A Keyword-based IP Tracking Method for Illegal Web Content Distribution Using Port Scanning on HTTP and HTTPS

Seyoung Jang, Byeongchan Park, Seok-Yoon Kim and Youngmo Kim*

Department of Computer Science & Engineering, Soongsil University, Korea
E-mail: Seyjang216@soongsil.ac.kr; pbc866@ssu.ac.kr; ksy@ssu.ac.kr; ymkim828@ssu.ac.kr
**Corresponding Author*

Received 05 March 2025; Accepted 23 April 2025

Abstract

The rapid expansion of online content distribution has led to a significant increase in copyright infringement, where unauthorized works are illegally shared through various web-based platforms. To fundamentally block these copyright-infringing websites, it is essential to accurately identify the IP address or physical location of the original server. However, most illegal content distribution sites utilize advanced security mechanisms, such as DNS resolvers, reverse proxies, and anonymization techniques, to conceal their true IP addresses, making direct tracking increasingly difficult. These evasive tactics allow illegal sites to continue operating while avoiding enforcement measures. To address this challenge, this paper proposes a keyword-based IP tracking method for identifying illegal web content distribution sites by leveraging port scanning on HTTP and HTTPS (ports 80 and 443). The proposed approach systematically detects and analyzes servers that provide unauthorized content by scanning network ports commonly used for web services. By correlating detected IP addresses with keyword-based filtering techniques, this method enables efficient tracking of illegal sites that actively

Journal of Web Engineering, Vol. 24_3, 457–472.
doi: 10.13052/jwe1540-9589.2435

hide their original server's IP address. Through experimental validation, the proposed method successfully pinpoints the IP addresses of illegal content distribution servers, even when they employ obfuscation techniques to mask their identity. This study contributes to enhancing copyright protection by introducing a web-based detection approach that integrates network security techniques, web engineering principles, and automated keyword analysis. Furthermore, the findings provide a practical solution for law enforcement agencies, copyright holders, and regulatory bodies to combat illegal web content distribution more effectively.

Keywords: Internet address, CDN (content delivery network), cloud service, web crawling, illegal site.

1 Introduction

According to a recent survey, the media market size is growing at an average annual growth rate of 2.4% and is expected to reach approximately KRW1248 trillion by 2027 [1]. However, at the same time, illegal content distribution sites that infringe copyright are continuously increasing, posing a major challenge to the media industry. The Korea Copyright Protection Agency's investigation, 'The Glory', which was released as a Netflix original series, ranked first in the Netflix TV series category worldwide, drawing attention not only in Korea but also around the world. However, at the same time, 'The Glory' was illegally distributed through several illegal content distribution sites. As such, copyright infringement has become serious due to the distribution of illegal content, and much popular content has been damaged. The scale of such illegal distribution is quite large, and the estimated damage is reported to be approximately KRW5 trillion [2]. In particular, the indiscriminate copying and distribution of illegal content has not only caused enormous economic losses to content creators and investors, but has also emerged as a serious problem that can hinder the motivation for creative activities. The importance of strong measures against illegal content distribution sites and copyright protection is emphasized so that the works that creators and production companies have worked hard to create can be recognized for their fair value. These illegal content distribution sites are actively utilizing network technologies such as CDNs (content delivery networks) and cloud storage to hide IPs and make tracking difficult. CDN services deliver content to users quickly and reliably through a network of edge servers distributed around the world. In particular, caching, a core function of CDNs, plays

a significant role in reducing the load on the original server, minimizing network delay, and delivering content to users quickly. Caching reduces direct traffic to the original server by storing content on edge servers, and also increases resistance to large-scale attacks such as DDoS. As a result, CDNs significantly improve the availability and performance of websites, providing high reliability to users. However, if these technical advantages are exploited, they can be used by illegal content distribution sites to hide or modify IP addresses, effectively obscuring the location of the original server. This contributes to the easy distribution of illegal content while avoiding IP tracking for stability and investigation, making it difficult for law enforcement agencies to track it. When an illegal content distribution site uses a CDN, only the IP address of the CDN edge server is exposed when searching for a domain, and the IP address of the original server is hidden. By exploiting these structural characteristics, illegal content distribution sites provide users with easy access to streaming services that infringe copyright, as well as much harmful information such as pornography, gambling, and drug trafficking. The Korean government operates various blocking methods, such as DNS (domain name service) blocking and SNI (server name indication) field blocking, to block access to illegal content distribution sites that provide illegal or harmful information [2]. However, illegal content distribution sites are evading crackdowns by using methods such as domain changes and automatic redirection to avoid existing DNS blocking methods [3]. To fundamentally block these sites, it is necessary to identify the IP address and physical location of the original server [4, 5]. Most illegal content distribution sites hide their IP addresses through security technologies such as DNS resolvers and reverse proxies. Measures are required to track operators of illegal content distribution sites and fundamentally block the operation of illegal content distribution sites by identifying the actual IP addresses of illegal content distribution sites in service while they are hidden [6]. To solve these problems, this paper proposes a keyword-based IP tracking method for illegal content distribution sites using port 80 and 433 scans. This method enables fundamental blocking of illegal content distribution and tracking of operators by identifying the actual IP of illegal distribution sites operating in a hidden state.

In the structure of this paper, Section 2 summarizes related research and Section 3 explains a keyword-based method for tracking IP addresses of illegal content distribution sites using port scanning. Section 4 presents experiments and results for the proposed method and Section 5 concludes the paper.

2 Related Research

Due to illegal content distribution sites, various copyright infringement issues have occurred and, in order to respond to this, the Korean government and private organizations are working to raise awareness of copyright among the public through education and campaigns on the problem of copyright infringement, and various attempts and technical response methods are being studied.

2.1 Analysis of the Main Page of Illegal Content Distribution Sites

Park et al. studied a method of utilizing information disclosed within a site, such as site logos, menus, categories, content, and advertising banners, to identify illegal content distribution sites [7]. The study conducted a detailed analysis of 51 illegal content distribution sites over a period of six months and found that the same advertiser identification code could be found in 99.3% of the advertising banners. This empirically proved the possibility of effective identification and blocking of sites whose domains are constantly changing. However, the proposed approach has a limitation in that it relies heavily on the advertiser identification code included in the advertising banner. If illegal content distribution sites remove these advertising banners or change the identification code, the methodology of this study may lose its practical effectiveness. This suggests that future research requires the development of identification criteria and supplementation of additional identification indicators.

Hwang et al. proposed a web crawler system for the purpose of blocking illegal content distribution sites [8]. The study designed a system that collects information from a link collection site of illegal content distribution sites, automatically detects URLs of illegal content distribution sites, and tracks promotional information of illegal content distribution sites on various SNS platforms to collect changed URLs. The accuracy of identifying illegal content distribution sites collected through the proposed web crawler was approximately 95%, accurately identifying 433 harmful sites out of a total of 456 samples. However, this approach has a structural limitation that it heavily relies on link aggregation sites in the process of identifying illegal content distribution sites. If these link aggregation sites are blocked or disappear, information collection may become difficult, which may hinder the sustainability of the system. Therefore, in future research, alternative information collection methods and identification technologies are required to reduce the dependence on link aggregation sites.

2.2 Correlation Analysis of Illegal Content Distribution Sites

Kim et al. presented an efficient and systematic technique to block illegal content distribution sites that infringe copyright [9]. This study monitored over 50 illegal content distribution sites for about a month and confirmed that mirror sites were created simply by modifying the domain address that they were quickly recreated even after being blocked. Through this, we found that simply blocking domain addresses is not a fundamental solution. In addition, the research process revealed that there are difficulties in effectively collecting URLs for illegal content, and in particular, the limitation of having to individually design URL collection programs depending on the structure of the illegal content distribution site was revealed. This analysis suggests the need for a more comprehensive and sustainable approach to blocking illegal content.

Lim et al. conducted an in-depth analysis of the structural characteristics of illegal content distribution sites, uncovering the interconnections among link-providing sites, intermediary sites, and actual video streaming sites [10]. Their study analyzed a total of 62 illegal content distribution sites and found that 53 of them provided illegal content through external links. By conducting a detailed analysis of the content links offered by each site, the study systematically derived the relationships between external servers that deliver the actual content. However, the study highlighted a limitation in that illegal content distribution sites frequently change domains to evade legal sanctions. Additionally, although the study employed Google Analytics IDs and directory information to analyze the relationships among sites, it identified a drawback wherein the accuracy of detection diminishes if operators modify this information. These findings emphasize the need for more robust and sustainable analytical techniques for identifying and blocking illegal content distribution sites.

3 A Keyword-based IP Tracking Method for Illegal Content Distribution Sites Using Port Scanning

This study proposes a keyword-based IP tracking method for illegal content distribution sites using port scans on ports 80 and 443, addressing the issue highlighted in previous research where frequent address changes of illegal distribution sites make them difficult to trace. Figure 1 illustrates the system architecture of the proposed keyword-based IP tracking method using port scans on ports 80 and 443.

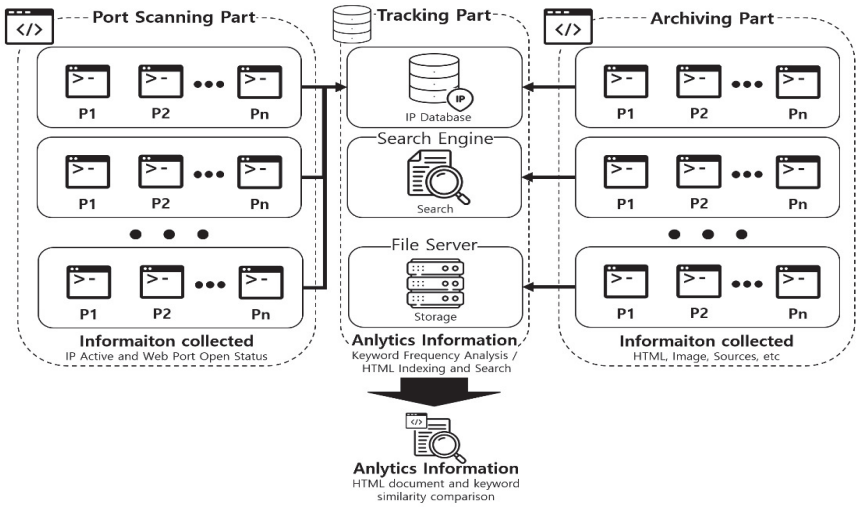

Figure 1 System architecture diagram.

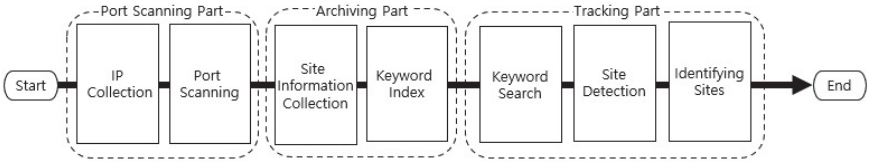

Figure 2 System flowchart.

It consists of a total of three parts: the port scanning part, which collects IP addresses and whether ports are active; the archiving part, which collects information such as the main page and HTML documents of the collected IP addresses and extracts keywords; and the tracking part, which detects and identifies illegal content distribution sites using the extracted keywords. The process proceeds in order as shown in Figure 2.

3.1 Port Scanning Part

The port scanning part consists of two stages: the IP collection stage and the port scanning stage. The IP collection stage is the stage where information is collected on approximately 110 million IPs registered with the Korea Internet & Security Agency (KISA) among the IPv4 system addresses. This scans the active ports and provides basic data for collecting information on the site. IP collection can provide general information about the components of the

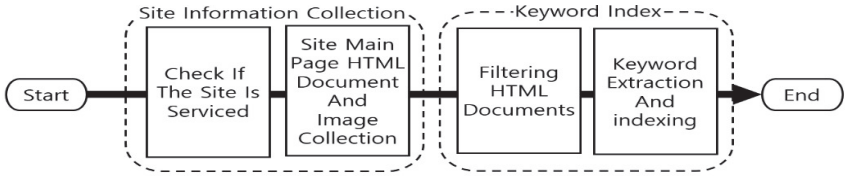

Figure 3 Archiving part flowchart.

target network. The port scanning step is the step to check whether the IP generated using the IP collection process is active and the port. Through the IP collection step, the activation status of each IP and the open status of specific ports are investigated for the IP addresses obtained. In particular, the activation status of ports 80 (HTTP) and 443 (HTTPS) related to web services is checked to identify the IP addresses where web servers are running. This port scanning process makes a significant contribution to identifying what kind of service the target network is providing, and this information is stored in the database as a valid web service target IP. The extracted IP addresses are sent to the database.

3.2 Archiving Part

The archiving part consists of the site information collection stage and keyword indexing stage and proceeds as shown in Figure 3.

In the site information collection stage, the HTML document and image of the main page for the active IP address extracted from the port scanning part are collected. Through this, the structural characteristics and visual elements of each site can be analyzed, and the basic information necessary to understand the site's operational purpose, user interface, content structure, etc. is obtained. At this time the analysis of the HTML document includes the site's structure, hyperlinks, JavaScript functions, metadata, etc. In the keyword indexing stage, unnecessary data such as tags are filtered out from the collected HTML documents and only major keywords are extracted. At this time, keywords are extracted from existing registered illegal content distribution sites, keywords are selected in order of high frequency of appearance, and related keywords are extracted from HTML documents based on these keywords. Figure 4 shows the extracted keywords as a word cloud. Through this process, key information related to illegal content is filtered and indexed in the search engine. The keywords indexed in this way play an important role in the tracking part that checks whether or not the site is an illegal content distribution site.

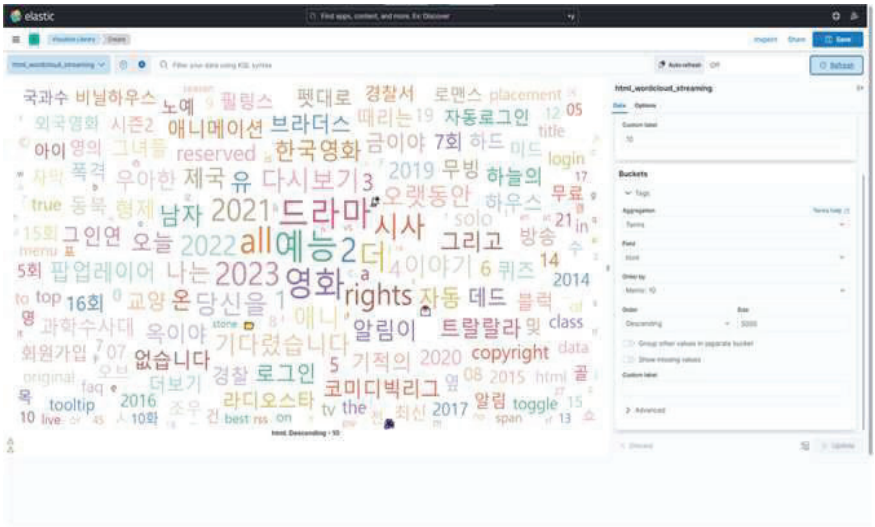

Figure 4 Keyword extraction using a word cloud.

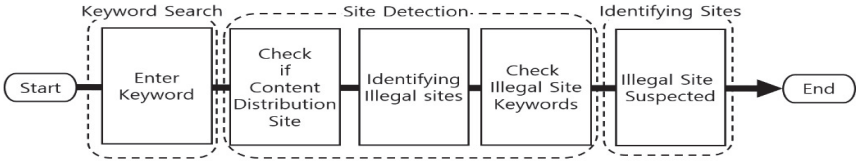

Figure 5 Tracking part flowchart.

3.3 Tracking Part

The tracking part consists of a keyword input stage, a site detection stage, and a site identification stage, and proceeds as shown in Figure 5.

In the keyword input stage, when a query keyword is entered, it is compared for similarity with the indexed keywords to determine whether it is a content distribution site. The site detection stage is the stage where illegal content distribution sites are identified and it is checked whether the content distribution site is legally registered. In the site detection stage, when it is determined that it is an illegal content distribution site, the site is accessed and compared and analyzed, and additionally the type and characteristics of the illegal content distribution site and the illegal content are identified. Even if the IP changes in the future, it can be easily tracked based on this. The site identification step is the step to confirm whether the illegal content

distribution site has been accurately identified in the site detection step. This process compares and analyzes the similarities between illegal content distribution sites accessed through domains and sites accessed through IP. At this point, the main page of the illegal content distribution site is compared, and invariant elements are extracted for comparison by excluding the variable elements of the illegal content distribution site.

4 Experiment on Tracking Methods for Illegal Content Distribution Sites

4.1 Experimental Environment

In order to experiment and verify the method of tracking IP addresses of illegal content distribution sites based on keywords using IP port scanning proposed in this paper, an experimental environment was established as shown in Table 1.

Currently, there are approximately 4.2 billion IPv4 addresses worldwide. Among these, around 3.6 billion are active IP addresses in actual use, excluding those reserved for special purposes. Approximately 110 million active IP addresses registered with the Korea Internet & Security Agency (KISA) were analyzed for activation status and subjected to port scanning for ports 80 and 443, which were then archived. Table 2 presents the status of active IP addresses and ports among the IP addresses registered with KISA.

Table 1 Experiment environment

CPU	Intel(R) Xeon(R) Platinum 8259CL
GPU	NVIDIA Geforce RTX 4090
RAM	32 GB
OS	Windows 10
SSD	100 GB

Table 2 IP status in use

Type	Active Status
Total IPv4	3,687,207,672
kor Total IPv4	112,501,248
Active IP	38,094,248
80 Port (http)	1,373,342
443 Port (https)	597,279

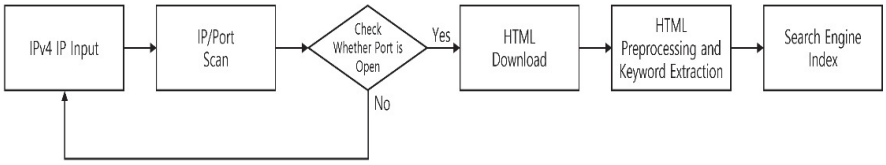

Figure 6 Flowchart of the IP address specification method.

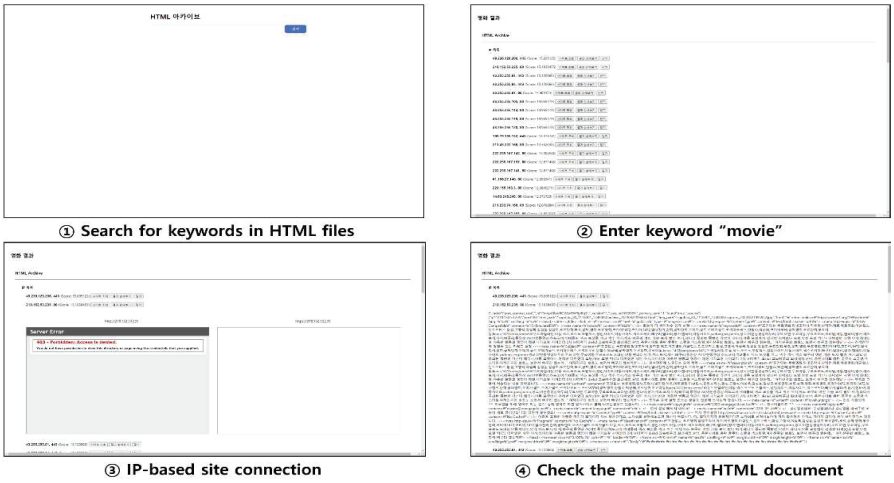

Figure 7 Archiving data analysis.

4.2 Experimental Method

To check the ports, an IPv4 address is first entered. To verify IP activation, ports 80 and 443 are scanned. The HTML of the main page of the active IP is downloaded, and keywords are extracted. The extracted keywords are indexed in a search engine. The system was configured as shown in Figure 6.

The HTML of the activated IPs was downloaded and indexed. From the indexed HTML documents, movie-related keywords were entered to extract IPs with similar keywords. Figure 7 illustrates the process of searching the archived data to identify similar HTML.

To identify illegal content distribution sites, keywords suspected to be site addresses were extracted from the HTML. These extracted keywords were then searched to verify information about the illegal content distribution sites. The process of keyword extraction and search is illustrated in Figure 8.

Figure 8 Keyword extraction and search.

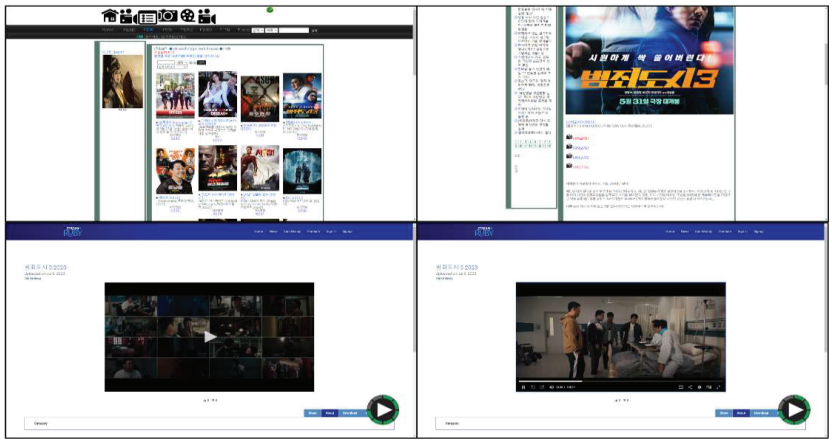

Figure 9 Check for illegal sites.

4.3 Experiment Results

Out of 1,373,342 active HTTP IPs, 8932 were flagged as potential illegal sites based on keyword similarity, yielding an estimated detection accuracy of 94.2% with a false positive rate of 3.8%. To verify the consistency between sites accessed via IP addresses and those accessed via domain names, searches were conducted using keywords related to illegal content distribution sites. As shown in Figure 9, it was confirmed whether videos on the illegal content distribution sites accessed via keywords could be played, and it was verified that the same videos were being played.

Figure 10 Comparison of IP connection and domain connection.

Figure 10 illustrates the comparison between websites accessed through IP addresses and domain names, as identified using the methodology proposed in this study. The results confirm that both accesses correspond to the same website.

5 Conclusion

This study proposes a keyword-based IP tracking method for illegal content distribution sites using port scanning on ports 80 and 443 to address the challenge of locating the physical addresses of illegal sites that evade enforcement by frequently changing URLs or employing redirection techniques. Additionally, to resolve the limitation identified in previous studies – where tracking becomes difficult when illegal distribution sites change their addresses – approximately 110 million IP addresses registered with KISA were scanned on ports 80 and 443 to verify their activation status. For active IPs, HTML files were downloaded to extract keywords, which were then used to trace the corresponding IP addresses. The extracted keywords were searched within the archived database to identify IPs suspected of being associated with illegal sites. Subsequently, the domains extracted from the corresponding HTML files were accessed, confirming that they belonged to the same sites. It is anticipated that the original server IPs of concealed illegal sites can be traced using technologies such as content delivery networks (CDNs), DNS resolvers, and reverse proxies. This approach is expected to facilitate more effective blocking and investigative tracking of illegal distribution sites that conceal their IP. However, the issue of the ease with which illegal distribution sites can still be created remains a significant challenge. Therefore,

future research should focus on strengthening penalties for operators of illegal distribution sites and developing effective policies. Future research will also explore integration with automated real-time blocking systems, legal enforcement collaboration, and international copyright policy frameworks. Therefore, automated approaches to suppress the creation of various types of illegal distribution sites should be explored.

Acknowledgment

This research project supported by Ministry of Culture, Sport and Tourism R&D Program through the Korea Creative Content Agency grant funded by the Ministry of Culture, Sport and Tourism in 2025(Project Name: Development of Copyright Technology for OTT Contents Copyright Protection Technology Development and Application, Project Number: RS-1375027563, Contribution Rate: 100%).

References

[1] Korea Copyright Protection Agency. (2024). Annual Repport on Copyright Protection 2024 (Comprehensive Edition)

[2] Korea Copyright Protection Agency. (2024). Annual Repport on Copyright Protection 2024 (Statistics section)

[3] I. J. Yoo, J. C. Lee, B. C. Park, S. Y. Kim and Y. M. Kim. (2022). A Method for Generating Signature Information to Determine Illegal Distribution of Cloud-based Streaming Video. Journal of Software Assessment and Valuation, 18(2), 239–246. DOI: 10.29056/jsav.2022.12.24.

[4] E. S. Choi, Y. M. Kim and M. C Park. (2023). Research on Methods of Feature Information Gathering for Identifying Illegal Copyright Infringement Sites. Journal of Software Assessment and Valuation, 19(3), 1–10. http://www.riss.kr/link?id=A108761284.

[5] J. W. Choi, G. Y. Choi and S. J. Lee. (2023). Tracing Copyright Infringement Activities through Illegal Streaming Device Protocol Analysis. Journal of digital forensics, 17(2), 62–72. DOI: 10.22798/kdfs.2023.17.2.62.

[6] C. Wan and Y. D Kim. (2021). A Study on the Search and Seizure of User Information in Cloud Computing Service. Law, 70(3), 155–189. DOI: 10.17007/klaj.2021.70.3.005.

[7] E. S. Choi, Y. M. Kim and M. C. Park. (2023). Research on Methods of Feature Information Gathering for Identifying Illegal Copyright Infringement Sites. Journal of Software Assessment and Valuation, 19(3), 1–10. DOI: 10.29056/jsav.2023.09.01.

[8] S. Y. Choo, Y. S. Hwang and S. J. Lee. (2021). Methods for Collecting Harmful Websites Using Web Crawling. Journal of Software Assessment and Valuation, 15(3), 127–138. DOI: 10.22798/kdfs.2021.15.3.127.

[9] C. H. Kim, H. J. Yu, S. Y. Kim and S. H Oh. (2022). Effecient Techniques to Block Copyright Infringement Illegal Streaming Sites. Journal of The Korea Institute of Information Security and Cryptology, 32(5), 837–844. DOI: 10.13089/JKIISC.2022.32.5.837.

[10] J. Y. Jang, K. D. Lim and S. J. Lee (2022). An Harmful site collection system using Characteristic of HTML and URL. Journal of digital forensics, 16(1), 54–63. http://dx.doi.org/10.22798/KDFS.2022.16.1.54.

Biographies

Seyoung Jang received his Bachelor's degree in 2018, received his Master's degree in computer engineering from Soongsil University in 2021, and has been conducting his Ph.D. in computer engineering from 2023 to the present. His research interests include copyright protection and utilization activation.

Byeongchan Park received his Bachelor's degree in 2015, his Master's degree in computer engineering from Soongsil University in 2018, and his doctorate in computer engineering from Soongsil University in 2023. His research interests include copyright protection and utilization activation.

Seok-Yoon Kim received his B.Sc. degree in Electrical Engineering from Seoul National University in 1980, his M.Sc. degree in ECE from the University of Texas at Austin in 1990, and his Ph.D. degree in ECE from the University of Texas at Austin in 1993. His research interests include system design methodology and copyright protection technology.

Youngmo Kim received his Bachelor's degree in computer engineering from Daejeon University in 2003, his Master's degree in computer engineering from Daejeon University in 2005, and his doctorate in computer engineering from Daejeon University in 2011. His research interests include copyright protection and utilization activation.