3.2Crawler search method

3.2.1 Breadth-first search

Breadth-first traversal is a search strategy widely used by crawlers. Its process is to have a URL queue first, pop up the URL in the queue, then extract the sub URL in the pop-up links, and put them back in the original URL queue to wait for pop-up. The URL that have been searched will be put into a table similar to the collection. Each time a new pop-up URL is processed, a judgment will be made first to see if there is a URL in the searched table. If there is one, skip it and carry out the next operation. The advantage of this operation is that it reduces the repetition rate without repeated crawling. The disadvantage is that the information update will not be timely after the page update, and each judgment will consume more resources and time. (Wang et al 2019)

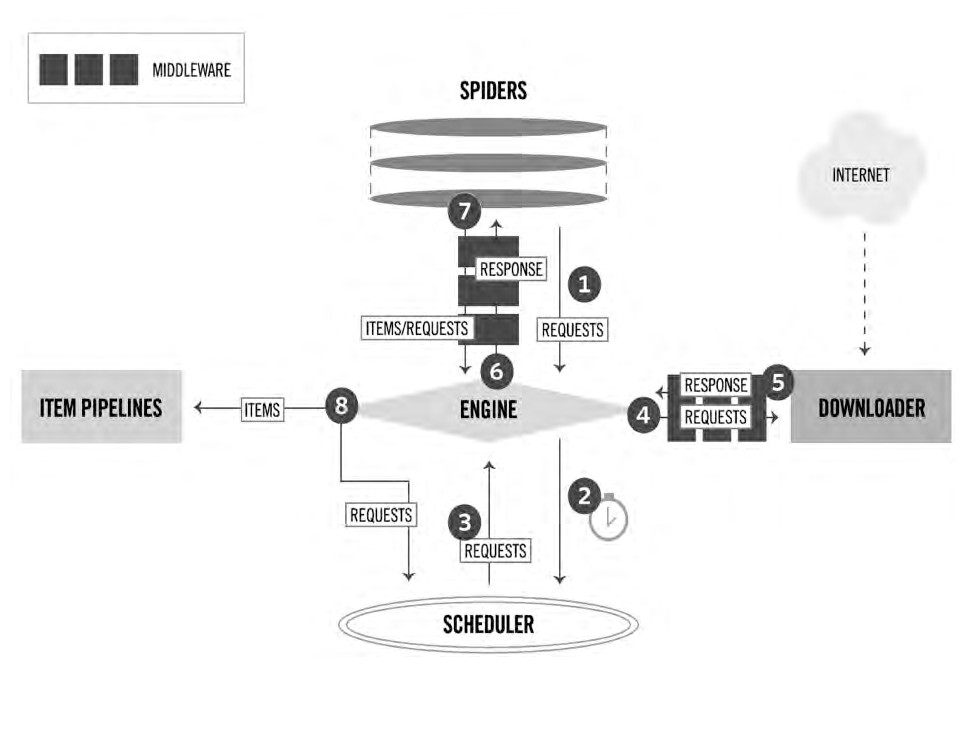
3.2.2 Depth-first traversal algorithm

Depth-first traversal is often used in early crawler development. The general process is as follows: Firstly, find the first hyperlink URL from an HTML page. Then extract the URL, and then extract the first URL within the URL. It is always a single chain mode to dig deep until there is no next URL, and then return to the first HTML interface, and perform the same operation from the second URL. That is to say, there will be a complete single chain search before the next capture. When there is no next step for all links, the search ends. (Wang et al 2019) The advantage of this approach is that it can find all the deep URL, but it also has disadvantages, because once the search starts, it may fall into a permanent deep search and cannot jump out. (Zhao et al 2017, Wang et al 2019)

3.2.3 Design and implementation of Crawler Based on scratch framework

Scrapy is a fast and high-level application framework written in Python for crawling website data and extracting structural data. At present, it is the most famous and widely used framework among all crawler frameworks. Scrapy is now capable of interacting with APIs in order to extract data. There are many reasons that contribute to the use of Scrapy as the WebCrawler. (Shi et al 2016)

Instead of designing a crawler framework from scratch, we can learn how to use Python's scratch framework simply and efficiently. (Liu 2018) Developers only need to develop a few specific modules to write a stable and efficient web crawler. (Sun et al 2019) Its detailed framework and operation process are shown in the figure. (Li et al 2017, Wang et al 2019)



3.2.3.1The running process of the Scrapy Crawler

The Engine gets URLs to scrape from the Spider and schedules them in the Scheduler, as Requests. Then, the Engine asks the Scheduler for URLs to crawl, as Requests, and send them to the Downloader, passing through the Downloader Middleware. Once the page finishes downloading, the Downloader generates a Response and sends it to the Engine. The generated response contains the copy of the static HTML of the web page. The information extracted is stored in Items that are data holders of the framework. Then through the use of the Item Pipeline, the Items or data can be saved in suitable Formats including but not limited to CSV or SQL database. (Bassam 2016, Shi et al 2016)

3.2.3.2Modules of the scrapy framework

1) Scrapy Engine: It is responsible for the transfer of regulatory data between modules in the system and calling corresponding functions to respond to specific events.

2) Scheduler : It is responsible for the unified management of all URL resources to be crawled. For example, insert the URL resources submitted by the receiving scrapy engine into the request queue. Then, the URL is taken from the queue and sent to the scrapy engine in response to the URL request from the scrapy engine.

3) Downloader Middlewares : It is responsible for delivering the URL request sent by the scrapy engine to the downloader module and the HTTP response sent by the downloader module to the scrapy engine.

4) Downloader : It is responsible for downloading the data on the web page and finally sending it to the crawler module through the scrapy engine.

5) Spider : It analyzes the data obtained from the downloader module, and then extracts the item or relevant URL resources.

6) Spider Middlewares : It is responsible for the input and output of the crawler module.

7) Item Pipeline : Process items extracted and sent by the crawler module through data cleaning, data validation, data persistence and other operations. (Cattell 2011, Li 2017)

3.2.4Strategy of anti-crawler Technology

Nowadays, many websites prohibit crawlers from crawling data. Websites use headers, user behavior, website directory, data loading and other ways to anti crawl, so as to increase the difficulty of crawling. Thus, there are several strategies:

3.2.4.1 Set download\_delay parameter

If the download waiting time is too long, the task of large-scale data grabbing in a short time will not be completed, and too short will increase the probability of being prohibited from crawling data. So we set DOWNLOAD\_DELAY = 2 in settings.py (Li 2017)

3.2.4.2 Disable cookies，

This prevents crawler behavior from being detected by sites that use cookies to identify crawler tracks. So we need to set COOKIES\_ENABLES = False. (Li 2017)

3.2.4.3 User-agent

User agent also refers to browser, including hardware platform, system software, application software and user's personal software preferences. (Chen 2016) Every browser and regular web crawler has a fixed user agent. Camouflage user agent can judge the category of the website visitors by violating rules. For camouflage browser and famous crawler, camouflage browser is more recommended. Compared with the crawler, the browser has no fixed IP and can be anyone, while the crawler has a fixed IP. Camouflage browser can improve multiple user agents. Each time a request is sent, a user agent can be randomly selected to set the code according to the specific needs. (Liu 2019) We need to set DOWNLOADER\_MIDDLEWARES = { ' scrapy.contrib.downloadermiddleware.useragent.UserAgentMiddleware':None,' HouseInfoSpider.spiders.rotate\_useragent.RotateUserAgentMiddleware' : 400，} in settings.py. (Li 2017)

Data storage

The crawled data can be stored locally or in the database.

Local storage

JSON files can be directly created in pipeline to write data, but the readability of JSON files is poor. Therefore, we can save the JSON file as a readable excel file after further processing. The extracted data table is a two-dimensional table structure. The Pandas library is the main data processing tool in Python. With the help of Dataframe, we can store two-dimensional table structure. First, build a dictionary, then a two-dimensional dictionary priceall, save it as a Dataframe, then build an index of Dataframe with a list, and finally save it as an excel file using the to\_excel() method of dataframe (Li et al 2018)

Database storage

The data model defined in scrapy establishes tables in the database, By using pymysql, a third-party module of python, to log in to the database, and executes SQL statements to insert data into the database. In order to achieve incremental crawling, a verification field needs to be created in the data table. (Li et al 2018)

3 Python-based Data analysis

3.1 Python Common Library

Python has Numpy, Pandas, Matplot lib, Scikit learn and other libraries with complete functions and unified interfaces, which can provide great convenience for data analysis. (Zhai 2018)

3.1.1 NumPy

This library, whose name means Numerical Python, actually constitutes the core of many other Python libraries that have originated from it. Indeed NumPy is the foundation library for scientific computing in Python since it provides data structures and high-performing functions that the basic package of the Python cannot provide. In fact, NumPy defines a specific data structure that is an N-dimensional array defined as ndarray. The knowledge of this library is revealed in fact essential in terms of numerical calculations since its correct use can greatly influence the performance of a computation. (Nelli 2015) Besides，Numpy also has the following features: (Zhai 2018)

1) It has functions that perform element level calculation on arrays and directly perform mathematical operations on arrays.

2)It can integrate C, C++ code into Python.

3) It can be used as a container for transferring data between algorithms.

3.1.2 Pandas

This package provides complex data structures and functions specifically designed to make the work on them easy, fast, and effective. This package is the core for the data analysis with Python. The fundamental concept of this package is the DataFrame, a two-dimensional tabular data structure with row and column labels.Pandas combines the high performance properties of the NumPy library to apply them to the manipulation of data in spreadsheets or in relational databases (SQL database). In fact, using sophisticated indexing it will be easy to carry out many operations on this kind of data structures, such as reshaping, slicing, aggregations, and the selection of subsets. (Nelli 2015) In addition, Pandas can also be used for data preprocessing, such as data consolidation, data cleaning, data standardization and data conversion. (Zhai 2018)

3.1.3 Matplot-lib

This package is the Python library that is currently most popular for producing plots and other data visualizations in 2D. Since the data analysis requires visualization tools, this is the library that best suits the purpose. (Nelli 2015) Matplotlib consists of four parts: (Li 2018)

(1) The basic figure type of Matplotlib;

(2) Adjust the style and color of figure;

(3) Add notes to the drawing (including coordinate axis range, length width ratio or coordinate axis, etc.);

(4) Other complex figures.

3.1.4 Scikit-learn

Scikit-learn is a simple and effective data mining and analysis tool. It is based on Numpy, SciPy and Matplotlib, and encapsulates some common algorithms. Its main modules are data preprocessing, model selection, classification, clustering, data reduction and regression and other machine learning algorithms, which can help users to quickly build models in the process of data analysis, and the model interface is unified, which is very convenient to use. (Zhai 2018)

3.1.5 WordCloud

The system can import Wordcloud to generate the specified word cloud, and remove keywords without substantive statistical significance in the continuous optimization process. (Guo 2018)

1李阿妮，张晓，张伯阳，等． 公有云存储系统性能评测方法研 究［J］． 计算机应用，2017，37( 5) : 1229 － 1235

5］赵亮，胡树煜，吴宇玲． 基于阿里云 ECS 的动物医学考试系统研 究［J］． 家畜生态学报，2017，38( 9) : 94 － 96．

3 郭丽蓉 基于Scrapy的科研数据分析平台设计 电子技术与软件工程 2018（12）：136-137

4 Shi, Z., Shi, M., & Lin, W. (2016). The Implementation of Crawling News Page Based on Incremental Web Crawler.

［12］ Cattell R．Scalable SQL and NoSQL data store ［J］．ACM SIGMOD Record，2011( 2) : 12－27．

[6] 陈利婷 . 大数据时代的反爬虫技术 [J]. 电脑与信息技 术 ,2016,24(6):60-61.