Review of data scraping and data mining research

Yuguang Wang, Dengyun Zhu, Bin Zhang, Qi Guo, Fucheng Wan*, Ning Ma

Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education Northwest Minzu University, Lanzhou, China

*Corresponding author e-mail: 185152565@xbmu.edu.cn

Abstract. In recent years, with the advent of the era of big data, the importance of data has also become more prominent. This article introduces the characteristics of today's Internet data from the background of big data, and the main method of data scraping-crawlers. In addition, the concepts of data mining, as well as hotspots and trends are briefly introduced. Highlights the importance of data scraping and data mining in today's Internet field.

Keywords: data scraping, data mining, big data, crawlers.

1. Introduction
With the development of Internet technology, the term big data has been mentioned more frequently. Artificial intelligence, the Internet, the Internet of Things, cloud computing, and our daily lives are generating more and more data, and data is already an essential part of our lives. In today's society, the application of big data is becoming more and more widespread and it continues to show its advantages in various fields. Such as e-commerce, O2O, logistics and other areas using big data for development. With the concept of big data, consumers can predict consumer behavior by obtaining massive amounts of data and then analyzing it.

In fact, for ordinary people, we may not feel the value of the data, and some scholars use the data to predict the market trend and obtain good results. It can be seen that big data is so important that related technologies such as data capture and mining have become important research topics.

2. Basic Characteristics of Internet Data
Data grabbing is a key step in obtaining big data information, but what are the characteristics of data on the Internet and what problems do you encounter during the actual crawling process? Therefore, it is necessary to analyze the characteristics of the data when analyzing the data on the Internet. In this way, there will be more targeted in the actual application process. The characteristics of the data are as follows.

2.1. Huge Amount of Data and Complex Types
With the development of the network, our understanding of data capacity is not limited to the KB and MB levels. As of 2012, the amount of data has jumped from the TB (1024GB = 1TB) level to the PB (1024TB = 1PB), EB (1024PB = 1EB) and even ZB (1024EB = 1ZB) levels. According to the research results of International Data Corporation (IDC), the amount of data generated globally in 2008 was
0.49ZB, in 2009 it was 0.8ZB, in 2010 it was 1.2ZB, and in 2011 it was 1.82ZB, which is equivalent to 5EB. According to IBM research, 90% of all data obtained by the entire human civilization are generated in the past two years. By 2020, the scale of data generated worldwide will reach 44 times today[1]. In the current network environment, the huge amount of data is a very significant feature. At the same time, the source of the data is also very complicated. According to the data source, it includes audio, video, pictures, and so on. Data division can also be divided according to data relationships, divided into structured, semi-structured and unstructured data. So if you want to extract useful information from such data, you need some technical means and related methods, such as crawler technology.

2.2. Data Reliability and Real-time
According to the characteristics of large amount of data and complex data types, we can know that the effectiveness and correctness of the data is particularly important when analyzing and fetching the data. If the data we analyze lacks reliability, the results obtained from our analysis may be inaccurate or even produce errors. As for the timeliness of data, we definitely hope to get the latest results through data analysis instead of outdated conclusions. Because the Internet environment is changing rapidly, the needs of users are changing at all times. Only by analyzing and grabbing real-time data can we truly reflect the actual situation of the user and the degree of preference for information. Such data analysis and grabbing is also the most effective [2].

3. Main Methods for Grabbing Data
Faced with the large amount of information on the Internet and the complexity of the data, crawlers targeting webpage information have emerged. A web crawler, also known as a web spider, is a program or script that automatically captures information on the World Wide Web in accordance with certain rules. The main workflow of the crawler includes: (1) imitating a browser to send a request through the URL of the web page (2) after obtaining the web page code, using a regular expression or BS to obtain a new URL or web data The mode output is saved. In general, a crawler mimics the process in which a browser sends a request to a server. The principle of the crawler is shown in Figure 1.

![Fig. 1 Schematic of the crawler](image)

Web crawlers can be roughly divided into general web crawlers, focusing on web crawlers [3]. A universal web crawler, also called a full web crawler, starts with one or more initial URLs, obtains the code for the initial page, and extracts the relevant URLs from the page and puts them in the queue until the stopping conditions of the program are met. Compared to general web crawlers, the workflow of focusing on web crawlers is more complicated. It needs to filter out some URLs that are not related to the topic through a certain web analysis algorithm in advance, and ensure that the remaining URLs are related to the topic to a certain extent. Put into the URL queue waiting to be crawled. Then according to the search strategy, select the URL to be crawled in the next step from the queue, and repeat the above
operations until the stopping conditions of the program are met. Focused web crawlers can crawl information that is more relevant to the topic, for example: in order to quickly obtain data in Weibo, you can use focused crawler technology to develop a tool to capture Weibo data [4, 5, 6].

In short, crawler technology is one of the main technologies applied in the Internet field.

4. Data Mining

In today's Internet environment, massive amounts of data provide us with great convenience but also bring certain negative effects, such as information overload and so on. The huge amount of information prevents us from refining accurate and effective information. Only by analyzing huge data and finding effective information hidden in it can we better obtain the convenience brought by the data. Due to the huge amount of data to be analyzed, traditional methods have been unable to meet the needs of today's data analysis, and it is difficult to achieve the expected results. Moreover, the data has the characteristics of timeliness. If the extraction process is too slow, it is often impossible to obtain the first-time data analysis results. The results can be imagined. At the same time, it cannot predict the future trends, and it is more difficult to dig deep Level of knowledge. Therefore, data mining technology came into being.

Because a large amount of data is stored in the database, data mining is also known as knowledge discovery of the database. There are many definitions of data mining. "Extracting from the data contains potentially useful information that was previously unknown and useful." Data mining related technologies have been vigorously applied in various fields, such as the analysis of DNA using data mining technology in biological research [7]. In short, data mining is a process of obtaining useful information from massive data. In recent years, data mining technology has attracted much attention from the industry. The data mining system prototype is shown in Figure 2.

![Figure 2. Data mining system prototype](image)

5. Hot Spots and Trends in Data Mining

With the development of the Internet, the huge amount of data in the Internet contains extremely rich value. Web mining refers to the application of data mining on the Web. It uses data mining to obtain useful information from the World Wide Web. It has been widely used in the fields of e-commerce, search engines and website design. Text mining refers to computer processing technology that obtains valuable information and knowledge from text data, mainly to obtain the relationship between text and grammar and sentences. Text mining is mainly applied in many aspects such as natural language processing and machine translation. In addition, there are aspects such as data flow mining and biological information data mining. Data mining has been widely used in the fields of finance, banking,
agriculture, manufacturing, retail, telecommunications, health care, education and biological sciences. The number of published data mining papers from 2014 to 2019 is shown in Table 1.

| Years | Number of documents |
|-------|---------------------|
| 2014  | 4686                |
| 2015  | 5033                |
| 2016  | 5870                |
| 2017  | 5478                |
| 2018  | 5582                |
| 2019  | 4551                |

Nowadays, the data mining technology has been developed for a long time and has become a hot topic at home and abroad. The future development trends in this field may have the following points:

1. Research on data mining algorithms. Although the data mining algorithm has made great progress after a long period of development, in the face of massive data, the efficiency and accuracy of the algorithm become more and more important. So algorithm research is still a subject worthy of study.

2. Data mining and privacy protection. The privacy and information security issues of data mining have attracted much attention. Misuse and abuse of data mining may lead to the leakage of user data, especially sensitive information. More and more people are worried about this. How to perform data mining without exposing user privacy will become a very interesting research topic [8].

3. Integration of data mining technology with other systems. For example, integration with database data warehouse system and integration with language model system. Only in this way can the data mining system be better applied to real life, complete the tasks of data mining outstandingly, and take advantage of data mining.

4. Preprocessing of large data. Data mining is obviously purposeful. In order to obtain better data processing results, the preliminary work of mining should be done well. With the emergence of a large amount of data, how to effectively and quickly do preprocessing needs further research.

5. Spatial and temporal data mining. Spatial databases are rich in data types. In addition, there is a temporal relationship between the data of a type of data set, which is called time series data. In the process of mining time series data, it must be considered that there is a temporal relationship between the data in the dataset [9].

6. Summary

In short, a web crawler is a program that automatically captures data. It obtains relevant information from the World Wide Web based on determined goals. Reptile technology has become one of the key technologies in the Internet field, so the level of reptile technology plays a vital role in the development of the Internet. Therefore, in the subsequent development process, we must pay attention to the development of crawler technology, pay attention to data capture and data analysis, comprehensively improve the data capture ability, and help the development of big data technology.

Data mining has received more and more attention from the academic community, and its development is of great significance to the development of Internet technology. It is also used more and more widely, such as in medical applications. But data mining is not a panacea. It is just an analysis method and tool. The results obtained need to be combined with the specific conditions of the industry, and sometimes with the relevant national laws and policies. Only by comprehensive analysis can we get correct data mining information.

In summary, data capture and data mining technologies are key technologies for us to obtain effective information in a big data environment. Only continuous development and improvement can be invincible in such an environment.
Acknowledgments
This work is supported by Natural Science Fund of GanSu Province (21JR1RA196) and the Fundamental Research Funds for the Central Universities (NO. 31920210020).

References
[1] Shi Xiaobo. Big Data: Seizing Opportunities and Preserving Value [N]. Science and Technology Daily, 2014-07-19 (002).
[2] Monday Cong. Application of reptile technology in the Internet field [J]. China Hi-Tech, 2019 (18): 64-65.
[3] Zhou Lizhu, Lin Ling. Review of Focused Reptile Technology Research [J]. Journal of Computer Applications, 2005 (09): 1965-1969.
[4] Zhou Zhonghua, Zhang Huiran, Xie Jiang. Python-based Sina Weibo data crawler [J]. Journal of Computer Applications, 2014, 34 (11): 3131-3134.
[5] Liu Jingjing. Research and Implementation of Weibo Oriented Web Crawler [D]. Fudan University, 2012.
[6] Wang Jing, Zhu Ke, Wang Binqiang. Summary of Weibo Research Based on Information Data Analysis [J]. Journal of Computer Applications, 2012, 32 (07): 2027-2029 + 2037.
[7] Zhu Yangyong, Xiong Yan. DNA sequence data mining technology [J]. Journal of Software, 2007 (11): 2766-2781.
[8] Qian Ping, Wu Meng. Overview of data mining methods for homomorphic encryption privacy protection [J]. Application Research of Computers, 2011, 28 (05): 1614-1617 + 1622.
[9] Jia Pengtao, He Huacan, Liu Li, Sun Tao. Survey of Time Series Data Mining [J]. Application Research of Computers, 2007 (11): 15-18 + 29.