International Journal of Science and Engineering Applications
Volume 12-Issue 08, 256 – 258, 2023, ISSN:- 2319 - 7560
DOI: 10.7753/IJSEA1208.1079

Multi-Dimensional Protection Law Adaptive Matching Algorithm for Citizens' Personal Information Under the Background of Intelligent Crawling Data

Zhaobin Pei
School of Marine Law and Humanities
Dalian Ocean University
Dalian, Liaoning, 116023, China

Abstract:The legal construction of the protection of personal information rights in the era of big data should focus on the core issues of who will protect and how to protect them. The regulation of responsible subjects should change from identifiable to risk controllability, and from fragmentation to systematization in legislative norms. In terms of multiple protection, the supervision, self-discipline and management are promoted in parallel, and the litigation jurisdiction spans the cyberspace and the real society, so as to realize the effective protection and rational use of personal information. In order to accurately analyze how different users like different products, solve the problem of reasonably recommending products to different users, and improve the accuracy of the recommendation algorithm and the conversion rate of advertisement placement, a web crawler technology using Python is designed to crawl massive advertising data information, It simulates the behavior of users clicking on advertisements, generates tag data, and implements an intelligent advertisement recommendation process combined with collaborative filtering algorithms.

Keywords: Intelligent Crawler, Multi-Dimensional Protection of Information, Matching Algorithm

1. INTRODUCTION

With the vigorous development of mobile Internet technology and big data technology, the means and degree of personal [1] information collection and utilization have been continuously upgraded. While enjoying convenient services such as user-customized services and market-oriented and accurate promotion [2] brought by big data, people also enjoy personal information. Experience the danger of personal information being leaked and violated. In the era of big data, personal information is particularly important. According to the "Opinions of the Central Committee of the Communist Party of China and the State [3] Council on Building a More Complete System and Mechanism for Market-Based Allocation of Factors" published in April 2020, data is included in the category of factors of production for the first time [4].

By the end of 2020, the number of netizens in my country has reached 989 million, and the Internet penetration rate has increased by 70%. [5] At present, the leakage of citizens' personal information based on the application of big data is becoming more and more serious, and the resulting illegal and criminal phenomena such as text message harassment, privacy leakage, violent debt forcing [6], and telecommunication fraud are emerging one after another, which not only affects the normal life of citizens [7]. With the rapid development of the Internet and the advent of the era of big data, the production and demand for data in all walks of life are rapidly increasing. How to efficiently collect, analyze and sort out the data information of interest from the massive data has become an indispensable part of our lives [8].

With the rapid development of computer technology, people have done a [9] lot of research work in the literature analysis system in the past ten years. Wang Yuefen et al. developed a software for statistical analysis of journal publication volume and keywords [10]. Zhang Mannian and others put forward the idea of constructing the evaluation and analysis system of scientific journals. Jiang Chunlin, etc. As the core content of the search engine system, web crawler can directly retrieve and process information for [11] the underlying organization in the Internet system and can directly affect the update of relevant data and information in the Internet system based on the fundamental level [12]. We need to look at the impact of big data, cloud computing, WeChat, smartphones, etc. from a new perspective. [13] In the information society, the protection of personal information needs to be regulated by law, and the advent of the Internet era has greatly increased the importance of data and network virtual property [14].

Although my country's protection of personal information has been [15] continuously enhanced, information violations have occurred from time to time. This reflects that my country's personal information [16] protection model still has problems such as outdated concepts and lagging systems. How to balance the relationship between information utilization and information protection has become one of the biggest legal problems at present. Establish a complete legal protection system and increase punishment for illegal and criminal acts. As far as the current legal system is concerned, in addition to the general provisions of the [17] Civil Law, the relevant implementation rules should be issued as soon as possible. Web crawlers can be simply divided into two categories: general crawlers and topic crawlers (focused crawlers). Universal crawler is to crawl all web page information on the Internet through search engines, such as Baidu, Google, etc [18].

Literature analysis can be divided into two levels: macro-level analysis and micro-level analysis: at the macro level, the system helps users understand the research overview under the retrieval theme; at the micro level, the system helps users understand the specific research content under the retrieval theme. We describe the research overview from the four perspectives of document time distribution, author distribution, subject distribution and keyword distribution. Web crawler technology is also called web spider technology, or it can be called web robot. The web crawler technology is mainly that the search engine efficiently downloads the

relevant web page information through the World Wide Web, and further collects information from the corresponding network system along the web page link.

2. THE PROPOSED METHODOLOGY

2.1 The Intelligent Crawler

The second is improper collection by network service providers or operators on social platforms. Such collection methods are more concealed, and they usually take advantage of their technological advantages to induce customers to provide unnecessary personal information without ordinary people noticing. The third is improper collection of software developers or software providers on the technology application platform. The rational man hypothesis originates from modern humanistic thought. the middle of the last century, humanism emerged in the United States. It takes people as the ultimate goal, emphasizes the dignity and value of people, allows people to express their thoughts and emotions freely, and realizes people's control over themselves.

Therefore, the right to self-determination of personal information is given to individuals through the means of information confirmation, so as to realize the control of personal information. It is recommended to set up a third-party agency to rate the company's personal information protection measures and regulatory measures, and make the rating results public, so that the public can know the company's protection of citizens' personal information. The design goal of the topic crawler is to collect topic-related web pages. The determination of the topic is to extract the representative topic feature items from the Internet, and then perform the topic similarity calculation with the web page to complete the calculation of the topic relevance of the web page.

In order to improve the accuracy of data acquisition. It can be seen from Table 1 that the literature quality evaluation index system we designed comprehensively considers the influence of literature, the influence of literature authors, the influence of literature publications, the influence of literature references and the influence of literature citations. The first-level index in the table is the average of its corresponding second-level index.

2.2 The Multi-Dimensional Protection of Information

In order to eliminate dimensional differences, all secondary indicators need to be standardized. The search engine system supported by web crawler technology is a relatively common information retrieval method and data query tool, which provides greater convenience for people's various network experiences in the new era. In the context of modernization, Internet technology continues to innovate and develop, and data storage forms show diverse characteristics. There are many types of personal information illegally provided in reality, including document information, address information, credit information, transaction information, as well as track information, health and physiological information, etc. All kinds of information” ① are at risk of being provided illegally, among which there are four serious forms of personal information being provided illegally.

The first is the situation where the information provided whereabouts is used by others as a crime, or where personal information is provided knowingly that it is used to commit a crime. Article 14 of my country's Personal Information Protection Law stipulates the principle of informed consent. This principle places the information subject at the center of information decision-making, assumes that individuals can manage personal information independently and rationally, and makes a structural allocation of information risks based on this principle, and the responsibility for information protection belongs to individuals. Enterprises protect citizens' personal information. Law enforcement agencies can also rely on legal provisions to supervise and inspect the protection measures for citizens' personal information formulated by enterprises.

2.3 The Matching Algorithm

Through external legal supervision, strengthen corporate self-discipline and allocate responsibility for the protection of citizens' personal information to each enterprise. In this paper, Taobao is used as the material website, the keyword collection is used to determine the theme, and the collection of more than 20,000 advertisement data is used as the basic data to achieve the coverage of most tag categories. In addition to picking the initial seed. The preprocessed literature data will flow through the analyzer module, which consists of three sub-modules. The literature overview sub-module is responsible for generating macro-level analysis results, the literature recommendation sub-module is responsible for generating micro-level analysis results, and the report generation sub-module is used to generate Literature analysis report. Under the web crawler technology, according to the initial URL target sequence, various information and related links in the network can be selectively explored and accessed, and the required information and data can be obtained smoothly.

In the process of applying web crawler technology to implement information capture, personal information is illegally used, which actually exceeds the reasonable use limit of "the law requires information rights holders to tolerate minor harm from the use of personal information by others" ②. Common situations include purchasing personal information for business promotion, and using personal information to register a Taobao account to perform “Taobao swiping”. The exclusive domination of information reflects the current information protection model's position of resolutely safeguarding personal dignity and personal freedom, and greatly promotes the economic and social development in the era of small data. After the occurrence of "after the sheep"-style supervision. For example, a courier company launched the "Feng Dian Slip", customers can fill in the sending and receiving information by scanning the QR code of the shipping slip, without filling in personal information on the shipping slip.

The intelligent advertisement recommendation system designed in this paper needs to analyze the HTML page text of Taobao.com, and uses the Beautiful Soup package of Python. The web page is parsed into a tag tree composed of many nodes, and various useful information is extracted on the basis of web page tags. All modules of the system are implemented by Python. Among them, the crawler module mainly uses the third-party asynchronous network request library aiohttp, the multi-process standard library multiprocessing and the third-party parsing library lxml of HTML pages. With the help of multi-process asynchronous crawler, it can achieve a speed of 50-200 articles/second of document data acquisition. Web crawler technology is an important means and core tool for network information collection and retrieval. The design principles related to web crawler mainly include the following. Entering the era of big data, the data value of personal information is reflected in the

way of quantitative change and qualitative change. "Although a certain type of personal information that is packaged and processed is not for the purpose of identifying an individual, the typed treatment after categorization processing will also cause serious damage to individuals. Infringement of the rights of the information subject" ③. However, the protection, sanction capability, and strike radius of existing laws and judiciary are all inadequate, and high-tech features are ignored. Personal information infringement means that the infringer collects, processes or uses personal information without the informed consent of the information subject. The current protection model holds that the legal interests of information are personal interests, individuals are the best decision-makers for their own interests, and the infringer is in the same legal status as the information subject, so the information subject decides whether to file a lawsuit. In the era of big data, if the circulation of personal data is completely prohibited, it will hinder social development to a certain extent, and the imbalance between supply and demand may lead to more personal information trading. Therefore, at the same time as legislative protection, it is recommended to formulate the "permissible level of dissemination of personal information". After simple data cleaning, the specific information on the collected web pages can be stored on the local server in different ways. The storage methods used in this article mainly include MySQL database storage and Excel table storage. To add the collected information to the database. The literature analysis report given by the system consists of introduction, literature overview and literature recommendation.

3. CONCLUSIONS

The current personal information protection framework is premised on the assumption of rational persons, the principle of informed consent is the core, the right to information self-determination is the main content, and civil litigation is the remedy. The current private law protection path is in line with the characteristics of information individualization in the era of small data. The web crawler based on Python language has become the mainstream tool for crawling data on websites. The request library can be used to obtain the content of the webpage, and the obtained html text can be parsed through bs4, and then the user can be obtained by simple data cleaning and collection using re regular expressions Really needed information data.

4. REFERENCES

[1]Liu Jingyu, Dai Pengcheng, Luo Guanhong. Multi-dimensional construction of citizens' personal information protection system [J]. People's Procuratorate, 2018(17):1.

[2] Yao Ye. Multidimensional Interpretation and Selection: An Analysis of the Intellectual Property Protection Path of Artificial Intelligence Algorithms [J]. Science and Technology and Law (English and Chinese), 2022(1):9.

[3] Zhu Yongli, Song Shaoqun. Research on Adaptive Coordination Protection System Based on Wide Area Network and Multi-Agent [J]. Chinese Journal of Electrical Engineering, 2006, 26(16):15-20.

[4] Zheng Tianying. Analysis of personal privacy protection in the context of artificial intelligence [J]. Market Weekly, Theory Edition, 2018(14):1.

[5] Li Hong. Examination of the predicament and path exploration of the protection mechanism of personal information administrative law——Based on the empirical analysis of 452 administrative judgments.

[6] Wang Yao, Wang Shixin, Zhou Yi, et al. Research on Brightness Temperature Difference Corrected Fire Point Detection Method Based on GF-4 PMI Data [J]. Spectroscopy and Spectral Analysis, 2021, 41(11):3595-3601.

[7] Chen Yuchao, Chen Yuzhuo. On Personal Data Protection in the Age of Intelligent Algorithms [J]. Foreign Economic and Trade, 2022(2):6.

[8] DING Shengrong, MA Miao, GUO Min. Application of Artificial Fish Swarm Algorithm in Adaptive Image Enhancement [J]. Computer Engineering and Applications, 2012, 48(2):185-187.

[9] Zhang Suli. Analysis of legal protection of personal information under the background of big data [J]. Journal of Jilin Radio and Television University, 2018(7):2.

[10] Dai Shaoqing. Research on the crime of infringing citizens' personal information [D]. Zhongnan University of Economics and Law, 2019.

[11] Huang Hongjian. Research on personal information protection from the perspective of "General Principles of Civil Law".

[12] Liu Zhenling. Personal information protection from the perspective of information disclosure [D]. Central South University, 2010.

[13] Yang Zhen. Legislative Research on Personal Information Protection [D]. Jilin University of Finance and Economics.

[14] Chi Qiong. Jurisprudence Research on Personal Information Protection [D]. Liaoning Normal University.

[15] Liu Yang. my country's Personal Information Protection Legislation and Literature Research Review [J]. Human Resources Development, 2014(4X): 2.

[16] Yu Shaoru, Lei Gang. The use of personal information and its boundaries in targeted poverty alleviation [J]. Journal of Beijing Institute of Technology: Social Science Edition, 2022, 24(1):138-151.

[17] Zhao Rui. Analysis of the application of the seventh national census data processing platform [J]. Statistics and Consulting, 2022(1):42-43.

[18] Zhong Huajuan, Fang Lu. Research on the Protection Path of Citizens' Personal Information in Epidemic Prevention and Control [J]. Postgraduates of Zhongnan University of Economics and Law, 2021(6):109-114.