Abstract: In recent years, there have been many cases of malicious crawling of network data by using web crawlers. These cases involve many legal issues such as the protection of citizens' personal information, the ownership of network data rights, etc. These cases have aroused widespread concern among network practitioners and legal professionals. Web crawlers will promote the effective circulation and application of network data, and then promote the healthy and rapid development of the Internet industry. But the malicious crawlers will seriously infringe on the rights and interests of the website, such as affect the normal operation, steal important data, and even constitute a criminal offense. Internet practitioners must abide by relevant laws and regulations, properly crawl data in accordance with the requirements of the robot protocols, and guard against the legal risks that may be brought by crawling behaviors.

Keywords: Web Crawler; Data Crawling; Malicious Crawler; Robot Protocol; Legal Risk

I. INTRODUCTION

Web Crawler, also known as Web Robot or Web Spider, is a computer program that can automatically and efficiently browse the Internet and grab the required data by simulating the behavior of web users. Technology is neutral, but the behavior of people who use technology can be divided into good and malicious, legal and illegal. Web crawlers can also be divided into good crawlers and malicious crawlers. Good crawlers are also called the legal crawlers. They crawl the web pages or the public interfaces on the premise of observing the robot protocols of the websites. Their crawling behavior will not affect the normal operation of the websites, and will not infringe on the personal information of website users and important data resources of the websites. The crawling behavior of the good crawlers is very popular because it can increase the traffic of related websites. For example, general web crawlers used by large search engines such as Baidu and Google will regularly search the entire network to provide users with corresponding web page information. But the malicious crawlers will break the crawling prohibition requirements of the website robot protocol (The crawling prohibition requirements are specified in the Disallow section of the robots.txt file), crawl the core data resources of the websites, and may greatly consume the performance of the crawled websites’ server. In this paper, we will discuss the legal issues of the malicious attack on network system and the malicious data crawling.

II. LEGAL ANALYSIS OF THE MALICIOUS ATTACKS BY WEB CRAWLERS

The malicious attack of the network system by web crawlers will take up a lot of network bandwidth, cause network congestion, make the server unable to process the user’s normal access request, and even cause a crash. With the development of the big data industry, web crawlers, especially malicious web crawlers, are becoming more and more rampant. China’s Internet Network Security Situation in the first half of 2019 shows that in the first half of 2019, the average number of malicious computer programs was about 9.98 million per day. MongoDB and Elasticsearch exposed serious security vulnerability issues. Users can access the database locally or remotely through the default port without any verification and perform any operations such as adding, deleting, modifying, etc. CC attacks on important websites and DDoS attacks from abroad are increasingly frequent in China.

The malicious attack of web crawlers also challenges the legal system of China. Article 21, items 1, 2 and 3 of Cyber Security Law of the People’s Republic of China require network operators to take technical measures to prevent cyber attacks and intrusions. Article 25 of this law stipulates that network operators should formulate emergency plans to deal with network security risks in a timely manner. Article 59 of this law provides the corresponding legal responsibilities. But this law does not stipulate the method, frequency, and legal responsibility of the crawling behavior of Web Crawler, which is a major defect in legislation. On May 28, 2019, the Cyberspace Administration of China (CAC) released Measures for Data Security Management (Draft for Comments). This draft defines the crawling behavior of web crawlers for the first time. Article 16 of the draft stipulates that network operators must not obstruct the normal operation of the websites by using automatic means to access and collect website data. If the crawling behavior of web crawlers seriously affects the operation of the websites, for example, if the traffic of automatic access collection exceeds one-third of the daily average traffic of the websites, the websites have the rights to request the user to stop automated access and collection, and the user should stop it. However, whether the limit of one-third of the daily average traffic is reasonable remains to be discussed. Moreover, the draft did not specify the responsibilities of the website operators in detail. The draft was not operational and needed further improvement.

The malicious attacks of web crawlers will not only cause civil torts, but also cause serious crimes. Article 285 of Criminal Law of the People’s Republic of China stipulates the crime of illegally invading the computer information system. If a perpetrator uses malicious crawlers to deliberately invade a computer information system in the fields of national affairs, national defense constructions, and cutting-edge technologies, he may constitute the crime of illegally invading the computer information system. Moreover, the crime is an abstract dangerous crime. Even if the perpetrator does not steal...
the data after entering the system and does not cause serious harmful results, he will also constitute this crime. Computer information systems in the fields of national affairs, national defense constructions, and cutting-edge technologies store a large number of national political, economic, military, and cutting-edge scientific and technological secrets which are related to the vital interests of China. Once these computer systems are maliciously invaded, state secrets will be in serious danger of being leaked. Article 286 of Criminal Law of the People’s Republic of China stipulates the crime of destroying computer information system. If a perpetrator invades the general computer information system, he may constitute the crime of destroying the computer information system. This crime is a result crime. It not only requires the perpetrator to intentionally carry out the illegal invasion, but also requires the invasion to cause serious consequences.

III. LEGAL ANALYSIS OF THE MALICIOUS CRAWLING DATA BY WEB CRAWLERS

Data, namely network data, refers to all kinds of electronic data collected, stored, transmitted, processed and generated through the network. Data is anonymous and impersonal. Although data is generated by the network behavior of many users, it is not relevant to specific users. Data is different from personal information of citizens. Personal information is related to a specific natural person. It is a variety of information that can identify a natural person’s personal identity and has the characteristics of identification. The protection of personal information of citizens in Chinese law has been relatively perfect. Article 111 of General Rules of the Civil Law of the People’s Republic of China stipulates that the personal information of a natural person shall be protected by law. Article 1034 of Civil Code of the People’s Republic of China, Article 1136 of the Civil Code, which will enter into force on January 1, 2021, has the same provisions. Only with the consent of users can network operators collect, use, process and transmit personal information. Article 253 of Criminal Law of the People’s Republic of China stipulates the crime of infringing citizens’ personal information. If a network operator steals personal information by using web crawlers, or illegally provide or sell personal information, he will constitute civil tort or even constitute the crime of infringing citizens’ personal information.

As the commercial value of data continues to be highlighted, big data companies and data trading platforms are developed rapidly. They create corporate and social benefits by collecting, analyzing, and trading data. Data can be divided into two categories: the public data and the core data. As for the public data, it is shared and open. Network operators can collect, mine and analyze the public data by using web crawlers without affecting the normal operation of the website network system. But for the core data, network operators need to abide by the robot protocol and cannot crawl without permission.

In recent years, some companies have crossed the red line of law and used malicious crawlers to collect data illegally. Lawsuits about illegal collection of data are constantly emerging. The case of Beijing Taou World Technology Co., Ltd. v. Beijing Micro Dream Network Technology Co., Ltd., which was finally adjudicated in 2016, is regarded as the first case of unfair data competition in China. Beijing Taou World Technology Co., Ltd. is the operator of Maimai software and Maimai website (website: https://maimai.cn/). Beijing Micro Dream Network Technology Co., Ltd is the operator of Sina Weibo, and is the filing person of www.weibo.com, www.weibo.com.cn, www.weibo.cn, and obtains the network culture business license. The court of second instance of the case held that Maimai illegally grabbed and used the user information of Sina Weibo. Maimai obtained the user information of Sina Weibo without authorization and displayed it in details in the Maimai app. Maimai’s behavior infringed the business resources of Sina Weibo. It was an unfair competition act prohibited by law and should bear corresponding tort liability. In the same year, the trial of Shanghai Hantao Information Consulting Co., Ltd. v. Beijing Baidu Netcom Science Technology Co., Ltd. was concluded. The court emphasized that the unauthorized use of other people’s data and information cannot be deemed as unfair competition, but should be considered with various factors to delineate the boundary of the behavior. The adjudication rules in this case not only encourage the market subjects to operate honestly, but also provide due judicial incentives for enterprises’ efforts in data collection, mining, and collation, and also reasonably determine the boundaries of the behavior of using other people's data. This is of great significance for maintaining a fair and healthy data competition market environment. Other cases, such as Shenzhen Goom Science Technology Co. LTD v. Wuhan Yuanguang Technology Co., Ltd., Taobao (China) Software Co., Ltd. v. Anhui Meijing Information Technology Co., LTD, also followed the same rules of adjudication.

As for the ownership of data, Chinese laws have not yet made specific provisions, and scholars have different views. However, through the analysis of the above cases, it can be concluded that the courts tend to believe that the core data has commercial value and can bring competitive advantages to network operators. So the core data should be protected by Law of the People’s Republic of China Against Unfair Competition. If someone uses web crawlers to crawl the core data of others without permission, it is an unfair competition behavior and should bear the corresponding civil tort liability.

IV. REASONABLE USE OF WEB CRAWLER TO AVOID LEGAL RISKS

Technology is a double-edged sword. The same is true for web crawlers. Web crawlers have created the prosperity of the big data industry, but they also face many legal risks. Network operators must use legal and reasonable methods to collect network data.

A. To follow the principle of legality

Laws such as General Rules of the Civil Law of the People’s Republic of China, Cyber Security Law of the People’s Republic of China and Criminal Law of the People’s Republic of China have made systematic provisions on the protection of citizens’ personal information. Network operators must collect and use personal information under the premise of citizen authorization. If network operators want to transfer personal information, they must obtain authorization again. It is forbidden to crawl other people’s core data with commercial value, otherwise it may constitute unfair competition.

B. To follow the principle of rationality

Network operators should abide by the requirements of the robots protocol of others’ websites, and cannot crawl the data that others’ websites prohibit crawling. During crawling
the public data, network operators must control the crawling frequency of the crawlers, which can not cause too much pressure to the crawling websites, affect the normal user access, and of course, prohibit DDOS attacks. Websites should also improve their own anti-climbing measures to strengthen the safety of websites.

In short, only by using web crawlers legally and reasonably can the Internet industry continue to prosper.

References


