Job Analysis System Based on Spark Platform

Taizhi Lv[a,*], Juan Zhang[b], and Yingying Chen[c]

College of Information Technology, Jiangsu Maritime Institute, Nanjing 211170, China

[a]lvtaizhi@163.com, [b]47679546@qq.com, [c]760846936@qq.com

*Corresponding author

Keywords: Job Analysis System, Spark, Distributed Crawl, Microservice, Vocational Education

Abstract: Big data technology is a new technological revolution after cloud computing, and widely used in medical, financial, transportation, education and other industries. Due to the lack of understanding of industrial development, the teaching content of higher vocational colleges is often unable to meet the needs of enterprises. In order to overcome this problem, a job analysis system based on big data technology is implemented. It collects the job information in time, and obtains skills which are suitable for employment by the spark platform. The talent training plan is updated in time, and the teaching quality is improved.

1. Introduction

With the expansion of higher education, vocational education has achieved a leap in scale, but the rapid development of higher vocational education has not received sufficient social recognition, and the overall level of higher vocational students is still poor. The development of information technology is “one thousand miles away”, which is one of the fields with the fastest rate of knowledge update, and professional talent training programs often cannot keep up with the development of the times. Through big data technology, timely grasp the professional corresponding post information, even if the professional settings are adjusted according to the data, adjust the talent training program, and update the teaching content, so that students can accept new knowledge faster and keep pace with the times. The major task of the IT major in higher vocational education is to train more IT-skilled talents who can adapt to society. Therefore, the teaching, teachers, scientific research and management of the school must focus on the task of talent training. However, the traditional education model is mainly based on transcendental education and does not pay attention to the cultivation of students' personality. At the same time, there is a large gap between domestic universities and the number of students and teachers is not very coordinated. Common classroom teaching modes make it difficult for teachers to fully and deeply understand the actual learning status of students. In the era of big data, the application of big data technology in higher education management can greatly improve the quality of higher education. School management decision-making can also use big data technology to play an important role in motivating and supporting decision-making. Big data can be found in data-to-data association rules, rather than proof rules. The main value is the inherent performance data found by each manufacturer to develop big data application ideas and provide some guidance for decision-making.

The number of fresh graduates has increased year by year, and employment pressure cannot be underestimated. In recent years, the employment problem of college students has still been one of the focuses of the country and society. The school has also introduced a large number of companies to recruit, but it has not solved the problem of difficult employment. Students are not unwilling to find employment, but are unable to find suitable jobs. .Some recruitment websites, such as 51 Job and Zhaopin, use mechanization to find jobs that require employment, and cannot intelligently recommend jobs based on the skills held by students. As a result, students have read countless recruitment information and failed to find a satisfactory one work. Therefore, if we can analyze the student's performance at school to measure the skills learned by the students, and then dig out the skills required for the job, the two can be a good recommendation for students.

The distributed crawler based on Redis represents the text content from the recruitment website into the format required for subsequent processing, builds a job big data analysis platform based on the Spark platform, and builds a visualization platform based on Microservices and Echarts, and extracts job related information for processing. The school offers courses for reference, and compares and analyzes the undergraduate's school performance and graduate employment information to recommend appropriate positions for students, which can well solve the difficult situation of school graduates' employment.

2. Distributed Crawl based on Redis

Aiming at the characteristics of the Zhilian recruitment website, a distributed crawling framework based on Python-Request-Beautiful Soup- Redis-Zookeeper-HBase was proposed to crawl and store recruitment post information.

2.1 Framewrok

Based on the Request and Beautiful Soup components, crawling of job data can be completed, but distributed crawling is not supported, which cannot meet the task of crawling massive thesis data. In order to meet the data crawling needs, a distributed crawler framework based on Python-Request-Beautiful Soup- Redis-Zookeeper was built [1-2]. The framework mainly includes three parts: a crawler system, a URL scheduling system, and a monitoring alarm system. The overall architecture is shown in the following figure.

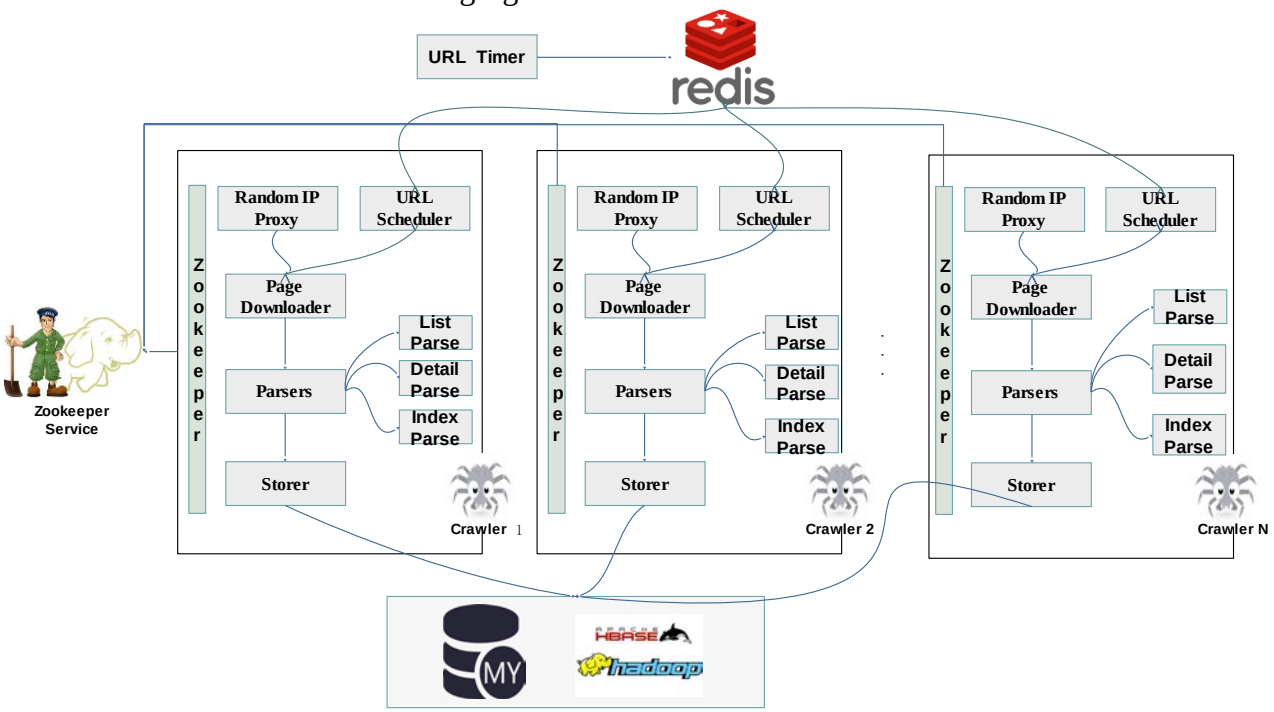

Fig .1 Distributed Crawl Framework

The crawler module is designed to download and parse web pages, which are distributed and run on different nodes. The URL scheduling module includes URL scheduler and redis database. The URL scheduler takes out the data in the URL queue according to certain policies for consumption, and the URL queue is stored in the redis database. The monitoring and alarm module is completed by zookeeper, which realizes the monitoring of crawler monitoring, including the client and the server. The client is installed on each crawler node, and the server is installed on the designated server to complete the monitoring of the crawler program.

2.2 Page Download

Similar to the common website crawling, the crawling of position data first searches the list page according to the position keyword, and then requests the details page according to the hyperlink in

the list. The system searches according to the position name as the keyword. The query entry is https://sou.zhaopin.com/, and the request parameters are determined by the developer tool of the browser.

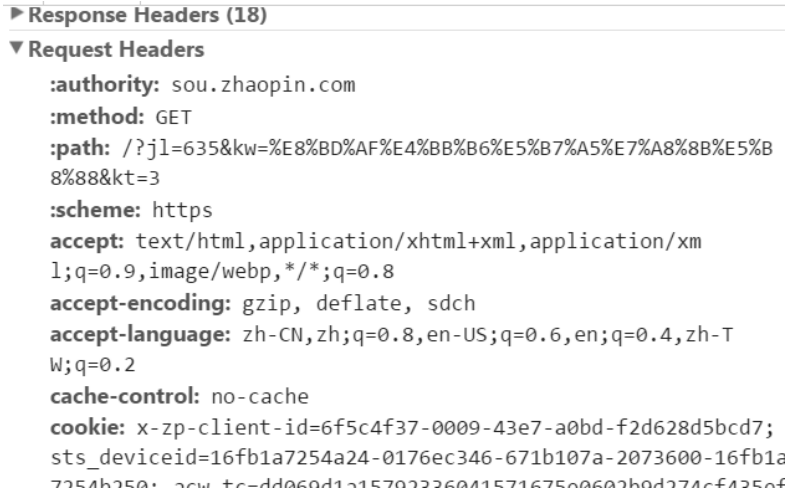

Fig 2. Request parameter diagram

2.3 Page parse

Through Beautifulsop and lxml parsing library, the position list page is resolved. In the list page, first find the table item where the list is located, then traverse all the div items under the div item with id = "listcontent", and obtain the position name, education requirements, company name, company treatment, company information and the link address of the detailed page through the div item. Analyze the relevant information such as skill requirements, job responsibilities, qualifications and work place in the detailed page.

2.4 Job analysis based on spark

Spark is one of the top projects of the Apache foundation. Spark provides high-level APIs in four languages, including Scala. Resilient distributed dataset (RDD) is the core abstract data structure in spark application [3-4]. All operations on data in spark are around RDD, such as creating RDD, converting RDD, RDD evaluation, etc. RDD based on spark realizes the cleaning and analysis of post data. In the data cleaning stage, the fields with null value, incomplete content and inconsistent content are supplemented and corrected, which is divided into five steps. Post skill point processing is to clean the data files collected in HDFS, and save the posts and skill points to the database. However, in this step, the cleaned file needs to be saved in lines as text and uploaded to HDFS for conversion into a sequence file in the form of key-value. The following steps are respectively: sequence file is converted into vector file, vector file clustering, clustering results are saved locally, analysis and extraction of local file skill points are saved to the database.

3. Visualization based on Microservices and Echarts

Microservice architecture style is a way of using a set of small services to develop a single application, maintaining a minimum of centralized management [5]. The platform adopts the microservice architecture, and divides the application into system management, basic data management, data display, web crawler and other micro services. Each Microservice completes the corresponding functions, and the lightweight rest interface is used between services to achieve communication. The overall architecture design of the system is shown in figure 3.

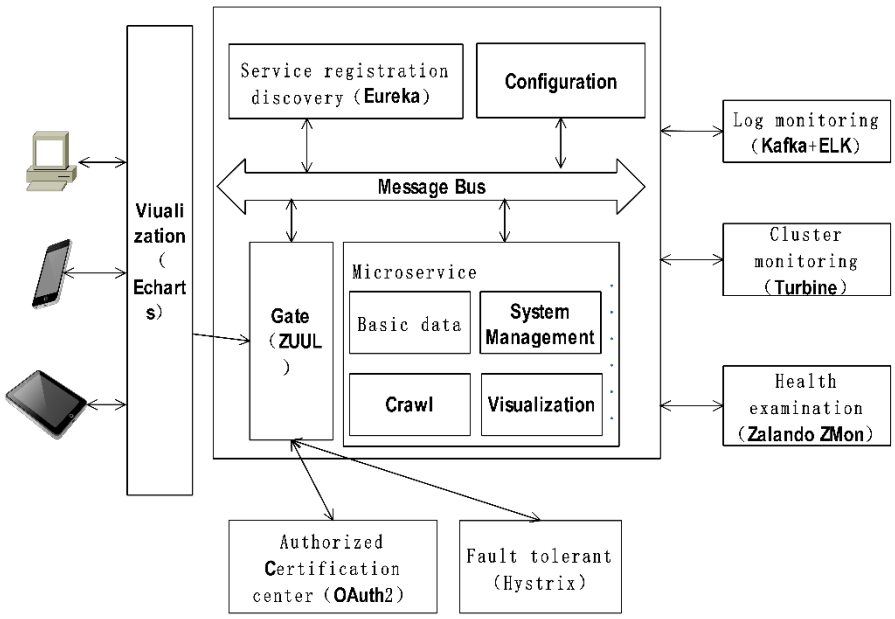

Fig 3. Microservice architecture diagram

The platform development adopts the Microservice architecture based on the spring cloud, which is separated from the front end and the back end. The front end uses the Echarts framework, and the back end uses the spring cloud + spring boot + Mybatis + MySQL and other technologies to realize the microservice discovery registration, configuration center, message bus, load balancing, circuit breaker, data monitoring, etc.

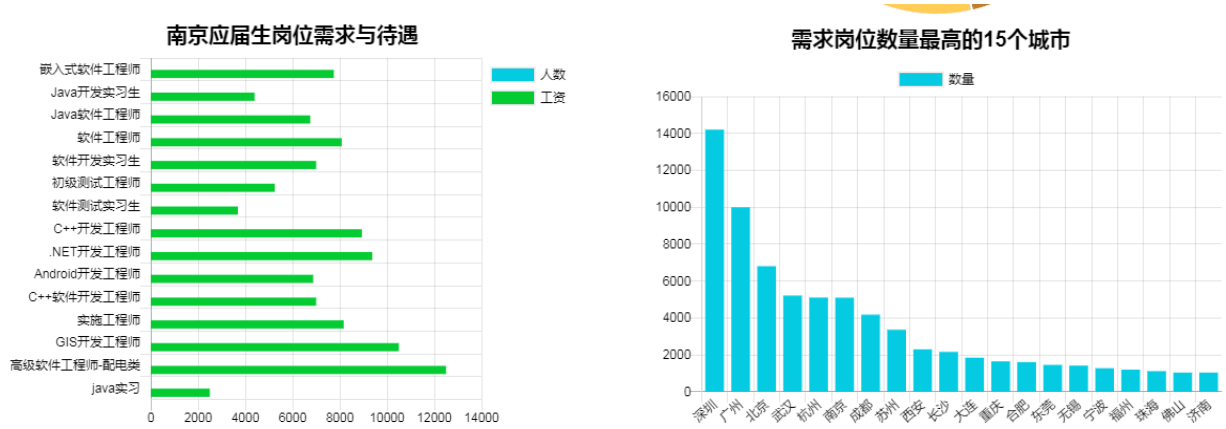

Fig 4. Visualization result

4. Application of the platform

The job analysis platform based on the recruitment website has been applied to the formulation of talent training program for 2020 level big data technology and application new majors, and the revision of talent training program for 2020 cloud computing technology and application, software technology and software engineering majors.

(1) Through the automatic crawling and analysis of big data related posts, it is determined that python programming foundation, data collection, data crawler, data visualization, Hadoop and spark platform construction, R language, etc. are the core skills of the post, and the core courses of related majors are determined according to these core skills.

(2) Through the automatic crawling and analysis of cloud computing technology related posts, it is found that container technology and automatic operation and maintenance technology are more and more enterprises as basic post skills, so based on this change, container technology and operation and maintenance courses are added to the talent training program.

(3) By automatically crawling and analyzing the relevant positions of software engineer, it is found that the front-end development position receives more and more attention, and the post salary is basically the same as the back-end engineer. The salary of engineers in the whole stack can reach more than 10000. Based on this, the front-end development related courses are added, including Vue and corresponding development

5. Summary

Before there is a job analysis platform based on recruitment website, the professional skills needed to master a job are realized by investigating some enterprises. In this way, the results are not comprehensive, and sometimes the results are not objective or even biased because the information given by the enterprise is more arbitrary.

In order to train students to meet the needs of the society, it is necessary to comprehensively and objectively grasp the current enterprise's occupational skills requirements for posts. First, through the web crawler technology, extract and save the information of relevant posts published by various enterprises on 51job.com, then use the word segmentation technology to segment the extracted post information, and finally obtain the core skills required by the post As an important basis for the formulation or revision of personnel training.

Acknowledgments

This work was financially supported by the funding of the Project of Philosophy and Social Science Research in Colleges and Universities in Jiangsu Province (2019SJA0650), the 13th Five-Year plan project for education science in Jiangsu province, big data collaborative innovation center of Jiangsu Maritime Institute.

References

[1] Feng Ye, Zongfei Jing, Qian Huang,etc. The Research of a Lightweight Distributed Crawling System[C]// 2018 IEEE 16th International Conference on Software Engineering Research, Management and Applications (SERA). IEEE Computer Society, 2018.

[2] Xiao X, Zhang W, Zhang H, et al. Scale-Adaptable Recrawl Strategies for DHT-based Distributed Web Crawling System [J]. 2010, 6289:91-105.

[3] Giannis Agathangelos, Georgia Troullinou, Haridimos Kondylakis, etc. RDF Query Answering Using Apache Spark: Review and Assessment[C]// 2018 IEEE 34th International Conference on Data Engineering Workshops (ICDEW). IEEE, 2018.

[4] Michael Armbrust, Tathagata Das, Joseph Torres, etc. Structured Streaming: A Declarative API for Real-Time Applications in Apache Spark[C]// the 2018 International Conference. 2018.

[5] Marco Gribaudo, Mauro Iacono, Daniele Manini. Performance Evaluation of Replication Policies in Microservice Based Architectures [J]. Electronic Notes in Theoretical Computer Science, 2018, 337:45-65.