Design and Implementation of Crawler Program Based on Python

Xiaoju Ma 1, *, Min Yan 2

1 School of Information Engineering, Shaanxi institute of international trade & commerce, Xianyang 712046, China;
2 IHPC, Agency for Science, Technology and Research, Singapore 138632, Singapore

*Corresponding author e-mail: Ma_Xiaoju@ihpc.a-star.edu.sg

Abstract. With the development of computer and network technology, we often get information through the Internet. However, it is difficult for us to obtain valuable information from massive amounts of data because of the large amount of network data and complex formats. At present, researches found that web crawler technology can be automatically obtained information from internet. In this paper we takes the crawling of second-hand housing information of Anjuke Xi’an as an example. According to the crawler principle and process, the structure of Anjuke's page is first analyzed, using requests to obtain web pages, lxml to analyze web pages and SQL Server 2017 to store data to design and implement a network. The crawler program collects and stores housing information in some cities in East China through this program, and finally analyzes the housing price trend through the collected data through Excel. The results show that this program can automatically obtain housing information from the Internet, which provides a data source for later data analysis.

Key words: Python; Big Data; Web Crawlers; Data Analysis.

1. Background
Housing prices are the hottest topic discussed today, and housing prices across the country have been rising in recent years. In the face of rising housing prices in various places, it is difficult for working-class people with average income to bear the pressure of buying houses. In order to relieve the pressure on housing, more and more people choose to buy second-hand houses with relatively low prices. During the Golden Week of November 2020, the second-hand housing market is active. According to data from the Shell Research Institute, the average daily transaction volume of second-hand housing in the 28 cities under monitoring has increased by 50% year-on-year [1].

Before buying a house, people search for the housing information they want to know in advance on relevant websites, such as lianjia.com, 58.com, anjuke.com, etc. The website of anjuke can lead domestic real estate information service platform for new houses, second-hand houses, rental houses, commercial real estate, overseas real estate, decoration, etc. The platform provides users with a secure, reliable, and convenient housing search service, while meeting the needs of developers and brokers for efficient online promotion. In the process of users browsing the website information, the platform can record the process data and perform big data analysis to grasp the user's viewing preferences, recommend interested
houses, highly praised intermediary companies and familiar with the community of real estate brokers provide services to users. Therefore, most people now get housing information from anjuke.com. However, in the face of huge and widely sourced network data, how to shorten the download time of web pages, how to analyze large amounts of data and find useful information, requires the use of web crawler technology.

The housing price of second-hand housing is affected by other social factors in addition to the area, geographic location, price, house type and other factors. Therefore, it is necessary to study the influencing factors behind the price of second-hand housing. To study the influence of these factors, it is necessary to obtain relevant data of second-hand housing first. This article takes the acquisition of second-hand housing information of Anjuke Xi’an as an example, and designs a Python-based crawler program to automatically obtain second-hand housing information from the Internet, and stores the acquired data in the database to dig out data hidden on the Internet for users. The distribution of all second-hand housing listings in Features such as area, orientation, floor, construction date, unit price, total price, community name, etc., help users choose suitable and cost-effective housing.

2. Introduction to Crawlers

2.1. Network Reptile Principles
Web crawlers, but also web spiders, web robots and other titles, is a set in advance in accordance with the rules of procedure to independently grasp the network information program script. Reptile technology can collect information materials on web pages quickly and widely, which is very convenient in obtaining web data. Therefore, it is widely used in search engines to achieve the purpose of collecting and processing network data [2].

2.2. Classification of Network Reptiles in Usage Scenarios
The main work of network reptiles is to automatically search for the required web page information, and download the information required by users from the network, usually can be divided into data collection, data collation, data storage three modules. Network reptiles are available in the general reptile, focused reptiles, and incremental reptile categories [3].

Universal reptile. Also known as a full-web crawler, a new URL obtained and downloaded pages starting with a single or multiple initial seed URLs, using a specific parser to remove HTML tags, saving the summary to the Web database, and entering a new loop until the required information is obtained [4].

Focus on reptiles. Focused reptiles are based on generic reptiles that capture specific local content on a page. For example: Baidu paste all the comments in the bar [5].

Incremental reptiles. Detecting data updates in a web site will only crawl the latest updated data in the site. A large increase in the number of web page downloads, a change in focus crawling, or new pages can reduce time consumption to a certain extent [6].

2.3. Network Reptile Process
The process of web crawlers can be divided into three parts: getting a Web page, parsing a Web page (extracting data), and storing data.

3. Design and Implementation of Crawepe Based on Python
In this paper, through the Python reptile program to climb all the second-hand housing source information on the Xi’an Network, the crawling fields include household type, area, orientation, floor, construction date, unit price, total price, community name. The crawl process of second-hand housing information in Xi’an can be summarized as 4 parts, namely URL request, page data resolution, multi-page crawl, data storage.
3.1. Page Structure Analysis
Just as browsing a web page using a browser, a web crawler needs to send a request to the server through URL. When the server receives the request, it is sent back to the appropriate page. The initial URL of this crawl is the URL of the second-hand housing listing in Xi’an City selected on the home page of the official website. The website is as follows:
https://xa.anjuke.com/sale/?from=navigation

Open the home page of the official website of the resident guest to select the second-hand housing source in Xi’an City, the home page as shown in Figure 1 now to grab Figure 1 Chinese box in the household type, area, orientation, floor, construction date, unit price, total price, community name. View the structure of the page through Google Chrome's developer tool, and the information you want is under the div tag with the value "property-content-info" of the property class.

3.2. Page Data Resolution
Python's requests library can simulate a browser URL to send http requests, and obtain web content through the get method of the request library. The official website does not allow crawling programs to obtain information. If you want to crawl website information, you need to disguise UA. This time imitating the Google browser, set the headers parameter in the get function. The parameter setting format is as follows:
headers = {
    User-Agent': "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/88.0.4324.96 Safari/537.36 Edg/88.0.705.56"
}

When the web page information is obtained, the page can be parsed. We use lxml to parse the page, and parse out the house type, area, orientation, floor, construction date, unit price, total price, and community name that need to be crawled this time.

Using the idea of recursion [7], combined with the etree () function in lxml, parse out the field information of house type, area, orientation, floor, construction date, unit price, total price, and community name, and store the data in the defined variables. The core code is as follows:

```python
div_list=response.xpath('//div[@id="__layout"]//section/div[3]/section[1]/section[2]/div')
for div in div_list:
    unit_type = div.xpath('./a/div[2]/div[1]/section/div[1]/p[1]/span[1]/text()')
    unit_type = unit_type + div.xpath('./a/div[2]/div[1]/section/div[1]/p[1]/span[2]/text()')
    unit_type = unit_type + div.xpath('./a/div[2]/div[1]/section/div[1]/p[1]/span[3]/text()')
    unit_type = unit_type + div.xpath('./a/div[2]/div[1]/section/div[1]/p[1]/span[4]/text()')
    unit_type = unit_type + div.xpath('./a/div[2]/div[1]/section/div[1]/p[1]/span[5]/text()')
    unit_type = unit_type + div.xpath('./a/div[2]/div[1]/section/div[1]/p[1]/span[6]/text()')
    area = div.xpath('./div[1]/section/div[1]/p[2]/text()')
    house_orientation = div.xpath('./div[1]/section/div[1]/p[3]/text()')
    floor = div.xpath('./div[1]/section/div[1]/p[4]/text()')
```

Figure 1. Homepage of Second-hand Housing Information in Xi’an
3.3. Crawl Multiple Pages of Data
Only a single page of housing information can be collected through the designed program. To obtain all the information of second-hand housing in Xi’an, it is necessary to capture multiple pages of information. Through the analysis of the URL of each page. It is observed that the URL of each page is different because the number after the parameter p is changing, and it is also found that the page number value corresponds to the number after p. Therefore, you can set a universal URL: https://nanjing.anjuke.com/sale/p%s/?from=navigation % page. The page number starts from 2, and different URLs can be constructed using recursion. This article crawls the first 100 pages of information from Anjuke’s second-hand housing in Xi’an. The core code is as follows:
```python
for page in range(2,100):
    new_url  = "https://xa.anjuke.com/sale/p%s/?from=navigation " % page
    get_house_info(new_url)
```

3.4. Data Storage
In Python development, data storage and reading are indispensable links, and there are many storage methods that can be used. Common methods include json files, csv files, MySQL databases, SQL Server databases, Redis databases, and Mongdb databases. This article uses SQL Server 2017 for the crawled data storage.
(1) Create a database: anjuke;
(2) Create a data table: house_info_xian;
(3) Write Python code to store the second-hand housing data of Anjuke Xi’an in the SQL Server 2017 database;
(4) View storage results.
Open the database and query the storage results. From the query results, this program crawls a total of 5940 second-hand housing information in Xi’an.

3.5. Program test
In this paper, the designed reptile program is used to test the information of second-hand housing sources in some cities in East China. The test results are shown in Table 1.

| The name of the city | URL | The number of listings to climb |
|----------------------|-----|--------------------------------|
| Hangzhou             | https://hangzhou.anjuke.com/sale/?from=esf_list | 3611 |
| Suzhou               | https://suzhou.anjuke.com/sale/?from=esf_list | 3584 |
| Wuxi                 | https://wuxi.anjuke.com/sale/?from=esf_list | 6001 |
| nanking              | https://nanjing.anjuke.com/sale/?from=esf_list | 5940 |
| Jinan                | https://jinan.anjuke.com/sale/?from=esf_list | 4981 |
| Qingdao              | https://qingdao.anjuke.com/sale/?from=esf_list | 5740 |
| Shanghai             | https://shanghai.anjuke.com/sale/?from=esf_list | 3679 |

4. Data Analysis
The main purpose of collecting data is to tap the hidden value existing in the data, therefore, this paper analyzes the information obtained from second-hand housing in Xi’an, in order to provide some reference for users when buying second-hand housing. Using Excel to analyze the collected data, the
main analysis of different household types in various areas of Xi’an city average house price information.

4.1. Data Processing
The price column of the acquired data has a "meta/m2" behind each price data. To analyze the average value of the house prices of different apartment types in each community, we need to process the data in this column. We use the search and replace function in Excel to change the value in the price column. "meta/m2" is removed. Then perform data visualization analysis on the processed data.

4.2. Data Visualization
We use the Excel data visualization function to analyze the crawled data. Table 2 is part of the results after using the Excel data pivot. From this table, users can see the average house prices of different housing types in different communities. By clicking on the corresponding fields, users can see the detailed housing information in the area. Figure 2 is a data perspective histogram.

### Table 2. Average House Prices of Second-hand Houses in Various Communities in Xi’an

| Neat Type | Average Price (meta/m2) |
|-----------|-------------------------|
| 1 bedroom | 15,000 | 16,000 | 17,000 |
| 2 bedrooms | 19,000 | 20,000 | 21,000 |
| 3 bedrooms | 23,000 | 24,000 | 25,000 |

Figure 2. Average Second-Hand Housing Prices in All Regions of Xi’an

5. Conclusion
The Python-based web crawler program designed in this article can quickly obtain housing information in different cities in a short period of time by changing the URL.

This procedure is not only limited to the acquisition of second-hand housing information in cities in East China, but can also be applied to the needs of buying and renting houses in other cities in China. Buyers and renters can control the housing market dynamics, predict its trends, and then make reasonable decisions. The next step of research focuses on the following aspects:

1. Research on multi-threaded crawler technology to achieve parallel crawling of website data and improve crawler efficiency;
2. Use the data analysis module in Python, and use the data analysis module provided by Python to complete the analysis of a large amount of data.
3) Track the dynamics of crawler technology, continuously learn new technologies, and optimize the program designed in this article later to make it more widely used.

**Acknowledgments**

This work is supported by the "Web Data Acquisition Technology Research Based on Python" under the 2020 Scientific Research Plan Project of Shaanxi Provincial Department of Education" (20JK0515).

**References**

[1] Bai Yang. Design and Implementation of Python Proxy IP Collection Crawler [J]. China New Communications. 2019, (1): 35-36.

[2] Zhao Cao, Rao Jiaodong. Crawling and Analysis of Second-hand Housing Data Based on Python [J]. Computer Knowledge and Technology. 2019, (19): 1-3.

[3] Wang Fucheng, Qi Ping. Python-based Information Crawling and Data Analysis of Second-hand Housing in Hefei [J]. Journal of Jiujiang University: Natural Science Edition. 2020, (3): 49-51.

[4] Liu Hang. Crawling and Analysis of Second-hand Houses in Chongqing Based on Python [J]. Computer Knowledge and Technology. 2019, (36): 6-7.

[5] Weng Shaofi, Liao Xiangyu, Zhu Guangyi, et al. Implementation of Distributed Crawler Collection Software Based on Scrapy [J]. Computer Knowledge and Technology. 2019, (20): 73-75.

[6] Sun Nuan, Cao Xiaoping, Liu Jun. Python-based Internet Recruitment Data Collection Technology [J]. Information and Computer: Theoretical Edition. 2020, (18): 161-163.

[7] Ai Peiyu, Yan Li. Research and Practice of Python-based Internet Data Crawling and Analysis [J]. Information and Computer: Theoretical Edition. 2019, (17): 129-130.

[8] Ren Xitao, Wang Jian. Research on Enterprise Marketing Innovation Based on Big Data Collection and Analysis Technology [J]. Chinese Market. 2020, (31): 116-120.

[9] Wen Zuocheng, Jia Xue. Python-based Web Crawling [J]. Computer Programming Skills and Maintenance. 2020, (12): 23-24.

[10] Dai Yuan, Zheng Chuanxinxing. Python-based Data Crawling and Analysis of Second-hand Houses in Nanjing [J]. Computer Times. 2021, (1): 37-40.