The Crawl and Analysis of Recruitment Data Based on the Distributed Crawler

Jiancai Wang[1(✉)] and Jianting Shi[2]

[1] Office of Academic Affairs, Hei Longjiang University of Science and Technology, Harbin 150022, China
154539860@qq.com

[2] School of Computer and Information Engineering, Hei Longjiang University of Science and Technology, Harbin 150022, China

Abstract. Because of the rapid development of Internet, how to efficiently and quickly obtain useful data has become an importance. In this paper, a distributed crawler crawling system is designed and implemented to capture the recruitment data of online recruitment websites. The architecture and operation workflow of the Scrapy crawler framework is combined with Python, the composition and functions of Scrapy-Redis and the concept of data visualization. Echarts is applied on crawlers, which describes the characteristics of the web page where the employer publishes recruitment information. In the base of Scrapy framework, the middleware, proxy IP and dynamic UA are used to prevent crawlers from being blocked by websites. Data cleaning and encoding conversion is used to make data processing.

Keywords: Distributed crawler · Scrapy framework · Data processing

1 Introduction

With the widespread of modern network, especially after the application of 5G, people are spending more time in searching useful information through piles of data. Therefore the distributed web crawler is adopted to search and obtain Internet data, which can greatly improve the search efficiency. The Internet has been thriving rapidly and changing people's life greatly for decades. According to China Internet development report 2019 issued by The sixth World Internet Conference held in Wuzhen, Zhejiang Province in Oct.20, 2019, China has 0.89 billion netizens and Internet penetration mounted 59.6% [1–4]. There are 5.23 million websites and 281.6 billion web pages. What's more, with the promotion of the commercial process of 5G, there will be a dramatic development in Big data, cloud computing, IOT and data size. The traditional way to find the relevant information is to use Internet search engine, but the efficiency is low if search from Big data, and it is also not conductive to data processing and analysis. Web Crawler, also called Web Spider, was originally developed by Matthew Grey from MIT in 1993 [5]. It is a contemporary to World Wide Web and it can't survive without the Internet by nature. If compare the Internet to a spider net, the web crawler is the crawling spider on the net. By

X. Jiang and P. Li (Eds.): GreeNets 2020, LNICST 333, pp. 162–168, 2020.
https://doi.org/10.1007/978-3-030-62483-5_18

Requesting URL address, the crawlers collect and analyze data by responded contents. For example, if the responded content is html, a DOM structure will be analyzed, parse it and adopt regular match; If the responded content is xml/json, a data objects will be converted and make further analysis.

The distributed crawler uses many computers and many crawlers to coincide with many urls. The distributed system can greatly improve program grabbing efficiency, fault tolerance, and its own Message Queue ensuring the disposable web node to url [6]. There are a great many web crawler researches up to now, and the first crawler is Wanderers [7]. Mercator, based on Java, was developed in 1999 with good scalability. The Apache Foundation published a Java crawler program Nutch [8]. People like Mehdi Bahrami applied distributed crawler to cloud platform. Every crawler, capable of storing a lot of unstructured data, stores results in Cloud Azure, even in Azure Blob. The studies of web crawler are relatively less and late in China. Shanghai University studied distributed crawler based on P2P in 2005. A study of Shanghai Jiaotong University called Igloo, further optimized the performance of distributed crawler system. Therefore, a distributed crawler with a high speed and a high level needs to be devised and applied in the actual situation.

2 Methods

Scrapy is a framework for web scraping developed by Python for scraping web sites and extracting structured data from pages. This framework is more efficient and does not need to consider the problems of multi-process and multi-threading, and the module classification is relatively detailed, which is widely used for data mining and automated testing. Scrapy uses Twisted which is an asynchronous framework to handle network communication. The architecture is clear and includes various middleware interfaces, which can send multiple requests at the same time. Scrapy is a popular Python event-driven networking framework written by Twisted.

2.1 Scrapy Framework

The overall architecture is roughly as follows in Fig. 1.

First, the program gives the initial URL in Spiders, and the engine will pass the URL to the Scheduler; Next, the URL is taken from the scheduler and passed to the Downloader through Requests; After downloading from the Internet, the downloader returns the corresponding response to the Spiders; Spiders will parse and generate two parts, one of which is data and the other is the new URL; The data part is pushed to the project Pipeline for processing data and then the new URL is assigned to the Scheduler and the cycle continues.

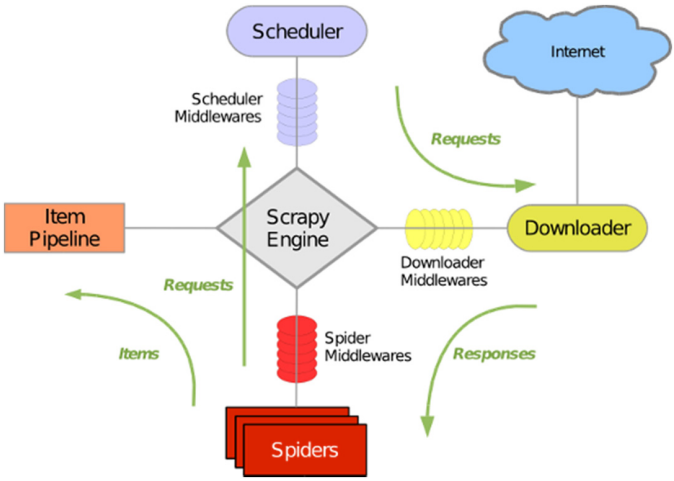

Fig. 1. Scrapy framework

2.2 Distributed Crawler Crawling Data

Open a command line window in PyCharm with Python 3.5, and set the crawler name in the project folder specified. Note: Generally, it is named after the website domain name when creating the crawler file Then there is the spiders folder and its directory structure. A projetc is designed in the crawler file. Define the crawled fields in the parse_item module, as shown in Fig. 2, the fields are necessary for crawling Python-related companies located in Nanjing in a website called "the future worry-free".

```
yield {
    "title" : title,
    "company_name" : company_name,
    "area" : area,
    "exp" : exp,
    "xue_li" : xue_li,
    "ren_shu" : ren_shu,
    "shi_jian" : shi_jian,
    "min_exp" : min_exp,
    "max_exp" : max_exp,
    "min_money" : min_money,
    "max_money" : max_money,
    "address" : address,
}
```

Fig. 2. Defined fields

2.3 Data Cleansing Processing

Data cleaning is the process of re-examining and verifying data, with the goal of removing duplicate information, correcting existing errors, and providing data consistency. The data in each web page is not necessarily the same format. There are many spaces, line breaks, tabs, fields may be misplaced, and the value is empty. Therefore, regular expressions are used to uniformly format the data. Data cleaning is a prerequisite for future data analysis. The specific methods such as using the 'split' method to split a piece of information, the 'strip' method to remove spaces, and then using the regular expression 'findall' to extract the specified content. Finally, these methods replace different data types with the same data type and unify the units. After processing the data, use the 'parse' method to extract the URL of the web page, use the 'for' loop to crawl from the first page, call the parse_item method each time to clean the data, and then parse the source code of the web page until it reaches the specified end. Up to one page. After all the data has been crawled, push them to the pipeline module, and open the database in the pipeline module to achieve the storage of the data. The format after the data cleaning process is completed is shown in Fig. 3.

	id	title	company_name	area	exp	xue_li	ren_shu	shi_jian	min_exp	max_exp	min_money	max_money	address
4	ObjectId("...	python开发...	南京富士通...	南京	无工作经验	本科	招2人	09-23发布	0	0	3.0	4.5	雨花台区文...
5	ObjectId("...	Python开发...	苏宁易购集...	南京	3-4年经验	大专	招5人	09-23发布	3	4	15.0	20.0	徐庄软件园...
6	ObjectId("...	python开发...	南京奥拓电...	南京	2年经验	本科	招1人	09-24发布	2	2	8.0	12.0	南京市雨花...
7	ObjectId("...	Perl/Pytho...	中盈优创资...	南京	3-4年经验	本科	招2人	09-24发布	3	4	12.0	24.0	建邺区嘉陵...
8	ObjectId("...	Python开发...	南京迈特望...	南京-秦淮区	无工作经验	本科	招2人	09-24发布	0	0	8.0	15.0	南京市汉中...
9	ObjectId("...	Python开发...	融资数智科...	南京-建邺区	2年经验	本科	招2人	09-24发布	2	2	10.0	15.0	白龙江东街...
10	ObjectId("...	Python开发...	大汉软件股...	南京	2年经验	本科	招若干人	09-25发布	2	2	6.0	12.0	南京市玄武...
11	ObjectId("...	Python自动...	迪原创新（...	南京-雨花台区	3-4年经验	本科	招若干人	09-25发布	3	4	8.0	15.0	南京市雨花...
12	ObjectId("...	Python开发...	南京中孚信...	南京-浦口区	无工作经验	学历不限	招若干人	09-25发布	0	0	10.0	15.0	浦口大道13...
13	ObjectId("...	平台软件工...	南京汇川工...	南京	无工作经验	学历不限	硕士	招1人	0	0	10.0	20.0	江宁区创研...
14	ObjectId("...	系统架构师(...	南京寰宇信...	南京-江宁区	5-7年经验	本科	招1人	09-25发布	5	7	15.0	25.0	秣周东路12号
15	ObjectId("...	Linux运维工...	北京源晨动...	南京-雨花台区	无工作经验	学历不限	招若干人	09-27发布	0	0	7.0	8.5	华为路北扇...
16	ObjectId("...	软件开发工...	谷游科技（...	深圳-南山区	无工作经验	本科	招3人	09-27发布	0	0	10.0	20.0	软件产业基...
17	ObjectId("...	web渗透工...	江苏苏源高...	南京-玄武区	1年经验	大专	招1人	09-27发布	1	1	8.0	12.0	太平桥南9号
18	ObjectId("...	软件工程师	南京三选纪...	南京-江宁区	无工作经验	本科	招若干人	09-27发布	0	0	10.0	15.0	秣周东路12...
19	ObjectId("...	数据处理工...	数据堂（北...	南京-雨花台区	2年经验	本科	招1人	09-27发布	2	2	6.0	10.0	花神庙邦宁...
20	ObjectId("...	HIL, 英语汽...	狮擎技术管...	南京-江宁区	无工作经验	招若干人	09-26发布	英语精通	0	0	10.0	16.666666...	南京江宁开...
21	ObjectId("...	IT运维工程师	帮维光电科...	南京-浦口区	无工作经验	学历不限	招1人	09-26发布	0	0	4.0	10.0	星火路17号...
22	ObjectId("...	Java高级开...	南京明鼎教...	南京	5-7年经验	本科	招1人	09-26发布	5	7	10.0	15.0	建邺路116...
23	ObjectId("...	数字IC设计...	南京芯视界...	南京-浦口区	无工作经验	学历不限	招1人	09-26发布	0	0	20.833333...	33.333333...	江北新区研...

Fig. 3. Data cleansing processing

3 Results and Discussion

3.1 The Deployment and Implementation of Scrapy-Redis

The Scrapy framework does not support distribution. In order to achieve distributed crawling, foreign software engineers have developed a distributed crawler framework based on redis, allowing crawlers to have distributed crawling capabilities. The principle of this module is somewhat similar to big data, that is, distributed work, multiple machines work together to complete a goal based on the same URL. The machine that

stores the URL list and uniformly manages the URL is called the master server, and other machines running crawlers become slaves.

Due to the queue mechanism of scrapy-redis, the links obtained by slaves node will not conflict with each other. In this way, after each slave completes the fetching task, the obtained results are summarized on the server. Experiments can be performed on the cluster. Without so many machines, multiple Linux virtual machines can be deployed on one physical machine. First, write a configuration file on the master to connect to the redis database, and copy the previously written crawler programs to the Linux virtual machine slaves. The crawling out of order is started and the initial URL of the website is run to crawl in redis-cli. Note that the URLs of the crawled web pages should be staggered. Finally, the data in redis is imported into mongodb. Figure 4 shows the process of distributed crawling. Figure 5 shows the data stored in the database after the crawling is completed.

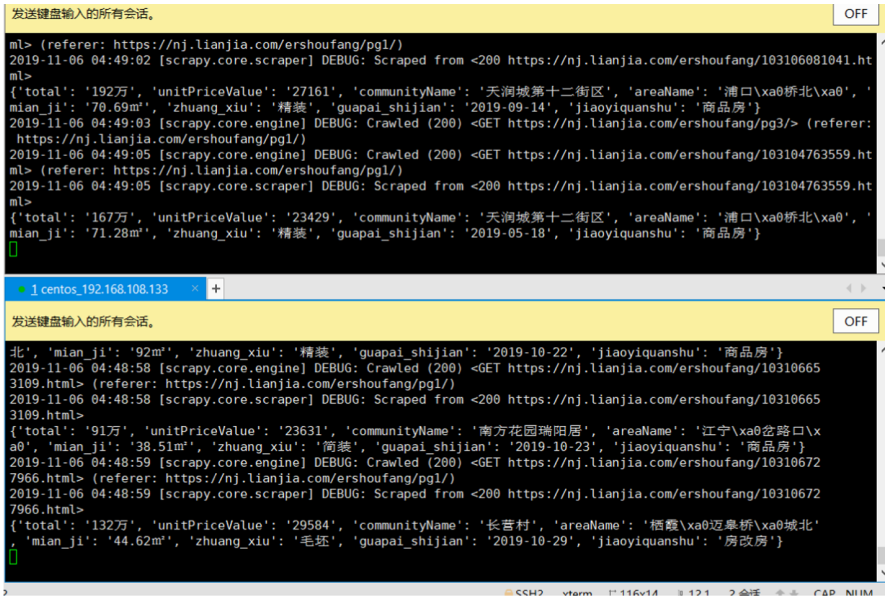

Fig. 4. Distributed crawling process

3.2 Data Visualization Analysis

Data visualization technology converts data into graphs and charts to provide a basis for decision making. The research of data visualization technology has developed rapidly and achieved corresponding achievements. An analysis of the salary situation of Nanjing IT companies, and analysis of the range of the highest and lowest wages are available. In the IT industry, the minimum wage is generally taken when starting a job, and most can get nearly 10 k. After working for a few years, most of the maximum salary can

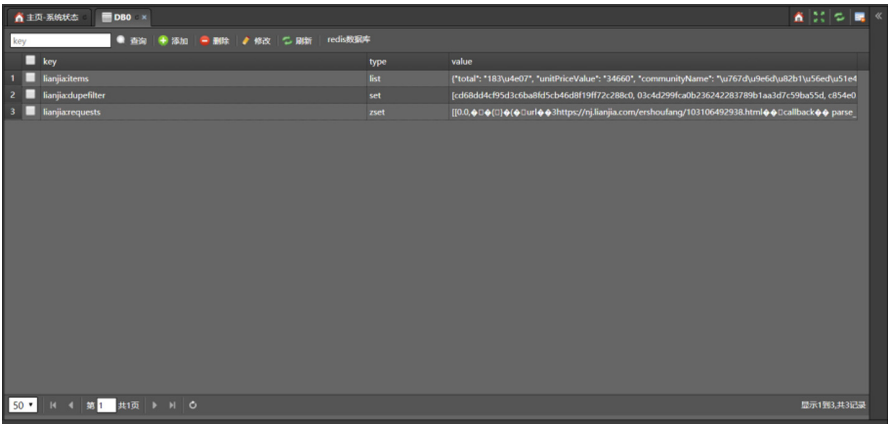

Fig. 5. Data stored in redis

reach 20 k, which is in line with the current average salary status. The data visualization analysis of salary is shown in Fig. 6.

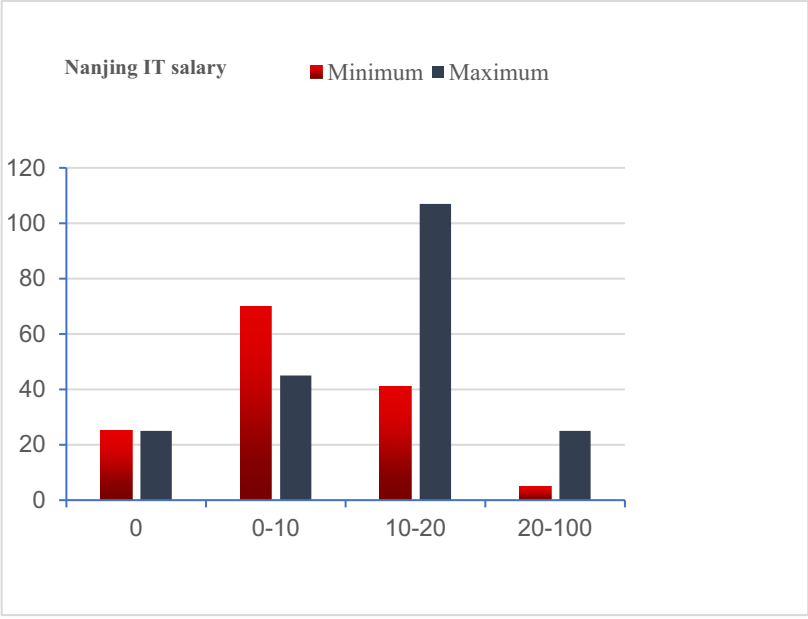

Fig. 6. The bar chart of Salary

The data visualization analysis of education is shown in Fig. 7. Undergraduates in the IT industry account for more than 50% of the undergraduates, with the highest proportion, while there are almost no PhDs, indicating that the IT industry is an industry for the youth only and attaches great importance to the application of technology.

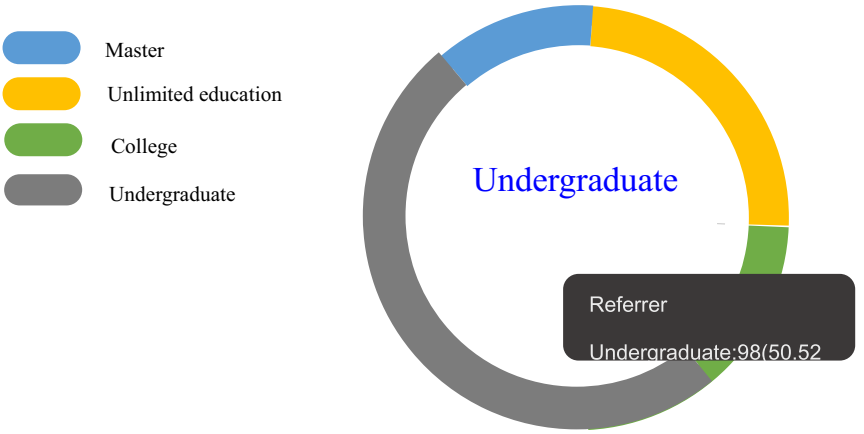

Fig. 7. The pie chart of Education

4 Conclusion

With the development of network technology, how to use crawler and data visualization technology to better understand users and their intentions is a key area in the network era. This paper studies how to collect relevant data from the Internet. Distributed technology makes data collection more efficient. It cleans and filters the collected data, and then displays the useful data and visual analysis to analyze and mine valuable information. There will make full use of the potential value of big data.

Acknowledgements. The work was supported by Hei longjiang Fundamental Research Funds for the Local Universities in 2018 (No. 2018-KYYWF-1189).

References

1. Jie, H.: China's Internet users reached 829 million and 5 G industrialization achieved preliminary results. CNNIC released a report (2019)
2. Biling, G.: 5G key technology and its impact on the internet of things. J. Wirel. Internet Technol. **4**(7), 30–32 (2019)
3. Yuezhong, S., Chao, C., Bin, S.: Analysis of communication technology and challenge of Internet of things in 5G network. J. Inf. Syst. Eng. **7**, 37–38 (2019)
4. Jingke, Z.: 5G key technology research. Inf. Commun. **3**, 261–263 (2018)
5. Sunguo, C.: Research on the implementation of web crawler in subject search engine. Comput. Knowl. Technol. **12**(17), 23–25 (2016)
6. Xiaohui, W.: Improvement of URL de duplication strategy for distributed web crawler. J. Pingdingshan Univ. **24**(5), 116–119 (2009)
7. Yuhao, F.: Design and implementation of distributed web crawler system based on scratch. Chengdu:University of Electronic Science and Technology of China (2018)
8. Hengfei, Z., Yuexiang, Y., Hong, F.: Research and optimization of nutch distributed network crawler. J. Front. Comput. Sci. Technol. **5**(1), 68–74 (2011)