Transformation of Scientific Research Projects into Teaching Content for Artificial Intelligence Majors in "Web Crawler" Unit

Yiyang Sun*, Shuxi Chen

School of Computer and Information Engineering, Nantong Institute of Technology, Nantong, China

*justsososun@sina.com, 835004954@qq.com

Abstract. Cultivating advanced specialized talents and promoting scientific and technological development are important functions of higher education, and the two complement each other and have unity. This paper practices the teaching concept of research feedback, integrating teacher research and teaching, and using project-based teaching methods to cultivate applied undergraduate talents. We organize the course content of "Python Programming" according to the training needs of artificial intelligence majors. This paper introduces the teaching design of the unit "web crawler". Through teaching practice, we have realized that integrating scientific research into teaching can improve teaching quality and better cultivate high-quality talents.

Keywords: higher education, applied undergraduate, Python Programming, artificial intelligence, web crawler.

1 Introduction

Method of applying scientific research to teaching refers to the introduction of the latest achievements, research methods, and ideas of teachers in scientific research into teaching, making the teaching content more rich, in-depth, and cutting-edge. Through research practice experience and case studies, students can better understand and master theoretical knowledge, improve teaching quality and level, and promote the cultivation of their research literacy and innovation ability. Method of applying scientific research to teaching is a teaching model that organically combines research and teaching, which can improve teaching effectiveness, promote discipline development, and cultivate more innovative talents[1-3].

Artificial intelligence is an emerging interdisciplinary field that is based on computer science and combines theories, methods, and technologies from multiple disciplines such as computer science, psychology, and philosophy. Its aim is to research and develop theories, methods, technologies, and application systems that can simulate, extend, and expand human intelligence. The research fields of artificial intelligence are

M. Yu et al. (eds.), *Proceedings of the 2024 5th International Conference on Big Data and Informatization Education (ICBDIE 2024)*, Advances in Intelligent Systems Research 182,
https://doi.org/10.2991/978-94-6463-417-4_9

extensive, including robotics, language recognition, image recognition, natural language processing, and expert systems[4-5].

"Python Programming" is a fundamental course for artificial intelligence majors, providing students with basic knowledge and skills in "Python Programming", including syntax, data types, control structures, functions, data structures, exception handling, and other aspects of python. Through this course, students will master the basic skills of "python programming" and be able to write simple programs to solve practical problems.

As an automatically executed program, the core task of web crawlers is to simulate the browser behavior of human users. Web crawlers use network protocols and structures to automatically access and extract webpage data, thereby collecting and storing website information. As a comprehensive module in "Python Programming", it not only covers the technologies of network communication and web page parsing, but also delves into knowledge in multiple fields such as data storage, processing, and analysis. Through in-depth learning and practice of web crawler technology, students majoring in artificial intelligence can better understand and apply this comprehensive tool, and cultivate a profound understanding of data mining and analysis, laying a solid foundation for future career development. The application of this technology is not limited to obtaining information, but also provides students with unique opportunities to process and analyze large-scale data through automation. So that they can better deal with complex problems in the real world and achieve more significant achievements in the field of artificial intelligence[6-9].

2 Fundamentals of Crawler Technology

2.1 Fundamentals of Web Page Technology

There is a unique URL for every web pages, images, videos, and other resources on the internet. When we enter the URL in the browser and press enter, we can view the page content. The process involves the browser sending a request to the server where the website is located. After receiving the request, the website server processes and parses it, ultimately generating response content and returning it to the browser. After the browser parses the source code of the webpage in the response, it presents the webpage, as shown in figure 1.

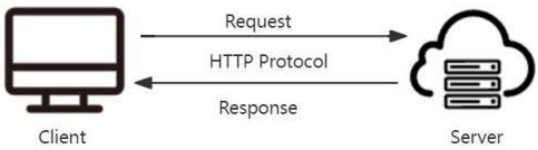

Fig. 1. Web page request process

2.2 Website composition

Generally, webpage content consists of HTML, CSS, and JavaScript. When crawling data from static HTML pages, it is necessary to simulate a browser sending a visit request to the website server to obtain its response content. The remaining task is to parse the webpage content, extract the required input, and store it in the database. The crawler process is shown in figure 2.

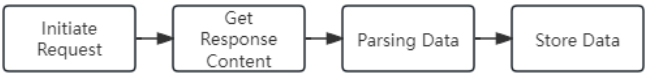

Fig. 2. Crawler process

Usually, websites set certain thresholds to block web crawlers, in order to reduce server load and protect their own data from being crawled. Common anti-crawler measures include access frequency verification, request header verification, etc. Therefore, a core technology of web crawlers is how to construct a network request that is as real as possible to mimic the behavior of browsers. In the process of accessing a website through a browser, it is not only necessary to specify the URL to be accessed, but also to consider the browser's own parameter information, such as User Agent, as well as key information such as Sessions or Cookies required for interaction with the website. Therefore, when configuring web crawlers to send access requests, it is necessary to carefully construct these parameter information to ensure the legitimacy and validity of the request.

User Agent is an important browser parameter that contains information about the browser's identity, such as the type, version, and operating system of the browser. By simulating legitimate User Agents, web crawlers can better simulate the access behavior of real users and reduce the risk of being detected as crawlers by websites. In addition, for websites that require user authentication or maintenance of status, information such as Sessions and Cookies is also essential. Their correct settings in the request header are key to ensuring user identity and maintaining login status.

Therefore, developers of web crawlers need to carefully construct access requests, including setting the correct User Agent, carrying necessary session and cookies information, in order to interact smoothly with the target website. By properly configuring these parameter information, web crawlers can more effectively obtain the required data, while reducing the probability of being blocked or restricted by websites, thereby improving the stability and availability of the crawler.

2.3 Data Extraction

After initiating a network request, the response content returned by the website can be obtained, and the remaining task is to parse the webpage and extract page data. Regular expressions can generally be used, but they are cumbersome and not convenient enough. There are many HTML parsing libraries in Python, such as LXML, Beautiful

Soup, PyQuery, etc. Using such libraries can greatly improve the efficiency of web data extraction[10-14].

2.4 Data Storage

After data extraction, it is necessary to consider persistently storing the obtained information to ensure the long-term preservation of the data and facilitate subsequent retrieval and analysis. There are various forms of data storage, and you can choose to save it in JSON format, CSV format, TXT text format, or directly store it in a database. In this process, students need to apply the database related technologies they have learned.

3 Course Implementation

3.1 Task Background

Currently, fresh graduates not only consider the correlation between the applied position and their major when seeking employment, but also factors such as the geographical location of the employment city and the salary and benefits of the position. In order to help graduates better understand the job market situation, we plan to collect some recruitment information from a certain employment website, and extract the job title, salary and benefits, educational requirements, company name, work location, etc. from the recruitment information. We need to analyze and display recruitment data from different dimensions, so that graduates can easily understand the job market situation and choose positions that are more suitable for them.

3.2 Task Implementation

Analyzing the source code of the target website page, it was found that all job data information is contained in each unique div box, which is also the main crawling object for our subsequent crawling work. Through careful analysis of the source code of the target website page, we found that all job data information is nested in each unique div box. This is our main crawling object, as these unique div boxes carry the key information we need. In each div box, we can find key data including job titles, company information, salary benefits, etc., which enables us to extract and analyze the required job information in a targeted manner. This structured page design provides clear goals for crawler work, allowing us to access and extract the required data in an orderly manner. This clear structure lays a solid foundation for our crawler tasks, enabling us to efficiently obtain and process job data, providing a data source for subsequent data processing and analysis work.

Because many recruitment websites have adopted a monitoring mechanism for Selenium, if the Selenium value is true when accessing web pages, the crawler can be easily detected, leading to crawling failure. However, setting the ChromeDriver startup

parameter can solve this problem, avoid Selenium value detection, and achieve the purpose of hiding Selenium. In order to improve the crawling efficiency of the crawler, the main program calls the Threading package, sets four groups of data keywords in a list at the same time, and circularly uses different threads to crawl the four keywords, thus realizing multi-threaded crawling.

View the webpage source code by accessing the job details page. Design crawlers using this information as crawling objects. Use the method of annotating the keyword primary key to crawl all valid information within the primary key range until the last position information is crawled. View the source code information as shown in figure 3.

Fig. 3. Original website data

By parsing the website source code and locating according to the attributes of the tags, you can find the required job information by using the Find_element_by_xpath() function in the Selenium library under different nodes. This involves the web knowledge that students have learned.

3.3 Task Expansion

In order to improve the running speed of the crawler, multi-threaded technology was adopted to achieve concurrent execution and achieve multi-threaded crawling. We call the Threading library in the main program to create multiple threads, store the search keywords in a list, distribute different search keywords to different threads at runtime, and start and run these threads at the same time, so as to improve the crawl rate of crawlers. To avoid errors when multiple threads defend the same piece of data, it is also necessary to use Threading. lock() to prevent this issue from occurring.

If the same IP address is frequently used for web crawling, the website will recognize it as a crawler and no longer return response data. To avoid a single IP address being denied access due to high access frequency, we can use an IP proxy to disguise the IP and make the server unable to recognize that the request was initiated by our local machine. Using an IP proxy server can build a bridge between the local machine and the server. At this point, the local machine does not directly make requests to the web server, but sends requests to the proxy server, which forwards them to the web server. Finally, the proxy server forwards the response content returned by the web server to the local machine. The process is shown in figure 4. During this process, the IP address

recognized by the web server is the proxy server's, which achieves the disguise of the IP address. In practical use, it is necessary to use more IP proxy servers for cyclic or random use.

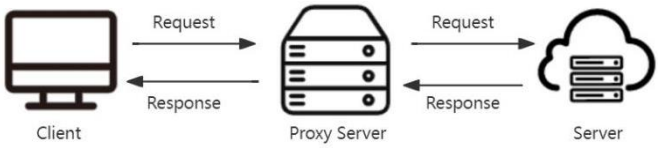

Fig. 4. Running process of IP proxy server

In order to store the data for future use, we choose to store the recruitment data in a relational database and store the recruitment information in a data table in order of fields. This involves the database knowledge that students have learned.

After data storage, high-frequency vocabulary can be displayed for unified type of job information, which needs to be divided into two parts: vocabulary decomposition and heat map production. Vocabulary decomposition uses the jieba library to perform Chinese word segmentation on recruitment information. The segmented text is stored in a list and sorted by frequency. The most 25 keywords are taken out for creating a word cloud, which is then created using the WordCloud library. The effect is shown in figure 5.

Fig. 5. Wordcloud of igh frequency words

Finally, perform a simple data analysis on the crawled recruitment data, visually display the salary levels of relevant positions and the salary benefits of enterprises of different scales, and use the Matplotlib library to achieve this. After extracting database information, create bar charts, pie and line charts. We have read recruitment information from different urban areas in Beijing and made statistics on the average recruitment salary in different urban areas. The bar chart shows that the average salary in remote urban areas is lower, as shown in figure 6. At the same time, statistics were conducted on the relationship between the size of the company's personnel and salary benefits, which was shown in the line chart. It can be seen that the larger the size of the personnel, the better the average salary, as shown in figure 7.

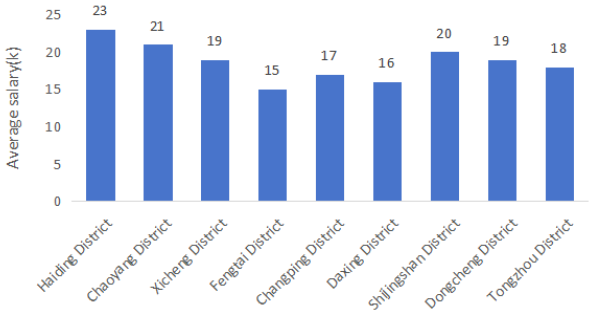

Fig. 6. Average salary in different district

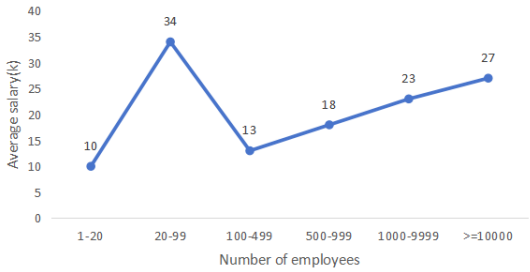

Fig. 7. Average salary under different company size

4 Summary

A complete web crawler unit involves various technical knowledge, such as Python programming, web development, multithreading technology, database technology, etc. If only explaining network data crawling separately, students will not be able to master the complete crawling process and crawling technology. For students majoring in artificial intelligence, learning Python programming language can help them master basic programming skills, understand how computers process data and perform tasks, as well as how to use various artificial intelligence libraries and frameworks. Using web crawlers can crawl large amounts of data from the internet, providing necessary datasets for machine learning and deep learning. In addition, web crawlers can also be used to automate the collection of industry information, competitor analysis, and other fields, improving data processing efficiency. Mastering web crawler skills can help students majoring in artificial intelligence better process and analyze data, and improve research efficiency. Python programming and web crawling are essential skills for students majoring in artificial intelligence. By learning and practicing these two skills, students can achieve better results and development in the field of artificial intelligence.

By integrating research projects into practical courses, students can gain a deeper understanding of the cutting-edge dynamics of computer science and master advanced research methods and skills through exposure to research projects. Meanwhile, students

can also enhance their innovation and problem-solving abilities by participating in scientific research projects. By introducing scientific research achievements and cutting-edge academic knowledge into classroom teaching, teachers can enrich teaching content and improve teaching quality. At the same time, teachers can also provide feedback through scientific research to help students better understand and master the basic concepts and principles of computer science. Computer related research projects usually require experimentation and practice, which can provide students with more practical opportunities. Through practice, students can better understand and apply the knowledge they have learned, and improve their practical and programming abilities.

Acknowledgement

(1)Universities Philosopthy and Social Science Research Projects of Jiangsu Province(No.2022SJYB1758 and No.2023SJYB1720).

(2)Universities Natural Science Research Projects of Jiangsu Province(No. 22KJB520032).

References

1. J. Li, “The Rationality of Research Backfeeding Teaching and Local Universities' Response Strategies,” Education Research, 2012, 33 (03): 53-56+70.
2. Z. Li and Q. Xie, “The Transformation of Scientific Research Projects into Professional Teaching Content - Taking the ‘Attention Mechanism in Computer Vision’ Unit as an Example,” Science and Technology Wind, 2023 (32): 37-40+169.
3. Q. Li, W. Xu and H. Shen, “Exploration and Practice of the ‘Research Backfeeding Teaching’ Model in Applied Undergraduate Colleges,” Computer Engineering and Science, 2019,41 (S1): 153-156.
4. H. Qin, J. Chen and S. Wang, “The concept and implementation methods of scientific research feedback teaching,” Education and Teaching Forum, 2019 (10): 233-235.
5. C. Miao, L. Chen and Q. Du, “Collaborative Innovation between Teachers and Students in Teaching through Scientific Research,” Computer Education, 2017 (03): 92-94.
6. X. Gao, F. Yuan and J. Fan, “Python based crawling program for collecting data from websites,” 6th International Workshop on Advanced Algorithms and Control Engineering (IWAACE 2022). Vol. 12350. SPIE, 2022.
7. Y. Wang, “Research on Python Crawler Search System Based on Computer Big Data,” 2023 IEEE 3rd International Conference on Power, Electronics and Computer Applications (ICPECA). IEEE, 2023.
8. T. Dhar, S. Mazumder, S. Dhar, et al. “An Approach to Design and Implement Parallel Web Crawler”,2021 International Conference on Smart Generation Computing, Communication and Networking (SMART GENCON). IEEE, 2021: 1-4.
9. K. Vayadande, R. Shaikh, T. Narnaware, et al. “Designing Web Crawler Based on Multithreaded Approach For Authentication of Web Links on Internet”,6th International Conference on Electronics, Communication and Aerospace Technology. IEEE, 2022: 1469-1473.
10. H. Salem and M. Mazzara, “Pattern matching-based scraping of news websites,” Journal of Physics: Conference Series. IOP Publishing, 2020, 1694(1): 012011.

11. X. Zhang, Y. Chen and J. Liang, “Research on Anti-crawler and Anti-Anti-crawler Technology,” 2022 International Conference on Informatics, Networking and Computing (ICINC). IEEE, 2022: 35-39.
12. F. Zhou and Y. Wang, “Exploring The Role of Web Crawler and Anti-Crawler Technology in Big Data Era,” 2022 11th International Conference of Information and Communication Technology (ICTech)). IEEE, 2022: 316-319.
13. J. Wang and J.Shi , “The Crawl and Analysis of Recruitment Data Based on the Distributed Crawler,” Green Energy and Networking: 7th EAI International Conference, GreeNets 2020, Harbin, China, June 27-28, 2020, Proceedings. Springer International Publishing, 2020: 162-168.
14. J.Gao, “Distributed Collection Method of Economic Growth Data Based on Cloud Computing”, 2022 6th International Conference on Wireless Communications and Applications (ICWCAPP). IEEE, 2022: 238-241.