Employment Data Analysis based on Python Crawler Technology

Li Li

College of Information Engineering, Yunnan Vocational College of Mechanical and Electrical Technology, Kunming 650000, China

Abstract. With the explosive growth of network information and the advent of the era of big data, it is of great significance to analyze and process employment data by using web crawler technology. This article takes Lagou.com as an example, uses crawler technology to collect data on the basis of Python and MySQL, and analyzes the collected employment data in various aspects, and uses these data analysis results to help college students in their employment and career planning. Provide reference basis, provide objective reference.

Keywords: Python; Crawler; Employment; Data Analysis.

1. Introduction

In recent years, with the progress of human society and the rapid development of the Internet industry, the amount of social information has been accumulating and growing explosively. It can be said that the era of big data has come in an all-round way. At the same time, as an important way for enterprises to recruit talents, online recruitment has the advantages of low cost, flexibility, strong pertinence, abundant resources and wide coverage, compared with offline recruitment with cumbersome process, time-consuming and narrow communication channels. It is usually the source of information that contemporary college students need to pay close attention to when seeking jobs. With the growth of online recruitment scale, a large number of recruitment websites appear, and the recruitment information contained in these websites also provides researchers with a large amount of data. Therefore, by analyzing the data of the recruitment website, we can explore the requirements of relevant enterprises for professional ability and talent demand, and at the same time provide a more objective reference for college students.

2. Related Technologies and Technical Indicators of the Research

2.1 Python

Python is a cross-platform, open source and free interpretive high-level dynamic programming language. It supports imperative programming, object-oriented programming and functional programming. It includes a perfect and easy to understand standard library and a rich third-party open-source library, so that users can easily complete development tasks.

2.2 Web Crawler

Since Matthew Gray's Wandered developed the first recorded web crawler at the Massachusetts Institute of Technology in early 1993, crawler technology has undergone more than 20 years of development, and the technology has become increasingly diverse. In order to meet the diverse needs of different users, many types of crawler systems have been created and developed. According to the system structure and implementation technology, web crawler can be roughly divided into general web crawler: whole web crawler, focused web crawler, incremental web crawler and deep web crawler. Crawler technology is quickly used in search engines or other related websites to get the content and other data of data websites in time. The web crawler can automatically collect all accessible web pages and their contents by setting, and the collected data can be further processed by the search engine, so that users can accurately obtain the required information at the first time.

According to the system structure and implementation technology, web crawler can be roughly divided into the following types.

2.2.1 Universal Web Crawler

Universal Web crawler, also known as the whole web crawler, extends its crawling objects from a batch of seed URLs to the whole network, which is mainly used by search engines or large web service providers. The crawling range and quantity of this kind of crawler are relatively large, and the requirements for crawling speed and storage space are relatively high, while the requirements for crawling page order are relatively low, so parallel work is usually adopted to deal with a large number of loan refresh pages.

2.2.2 Focus on Web Crawler

Focused web crawler, also known as topic web crawler, refers to selectively crawling web crawler related to predefined main pages. Compared with the general web crawler, the focus crawler only needs to crawl the pages related to the topic, which greatly saves hardware and network resources, can update the saved pages faster, and meet the needs of specific people for information in specific fields.

2.2.3 Incremental Web Crawler

A crawler that incrementally updates crawled webpages or crawls only newly generated or changed webpages. This mechanism can ensure that the crawled pages are as new as possible to some extent. Compared with other web crawlers that periodically crawl and refresh pages, incremental web crawlers only crawl newly generated or updated pages when needed. However, unchanged pages are not crawled, which can effectively reduce the amount of data downloaded and update crawled pages in time, reducing the waste of time and storage space. However, the disadvantage is that the algorithm is more complex and difficult to implement.

2.2.4 Deep Web Crawler

According to the way of existence, Web pages are divided into two types: surface pages and deep pages. Surface pages are pages that can be indexed by traditional search engines, mainly static pages that can be reached by hyperlinks. Deep pages are web pages that can't be obtained through static links, are hidden behind the search form, and can only be obtained after users submit keywords. The most important part of deep web crawler is form filling.

2.3 Data Cleaning

Data cleaning is the last procedure to find and correct identifiable errors in data files, including checking data consistency and dealing with invalid and missing values. Different from the questionnaire review, the data cleaning after entry is generally completed by computer rather than manually. The data cleaning path is divided into six stages: pretreatment stage, removal or completion of missing data, removal or modification of data with wrong format and content, removal or modification of data with logical errors, removal of unnecessary data and relevance verification.

2.4 Data Analysis

Data analysis is to analyze a large amount of data collected by appropriate statistical analysis methods, and summarize, understand and digest them, so as to maximize the function of data and play the role of data. Data analysis is the process of studying and summarizing data in detail in order to extract useful information and form conclusions.

3. Necessity of Research

For colleges and universities, the most important job is employment. The future development of students is closely related to this job, and the reputation of the school will also be directly affected by it. This year, 9.09 million college graduates, an increase of 350,000 year-on-year, the number of

graduates hit a record high. At present, the employment situation is complicated and severe due to the superimposed influence of multiple factors such as the new crown pneumonia epidemic and economic downward pressure. In order to actively respond to this year's complex and severe employment situation, the Ministry of Education, in conjunction with relevant departments, has focused on promoting the employment of college graduates. At present, the Internet is becoming more and more developed, various kinds of employment information are spreading faster and faster, and there are more and more types of employment information on the recruitment platform. In the employment work of colleges and universities, various network platforms are used to broaden the channels of obtaining employment information. For graduates to find a good job, the release of various recruitment information and consultation on employment information will be carried out through various online communication software and relevant employment websites. Such employment initiatives provide rich data resources for employment analysis. Therefore, through web crawler technology, multi-dimensional analysis of employment data information is particularly important to promote employment in Colleges and universities.

For graduates, with the explosive growth of network information and the advent of the big data era, it is of great significance to analyze and process employment information by using web crawlers. When college graduates are approaching graduation, there will be a lot of confusion more or less, and it is difficult to make the right choice among numerous job information. Information retrieval through the Internet will get a lot of information that graduates don't need, and only through manual screening, summary and comparison, can they finally get the information they want.

Through extensive access to relevant information at home and abroad in the early stage, it is found that most of the research points on employment in colleges and universities are from the perspective of pedagogy. For example, the thinking of the employment service mechanism in colleges and universities, the ways of high-quality employment for college graduates, the implementation mechanism and path selection of precise employment in colleges and universities, and the exploration of reasonable employment expectations of college graduates, etc. In the process of analyzing the graduates' employment situation in colleges and universities, due to the huge amount of data, the accuracy of the original analysis method is poor. With the increasing employment pressure, we can accurately and comprehensively obtain the data of mainstream recruitment websites through web crawler technology, clean the data, analyze the data from multiple dimensions, and finally form a comprehensive research report. It can help college students avoid employment risks and correctly understand their own value. It has very important research value. It can not only provide relevant information for colleges and universities to reasonably allocate educational resources, but also help college employment departments to promote students' high-quality employment.

4. The Content and Significance of the Research

This paper will develop a set of special crawlers for recruitment and employment. Taking the mainstream recruitment website as an example, this paper will study and design the crawler program for crawling the information of the mainstream recruitment website, obtain the website recruitment information, clean and analyze the obtained recruitment information, visualize the data, and finally form a comprehensive research report. The application value of employment in colleges and universities has the following three points:

The first is to organize and analyze employment data by means of web crawlers, data cleaning, data analysis, etc., which can predict the general direction of social development, so that college students can have a reasonable employment plan and better meet social needs in the future.

The second is to provide relevant information for the reasonable allocation of educational resources for colleges and universities. According to the existing major and market demand, optimize the training scheme of professional talents. For those industries that are developing rapidly in the future, colleges and universities should set up related majors and train a large number of professionals.

This not only meets the needs of relevant talents in this industry, but also ensures the employment of students.

Third, it is helpful to promote the informatization construction of employment. We are in an era of globalization. By analyzing a large amount of employment information, we can give targeted guidance to students in employment, make the employment guidance work in colleges and universities more perfect, and make the construction of employment informatization complete as soon as possible.

5. Employment Data Analysis

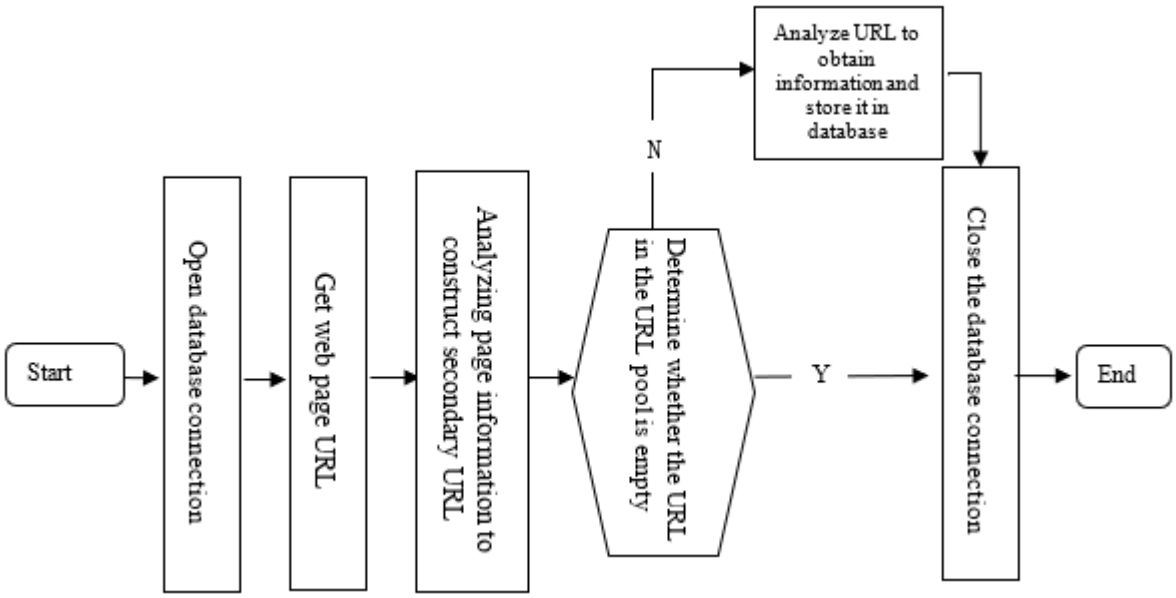

Fig 1. The storage methods

This paper analyzes the storage methods of Lagou.com's recruitment information for key data acquisition research, develops based on Python technology, and comprehensively uses multiple Python network function libraries. And get information about work location, work experience, academic requirements, financing stage, company size, industry area, salary and other information from it. Finally, save the data in the database for statistical analysis and provide data for later employment recommendation.

This article uses the data published on Lagou since 2021 as the search scope, and crawls all the job recruitment information data related to Python. In the process of information collection, the search range URL is used as the target URL of the web crawler, and the requests library is used to download the target web page to obtain the instantiated object. The obtained information is parsed using the bs4 library to parse the web page to obtain the required information, and assign values to variables respectively. Use the time library to control the frequency, and stop for 30 seconds every time a page is crawled to avoid high frequency access. After the information collection is completed, connect to the MySQL database and use the for loop to traverse the data row by row and store it into the database.

The crawler collected more than 10,000 pieces of recruitment information. In order to ensure the reliability of data and the authenticity of research, it is necessary to preprocess the selected samples. When processing samples, it is necessary to remove or complete missing data, remove or modify data with logical errors, and remove unnecessary data.

For employment data analysis, the most important thing is salary analysis. For each element of the extracted salary list, use regular expression to match the first two numbers, that is, the upper and lower salary limits of this job, find an average value, traverse the whole list, and make statistics on salary distribution. At the same time, by positioning and including character segments, the work and corresponding salary of each region can be extracted, and then the average salary of each region can

be analyzed through job statistics of each region. On this basis, the salary level of each industry is calculated according to the number of posts in each industry and the corresponding average salary.

The analysis results show that the salary level of "Python" post is generally satisfactory. Mainly concentrated in the monthly salary of 10 ~ 25 K, the monthly salary above and below 20 K accounted for about 50% respectively, and the overall income level remained at a relatively high level. Some people can get a high salary of 40k/ month. The median salary distribution in Shanghai is about 22k, ranking first in China. Secondly, the median of Beijing and Suzhou is about 20k, while the median of Jinan is only about 8k. It can be seen that the development of Python is mainly concentrated in Beijing, Shanghai, Shenzhen, Suzhou and other regions. A bachelor's degree is a stepping stone to the Python industry. If you have rich work experience and strong technical capabilities, you will also have advantages.

6. Conclusion

In this paper, through the research of web crawler technology, taking Python-related posts of Lagou.com as an example, we can deeply understand the concept of web crawler, crawler strategy and other aspects. By obtaining the target page information, analyzing the URL in the page, and at the same time filtering the constructed URL again, and storing it in the database. On this basis, the data will be deeply mined and a series of data analysis methods will be used, to obtain a series of important information about the salary of Python jobs, recruitment needs, etc., to provide useful reference and reference for the employment of college students.

References

[1] Zuo Weigang. Research on Web Crawler of News Aggregation System Based on Python [J]. Journal of Changchun Normal University, 2018, 37(12): 29-33.

[2] Xing Li, Wu Maonian. Design and Application of Recruitment Theme Crawler [J]. Computer Knowledge and Technology, 2018, 14(25): 73-75.

[3] Zheng Dingchao, Ma Shaoqiu. Research and Design of Web Crawler [J]. Computer Knowledge and Technology, 2018, 14(25): 43-45.

[4] Yang Guozhi, Jiang Yefeng. Design and implementation of data collection system for focused web crawlers based on Python [J]. Science and Technology Innovation, 2018(27): 73-75.

[5] Chen Le. Web crawler technology based on Python[J]. Electronic World, 2018 (16): 163-165.

[6] Tian Xiaoling, Fang Yuan, Jia Minzheng, etc. Keyword web crawler design based on data analysis[J]. Journal of Beijing Polytechnic Institute, 2018, 4(17): 38-45.