



KOREAN SOCIETY FOR INTERNET INFORMATION

The 15th Asia Pacific International Conference on Information Science and Technology (APIC-IST 2020)

July 5-7, 2020, Sheraton Seoul Palace Gangnam Hotel, Korea

<http://www.apicist.org/2020>

Proceedings of APIC-IST 2020

| Organized by |

**Korean Society for Internet Information (KSII)
The University of Queensland (Australia)**

| Sponsored by |



Contents



2-3	A study on the genetic algorithm for solving the job shop scheduling problem Fuladi Shubhendu, Chang Soo Kim (Pukyong National Univ., ROK)	44-46
2-4	Implementation of Subjective Visual Vertical Testing System with Virtual Reality Hongly Va, Sungjin Lee (Soonchunhyang Univ., ROK), Soo Kyun Kim (Jeju National Univ., ROK), Min Hong (Soonchunhyang Univ., ROK)	47-49
2-5	VR based random dot kinematogram for Mild cognitive impairment Jun Ma, Eek-Sung Lee, Min Hong (Soonchunhyang Univ., ROK)	50-52
2-6	TestHub - Towards An Ecosystem for software and hardware testing using AI and ML Tahir Hussain, Mir Hassan Qazi, Zahoor Ahmad, Isma Farah Siddiqui (Mehran Univ. of Engineering & Technology, Pakistan)	53-61
3-1	A Detection of Webtoon Infringement Site Using Page Source Analysis Eui-Jin Kim, Seul-Ki Choi, Jin Kwak (Ajou Univ., ROK)	62-65
3-2	Drive-by Download Detection Method with Traffic Analysis Hee-Seung Lee, Jin Kwak (Ajou Univ., ROK)	66-69
3-3	Video Steganography-based Botnet Communication in Telegram Messenger Minkyung Kwak, Youngho Cho (Korea National Defense Univ., ROK)	70-73
3-4	Research on Cyber Intelligence Preparation of the Battlefield Visualization Plan using BGP Archive Data Jaepil Youn (Sejong Univ., ROK), Oh Haengrok (Agency for Defense Development, ROK), Jiwon Kang, Dongkyoo Shin (Sejong Univ., ROK)	74-80

A Detection of Webtoon Infringement Site Using Page Source Analysis

Eui-Jin Kim¹, Seul-Ki Choi² and Jin Kwak³

^{1,3}Department of Cyber Security, Ajou University
Suwon, Republic of Korea

²ISAA Lab., Institute for Information and Communication, Ajou University
Suwon, Republic of Korea

[e-mail: dmlwls0403@ajou.ac.kr, skchoi.isaa@gmail.com, security@ajou.ac.kr]

*Corresponding author: Jin Kwak

Abstract

Recently, K-content has been making inroads into overseas markets due to the globalization of Korean culture. Among them, as the original webtoon K-drama and K-movie are expanding into overseas markets, the size of the webtoon market is increasing. Along with the growth of the webtoon market, however, the ‘Webtoon Infringement Market’ is also on the rise with the emergence of Webtoon-infringing sites. Webtoon infringement sites, which appeared with the paid version of webtoons, appeared in 2013, and the amount of damage caused by this increased steadily until 2018. To solve this problem, the joint investigation team of the private organizations and government has eliminated the ‘Bamtoki’ site, Korea’s largest webtoon infringement site, and reduced the size of the webtoon infringement market, but webtoon infringement sites that illegally taken images of new webtoons, as well as images of webtoons that have already been illegally taken, have continued to emerge. Therefore, to prevent damage caused by webtoons infringement sites, we would like to propose ways to analyze page sources to determine and detect webtoons infringement sites.

Keywords: Webtoon, Infringement, Copyright

1. Introduction

Recently, K-Contents have been making inroads into overseas markets due to the globalization of Korean culture. K-drama and K-films also made inroads into foreign markets, original from webtoon [1]. The overseas webtoon market continues to grow, translating into English, Indonesian, Chinese and Thai [2].

In general, the profit structure of webtoons is that webtoons are provided free of charge on legal site and unlike comics, they do not generate income from the concept of royalty [3]. Thus, writers have been able to earn additional income through part of the webtoon paid since 2013.

Along with the paid-for webtoons, a site that

allows users to view paid webtoons for free has emerged, which is called a ‘webtoon infringement site.’ Webtoon infringement sites have started to increase since 2015 after their appearance in 2013. In 2018, the joint investigation government agencies and private organizations achieved results such as judicial proceedings of eight infringement sites including ‘Bamtoki’ and closure of 12 infringement sites.

However, similar Webtoon infringement sites continue to emerge as data on webtoons that have already been illegally defrauded have been released to an unspecified number of people. The amount of damage caused by Webtoon infringement sites is about 939.9 billion won as of 2018 and if the damage continues, the wrong perception of illegal use of the work will spread, the damage to the company will be serious and

webtoon industry will lose diversity [4].

Therefore, we deal with the size of damage to the webtoon market in section 2, explain the proposed scheme with suggestions in section 3 and conclusion in section 4.

2. Related Works

2.1 Size of damage to the webtoon market

In 2017, the size of webtoon market infringement was calculated using the annual page view (PV), effective click rates and paid sales per PV. As of 2017, the annual PV of the webtoon infringement site was about 5.1 billion views and the effective click rate is the ratio of PVs that clicked 'paid episode' within the infringement site divided by the total traffic and the constant is 59.87%. The amount of paid sales per PV represents the average amount of paid sales for one episode of webtoon and is applied differently for each type of webtoon. Paid sales per PV vary depending on the type of webtoon, which varies in price depending on adult and non-adult webtoons. The average amount of non-adult webtoons is 212.8 won and the average amount of adult webtoons is 385.8 won. The ratio of non-adult webtoons to adult webtoons is 35:65 based on the representative webtoons encroaching site, which is approximately 993.9 billion won [4]. Table 1 is a formula for calculating the size of infringement.

Table 1. Infringement Size Formula

$$\text{Infringement Size (V)} = \text{Annual Page View} \times \text{Effective Click Rate} \times \text{Paid per Page View}$$

3. Proposed Scheme

This section presents detection methods for detecting webtoons infringement sites based on page sources and the algorithm is as shown in Fig. 1.

3.1 Crawl Target Site

This is to determine whether the 'input site' is legal or infringement.

Step 1-1. Check Illegal Site

First, the 'input sites' are crawled and then it is determined whether the site is 'illegal or harmful information sites' that already judged. This process is to check whether the 'input site' has already been judged illegal sites by the government and the organizations.

Step 1-2. Check Traffic Detection Tech.

In the process of crawling the 'input site', it is determined whether data can normally get because of abnormal traffic detection technology such as 'DDoS Cloudflare'.

Step 1-3. Check Registration Number

This is the process of checking whether a business registration number exists on the 'input site'. It is determined whether the infringing site is using a feature in which the business registration number does not exist in the infringing site.

3.2 Check Keyword

This is to determine whether 'input site' is suspected of being infringed by checking keywords that exist on the webtoon infringement site in the crawled site.

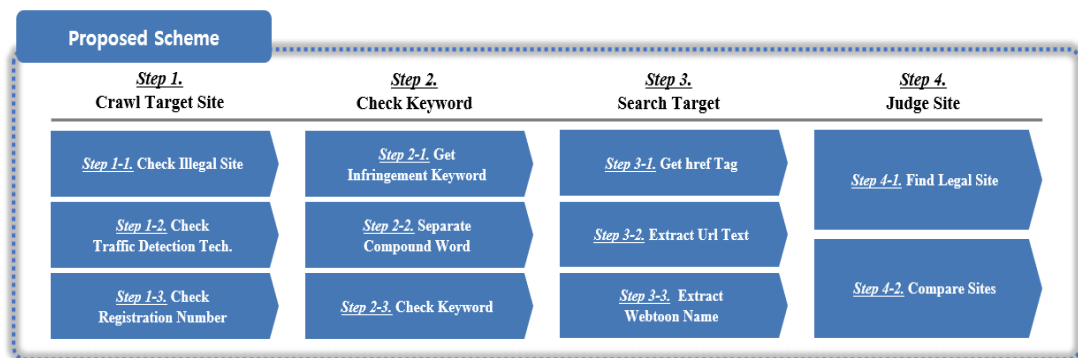


Fig. 1. Proposed Scheme Overview

Step 2-1. Get Infringement Keyword

This is a process of checking the keywords existing on the infringing site and if the keyword is found, it is determined as a suspected infringement site. When crawling the page source, considering the case of a compound word containing the keyword, the texts are brought to the front and back word boundaries using regular expressions.

```
for keyword in keyword_list:
    regex = re.compile(r"\w*{}\w*".format(keyword), re.I)
    matching = regex.findall(str_html)
    infringe_keyword = []
```

Fig. 2. Part of ‘Get Infringement Keyword’ Code

Step 2-2. Separate Compound Word

Separate the compound word using the word segment of python with the string imported from Step 2. By separating the compound words, you can simply lower the false detection rate of the work than the original task of identifying the keyword.

```
for match in matching:
    seg_match = []
    load()
    seg_match = segment(match)
    if keyword in seg_match:
        infringe_keyword.append([match])
```

Fig. 3. Part of ‘Separate Compound Word’ Code

Step 2-3. Check Keyword

Keywords can be checked again to determine which sites are suspected of infringement with a higher detection rate than the general keyword detection process.

3.3 Search Target

This is the process of extracting the name of webtoon from a suspected webtoon infringement site.

Step 3-1. Get href tag

Web pages are organized in the form of ‘Webtoon_Infringement_Site/Internal_url’ on webtoons infringement sites. Because the format of the href tag on all webtoon infringement sites is not the same, it moves to the url inside the ‘a href tag’ to extract the webtoon name.

Step 3-2. Extract Url Text

Using the selenium module of python from the moved internal url, copy the text displayed on the web page to the clipboard and extract it and then remove the unnecessary text.

```
actions = ActionChains(driver)
actions.keydown(Keys.CONTROL).send_keys('a').perform()
actions.keydown(Keys.CONTROL).send_keys('c').perform()
root = Tk()
result = root.clipboard_get()
```

Fig. 4. Part of ‘Extract Url Text’ Code

Step 3-3. Extract Webtoon Name

Use the Konlpy(Python Korean NLP) to obtain high-frequency words from the extracted text. Since ‘a href tag’ url of the infringing site publishes a series of webtoons, using the features that show the most names of webtoons, the highest frequency of words is judged by the name of webtoons by combining them.

```
Okt = Okt()
morphs = okt.morphs(text)
count = Counter(morphs)
sentence_list = count.most_common(5)
```

Fig. 5. Part of ‘Extract Webtoon Name’ Code

```
(‘혼자’, 116)
(‘만’, 116)
(‘레벨’, 116)
(‘업’, 116)
(‘나’, 115)
혼자 만 레벨 업 나
```

Fig. 6. Result of ‘Search Target’

3.4 Judge Site

This is to compare the name of the webtoon extracted in Section 3.3 with a legal site.

Step 4-1. Find Legal Site

Search for webtoon names extracted from section 3.3 on the search engine where you can find the webtoon’s legal site.

Step 4-2. Compare Sites

By comparing the ‘input site’ with the legal site retrieved in Step 4-1, if the ‘input site’ is not a legal site, it can be determined whether or not it is infringed.

4. Conclusion

In this paper, we analyzed how to crawl the page source to reduce the damage caused by the webtoon infringement and then to detect whether the webtoons were violated based on the proposed scheme. As a result of applying the proposed scheme based on the website 'T' that infringes webtoons, the webtoon name was extracted as shown in Fig. 7. After searching the search engine, it was confirmed that the 'input site' was different from the 'K' site, which is a legal site and it could be judged as a webtoon infringement site.

In future work, the experiment will be carried out through regularization and automation of algorithms.

References

- [1] Korea Manwha Contents Agency, "Southeast Asia Cartoon Market Settlement", *Overseas Cartoon Industry Report*, p. 5, Dec. 2018.
- [2] Korea Creative Content Agency, "Webtoon Business Survey", *KOCCA 19-49*, p. 32, Dec. 2019.
- [3] Korea Creative Content Agency, "Current status and characteristics of webtoons and how to activate webtoons-based OSMU", *KOCCA FOCUS 12-09*, p. 12, Aug. 2012.
- [4] Korea Creative Content Agency, "A Survey on the illegal Distribution of Cartoon Webtoon", *KOCCA 18-27*, p. 55, Dec. 2018.
- [5] Intellectual Property Office, "Online Copyright Infringement Tracker", *Kantar Media*, Jun. 2018.
- [6] European Union Intellectual Property Office, "Online Copyright Infringement in the European Union", *Trend in Digital Copyright Infringement in the European Union*, Nov. 2019.