set consisted of 157 sites, including 34 torrent sites, 71 video streaming sites, and 52 webtoon sites. The experiment was also conducted with a total of 314 datasets by adding 157 legitimate sites unrelated to copyright infringement. Furthermore, **Fig. 8** illustrates the piracy site and data set of the legitimate site used in the proposed technique, and **Table 12** lists the number of legitimate and piracy sites used in the proposed technique.



No	Category	Site_url	Is Abnormal
1	Torrent	https://txtorrents2.com	O
2	Torrent	https://linkzoa.org	O
3	Torrent	https://linkpan21.com	O
4	Torrent	https://www.mrto12.net	O
5	Torrent	https://utofile.net	O
6	Torrent	https://jujutorrent13.com	O
7	Torrent	https://www3.todawa.com	O
8	Torrent	https://torrentmax.cc	O
9	Torrent	https://torrentmu1.net	O
10	Torrent	https://torrent.vet	O
11	Torrent	https://torrentview15.net	O
12	Torrent	https://www.torrentp.com	O
13	Torrent	https://torrent-you8.site	O
14	Torrent	https://torrentzozo.com	O
15	Torrent	https://torrentgee16.com	O
16	Torrent	https://torrentqq36.com	O
17	Torrent	https://torrenthaja4.com	O
18	Torrent	https://torrenthall24.com	O
19	Torrent	https://torrent-who8.site	O
20	Torrent	https://torrent-world.site	O
21	Torrent	https://www.torrentboza12.com	O
22	Torrent	https://torrenttt03.me	O
23	Torrent	https://torrentdodo.com	O
24	Torrent	https://torrenti.co.kr/	O
25	Torrent	https://sektorrent9.org/	O
26	Torrent	https://ttobogo.net/	O
27	Torrent	https://torrentkk5.com/	O
28	Torrent	https://tor.tenpro.me/	O
29	Torrent	https://www.ttking05.me/	O

No	Category	Site_url	Is Abnormal
1	Video	https://19.atkor.net	O
2	Video	https://56.atkor.net/	O
3	Video	https://b01.koreanz.fun/	O
4	Video	https://a25.koreanz.fun/	O
5	Video	https://hhh5.wagle.club/	O
6	Video	https://hon4.wagle.club	O
7	Video	https://ja6.zalba.club/	O
8	Video	https://jo6.zalba.club	O
9	Video	https://movie3.oguogu.net/	O
10	Video	https://oqu17.o9o9tv.com/	O
11	Video	https://www.redspot.tv	O
12	Video	https://www.tvnamu-c.xyz	O
13	Video	https://www.tvnanamu.com/	O
14	Video	https://www.tvnanawa.com/	O
15	Video	https://www.tv-nawa1.fun	O
16	Video	http://a48.kkingtv24.com/	O
17	Video	http://daboatv365.com	O
18	Video	http://kotune.net/	O
19	Video	http://krvideo.net	O
20	Video	http://wangtv.co.kr	O
21	Video	http://www.goratv00.com/	O
22	Video	http://www.ikoreantv.com	O
23	Video	http://www.ktownstory.com/	O
24	Video	https://133133.tv/	O
25	Video	https://a45.linktv.biz/	O
26	Video	https://v21.danbimovie.co/	O
27	Video	https://a8.sonagitv.net/	O
28	Video	https://b1.hobaktv.xyz/	O
29	Video	https://baroboatv17.site/	O

No	Category	Site_url	Is Abnormal
1	Webtoon	http://11toon.com	O
2	Webtoon	https://all-32.net	O
3	Webtoon	https://www.ntoon29.com	O
4	Webtoon	https://newtoki66.com	O
5	Webtoon	https://wfwf98.com	O
6	Webtoon	https://www.racetoon31.me	O
7	Webtoon	https://manamoa53.net/	O
8	Webtoon	https://marumaru.zone	O
9	Webtoon	https://5.mkmk20.com	O
10	Webtoon	https://buzztoon53.com	O
11	Webtoon	https://adulti130.com	O
12	Webtoon	https://5rs-wc12.com	O
13	Webtoon	https://copytoon92.com	O
14	Webtoon	https://toonsarang.gripe	O
15	Webtoon	https://tkor.lol	O
16	Webtoon	https://funbe.cloud	O
17	Webtoon	https://kktoon5.com	O
18	Webtoon	https://webtoon11.me	O
19	Webtoon	ww1.weani.net	O
20	Webtoon	https://l-community.net	O
21	Webtoon	http://www.qquing.net	O
22	Webtoon	https://manhwa-manga4.bid/	O
23	Webtoon	https://nicelink3.com	O
24	Webtoon	https://nicelink3.com	O
25	Webtoon	https://ranktoon.com/	O
26	Webtoon	https://marveltoon13.com/	O
27	Webtoon	https://rktoon3.com	O
28	Webtoon	https://metoon27.com/	O
29	Webtoon	https://namedtoon33.com	O

Fig. 8. Example of a piracy site data set

Table 12. Number of piracy sites and legitimate sites

Category	Piracy Sites	Legitimate Sites	Total
Torrent	34	34	68
Video Streaming	71	71	142
Webtoon	52	52	104
Total	157	157	314

4.3 Analysis of Accuracy

We used the F1 score to verify the accuracy of the proposed technique, and we derived the confusion matrix for each site to determine the F1 score.

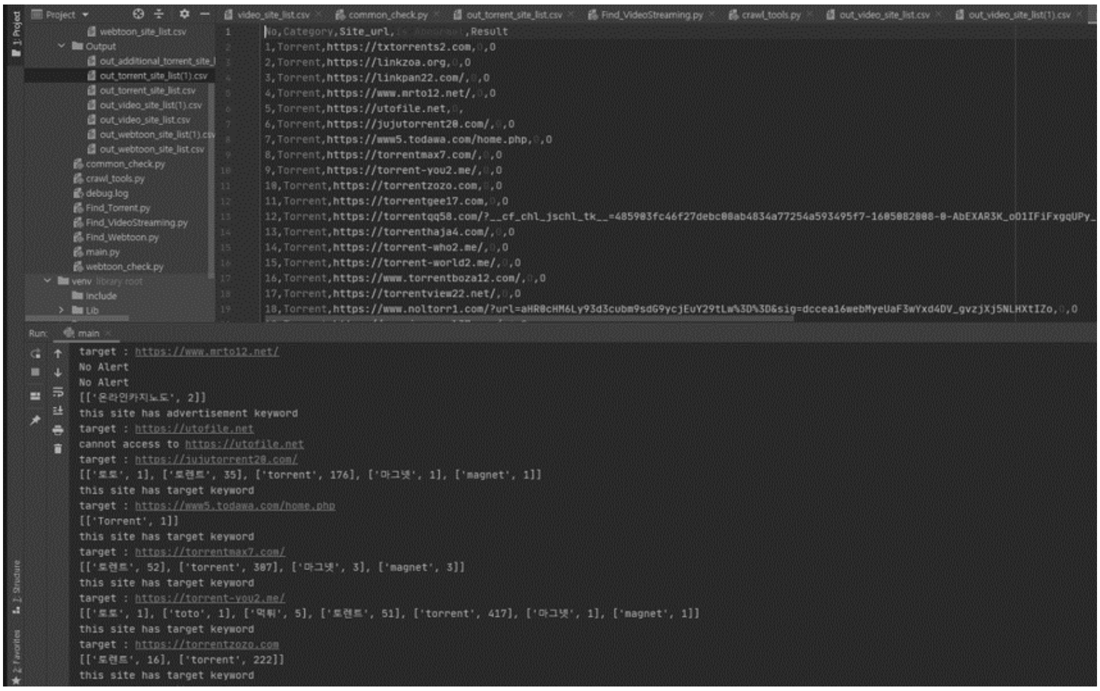
4.3.1 Analysis of Torrent Sites

The proposed intelligent detection technique was applied to 68 sites with a 1:1 ratio of 34 torrent sites and 34 legitimate sites. In the experiment, 33 cases were judged as suspected piracy sites, and one case was judged as a legitimate site. There are two cases in which legitimate sites are suspected of piracy and 32 cases in which legitimate sites are judged as legitimate sites. The precision, recall, and F1 score determined using the aforementioned values are given in (4), (5), and (6). As calculated in (4), the precision is 94.2% and corresponds to the ratio of actual piracy sites to those that are determined as suspected of piracy. As calculated in (5), the recall is 97.0% and corresponds to the ratio of actual piracy sites to those that are determined as piracy sites. The F1 score calculated in (6) determines the accuracy of the proposed technique for the torrent site, which corresponds to 95.5%. Fig. 9 illustrates the torrent site's experimental process. Table 13 presents the confusion matrix for the torrent site.

$$Precision = \frac{TP}{TP+FP} = \frac{33}{33+2} = 0.942 \tag{4}$$

$$Recall = \frac{TP}{TP+FN} = \frac{33}{33+1} = 0.970 \tag{5}$$

$$F1\ score = 2\ \times \frac{Precision\ \times Recall}{Precision\ + Recall} = 2\ \times \frac{0.942 \times 0.970}{0.942 + 0.970} = 0.955 \tag{6}$$



```
No,Category,Site_url,Is_Abnormal,Result
1,Torrent,https://txtorrents2.com,0,0
2,Torrent,https://linkzoa.org,0,0
3,Torrent,https://linkpan22.com/,0,0
4,Torrent,https://www.mrto12.net/,0,0
5,Torrent,https://utofile.net,0,
6,Torrent,https://jujutorrent20.com/,0,0
7,Torrent,https://www5.todawa.com/home.php,0,0
8,Torrent,https://torrentmax7.com/,0,0
9,Torrent,https://torrent-you2.me/,0,0
10,Torrent,https://torrentzozo.com,0,0
11,Torrent,https://torrentgee17.com,0,0
12,Torrent,https://torrentqq58.com/?__cf_chl_jschl_tk__=485983fc46f27debc08ab4834a77254a593495f7-1605082808-0-AbEXAR3K_oO1IFiFxgqUPy_
13,Torrent,https://torrenthaja4.com/,0,0
14,Torrent,https://torrent-who2.me/,0,0
15,Torrent,https://torrent-world2.me/,0,0
16,Torrent,https://www.torrentboza12.com/,0,0
17,Torrent,https://torrentview22.net/,0,0
18,Torrent,https://www.noltorr1.com/?url=aHR0cHM6Ly93d3cubm9sdG9ycjEuY29tLw%3D%3D&sig=dccea16webMyeUaF3wYxd4DV_gvzjXj5NLHXtIZo,0,0
```

```
target : https://www.mrto12.net/
No Alert
No Alert
[['온라인카지노도', 2]]
this site has advertisement keyword
target : https://utofile.net
cannot access to https://utofile.net
target : https://jujutorrent20.com/
[['토토', 1], ['토렌트', 35], ['torrent', 176], ['마그넷', 1], ['magnet', 1]]
this site has target keyword
target : https://www5.todawa.com/home.php
[['Torrent', 1]]
this site has target keyword
target : https://torrentmax7.com/
[['토렌트', 52], ['torrent', 307], ['마그넷', 3], ['magnet', 3]]
this site has target keyword
target : https://torrent-you2.me/
[['토토', 1], ['toto', 1], ['먹튀', 5], ['토렌트', 51], ['torrent', 417], ['마그넷', 1], ['magnet', 1]]
this site has target keyword
target : https://torrentzozo.com
[['토렌트', 16], ['torrent', 222]]
this site has target keyword
```

Fig. 9. Experimental process of the torrent site

Table 13. Confusion matrix of the torrent site

Torrent		True Condition	
		Positive	Negative
Predicted Condition	Positive	33	2
	Negative	1	32

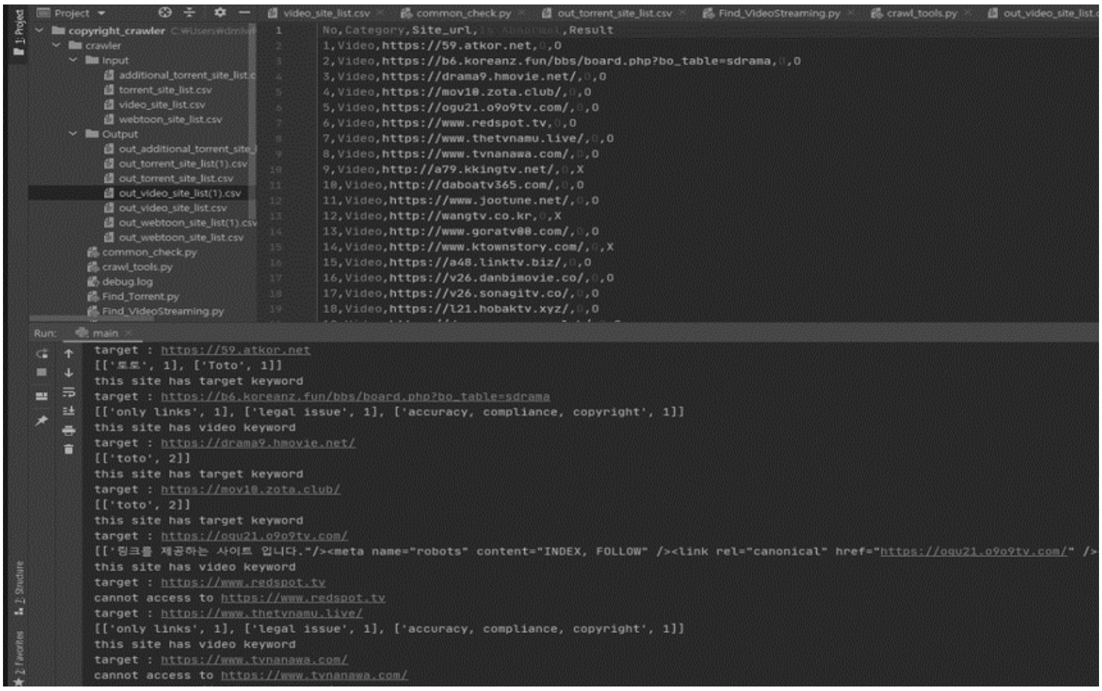
4.3.2 Analysis of Video Streaming Sites

The intelligent detection technique is applied to 142 sites in a 1:1 ratio of 71 video streaming sites and 71 legitimate sites, and 64 cases were judged as suspected piracy sites, and 7 cases were judged as legitimate sites. There are 5 cases in which legitimate sites are suspected of piracy and 66 cases in which legitimate sites are judged as legitimate sites. The precision, recall, and F1 score derived using the values are as shown in (7), (8), and (9). As calculated in (7), the precision is 92.7% and corresponds to the ratio of actual piracy sites relative to those determined to be suspected of piracy. As calculated in (8), the recall is 90.1% and corresponds to the ratio of the actual piracy sites to those determined as piracy sites. The F1 score calculated in (9) indicates the accuracy of the video streaming site's proposed technique and corresponds to 91.3%. **Fig. 10** illustrates the experimental process for the video streaming site. **Table 14** presents the confusion matrix for the video streaming site.

$$Precision = \frac{TP}{TP+FP} = \frac{64}{64+5} = 0.927 \tag{7}$$

$$Recall = \frac{TP}{TP+FN} = \frac{64}{64+7} = 0.901 \tag{8}$$

$$F1\ score = 2\ \times \frac{Precision\ \times Recall}{Precision\ + Recall} =\ 2\ \times \frac{0.927 \times 0.901}{0.927 + 0.901} = 0.913 \tag{9}$$



```
No,Category,Site_url,Is_Abnormal,Result
1,Video,https://59.atkor.net,0,O
2,Video,https://b6.koreanz.fun/bbs/board.php?bo_table=sdrama,0,O
3,Video,https://drama9.hmovie.net/,0,O
4,Video,https://mov10.zota.club/,0,O
5,Video,https://ogu21.o9o9tv.com/,0,O
6,Video,https://www.redspot.tv,0,O
7,Video,https://www.thetvnamu.live/,0,O
8,Video,https://www.tvnanawa.com/,0,O
9,Video,http://a79.kkingtv.net/,0,X
10,Video,http://daboatv365.com/,0,O
11,Video,https://www.jootune.net/,0,O
12,Video,http://wangtv.co.kr,0,X
13,Video,http://www.goratv88.com/,0,O
14,Video,http://www.ktownstory.com/,0,X
15,Video,https://a48.linktv.biz/,0,O
16,Video,https://v26.danbimovie.co/,0,O
17,Video,https://v26.sonagitv.co/,0,O
18,Video,https://l21.hobaktv.xyz/,0,O
```

```
target : https://59.atkor.net
[['토토', 1], ['Toto', 1]]
this site has target keyword
target : https://b6.koreanz.fun/bbs/board.php?bo_table=sdrama
[['only links', 1], ['legal issue', 1], ['accuracy, compliance, copyright', 1]]
this site has video keyword
target : https://drama9.hmovie.net/
[['toto', 2]]
this site has target keyword
target : https://mov10.zota.club/
[['toto', 2]]
this site has target keyword
target : https://ogu21.o9o9tv.com/
[['링크를 제공하는 사이트 입니다."/><meta name="robots" content="INDEX, FOLLOW" /><link rel="canonical" href="https://ogu21.o9o9tv.com/" />
this site has video keyword
target : https://www.redspot.tv
cannot access to https://www.redspot.tv
target : https://www.thetvnamu.live/
[['only links', 1], ['legal issue', 1], ['accuracy, compliance, copyright', 1]]
this site has video keyword
target : https://www.tvnanawa.com/
cannot access to https://www.tvnanawa.com/
```

Fig. 10. Experimental process of the video streaming site

Table 14. Confusion matrix of the video streaming site

Video Streaming		True Condition	
		Positive	Negative
Predicted Condition	Positive	64	5
	Negative	7	66

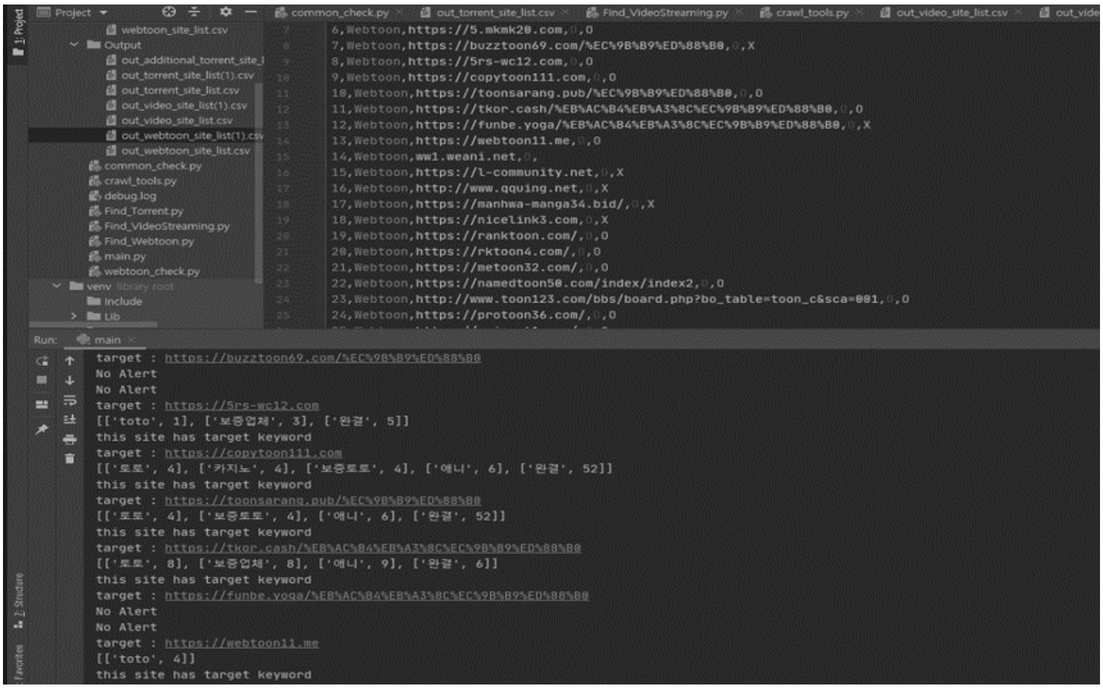
4.3.3 Analysis of Webtoon Site

The intelligent detection technique was applied to 104 sites in a 1:1 ratio of 52 webtoon sites and 52 legitimate sites, and 47 cases were judged as suspected piracy sites, and 5 cases were judged as legitimate sites. Legitimate sites are suspected of piracy in 2 cases, and legitimate sites are judged as legitimate sites in 50 cases. The precision, recall, and F1 score derived using the values are given in (10), (11), and (12). As calculated in (10), the precision is 95.9% and corresponds to the ratio of actual piracy sites relative to those determined to be suspected of piracy. As calculated in (11), the recall is 90.3% and corresponds to the ratio of the actual piracy sites judged to be piracy sites. The F1 score calculated in (12) indicates the accuracy of the proposed technique for the webtoon site and corresponds to 93.0%. **Fig. 11** illustrates the experimental process for the webtoon site. **Table 15** presents the confusion matrix for the webtoon site.

$$Precision = \frac{TP}{TP+FP} = \frac{47}{47+2} = 0.959 \tag{10}$$

$$Recall = \frac{TP}{TP+FN} = \frac{47}{47+5} = 0.903 \tag{11}$$

$$F1\ score = 2 \times \frac{Precision \times Recall}{Precision + Recall} = 2 \times \frac{0.959 \times 0.903}{0.959 + 0.903} = 0.930 \tag{12}$$



```
6,Webtoon,https://5.mkmk20.com,0,0
7,Webtoon,https://buzztoon69.com/%EC%9B%B9%ED%88%B0,0,X
8,Webtoon,https://5rs-wc12.com,0,0
9,Webtoon,https://copytoon111.com,0,0
10,Webtoon,https://toonsarang.pub/%EC%9B%B9%ED%88%B0,0,0
11,Webtoon,https://tkor.cash/%EB%AC%B4%EB%A3%8C%EC%9B%B9%ED%88%B0,0,0
12,Webtoon,https://funbe.yoga/%EB%AC%B4%EB%A3%8C%EC%9B%B9%ED%88%B0,0,X
13,Webtoon,https://webtoon11.me,0,0
14,Webtoon,ww1.weani.net,0,
15,Webtoon,https://l-community.net,0,X
16,Webtoon,http://www.qquing.net,0,X
17,Webtoon,https://manhwa-manga34.bid/,0,X
18,Webtoon,https://nicelink3.com,0,X
19,Webtoon,https://ranktoon.com/,0,0
20,Webtoon,https://rktoon4.com/,0,0
21,Webtoon,https://metoon32.com/,0,0
22,Webtoon,https://namedtoon50.com/index/index2,0,0
23,Webtoon,http://www.toon123.com/bbs/board.php?bo_table=toon_c&sca=001,0,0
24,Webtoon,https://protoon36.com/,0,0
```

```
target : https://buzztoon69.com/%EC%9B%B9%ED%88%B0
No Alert
No Alert
target : https://5rs-wc12.com
[['toto', 1], ['보증업체', 3], ['완결', 5]]
this site has target keyword
target : https://copytoon111.com
[['토토', 4], ['카지노', 4], ['보증토토', 4], ['애니', 6], ['완결', 52]]
this site has target keyword
target : https://toonsarang.pub/%EC%9B%B9%ED%88%B0
[['토토', 4], ['보증토토', 4], ['애니', 6], ['완결', 52]]
this site has target keyword
target : https://tkor.cash/%EB%AC%B4%EB%A3%8C%EC%9B%B9%ED%88%B0
[['토토', 8], ['보증업체', 8], ['애니', 9], ['완결', 6]]
this site has target keyword
target : https://funbe.yoga/%EB%AC%B4%EB%A3%8C%EC%9B%B9%ED%88%B0
No Alert
No Alert
target : https://webtoon11.me
[['toto', 4]]
this site has target keyword
```

Fig. 11. Experimental process of the webtoon site

Table 15. Confusion matrix of the webtoon site

Webtoon		True Condition	
		Positive	Negative
Predicted Condition	Positive	47	2
	Negative	5	50

4.3.4 Analysis of detection technique results

The F1 score for torrent sites derived from **Table 13** and approximately corresponded to 95%. The precision was approximately 94%, and recall was approximately 97%. Piracy sites were not detected through the intelligent detection technique proposed in the study because infringement keywords and illegal advertisements do not exist in torrent sites and there are cases in the form of blogs.

The F1 score for the video streaming site derived from **Table 14** and indicated an accuracy of approximately 91%. The precision approximately corresponded to 92%, and the recall approximately corresponded to 90%. Piracy sites were not detected through the intelligent detection technique because there were cases in which the service can be used only by logging in to the site and there were no infringement keywords or illegal advertisements. This was because there were cases where they were configured almost identically.

The F1 score for the webtoon site was derived from **Table 15** and indicated an accuracy of approximately 93%. The result approximately corresponded to 96% for precision and approximately 90% for recall. The reason as to why the piracy site was not detected in the intelligent detection technique proposed in the study is that it was operated in the form of a blog rather than a general website, or the access to crawling was prevented by using a CAPTCHA in the form of a picture to distinguish users and bots. Although research on bypassing the picture-type CAPTCHA used in webtoon sites is ongoing, there are limitations and difficulties in bypassing them, and thus the proposed technique could not be verified in these cases [11]. **Table 16** lists the precision, recall, and F1 score for each site derived from **Table 13-15**.

Table 16. F1 score for the proposed technique

	Precision	**Recall**	**F1 score**
Torrent	0.942	0.970	0.955
Video Streaming	0.927	0.901	0.913
Webtoon	0.959	0.903	0.930

5. Conclusion

In this study, the features of piracy sites were analyzed, and an intelligent piracy site detection technique was proposed based on the features of piracy sites. The proposed intelligent piracy site detection technique with high accuracy was applied to a total of 314 sites, with piracy sites and legitimate sites in a 1:1 ratio. The detection results were derived in the form of a confusion matrix for each site, and the F1 score was calculated for accuracy verification.

The F1 score results of the experiment indicated that the application of the intelligent detection technique to torrent sites resulted in an accuracy of 95%, 91% for the video streaming sites, and 93% for the webtoon sites. The results indicate the high accuracy of the proposed technique. Hence, the intelligent piracy site detection technique proposed in the study results in high accuracy and can be used to detect piracy sites. Furthermore, the technique contributes to decreasing the losses for copyright holders due to copyright infringement.

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] 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)

[2] Korea Copyright Protection Agency, “English Version of C STORY 2016,” *KCOPA Report*, Seoul, Korea, Dec. 2016. Article (CrossRef Link)

[3] S. 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)

[4] Department of Communications, "Online Copyright Infringement Research," *A Marketing Report*, Australia, June 2015. Article (CrossRef Link)

[5] Korea Copyright Protection Agency, "2017 Annual Report on Copyright Protection in Korea," *Korea Copyright Protection Agency*, Seoul, Korea, June 2017. Article (CrossRef Link)

[6] M. S. Shin, M. R. Yong, and Y. J. Lee, "Study on Preventing Copyright Infringement through Blocking Advertisements of Illegal Copyrighted Websites," *The Journal of the Korea Contents Association*, vol. 20, no. 7, pp. 331-341, July 2020. Article (CrossRef Link)

[7] M. Wieting and J. Seldeslachts, "Advertising in illegal markets evidence from online gambling in the Netherlands," M.S. thesis, Dept. Economics and English, University of Amsterdam, Amsterdam, Netherlands, Aug. 2019.

[8] J. H. Lee and H. K. Lee, "A Study on Korean Emotion Index Using F1_score," *The Journal of Internet Electronic Commerce Research*, vol. 20, no. 1, pp. 131-145, Feb. 2020. Article (CrossRef Link)

[9] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Region-Based Convolutional Networks for Accurate Object Detection and Segmentation," *2016 IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 1, pp. 142-158, Jan. 2016. Article (CrossRef Link)

[10] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436-444, 2015. Article (CrossRef Link)

[11] S. Sivakorn, I. Polakis, and A. D. Keromytis, "I am Robot: (Deep) Learning to Break Semantic Image CAPTCHAs," in *Proc. of 2016 IEEE European Symposium on Security and Privacy*, pp. 388-403, May 2016. Article (CrossRef Link)

Eui-Jin Kim is a postgraduate student in master's course at Dept. of Computer Engineering in Ajou University, Republic of Korea. He received the B.S. degree from Ajou University, Republic of Korea. His research interests include Copyright protection, Cryptographic protocols, Network security.

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,