Analysis of Regional Talent Demand based on Recruitment Website Text Mining

Longmao Hu and Bingwu Fang*
Anhui Finance & Trade Vocational College, hefei, Anhui, China
hulongmao@163.com, bingwufang@163.com
*corresponding author

Abstract—Mining the characteristics of talent demand through the recruitment information released by the recruitment website can provide a useful reference for major setting and talent cultivation in higher vocational colleges. The data with college degree under the category of “computer/Internet/communication/electronics” from 51job recruitment website were crawled. After data preprocessing, 32,971 computer recruitment data of higher vocational colleges in the Yangtze River Delta region were obtained. The data were analyzed from three aspects: position demand, salary level and core skills by using data statistics and B-IDF weight method. The study found that the demand of each position varies greatly, the median salary level has a certain difference, and the three-quarters of the difference is large. Each position emphasizes hands-on ability, and the artificial intelligence position requires a certain theoretical basis.

Keywords: recruitment website; yangtze river delta; higher vocational college; computer; talent demand

I. INTRODUCTION

Higher vocational colleges should cultivate high-quality technical and skilled personnel who serve regional development [1]. As one of the regions with the most active economic development, the Yangtze River Delta region plays an important supporting and leading role in information technology. Understanding the demand scale, salary and core skill requirements of computer talents in higher vocational colleges in the Yangtze River Delta region plays an important guiding role in the setting of computer related majors and personnel training in higher vocational colleges in the region.

Researchers mainly use the following two methods to obtain talent demand. 1) Interview or questionnaire survey. Through interviews with some laboratory directors and researchers, Kim Y et al. [2] found that eScience professionals must possess six major duties with respect to “data”(including collecting primary data, collecting secondary data, storing data, managing data, analyzing data, and presenting data), as well as skills in using computers and software. Huang [3] visited some big data enterprises in Beijing Zhongguancun and Fujian southeast big data Industrial Park to understand the job requirements of enterprises for big data of Higher Vocational Education. Through in-depth discussion with experts of Boya big data research institute, he divided the core competencies of big data and defined the training requirements of big data operation and maintenance engineers, data analysis engineers and big data development engineers. Chi et al. [4] surveyed 67 representative enterprises in the cloud computing and big data industries across the country, and found that higher vocational personnel have become the main force of cloud computing, and their jobs are mainly distributed in sales, technology and service positions, while only 12.22% of big data talents are in higher vocational education. Through the matching analysis of talent demand and talent training in vocational colleges, it is found that the professional setting of cloud computing and big data basically meets the development needs of the industry. The professional counterpart rate of cloud computing technology in the past three years All of them are more than 82%, while the counterpart rate of big data technology is basically around 20%. The main reason is that the focus of college training goals is not enough, resulting in a lack of students' mastery of core technologies. Such methods take a long time to obtain data, and the cost is also high. There are many subjective factors in the data. The quality of the data depends on the design of interview questions or questionnaires, and the amount of data is small. It is difficult to truly and comprehensively reveal the talent needs of enterprises.

2) Mining online recruitment information. Sodhi et al. [5] collected some data of operations research majors on recruitment websites, and studied the differences in the demand for operations research skills in different industries by constructing a dictionary of job and skill keywords, and using statistical and content analysis methods. Wei [6] captured 5,691 online recruitment data of accounting positions in Hunan Recruitment Network for 6 years, and analyzed the actual needs of the accounting talent market in Hunan Province from the aspects of talent education requirements, knowledge ability requirements, enterprise scale, registered capital, and industry distribution. This paper analyzes the knowledge and qualities that accounting talents should have with the help of tag cloud and semantic network, and puts forward the countermeasures for the cultivation of entrepreneurial talents of accounting in higher vocational colleges. Yang et al. [7] crawled 11632 valid data from the Internet financial industry from the Zhaoxin website, extracted the post and skill entities from the recruitment text, and constructed the post skill network model; Based on the modularity, the network is divided into technology R & D, business operation, financial sales and project management. The correlation between skills is calculated by using the aggregation coefficient. It is found that a small number of...
posts and skills play a major role in supporting the industry, and different types of post associations have different emphasis on the requirements of non cognitive skills. This kind of research pays attention to the role of the massive data of online recruitment in mining talent needs, but the mining is not deep enough, and only stays in the statistical description of online data, or the processing of job texts is relatively simple, and fails to truly reflect the core skills of talents.

Starting with the online recruitment information, through statistical analysis and natural language processing methods, this paper analyzes the post demand and salary level of computer majors in Higher Vocational Colleges in the Yangtze River Delta, and excavates the core skills from the post text, in order to provide the basis for the setting of computer majors and talent training in higher vocational colleges in the region.

II. DATA ACQUISITION AND RESEARCH METHODS

A. Data Sources and Collection

The data in this paper comes from the 51job recruitment website.

The description of data collection is as follows: (1) The Scrapy crawler framework is used to write the crawler program, the rendered HTML page content is obtained through the Splash engine, and the text information to be obtained in the page is parsed by the xpath and css parser. On February 27, 2021, the crawler crawled all the data under the category of "computer / Internet / communication / Electronics", with a college degree on 51job website. A total of 300103 data were crawled and saved as CSV files. (2) In 51job.com, each position category includes several second-level positions, and each second-level position includes several third-level positions. For example, the category of "computer / Internet / communication / Electronics" includes 15 second-level positions such as back-end development, mobile development and artificial intelligence, while the back-end development of second-level positions includes 18 third-level positions such as software engineer and java development engineer. Each recruitment data can belong to multiple three-level posts (such as web front-end development and software engineer). In this paper, the crawler program crawls all the data according to the second-level positions. Therefore, if a recruitment data corresponds to multiple three-level positions and belongs to the same second-level position, this data will only appear once; otherwise, this piece of recruitment data will appear repeatedly according to the different second-level positions which three-level positions belong to. The multi-label classification of recruitment data is reasonable, so the crawler program in this paper does not perform deduplication processing. If a certain recruitment data belongs to two different second-level positions, it will appear twice in the csv file. (3) Crawling fields include 12 fields, including second-level position category name, position name, salary, company, position profile, benefits, position details, three-level position category name, keyword, company nature, number of employees and industry.

B. Research Methods

The research method in this paper is mainly divided into the following steps. First, program a web crawler to collect recruitment website data; second, preprocess the acquired data, including data cleaning, transformation and other steps; third, use statistical analysis methods to visualize position requirements and salary levels; Fourth, the position details (also called position text) are tokenized, and the core skills are mined using the Boolean-inverse document frequency method.

1) Data preprocessing

Due to the concurrency and complexity of the crawling process, there are duplicate information or even wrong information in the recruitment text, which is cleaned in two steps. First, read the csv file, and use pandas' deduplication function to remove duplicate information; second, program to replace non-utf8 characters in the text with spaces (expressed as " ")

In order to meet the research purpose of this paper, the following data transformations are carried out. First, adjust the "big data development engineer", "Hadoop engineer" and "crawler development engineer" in the third-level positions to the second-level position "data" category. Second, when the recruitment data belongs to different second-level positions, there will be multiple rows in the table, and the second-level positions to which they belong are spliced into one field with spaces, so that the recruitment data is merged into one row. Third, use regular expressions to extract three fields from the position profile field: City, whether experience is needed and the number of recruits. For example, the extraction result of "Shanghai Pudong New Area | 3-4 years of experience | College | recruit 1 person | 02-27 release" is "Shanghai", "experience required" and "1". In order to facilitate the subsequent quantitative processing of post requirements, several people extracted from the number of recruiters are replaced with three. From the 51job response page, manually extract the list of cities in the Yangtze River Delta. If the extracted city is not in the list, change the extraction result to "non-Yangtze River Delta city"; otherwise, keep the city name unchanged. Fourth, in the salary field, the salary given by the range is converted into the average of the upper and lower limits of the range, and the payable time such as years and days is converted into monthly salary. Fifth, detect the English terms in the post text, and convert the case into a standard format (such as python, PYTHON is converted to Python). After the above 5-step transformation, designate the city as not a "non-Yangtze River Delta city", and designate the secondary positions as "back-end development", "mobile development", "front-end development", "artificial intelligence", "game", "Testing", "Operation and Maintenance/Technical Support", "Data", "Product" and "Technical Management", screened out the recruitment data of higher vocational computer in the Yangtze River Delta region, a total of 32971 pieces of data.

2) Extraction of core skills based on B-IDF

After the text is divided into words, words that appear frequently in certain types of documents and low in other documents are usually the core vocabulary of such documents, which can be calculated by TF-IDF (document frequency
inverse document frequency) [8]. It has been observed that position texts generally include position requirements and qualifications. In position requirements, core skills often appear many times; while in qualifications, core skills (such as computer language, framework or platform to be mastered, generally English vocabulary) tends to appear only once. For example, the recruitment data for an "artificial intelligence" position is:

Position Requirements: 1. Provide technical solutions (2D and 3D vision) for customers using the company's machine vision products; ... 3. Responsible for technical evaluation and software development of machine vision projects.

Qualifications Required: 1. College degree or above, ..., 4. Experience in developing machine vision development kits such as VisionPro, OpenCV, Halcon is preferred; ...

In this job posting, "vision" appears three times, "VisionPro", "OpenCV", and "Halcon" appear once each, and these four are core skills. Drawing on the idea of TF-IDF, the weight of the words in the position text is defined as B-IDF (Boolean-Inverse Document Frequency). The calculation formula of the weight of the i-th word in the j-th type of position is:

\[ w_{ij} = b_{ij} \times \text{idf}_i = b_{ij} \times \log \left( \frac{N}{n_i} + 0.01 \right) \]  

(1)

where \( b_{ij} \) represents the Boolean frequency of the i-th word, if the i-th word appears in the j-th category, \( b_{ij} \) is 1, otherwise 0; \( \text{idf}_i \) represents the inverse document of the i-th word Frequency, \( N \) represents the total number of recruiting data, and represents the number of recruiting data where the i-th word appears. Represents the total number of position documents (i.e., the number of position categories), and \( n_i \) represents the number of position documents in which the i-th word appears.

III. EXPERIMENT AND RESULT ANALYSIS

The experiment uses the Python development environment, uses Scrappy as crawler framework, and uses jieba word segmentation as Chinese word segmentation tool, which supports custom dictionaries.

A. Recruitment demands analysis

1) Recruitment demands for second-level positions

The recruitment data is classified according to 10 second-level positions to count the number of recruits, and the results are shown in Figure 1.

**Fig. 1. Recruitment demands for second-level positions in computer category**

Figure 1 shows that the demand for operation / technical support and back-end development posts is the largest, accounting for about 60% of the total demand, while the demand of artificial intelligence for higher vocational students in recent years is less than 2%. Vocational colleges should strengthen the training of traditional majors and cautiously set up new majors such as artificial intelligence.

2) Recruitment demands for third-level positions

The 10 second-level positions in the computer category can be refined into 116 third-level positions, and their recruitment requirements are shown in Figure 2.

**Fig. 2. Recruitment demands for third-level positions in computer category**

Figure 2 shows that the sum of the top 10 three-level positions with the largest demand, such as software engineer, technical support / maintenance engineer and Java development engineer, accounts for 49% of the total demand, while the sum of the remaining 106 posts is 51%. The demand for three-level Posts conforms to the 80/20 rule, and higher vocational colleges need to avoid minority needs when setting the direction of professional training.

3) Recruitment demands of each city

The number of recruits is counted according to the cities in the Yangtze River Delta region, and the results are shown in Figure 3.
Figure 3 shows that the recruitment demand in Shanghai exceeds one third of the total demand. The top 10 cities such as Shanghai, Nanjing and Hangzhou account for 89% of the total demand, while the rest only account for about 11%. The more developed the economy, the greater the demand for computer talents.

B. Salary Distribution

The salary distribution of each position is not uniform. This paper uses the boxplot method to display the salary distribution. The number on the horizontal line in the box represents the median salary, which measures the average level of salary for each position. The median can effectively reduce the impact of outliers on the average level. The lower line of the box represents the first quartile (Q1) of salary, and the upper line of the box represents the third quartile (Q3) of salary. The difference between the two is called the interquartile range (IQR, interquartile range); the upper and lower T-shaped boxes of the box are the inner limits, and their values are Q3+1.5IQR and Q1-1.5IQR respectively; the values at Q3+3IQR and Q1-3IQR are called outer limits [9]. Data within the inner and outer limits are called mild outliers, and data outside the outer limits are called extreme outliers. This paper first finds extreme outliers through boxplots and removes them, then uses the remaining data to draw boxplots, and plots the average value on the boxplots (represented by green triangles).

1) Position salary analysis

The salary distribution is calculated according to the 10 second-level positions, and the results are shown in Figure 4.

Figure 4 shows that the median salary of mobile development, frontend development, and game positions is the highest (all are 12,500 yuan), while the median salary of operation/technical support and data is the lowest (both are 7,000 yuan). On average, there are certain differences in the income of different second-level positions, and the salary level of positions with lower technical content (such as operation/technical support, testing) is also lower. The special case is the data post. It is possible that the higher vocational level is mainly engaged in posts with low technical content such as data collection and labeling. Among the three types of positions with the highest median, in terms of average value, the game is the highest, followed by mobile development, and the frontend development is the lowest. The salary dispersion above the median of game and mobile development is greater. For these two types of positions, the higher the skills level, the more likely it is to get a higher salary.

2) City salary analysis

There are many cities in the Yangtze River Delta region, and the 10 cities with the highest median salary are selected. The results are shown in Figure 5.

Figure 5 shows that there is not much difference in median salary. Shanghai has the highest median salary and average salary, followed by Hangzhou. Shanghai, Hangzhou, Nanjing and Suzhou have more high salaries, and in terms of development space, they are the first choice for graduates of higher vocational colleges.

C. Analysis of Position Core Skills

The word segmentation tool cannot recognize some computer professional terms in the position text, such as "computer vision". In CNKI, the name of the second-level position in the computer category is used as the subject keyword to search, and CNKI returns the main subject and secondary subject of the search, and adds these subject to the custom dictionary of jieba word segmentation.

1) Comparison of B-IDF and TF-IDF postion core skills extraction methods

In order to compare the difference between the B-IDF method proposed in this paper and the traditional TF-IDF method, we select the backend development secondary positions to extract the top 20 postion core skills, and the results are shown in Figure 6.
As can be seen from Figure 6, 16 of the 20 core skills extracted by the two methods are the same. The 16 same words can better reflect the core skills required for back-end development. The difference is that for the remaining four core skills, B-IDF considers JavaScript, C#, language, and coding, while TF-IDF considers technology, project, system, and front-end. Obviously, JavaScript and C# are the programming languages required in the back-end development, while language and coding emphasize that the back-end development needs to master a certain language and coding ability; Technology, projects and systems are common skills that other second-level positions need to master except operation/technical support, testing and technical management. They do not belong to the unique core skills of backend positions, while the frontend belongs to the frontend development positions. From the comparative analysis of the experimental results, it can be seen that the B-IDF method proposed in this paper can better extract the post core skills.

2) Analysis of position core skills

In traditional majors, the demand for operation/technical support and back-end development of second-level positions is the largest, and the median salary of back-end development is much higher than that of operation/technical support. Therefore, back-end development is selected to investigate its core skills. In order to further investigate the skill requirements of three-level positions, the software engineer post with the greatest demand in back-end development is selected. Among the emerging majors, choose the most popular second-level position of artificial intelligence in recent two years, and also choose the third-level position of image algorithm engineer subordinate to artificial intelligence to investigate the core skills. In the position text, calculate the b-idf value of vocabulary by position, and sort the vocabulary according to its b-idf value from large to small, as shown in Table 1.

As shown in Table 1, (1) the core skills of the post of back-end development and software engineer mainly include five aspects: framework (spring), database (MySQL or Oracle), language (Java, C#, JavaScript), project development (development, code, programming, design, software development, requirements) and document writing (writing, document). Software engineers generally also need to master the skills of website development (CSS, Web). The core skills of artificial intelligence mainly include machine vision (vision), algorithm, production line implementation (automation, camera, light source, debugging, model selection, lens, industry, electrical, electromechanical), software tool library (Halcon, OpenCV) and coding implementation (C++, C#, programming). The core skills of image algorithm engineer mainly include algorithm (deep learning, machine learning, pattern recognition), software tool library (Halcon, OpenCV), machine vision (vision, computer vision), image processing (image, camera, detection, night shooting, beauty) and coding implementation (C++, Python, programming). (2) All four positions emphasize hands-on ability. For example, back-end development and software engineer positions require project development and document writing ability, artificial intelligence positions require production line implementation and coding ability, and image algorithm engineer positions require coding ability. (3) Both artificial intelligence and image algorithm engineering positions require the mastery of relevant algorithms, which brings challenges to the training of students at the higher vocational level. (4) The main field of artificial intelligence positions is still computer vision and the combination with hardware (production line implementation, robotics).

IV. Conclusion

This paper crawls the recruitment information with college degree under the category of "computer / Internet / communication / Electronics" published on 51job website, and obtains the recruitment data of computer at higher vocational level in the Yangtze River Delta through
preprocessing. By analyzing the recruitment demand and salary of positions and cities, it is found that there are great differences in recruitment demand, the median salary level has a certain difference, and three quarters have a great difference. Mining the core skills from the post text, it is found that higher vocational students are generally required to have hands-on ability.

1) Position demand. The demand for operation / technical support and backend development positions is the largest, and the sum of the two accounts for about 60% of the total demand; Emerging positions such as data and artificial intelligence account for a small proportion, accounting for 3.1% and 1.5% respectively. Higher vocational colleges need to set up courses reasonably, firmly grasp the employment market of traditional majors, and cautiously set up emerging majors. According to the analysis of cities, Shanghai has the largest demand, accounting for about 1/3, and Nanjing, Hangzhou and Suzhou account for about 10% respectively; The 10 cities with the largest demand account for 90% of the total demand. Higher vocational colleges need to encourage students to develop in big cities and cultivate students' stress resistance to adapt to the pace of life and competitive pressure in cities.

2) About salary. The median salary of mobile development, frontend development, and game positions is the highest, and the median salary of operation/technical support, data is the lowest, and the difference between the highest and lowest median is 5,500 yuan. In terms of three-quarters of the figures, the game is about 10,000 yuan higher than the operation and maintenance / technical support, and the maximum salary difference is even greater. If higher vocational students want to get high salary, they need to study in posts with strong professionalism and high technical level. Among the top 10 cities, there is a certain difference in median salary, and Shanghai has the highest median salary; three-quarters and the highest salary in Shanghai are significantly higher than other cities.

3) Core skills. Whether it is a second-level position or a more specific third-level position, higher vocational students are required to have hands-on ability. Artificial intelligence-related positions also require mastering algorithms, which requires students to have a certain theoretical foundation and strong understanding ability. Higher vocational colleges need to set up mathematics courses that can support students' understanding of artificial intelligence.

ACKNOWLEDGMENT

The research was supported by Major Project of National Social Science Fund (Grant No. 19ZDA364), Major Research Projects of Natural Science of Colleges and Universities in Anhui Provincial (Grant No. KJ2021ZD0175) and Key Research Projects of Natural Science of Colleges and Universities in Anhui Provincial (Grant No.KJ2020A1120).

REFERENCES