The Design and Implementation of Academic Search System based on crawler

Yu Zhuang [1, a], Taizhi Lv[1, b] and Jingwen Han[1, c]

[1] School of Information Technology, Jiangsu Maritime Institute, Jiangsu Nanjing 211170, China

[a]2200307645@163.com, [b]lvtaizhi@163.com, [c]1604363663@qq.com

Abstract

The paper proposes an important index to evaluate teachers'scientific research ability. The research and development of an analysis platform for teachers' scientific research ability oriented to HowNet literature can provide data support for better promoting the development of school scientific research work. The project builds an automated analysis platform, uses crawler technology to obtain literature data, based on large data technology to analyze various indicators, and uses visualization technology to show the results.

Keywords

Python; Crawler; SSH; MySQL.

1. Introduction

As we all know, the three basic functions of contemporary universities are personnel training (teaching), academic research (scientific research) and social services. Academic papers are the criteria to measure teaching achievements, the condensation of academic views and the crystallization of painstaking efforts, and also the important achievements of education, which directly reflect the quality of higher education. The latest research results show that the scientific research achievements of university teachers usually represent the scientific research level of a country. At present, it is an urgent problem to transform scientific and technological achievements into productive forces on the basis of theoretical and practical research.

Scientific research ability is the core index reflecting the strength of running a university, and it is also an important guarantee for the professional construction of a university. With the rapid development of higher vocational education in the past 20 years, the scientific research consciousness, ability and achievements of higher vocational colleges have been greatly improved [1]. However, the development of scientific research ability is not balanced due to the different development basis, development environment and nature of running schools. Some schools have developed very fast, and their scientific research ability even surpasses that of ordinary undergraduate colleges; some schools have developed slowly, but their scientific research ability is still significantly improved compared with the original foundation of secondary technical schools or staff universities; some schools have stagnated or even retrogressed in some aspects due to the restrictions of subjective and objective conditions. By analyzing the current situation of teacher scientific research work from CNKI（China national knowledge internet）, this paper objectively evaluates teachers scientific research ability and finds out the bottleneck that restricts teachers scientific research ability. It is of positive significance to determine the development strategy of school's scientific research work and accelerate the improvement of scientific research ability.

2. Crawler Implementation

The crawler is based on Python language and uses the school name as a search term to crawl the corresponding college papers. The crawled content includes the title of the paper, author, published journal, publishing time, number of introductions, downloads, abstracts, publishing institutions, fund projects, whether SCI index, whether EI index, whether Chinese core, whether CSSCI, whether CSCD. The whole crawling process is show in Fig.1.

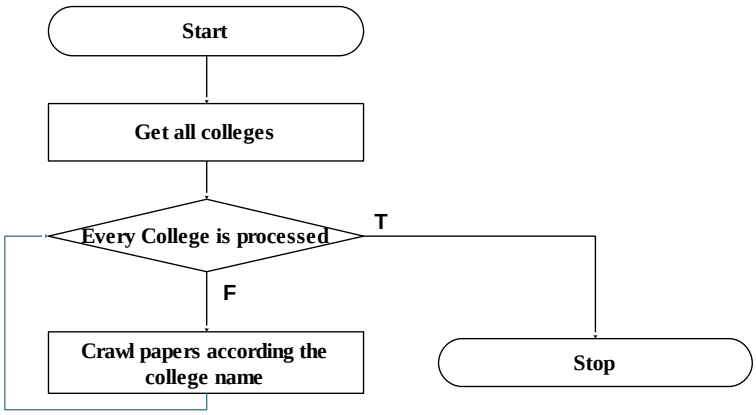


Figure 1. The whole flowchart

Python provides a convenient library to help implement many crawler functions [2]. In this system, Requests library are used to interact with the website and obtain the source code of the web page, and Beautiful Soup 4 is used to process the source code of the website to obtain the required content [3-4].

According to the name of the school, the paper of the corresponding school is crawled. First, the school name is sent as a search word through the Request library, and then the get request is sent according to the returned results. The data of the first page can be obtained. On the first page, the total number of records and pages can be obtained, and the data of each page can be crawled through. Each page is parsed by Beautiful Soup 4, and the names of papers, downloads, authors, units, journals, downloads and citations obtained from the list are stored in MySQL database. It also needs to visit the detailed description page of the corresponding paper according to the link, obtain the corresponding abstract, keywords, funds and other information of the paper, and update it to the database. When all papers have been crawled, the paper is searched through the name of the paper and whether SCI, EI, Chinese core, CSSCI and CSCD. If the paper appears in a list of certain types, indicating that the corresponding papers are searched by SCI, EI, Chinese core, CSSCI or CSCD, the database is updated, and the corresponding logo is set to 1. The process flow is shown in Figure 2.

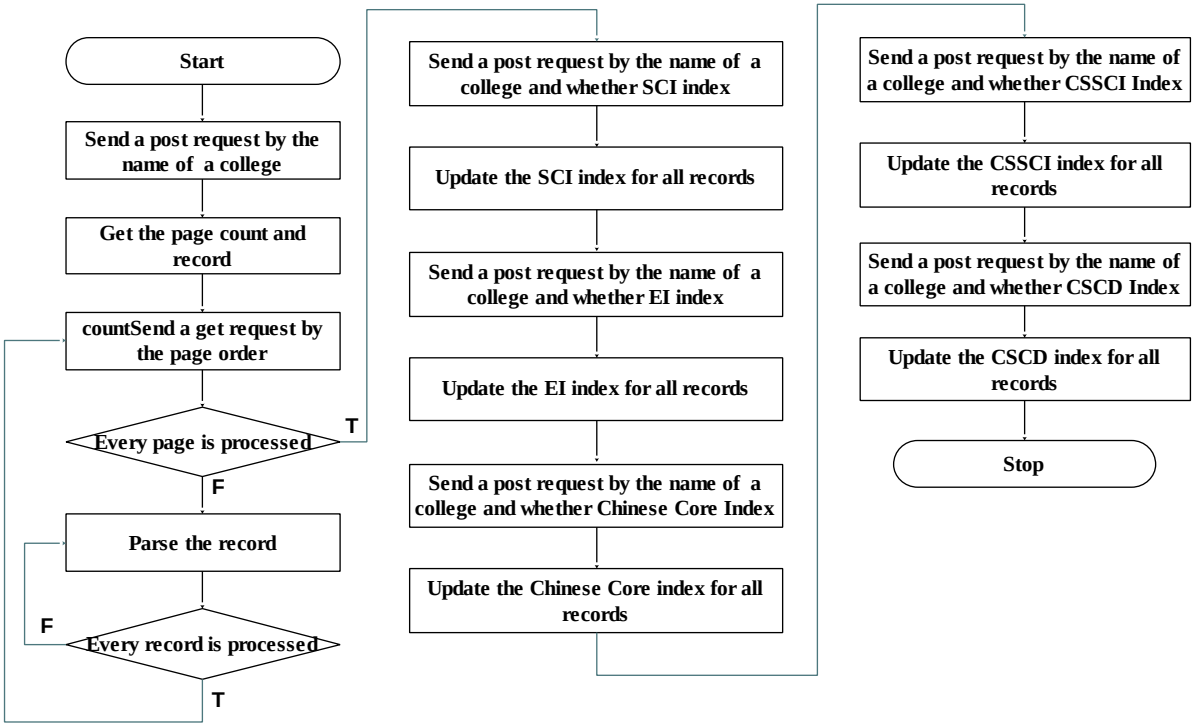


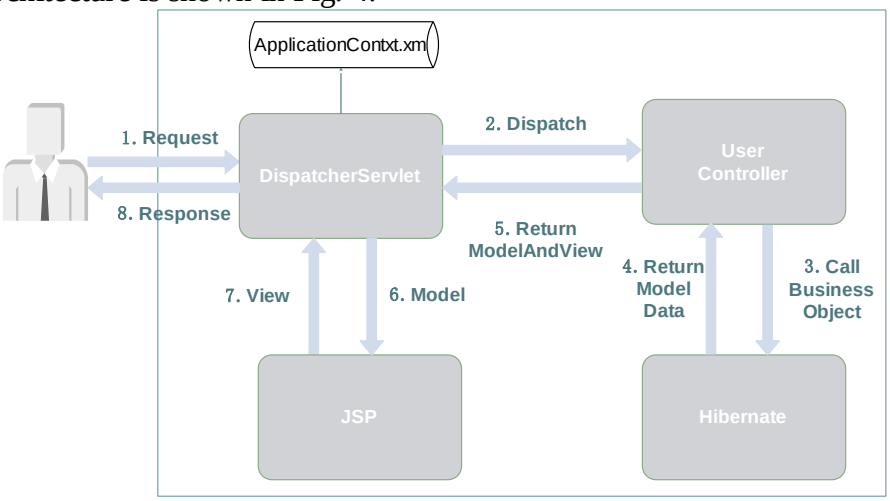
Figure 2. The crawler flowchart according the name of a college

3. Visualization Implementation

Data visualization can help us understand the distribution, trend, relationship, comparison and composition of data. Help decision makers discover hidden features in large amounts of data when they review it. The project visualizes the stored data in the form of Web.

The data visualization module is implemented by the MVC (Model View Controller) model and based on the SSH (Spring, Spring MVC, Hibernate) framework [5]. Spring MVC carries on the process control, Spring carries on the business circulation, Hibernate carries on the encapsulation of the database operation [6-8].

The platform architecture is shown in Fig. 4.



In SSH framework, Spring MVC is a standard MVC Web tier framework with a front-end controller Dispatcher Servlet for overall scheduling. The Dispatcher Servlet receives requests and forwards them to Spring's MVC controller, the business controller. A Controller is responsible for handling this request, which is typically handled by calling the Hibernate for business logic. When the controller processes the request, it encapsulates the business processing results into a model. In order to make the model of the processing results better displayed on the page, the controller also specifies the view corresponding to the JSP page. The Dispatcher Servlet eventually returns the page to the user

4. Conclusion

This paper proposes a platform for the analysis of teacher scientific research ability, which is based on big data technology. It realizes a statistical analysis method of teachers' scientific research ability, which is different from the traditional methods. JSON page parsing technology based on DOM tree and malicious URL detection technology are synthetically used to realize the automatic crawling function of KWN document data; the document data analysis method based on big data technology is adopted to design effective indicators, through data mining and processing of document data, important data patterns are found; and the visual expression of Web form is adopted to enhance users knowledge of data.

Acknowledgements

This work was financially supported by practical innovation project for college students of Jiangsu Maritime Institute, the higher vocational scientific research subject of computer national computer basic education institute (2018-AFCEC-265), the funding of Jiangsu QingLan outstanding young teacher project and the funding of professional leader high level study project for Jiangsu higher vocational institute teachers .

References

[1] Velde C. Employers' perceptions of graduate competencies and future trends in higher vocational education in China[J]. Journal of Vocational Education & Training, 2009, 61(1):35-51.
[2] Ren Y, Ren Y. A Framework of Petroleum Information Retrieval System Based on Web Scraping with Python[C]//2018 15th International Conference on Service Systems and Service Management (ICSSSM). IEEE, 2018: 1-6.
[3] Xu Z, Liu P, Zhang X, et al. Python predictive analysis for bug detection[C]//Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering. ACM, 2016: 121-132.
[4] Zheng C, He G, Peng Z. A Study of Web Information Extraction Technology Based on Beautiful Soup[J]. JCP, 2015, 10(6): 381-387.
[5] Lv T, Zhang J, Chen Y. Research of ERP Platform based on Cloud Computing[C]//IOP Conference Series: Materials Science and Engineering. IOP Publishing, 2018, 394(4): 042004.
[6] Srai A, Guerouate F, Berbiche N, et al. Applying MDA approach for spring MVC framework et al[J]. International Journal of Applied Engineering Research, 2017, 12(14): 4372-4381.
[7] Fisher P, Murphy B D. Architecting Your Application with Spring, Hibernate, and Patterns[M]//Spring Persistence with Hibernate. Apress, Berkeley, CA, 2016: 1-16.
[8] Cosmina I, Harrop R, Schaefer C, et al. Using Hibernate in Spring[M]//Pro Spring 5. Apress, Berkeley, CA, 2017: 355-392.