



ICONI



KOREAN SOCIETY FOR INTERNET INFORMATION

The 12th International Conference on Internet (ICONI 2020)

Dec. 13-16, 2020, Landing Convention Center,

Jeju Shinhwa World, Korea

<http://www.iconi.org>

Proceedings of ICONI 2020

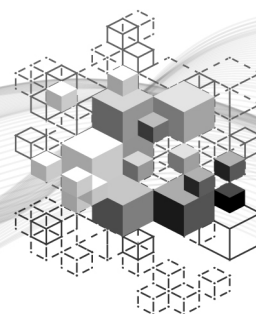
| Organized by |

Korean Society for Internet Information (KSII)

| Sponsored by |



Contents



9-3	Automation Mechanism of Detecting Suspected Piracy Video Streaming Sites Minkyung Lee, Deuk-Hun Kim, Jin Kwak (Ajou Univ., ROK)	392-396
9-4	Prediction Method for New Domain of Suspected Copyright Infringement Site Hae-Seon Jeong, Deuk-Hun Kim, Jin Kwak (Ajou Univ., ROK)	397-402
9-5	Automated Detection Architecture for Piracy Sites Based on Link Collection Sites Eui-Jin Kim, Deuk-Hun Kim, Jin Kwak (Ajou Univ., ROK)	403-406
9-6	Usability and Performance Evaluation of Medical Institution Websites in Malaysia Nor Azman Ismail, Mohd Ainal Farhan Mohamad Johari, Mohammad Ridzwan Syah Irwan, Mohamad Faiz Hakimi Ishak, Su Elya Namira Mohamed (Universiti Teknologi Malaysia, Malaysia)	407-413
9-7	Exploring Levels of Credit Risk in P2P Lending: A Quantitative and Qualitative Emotion Mining Method Min Gyeong Kim, Francis Joseph Costello, Tolchinski Rita, Kun Chang Lee (Sungkyunkwan Univ., ROK)	414-419
9-8	Sentiment Analysis: Study of Social Behaviour of Youth in Different Social Class Nor Azman Ismail, Siti Noorafishah Abd Latif, Su Elya Namira Mohamed (Universiti Teknologi Malaysia, Malaysia)	420-426
9-9	A Novel Approach for Airport Recommendation Using the Comparison of Hierarchical Regression, PLS-SEM, and Bayesian Network-Based PSEM Cheong Kim, Dong Eun Lee, Kyun Sun Eo, Kun Chang Lee (Sungkyunkwan Univ., ROK)	427-432
9-10	C3BR: A Novel Multi-Label Classification Ensemble Mechanism Jong Yoon Won, Chang Min Kang, Dong Eun Lee, Kun Chang Lee (Sungkyunkwan Univ., ROK)	433-438

Prediction Method for New Domain of Suspected Copyright Infringement Site

Hae-Seon Jeong¹, Deuk-Hun Kim² and Jin Kwak^{3*}

¹ISAA Lab., Department of AI Convergence Network, Ajou University
Suwon, Republic of Korea

²ISAA Lab., Department of Computer Engineering, Ajou University
Suwon, Republic of Korea

³Department of Cyber Security and Department of AI Convergence Network, Ajou University
Suwon, Republic of Korea

[e-mail: haeseon0363@ajou.ac.kr, kimdh1206@ajou.ac.kr, security@ajou.ac.kr]

*Corresponding author: Jin Kwak

Abstract

A suspected copyright infringement site, which is suspected of illegally sharing copyrights and violating copyright holders' rights, is monitored by the Korea Copyright Protection Agency to reduce copyright holders' damage and protect copyright. However, suspected copyright infringement site bypass monitoring by using new domains to change their access address. Therefore, efficient monitoring of suspected copyright infringement sites requires predictions of new domains to be created and the operation start date of the site using a new domain. However, the frequency from creating new domains to operating sites varies from site to site. Also, it is difficult to identify the duration of the closed site, which has a previous domain, because it is no longer accessible. However, you can determine the start date of the operation by using an index search which prints the posting date of sub-webpages. Also, as suspected copyright infringement site follows the rule of creating a domain, the new domain can be predicted. Therefore, in this paper, we increase the efficiency of monitoring by proposing a method of predicting new domain and the start date of operation by utilizing the results of the index search.

Keywords: Copyright, Suspected copyright infringement site, Prediction method

1. Introduction

Copyright infringement sites illegally share works without permission from copyright holders. As the demand for copyright infringement sites continues to grow, with free access to copyrighted works, the extent of copyright holders' damage is also increasing.

A site that is suspected of illegally sharing works and needs to be monitored is called a suspected copyright infringement site. To block the operation of suspected copyright infringement sites and the use of a site by users, the Korea

Copyright Protection Agency lists the domains of suspected copyright infringement sites and monitors those domains continuously [1].

Copyright infringement site has steps over time which is defined as lifecycle to respond to monitoring. Suspected copyright infringement site responds to monitoring by blocking access requests or performing redirection to another site. Also, it responds to monitoring by changing the access address to a new domain. However, even if a suspected copyright infringement site block access requests, monitoring can be carried out by bypassing them using VPN, etc. Also, if a suspected copyright infringement site is

redirected to another site after a request for access, monitoring can be performed by analyzing the response domain.

However, if a suspected copyright infringement site creates a new domain and uses it to operate the site, it will be difficult to secure a target domain to monitor. Monitoring can be carried out quickly if the new domain of the suspected copyright infringement site and the operation start date of the site using it can be predicted. Each site has a difference in the cycle from the creation of a new domain to the start date of the site's operation, which can be predicted from the start date of previous domain operations. However, it is difficult to obtain information about the operation start date of the previously closed domain using a search. Although, you can obtain a variety of information within sub-webpages of the site which is already closed by using index search. Thus, it is possible to predict a new domain based on the domain creation rules of copyright infringement sites and to predict the operation start date based on index search results.

Accordingly, in this paper, the monitoring response techniques are analyzed through the lifecycle of the suspected copyright infringement site, and the domain creation rules and index search results of the suspected copyright infringement site are analyzed in Chapter 2. In Chapter 3, we propose a method to predict new domain, creation date, and operation start date using domain history and finally concludes in Chapter 4.

2. Related Works

2.1 Analysis of Lifecycle Model

The lifecycle model for suspected copyright infringement sites is defined as five steps of creation, operation, response, change, and closure as shown in Table 1 [2]. The creation is the step where the domain is first registered and the site is created, and the operation is the step where the site is operated providing copyright infringement contents. The action is a step that allows only bypass access to both the latest and previous versions of the domain. The change is a step when it is redirected to another suspected infringement site. Action and change are steps to respond to monitoring, but monitoring can be performed if you access through VPN or

response domains are identified. The closure is a step where services are no longer provided and operations are discontinued.

Table 1. Steps of lifecycle model

Step	Description
Creation	The Site is created through domain registration
Operation	Operators or users provide copyright infringement contents
Action	Bypass access is available when the domain of a suspected copyright infringement site is updated to the latest version.
	If the domain of a suspected copyright infringement site is not updated to the latest version, normal access is not possible but only bypass access is possible.
Change	When the domain of the suspected copyright infringement site is updated to the latest version, it is accessible without bypassing it and then redirected to another site
	When the domain of the suspected copyright infringement site is not updated to the latest version, normal access is not possible, but only bypass access is available and redirected to another site
Closure	The site does not provide copyright infringement contents anymore and no more operates the site

However, to prepare site blockage and monitoring, the suspected copyright infringement site creates a new domain and use the domain to resume operation of the site. Therefore, monitoring suspected copyright infringement site requires predicting new domains and predicting the start date of operation of the sites that use a new domain.

2.2 Analysis of Rules for Generating Domain

Many suspected copyright infringement sites generate new domains by following the rules in Fig. 1. A domain consists of a sub-domain and a top-level domain, depending on the location of the elements created when separate a domain by “.”. The element on the right becomes the top-level domain, and the element on the left becomes the sub-domain. There may be several sub-domains, and the more they are located on the left, the lower the sub-domain. Suspected

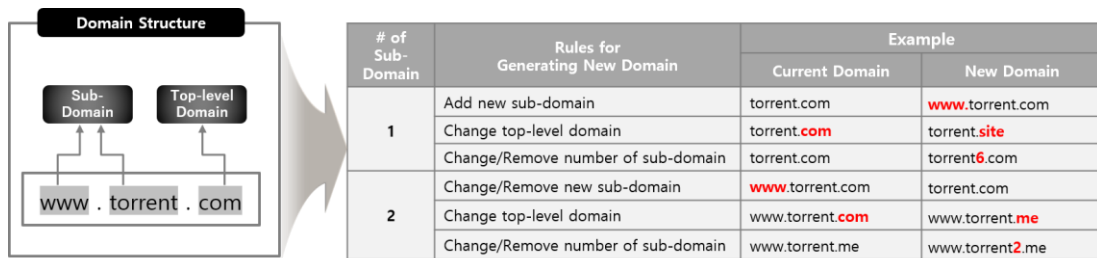


Fig. 1. Rules for generating domain

copyright infringement site generates a new domain by changing the top-level domain or deleting sub-domain. Also, if a number comes before or after a sub-domain, increase the number to create a new domain. You can create a new domain by mixing two or more of the new domain creation rules [3]. Therefore, domain generating rules allow you to predict past domains or predict new domain.

2.3 Analysis of Index Search

The copyright infringement site has a different operating cycle for each site. To know the operating cycle of a site, you need information about the start date of the site's operation in the previous domain. However, for previous domains that are currently closed and inaccessible, it is difficult to retrieve relevant information on the start date of operation. However, Google's index search function allows users to collect information on closed sites about the start date of the site's operation in previous domains. Google performs indexing on sub-webpages of multiple domains. Google provides an index search with the "site:" operator which prints information about the sub-webpage indexed [4].

Suspected copyright infringement site posts the name of the post and the date of posting on the webpage when uploading the work as shown in Fig. 2. The following is a detailed description of the information that exists within the result of the index search.

- "site:" Operator

Perform index search about a domain

- Search Result

Saved information when the domain was indexed

- Sub-webpage

sub-webpage with URL starting with the searched domain



Fig. 2. Result of index search

- Posting Date

The Posting date of a work of art in sub-webpage with URL starting with the searched domain

The posting date means that the domain searched was operated during the period, so the earliest posting date can confirm the operation start date of the site using the domain.

3. Proposed Method

This chapter proposes a method to predict the new domain and the operation start date of the site, based on the domain history of the suspected copyright infringement site and the cycle of the start date of operation utilizing it.

In section 3.1, We generate a list of domain history for suspected copyright infringement sites based on domain generation rules. In section 3.2, we analyze the domain creation date and the site's operation start date for the domain change list. Based on this, in section 3.3, we propose a method to predict the new domain, the creation date of it, and the start date of site operation using it. Details of the proposed method are as follows.

3.1 Generation of Previous Domain History

Create a domain history first to analyze the creation dates and start dates of the previous domains of suspected copyright infringement site.

By using the domain creation rules analyzed in Section 2.2, we generate a list of possible domain changes. The detailed process for creating a domain history of a suspected copyright infringement sites is as follows.

Step 1. Decrease/Delete Number in Sub-domain

The suspected copyright infringement site creates a new domain by increasing the number in the sub-domain. Therefore, it can be expected that a number smaller than the current number in the sub-domain exists within the previous domain, or that a number does not exist. Therefore, the domains created by reducing the number in the sub-domain by 1 are added to the domain history and the domain that has removed the number in the sub-domain also added to the domain history as shown in Fig. 3.



Fig. 3. Example of domain history when two sub-domains are present in the current domain

Step 2. Remove Sub-domain of Left-side

The suspected copyright infringement site creates a new domain when it has one sub-domain by adding a new sub-domain on the left-side. Therefore, if two sub-domains exist, it can be expected that no sub-domain on the left would have existed in the previous domain. Therefore, if there are two sub-domains, the domain that has removed the sub-domain located on the left-side is added to the domain history as shown in Fig. 4.



Fig. 4. Example of Domain history updating

Step 3. Check Top-level Domain

The suspected copyright infringement site creates a new domain by changing the top-level

domain. Therefore, after combining the top-level domains used by suspected copyright infringement sites, only the domain where the index search results exist is added to the domain history as shown in Fig. 5.

After generating a domain change history through the above process, we analyze the domain creation date and site operation start date.



Fig. 5. Process of top-level domain check

3.2 Analysis of Domain Operation Period

Each suspected copyright infringement site has a different cycle from creating a new domain to using it for operation. Therefore, to predict the operation start date of a site using a new domain, an analysis of the domain creation date and the start date of operation of the domains in the history is required. The detailed analysis process for this is as follows.

Step 1. Check Domain Creation Date

As the domain creation date can be checked in the “Creation Date” of the WHOIS response, we request the domain creation date from the WHOIS server. The example of Fig. 6 shows that the creation date for a domain with increased number is up to date.

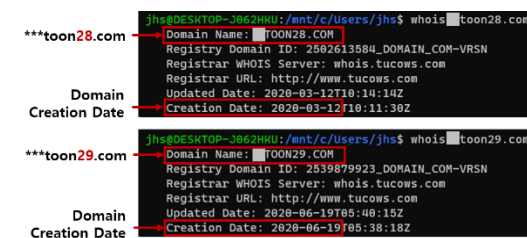


Fig. 6. Domain creation date in whois response

Step 2. Check Posting Date by Index Search

By performing an index search with the “:site” operator, you can check the posting date of posts within the suspected copyright infringement site. Therefore, the posting date can be determined by matching the regular expression which means the date for the index search results. The earliest date of the results is identified as the start date of operation.

3.3 Prediction Method for Next Domain

Based on the creation date and the operation start date of the previous domains, we make predictions about the new domain that will be created in the future, as well as the creation date and start date of operation. We secure the target list to monitor first and then monitor a new domain during the predicted period.

When the creation of a new domain is detected, we monitor the operation start of the suspected copyright infringement site during the predicted period. The symbols for predicting the creation date of the domain and the operation start date of the site are shown in Table 2. The detailed process of predicting the method is as follows.

Table 2. Symbols for prediction of new domain use

Symbol	Description
$N - 1$	The number of previous domains
i	The sequence number of a domain ($1 \leq i \leq N - 1$)
c_i	Creation date of a previous domain ($1 \leq i \leq N - 1$)
o_i	Operation date of a previous domain ($1 \leq i \leq N - 1$)
c_N	Creation date of a current domain
o_N	Operation start date of a current domain
c_{N+1}	Creation date of a new domain
o_{N+1}	Operation start date of a new domain
cd_i	Difference between the domain creation dates, $c_{i+1} - c_i$
od_i	Difference between the domain creation date and the operation start date, $c_i - o_i$
m_c	Mean of cd_i , $\frac{\sum_{i=1}^{N-1} cd_i}{N-1}$
m_o	Mean of od_i , $\frac{\sum_{i=1}^N od_i}{N}$
σ_c	Standard deviation of cd_i , $\sqrt{\frac{\sum_{i=1}^{N-1} (cd_i - m_c)^2}{N-1}}$
σ_o	Standard deviation of od_i , $\sqrt{\frac{\sum_{i=1}^N (od_i - m_o)^2}{N}}$
P_c	The predicted range of c_{N+1} ($c_N \leq P_c \leq c_N + m_c + \sigma_c$)
P_o	The predicted range of o_{N+1} ($c_{N+1} \leq P_o \leq c_{N+1} + m_o + \sigma_o$)

Step 1. Predict New Domain List

The process of creating a new domain list is similar to the process of creating a domain history in Section 3.1. Generate a list of new domains, by increasing the current number A ($1 \leq A$) in the sub-domain numbers from $A + 1$ to $A + 10$. If there are two sub-domains, add a new domain that removed the left-side sub-domain to the new domain list. Next, for each new domain, the domains created by changing the top-level domains used by suspected copyright infringement sites are added to the new domain list.

Step 2. Predict Domain Creation Date

For the history domain changes of one suspected copyright infringement site, obtain the mean (m_c) of the differences (cd_i) between the domain creation dates (c_i) and obtain the standard deviation (σ_c), which is the average for the difference between cd_i and m_c . The new domain creation date (c_{N+1}) may be the same as the current version of the domain creation date (c_N), and there is a high probability that a new domain will be created for a period up to $c_N + m_c + \sigma_c$ based on the domain history. Therefore, the period from c_N to $c_N + m_c + \sigma_c$ is set to the predicted range of new domain creation date (P_c). We monitor the creation of a new domain through WHOIS for the period corresponding to P_c . If the new domain creation is detected, monitor the start of site operations using the new domain through the following process.

Step 3. Predict the Operation Start Date

If creation of the new domain is detected, the site operation is monitored for a period of that range by predicting the scope of the start date of the site's operation.

For the history domain changes of one suspected copyright infringement site, obtain the mean (m_o) of the differences (od_i) between the domain creation date (c_i) and the operation start date (o_i) and obtain the standard deviation (σ_o), which is the average for the difference between od_i and m_o . The operation start date of the new domain (o_{N+1}) may be the same as the creation date of the new domain (c_{N+1}), and there is a high probability that the operation start date of the new domain will be created for a period up to $o_N + m_o + \sigma_o$ based on the domain history. Therefore, the period from c_{N+1} to $c_{N+1} + m_o + \sigma_o$

σ_o is set to the predicted range of operation start date of the new domain (P_o). We monitor and detect the operation start through characteristics of the suspected copyright infringement sites.

- [4] Google Search Quality team, “Search Engine Optimization (SEO) Starter Guide”, *Google*, 2020.

4. Conclusion

The suspected copyright infringement site responds to monitoring by generating a new domain and using it as an access address for the site. Thus, to monitor suspected copyright infringement sites efficiently, new domain and operation start date using it must be predicted.

In this paper, we predicted a new domain name through the rules for generating domain and predicted the new domain creation date and the operation start date by using result of index search.

By using this prediction method, we can quickly detect the start of a suspected copyright infringement site operation using the new domain. Thereby, we can minimize the unauthorized work-sharing and the resulting damage to copyright holders.

In the future, we suppose to improve the accuracy of previous domain history by subdividing the rules for adding sub-domain when forming a new domain. Also, we plan to experiment on the proposed prediction method to verify the predicted results.

References

- [1] Korea Copyright Protection Agency, “2020 Annual Report on Copyright Protection”, *Korea Copyright Protection Agency*, pp. 1-174, Apr. 2020.
- [2] D.H. Kim, H.S. Jeong and J. Kwak, “Development of Lifecycle Model for Copyright Infringement Site”, *Journal of The Korea Institute of Information Security & Cryptology*, Vol. 30, No. 1, pp. 101-121, Feb. 2020.
- [3] H.S. Jeong and J. Kwak, “Automatic Detection Method for Life Cycle Step of Suspected Copyright Infringement Site”, *Proceedings of Summer Conference on Information Security and Cryptography*, Vol. 30, No. 1, pp. 477-480, Jul. 2020.