Data Acquisition on Network Public Opinion of Emergencies Based on Web Crawler Technology

Yanmei Wang[a], Yongchang Ren[b]

College of Information Science and Technology, Bohai University, Jinzhou, 121013, China

[a]19296571@qq.com, [b]rycryc@sina.com

Keywords: big data; emergencies; network public opinion; data acquisition; Web crawler technology; data acquisition process

Abstract: Emergency network public opinion refers to the sum of opinions, feelings, and attitudes issued by cyber citizens. In the big data environment, rapid analysis of massive network data, establishment of public opinion monitoring and guidance mechanisms, and decision support for managers have become a new development direction. Data collection is an important task. Web crawler technology is a convenient and quick effective method to collect network data. Guided by the basic principles of web crawler technology, this paper analyzes the data collection process of emergency network lyrics, and uses the breadth-first algorithm to design the information acquisition process based on web crawler technology, which provides complete solution of data collection for emergency network public opinion.

1. Introduction

Emergency network public opinion refers to the sum of opinions, feelings, and attitudes issued by cyber citizens after an emergency. The application and development of big data has brought far-reaching influence to all aspects of society. The research on social public opinion under the big data scenario has become a hot topic for governments, enterprises and scientific research institutions. In the big data environment, rapid analysis of massive network data, establishment of public opinion monitoring and guidance mechanisms, and providing decision support for managers have become hot and difficult issues for research. Different from the traditional social public opinion analysis, the social public opinion analysis in the era of big data pays more attention to the collection, storage and cleaning of a large amount of network data, and combines text mining technology to obtain public opinion information from a large number of low-value density data.

The data acquisition system integrates signals, sensors, actuators, signal conditioning, data acquisition equipment and application software. In the era of data explosion, the types of data are complex and diverse, including structured data, semi-structured data, and unstructured data. Big data collection is the entry point for big data analysis and an important part of big data technology. This paper is based on the research of web crawler technology, and provides support for building an emergency network lyric platform in big data environment.

2. Basic principles of Web crawler technology

Web crawler is an automatic web page extraction program that downloads web pages from the World Wide Web for search engines and is an important component of search engines. Web crawlers can be divided into two categories according to their working methods: In the first place, centralized web crawlers. First, crawling the initial configured URL seed collection, which can collect the page html corresponding to the URLs through the crawler main program, send the html to the page content analysis module, extract the information, save the valid information, and obtain the new information in the page. The connection is pointed to and saved to the task URL collection for subsequent crawling. The collection efficiency of centralized web crawlers is relatively lower. In addition to simple services, it is unable to meet the business needs of large data collection. The

second is the distributed web crawler, which can be seen as a combination of multiple centralized web crawlers. A centralized web crawler is a node in a distributed system. Multiple network crawlers in the system work together, and each node can communicate with each other. The distributed network crawler can increase the collection efficiency of the network crawler in multiple times, because it is composed of a centralized network crawler node. As long as the information communication between the nodes is handled well, a crawler system with reasonable performance can be constructed.

The traditional crawler starts from the URL of one or several initial webpages and obtains the URL on the initial webpage. During the process of crawling the webpage, the new URL is continuously extracted from the current page into the queue until a certain stop condition of the system is satisfied. The workflow of focusing on crawlers is more complicated, and it is necessary to filter the links irrelevant to the topic according to certain webpage analysis algorithms, retain useful links and put them into the queue of URLs waiting to be crawled. Then, according to a certain search strategy, the URL of the webpage to be crawled next is selected from the queue, and the above process is repeated until a certain condition of the system is reached. In addition, all web pages crawled by the crawler will be stored by the system, analyzed and filtered, and indexed for later query and retrieval. For the focused crawler, the analysis results obtained by this process are also feedback and guidance may be given to future crawling processes. The basic principle of web crawler technology is shown in Fig. 1.

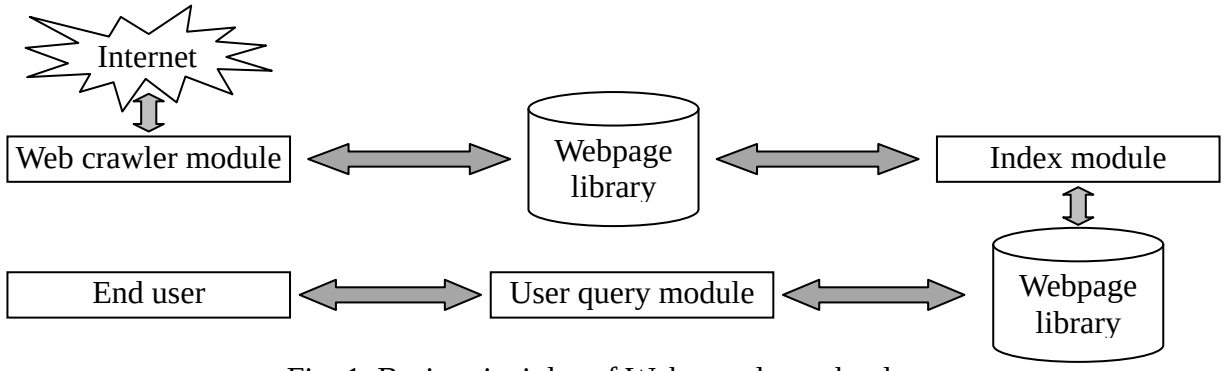

Fig. 1. Basic principles of Web crawler technology

3. Workflow of Web crawler technology

Web crawlers are an important part of the search engine crawling system. The main purpose of the crawler is to download web pages on the Internet to a local image to form a mirrored backup of the web content. The basic workflow of the web crawler is shown in Fig. 2.

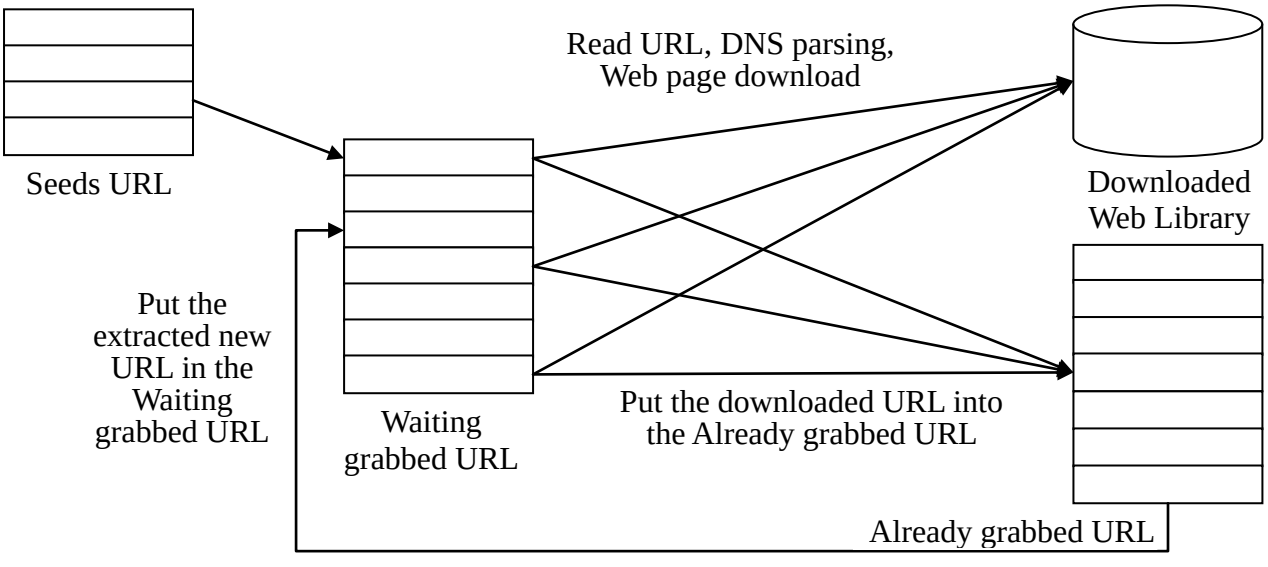

Fig. 2. Workflow of Web crawler technology

For the workflow shown in Fig. 2, it is divided into four steps, which are briefly described as follows: Step 1, select a part of the seed URL; Step 2, put the selected seed URL into the URL queue to be crawled; Step 3, from The URL to be crawled is retrieved from the URL queue to be crawled, the DNS is resolved, the IP of the host is obtained, and the webpage corresponding to the URL is downloaded and stored in the downloaded webpage library. In addition, the URLs are placed in the crawled URL queue. Step 4, analyze the other URLs in the content of the captured webpage, and put the URL into the URL queue to be crawled, so as to enter the next loop.

4. Data Acquisition Process of Emergencies Network Public Opinion

A site map is a container for all links to a website, and generates navigation web files based on the structure, framework, and content of the website. Many websites have deep connection levels, and crawlers are difficult to crawl. Site maps can easily crawl crawling websites. By crawling the website pages, the website structure is clearly understood. The website maps are generally stored in the root directory and named sitemap, guided the crawler and increased the inclusion of important content pages on the website. Sitemaps are good for improving the user experience, pointing the way for website visitors, and helping lost visitors find the page they want to see. The site-based emergency network data collection process is shown in Fig. 3.

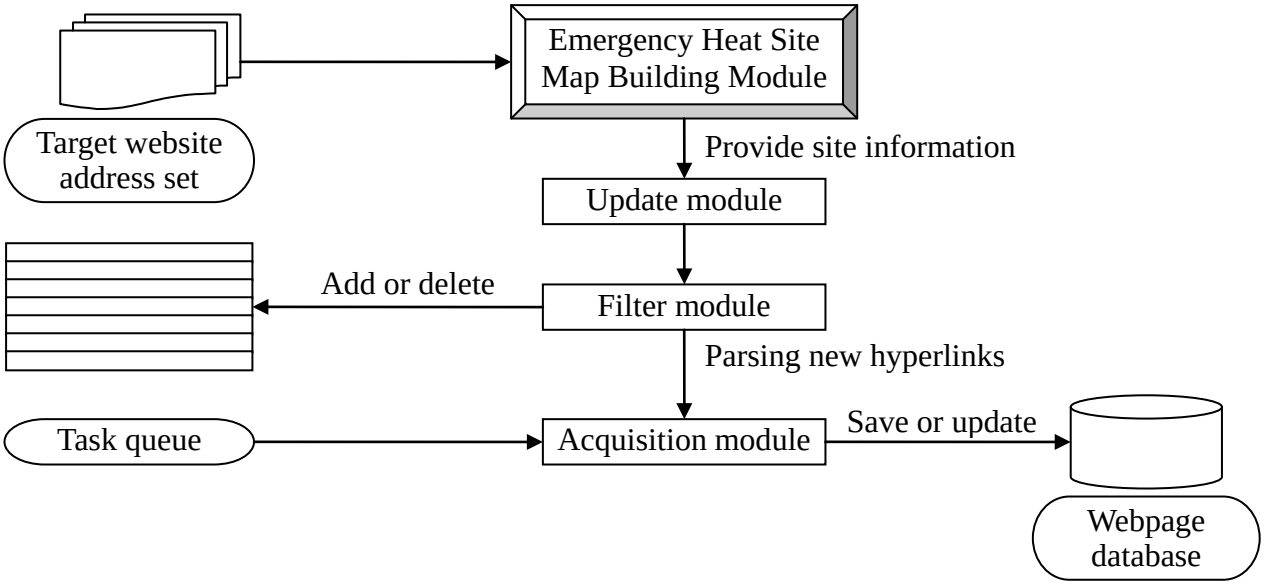

Fig. 3. Data acquisition process of emergencies network public opinion

For the data collection process shown in Fig. 3, it is divided into three steps: Step 1, using the emergency site map building module to construct a site map containing the hotspots of each section of the website; Step 2: Update the module and the filtering module updates and optimizes the task queue under the guidance of the site map; in step 3, the collection module sequentially takes the task from the task queue, collects the webpage and updates the webpage library, and simultaneously parses all the hyperlinks contained in the webpage, and delivers Follow-up acquisition of the filter module. The main module functions are briefly described as follows:

(1) Filter module. The core task is to calculate the collection priority of the hyperlinks related to the emergency, and join the task queue in order. If the task queue capacity is full, delete some lower priority hyperlinks to ensure the correlation with the incident. Web pages with higher and higher heat in the block are preferentially collected.

(2) Upgrade module. The upgrade module periodically analyzes the data of the target website, updates the site information of the target website, and submits it to the filtering module as a basis for task management, including structural update and content update. The structure update refers to periodically analyzing the structure of the target website to detect whether the website section has increased, decreased or changed. The content update refers to periodically analyzing the data of the

target website and detecting the existing sections of the website, whether the heat of the incident has changed.

5. Implementation Flow of Information Acquisition based on Web Crawler Technology

In the crawler system, the URL queue to be crawled is a very important part. It is also an important issue to queue the URLs in the URL queue in order, because this involves first crawling which page, then grabbing which page to take. The method of determining the order of URLs is called a crawling strategy. Commonly used include depth first and breadth first.

The breadth-first crawling process starts with a series of seed nodes, extracts the "child nodes" (that is, hyperlinks) from these pages, and puts them into the queue to fetch them in turn. The processed links need to be placed in a table. Before each new link is processed, you need to see if the link already exists in the table. If it exists, the proof link has been processed, skipped, and unprocessed, otherwise the next step is processed. Breadth-first traversal is one of the most widely used crawling strategies in crawlers. The main reason for using breadth-first search strategy is threefold: First, important web pages are often close to the seeds. For example, opening a news website is often the hottest. The news, as it continues to deepen, the importance of seeing pages is getting lower and lower. Second, the actual depth of the World Wide Web can reach up to 17 layers, but there is always a short path to a certain page. The breadth-first traversal will reach this page as quickly as possible. Third, breadth priority is conducive to the cooperation of multi-reptiles. Multi-reptile cooperation usually first captures the internal links. The sealing is very strong.

Depth-first search is a method that is used much more early in the development of crawlers. The goal is to reach the leaf nodes of the structure being searched. In an HTML file, when a hyperlink is selected, the linked HTML file will perform a depth-first search, that is, before searching for the remaining hyperlink results, a separate chain must be completely searched. Depth-first search along the hyperlink on the HTML file, go to no longer deep, and then return to an HTML file, and then continue to select other hyperlinks in the HTML file. When there are no more hyperlinks to choose from, the search has ended. The advantage is that it can traverse a Web site or a deep nested collection of documents; the disadvantage is that because the Web structure is quite deep, it may cause a situation that once it enters, it will never come out.

The breadth-first strategy can be collected simultaneously using multi-threading, which improves the collection efficiency. Therefore, this paper uses the breadth-first strategy for information collection, and the information acquisition implementation process is shown in Fig. 4.

For the information collection implementation process shown in Fig. 4, a brief description is as follows: After the system starts, the data is initialized first, then the information retrieval thread is started, and the acquisition is started. The acquisition process is divided into three steps:

In the first step, the URL address is analyzed and extracted according to the content of the query.

In the second step, the URL data is stored in the cache. If the cache reaches the predetermined value, go to step 3; otherwise, determine the crawl depth reaches the predetermined value or not, and if so, move to step 3, otherwise, move to step 1.

In step 3, the cached data is stored to the database. If you want to exit the system, end the acquisition, otherwise, move to step 1 and continue the acquisition.

Acknowledgement

This work is supported by 2018 annual natural science foundation in Liaoning province (20180550541): Research on Analysis Technology and System for Internet Public Opinion of Emergency in Big Data Environment.

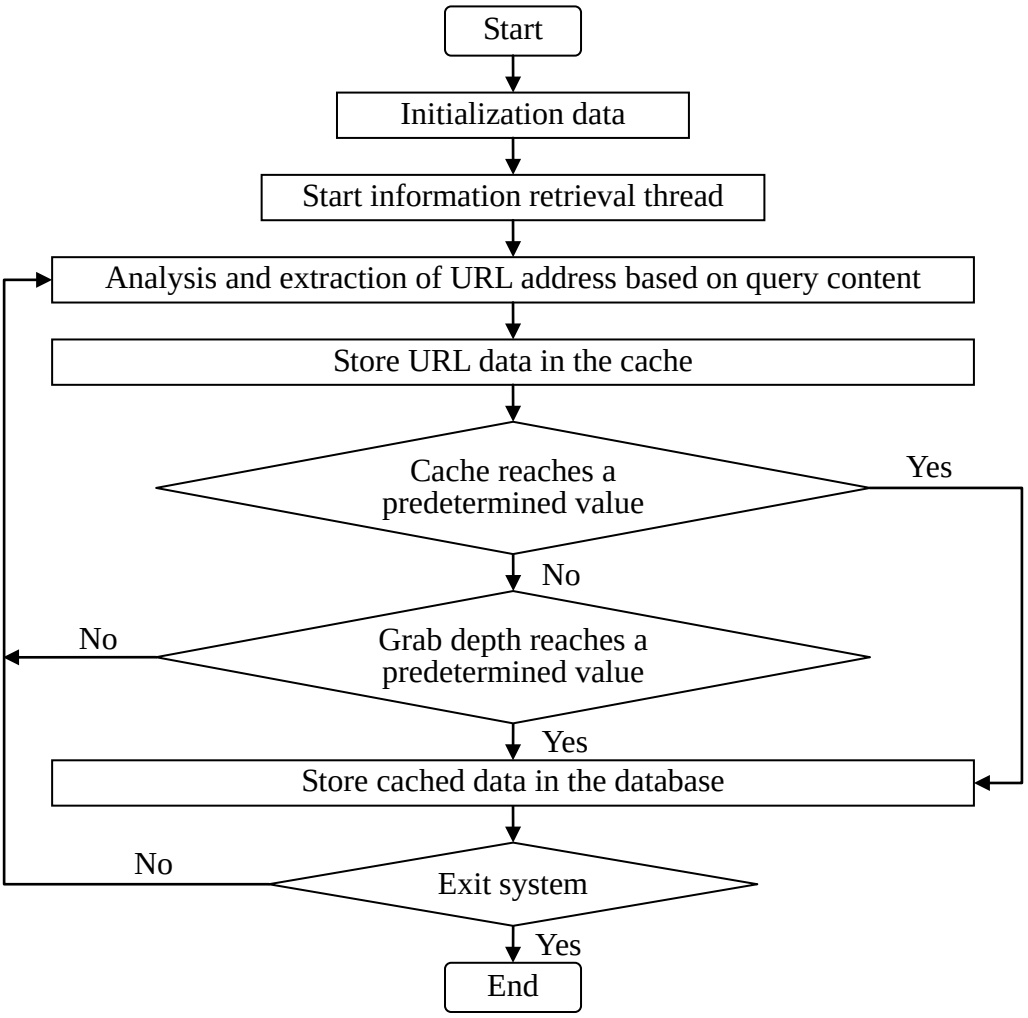

Fig.4. Implementation flow of information acquisition based on web crawler technology

References

[1] S. F. Lu, "Design and Implementation of Web Crawler System Based on Python," Computer Programming Skills & Maintenance, vol. 26, no. 2, pp. 26-27, 2019.

[2] X. F. Ge, J. Liu, "On Web Crawler Software Module Design," Journal of Heihe University, vol. 9, no. 10, pp. 209-210, 2018.

[3] M. J. Zhang, "Design and Implementation of Network Public Opinion Data Acquisition System Based on Web Spider," Modern Computer, vol. 32, no. 18, pp. 72-75, 2015.

[4] Y. R. Shen, Z. G. Wei, M. Liu, "Research and Application of Information Data Acquisition Method Based on Subject Network Crawler," Electronic Technology & Software Engineering, vol. 5, no. 7, pp. 168-169, 2016.

[5] Z. Li, "Governance of Internet," Youth Journalist, vol. 77, no. 2, pp. 76-77, 2017.

[6] Y. Yu, "Microblog-oriented crawler data acquisition," Information System Engineering, vol. 30, no. 12, pp. 36-37, 2017.