



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

Automated Detection Architecture for Piracy Sites Based on Link Collection Sites

Eui-Jin Kim¹ Deuk-Hun Kim¹ and Jin Kwak^{2*}

¹ISAA Lab., Department of Computer Engineering, Ajou University
Suwon, Republic of Korea

²Department of Cyber Security, Ajou University
Suwon, Republic of Korea

[e-mail: dmlwls0403@ajou.ac.kr, kimdh1206@ajou.ac.kr, security@ajou.ac.kr]

*Corresponding author: Jin Kwak

Abstract

With the recent development of smart devices and media services, people can use digital content, including movies, dramas, and music. The scale of the domestic copyright market continues to increase, and the scale of overseas copyright exports is also increasing due to the overseas export of K-Contents. However, as the copyright market scale has increased, piracy sites have emerged, and piracy sites have also increased. To prevent this, public organizations are in the process of determining whether there is a suspicion of copyright infringement and blocking piracy sites, but there are difficulties in finding the continuously created piracy sites. Moreover, the new piracy sites and existing piracy sites are updated on the link collection sites and provided to clients. Therefore, in this paper, I propose the advanced automatic detection architecture of piracy sites based on the link collection site.

Keywords: Copyright, Piracy Site, Link Collection Site

1. Introduction

Recently, media services have been developing due to IoT(Internet of Things) technologies, including smart devices, and the ‘snack culture’ that consumes cultural content in a short time [1]. As a result, individuals have increased their access to digital content, and the copyright market is also increasing. Also, due to the overseas expansion of webtoon-based K-drama and K-films, the scale of the copyright market is increasing due to the export of webtoons, movies, and dramas to overseas markets [2]. With the increase in the scale of the copyright market, piracy site’s scale has increased, and the damage caused by piracy sites has also increased [3]. To prevent damage caused by copyright infringement, public institutions are making

efforts to block copyright infringement sites. Still, it is difficult to prevent copyright infringement because the infringement site operator continues to operate the site through domain changes or creates and operates a new infringement site [4]. These domains changed piracy sites, and newly created piracy sites are updated on link collection sites and provided to clients, but research on link collection sites is insufficient.

In this paper, chapter 2 deals with the existing research on site detection techniques, and chapter 3 deals with the feature of link collection sites. Chapter 4 proposes advanced automated detection architecture for piracy sites, chapter 5 deals with the analysis of the proposed techniques, chapter 6 deals with conclusions.

2. Site Detection Techniques

2.1 Detection Technique for Blacklist Site

Blacklist detection techniques warn Web browser users when they access blacklisted websites. Blacklist detection techniques have the advantage of being easy to implement algorithms and low detection rates. Solutions using the Blacklist Detection Method include Google's 'Safe Browsing' and Microsoft's 'Phishing Filter'. Blacklist detection techniques were tested on phishing sites by researchers at the University of Vienna, showing a high detection rate of 90.23% [5]. However, there is a disadvantage of not detecting sites that are not registered, and they are taking follow-up measures such as blacklisting them after blocking the site [6]. It is also difficult to detect blacklists in Korea due to the lack of its blacklists collection and sharing system compared to overseas countries [5].

2.2 Search Technique for Similar Domain Site

A similar domain search technique is a detection technique that uses registering a similar domain when creating a piracy site using domain address keywords and is used when there is a certain pattern in the domain. However, a disadvantage cannot be detected if the Zone file of the Top-Level Domain is not open to the outside world or if the domain pattern changes. Johns Hopkins University and Google researchers tested 2,508 URL(Uniform Resource Locator)s with 'Local Regression Technique', 'Domain-based Feature Analysis', and 'URL Keyword Comparison Method'. As a result, 95.8% of the detection rate and 1.2% of the false detection rate were demonstrated to be effective. However, even a slight change in the URL pattern can reduce the detection rate [5].

Existing detection methods do not reflect new infringement sites and lower detection rates when changing the infringing site's domain. On the other hand, the link collection site already reflects the new and existing infringement site's domain changes. Therefore, it is necessary to research link collection sites.

3. Feature of Link Collection Site

In this chapter, we describe the features of the link collection site. Fig. 1 shows an example of a link collection site.

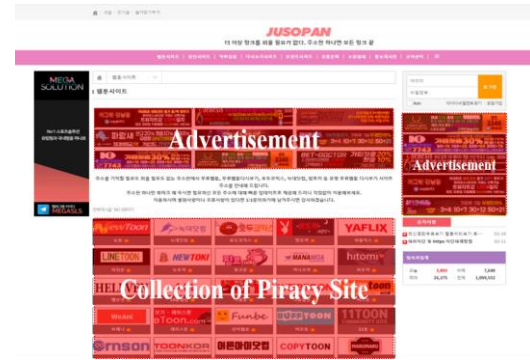


Fig. 1. Example of link collection site

Link collection sites are sites that list and provide piracy sites to clients. Link collection sites generate revenue by posting illegal advertising banners inside the site, and some piracy sites exist in the form of banners on the website's internal frame. Accessing the URL associated with the piracy site banner links to a page that can derive the piracy site's name and domain address called the Internal URL. Fig. 2 represents the Internal URL.

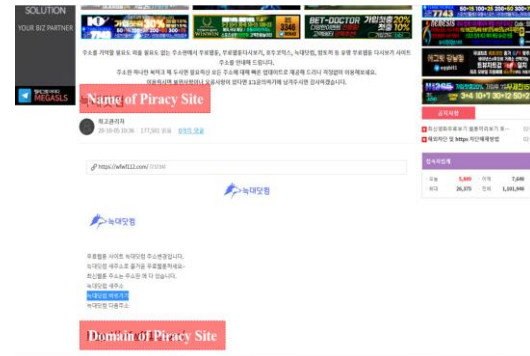


Fig. 2. Example of internal URL

Link collection sites offer Torrent, video streaming, and webtoons as well as pornography and gambling, and they post site links reflecting domains that change periodically due to the short life cycle of the piracy site. Also, new piracy sites and previously closed sites, are being updated on link collection sites.



Fig. 3. Proposed scheme overview

4. Proposed Scheme

This chapter, proposes advanced automated detection architecture for piracy sites using link collection sites, and Fig. 3 is a picture of the proposed scheme.

Step 1. Access Link Collection Site

First, access the link collection site. Using the features of multiple infringement sites in the form of banners on the link collection site page, find the 'a href' tag that connects to the banner link inside the frame. Exclude the URL corresponding to the tag 'a href' from links to the external URL other than the link collection site. The external URL includes an ad banner link, a notice board link, etc., and extracts an internal URL that can lead to piracy sites by conducting a refinement process that excludes external URLs.

Step 2. Access Internal URL of Link Collection Site

Access the internal URL derived from Step 1 and derive the URL accessible directly to the

infringing site. This process uses the features in which the piracy site's name and domain address exist on the Internal URL page.

Step 3. Create Data Set of Piracy Site

The data set is created by storing the piracy site's domain address derived from Step 2 in the 'csv' format. This process derives all the data sets for torrent, video streaming, and webtoon piracy sites, which database piracy sites.

5. Analysis of Proposed Scheme

In this chapter, we deal with the experimental results of the proposed scheme in chapter 4. The experimental environment for implementing the proposed scheme is, as shown in Table 1. Fig. 4 is a code that implements the proposed scheme.

Table 1. Experiments environment

Operating System	Windows 10 Pro
CPU	Intel(R) i5-10400
RAM	32 GB
Python Version	version 3.8.5

```

8 options = webdriver.ChromeOptions()
9 options.add_argument('headless')
10 options.add_argument('--disable-gpu')
11 options.add_argument('Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/85.0.4183.121 Safari/537.36')
12 driver = webdriver.Chrome('chromedriver.exe', chrome_options=options)
13 driver.implicitly_wait(3)
14
15 driver.get('Piracy Site URL')
16 html = driver.page_source
17 soup = BeautifulSoup(html, 'html.parser')
18 notices = soup.select('specific tag')
19 internal_URL = []
20 internal_URL = re.findall(r'href=[\'"]?([^\"]*>+)', str(notices))
21 driver.quit()
22
23 link = []
24 for real in internal_URL:
25     html = requests.get(real)
26     soup = BeautifulSoup(html.text, 'html.parser')
27     notices = soup.select('specific tag')
28     link.append(re.findall('href=[\'"]?([^\"]*>+)', str(notices)))
29
30 link = list(filter(None, link))
31
32 f = open("webtoon.csv", "w", encoding="UTF-8")
33 for i in range(len(link)):
34     f.write("\n".join(str(link[i])))
35 f.close()

```

Python Selenium Driver

Process of Extracting Internal URL

Process of Extracting Piracy Site

Create Data Set of Piracy Site

Fig. 4. Proposed algorithm overview

❑ Python Selenium Driver

In implementing this algorithm, abnormal traffic is sometimes prevented at the link collection site, so the python selenium module was used to solve this problem.

❑ Process of Extracting Internal URL

In extracting the internal URL, we used the features that exist in the form of 'a href' within a specific tag.

❑ Process of Extracting Piracy Site

In extracting the piracy site URL, the piracy site was extracted by accessing the inside of the internal URL and using the features of the URL within a specific tag.

❑ Create Data Set of Piracy Site

In this process, the extracted piracy site was saved in csv format as a database.

The results using the proposed scheme in chapter 4 are shown in [Table 2](#), and the proposed scheme has been used to derive 84 torrent sites, 73 video streaming sites, and 79 webtoon sites that infringe copyright.

Table 2. Result of the proposed scheme

Category	Number of Detected Sites
Torrent	84
Video Streaming	73
Webtoon	79
Total	236

However, even if piracy sites are derived from illegally operated link collection sites, some limitations include normally posting sites, not piracy sites, and blocked sites. It is also necessary to refine the proposed scheme, such as the presence of several same URLs within the internal URL.

6. Conclusion

In this paper, we proposed an architecture for automatically detecting continuously appearing piracy sites to prevent copyright infringement due to increased copyright market size and diversification of content. Based on the proposed detection techniques, the experiment results were also derived from the Data Set. The detection architecture of piracy sites proposed in this paper can respond to continuously appearing piracy

sites.

References

- [1] D.H. Kim, H.S. Jeong, and J. Kwak, "Development of Lifecycle Model for Copyright Infringement Site", *Journal of The Korea Institute of Information Security & Cryptology*, Vol. 30, No. 1, pp.101-121, Feb. 2020.
- [2] Korea Manwha Contents Agency, "Southeast Asia Cartoon Market Settlement", *Overseas Cartoon Industry Report*, p.5, Dec. 2018.
- [3] E.J. Kim, S.K. Choi, and J. Kwak, "A Detection of Webtoon Infringement Site Using Page Source Analysis", *KSII The 15th Asia Pacific International Conference on Information Science and Technology (APIC-IST) 2020*, Jul. 2020.
- [4] Korea Creative Content Agency, "A Survey on the illegal Distribution of Cartoon Webtoon", *KOCCA 18-27*, p.43, Dec. 2018.
- [5] J.H. Sa and S.J. Lee, "Real-time Phishing Site Detection Method", *Journal of the Korea Institute of Information Security & Cryptology*, Vol. 22, No.4, pp.819-825, Aug. 2012.
- [6] Y.G. Jung, "A Study on Blocking Criminal Websites Based on Internet Domain Registration Data of Phishing Sites", *Journal of Digital Forensics*, Vol. 9, No.2, pp.79-96, Dec. 2015.