Journal of Commercial Law Review
2024, VOL. 8, NO. 3, 18-24
DOI: 10.47297/wspjclrWSP2516-249703.20240803

The historical development of web crawlers and the rules of anti-unfair competition law

ZiLin Qu

College of Law, China Jiliang University, Hangzhou 310000, Zhejiang, China

ABSTRACT
Since the origin of web crawler in 1993, it has been 30 years of development history. The convenience of its technology has a positive effect on improving the quality of public life and promoting the prosperity of market economy. But it also brings a series of new problems such as privacy violation, digital crime and so on. Among them, whether the web crawler behavior constitutes the unfair competition behavior that the competition law needs to regulate has become a hot topic in the legal circle in recent years. From the perspective of competition law, this paper combs and compares the history of judicial regulation of web crawler behavior both inside and outside the domain in detail, and analyzes the change of thinking from emphasizing data protection to emphasizing data circulation and the measurement of interests of operators, consumers and the market.

KEYWORDS
Data capture; Web crawler; Anti-unfair competition law; History of legal system

1 The basic principle of web crawler and its historical development

1.1 Definition of web crawler and its mode of operation

Web crawlers, also known as web bots and web spiders, are software tools used to automatically access and copy the content of web pages. They perform information collection and data mining tasks on the Internet, and are widely used in search engines, data analysis and information retrieval. The working principle of web crawlers is based on an automated process that is able to autonomously navigate the World Wide Web, discovering and gathering all kinds of valuable information.

The basic operation mechanism of web crawlers can be divided into several steps. First, the crawler will start with one or more initial web pages (torrent pages) and get all the links on that page by downloading the content of the page. These links are parsed and stored in a waiting queue. The crawler will then select a new link from the queue to access according to the predetermined strategy, and repeat the above process, constantly expanding its access range. In this process, the crawler can operate according to the set rules (such as only visiting a certain type of web page, complying with the robots.txt protocol, etc.) and stop conditions (such as visiting a certain number of web pages, reaching a specific depth, etc.) to ensure efficiency and legitimacy.

The application scenarios of web crawlers are very wide. Search engine is one of the most typical applications, such as Google, Bing and other search engines rely on web crawlers to capture a large number of web pages on the Internet, and then index these pages, so that users can quickly find relevant information when searching. E-commerce sites also use web crawlers to monitor

competitors' goods and prices and analyze market trends. Web crawlers also play an important role in academic research, social public opinion monitoring, network security and other fields. For example, researchers can collect large amounts of text data through web crawlers for natural language processing and machine learning experiments; Government agencies can use crawlers to monitor public sentiment on social media; Cybersecurity experts use crawlers to detect and track malicious websites and vulnerabilities.

The design and implementation of web crawlers need to consider several technical factors. The first is page parsing technology, crawlers need to be able to correctly identify and extract useful information in the web page, which usually involves HTML parsing, XPath, CSS selectors and other technologies. The second is queue management and scheduling strategy. Crawlers need to efficiently manage the queue of links to be accessed to avoid repeated access and falling into loops. Third, it is also necessary to deal with anti-crawling mechanisms. In order to protect their data, many websites will set anti-crawling measures, such as IP blocking, dynamic content loading, etc., and crawlers need to design corresponding strategies to deal with these challenges.

Web crawlers, while powerful, face a number of ethical and legal issues. When crawling data, the crawler must comply with the robots.txt protocol of the website, respect the access rules of the website, and avoid excessive burden on the server. Data collected by crawlers may also involve privacy and copyright issues, and relevant laws and regulations must be strictly observed.

1.2 History of web crawler technology and evolution of related legal issues

The history of Web crawlers began in 1993, when Matthew Gray of MIT created the first web crawler, the World Wide Web Wanderer. In the mid-1990s, web crawler became an important tool for search engine indexing. For example, WebCrawler, which came out in 1994, became the first engine that could index complete web content with crawler. The year 2000 saw significant advances in web crawler algorithms and technology, with Google's PageRank algorithm able to assess the importance of web pages based on the number and quality of links, and achieving 100 page downloads per second at its peak.[1] In the 2010s, with the development of big data and analysis technology, web crawlers began to be used in content aggregation and data mining, mining useful information from a large amount of data to provide decision support for decision makers. From the 2020s to the present, it has played a key role in artificial intelligence, machine learning, big data analysis and other fields, and its functions are no longer limited to search engines, but are also used in various online services such as market analysis, social media monitoring, automated content creation and so on.

With the development of technology and application fields, the negative impact of web crawlers has gradually expanded. In the early stage of the development of web crawlers in the 1990s, the network infrastructure was fragile, and a large number of crawler visits often led to excessive consumption of resources during the service period, and even inadvertently triggered similar denial of service (DoS) attacks, interrupting the normal operation of the server. At this stage, crawlers, although relatively primitive, began to push the boundaries of personal privacy, especially in the crawling of public forums and web posts. In the 21st century, with the rise of e-commerce and social media, web crawlers began to collect user data on a large scale, exacerbating the problem of privacy violations. At the same time, earlier unintentional Dos attacks, this stage also appeared a lot of bad crawlers using network security vulnerabilities to carry out malicious attacks, such as distributed denial of service (DDoS attacks). In addition, the theft of original content by crawlers has sparked disputes in the field of copyright and intellectual property, Notable cases include Healthcare Advocates, Inc. v. Harding Earle Follmer & Frailey (2007) in the United States and Zhejiang Fanyi Company v. Baidu (2006) in China for copyright infringement. As web crawler technology matures in

the 2010s, the legal issues it raises become more complex. The prevalence of personalized advertising and data analytics has exacerbated the problem of personal privacy violations, and crawlers can obtain information from multiple sources such as e-commerce platforms, human resources, educational websites, and third-party data. In the commercial sector, its impact on e-commerce websites and online services is increasingly prominent, including price manipulation, theft of market research data, and so on.

2 The double-sided effect of web crawler in modern society

With the increasing perfection of web crawler technology in the 2020s, its role in social development is becoming more and more important, but it usually presents positive and negative effects. The following mainly analyzes the positive and negative impacts of web crawlers in modern society, and discusses them from three perspectives: consumers, operators and society.

2.1 The positive impact of web crawlers

From a positive point of view, first of all, it "improves the user experience of consumers to a certain extent and enrichis consumers' choices." For example, travel websites can capture price data from different airlines and hotels to enable users to make better value choices. E-commerce platforms provide users with more personalized product recommendations based on users' browsing history, purchase history and preferences; In addition, the application of data capture technology in health care, education, employment and other areas of people's livelihood is also becoming more and more extensive, such as the epidemic data visualization platform that provides real-time information during the epidemic[2], and the recruitment information crawling that promotes the flow of information between job seekers and recruiters and reduces information asymmetry[3]. For the crawled operator, when the network products and services provided by the crawled party and the crawled party do not have a direct competition relationship, but constitute a "complementary", the transmission efficiency of data can attract more traffic for the crawled party[4]. In terms of social governance, web crawlers provide a real-time and effective source of data for political decision-making. For example, the price statistics department uses web crawlers to update the latest developments in the economy and society in real time, and uses big data technology to improve the compilation of price indexes. In addition, the greater value gained by data in circulation, sharing, and innovation has played a significant role in improving the efficiency of the market economy. This can also be seen from the fact that in recent years, courts have gradually included innovative business models, industry development efficiency and the impact on market competition order into the consideration of the legitimacy of crawling behavior.

2.2 Negative effects of web crawlers

In the contemporary highly developed information technology environment, the negative effects caused by web crawlers can not be ignored. For consumers, because it can collect a large amount of personal sensitive information indiscriminately, it leads to widespread privacy disclosure, user portrait generation and price discrimination. At the same time, the abuse of personal data collected by this technology even provides new tools for downstream crimes, such as the network routine loan where crawlers invade borrowers' mobile phones to obtain private information, and AI face change for fraud activities. Some companies also use special methods to avoid legal risk, such as credit reporting companies using data labels instead of showing the amount of tax paid by users to cover up their malicious crawling of "non-desensitized" information.

For operators, the impact of early crawling behavior leading to increased service period burden is relatively weakened, because now most websites have the ability to bear large-scale crawling behavior, only a few small websites with immature technology still have the problem of weak affordability. Correspondingly, the crawling value and probability of small websites are also low. The most sensitive part of the current operators to data capture is the adverse impact on their market position and economic efficiency. Because the Internet platform development model is based on the "two-sided market" paradigm (that is, the benefit of one participant joining the platform depends on the scale of the other participant joining the platform[5]), the value of the platform will increase with the increase of the number of users using the platform service or related products[6]. For example, in the case of Beijing WeboTV Technology Co., Ltd. v. Beijing Chuanggrui Culture Media Co., LTD., the court held that the behavior of Beijing Chuanggrui Company to grab short videos and user information of Beijing Webod Company's Tiktok APP weakened the competitive advantage of Webod company, caused substantial damage to its competitive interests, and constituted unfair competition.

If the web crawler is not timely regulated by the law, it will also have a great negative impact on the market economy. By free-riding, the climber makes use of others' business achievements, goodwill and other competitive advantages to seek competitive benefits for himself without making necessary input[7], which is easy to strike the innovation power of the climber and destroy the fair competition order of the market economy.

3 Judicial historical regulation of web crawler from the perspective of competition law

With the update and development of web crawler technology and the aggravation of its social harm, the regulation of web crawler in domestic and foreign judicial practices has gradually expanded from the field of competition law, copyright law and other civil and commercial laws to the field of criminal crime. Considering that the digital economy is the key factor that determines China's leading position in the global industrial competition, the "interest theory" attribute of the protection object of the anti-unfair competition law makes it an important means to protect the data rights and interests whose right boundaries are not clear at present[8]. The following is mainly from the perspective of competition law to sort out and compare the historical regulations of judicial on web crawler behavior inside and outside the domain.

3.1 From the EB case to the LinkedIn case: Evolving U.S. judicial attitudes toward data rights

In the context of analyzing the historical legal regulation of web crawler behavior in the United States, we observed from the eBay v. Bidder's Edge case (referred to as EB case) in 2000 to the hiQ Labs Inc. v. LinkedIn Corp. case in 2017. The time span of this case (referred to as the LinkedIn case) is a remarkable evolution of the attitude of the US courts to data rights. The EB case of 2000 marked the first US judicial intervention into web crawlers. In the case, auction data aggregator Edge used a crawler to collect data from online auction company eBay's website, which eBay argued constituted an intrusion into its property. Although Edge believed that the auction information on eBay was a public resource created by users and eBay had no right to restrict its crawling behavior, the court finally ruled that the content on eBay was private property and Edge's behavior constituted an interference in the possession of movable property.

However, in Intel v. Hamidi (2003), the California Supreme Court significantly reinterpreted the

application of the legal theory of "trespass to chattels," implying that unauthorized electronic communications (such as data crawling activity) should not be considered an infringement of chattels. Unless the use of the computer system is physically impaired or deprived. This decision significantly weakens the legitimacy of legal sanctions against data crawling in the EB case and provides greater freedom for data flow. A similar position was later upheld by a federal appeals court in White Buffalo Ventures.

The LinkedIn case in 2017 further demonstrated the changing attitudes of U. S. courts toward data access rights. In the case, hiQ, a human resources data analytics company, used crawlers to crawl LinkedIn users' public data, and when LinkedIn tried to stop hiQ from scraping the data, hiQ filed a lawsuit seeking legal relief. Although the case was settled after several rounds of injunctions were confirmed and lifted, in the last court decision, the court made it clear that hiQ's data crawling was not illegal under the CFAA (Computer Fraud and Abuse Act) for "unauthorized access" to data from public websites. Writing for the 9th Circuit, Judge Marsha Berzon said that "giving companies like LinkedIn 'free rein' over who can use public user data may create an 'information monopoly' that harms the public interest."

The changing attitudes of the United States towards data access rights in the digital age are reflected in the different positions of the judicial authorities in dealing with the EB and LinkedIn cases: from emphasizing the private nature of information resources to recognizing the public nature of data. In recent years, the rapid development of Internet technology and the digital trend of business models, as some scholars put it, "everything is being transformed into data, which means that opportunities are also adjusting in the direction of model innovation and efficiency change. " The realization of data value is in circulation, through data sharing to achieve the appreciation of production factors. This also constitutes the underlying logic of the US judicial attitude to data rights change.

3.2 From the thinking of judgment to the change of law: The development of the path of regulating web crawler in our country

In the regulation of web crawlers, Chinese laws have a similar development trend as the United States from emphasizing data protection to promoting data circulation, which is mainly reflected in the transformation of three judgment ideas in judicial practice and the adjustment of legal provisions at the legislative level.

In the early days, when dealing with web crawler cases, the court focused on the three-step analytical framework of determining competition, damage fact and causation to determine the infringement. Over time, courts began to recognize the inevitability of loss of competition in a market economy: "Competition means that one party's interests will eventually be harmed." At the level of substantive elements, most cases have begun to follow the model of "operator damage + violation of the principle of good faith and business ethics". In recent years, some courts have begun to introduce broader considerations such as consumer rights, social public interests, and the efficiency of the market economy into their judgments. Some scholars believe that this also has a causal relationship with the changes of legal texts.

In addition, at the legislative level, China currently mainly regulates web crawlers according to the second general clause and the twelfth "Internet Article" of the "Anti-Unfair Competition Law" (hereinafter referred to as the "Anti-Law"). However, because the concept of "Internet clauses" is vague and unclear, the court is still inclined to judge according to the general provisions. However, the general provisions also have some problems such as weak pertinently and vague boundaries of business ethics, so many scholars believe that it is difficult to find a basis for crawling data from the specific provisions of the current anti-unfair competition law. To this end, the legislative department

in recent years is also actively exploring to improve the regulation of web crawler laws. The Interpretation of the Supreme People's Court on Several Issues concerning the Application of the Anti-Unfair Competition Law of the People's Republic of China (Draft for Comment) published on August 19, 2021 (hereinafter referred to as the Draft of the Anti-Law Interpretation) added two new provisions that can be identified as the "Internet clause" in the anti-law act (but it was not included in the final publication, Reflecting the complexity of the reptilian problem that requires deeper consideration). And the General Administration of Market Regulation announced on November 22, 2022, the "Anti-Unfair Competition Law of the People's Republic of China (Revised draft for Comment)" (hereinafter referred to as the new anti-Law opinion draft) added seven new Internet articles, of which there are 15 general Internet clauses, 18 commercial data clauses, and 21 comprehensive considerations related to web crawlers. These changes indicate the concern of the legislation on web crawlers and the possible direction of future legal adjustments.

Comparing the legal provisions or draft provisions in three different time periods, it can be found that the length of the content has increased significantly. In the setting of the anti-law framework, from the original anti-law "general provisions of competition law + Internet article" to the new anti-law draft more targeted "Internet general provisions + commercial data provisions + comprehensive consideration factors". Further analysis, we can observe that the essence behind the change of provisions lies in the relaxation of the web crawler regulation. First of all, the "principle of good faith + business ethics" in the current article 2 of the Antilaw interpretation has become one of the elements of the illegality of the conduct in the draft of the antilaw interpretation opinion and one of the elements and comprehensive considerations of the commercial data provisions of the new antilaw opinion. This change may help solve the problem that the high degree of abstraction and uncertainty of business ethics in the Internet field may excessively expand the discretion of judges and even hinder active competition[9]. Secondly, the subject of consideration has expanded from the original operator to consumers and the market, showing the diversity and inclusiveness of interest measurement. In addition, the anti-French Internet clause "impedes, destroys... The provision of "normal operation", the interpretation of the three elements of "unable to operate" in the new anti-law opinion draft and the two preconditions of limiting "substantive substitution" indicate that the legislature sees the problem of low application rate of the current law due to the lack of clarity in the description and the lack of consistency in the new behavior. At the same time, the boundary of commercial data is strictly divided, requiring the party to take corresponding technical management measures, and emphasizing its responsibility of self-protection. Finally, the non-discrimination principle and the impact on technological innovation introduced into the comprehensive consideration of the new anti-law draft also increase the possibility of defense of the creeper to a certain extent.

Generally speaking, in the two major fields of justice and legislation, China's management strategy for web crawlers is gradually showing a trend of easing. This change is reflected in a gradual shift from an initial reliance on more vague "integrity and business ethics" guidelines and protection around a "single operating entity" to a more integrated approach. This new approach not only encompasses multiple stakeholders and domains, but also takes into account a diversity of values. This transformation is more in line with the characteristics and needs of China's socialist market economy in the era of big data.

4 Conclusion

This paper reviews the development history of web crawler technology, reproduces its dual influence in different historical periods, and further discusses the evolution of attitudes towards

data crawling behavior in judicial practice at home and abroad. Faced with the increasingly complex problem of new web crawlers, the legislature is also actively promoting the formulation of a new Anti-Unfair Competition Law. However, the alienation of the web crawler + model, the concealment of technical means and the combination with downstream crimes have also brought greater regulatory difficulty and legal challenges. After the introduction of the new law, the court's control of the boundary of legal articles, the application relationship with other departmental laws, and how to better balance the interests of operators, consumers, and the market are all areas that deserve further attention and deliberation in the future.

References

[1] Seyed M. Mirtaheriet al., A Brief History of Web Crawlers 11-12 (New York: Marting's Press2013).

[2] Xiangfei Li, Ke Han. Visualization analysis of epidemic data based on Flask framework [J]. Computer Age,2021(12):60-63+68.

[3] Fang Wang. Information crawling and data Analysis of recruitment website based on Python [J]. Information Technology and Network Security,2019,38(08):42-46+57.

[4] Bing Chen. Protection and Competition: the function realization of competition law governing data crawling behavior [J]. Forum on Political Science and Law, 201,39(06):18-28.

[5] Mark Armstrong "Competition in Two-Sided Markets",The Rand Journal of Economics,Vol. 37,2006,pp. 668-691.

[6] Katz Michael&Carl Shapiro. Network Externality, Competition and Compatibility[J].American Economic Review,1985 (75):424-440.

[7] Xiping Zhou. Application of "Free rider" theory in commercial logo protection: from key words Unfair competition cases [J]. Law,2017(05):126-138.

[8] Jin Sun, Tao Feng. Judicial Review of unfair competition Disputes related to data capture in digital Age [J]. Journal of Applicable Law,2022(06):112-120.

[9] Zhaoyang Li. Reflection and revision of data capture regulation from the perspective of Anti-Unfair Competition Law [J]. Journal of Jinan (Philosophical and Social Sciences Edition), 21,43(06):65-76.