Design of Four Web Crawlers based on Python

Pan Liu[1, *], Yihao Li[2], Xuankui Zheng[1], Shili Ai[1], and Wenjie Zhang[1]
[1]Faculty of Business Information, Shanghai Business School, Shanghai 201400, China
[2]School of Information and Electrical Engineering, Ludong University, Yantai, China
*corresponding author

***Abstract*—The data on websites is an important source of data for both big data analysis and machine learning. Due to the limitation of data crawling on some websites, the general web crawler will be invalid. To facilitate the crawling of data in websites with different structures, this paper introduces four types of web crawlers. Then, based on some third party libraries developed for Python, the corresponding Python programs are designed respectively for these four web crawlers. This paper provides a technical guide for those researchers who want to construct web crawlers quickly.**

Keywords-web crawler; website; data source; Python; data collection

I. INTRODUCTION

In the digital period, data has become a core competitiveness of countries, enterprises, and business groups [1, 2]. These data can be analyzed or learned to get some meaningful data analysis results. Thus, how to obtain these data has become a critical task [3]. The data in websites, such as the customer reviews [4], the News media data [5], and the BBS data [6], can be downloaded and studied by researchers with a web crawler [7, 8].

A web crawler, known as a spider robot, is a program that can grab web page data according to users' needs or under certain rules. A classic web crawler can automatically traverse the hyperlink structure of the web, then locate and retrieve information on web pages. Generally speaking, starting at one page of the website, web crawlers read the content of the page to find hyperlinks in the page, and then follow those hyperlinks to the next page until all the pages have been visited.

In the past, web crawlers were divided into four types [9]: general purpose crawler, focused web crawler, incremental web crawler, and deep web crawler. Each crawler has its own designs, which make it difficult for many non-professionals to develop a proper web crawler to extract data from websites with different structures. This paper uses the Python programming language to design four programs of the four web crawlers. The programs we designed includes third-party libraries, web page fetching methods in different website structures, and data storage in CSV format. Our study provides a fast guide for researches to develop a web crawler using Python.

II. PROCESS OF WEB CRAWLER

Fig.1 shows a flow of a web crawling process based on Python. This process can be divided into six steps as follows:

1) Import some third-party libraries,
2) Send an HTTP request,
3) Set a URL of a website for data crawl,
4) Download the page from the URL,
5) Extract data from the page, and
6) Store data to a data repository.

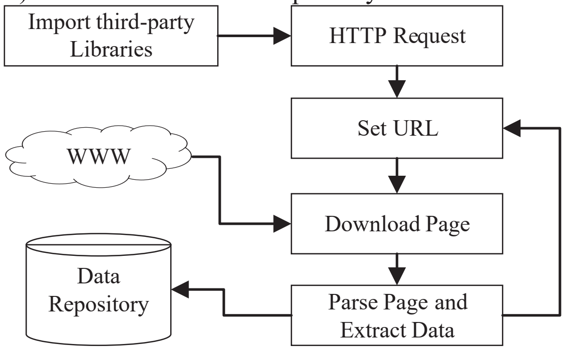


Fig. 1. Flow of a web crawling process based on Python

We first need to import these third-party libraries to a crawler developed by Python. Then, we can send an HTTP request to the target site through the HTTP library. The request consists of some additional headers, which usually contain the name and version of the operating system and browser used by the client. If the server can respond normally, the crawler will get a response that is the content of the web page. The type of the response may be an HTML page, JSON string, and binary data, such as pictures and videos. Then, we need to set a URL of a website for data crawl. Because breakpoints are set on the web pages of certain websites, HTTP links cannot locate the pages of these websites. Therefore, we need to set the parameters in the URL to realize the location of the pages. Next, we need to extract the data from the web page. Finally, we need to store the extracted data.

III. IMPLEMENTATION OF FOUR CRAWLERS

A. General purpose crawler

The target data of general purpose crawler are huge, and its crawler range is also very large. Because this type of web crawler can capture massive data, it requires high-performance equipment to run. The general purpose crawler is mainly applied to large search engines or large data providers, so it has very high application value.

B. Focused web crawler

Focused web crawler is a kind of crawler that selectively crawls web pages according to pre-defined themes. This type of web crawler does not locate target resources in the whole Internet like general web crawler, but locates the crawled target web pages related to a pre-defined theme. In addition, the bandwidth resources and server resources required by focused web crawler can be greatly saved. Focused web crawlers are mainly used in crawling specific information, and mainly provide services for a specific type of people.

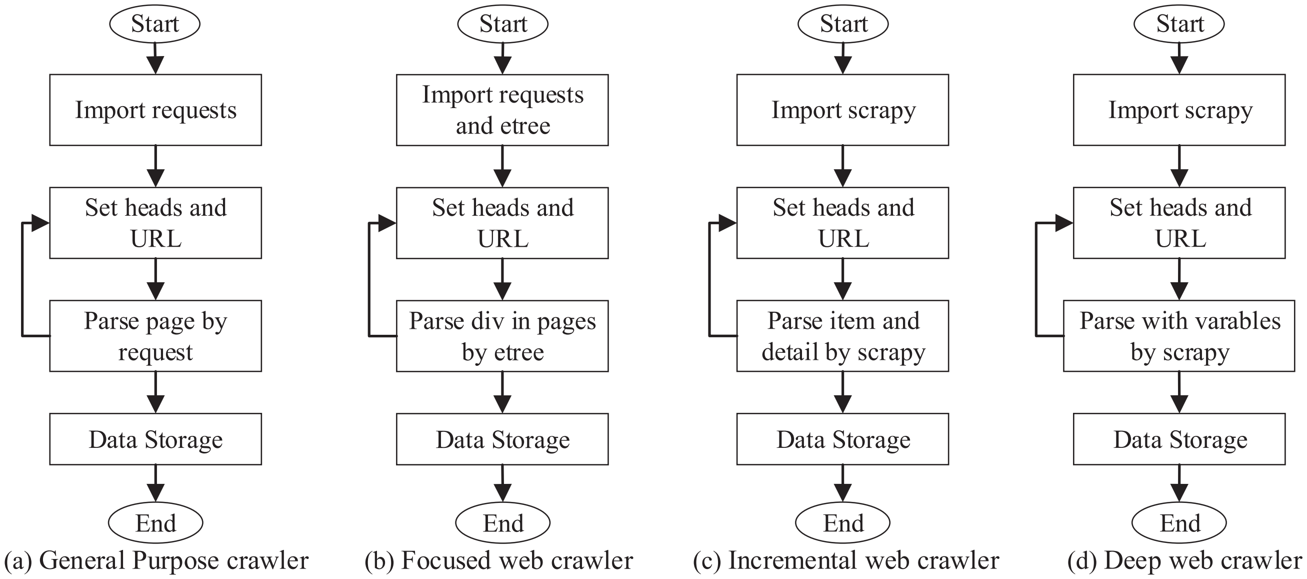
C. *Incremental web crawler*

When crawling web pages, the incremental web crawler only crawls web pages whose content has changed or newly generated web pages, and will not crawl pages whose content has not changed. Incremental web crawlers can, to a certain extent, ensure that the crawled pages are as new as possible.

D. *Deep web crawler*

On the Internet, web pages can be subdivided into surface pages and deep pages. The so-called surface page refers to a static page that can be reached by using a static link without submitting a form. The deep page is hidden behind the form and cannot be obtained directly through a static link. They are only available after entering some keywords. On the Internet, the number of deep pages is often much larger than the number of surface pages, so it is necessary to find a way to crawl deep pages.

Figure 2 shows four activity diagrams of the above four web crawlers. Their Python programs can be downloaded from https://pan.baidu.com/s/11d12Y-JN62kvGmgB1qL6oQ with password PAN1.



(a) General Purpose crawler (b) Focused web crawler (c) Incremental web crawler (d) Deep web crawler

Fig. 2. Four activity diagrams of four types of web crawlers

IV. CONCLUSION

It is one of the main ways to get data from web pages by designing a web crawler. This paper discusses the characteristics and advantages of four common types of web crawlers. Then, a Python-based web crawler processing flow is introduced. Next, we design four Python programs for four kinds of web crawlers. These designed programs can be used by researchers to quickly develop their own web crawlers for collecting data from web pages in website with different structures.

REFERENCES

[1] E. Raguseo, F. Pigni, and C. Vitari, "Streams of digital data and competitive advantage: The mediation effects of process efficiency and product effectiveness," *Information & Management,* vol. 58, no. 4, p. 103451, 2021.

[2] P. Liu, J. Yan, and X. Song, "Three Classifications of Big Data-based Software Testing," in *2021 8th International Conference on Dependable Systems and Their Applications (DSA)*, 2021: IEEE, pp. 751-752.

[3] J. Sun, Y. Zhao, P. Liu, J. Li, and H. W. Zhai, "Personalized Recommendation System of Web Academic Information Based on Big Data and Quality Monitoring Technology," in *International Conference on Data Mining and Big Data*, 2021: Springer, pp. 263-274.

[4] Y. Zhang, J. Wang, and X. Zhang, "Personalized sentiment classification of customer reviews via an interactive attributes attention model," *Knowledge-Based Systems,* vol. 226, p. 107135, 2021.

[5] B. Abu-Salih, P. Wongthongtham, D. Zhu, K. Y. Chan, and A. Rudra, "Sentiment Analysis on Big News Media Data," in *Social Big Data Analytics*: Springer, 2021, pp. 177-218.

[6] B. P. Edwards and A. C. Smith, "bbsBayes: An R package for hierarchical Bayesian analysis of North American breeding bird survey data," *Journal of Open Research Software,* vol. 9, no. 1, 2021.

[7] K. Prabha, C. Mahesh, and S. Raja, "An enhanced semantic focused web crawler based on hybrid string matching algorithm," *Cybern. Inf. Technol,* vol. 21, no. 2, pp. 105-120, 2021.

[8] M. ElAraby and M. Shams, "Face retrieval system based on elastic web crawler over cloud computing," *Multimedia Tools and Applications,* vol. 80, no. 8, pp. 11723-11738, 2021.

[9] M. A. Kausar, V. Dhaka, and S. K. Singh, "Web crawler: a review," *International Journal of Computer Applications,* vol. 63, no. 2, 2013.