Efficient Illegal Contents Detection and Attacker Profiling in Real Environments

Jin-gang Kim[1], Sueng-bum Lim[1], Tae-jin Lee[1*]
[1] Department of Information Security, Hoseo University
Asan, South Korea
[e-mail: krch9707@naver.com, gksrkdmftkfkd@gmail.com, kinjecs0@gmail.com]
[*]Corresponding author: Taejin Lee

Received February 28, 2022; revised May 18, 2022; accepted May 31, 2022; published June 30, 2022

Abstract

With the development of over-the-top (OTT) services, the demand for content is increasing, and you can easily and conveniently acquire various content in the online environment. As a result, copyrighted content can be easily copied and distributed, resulting in serious copyright infringement. Some special forms of online service providers (OSP) use filtering-based technologies to protect copyrights, but illegal uploaders use methods that bypass traditional filters. Uploading with a title that bypasses the filter cannot use a similar search method to detect illegal content. In this paper, we propose a technique for profiling the Heavy Uploader by normalizing the bypassed content title and efficiently detecting illegal content. First, the word is extracted from the normalized title and converted into a bit-array to detect illegal works. This Bloom Filter method has a characteristic that there are false positives but no false negatives. The false positive rate has a trade-off relationship with processing performance. As the false positive rate increases, the processing performance increases, and when the false positive rate decreases, the processing performance increases. We increased the detection rate by directly comparing the word to the result of increasing the false positive rate of the Bloom Filter. The processing time was also as fast as when the false positive rate was increased. Afterwards, we create a function that includes information about overall piracy and identify clustering-based heavy uploaders. Analyze the behavior of heavy uploaders to find the first uploader and detect the source site.

Keywords: Bloom Filter, Profiling, Heavy Uploader, OSP, Illegal Content

1. Introduction

The content market has rapidly changed from the offline market to the online market in the past, and it is safe to say that all of them are online markets these days. In addition, mobile devices such as smartphones and tablet PCs are portable and use contents by wirelessly connecting to the Internet, so the mobile environment has rapidly increased. Accordingly, the distribution of copyrighted content is moving from P2P and Web Hard to Bit Torrent and Mobile Web Hard in the mobile environment. According to statistics from the Korea Copyright Protection Agency, the number of online illegal copies used was 1.877 billion, accounting for 90% of the total illegal copy usage. As for the copyright infringement rate by content, movies were the highest at 22.9% and music at 20.3% [1]. Recently, there are cases in which artificial intelligence creates creations. This is because it is difficult to grant rights to content created by artificial intelligence and it is not clear to ask the person who designed the artificial intelligence. Therefore, content created by artificial intelligence are not created by humans and are therefore more vulnerable to the distribution of illegal content [2].

In the past, distributors of illegal content infringed copyright by copying and distributing content by an unspecified majority. Therefore, for copyright protection, filtering-based copyright protection technologies such as search word-based filtering, hash-based filtering, and feature-based filtering to which an unspecified majority are applied were used. By securing a feature database that can identify content, it can be usefully used to search and block illegal distribution of content in a special type of OSP such as web hard [3-5]. **Fig. 1** shows the characteristics of each filtering technology. Filtering technologies for copyright protection have disadvantages that can be easily bypassed by changing titles, combining words, and spacing. Recently, illegal content distributors are changing from an unspecified majority to a specific minority, and a specific minority is distributing the same illegal work using multiple OSP and multiple ID. Therefore, OSP's technical measures were mandatory and used to filter copyrighted content, but it is not enough to eradicate illegal distribution [6].

A certain minority of people who distribute illegal content are referred to as Heavy Uploaders. A Heavy Uploader is a person who distributes a work in bulk on internet sites such as web hard without the permission of the copyright holder. The contribution of this paper is as follows.

- Bloom Filter based illegal work detection and positive error removal algorithm presented
- Extraction of Feature Engineering features by OSP/ID for Heavy Uploader similarity analysis
- Detect the first uploader and analyze the source site through the analysis of the movement status of Heavy Uploader of illegal content

If you track and prevent further distribution through these technologies, numerous copyrighted content will be protected. In this paper, after analyzing the distribution environment of illegal content in the online environment, Chapter 2 examines related technologies and trends for copyright protection, Chapter 3 discusses the profiling technology for distributing illegal content, and Chapter 4 The results of the analysis of tracking and profiling of the distributors of illegal content are reviewed, and the conclusion is concluded in Chapter 5.

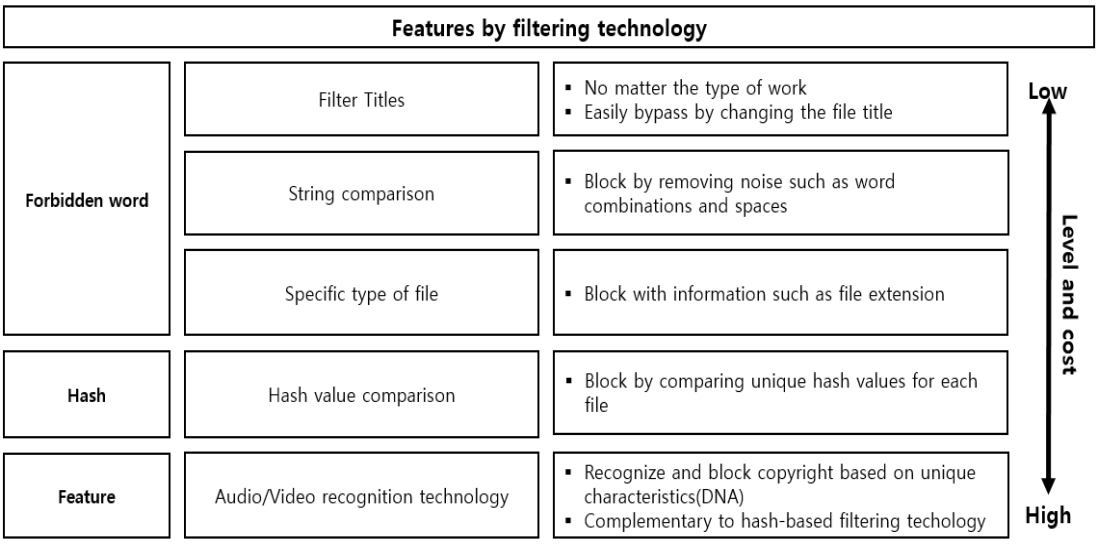

Fig. 1. Features by filtering technology

2. Related Research

Digital rights protection technologies that appeared in 1990 include digital watermarking and DRM (Digital Rights Management). Digital watermarking inserts a unique mark that can prove one's identity, such as fingerprinting, into the content so that even if the content is distributed, ownership can be verified through the mark embedded in the content of illegally copied people [7]. In the case of digital rights management, there is a technology that controls the use of digital content by restricting the content to be used only for its intended purpose. Alternatively, research combining block-chain technology for copyright protection is in progress. Content is registered in the system using chain link, and chain link creates and verifies blocks. **Fig. 2** is the structure diagram of the copyright management system using the corresponding chain link[8].

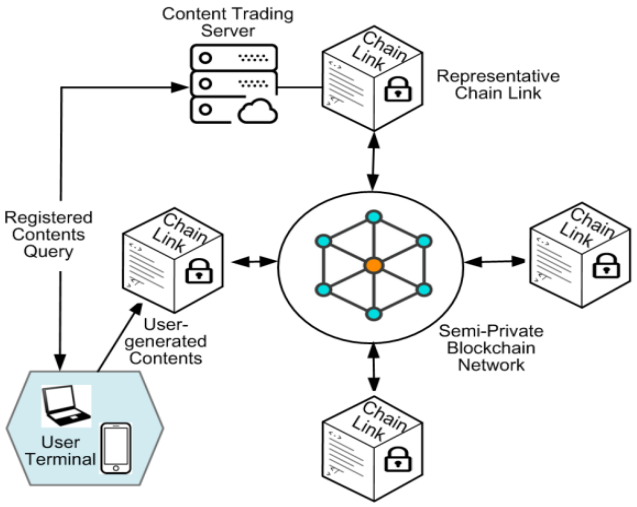

Fig. 2. Chain Link Copyright Management System

By connecting the block-chain and network, content transactions between authorized users are possible, and blocks are periodically created for registered content through chain links, and research on block-chain copyright protection is in progress [9]. However, there are several problems in the practical application of block-chain technology. The first block chain cannot change the transaction details already recorded in the block, but it is impossible to check the forgery or falsification of the authenticity of the work being recorded before it is recorded. In the case of large-capacity content such as the second video, since the data capacity is so large that it cannot be contained in a block, only transaction details and contents are recorded in the block, and the actual data has no choice but to use the server. Therefore, block-chain technology is delayed by technical limitations and uses filtering technology [10-13].

A representative example is the Copyright Protection Center's illegal work tracking system. By combining illegal reproduction monitoring information and search technology with emergency response content, when uploading content from portals, web hard, P2P sites, etc., it is judged as illegal content and blocked by comparing strings with databases with copyrights. As the number of content services through the mobile web as well as the PC environment is increasing, monitoring is being carried out [14].

In addition, fingerprinting-based filtering also exists, and OSP is a technology for blocking illegal copying and transmission of content online, including work recognition measures, search restrictions, transmission restrictions, and sending warning messages. **Fig. 3** shows a conceptual diagram of fingerprinting-based filtering. As a profiling technique, there is a technique that specifies the uploader with the torrent download history through IP. The torrent download history on the web is not deleted, and the torrent download history is checked through the IP suspected of being an illegal uploader. In the case of torrent, upload and download are simultaneous, so the collected history can be used as evidence necessary for legal action. However, there is a disadvantage that the corresponding IP must be known.

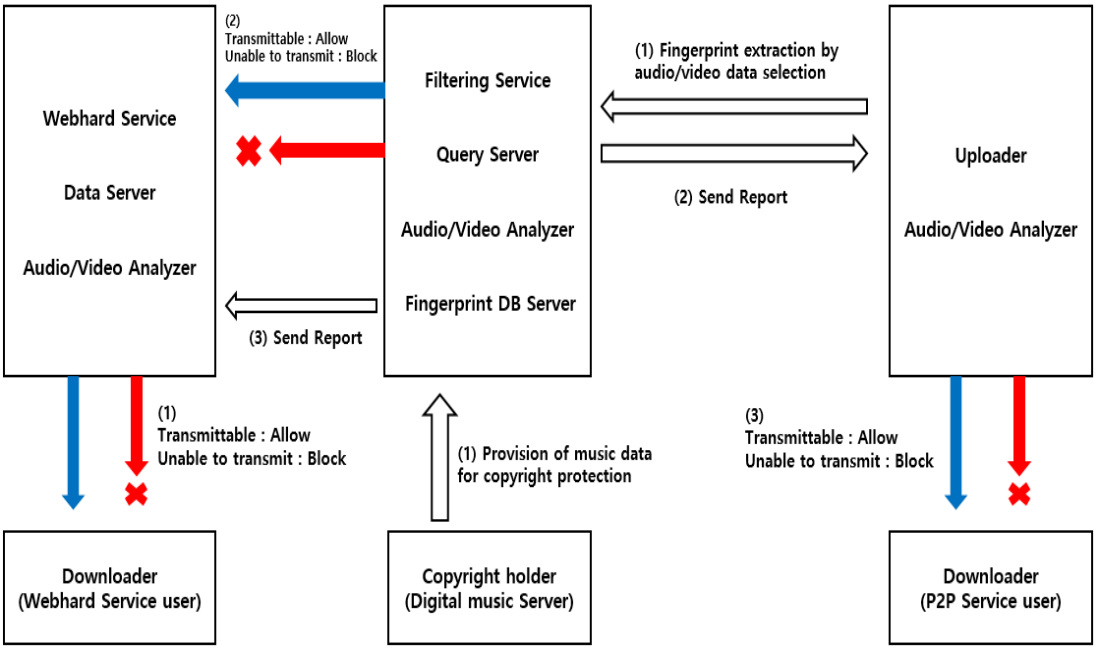

Fig. 3. Technical action conceptual diagram (Filtering based on fingerprinting)

3. Proposal Model

3.1 Overview

There is a difference in the fact that it is possible to detect bypassing filtering, which is the current technology for classifying illegal content, and that it is possible to detect a Heavy Uploader through profiling even if there is no IP. **Fig. 4** shows the profiling technique for distributing illegal content proposed in this paper. Data crawled from OSP and public data with copyright are used, and modified text normalization processing and word extraction are first performed to detect illegal content. The extracted word uses Bloom Filter to detect similar contents as a bit-array. Detected contents have false positives, and a score is generated by comparing actual words to eliminate false positives. Because word is inclusive, illegal content is detected through Score. Next, profiling of the illegal content is carried out through the detected illegal content. For the proposed profiling of illegal content distributors, a feature containing information on the overall illegal content for each OSP/ID is created. The generated characteristics are used to identify the Heavy Uploader. In addition, the same Heavy Uploader can be inferred by automatically tracking clustering-based distributors, and the movement status can be analyzed along with the number of uploads per week of Heavy Uploaders by analyzing the behavior of the distributors of illegal content. Therefore, according to the distribution flow of illegal content, a large number of illegal content distributors are identified, not an unspecified majority, and the first content uploader and source site are detected for this. It is expected that copyright damage can be minimized if the original content uploader and source site are analyzed and blocked.

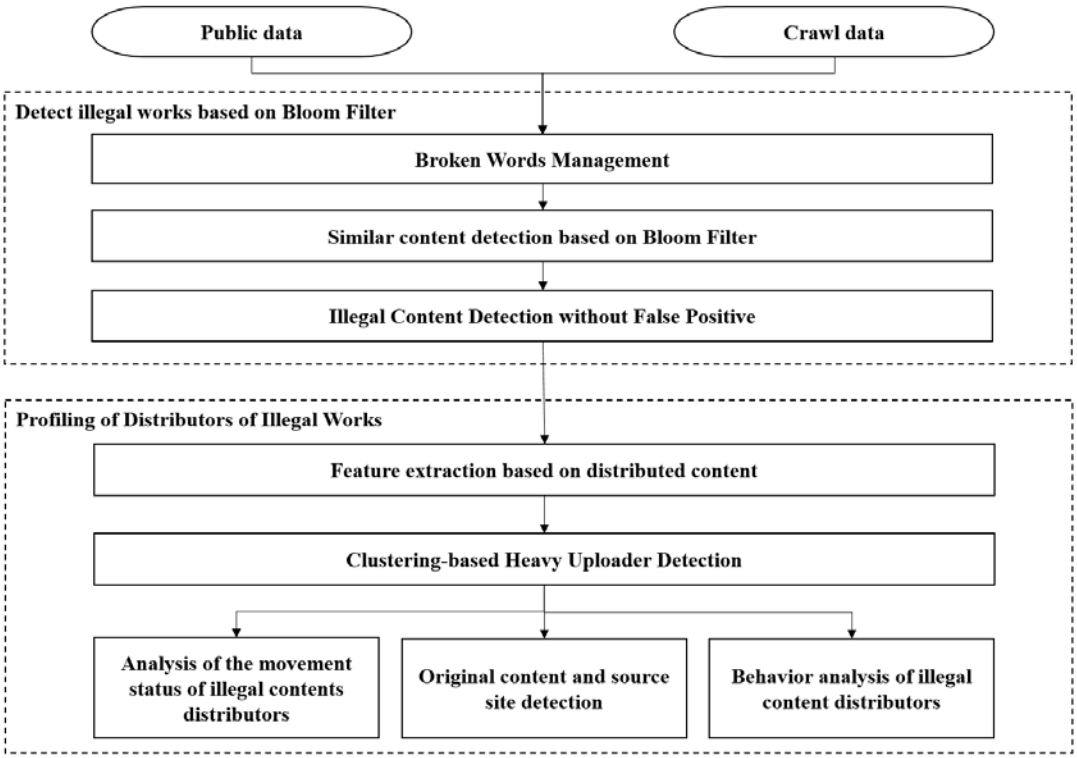

Fig. 4. System Architecture

3.2 Efficient Illegal Contents Detection

3.2.1 Broken Words Management

The profiling technology for distributing illegal content uses data identified as illegal content. In order to use data identified as illegal content, the author used data crawled from OSP and public data in which copyright exists. The normalization process is essential because the crawling data collected from OSP bypasses the existing filtering method because noise exists from special characters, spaces, and separation and combination of Korean consonants and vowels in the titles of content that are being distributed. In text normalization, unnecessary words are removed, spaces are removed, Hangul conversion, English conversion, etc. are generated and normalized [15]. Also, public data is normalized for effective filtering. The normalization process is shown in **Fig. 5**. It operates in the same step and normalizes it by converting it into an appropriate character or word corresponding to the transformed character or word. Since alphanumeric normalization includes various numbers such as dates and prices, such as '2O2O' and '1O8OP', if the leading and trailing characters are numbers based on the alphabet, the corresponding alphabetic character is changed to a number. The Hangeul normalization process transforms and combines consonants and vowels into appropriate consonants and vowels. Converts original characters such as 'ⓐ' and '①' to 'a' and '1', 'O' to 'ㅇ' and 'i' to 'ㅣ'. Alternatively, a pattern combining 'ㅇ' to 'a' is also included. However, there is a possibility that the type that is not combined among the combination patterns is an English letter, so it is converted back to the reverse. In addition, since normalization is impossible when English is mixed, normalization is performed with a complex pattern including the latest type. Finally, all words that are not necessary for the search, such as stopwords, numbers, and English letters, are removed. After going through the normalization process, the word is extracted through the 'Komoran' library.

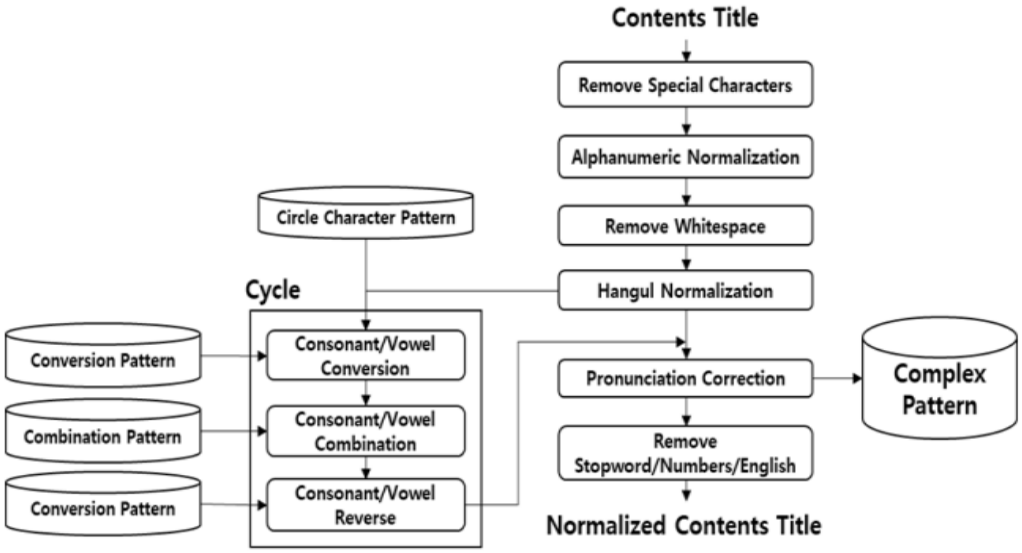

Fig. 5. Broken words management process

3.2.2 Similar content detection based on Bloom Filter

For normal data, you can compare words immediately after going through the normalization process, but in the case of crawled content data, it is often bypassed, and there are cases where there are no words in the content that it represents. Therefore, the existing similar search is impossible. To solve this problem, similar words are detected by transforming them into bit-arrays through Bloom Filter. Create a fixed bit-array for each content by using the word hash function. SHA-256 is used and the bit of the corresponding index is set to 1 through a fixed big modular operation of the hash value of the word generated for each content. The larger the fixed size, the more bits are not duplicated, and the fixed size is set to 10,000 to maintain uniqueness for each keyword. Bloom Filter is performed by making it consist of 10,000 unique bits for each content. Bloom Filter is a probabilistic data structure that determines whether an element belongs to a set or not [16-17]. When it is determined that an element is included in the set, it is possible for a positive error to occur, but conversely, it has a characteristic that false negatives do not occur. **Fig. 6** is the logic for Bloom Filter.

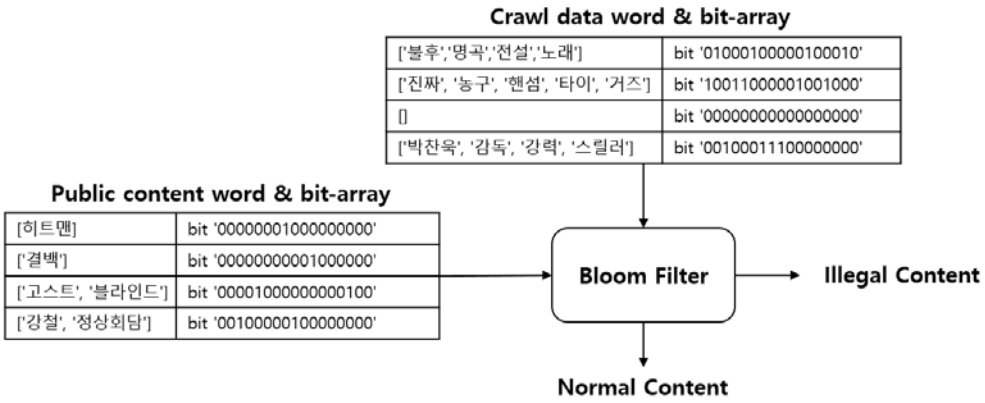

Fig. 6. Bloom Filter Logic

3.2.3 Illegal Content Detection without False Positive

The false positive of Bloom Filter can be adjusted by adjusting error_rate among parameters. The closer the error rate is to 0, the lower the false positive rate, but it increases the processing time and takes longer. On the other hand, the closer the error rate is to 1, the more false positives are extracted, so the false positive rate increases, but the processing performance decreases and becomes faster. Ultimately, the false positive rate and processing time are trade-offs. In this paper, the next step is to find an appropriate balance in these relationships. If it is judged to be illegal content through Bloom Filter, the word of the actual copyright data is compared with the word of the extracted similar content. Since more words of similar contents generally appear than the maximum number of words of copyright data, it is divided into cases where the word of similar contents is less than or exceeding the word of copyright data. If it exceeds half of the maximum number of copyright data words, it is detected as illegal content, otherwise it is detected as normal content. If the word of similar content is less than the word of copyright data, the score is measured and it is judged as exceeding the threshold. The score is calculated by dividing the same number of words by the number of words of similar content. **Fig. 7** is the word detection pseudo-code included in the Bloom Filter result. We increase the error rate of Bloom Filter to increase the false positive rate, detect whether it contains an actual word, eliminate the false positive rate and speed up the processing performance.

Algorithm1– Detect word inclusion in Bloom Filter results

Description. This function compares the actual word to the Bloom Filter result and detects illegal content according to the score.

SET crawl data : Fixed Bit-array Size
SET public data : Array of Extracted Keywords
SET crawl len : crawl data number of words
SET public len : public data number of words
Require : Proceed only when Bloom Filter results are extracted

```
1.  Int count, Score =0                          # 0 initialization
2.  Array Vector[]=0                             # 0 initialization
3.  for crawls data in crawl data:
4.      for publics data in public data:
5.          if crawls data == publics data:
6.              count += 1
7.  Score=count/len(crawl data)                  # Score calculation
8.  if len(crawl data) <= public len:
9.      if Score >= 0.5:
10.         Vector.append()
11. else:
12.     if Score*crawl len >= public len/2:
13.         Vector.append()
14. Return Vector
```

Fig. 7. Detect word inclusion in Bloom Filter results Pseudo-Code

3.3 Attacker Profiling

3.3.1 Same OSP/ID Based Attacker Profiling

For profiling the distributor of illegal content, the work-based feature is extracted as a result of Section 3.2. By utilizing the metadata of data considered to be illegal content and distributing content, the features based on the distributing content are extracted through feature engineering for each dissemination site (OSP)/distribution publisher (ID), and the number of distributed content and the number of illegal content are measured. Feature engineering can generate N feature values by calculating a hash value through a normalized work title-based hash function and adding index 1 to the modular operation result. **Fig. 8** shows the pseudocode of feature engineering. Accordingly, a feature set obtained by extracting N features for each OSP/ID, and the number of circulated content and illegal content are obtained. If the same work is distributed by OSP/ID through work-based feature extraction, the same feature sets are created. Therefore, if similar content are distributed, the key is to create similar feature sets.

Algorithm2– Feature Engineering

This is a feature engineering algorithm for copyright profiling.

SET S : Fixed Feature size
SET Contents : Array of Contents
SET HASH : SHA-256 HASH FUNCTION

Require : OSP and ID are meta information of content (Sort by OSP/ID)

```
1.  Array Features= [0 for i in range(S)]    # 0 initialization

2.  for content in Contents:
3.      if (content == ID) and (content == OSP):
4.          Index=HASH(content) Mod S
5.          Feature[Index]+=1      # Add feature
6.  Return Features
```

Fig. 8. Feature Engineering Pseudo-Code

In this chapter, we identify Heavy Uploader who distribute content in bulk to Internet web hard drives and take profit for profit through the OSP/ID-specific feature set extracted in Section 3.3.1. Based on the data containing the number of circulated content and the number of illegal content by OSP/ID, if the number of illegal content compared to the number of circulated content among the OSP/ID is 10% or more, it is considered as a single Heavy Uploader. In contrast, the clustering-based method of detecting the same Heavy Uploader considers the OSP/ID that similarly distributes a similar work as the same Heavy Uploader. Clustering uses distance-based K-means algorithm to group. The K-means algorithm groups by randomly designating k group centro ID, moving the group centroid until the distance between the group centroid and the data is minimized, and repeating it until it no longer changes. If the number of clusters is small, many OSP/ID will be judged as the same Heavy Uploader, and if the number of clusters is high, the same Heavy Uploader will be viewed as different people. Pay attention to this part and select the number of clusters appropriately to obtain cluster results for each OSP/ID.

The more similar the content distributed by OSP/ID, the more similar the feature set is, so belonging to the same cluster is the key. In other words, OSP/ID belonging to the same cluster means that the distributed content are similar. In fact, if the work distributed by the OSP/ID presumed to be the same person is verified, the title is the same or only specific characters or words are modified and distributed. If it is presumed to be a different person, the same work is distributed under a completely different type of title.

3.3.2 Behavior Analysis by Attacker Group Upload Time

In order to analyze the movement status of the distributors of illegal content, the number of circulated contents per week was calculated based on the distribution date of the content by OSP/ID. The analysis period is automatically calculated based on the distribution date of the work and was recorded from May 2017 to July 2020. Through this, meaningful results were drawn as it was possible to determine the distribution time of the illegal material distributor, the amount of the distributing material, the rest period without distribution, and when the distribution started again. **Fig. 9** shows an example of the movement status of the distributor of illegal content. Most of the OSP/ID of the same cluster that are actually classified can see slightly modified ID with the naked eye. In addition, it is possible to presume the same person by distributing the same work with a completely different ID in the same OSP at a similar time.

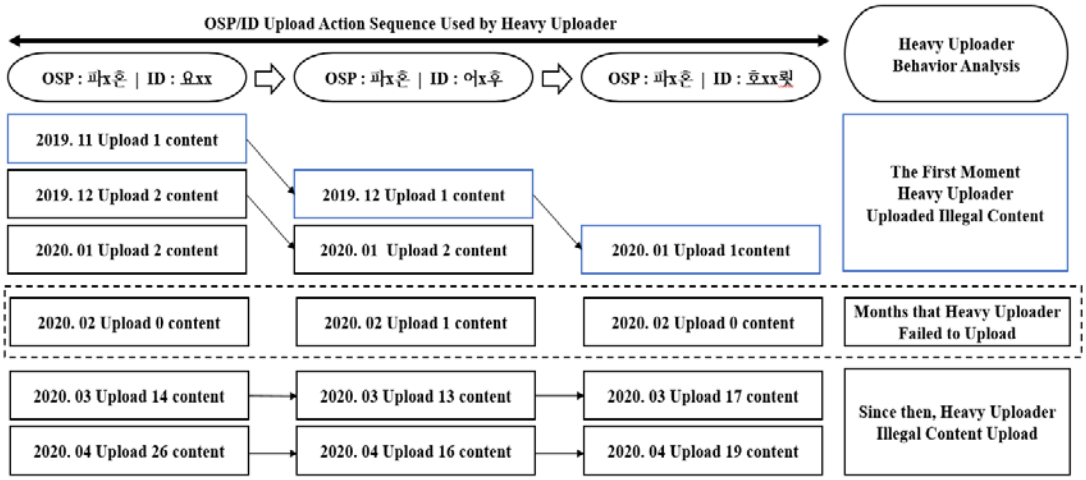

Fig. 9. Heavy Uploader Behavior Analysis

Profiling of illegal content distributors extracts features by OSP/ID through feature engineering from all distributed content, identifies single/identical Heavy Uploaders, and provides movement status analysis results accordingly. After that, all analysis results are synthesized to analyze the behavior of the distributor of illegal content. **Fig. 10** shows three types of behavioral analysis of illegal material distributors through profiling of illegal material distributors.

Most recently, distributors of illegal content are Heavy Uploaders and profit from distributing large amounts of illegal content by distributing them under different ID in the same OSP, under the same ID in different OSP, or under different ID in different OSP. For example, in the same OSP, a number is added to a specific ID and repeated to distribute mass illegal content. This can be estimated as the same person with the naked eye, and can be identified as the same cluster. Through the profiling of the distributor of illegal content proposed so far, as a result of the identification of Heavy Uploaders and the analysis of the movement status of the distributors of illegal content, it detects the distributors of illegal content according to the type of behavior of the distributors of illegal content.

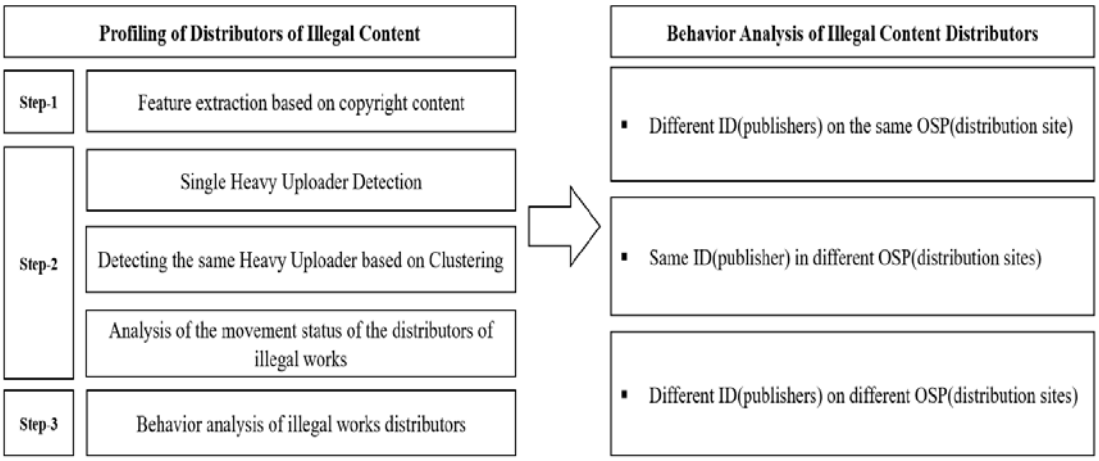

Fig. 10. Types of illegal work distributor behavior analysis

4. Experiment result

4.1 Dataset

There are two datasets used in the experiment: data to confirm the detection rate of illegal content, and a version with large-scale data added for profiling. The first data to confirm the detection rate of illegal content are 100 copyrighted contents and 100,000 data extracted from 2 web hard drives. The second includes the first data, and 185,145 crawled data collected from 47 OSP and 17,704 public data with copyrights were used. Fields of crawled data include title, OSP, ID, URL, publication date, and unique number. Various fields exist in public data with copyright to be compared, but only the title of the work is used. The noise of crawling data is mixed, and it is normalized to compare with public data.

4.2 Efficient Illegal Contents Detection

4.2.1 Similar content detection based on Bloom Filter

In the experiment, the first data set is used to detect similar content based on Bloom Filter. We use the first dataset instead of the second dataset because the word contains 100%, but we have to check whether it is actually detected correctly or not. The proportion of similar content depends on the parameter selection. We conducted an experiment to select an appropriate parameter for detecting illegal content based on a bloom filter. The adjusted parameters include capacity and error_rate. As a result of adjusting the first error rate, when it was 0.1, the detection rate was very high and the false positive rate was low, but the processing performance per piece was 2.19 seconds, which was the slowest. As a result of increasing the error rate, the accuracy gradually decreases, but it can be seen that the processing performance increases. For the second capacity, the default is 10,000. As a result of reducing the capacity to 1000 or 5000 or increasing it to 20,000 or 100,000, the detection rate increased and the processing performance also increased. **Table 1** shows the experimental results to increase the processing performance by optimizing the parameters. Among them, capacity 10,000 and error rate 0.8 indicated in red show the best performance and are used as parameters for the next illegal content detection rate measurement.

Table 1. Optimized parameter detection results

No	Processing Time	error_rate	capacity	No	Processing Time	error_rate	capacity
1	2.19 sec	0.1	10,000	8	1.90 sec	0.65	10,000
2	1.96 sec	0.5	10,000	9	1.88 sec	0.7	10,000
3	1.93 sec	0.55	10,000	10	1.86 sec	0.8	10,000
4	2.03 sec	0.58	10,000	11	2.03 sec	0.8	1,000
5	1.87 sec	0.6	10,000	12	2.03 sec	0.85	10,000
6	1.98 sec	0.6	5,000	13	1.95 sec	0.9	10,000
7	1.96 sec	0.6	20,000	14	1.88 sec	0.99	100,000

4.2.2 Illegal Content Detection without False Positive

We measured the illegal content detection rate based on the capacity 10,000 and error rate 0.8, which had the fastest processing performance. STEP-1 measuring detection rate by simply applying Bloom Filter. STEP-1 detected by increasing the false positive rate, so the processing performance was fast, but the detection rate was low in the 80% range. STEP-2 removes false positives and detects illegal content. As a result, it shows a detection rate of 95% and shows fast processing performance. As a result of comparing the actual data, STEP-1 includes false positives like Crawl Data. However, in STEP-2, only the same data as Public Data is displayed in Crawl Data, and Crawl Data, which is a false positive, does not appear. **Table 2** shows the detection rate and processing performance results for each STEP.

Table 2. Detection results and example data for each STEP

Step	Detection Rate	Processing Time	Public Data	Detected Illegal Data
STEP-1	80%	1.86 sec	너는 달밤에 빛나고	[금의위 비검수춘도]최후의 계승자들이 벌이는 처절한 대결 imm
				[너는달밤에빛나고]멈췄던 시간이 너로 인해 움직이기 시작했다
				[너는달밤에빛나고]멈췄던 시간이 너로 인해 움직이기 시작했다
				[금의위 비검수춘도]최후의 계승자들이 벌이는 처절한 대결 imm
				[금의위 비검수춘도]최후의 계승자들이 벌이는 처절한 대결 imm
			갱 (GANG)	**[갱]미치게 싸울 준비 되었는가 imm**
				보험왕의 m 자 실적의 비밀.. 가입하면 저를 가질수 있어요 고갱님
				보험왕의 m 자 실적의 비밀.. 가입하면 저를 가질수 있어요 고갱님
				[[히.스.토.리.오.브.더.켈.리.갱]] - 호주의 실존하는 갱단의 역사
				갱 (GANG 액션 2019)FHD.1080P
STEP-2	95%	1.88 sec	너는 달밤에 빛나고	**[너는달밤에빛나고]멈췄던 시간이 너로 인해 움직이기 시작했다**
				[너는달밤에빛나고]멈췄던 시간이 너로 인해 움직이기 시작했다
			갱 (GANG)	**[갱]미치게 싸울 준비 되었는가 imm**
				갱 (GANG 액션 2019)FHD.1080P

4.3 Attacker Profiling

4.3.1 Same OSP/ID Based Attacker Profiling

For a large dataset environment for profiling, we use the second dataset. Based on the data determined to be illegal content, 3,113 pairs of OSP/ID were created, and feature sets containing information on the overall content by OSP/ID, the number of illegal content, and the number of distributed content are generated through feature engineering. The feature engineering results are shown in **Table 3**. Heavy Uploader identification was clustered using the feature set generated by OSP/ID. OSP/ID belonging to the same cluster is assumed to be the same Heavy Uploader, which means that the distributed content are similar. To prove this, the content distributed by OSP/ID presumed to be the same Heavy Uploader were identified. **Table 3** shows examples of information about content distributed by the same cluster, presumed to be the same Heavy Uploader. In fact, OSP/ID belonging to the same cluster are distributing the same title. As a result, it represents a Heavy Uploader that distributes the same work with multiple ID on the same OSP. Also, it proves that feature engineering for the same Heavy Uploader is effective. **Fig. 11** is the title of illegal content uploaded by ID used by Heavy Uploader that is judged to be the same person through Clustering.

Table 3. Example of the feature engineering results

Group	ID	0	1	2	3	4	…	95	96	97	98	99	Count	Illegal Count
Attacker Group A	큐 xx	0	5	6	10	2	-	8	3	9	3	7	532	463
	엔 xx	0	2	5	12	2	-	11	4	0	1	0	291	325
Attacker Group B	ixxxx3	3	3	7	1	3	-	3	1	3	3	2	1224	299
	ixxxx7	4	7	4	3	1	-	1	2	2	4	1	916	204
Attacker Group C	cxxxx5	0	1	2	1	1	-	1	1	1	2	1	814	147
	cxxxx3	0	1	2	1	1	-	2	2	1	2	1	472	79
Attacker Group D	더 xxx	4	0	0	0	26	-	0	39	0	0	0	241	196
	텐 xx	4	0	0	0	23	-	0	34	0	0	0	193	147
Attacker Group E	요 xx	0	24	0	21	0	-	0	0	0	2	0	1901	86
	어 xx	0	13	0	18	0	-	0	0	0	4	0	2075	59
	호 xxx	0	17	0	11	0		0	0	0	5	0	2022	66

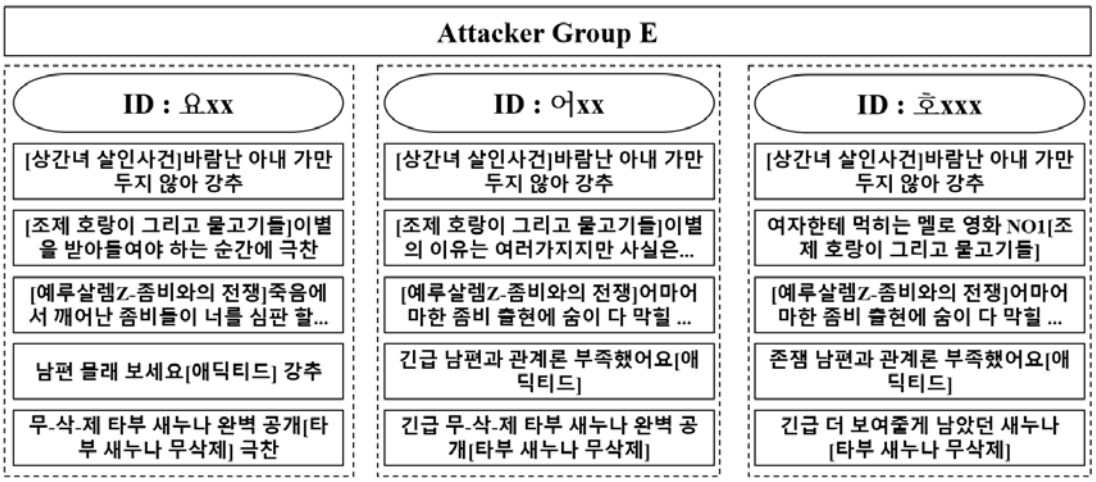

Fig. 11. Example of uploaded content by Attacker Group ID

4.3.2 Behavior Analysis by Attacker Group Upload Time

We classified three types of behavior analysis results. There are uploaders who distribute files with two or more ID on the same OSP, uploaders who distribute files with the same ID on different OSP, and finally, uploaders that distribute files with different ID on different OSP. Detected uploaders have the following three characteristics.

- The time period for starting the upload is the same, and the time period for stopping uploading the content is similar.
- The blank time period is the same when uploading content continuously and not uploading in the middle.
- Continue uploading with the Alpha ID, then stop and start uploading with the Beta ID.

Type 1 is a Heavy Uploader that uploads from the same OSP to a different ID, Type 2 is a Heavy Uploader that uploads to a different OSP with the same ID, and Type 3 is a Heavy

Uploader that uploads to a different OSP with a different ID. They were classified into the same class, and similar contents were uploaded as a result of clustering, and as a result of post-date-based behavior analysis, the upload and non-upload periods were the same in a similar period. The publication date-based measurement is the result of monthly breakdown of the number of illegal content uploaded by Heavy Uploader from May 2017 to July 2020. When we compared the uploaded content titles of the actually detected Heavy Uploaders, it was confirmed that they were uploaded with similar titles. **Fig. 12** is the analysis result of upload time by attacker group.

	OSP	ID	Count	Illegal Count	2020 July	2020 June	2020 May	2020 April	2020 March	2020 Feb	2020 Jan	2019 Dec	2019 Nov	2019 Oct	...	2018 Feb	2018 Jan	2017 Dec	2017 Nov	Type
Attacker Group A	파xx	큐xx	532	463	0	0	309	145	0	7	0	0	0	0	-	0	0	0	0	1
	파xx	엔xx	291	325	0	0	0	256	60	0	0	0	0	0	-	0	0	0	0	1
	미xxxx	bxxx3	1224	299	0	0	0	6	6	1	8	11	7	7	-	1	1	21	33	1
	미xxxx	bxxx7	916	204	0	0	0	0	0	0	0	2	2	0	-	0	0	21	84	1
Attacker Group B	파xx	요xx	1901	86	3	15	23	26	14	0	2	2	1	0	-	0	0	0	0	1
	파xx	아xx	2075	59	2	13	11	16	13	1	2	1	0	0	-	0	0	0	0	1
	파xx	호xxx	2022	66	5	11	23	9	17	0	1	0	0	0	-	0	0	0	0	1
	파xxx	hxxxx9	68	34	0	0	0	23	11	0	0	0	0	0	-	0	0	0	0	2
	위xxx	hxxxx9	73	34	0	0	1	26	7	0	0	0	0	0	-	0	0	0	0	2
Attacker Group C	파xxx	호xxx	343	318	0	0	71	179	68	0	0	0	0	0	-	0	0	0	0	2
	위xxx	호xxx	342	305	0	0	69	162	74	0	0	0	0	0	-	0	0	0	0	2
	위xxx	디xxxx	241	196	0	0	23	107	66	0	0	0	0	0	-	0	0	0	0	3
	파xxx	텐xx	193	147	0	0	14	81	52	0	0	0	0	0	-	0	0	0	0	3
Attacker Group D	파xx	Cxxxxxx2	1015	254	0	7	16	0	231	0	0	0	0	0	-	0	0	0	0	3
	미xxxx	맘xxx	824	264	0	0	6	0	258	0	0	0	0	0	-	0	0	0	0	3

Attacker Group A : While uploading with ID 엔xx, change account to ID 큐xx
Attacker Group B : Start uploading with ID 요xx, then create multiple IDs on the same OSP and upload illegal contents with 3 IDs
Attacker Group C : Upload the same time zone to another OSP with the same ID 호xxx, stop uploading from June
Attacker Group D : Upload the same content with different OSP and different ID only in the same time zone, upload both IDs in April Stop

Fig. 12. Upload Time Analysis Result by Attacker Group

Based on the uploaders classified into the same class, the first posted post in the order of posting date is determined as the original content. The original uploader and the original site are detected through the metadata of the original post. It is not just the order in which the content was uploaded first, but the source site and the original upload ID that were uploaded first among OSP/ID judged to be the same person. In most of the results, it is possible to check that illegal contents uploaded once are uploaded multiple times with different OSP and different ID by changing only the ID to be similar. In addition, it is possible to determine the importance of the source site by dividing the importance in the order of the highest ratio of illegal contents through the ratio of the total number of contents to the number of illegal contents.

5. Conclusion

With the development of the Internet, contents can be accessed quickly and easily online not only on PCs but also on mobile environments. Currently, there is an illegal copy tracking management system that monitors and responds to the distribution of illegal content, but most heavy uploaders insert noise such as spacing, character transformation, and insertion of special characters to bypass the existing filtering system. In the past, illegal content was distributed by an unspecified majority, but recently, the number of mass-distributing mass distributions for commercial purposes is increasing. Since it is impossible to detect illegal works through similar search in such an environment, this paper studied illegal works detection with false

positives but no false negatives by applying Bloom Filter. False positives were solved by detecting illegal works with Score when actual words were compared and included in the results to which Bloom Filter was applied. As a result, it was possible to obtain higher accuracy and faster processing time than before. In addition, original content and source sites were detected through Heavy Uploader profiling. If the source site's Heavy Uploader OSP/ID is blocked, copyright damage is expected to be greatly reduced. In addition to the proposed methods going forward, we will continue our research on copyright protection targeting Heavy Uploaders.

Acknowledgements

This research was supported by the MISP(Ministry of Science, ICT & Future Planning), Korea, under the National Program for Excellence in SW supervised by the IITP(Institute of Information & communications Technology Planing & Evaluation)(2019-0-01834) and supported by the 2021 Copyright Technology Development Project of the Ministry of Culture, Sports and Tourism and the Korea Copyright Commission.(No.2019-PF-9500)

References

[1] J. Kim, C. Hwang, and T. Lee, "Research on heavy uploader profiling technology for illegal content," *Journal of Internet Computing and Services*, vol. 22, no. 3, pp. 75-83, 2021. Article (CrossRef Link)

[2] S. Son, "Copyright Protection on Artificial Intelligence generated Content," *Journal of the Korea Association For Informedia Law*, vol. 20, No. 3, pp. 83-110, 2017. Article (CrossRef Link)

[3] B. Kim, H. Oh, "A Feature-Based Retrieval Technique for Image Database," *The Transactions of the Korea Information Processing Society*, vol. 5, no. 11, pp. 2776-2785, 1998. Article (CrossRef Link)

[4] Y. Kim, D. Shin, "Feature-Based Filtering Technology Performance Evaluation Trend," *Korea Institute of Information Technology Magazine*, vol. 11, no. 2, pp. 1-7, 2013. Article (CrossRef Link)

[5] Y. Oh, G. Jang, H. Kwon, J. Lim, "A Study on the Copyright Protection Liability of Online Service Provider and Filtering Measure," *Journal of the Korea Institute of Information Security & Cryptology*, vol. 20, no. 6, pp. 97-109, 2010. Article (CrossRef Link)

[6] Y. Kim, "Improvement Plan for Special Types of Online Service Provider Responsibility System - Focusing on Technical Measures and Corrective Order and Correction Recommendation System -," *Dankook Law Review*, vol. 40, no. 1, pp. 213-236, 2016. Article (CrossRef Link)

[7] S. Oh, "A Study on the Copyright New Service Model Using Block-chain Technology," *Korea Copyright Commission*, pp. 1-174, Jan. 04. 2018.

[8] J. Hwang, H. Kim, "Block-chain-based Copyright Management System Capable of Registering Creative Ideas," *The Korea Society of Science & Art*, vol. 20, no. 5, pp. 57-65, Oct. 2019. Article (CrossRef Link)

[9] J. Lee, "A Study on Music Copyright Management Model Using Block Chain Technology," *The Korea Society of Science & Art*, no. 35, pp. 341-351, Sep. 2018. Article (CrossRef Link)

[10] J. Kim, J. Nam, "Analysis of illegal content filtering technology trends," *Broadcasting and Media Magazine*, vol. 12, no. 4, pp. 53-63, 2007. Article (CrossRef Link)

[11] C. Li, J. Lu, Y. Lu, "Efficient Merging and Filtering Algorithms for Approximate String Searches," in *Proc. of IEEE 24th International Conference on Data Engineering*, pp. 257-266, Apr. 2008. Article (CrossRef Link)

[12] J. Kim, J. Kim, J. Kim, Y. Chin, "Development of the filtering technology of illegal IPTV contents," *Korea Institute of Information & Telecommunication Facilities Engineering*, pp. 108-111, Aug. 2009. Article (CrossRef Link)

[13] H. Son, K. Kim, Y, Lee, "A File Name Identification Method for P2P and Web Hard Applications through Traffic Monitoring," *Journal of KIISE*, vol. 37, no. 6, pp. 477-482, 2010. Article (CrossRef Link)
[14] Y. Suh, W. Yoo, Y. Kim, W. Kim, "A Study of Copyright Infringement and Technology Measures in a Mobile Environment," *Korea Institute of Information Technology Magazine*, vol. 13, no. 1, pp. 19-25, 2015. Article (CrossRef Link)
[15] C. Hwang, J. Ha, T. Lee, "Modified File Title Normalization Techniques for Copyright Protection," *journal of Information and Security*, vol. 19, no. 4, pp. 133-142, Oct. 2019. Article (CrossRef Link)
[16] S. Kim, J. Kim, "An Analysis on the Error Probability of A Bloom Filter," *Journal of The Korea Institute of Information Security & Cryptology*, vol. 24, no. 5, pp. 809-815, Oct. 2014. Article (CrossRef Link)
[17] C. Hwang, J. Kim, Y. Lee, H. Kim, T. Lee, "High-Speed Search for Pirated Content and Research on Heavy Uploader Profiling Analysis Technology," *Journal of the Korea Institute of Information Security & Crypthlogy*, vol. 30, no. 6, pp. 1067-1078, 2020. Article (CrossRef Link)

Jin Gang Kim is studying for a master`s degree in information security at Hoseo University. He received a bachelor`s degree in information security from Hoseo University. His research interests include malware analysis, machine learning and anomaly detection.

Sueng Bum Lim is studying for a bachelor`s degree in information security at Hoseo University. His research interests include malware analysis, machine learning and anomaly detection.

Tae Jin Lee graduated from Postech Computer Engineering Department in 2003 and graduated from Yonsei University in 2008 and Ajou University in 2017. He worked at Korea Internet Security Agency from 2003 to 2017 and he has been worked in Hoseo University since 2017. His research area are intrusion tolerance technology, VoIP/Wibro security, malware distribution detection/analysis, email security, cyber black box, and malware profiling and mobile payment fraud detection. His current main interests are artificial intelligence, malicious code analysis, intrusion detection.