Feature Analysis and Detection Techniques for Piracy Sites

Seul-Ki Choi[1] and Jin Kwak[2*]
[1] ISAA Lab., Department of Computer Engineering, Ajou University, Suwon, Republic of Korea
[e-mail: skchoi.isaa@gmail.com]
[2] Department of Cyber Security, Ajou University, Suwon, Republic of Korea
[e-mail: security@ajou.ac.kr]
*Corresponding author: Jin Kwak

Received December 6, 2019; revised January 28, 2020; accepted March 11, 2020; published May 31, 2020

Abstract

In recent years, digital content has become easily accessible because of internet technology. Representative examples of such digital content include various types, such as music, TV, (program, sport, drama etc.) and films. However, there are cases where internet technology is used in illegal ways without the authorization of the copyright holder for digital content. Such actions have a direct impact on copyright owners' profits and further affect the development of the digital culture industry adversely. Therefore, in this study, we analyze features to detect piracy sites that cause copyright infringement. Further, we develop a piracy site detection crawler based on these features and present the analysis of its performance.

Keywords: Copyright, Infringement, Piracy sites

1. Introduction

With the development of IT technology, the user's PC performance and communication technology have advanced rapidly, making it easy to use contents such as TV program, Film, and Music. High-definition video contents can be easily downloaded or provided as a streaming service by using the rapidly developed data communication technology. Music files and software can also be easily carried by individuals, such as content in video formats. Among such contents, there may be a case where a copyright owner of the contents exists. If a copyright owner exists for the content, it must be approved by the copyright owner when the content is to be used.

However, piracy sites that illegally copy and distribute such contents without the permission of the copyright owner are rapidly spreading. In particular, sites that use video technology to provide video and audio content in real time using streaming technology or to exploit network technologies such as distribution using BitTorrent technology are emerging. BitTorrent's original function is based on peer-peer communication, a protocol developed to easily and quickly receive specific data among multiple users. It is a situation that is being exploited to generate copyright infringement by using the features of the communication technology.

In addition, there are many cases of illegally recording and recording a service provided in a streaming form to illegally copy video and sound files without the consent of the copyright owner. In recent years, copyright infringement on webtoons, which has attracted attention from many people and has increased in size of the market, also occurs. Illegal sites, which illegally copy and post paid webtoon contents for free, are also increasing.

These cases of copyright infringement act as a deterrent to the development for the digital content culture industry, such as hindering copyright owners' profit activities. Researches are underway to detect such piracy sites. As a result of the studies, there have been studies that detect the piracy site reopening by analyzing the source code form of the piracy sites. In addition, there has been a study to detect the case where the site is reopened by analyzing the features of the piracy sites visually.

The crackdown on piracy sites continues, but the speed of their creation is much faster than the crackdown on piracy sites. In other words, in order to solve this situation, a technique for finding them at the speed at which piracy sites are generated and spread is required. Therefore, in this study, we analyze the characteristics of piracy sites that cause copyright infringement. We then use the analysis results to develop the crawler for detection of piracy sites and measure their performance.

This manuscruipt is organised as follows. In section 2, we review related research on the types of digital content and trends on piracy sites. In section 3, we analyze features to detect piracy sites. In section 4, we analyze the experimental results for piracy site detection, and finally, we present the conclusions in section 5.

2. Related Work

2.1 Types of Digital Contents

Digital content is a form that can be created or used using IT technology. Intellectual Property Office classifies major types of digital content as presented in the **Table 1**.

Table 2. Representative Types of Digital Content

Category	Definition
Music	Music tracks or albums (excluding online radio stations)
Films	Films (full length)
TV programs	TV programs (including live sports)
Computer software	Computer software (excluding mobile phone apps, and patches/upgrades to software already owned)
Books	e-books
Video games	Video games (excluding patches and upgrades)

In addition, there are three types of behaviors for users to use digital content [1]:

- Streamed or accessed: A method of viewing, listening, or playing digital content directly through internet technology without owning the original digital content directly on the user's device.
- Downloaded: A method of directly downloading the original digital content using internet technology and owning it directly on the user's device.
- Shared: A method of uploading original digital content by using internet technology to download, or to stream by publishing digital content.

Table 3 shows the distribution volume by digital content type.

Table 4. Volume of digital content consumed [1]

Category		Total	Physical format	Digital format
Music	Volume	507m	54m	453m
	% of total	57%	11%	89%
Films	Volume	96m	23m	73m
	% of total	11%	24%	76%
TV programs	Volume	160m	8m	152m
	% of total	18%	5%	95%

Computer software	Volume	19m	4m	15m
	% of total	2%	24%	76%
Books	Volume	75m	57m	18m
	% of total	8%	76%	24%
Video games	Volume	40m	18m	22m
	% of total	4%	45%	55%
Total	**Volume**	**896m**	**163m**	**733m**
	% of total		**18%**	**82%**

In this case, the physical format means that the digital content is included in a physical media such as CD and DVD. It can be seen that most digital content is distributed in digital format, and not in a physical form.

In particular, it can be seen that most of music, TV programs, and films, which occupy 86% of digital content, are distributed in digital format.

2.2 Trends of Piracy Sites

We may use copyrighted digital content through lawful use. However, the European Union Intellectual Property Office (EUIPO) categorizes and defines the types of digital content that are used in an illegal way, ignoring the copyright [2, 3].

Table 5. Type of copyright infringement [2, 3]

Type of infringement	Description
Physical infringement	A method of illegally copying digital content contained on physical media (eg. CDs and DVDs)
Internet infringement	A method of using internet technology to copy and distribute digital content illegally to other internet users
Signal theft	Receiving illegally supplying cable TV and radio signals without authorization
Broadcast piracy	Broadcasting programs that are either legal or illegal, without the permission of the copyright owner
Unauthorized public performance	Activities to show the digital content without the permission of the copyright owner

In addition, EUIPO classifies copyright infringement into four types based on internet technology[3]:

- Streaming: It refers to a method of providing digital content directly to the user through online streaming, and it is not authorized by the copyright holder. It provides a function to search various kinds of contents directly. It can be classified into a case where the infringing content is directly hosted and a link to an external host is provided.

- Download: It refers to a method of directly downloading and providing digital content to a user's device, and it is not authorized by a copyright holder. Like streaming methods, these methods often provide the ability to directly search for various types of content. They often own the allegedly infringing content and host the content directly to the user.
- Stream ripping: It refers to a method of extracting audio and video streaming content in a downloadable form. When the user enters the URL of the streaming content, the content is converted for download in audio and video format.
- Torrent: Torrent is a peer-to-peer download process that does not download content directly from a specific server, but instead provides a content file from another torrent user who has the same content. The torrent site therefore provides a seed file or magnet link to use the torrent protocol.

The figure below shows the usage rate of devices used for copyright infringement by contents in 28 countries (EU28).

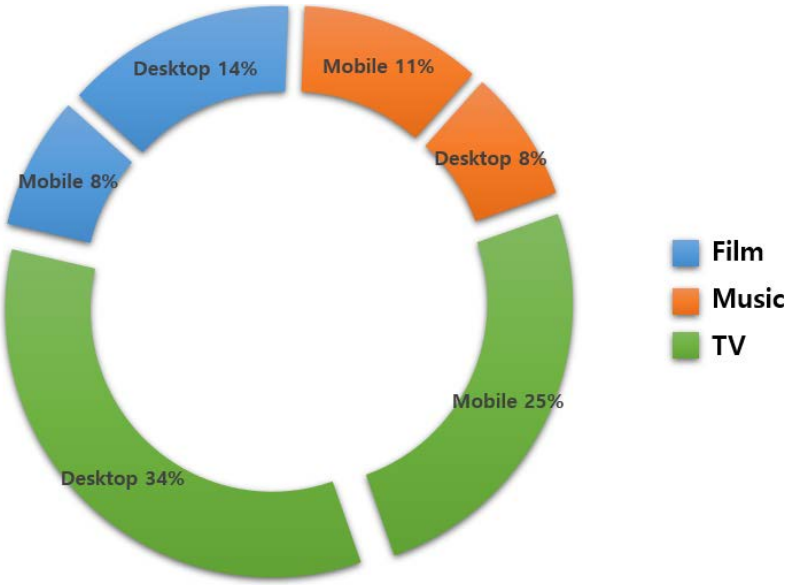

Fig. 1. Device usage for copyright infringement by content type [2]

Overall, desktop (56%) and mobile (44%) have similar usage rates, and music content shows that copyright infringement using mobile devices is higher.

The figure below shows trends in the types of piracy sites (streaming, download, torrent, and ripper) that internet users access monthly [2].

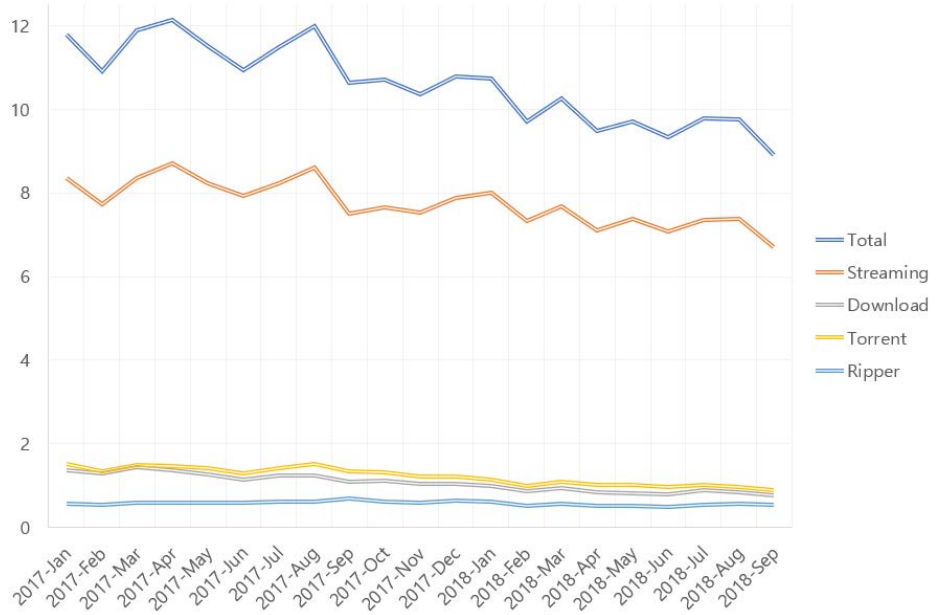

Fig. 2. Piracy trends by access method. Average accesses per internet user per month, EU28 [2]

From **Fig. 3**, it can be observed that piracy sites that infringe copyrights are generally preferred.

The figure below shows the piracy trends in film content [2].

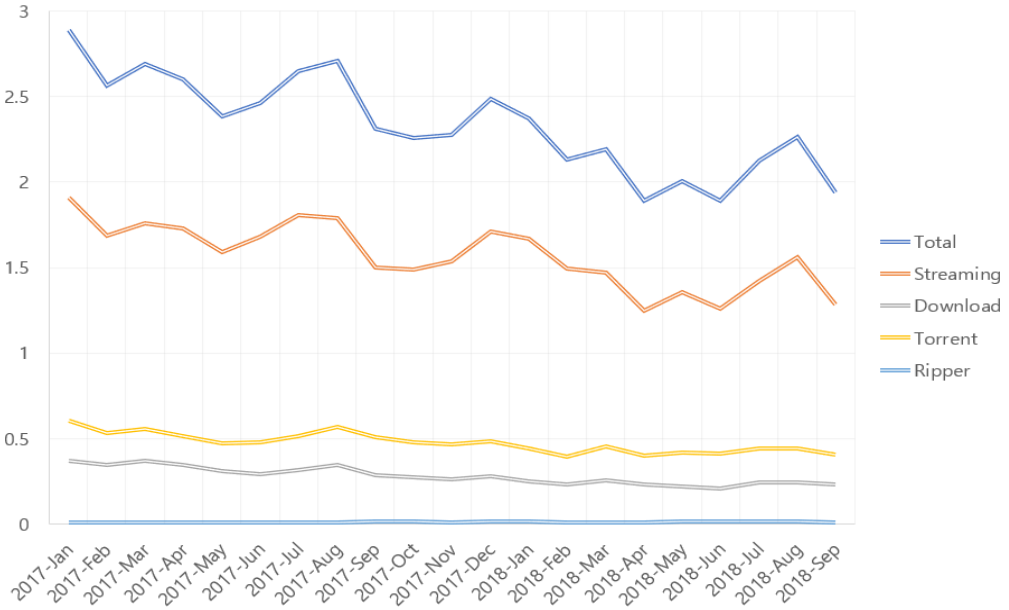

Fig. 4. Film piracy trends. Average accesses per internet user per month, EU28 [2]

Copyright infringement on film contents shows that the piracy site in the form of streaming is preferred.

The figure below shows piracy trends in TV content [2].

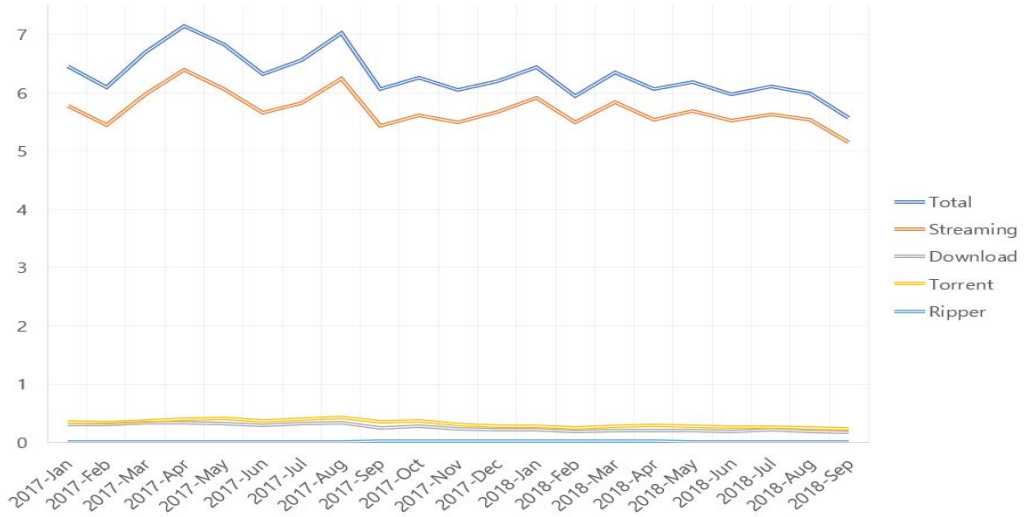

Fig. 5. TV piracy trends. Average accesses per internet user per month, EU28 [2]

Copyright infringement on TV content, similar to film content, is more common on piracy sites in the form of streaming. In particular, the streaming use rate is higher than other methods.

Finally, the figure below shows the piracy trends for music content [2].

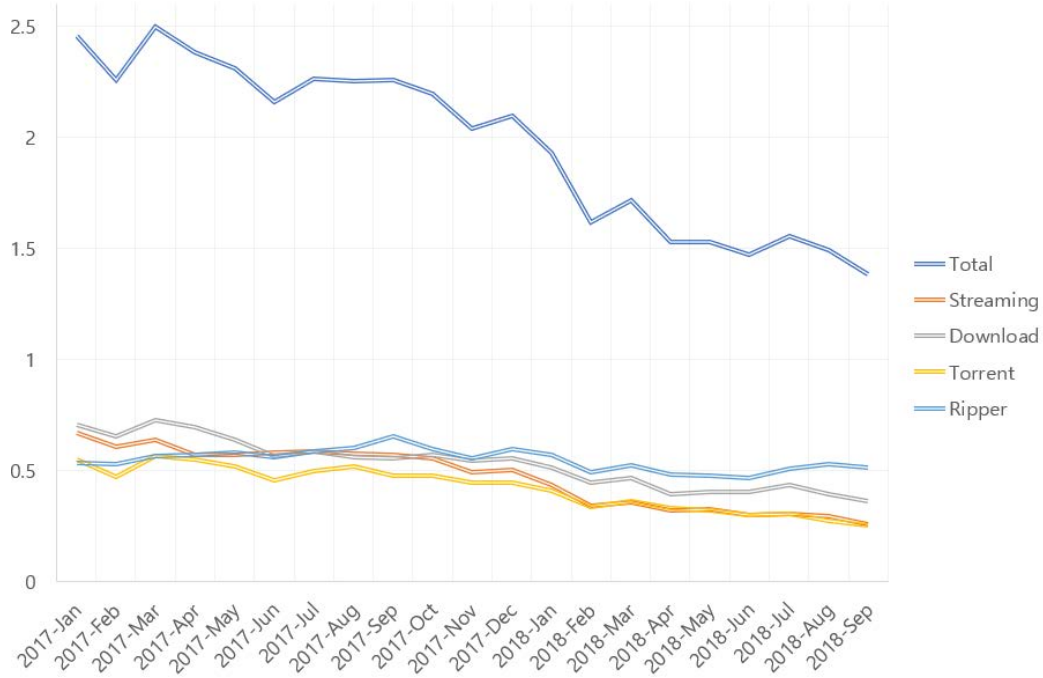

Fig. 6. Music piracy trends. Average accesses per internet user per month, EU28 [2]

In the case of music contents, it can be seen that four piracy methods are used evenly. The table below shows the piracy access types for each type of content.

Table 6. Total piracy by access type [2]

Category	Piracy Method by access type			
	Streaming	**Torrent**	**Download**	**Ripper**
Film	67.9%	20.3%	11.2%	0.6%
TV	92.0%	4.4%	3.3%	0.3%
Music	20.7%	20.3%	26.9%	32.1%
Total	**75.3%**	**10.4%**	**8.8%**	**5.5%**

As a result of analyzing the trend of copyright infringement, we can see that the piracy site using the streaming method is the most used, followed by torrent, download, and ripper methods.

3. Feature Analysis of Piracy Sites

3.1 Torrent Sites

Torrent sites are those that provide torrent files or magnet addresses that are essential for sharing data through the BitTorrent client. These torrent sites provide the ability to illegally download various kinds of works such as music, films, TV content (such as programmes and drama), publications, games, and software.

The figure below shows the main screens and an example of the torrent sites in Korea.

Fig. 7. Example of torrent site (Korean)

Most of the content posted on the site consists of titles related to the works, and various menus are provided for easy searching by filtering by the type of content.The figure below shows an example of the features found mainly in torrent sites.

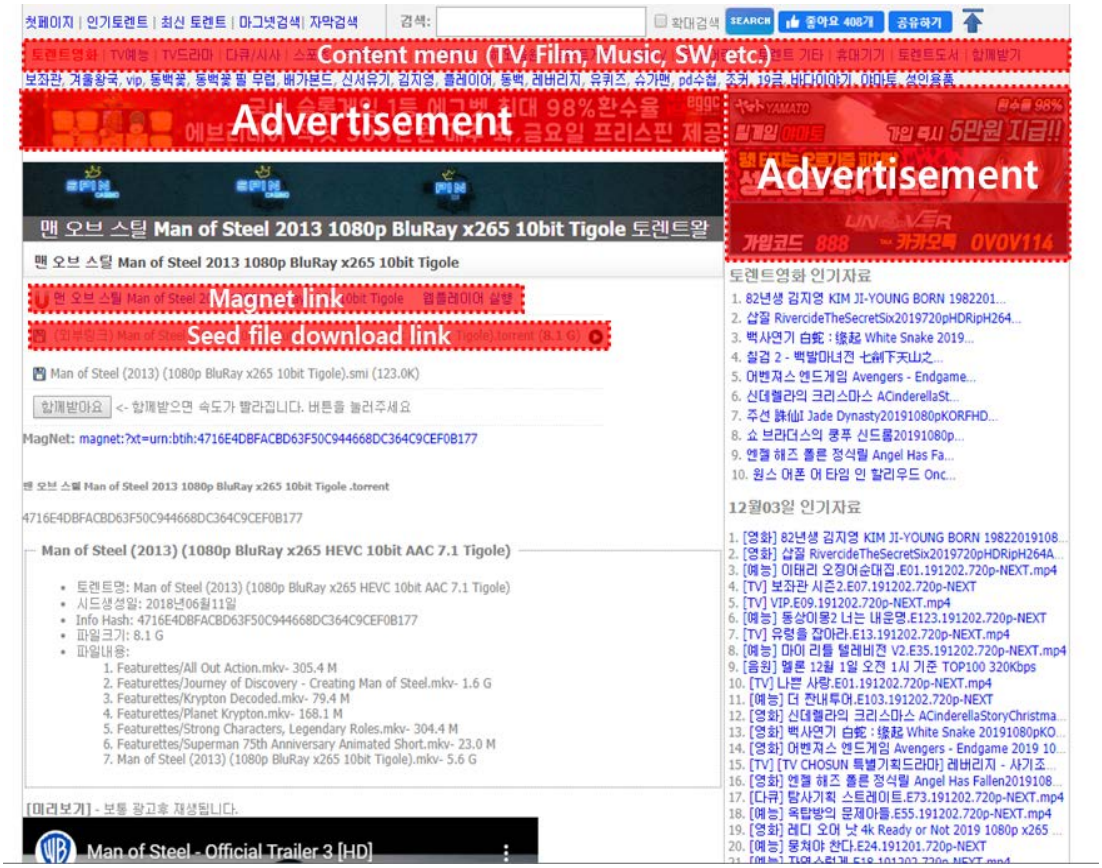

Fig. 8. Example features of torrent sites

Features for detecting torrent sites are as follows:

- Content menu and list: Various contents with violated copyrights are included in the torrent sites. It includes a list of digital content, including recently aired TV shows, latest films, music, games, and software.

- Advertisement: Most of the illegally operated piracy sites, as well as the torrent sites, contain advertising. These advertisements are used as the main source of revenue for piracy sites. They commonly contain adult content and illegal gambling advertisements.

- Magnet link: It provides a magnet link to use the torrent client. Users can download content using magnet links and torrent client programs.

- Download link (for seed file): It provides a seed file for using the torrent client. The site may have its own seed file, or indirectly provide a link to download the seed file through a file sharing site, such as filetender.

3.2 Video Streaming Sites

The video streaming site provides video content such as film, TV programs and sports. **Fig. 9** and **Fig. 10** show examples of video streaming sites.

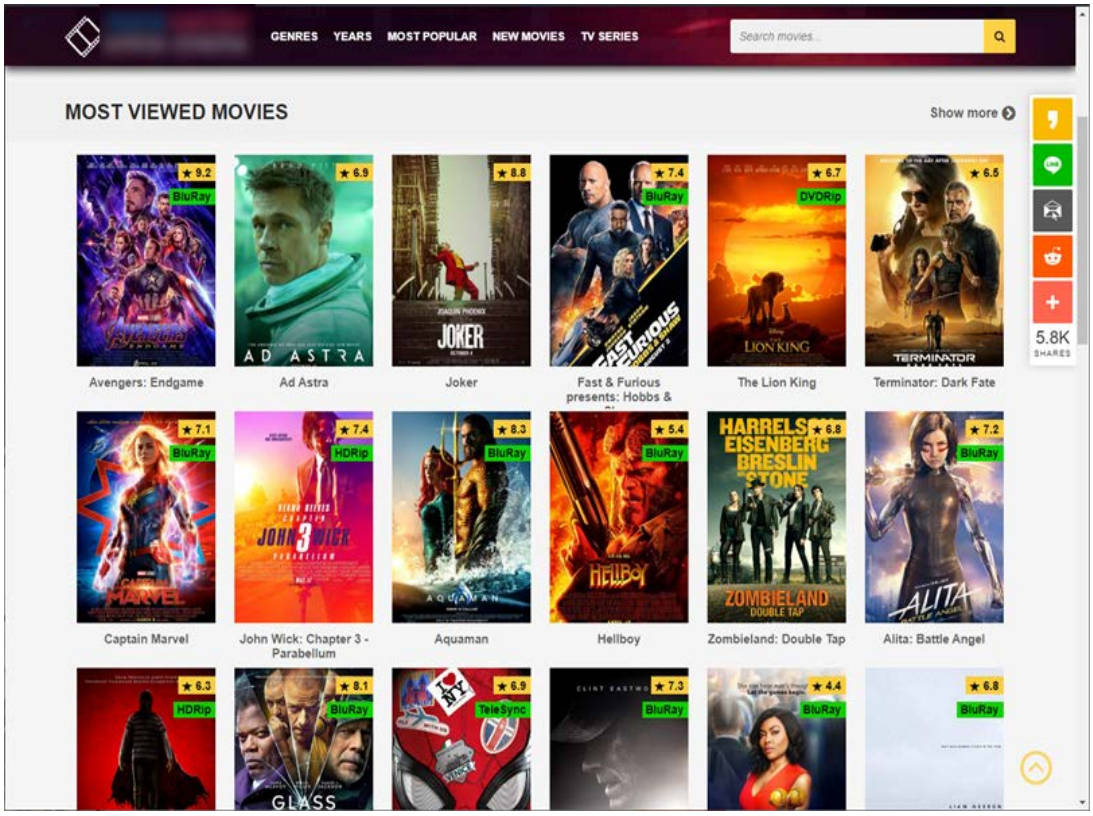

Fig. 11. Example of video streaming site

Features for detecting such video streaming sites are as follows:

- Content list: Within the video streaming site, as shown in **Fig. 12**, you can easily check the list related to the video content such as film or TV show.

- Advertisement: Advertisements are included in various parts of the video streaming site, and when a user selects the video content that they want to watch, there is a case of directly connecting to the advertisement site.

- Streaming server list: There is also a case where a link to another streaming server is provided without directly streaming video content. Even if one streaming server is blocked, a list of two or more streaming servers can be provided for the user to select the desired

quality. Certain sites do not stream directly and only provide links to other streaming sites, thereby attempting to avoid liability for copyright infringement implying that such cases are not classified as copyright infrinement.

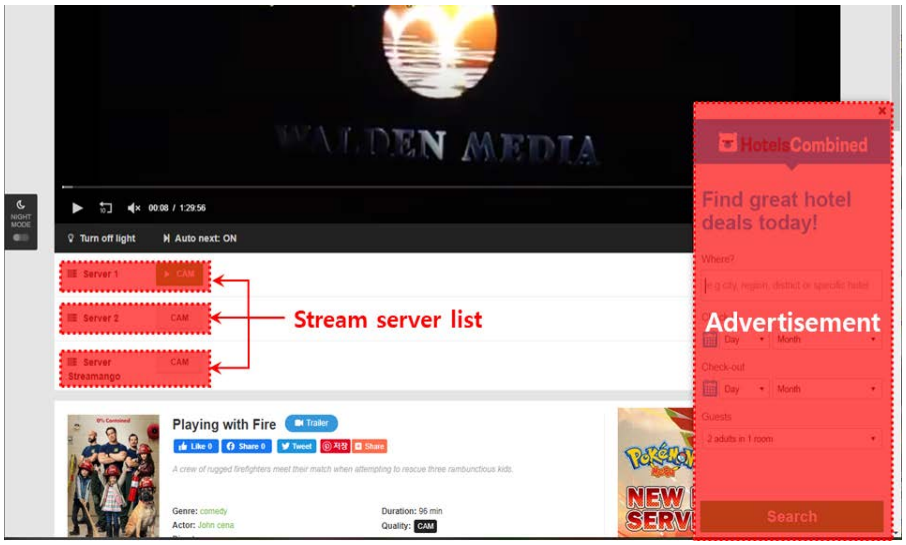

Fig. 13. Example features of video streaming site

3.3 Webtoon Posting Sites

Webtoon is a compound word of ‘web’ of the world wide web and ‘toon’ of cartoon. These webtoons are comics that are serialized on portal sites or on their own. The famous works of the webtoon are translated into foreign languages and serialized on foreign comic sites, and films and TV dramas are produced based on the webtoons. There are sites that illegally copy webtoons provided in the form of images and publish them illegally without the permission of the copyright holder. The figure below shows an example of the webtoon illegal posting site.

Fig. 14. Example of webtoon posting site (Korean)

Features for detecting sites that illegally publish webtoons are as follows:

- Content list: It contains information about the titles of webtoons that are in serial or complete in serialization on the webtoon publishing site.

- Advertisement: Like other piracy sites, advertisements are included through the whole site.

- Proper noun: Webtoons, unlike other content, are visual (letters and pictures) and have static characteristics. For example, in the case of a torrent site, the download is completed through a torrent client before it is available. You cannot check the contents until the download is completed. In addition, in the case of various streaming sites, the content is delivered to the user as the streaming player time passes. However, in the case of webtoons, every content is posted visually within the site. Therefore, proper nouns (such as names of characters) used only in the webtoon are included and can be easily identified.

- Watermark: There are cases of advertising and promotion within webtoons in the form of a watermark.

Fig. 15. Example features of webtoon posting sites

4. Experiments on Detecting Piracy Sites

4.1 Detection Process

Fig. 12 below shows the site crawling and analysis process for the experiments proposed in this paper. The site crawling and piracy site detection process has three phases, except for the visit site phase.

The first phase is the suspicion step, which collects the ads included in the visited sites and analyzes the characteristics of the ads. If the advertisement on the site has a lot of illegal characters, it is a step that is suspected as a piracy site. The second phase is to identify the type of piracy site by analyzing the content menus included in the site. Finally, the third phase is to analyze whether the piracy site features are included according to the type of piracy site identified in the second phase.

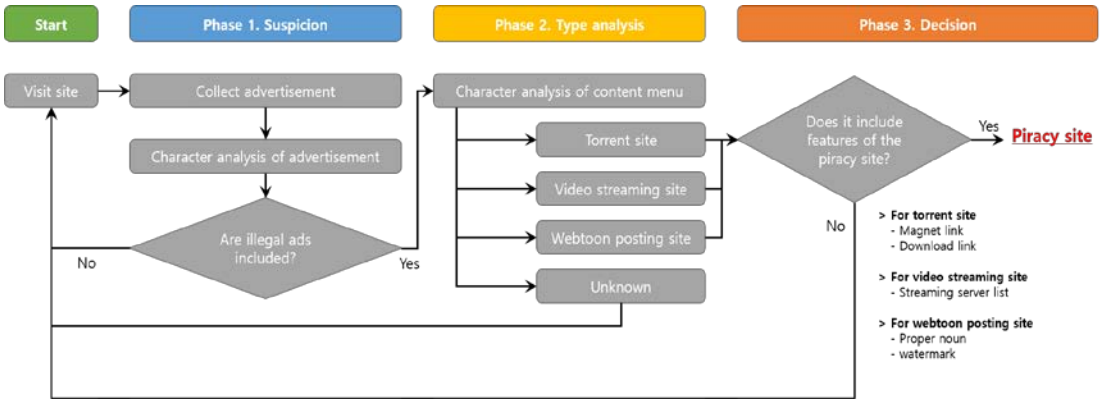

Fig. 16. Overall process for dectection piracy site

4.2 Design of Experiments

A crawler that extracts features for torrent sites, video streaming sites, and webtoon publishing sites is implemented using the Python development language. Also, to measure the performance of the crawler, we mixed the list of piracy sites by type and the list of normal sites. In particular, in the case of video streaming site and webtoon publishing site, sites that are legally operated under the permission of the copyright holder were also used as the test targets for the performance measurement of the crawler. The number of illegal and legal sites used in the experiment is listed in the **Table 7**.

Table 8. Number of illegal and legal sites used in the experiment

Category	Number of illegal sites	Number of legal sites
Torrent sites	23	23
Video streaming sites	14	14
Webtoon posting sites	20	20
Total	**57**	**57**

4.3 Results of Experiments

Tables 6–8 below show the confusion matrix values for piracy site detection experiments based on the characteristics of each piracy site analyzed in this study.

Table 9. Confusion matrix of torrent sites detection

		True condition	
		Condition positive	**Condition negative**
Predicted condition	**Predicted condition positive**	23	2
	Predicted condition negative	0	21

Table 10. Confusion matrix of video streaming sites detection

		True condition	
		Condition positive	**Condition negative**
Predicted condition	**Predicted condition positive**	13	1
	Predicted condition negative	1	13

Table 11. Confusion matrix of webtoon posting sites detection

		True condition	
		Condition positive	**Condition negative**
Predicted condition	**Predicted condition positive**	20	4
	Predicted condition negative	0	16

Table 12 below shows the overall results of this experiment based on the data in **Tables 6–8**. Through the results, we analyze the performance of this study.

Table 13. Detection performance measurement data of experiment result

Performance indicator	Formula	Torrent	Video streaming	Webtoon posting
Accuracy	(TP+TN) / (TP+TN+FP+FN)	0.957	0.929	0.900
Precision	TP / (TP+FP)	0.920	0.929	0.833
Recall	TP / (TP+FN)	1.000	0.929	1.000
False Alarm	FP / (FP+TN)	0.087	0.071	0.200
F1 Score	2 * (Precision*Recall) / (Precision + Recall)	0.958	0.929	0.909

In case of torrent site detection, the obtaind accuracy is approximately 95% and precision is 92%. In particular, as the recall value is 100%, it can be confirmed that no torrent sites can be detected. The false alarm also shows 8%.

In case of video streaming site detection, accuracy is 92.9% and precision is 92.9%. In addition, the detection of the video streaming sites, including recall values of 92.9% and false alarms of 7%, also indicates an appropriate level.

In case of webtoon posting sites, the accuracy is 90% and the precision is 83%, which is slightly lower than the detection rate for torrent and video streaming sites. However, it can be seen that there is no detection of webtoon posting sites through 100% recall value. On the other hand, it can be seen that the false alarm for detecting a legal site as an illegal webtoon posting site is 20%.

5. Conclusion and Future Research

In this study, we analyzed the trends of piracy sites that cause copyright infringement and analyzed the characteristics of those sites. Based on the analyzed features, we developed a detection crawler for torrent, video streaming, and webtoon posting sites that cause copyright infringement. As a result, we found that the total performance of the crawler in detecting of the torrent, video streaming, and webtoon posting sites is over 90% accuracy. In particular, since the recall value is 1.0 for the torrent and webtoon posting sites, it was confirmed that there is no undetection case. However, the results showed that the detection of the legal sites by the crawler as illegal was also somewhat higher for webtoon posting sites compared to other types. Therefore, in order to reduce such false positives, we are going to conduct research that performs the features of webtoon posting sites more precisely.

In future, we plan not only refining feature extraction for webtoon posting sites, but also improving crawler functionality. In this study, information collection was not smoothly performed for sites with anti-DDoS function and sites using dynamic expression based on javascript. In addition, for sites that connect to advertising sites through multiple redirections,

there were also difficulties in information collection. We plan to develop a crawler that automatically resolves these cases in our future works.

Acknowledgement

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

References

[1] Intellectual Property Office, "Online Copyright Infringement Tracker: Latest wave of research (March 2018)," *The Intellectual Property Office*, May, 2018. Article (CrossRef Link)

[2] 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*, November, 2019. Article (CrossRef Link)

[3] Walls W.D., "Corss-country analysis of movie piracy," *Applied Economics*, Vol. 40, Issue 5, pp. 625-632, April 11, 2008. Article (CrossRef Link)

[4] Bram Cohen, "Incentives Build Robustness in BitTorrent," pp. 1-5, May 22, 2003. Article (CrossRef Link)

[5] Zhaotian Li, Yuesheng Zhu, Guibo Luo, Biao Guo, "A New Copyright Protection Scheme for Depth Map in 3D Video," *KSII Transactions on Internet and Information Systems(TIIS)*, Vol. 11, No. 7, July 30, 2017. Article (CrossRef Link)

[6] Daniel Yue Zhang, Qi Li, Herman Tong, Jose Badilla, Yang Zhang, Dong Wang, "Crowdsourcing-Based Copyright Infringement Detection in Live Video Streams," in *Proc. of 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, August 28-31, 2018. Article (CrossRef Link)

[7] Gummaluri Sai Kumar, Gudipudi Manikanta, B. Srinivas, "A novel framework for video content infringement detection and prevention," in *Proc. of 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, August 22-25, 2013. Article (CrossRef Link)

[8] Daniel Yue Zhang, Lixing Song, Qi Li, Yang Zhang, Dong Wang, "StreamGuard: A Bayesian Network Approach to Copyright Infringement Detection Problem in Large-scale Live Video Sharing Systems," in *Proc. of 2018 IEEE International Conference on Big Data (Big Data)*, December 10-13, 2018. Article (CrossRef Link)

[9] Ken Rudman, Mathieu Bonenfant, Mehmet Celik, Joe Daniel, Jaap Haitsma, Jean-Paul Panis, "SMPTE Periodical - Toward Real-Time Detection of Forensic Watermarks to Combat Piracy by Live Streaming," *SMPTE Motion Imaging Journal*, Vol. 125, Issue 1, pp. 34-41, February 08, 2016. Article (CrossRef Link)

[10] Juan Eh, Lin Qiao, "Intellectual Property Risks and Protection Mechanisms of Big Data," in *Proc. of ICBDR 2018 Proceedings of the 2nd International Conference on Big Data Research*, pp. 38-42, October 27-29, 2018. Article (CrossRef Link)

[11] Abderrahim Abdellaoui, Habiba Chaoui, "Copyright Protection in the External Private Cloud using a Multi-Agent System and Chebyshev Polynomials," in *Proc. of ICCMB '18 Proceedings of the 2018 International Conference on Computers in Management and Business*, pp. 67-73, May 25-27, 2018. Article (CrossRef Link)

Seul-Ki Choi is postdoctoral researcher at Institute for Information and Communication in Ajou University, Republic of Korea. He received the Ph.D. degree from Ajou University, Republic of Korea. His research interests include IoT Security, Vulnerability & Malware analysis and Cryptographic protocols.

Jin Kwak is a professor at 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 Cryptographic protocols, Applied security mechanisms for Cloud and Big Data system and so on.