DS NEWSLETTER savoir faire
NEWSLETTER – LEGAL INFORMATIONS
PROPOSED BY DS AVOCATS

Web Crawlers' Data War

ASIE

INTRODUCTION

In this era of Data Economy, companies are required to exploit large volumes of data which, once analyzed, can have a significant impact on their productivity and profitability. In recent years, for example, financial technology (Fintech) has been greatly driven by big data but also under heated discussion in terms of compliance. The use of big data implies to always keep in mind the following question: what is the source of data and the way of data collection?

In this regard, Web crawlers as a strong tool of data collection should not be ignored.

Web crawler, also known as "web spider" or "web robot", is a program or script that automatically grabs new Internet data according to certain rules. The main function of Web crawler is to automatically browse the network, grab information, and build index.

The first Web crawler was born in the United States in 1993. Nowadays, well-known Web crawlers are Google's Googlebot, Baidu's Baiduspider, and Apple's Applebot. In short, no Web crawlers, no search engines!

Developed for almost 30 years, now the Web crawler plays a more critical role in the battle of data. Companies may use it directly to collect data from the Internet (pictures, videos, texts, etc.), or use the services of providers that provide data derivatives services/products by using Web crawler. In reality, compliant use of Web crawlers will improve efficiency. On the other hand, uncontrolled use of Web crawlers can lead to civil and even criminal liability for companies.

THE "CODE OF CONDUCT" OF WEB CRAWLERS

As a data "pet", Web crawler crawls data for a living. So, can Web crawlers freely crawl any data? If your answer is yes, the following use cases may change your mind.

In order to restrict the web crawler's arbitrary crawling, in 1994, a network engineer in Holland formulated a "Code of Conduct" for the Web crawler. When the Web crawler reaches a website, the website owner can display the "Code of Conduct" on the website to restrict the Web crawler (which data can be crawled and which data cannot be crawled). Before crawling data, Web crawlers should carefully read the "Code of Conduct" and abide by it. If the Code of Conduct does not offer any data to the Web crawler, the Web crawler should leave without touching any data. In practice, the above-mentioned "Code of Conduct" is called "*Robots Exclusion Protocol / Robots Protocol*" (mostly known under the terminology "robots.txt"), which is also one of the prevailing anti-crawling strategies.

However, *Robots Protocol* is not an effective technical measure, but like a "Notice" posted on the "door" of the website. Therefore, Robots Protocol can only effectively instruct the web crawlers with good faith, but it cannot prevent malicious web crawling activities. Not surprisingly, legal issues caused by the illegal Web crawlers arise.

Use case 1: The first case occurred in the United States in 1999. The web crawler of Bidder's Edge's website violated the Robots Protocol set by eBay website and crawled all kinds of commercial data of eBay website. In the end, the court prohibited Bidder's Edge from collecting eBay's website data through technical means (including web crawlers) without the permission of eBay.

Use case 2: The second case was in 2017 and between LinkedIn and HiQ. LinkedIn has more than 600 million users worldwide and provides different level of privacy protection strategy for its users. Users can choose to display their personal data to the public/to the contacts/to the contact network.

When users choose to go completely public, their personal data will be accessible to non-LinkedIn users and searching engines like Google. On the other side, HiQ is a company providing human resources monitoring service and "talents map" for employers, which is mostly based on the data collected from LinkedIn. After being put on notice to cease its actions, HiQ sued LinkedIn for an injunction to prevent LinkedIn from blocking access to its users' public data. HiQ was supported by the court. Indeed, the court found that LinkedIn's ban on access to public profile data was not conducive to a healthy market and had the effect of distorting free competition.

It is evident that the above two cases end differently, and the difference lies in how the technology is used, the competition relationship between companies in question and importantly the stage of development of data economy. Thus, the compliant use of Web crawler gradually becomes a case-by-case issue, and the story can be much different when it comes to other jurisdictions.

WEB CRAWLER IN CHINA

According to Article 127 of the *PRC Civil Code*, «if the law has provisions on the protection of data and network virtual property, such provisions shall prevail.» However, there are no other laws and regulations directly related to the protection of enterprise data assets in China. Does it mean that web crawlers are outside the law in China?

In fact, the *PRC Anti-Unfair Competition Law* was already applied to cases involving Web crawlers and several companies have been found guilty of unfair competition (see Use case 3 below).

In practice, the existing civil cases show that the legal issues caused by web crawlers are mainly about unfair competition, and the court will protect mostly the party being illegally crawled. It is because big data, artificial intelligence, Internet of things and other technologies are still in the initial stage of development in China: there is a desire to protect the investments made by the players in this field.

Use case 3: In 2016, Baidu (Chinese search engine) was sued by Dazhong Dianping (consumers' review site) for illegal crawling basic information of merchants and the comments from consumers, which violated Robots Protocol, caused substantial replacement and brought more traffic to Baidu. The court ruled that Baidu's conduct constitutes unfair-competition and shall pay damages of 3.23 million RMB to Dazhong Dianping.

In addition, it is worth noting that the consequences of improper use of Web crawler can be more than the compensation of civil disputes. When it is used as a tool of crime, it may constitute the crime of illegal collection of computer information system data, the crime of destroying computer system and so on (see Use case 4 below). When the data collected by a Web crawler involves copyright and personal information issues, it may also constitute the crime of infringing on copyright and the crime of infringing on personal information of citizens. In such cases, the relevant persons in charge would be also held legally liable and punished. There is no doubt that such record will adversely affect the social credit of the company itself and the social credit of the relevant executives.

Use case 4: In 2017, the first Chinese criminal case happened between a start-up company Xiutao (engaging online video and ecommerce) and Toutiao (video platform from Bytedance Group). The interesting part is that two of the founders of Xiutao were from Toutiao. One of the executives was in charge of product and operation while the other one was in charge of technical department in Toutiao. Because Xiutao illegally crawled video data of Toutiao by user agent/avoidance of the anti-crawling technical measures implemented by Toutiao, Xiutao and several of its executives were sued for committing the criminal crime of illegally obtaining computer information system data. In this case, not only Xiutao was ruled to pay damages, the relevant executives in charge of Xiutao, were arrested and then sentenced to a fixed term of imprisonment.

THE FUTURE OF WEB CRAWLERS

At present, the Web crawler has become the basic tool used by big data enterprises. Compliant and reasonable use of this technology can help enterprises reduce costs and increase efficiency. However, many enterprises and individuals still think that «public data» can be crawled freely. Besides, when they

use Web crawler or receive data derivative products (such as industry / product reports) provided by a third party, they are likely to ignore the compliance issues involved in the use of Web crawler or other technical tools.

Therefore, in order to reduce legal risks, we suggest that enterprises and individuals seriously consider at least the following aspects before using Web crawler and/or other technology to collect data:

- Robots Protocol;
- Personal information and privacy protection with the principle of data minimization;
- Copyright protection;
- Competition relationship among parties;
- Whether the crawled data and the way of using such data comply with the principle of necessity.

For any additional information please contact:

ZHANG Beibei
Associate - Shanghai Office
beibeizhang@dsavocats.com

Isabelle DOYON
Lawyer- Shanghai Office
doyon@dsavocats.com