

Technical Types of Data Crawling and its Legal Regulation

Haohan Lou, Meihong Ji*, Ruoxuan Weng, Shuting Liu

Jiangxi University of Finance and Economics, Nanchang, Jiangxi, 330000, China

*Email: 513457487@qq.com

Abstract

The progress of network crawling technology has given birth to data crawling technology in different links. Economic development and technological progress have led to frequent data disputes, which involve many legal issues such as the protection of citizens' personal information and the definition of ownership of network data. This paper combines the analysis of technology types and application links with the scene analysis, that is, the vertical technical penetration and the horizontal typical scene analysis. It focuses on the analysis of technology types in three links: data environment entry, data acquisition and collection, and data utilization, as well as the regulatory thinking of enterprise data and personal information crawling, and puts forward the path of reasonable government data disclosure and enterprise data rights. Horizontal coordination, rational demarcation and multi-dimensional co-governance.

Keywords

Data Crawling; Personal Information; Data Rights; Dual Dimension; Horizontal and Vertical Coordination.

1. Introduction

The arrival of the digital age means great changes in a series of links such as data storage, circulation and acquisition. Data, whose significance has gone far beyond mere recording or calculation, has become an increasingly important factor of production; The explosive growth of information has created conditions for the efficient development of the information age. However, on the other hand, from the national defense and data security between countries to the business data of every enterprise and even the information and privacy of every citizen, the identification, acquisition and management of these data are also facing great challenges.

2. Proposing the Problem

With the development of big data technology and digital economy, new data circulation mechanisms have emerged, including data transaction based on contracts and automatic data extraction based on crawlers. The game between data crawling and anti-crawling reflects the fierce competition and rapid development of digital economic entities under the guidance of digital technology, and also gives birth to a series of complex data disputes, and its legal regulation is gradually complicated. What is the relationship between the technical type of data crawling and legal regulation? How to divide the legal boundary of data crawling? Is there a more systematic and three-dimensional basis and thinking?

3. Technical Types and Application Links of Data Crawling

Data crawling technology is a common network information search and acquisition technology in the information age. At present, the types of data crawling technologies used in judicial adjudication and market competition are complex and diverse, but they are always included in

three links: data environment entry, data acquisition and collection, and data utilization. They can be classified and summarized so as to better summarize the laws of technology application. [3].

The key to data entry is the knowledge and consent of the data provider. If the actor bypasses or breaks through the protection barrier set by the system through technical means without authorization, and illegally obtains the computer information data of ordinary users, it constitutes the crime of illegally obtaining computer information system data. [2] In the case of unfair competition dispute between Shenzhen Gumi Technology Co., Ltd. and Wuhan Yuanguang Technology Co., Ltd.[14], Yuanguang Company prevents the ban by constantly changing the IP address of the crawler. Long-term and high-frequency access will overload the target server, and being blocked by IP means that a certain IP has exceeded the normalized access frequency. Continued access or the anti-crawling strategy of setting proxy IP will only have a negative impact on the normal operation of the target server, so it should bear corresponding legal responsibilities. However, the functions and requirements of different servers are different, and some web pages have demand for "page views" and "visits". A proper amount of crawler access frequency is beneficial to the operation of the target enterprise, which is superior to the competition, so IP ban will not be adopted. At this time, the pressure brought by web crawler to the server cannot be evaluated by the limit of IP ban, and the protection principle of maximizing benefits is emphasized, that is, as long as the sum of the benefits brought by crawling behavior to the crawling party and the crawled party is greater than the crawler access frequency. [13].

In the process of data acquisition and collection, unlike traditional web crawlers, the running process of focused crawlers is more complicated than that of traditional crawlers. Through a certain webpage analysis algorithm, relevant links are screened and put into the URL queue waiting for crawling, which makes the crawling results more targeted, improves the crawling efficiency and reduces the resource consumption. The focused crawler technology with built-in search engine used in "copyright infringement case of Li Jinbo, etc." is a focused crawler that crawls related web resources directionally. The search engine technology is used to automatically convert the internet web page format on the computer side, and the transcoding technology is used to solve the congenital defects of mobile terminals such as mobile phones.

In the process of data utilization, it focuses on the technical behavior of data crawlers in data utilization. In the case of "copyright infringement crimes such as Beijing Dingyue Literature Information Technology Co., Ltd.", Beijing Dingyue Literature Information Technology Co., Ltd. crawled the genuine e-books without the permission of right companies such as Palm Reading Technology Company and Beijing Fantasy Vertical and Horizontal Company, and then displayed them in more than 10 apps promoted and operated for others to visit and download and read, and made profits through advertising revenue and paid reading, which reflected the legal judgment of data crawling technology in data utilization [1].

4. Legal Analysis of Crawling Application Object Scenarios

4.1. Government Information Disclosure and Personal Information Protection[9]

As an important data subject to obtain personal information, the government's administrative management, data disclosure, and convenience for the public are closely related to the protection of personal information. In the context of market competition, i.e. the platform economy, the collection and use of personal information in various forms is more often expressed in terms of informed consent of individuals and the introduction of the principle of proportionality. In the scenario of government information disclosure, the legality of data crawling is more closely related to the government's management status and webpage settings.

Individuals have less freedom and flexibility in the government's collection of information, and the government's administration requires citizens to provide accurate and comprehensive personal information to the government, so the government should take more responsibility in the disclosure of information. [11].

By crawling a government information disclosure website, it was found that when information about citizens listed in the default list was made public, ordinary users entered the webpage with the necessary and reasonable privacy treatment of the ID numbers in the list information for the protection of citizens' personal information, but the persons concerned were given the full ID numbers after crawling the webpage information using web crawler technology. Does this mean that the crawling of personal information from government information disclosure websites is legal and reasonable? not. There is no privilege in data crawling, and government information disclosure should be open to all social groups, including ordinary users and crawlers. [10] In reality, the phenomenon of "double standards" is still common, and data crawling and its regulation have not received sufficient attention in government information disclosure. There is a certain degree of negligence in government information disclosure and management, and a lack of responsibility for the protection of citizens' personal information.[5]. The regulation of data crawling in the field of personal protection needs to balance the conflict between the government and the public interest it represents, the interests of enterprises, and the interests of individuals. When the information involved in privacy is not related to the public interest, the right to privacy should be protected from infringement in the process of government information disclosure, and the public interest and personal interest should be reasonably delineated with the help of the necessary interest weighing mechanism, to maximize the protection of privacy in government information disclosure.

4.1.1. The Path to Enterprise Data Rightsizing

The attribution of enterprise data is the core issue in data disputes such as data crawling between enterprises. Most of the courts in China deal with enterprise data disputes in the evaluation system of "unfair competition", and this practice tends to be rooted in the avoidance of the issue of enterprise data attribution and the vacancy in the field of data legislation, which not only cannot fundamentally determine the points and stop the disputes but also will make the resolution of data disputes fall into the vicious circle of "fault" definition. "This trend in practice stems from the avoidance of the issue of ownership of enterprise data and the absence of data legislation [5].

Data typing is the premise and key to the issue of data ownership. The author believes that the typing of data in the Internet economy should be based on the platform operation and data generation mechanism, relying on big data algorithms and combining the principle of proportionality to achieve the measurement of enterprise data rights. Data rights are different from data empowerment, and cannot be understood as mechanically setting rights, because it is difficult to clearly define the content of specific data rights, and different subjects are prone to conflict over the same data rights in different scenarios, so the rights are defined dynamically, i.e. the degree of legitimacy of data attribution is measured [6].

4.2. Enterprise Data Crawling Dispute Adjudication Standards

Taking Internet e-commerce platforms as an example, the algorithmic strategy of mainstream e-commerce platforms lies in: collection - processing, analysis - production, upgrading, and pushing, with the generation mechanism starting with algorithmic collection followed by algorithmic creation. The former includes company-owned data and users' personal information, while the latter includes users' usage information, i.e. digital traces and data products [7].

4.2.1. Algorithmic Collection Process

The enterprise's data i.e. enterprise name, operating costs, and other initial data belong to the enterprise data property rights; the operation of the Internet e-commerce platform cannot be separated from the huge database, and the user's personal information is obtained through the User Service Agreement. The data algorithm characteristics of the process determine the data disputes to be dominated by the level of big data subjects and personal information protection, concentrating on the user's privacy, enterprise competition rights and interests, and public interests. At this time, data sensitivity is strong, intellectual property properties are weak, and data collection has not substantially changed the properties of the original data, therefore, except after signing the Developer Agreement and authorizing third-party subjects to use part of the basic data, the data should not have a large public nature, so as not to violate the user's personality rights and interests.

4.2.2. The Algorithm Creation Process

The analysis of data and the innovation of data products in the big data environment are crucial for the development and growth of e-commerce platforms. The path of platform progress is mined in the massive user browsing data and comment statements, which includes the similarity calculation of comment statements and the distributed execution efficiency and clustering analysis of relevant clustering algorithms, product attribute word identification, generation of evaluation summaries, data analysis models and website update strategies, etc. Through technical means to mine the useful information in the massive information and realize the innovation based on algorithm collection, thus expanding the consumer base of e-commerce. [8] In the algorithm creation stage, the fruits of the data subject's labor make the attributes and institutions of the data transform, and the personality rights and sensitivity of the data diminish or even disappear, replaced by the attributes of the new data products, i.e. the property rights attributes of the data subject.

4.2.3. Differences in Adjudication Standards

Distinguishing the objects of data disputes according to the algorithm generation segment, which embodies distinct characteristic attributes in the processes of collection and creation, the two types of data disputes mentioned above should be measured with the principle of proportionality for the rightsizing of corporate data. The principle of proportionality emphasizes appropriateness, balance, and necessity, and requires different weightings in different processes. At the level of algorithm creation, enterprises have gradually developed some tools to help delineate boundaries in the course of competitive games, such as the "Robots Agreement", but due to the large autonomy and lack of legislation, there is legal uncertainty, and it is necessary to combine the above classification and proportionality principle with specific scenarios to reasonably delineate boundaries.

For example, in the Taobao v. Fairview case [Hangzhou Railway Transport Court (2017) Zhe 8601 Min Chu No. 4034 Judgment, Hangzhou Intermediate People's Court (2018) Zhe 01 Min Final 7312 Judgment...] The "data product" in the case belongs to the algorithm creation process, which can firstly be tilted towards the property competition rights of the enterprise platform, focusing on the use of data by third parties to achieve a balanced legal regulation of data products. Secondly, the case involves two types of rights and interests of the Internet platform on the use of data by users, i.e. the right to use digital trace resources and the property rights and interests of the e-commerce platform, and it should be clear that the protection of users' personal information contains a higher degree of personality attributes than property attributes, giving priority to the protection of platform users' data security, and thus achieving a balance of the need for legal protection [4,15].

5. The Relationship between Data Crawling Technology Analysis and Legal Scenario Analysis

Purely technical analysis or a combination of law and technology without scenario-based application does not provide a comprehensive and general solution to this problem. The legal regulation of data crawling is the result of both a legal and technical orientation and is the product of both a technical and a scenario-based analysis. The analysis of "technology + law" and "scenario-based + law" should occupy the same weight in the consideration of legal risks, among which the problem of data crawling is ultimately a problem arising from technological development, and therefore the technical analysis runs through the legal regulation of it, regardless of the in contrast, the scenario-based analysis focuses on horizontal comparison and contrast, extracting features for different industries, fields and data types, taking into account the specificity of the scenario, to achieve the horizontal extension of legal regulation.

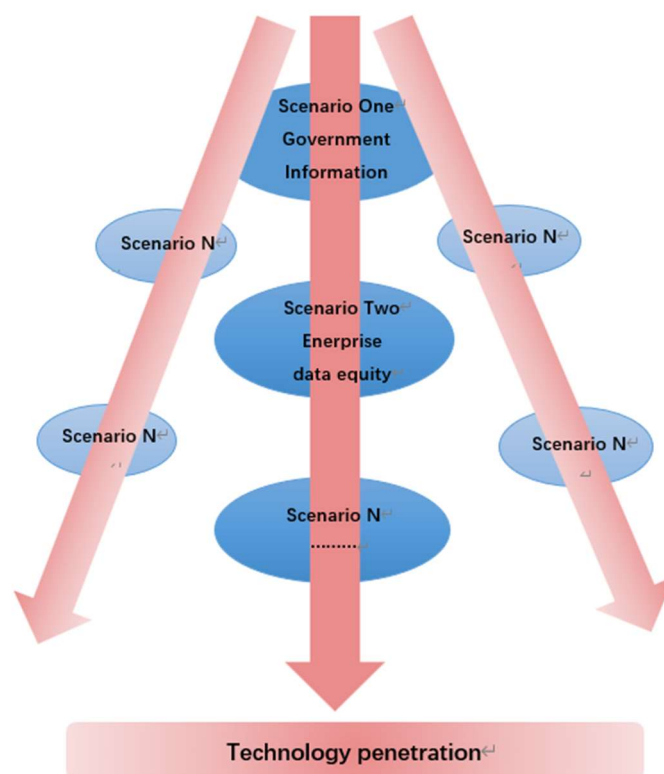


Figure 1. Relationship structure diagram of data crawling technology analysis and legal scenario analysis

6. Reflection on the Border between Technical Types of Web Crawlers and Legal Regulation

In summary, from the horizontal dimension, the characteristics of the industry sector are finely analyzed from a scenario-by-scene consideration in the scenario turn, the organic combination of scene and law is realized; vertically, the integration of technical analysis and technical types is insisted on throughout, realizing the three links from the refinement of the technology applied in specific scenes to the entry of data environment, data acquisition and integration, and data utilization, thus realizing the organic combination of technical analysis and legal analysis. At this point, a more complete regulatory chain has been formed in both horizontal and vertical dimensions, achieving a legal regulatory idea of technology verticalization,

scenario horizontalization, vertical and horizontal coordination, and interlacing and balancing. Although this idea is not enough to achieve specific weighing calculations to provide a clear basis for delineating the boundaries of more specific cases, it can provide a dual-dimensional direction for thinking about the delineation of the legal boundaries of web crawling technology.

Acknowledgments

The funding support project of the 2022 College Students' Innovation and Entrepreneurship Training Plan, "the types of data crawling technologies and their legal regulations" (202210421121).

References

- [1] Yan Huizhen. Design and implementation of news propaganda information system based on data crawling [D]. North Central University, 2021. DOI:10.27470/d.cnki.ghbgc.2021.000936; [1] Chen Bing. Protection and competition: The realization of competition law functions in governing data crawling behavior[J]. Political and legal forum, 2021, 39(06): 18-28.
- [2] Li Chan. Type distinction and governance model of malicious data crawling crime[J]. Crime Research, 2022(01): 79-88.
- [3] Zhai Wei, Liu Yino. The boundary of legality of public data crawling from the perspective of anti-unfair competition law [J]. Price theory and practice, 2021(08): 69-73+149.
- [4] Permission. The legitimacy of data crawling and its boundaries [J]. Chinese jurisprudence, 2021(02): 166-188.
- [5] Zhang Shuqing. Footprints and roads: Personal information protection and attribution of big data rights and interests[J]. Electronic intellectual property, 2018(11): 12-22.
- [6] Ji Leilei. The judicial dilemma of enterprise data protection and the dimension of breaking the game: the road to typological confirmation of rights [J]. Law Forum, 2022, 37(03): 109-121.
- [7] Tan Wenbin. Big data acquisition and application of product reviews based on e-commerce platform [D]. Guizhou Normal University, 2016.
- [8] Wang Binyang. Research on the innovation path of business model of e-commerce enterprises under the influence of big data [D]. Northeast University of Finance and Economics, 2016.
- [9] Yang Yiping, Chen J. Protection of personal information in government information disclosure in the era of big data [J]. Journal of Beijing Industrial Vocational Technology College, 2022, 21(01): 98-102.
- [10] Kuang Yuting. Legal regulation of web crawlers under open government data [J]. Journal of the Party School of Xiamen SAR, 2019(06): 73-77.
- [11] Data security is the lifeline of digital government. <http://www.northnews.cn/news/3265/0614/2003543.html>.
- [12] Typical cases of public interest litigation on personal information protection by procuratorial organs. https://www.spp.gov.cn/spp/xwfbh/dxal/202104/t20210422_517106.shtml.
- [13] Jiang Ying. On the Types and Limits of Criminal Law Involvement in Network Data Crawling[J]. Zhejiang Social Science, 2021(10): 37-45+156. DOI:10.14167/j.zjss.2021.10.005.
- [14] Civil judgment (2017) Yue 03 min Chu No. 822 of Shenzhen intermediate people's court, Guangdong Province.
- [15] Hangzhou Railway Transportation Court (2017) zhe 8601 min Chu No. 4034 judgment, Hangzhou intermediate people's Court (2018) zhe 01 min Zhong No. 7312 judgment.