Journal of ICT in Education (JICTIE)
ISSN 2289-7844 / Vol. 8 / Issue 2 / 2021 / 118-125

Preliminary Analysis of Data Science Talents in China's Online Recruitment Market Using Web Scrapping Tool

Yang Shiwei, Ashardi Abas*

Faculty of Art, Computing and Creative Industry, Sultan Idris Education University; 21778603@qq.com
Faculty of Art, Computing and Creative Industry, Sultan Idris Education University; ashardi@fskik.upsi.edu.my

** Correspondence author*

To cite this article (APA): Shiwei, Y. & Abas, A. (2021). Preliminary Analysis of Data Science Talents in China's Online Recruitment Market Using Web Scrapping Tool. *Journal of ICT in Education, 8*(2), 118-125
https://doi.org/10.37134/jictie.vol8.2.11.2021

To link to this article: https://doi.org/10.37134/jictie.vol8.2.11.2021

Abstract

China implements a big data strategy, accelerates the construction of a digital China, and data science has entered a new and dynamic era. There is an increasing demand for data science talents from all walks of life. The main purpose is to obtain the market demand for data science talents, understand the demand for the data science talent market, and analyze the demand for data science talents. The research method uses Web Scrapping Tool to grab data science talent demand information from major domestic recruitment websites and then obtain data science talent demand through information analysis. Major findings preliminary assessed the data talent market in China from the aspects of geographic demand for data talents, company size, industry demand, and salary. Future research can use crawling information to join machine learning algorithms, in-depth study of the internal connection of China's data science talent needs, and provide suitable training programs for universities.

Keywords: data science, recruitment websites, Web Scrapping, grab data, demand analysis

INTRODUCTION

With the integration of Internet technology and modern society's production and lifestyles, massive amounts of data worldwide have gathered, making its growth show a blowout trend. People feel the existence of data all the time from all aspects of daily life. The use and development of data have had a major impact on economic development, social governance, state management, and people's lives. At the same time, Data Science, a term that makes people feel familiar and unfamiliar, once again appeared in people's vision. Karpatne and Atluri (2017) support that the start of twenty-first century may well be remembered in history as the golden age of data science.

Received: 21 July 2021; Accepted: 17 August 2021; Published: 27 August 2021

Data science is a broad sense is not only the study of scientific methods for data, such as mathematical statistics, machine learning, data mining, etc. but also the application of data theories and methods to the study of other science and technology, including actuarial science, business Intelligence, industrial statistics, etc.

In the action of implementing the big data development strategy, local governments at all levels actively responded to the call of the country, actively participated in the development of the data industry, continuously strengthened the application of big data technology, consolidated the training of data science talents, and steadily promoted the construction of a strong data. Guizhou Province, Zhejiang Province, Guangdong Province, Shanghai, Chongqing City, Beijing, and other provinces have successively formulated big data development policies.

Data science has only begun to develop in China in recent years. Although it has developed rapidly, there is very little systematic analysis of this industry. Therefore, a market analysis of the development needs of the data science industry is fundamental.

LITERATURE REVIEWS

Data mining is a process that analyzes a large amount of data to find new and hidden information that improves business efficiency. Various industries have been adopting data mining to their mission-critical business processes to gain competitive advantages and help business grows.

Gao (2015) is based on historical data on college graduates' employment and employment guidance. Try to find out the useful information hidden in the employment history data through data mining of the employment history data of college graduates.

De Gagne et al. (2021) proposed that learning data mining analysis has become a new process and tool to improve learning performance. As a tool, it can be used to measure, collect, analyze and report data in order to understand and optimize learning, students and their background.

Mabić et al. (2017) introduced how the application of data mining techniques can promote curriculum development in higher education. The development of a good curriculum is vital to higher education institutions because a good curriculum attracts new students, improves the quality of students, and recruits institutions with high power and quality to increase visibility. The order of the courses has a great influence on students' learning. Successfully realize the planned learning outcomes.

RESEARCH METHODOLOGY

Among online recruitment companies, Zhaopin's market share accounted for 30.7%, 51job has 31.8%, and the remaining other websites accounted for 37.5% of the market. Although there has been an influx of new recruitment websites in recent years, the growth of new recruitment websites will still take time, and the market will continue to have a dual giant distribution pattern in the future.

The sample selection for this study will be the recruitment data of data science talents from the twin giants 51job and Zhaolian recruitment website.

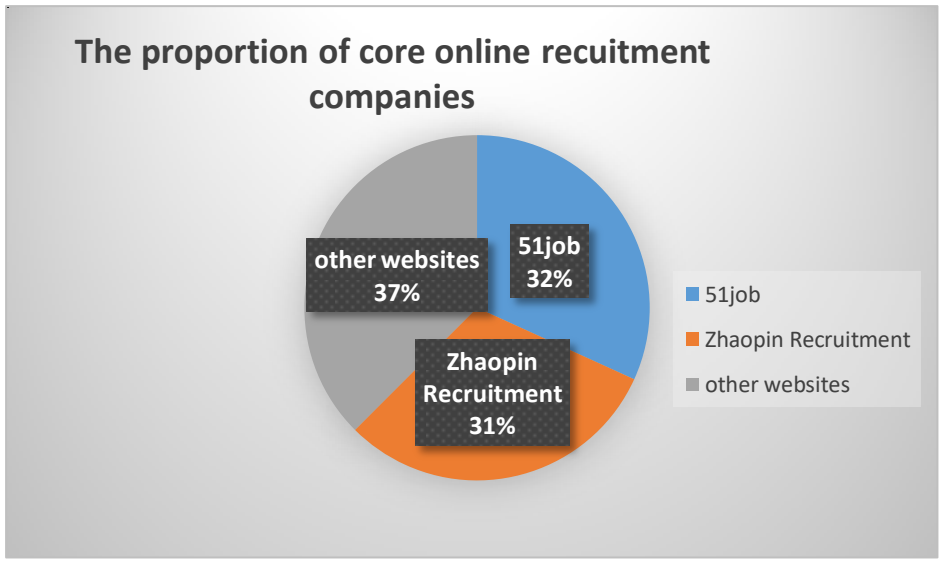

Figure 1: The proportion of core companies

This research will use data mining technology to crawl the data of two major recruitment websites in China, store the acquired data in the database, and then conduct data analysis on the relevant information of data science talents. Through the analysis, the characteristics of the demand for data science talents are obtained.

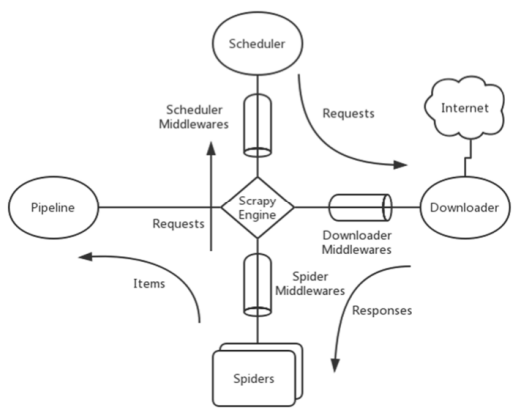

Figure 2: Figure Scrapy Structure

The Scrapy crawler framework is a powerful automatic crawler framework developed and packaged in the python language. This is also one of the powerful community library functions of the python language. Targeted data crawling is carried out through the functions provided by the Scrapy framework. In the Scrapy project, fill in the customized crawler rules, and then run scrapy, you can quickly obtain the required web page data. The powerful functions of Scrapy benefit from its structure. It consists of 8 parts in total, which are 8 components as shown in the figure2 which is Scrappy 0.22 documentation:

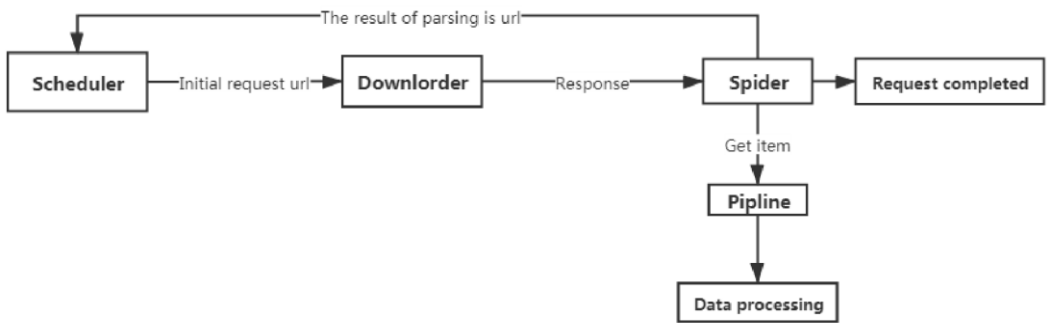

Figure 3: The process of Scrapy crawl

The data flow of the scrapy crawler is controlled by the Scrapy engine as a whole. Configure start_urls to specify the initial target for crawling. After the engine receives the URL target address initially crawled, it dispatches network requests and responses through the dispatcher component. The download module receives the request body and initiates a data request to the specified address on the Internet. And send the final response to the scrappy engine by downloading the middleware. After the engine receives the response data, it passes it to the spider module through the spider middleware, executes the callback function defined in the spider, and parses the data body. The information entity obtained in the data volume analysis process will continue to be pushed to the pipeline module for further processing. After the pipeline component receives the entity passed by the spider, it can get useful information and store it in the database. The continuous loop crawler process is processed, and the data request process is executed when the response body is completely parsed. After the subsequent pipeline data processing is completed, the entire crawler project process execution is completed.

Traditional relational databases have great difficulty in storing data formats of different complexity on different websites. Therefore, in data crawlers, more will choose NoSQL non-relational database to store the desired data. The non-relational database is different from the relational database. The NoSQL database can store all kinds of data, and at the same time, has high scalability and high availability. The data model is more flexible, the data read and write is easier, the data has no relationship, and it is convenient for data expansion. Among the currently popular NoSQL databases,

select the MongoDB database to store the crawled data for subsequent data mining operations. The content stored in the MongoDB database through Pymongo after the recruitment website data is crawled through the web crawler. These contents mainly include company name, job title, salary, work location, academic requirements, job skill requirements, etc.

FINDINGS AND DISCUSSION

After pre-processing the data science talent recruitment data, this research briefly analyzes the current demand for data science talents from three different perspectives of the company's industry, company size, city location, and average salary.

The data obtained from the recruitment website in this study covers data science talent demand information in 34 cities in 23 provinces, 1 autonomous region, and 4 municipalities. Figure3 plots a pie chart of the distribution of the number of data science talents in major cities. Proportionally, Beijing, Shanghai, Guangzhou, and Shenzhen account for 79% of the national demand for data science talents and are the four regions with the highest demand.

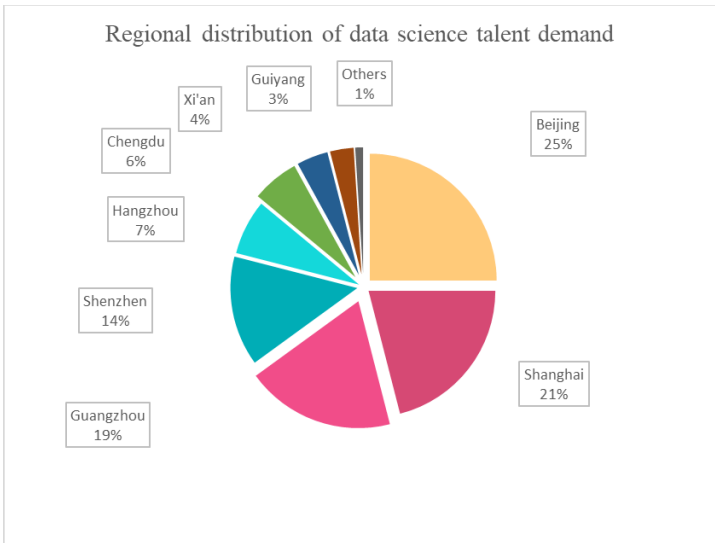

Figure 4: Regional distribution of data science talent demand

It is not difficult to find that the distribution of data science talents also conforms to the Pareto law through observation. Beijing is the capital of China and the country's political, cultural, and technological innovation center. The demand for data science talents also ranks first in the country, accounting for 25% of the overall demand. Shanghai, China's economic and financial center, known as the "first-tier city in the world," followed closely behind with 21% of demand as China's largest economic provinces, Guangdong Province, Guangzhou (19%), and Shenzhen (14%) are among the top cities in demand for data science talents. As the representative city of the Jiangsu, Zhejiang, and

Received: 21 July 2021; Accepted: 17 August 2021; Published: 27 August 2021

Shanghai free shipping areas, Hangzhou has become the fifth place in the talent demand list with the desire of many start-up companies for data science talents. As the "big data city" Guiyang, data science talents are also in great demand.

The salary of data science talents is generally relatively high, mainly concentrated in the range of 10k-20K. However, because human factors cause many outliers in the original salary, the salary analysis will be based on the average salary after the common logarithm processing. The average salary is taken from the average of the highest and lowest salary offered by the position. The following analysis of the salary and treatment of data science talents is mainly carried out from the logarithmic average salary distribution and its relationship with company category and company size.

As shown in Figure 4 below, the larger the company's size, the higher the corresponding logarithmic average salary, especially for foreign capital and joint ventures. This is also a signal to data science talents. In choosing a career, you must learn more about the scale and culture of the company and use this as a prerequisite to screen out the employment opportunities provided by high-quality companies.

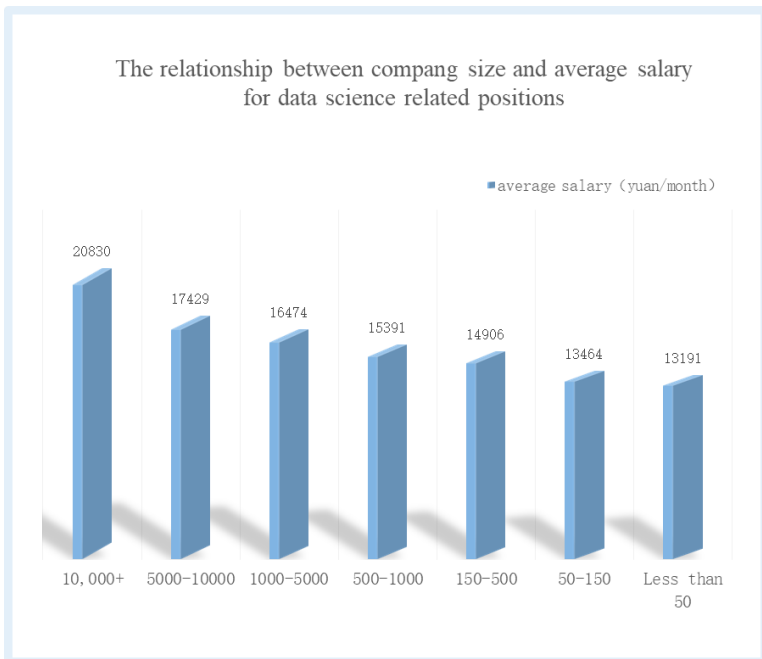

Figure 5: The relationship between company size and average salary for data science related positions

From the relationship between the company industry and average salary in the figure5 below, it can be seen that the Internet and e-commerce industries have the highest treatment, followed by accounting and auditing industries. In addition, the treatment of instrumentation and industrial automation is also good. The salary of trust, guarantee, marketing, and the pawn is relatively low, but it also exceeds the level of 1K per month. Generally speaking, the salary level of any industry is considered to be a relatively high salary level in China.

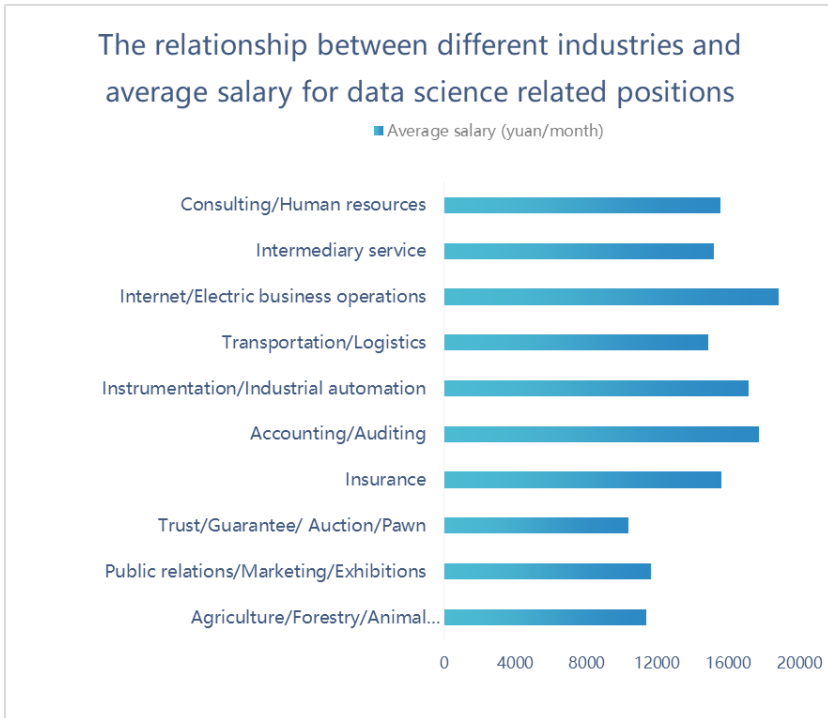

Figure 6: The relationship between different industries and average salary for data science-related positions

SUMMARY

The main content of the research is to conduct descriptive statistical analysis on the structured data of data science talent demand from four aspects: city location distribution, company type, company size, and average salary. In the analysis process, it can be found that the demand for data science talents is mainly concentrated in the emerging hot industries such as the Internet and e-commerce in Beijing, Shanghai, Guangzhou, and coastal cities. The average salary level of data science talents is relatively high, and the salary of entrepreneurial companies is not high. And small and medium-sized private companies have more job opportunities. Future research can also use machine learning

clustering algorithms to conduct in-depth research on the skill requirements of data talents through the crawled data talent information and give a reasonable training plan for schools to train data science talents.

REFERENCES

De Gagne, J. C., Cho, E., Yamane, S. S., Jin, H., Nam, J. D., & Jung, D. (2021). Analysis of cyber incivility in posts by health professions students: descriptive twitter data mining study. *JMIR Medical Education*, *7*(2), e28805. https://doi.org/10.2196/28805

Gao, L. (2015). Analysis of employment data mining for university student based on WEKA platform. *Journal of Applied Science and Engineering Innovation*, *2*(4), 130-133

.

Karpatne, A., Atluri, G., Faghmous, J. H., Steinbach, M., Banerjee, A., Ganguly, A., ... & Kumar, V. (2017). Theory-guided data science: A new paradigm for scientific discovery from data. *IEEE Transactions on Knowledge and Data Engineering*, *29*(10), 2318-2331. https://doi.org/10.1109/TKDE.2017.2720168

Mabić, M., Dedić, F., Bijedić, N., & Gašpar, D. (2017). Data mining and curriculum development in higher education. In *International Conference on Information Technology and Development of Education*–ITRO. Pp. 1-6.