Summary of web crawler technology research

Linxuan Yu [1], Yeli Li[2], Qingtao Zeng[3], Yanxiong Sun[4], Yuning Bian [5] and Wei He[6]

[1]Beijing Institute Of Graphic Communication, Beijing, 102600, China, 951101614@qq.com

[2]Beijing Institute Of Graphic Communication, Beijing, 102600, China, 1605872754@qq.com

[3]Beijing Institute Of Graphic Communication, Beijing, 102600, China, 954276545@qq.com

[4]Beijing Institute Of Graphic Communication, Beijing, 102600, China, tongxin_yxsun@163.com

[5]Beijing Institute Of Graphic Communication, Beijing, 102600, China, 774200483@qq.com

[6]Beijing Institute Of Graphic Communication, Beijing, 102600, China, 871613503@qq.com

Abstract. With the continuous development of network information technology, there is a large amount of unstructured data called big data on the network. Human resources to collect information laborious, so web crawler technology came into being. This paper explores the basic principle and characteristics of web crawler and the classification of current popular crawler, introduces the key technology of crawler, compares two search strategies and the current application of crawler. Finally, the future research direction of web crawler is introduced.

1. Introduction

With the emergence of network technology, network data has reached a considerable amount. There are all kinds of big data on the Internet, and the Internet is a collection of these huge data. However, this data is not easily stored in a local database for access and processing. The search tools that people are using in daily life, such as Google, can only provide rough search results for people, but cannot provide accurate information. In order to make up for the defects of general search engines, a search tool that can obtain information directionally -- vertical search engine has emerged. Web crawlers play an important role in collecting network data. A web crawler is a computer program that traverses hyperlinks and indexes them. As the core part of the vertical search engine, how to make crawlers more accurate and faster to grab information has become an important research direction in the field of crawlers, which has attracted extensive attention from many researchers at home and abroad [1].

This paper mainly introduces the basic principle, system structure, classification, key technology of current web crawler, and the main application of current web crawler. The principle and key technology of each classification are introduced in detail.

2. Introduction to web crawler

2.1. Basic principle of crawler

The web crawler can traverse the information on the web page by itself, and the search engine is inseparable from the web crawler. The most important role of the web crawler is to crawl in the big data of the Internet, find effective information, and store the needed information data into the local database. The realization principle and process of crawler are shown in figure 1.

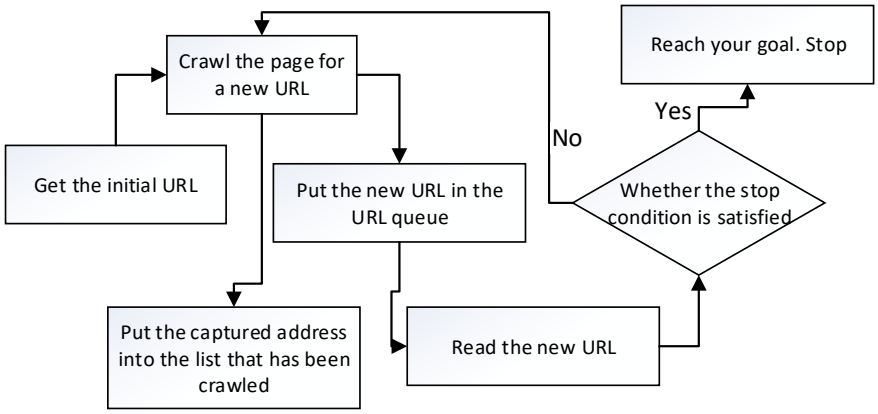

Figure1. Realization process principle of crawler

Crawlers mainly include downloaders, information extractors, schedulers and crawl queues. Scheduler will seed URL provided to download, then downloader get page information from the Internet to find and send information extractor, extractor strategy according to the instruction from information extraction to obtain information and the next level in the URL, then the next level URL to a waiting queue, waiting queue to go to submit the URL of the heavy, filtering and sorting operation into the list, after waiting for the scheduler calls [2-3].

2.2. Several characteristics of crawler

•distributed -- can be synchronized across multiple machines in a distributed environment.

•scalability -- crawling is slow due to the large amount of data, which can be improved by adding additional machines or increasing network bandwidth.

•performance and efficiency -- a web crawler that traverses a site for the first time can download all available files, enabling system resource utilization efficiency, i.e. processor, storage, and network bandwidth.

•quality -- web crawlers should give priority to obtaining high-quality pages that users need, and improve the accuracy of obtaining pages and reduce the acquisition of other pages.

•freshness -- keep search engines fresh, crawling their data independently based on the change frequency of each page and database, and crawling new urls that are randomly created or updated. For example, news, updated novels [4].

•extensibility -- designed to be extensible crawlers, with crawlers set up in a modular architecture. So as to adapt to the new data format and acquisition protocol [5].

3. Classification of web crawlers

According to their different characteristics, crawlers can be divided into generic web crawlers, focusing web crawlers, incremental web crawlers, distributed crawlers, parallel crawlers, traditional web crawlers and Internet of things web crawlers.

3.1 Generic web crawler

Generic web crawlers are also known as traditional crawlers. Traditional crawlers grab all documents and links related to the topic. Generic web crawlers retrieve a large number of pages from various fields

from the web. To find and store these web pages, a generic web crawler must run for a long time and consume a lot of hard disk space. For example, Google's PageRank algorithm returns pages conforming to search criteria from 25 billion documents on the network [6].

3.2 .focus web crawlers

A focus web crawler is also called a topic web crawler. Unlike general crawlers, focused crawlers only crawl specific web pages, which can save a lot of time, disk space, and network resources. As the saved pages are fewer and better, the update speed is faster, which is more suitable for some enterprises and individuals to collect some specific information [7]. The difference between focus crawler and general crawler lies in two modules to filter web links: web page decision module and URL link priority ranking module to filter web pages.

• web page judgment module: when the crawler crawlies to obtain specific content, the web page relevance evaluator starts to compare the relevance between the content in the web page and the pre-given topic. If the relevance of the web page does not reach the previously set threshold, it will abandon the web page to maintain the high accuracy of obtaining the web page.

•URL link priority ranking module: this module is mainly used to compare the degree of relevance between the URL resolved and a given topic. The module prioritized the links according to the authority of the links to the content and the number of citations of the links. Sort by priority and remove links that are too low in priority.

3.3.Incremental web crawlers

The difference between incremental crawler and general crawler mainly lies in different search strategies. For general crawler, after the completion of one traversal, it needs to update the data, conduct a new traversal of the whole network according to the previous traversal form, and then replace the previous results. Incremental crawler adopts a new mechanism to update the data. Based on the previous results, it marks the existing collection. When updating the data, it only obtains the data of expired pages through the marking information of the existing data, and replaces the past pages with new pages, and other unexpired information is not understood. It can greatly improve crawling efficiency, reduce the occupation of physical memory, and greatly improve the rate of data update.

The incremental crawler has a scheduler that signals the web page and database to crawl again at a specified interval based on some refresh policy. Focus on the time interval between two changes to the same database, and then crawl their data independently based on the change frequency of each web page and each deep web database. Instead of thinking about the common time for the next incremental crawl of all web pages and databases.

Incremental crawler is mainly divided into the following three situations:

• websites with new pages, like new chapters in novels, daily news, etc.

Method: determine if the URL has been crawled before sending the request.

Advantages and disadvantages: cannot get the content of the page changes, but because there is no need to crawl the url to send a request, so the pressure on the server is relatively small, faster, suitable for crawling new pages

• websites where page content will be updated.

Method: after analyzing the content, judge whether this part of the content has crawled before

Advantages and disadvantages: you can sense whether the content of each page has changed or not, and you can get the content added or changed on the page, but the speed is relatively slow due to the request for each url, and the pressure on the website server is also relatively large

• determine if the content already exists in the storage medium when writing to it (final safeguard).

3.4.Distributed Web Crawlers

Distributed web crawlers run on groups of computers, each of which runs a focused crawler. The core problem of distributed crawler is how to coordinate and manage the work between each node. Avoid repeated work on each machine, so that distributed groups can run efficiently and steadily. So the key

problem is to solve the communication coordination between machines. Distributed crawlers can be divided into the following three types:

• master slave mode: this mode is hosted by one machine and controls the operation of the whole group. The host is responsible for managing the list of urls to be crawled, maintaining communication with each machine, issuing tasks to each machine and monitoring the working status of each machine to ensure the normal operation of each machine. Each slave machine only needs to complete its own task and report the result to the host machine. There is no need to communicate between the slave machines.

• autonomous mode: this mode has no host to control the release task, and mainly ensures the normal operation of distributed crawler through the communication between each machine, which is relatively simple to realize. There are two forms of communication, circular communication means all machines form a circular structure, one-way transmission of information, relatively simple. Full unicom communication means that each machine has to communicate with other machines to form network communication, which is relatively complex.

• mixed mode: combined with the characteristics of the above two modes, mixed mode has the host responsible for task assignment to other machines, but other slave machines can also communicate with each other and have task assignment function, and the failed task assignment will be secondary assigned by the host.

3.5.Parallel Crawler

It is difficult to retrieve all the Internet data with a single crawler process, so the crawler process must be parallel to complete the crawler process in the shortest time. This type of crawler is called a parallel crawler. The main goal of parallel crawler design is to maximize parallel performance and minimize parallel consumption. Parallel crawlers can be embedded or distributed. Parallel distributed crawlers can communicate over local or wide area networks. Currently, the improved UbiCrawler[8] is a distributed crawler, which can run on any type of network and greatly reduce the time consumed in the crawling process.

3.6.Iot Web Crawler

Iot devices can be public Iot and special Iot. Public Internet of things devices connect to the Internet. In common Iot devices, Iot sensors/devices (nodes) are accessed via HTTP.

Shodan search engine looks for devices connected to the Internet using service banners (a block of text about the service being performed). Devices connected to the Internet such as printers are controlled remotely; Web cameras and other devices can be detected by Shodan [9]

Shodan Collect data from the following ports: HTTP, FTP.SSH. Telnet. SNMP. SNMP. SIP. RTSP. Shodan crawlers discover collect metadata about the devices, while Google crawls for websites.

4. Technical strategies applied in web crawlers

4.1.Webpage Acquisition And Analysis

The basic principle of Web crawler is to simulate the browser to make HTTP request, and the crawler sends the request to the Web server through HTTP request. After obtaining the response from the server, the crawler analyzes and stores the Web page, and completes the crawling work of the crawler system.

Web page parsing is mainly a process of web page de-noising. In the Internet, all kinds of information of web pages are stored in the framework of HTML. Web page denoising is mainly the extraction of web content text. When the theme crawler extracts the content in the web page, it needs to parse the HTML structure of the page to extract effective information from the page. Common methods include parsing the HTML structure through Beautiful Soup and extracting text data using regular expressions.

4.2.Data Storage Method

Crawler access to the data, generally choose two storage methods: storage to the local file or save to the database. Small amount of data can be directly saved to the local data, large amount of data generally choose to save to the database.

The database can choose to use the Redis database, which is a high-performance key-value database. Redis database has the characteristic of disorder and no repetition. We can through the URL or page content fingerprints as a key in the Redis collection in the database, using the collection does not achieve the purpose of to heavy repeatability, every time the crawler to deal with the URL or the page will go first to check whether the Redis database already exists, because the Redis database on key - the value stored in the form of, so the speed of this step will be very considerable, Secondly, Redis can persist the contents in memory to disk, and every operation is atomic, which guarantees the reliability of crawler, that is, crawler will not lose data due to unexpected stop.

4.3.Web Search Strategy

There are three search strategies: depth-first strategy, breadth-first strategy and optimal search strategy.

• depth-first strategy: depth-first strategy is generally adopted for those with small depth. The depth-first search strategy is executed from the beginning to enter and analyze the initial page before jumping to the next link, traversing a path and traversing the next path. There is a disadvantage in the web crawler system that USES the depth-first strategy, that is, the links provided by some portal sites are often of high value, and the value of the web page is constantly weakened as the web page continues to crawl. In addition, the structure of the Internet is so complex and deep that crawlers keep digging deep, which can lead to problems.

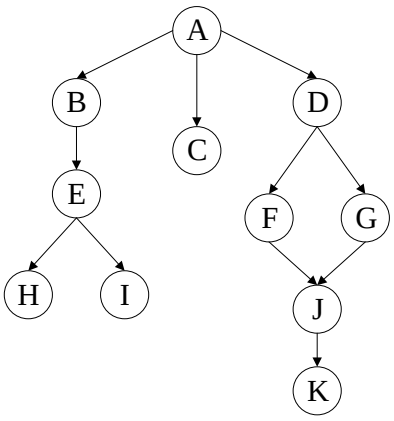

Figure2. Network link simplified diagram

The network structure is shown in figure 2, and the fetching path is shown in table 1.

Table 1. Depth-first crawl path

Path number	Through the path
1	A→B→E→H
2	A→B→E→I
3	A→C
4	A→D→F→J→K
5	A→D→G→J→K

• breadth first strategy: breadth first search strategy is used in more general cases. Mainly around the original seed within a certain distance of the value of the web page is higher, so from the beginning of the web page to crawl the link to the web, and then choose one of them, continue to dig. Breadth first search strategy allows search according to tree split level, if the current level of search is not complete, will not move to the next level of search. In this regard, breadth first search strategy is blind search, it

will search the whole knowledge area, and make the efficiency reduced. But if you want coverage as the core, breadth first search strategy is a good choice. Under the network in figure 2, the fetching path is shown in table 2.

Table.2 Breadth-first crawl path

Path number	Through the path
1	**A**
2	**B→C→D**
3	**E→F→G**
4	**H→I→J**
5	**K**

• best priority search strategy [10]: according to a specific algorithm, the best priority search strategy calculates the similarity between candidate web pages and target web pages and performs the first crawl. It only accesses the passed web page analysis algorithm for prediction and evaluation, and it also has corresponding problems. Many related web pages may be ignored in the path of crawler track, because the best priority strategy is local optimal strategy rather than global optimal strategy, so some subject-related web pages will be ignored in the crawling process.

5. Application of web crawler

5.1.Network Information Field

Every time the user searches, browses and accesses information online, the most important thing is the freshness of information or whether it is related to their own needs. All these works are based on crawler technology. Crawler obtains data, conducts analysis and processing, and feeds back to users, so that users can obtain relevant information.

5.2.Financial Sector Field

Web crawlers get information about the financial sector or competitors for businesses. However, the most important thing is to obtain the information of customers, and analyze the browsing information and habits and preferences of customers through the information mining and analysis of the web crawler.

5.3.Network Security Field

When detecting whether a document is a malicious document, a large number of secure documents should be collected for analysis to obtain the characteristic information that can identify the secure document and form the feature library. When testing the file to be tested, analyze and compare the file to the features in the feature library, and finally return the result of whether the file is safe or not. When collecting the security information characteristics of the training sample, it can be realized through crawler technology [11].

6. Summary and prospect

So far, researchers have done a lot of research on the theme web crawler, but there is still a lot of room for research on the performance of the theme crawler, which can be divided into the following points:

a) Web crawlers are all fixed search strategies. Faced with the different forms of web page organization among different websites in the Internet, fixed search modes cannot be effectively crawled. How to improve the performance of theme crawlers by integrating crawling rules remains to be studied.

b) Building a topic crawler by using web content and link context for a broad topic can effectively calculate the relevance of the topic. However, there are some limitations for a more detailed topic, so improving the selection of topic feature words from the semantic perspective has become a hot research topic crawler technology in the future.

c) For the protection of website information, a set of anti-crawler strategy is designed to prevent crawlers from grasping data. As for the anti-crawler strategy, researchers introduce advanced crawlers to obtain massive information. However, the more advanced crawlers have higher development costs, so it remains to be studied whether a low-cost crawler can be designed.

Acknowledgments
Our thanks to Beijing science and technology innovation service capability construction project (PXM2016_014223_000025) and Beijing Institute of Graphic Communication 2018 R&D Project (Ec201802).

References
[1] Pan XY, Chen L, Yu HM, Zhao YJ, Xiao KN, (2019) Survey on research of themed crawling technique. Application Research of Computers.Vol.37, No.5
[2] Rungsawang A, AngKawattanawit N. (2005) Learnable topic-specific web crawler. Journal of Network&Computer Applications, 28(2):97-114.
[3] Kozanidis L. (2008) An Ontology-Based Focused Crawler. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 376-379.
[4] G. Pavai1, T. V. Geetha. (2016) Improving the freshness of the search engines by a probabilistic approach based incremental crawler. Springer Science+Business Media New York.19:1013-1028.
[5] Deka, GC, (2018) NoSQL Web Crawler Application. Advances in Computers.109:77-100.
[6] D. Austin,(2017) How does Google find your needle in the haystack of the web, http://www.ams.org/samplings/feature-column/fcarc-pagerank.
[7] Sun LW, He GH, Wu LF (2010), Research on web crawler technology, Store brain knowledge and technology,06(15):4112-4115.
[8] P. Boldi, B. Codenotti, M. Santini, S. Vigna(2017) UbiCrawler: a scalable fully distributed web crawler, http://vigna.di.unimi.it/ftp/papers/UbiCrawler.pdf.
[9] Cybrary. IT(2017), Shodan: the hacker's search engine, https://www.cybrary.it/0p3n/intro-shodan-search-engine-hackers/
[10] Cho J, Garcia-Molina H, Page L. (1998) Efficient crawling through URL ordering. Computer Networks and ISDN Systems, 30(1-7): 161-172.
[11]P.Pantel, D, Lin,(2002), Document clustering with committees, SIGIR Forum (ACM Special Interest Group on Information Retrieval):199-206.