Research on the Infringement of Personal Information by Web Crawlers Based on Legal Regulation

Qingyuan Liu\(^1\) and Feng’e Huo\(^2\)

(\textit{Corresponding author: Feng’e Huo})

Office of Academic Affairs, Cangzhou Normal University\(^1\)
School of Politics and History, Cangzhou Normal University\(^2\)
Cangzhou, Hebei 061001, China
Email: e944070\textunderscore 163.com

(Received Nov. 30, 2018; Revised and Accepted Oct. 11, 2022; First Online Oct. 15, 2022)

Abstract

Web crawlers have been widely used to collect data and information efficiently. This paper first analyzed the application of web crawlers in information crawling, introduced the topic crawler and HITS algorithm, and proposed an improved HITS algorithm to improve the crawling accuracy of web pages. The experimental analysis verified that the improved algorithm was reliable in enhancing the crawling accuracy and always had a crawling accuracy above 0.7 when crawling different topics, higher than the traditional HIS algorithm. Then, this paper analyzed the infringement of personal information by web crawlers, described some relevant cases, pointed out the shortcomings of existing laws, and gave some suggestions. This work supports further improving the rule system needed for crawler applications and punishing personal information infringement.

Keywords: Legal Regulation; Personal Information; Privacy Protection; Web Crawlers

1 Introduction

With the rapid development of technology, information has become an important resource that plays a huge role in many fields\(^9\); therefore, access to information has become an increasingly essential issue. Web crawlers refer to a technology that crawls specific data according to preset algorithms and procedures\(^12\), which is a good way to obtain information\(^3\). With the progress of the Internet, web data is increasing, and web crawlers can help users to crawl the information needed from a massive number of pages\(^15\).

At present, web crawler technology is in continuous development and has been well-used in many fields\(^1\). Luo \textit{et al.}\(^4\) crawled the data of Baijiu in the Tmall platform based on web crawler technology and analyzed the impact of the seller’s credit, brand, and service on the sales of goods. Yu \textit{et al.}\(^11\) analyzed the web crawler technology and applied the PageRank algorithm in a topic crawler to build a vertical search engine. Surahman \textit{et al.}\(^8\) used a web crawler to subdivide the online products of e-marketplaces, including Tokopedia, Shopee, and Bukalapak, and obtained a success rate of 79%.

Pramudita \textit{et al.}\(^7\) developed a multi-threaded web crawler application and also adjusted the web structure to make it crawl web information more efficiently. However, while sharing and utilizing data and information, the problems of information leakage, data misuse, and privacy violation brought about by web crawlers\(^6\) have provoked thought. Individual users cannot enjoy the value of big data but must face the risk of privacy infringement. This paper analyzed the application of web crawlers in information crawling, explained the violation of personal information by web crawlers through specific cases, analyzed the illegality of web crawlers from the perspective of laws, and proposed some suggestions based on the shortcomings of existing laws. The present study provides a theoretical basis for the reasonable and legal use of web crawlers in practice.

2 Application of Web Crawlers in Information Crawling

In the era of big data, the collection, arrangement, and application of web data are beneficial to improve the efficiency of obtaining information and can help various enterprises to catch the market trends and grasp the user behavior to create greater benefits\(^10\). The process of crawling information with web crawlers is as follows. Crawlers access through initial Uniform Resource Locators (URLs) to download and save the content, parse
The workflow of a topic crawler

According to Figure 1, the topic crawler discards the pages that do not match the topic and then stores the pages that match the topic when crawling the web content, and this process is repeated over and over until it reaches the criteria set by the system. The HITS algorithm is an important algorithm for crawling pages in the web crawler, which assigns two attributes to every page: hub and authority. The authority page refers to the page that best matches the keyword, and the hub page is the authority page where multiple links exist. When crawling pages, the HITS algorithm can sort the pages related to the topic according to the authority value from highest to lowest. The principle of the HITS algorithm is to determine the web subgraph $G = (V, E)$ related to the topic and calculate the hub and authority values for every page. It is assumed that there are $n$ nodes in $G$. The authority and hub values of node $i$ are written as $a_i(v)$ and $h_i(v)$. The specific calculation formulas are:

$$a_i(v) = \sum_{(w,v) \in E} h_{i-1}(w),$$

$$h_i(v) = \sum_{(v,w) \in E} a_{i-1}(w).$$

To ensure the invariance of the results, after every calculation, $a_i(v)$ and $h_i(v)$ are normalized, and the formulas are written as:

$$a_i(v) = \frac{a_i(v)}{\sqrt{\sum_{q \in n} [a(q)]^2}},$$

$$h_i(v) = \frac{h_i(v)}{\sqrt{\sum_{q \in n} [h(q)]^2}}.$$

In web analysis, the HITS algorithm still has some shortcomings. First, it tends to ignore new pages, which is because the newly appeared pages are not easily found by the HITS algorithm due to fewer links to other pages; second, the HITS algorithm also tends to cause the problem of topic drift: if there are some pages that point to a lot but are not strongly related to the topic, the HITS algorithm also tends to give them high authority values. This paper proposed an improved HIT algorithm based on weight value $p(f)$, which is a value related to the modification time of a web page and the number of comments on the web page. For a web page, if there is a large difference between the query time and the time of its last modification, it means that the reference value of this web page is small; if the number of its comments (more than ten words as a valid comment) decreases, it means that the value of the web page decreases, which can be described by:

$$f(t, k) = \begin{cases} t, & m < 10 \\ \frac{t}{\log_{10} m}, & m \geq 10, k > 0 \end{cases}$$

where $t$ refers to the difference between the time of crawling a page and the last modified time of the page (unit: month), $m$ refers to the number of comments on the page, and $k$ is the log bottom number. Ultimately, weight value $p(f)$ is calculated by:

$$p(f) = \begin{cases} 1, & t \leq 1 \\ \frac{t}{\log_{10} m}, & t > 1 \end{cases}$$

Based on $p(f)$, the calculation formulas of $a_i(v)$ and $h_i(v)$ of the improved HITS algorithm are updated as:

$$a_i(v) = \sum_{(w,v) \in E} h_{i-1}(w) \times p(f),$$

$$h_i(v) = \sum_{(v,w) \in E} a_{i-1}(w) \times p(f).$$

To understand the performance of the improved HITS algorithm, this paper conducted an experiment on Windows 10 operating system. Let the number of topic-related pages crawled be $N$ and the number of all web pages crawled be $T$. The performance of the topic crawler was evaluated in terms of the crawling accuracy rate, which was calculated by:

$$\text{Accuracy} = \frac{N}{T}.$$
It was seen from Figure 2 that when the number of crawled pages was 1000, the crawling accuracy of the improved HITS algorithm was 0.05 higher than that of the HITS algorithm (0.72 vs. 0.67); when the number of crawled pages reached 2000, the crawling accuracy of the two algorithms reached the highest value, 0.69 and 0.75 respectively, and the crawling accuracy of the improved HITS algorithm was 0.06 higher than that of the HITS algorithm. Then, as the number of crawled pages increased, the number of topic-independent pages crawled by the algorithm also increased, leading to a decrease in the crawling accuracy. When the number of crawled pages reached 5000, the crawl accuracy of the HITS algorithm dropped to 0.58, and the crawl accuracy of the improved HITS algorithm dropped to 0.72, but it was still 0.14 higher than that of the HITS algorithm. The experimental results demonstrate the reliability of the improved HITS algorithm in crawling web content. The improved HIS algorithm could ensure good crawling accuracy even when the number of web pages crawled was large.

To further understand the performance of the improved HITS algorithm, the performance of the two algorithms was compared when crawling different topics, and the results are shown in Figure 3.

It was noticed from Figure 3 that the crawling accuracy of the improved HITS algorithm was always higher than that of the HITS algorithm in crawling different topics. The maximum and minimum values of the crawling accuracy of the HITS algorithm were 0.71 and 0.65, while the crawling accuracy of the improved HITS algorithm was always above 0.7, with a minimum value of 0.74 and a maximum value of 0.77. These results further proved the excellent performance of the improved HITS algorithm in crawling web pages.

### 3 Violation of Personal Information by Web Crawlers

According to the previous section, it is found that web crawlers have significant advantages in crawling information; however, as crawlers are more and more widely used, there are also many abuse cases of crawlers. The improper use of web crawlers not only consumes computer resources excessively but also leads to the leakage of information, which also includes personal information.

Personal information contains many human behavior data, such as personal preferences, occupation, health, etc. It is also closely related to national security, and the illegal crawling and use of personal information by web crawlers is an act of infringement. The infringement acts of web crawlers on individual users include:

1. Creating crawler software and selling it to others for profit;
2. Creating crawler software to crawl personal information;
3. Purchasing the right to use crawler software and crawling personal information;
4. Purchasing the right to use crawler software to crawl personal information and selling it for profit;
5. Serving on a platform company that uses crawlers to obtain user information and obtain personal information to sell for profit.

The number of cases involving web crawlers from 2013 to 2020 is shown in Table 1.

It was seen from Table 1 that the number of cases concerning web crawlers had seen a significant increase in recent years, increasing from 15 cases in 2019 to 27 cases in 2020. The distribution of applicable charges is shown in Table 2.

Some examples of cases related to violating citizens’ personal information are as follows. In 2016, Weimeng company, the operating entity of Weibo, sued TalkU company (the operating entity of Maimai software). During the cooperation period, Weimeng company inferred that the defendant had illegally captured information about Weibo users through crawler technology and that Maimai still used a large amount of basic information about Weibo
users in the months after the cooperation between the two parties ended in August 2014. The court found that the parties had a competitive relationship and constituted unfair competition. From 2019 to 2020, without the permission of Company E, Company Z illegally obtained the order information, merchant information, and other data of Company E through a crawler program, and Company Z was prosecuted for the crime of illegally obtaining computer information system data.

In April-May 2020, the suspects Chen and Wu used crawler programs and other means to illegally obtain more than 90,000 personal information, including names, ID cards, cell phone numbers, etc. They were suspected of violating personal information and were sentenced to fixed-term imprisonment and a fine.

In 2021, a Shanghai information technology company was prosecuted for infringing on citizens’ personal information by crawling personal information such as ID card, social security, provident fund, credit card, and communication records more than 3.08 million pieces through crawling technology without obtaining approval from the relevant state departments and obtaining illegal gains more than 17.5 million yuan.

4 Legal Regulation of Web Crawlers

Legal regulation is the main regulatory method for the infringement of personal information by web crawlers. However, at present, there is no clear domestic regulation for the infringement of personal information by web crawlers.

For the protection of personal information, the content about whether it is illegal for crawlers to crawl personal information is scattered in many laws:

1) The content about the principles of handling personal information in the Civil Code;

2) The limits of operators in crawling and using personal information through crawler technology in the Law of the People’s Republic of China Against Unfair Competition;


In the current legal regulation, there are also some issues that need to be addressed.

1) The interpretation of the provision of “violation of relevant state regulations” in the ‘crime of infringing citizens’ personal information’ is unclear.

2) There is an expansion of judgment on the object of the act of infringement, which deviates from the original purpose of protecting the legal interests of the crime of infringement of citizens’ personal information.

3) The criteria for judging “illegal access” are rough and vague, making it difficult to deal with the increasingly complex illegal acts of web crawlers.

4) There is a lack of Detailed and specific regulations for the determination of the infringement results of the act, which cannot meet the requirements of reality.

The legal regulation of web crawlers is conducive to protecting personal information but restricts the development of network technology. In order to balance the protection of personal information and the flow of data value, the following suggestions are proposed for the current legal regulation:

1) Further clarifying the specific meaning of “violation of relevant state regulations”;

2) Further clarifying the scope of citizens’ personal information in the criminal law, and personal information that has undergone multiple transformations or needs to be combined with multiple other information before identification should not be included in the scope of protection;
3) Further regulating the access rights of crawler technology and clarify the criteria for judging the act of “illegal access”: the act of publicly crawling data under the premise of following the robots agreement does not constitute an illegal act, but the act of forcibly crawling others’ data in violation of the robots agreement may constitute unfair competition;

4) Further clarifying the infringement of the legal interests of crawling personal information with crawlers and determining which behaviors need to be regulated and punished by criminal law.

5 Conclusion

This paper studied web crawlers. First, an improved HITS algorithm was designed, and a high crawl accuracy of the algorithm on web crawling was obtained through experimental analyses, which proved the reliability of the method for information crawling. Then, the infringement of web crawlers on personal information was analyzed, and several relevant cases were introduced. Finally, the current legal regulation shortage was pointed out, and some suggestions were given for balancing personal privacy protection and encouraging technological innovation.

References


Biography

Qingyuan Liu, born in November 1991 in Cangzhou, Hebei Province, has received the master’s degree. She is a lecturer. Her main research direction is civil and commercial law.

Feng’e Huo, born in September 1978 in Haixing, Hebei Province, has received the master’s degree. She is an associate professor. Her main research direction is intellectual property law.