Analysis of data crawling and visualization methods for recruitment industry information

Song Zihan¹, Yang Yanhong¹* and Guo Hongtaí¹

¹ School of Information Engineering, Beijing Institute Of Graphic Communication, Beijing, China
*Corresponding author’s e-mail: yangyanhong@bigc.edu.cn

Abstract: With the popularity of the Internet, major job boards have become the main way for job seekers to obtain information, but the current job boards generally lack clear information display, overall industry trend analysis and data visualization, which makes job seekers fall into confusion. In this paper, we first crawl the job boards based on Scrapy framework, store the data in EXCEL after cleaning, explore the potential connection between different data through Apriori algorithm, and finally visualize the different job information in a targeted way. In this paper, we take the job information of printing industry as an example for implementation, extract key information, and propose a suitable visualization method for different types of job information. The method proposed in this paper can not only help job seekers understand the current talent demand of the industry in a simple, intuitive and fast way, but also has some guiding significance for the talent training program of universities.

1. Introduction

Being in the information age, online recruitment has become the mainstream way for companies to recruit talents, which has the advantages of high flexibility and targeting, but also has the disadvantages of numerous information and low efficiency. For job seekers, it is difficult to understand the industry's demand for talent from various job postings. How to visualize the recruitment information clearly and easily to show the needs of the industry is of great practical value for job hunting.

In this paper, taking recruitment in printing industry as an example, the crawler technology based on Scrapy framework[1] is used to collect recruitment information. Jieba word segmentation[2] is used to clean data. Apriori algorithm[3] is used to mine the hidden association and value of massive data. Finally, different types of data are displayed in a multi-dimensional way by matching visualization methods. Thus, the employed people can understand the talent demand and current situation of the industry simply, quickly and clearly.

2. Data collection

2.1. The principle of Scrapy framework

Scrapy provides a simple and convenient framework for data crawling, and is the mainstream framework for crawlers. The process of crawling information using Scrapy framework can be roughly described as follows: the crawler sends the URL to be processed to the engine, the engine sends a request to the scheduler, and then sends the processed request to the downloader, and then provides the resulting content to the crawler, which then performs further processing, and finally gets the next level URL or puts the resulting data into the corresponding database.
2.2. Data Collection Implementation

In this section, we introduce the data collection process of China Printing Talent website, and the obtained data are only used for academic analysis. The crawler is designed based on the Scrapy framework.

First of all, analyze the URL setting of the China printing talent website, take the business management position as an example, the url of the management position is "http://www.pjob.net/public/J1100_T1_P{{}}", the difference between different pages is that the page number in "{}" changes, so that the list of URL is obtained, and the page is obtained by get_page function.

Then we analyze the composition of the web page, and write regular expressions to get the required information from the web page. In this paper, the information to be collected is post name, work location, salary level, post requirements, company name and post category. Based on this, regular expressions are written to filter the key information. The code of regular expression matching is shown in Figure 1. Finally, traverse the URL list, get the corresponding page, call the regular expression for data matching, and store it in EXCEL.

```python
def get_info(page):
    title = re.findall("<h3>(.*?)</h3>\n", page, re.S)
    item = re.findall("<a.*?>(.*?){3}</a>\n", page, re.S)
    data = []
    for title in titles:
        for item in items:
            data.append({title: item})
    return data
```

Figure 1  Regular Expression Matching

2.3. Data Cleaning

Before data visualization, the collected data should be cleaned, mainly including data normalization, merging duplicate items, removing deactivated words and deleting illogical data items. In this paper, we use Jieba word separation for data cleaning, because Jieba word separation is rich in functions and easy to operate.

First, unify the data with different names but the same meaning. For example, "Beijing" and "Beijing City" have the same meaning, so they can be unified as "Beijing". Second, use the jieba library in Python to split the words and remove the deactivated words such as "the", "land", "we", etc. In addition, it is also necessary to eliminate data that do not make sense and values that deviate significantly from the rest of the data. For the salary data, the data item "face-to-face" was excluded before the visualization. For the data in the given salary range, the midpoint of the interval is selected as the value of this item.

3. Visualization method design

The content of recruitment information is complex and extensive, and the visualization method commonly used in the current market is to display charts classified by information categories, such as salary level, geographical distribution, etc. Although this visualization method has certain practicality, it wastes the hidden value under the massive data. Nor can the employed have a simple and clear understanding of the current industry. In this paper, Apriori algorithm will be used to explore the hidden internal relations and key points between recruitment data, and design corresponding visualization methods for different types of data.

3.1. Visualization design of association rules

A total of 9888 pieces of data are crawled on the China Printing Talent website, and each piece of data includes information such as job name, job category, work location, salary level, job requirements and
company name. These information seem to be relatively independent, but in fact there are many hidden relationships. Through the Apriori algorithm, we can dig deeper into the hidden connections and values in the job information, and then present them in the form of network graphs, which can provide new ideas for job seekers.

Apriori algorithm is one of the mainstream algorithms in data mining, and its main algorithm process is divided into two steps, firstly, matching frequent items that meet the support threshold, and secondly, finding strong association rules that meet the confidence threshold, and finally getting the association rules that meet both the minimum support and the minimum confidence.

The support represents the probability that a piece of data contains both keyword A and keyword B. The formula is shown below.

\[
\text{Support}(A \rightarrow B) = \frac{n(A \cap B)}{N}
\]  

(1)

The confidence level represents the probability of having keyword B given that keyword A is already available, and the formula is shown below.

\[
\text{Confidence}(A \rightarrow B) = \frac{n(A \cap B)}{N_A}
\]  

(2)

\(n(A \cap B)\) denotes the confidence of keyword A and keyword B, \(n(A \rightarrow B)\) denotes the confidence of keyword A and keyword B, \(n(A, B)\) is the number of data containing both keyword A and keyword B, \(N\) denotes the total number of data, \(N_A\) denotes the number of data containing keyword A.

Since the amount of data relied on in this paper is not sufficient and the data contains many categories, if the support threshold and confidence threshold are set too high, the accurate association rules cannot be obtained, thus making the Apriori algorithm meaningless. Combining the characteristics of the crawled printing industry job information to show the representative association rules in the job information, here the support threshold is set to 0.01 and the confidence threshold is set to 0.7 to get the association rules with confidence over 70% as shown in Table 1, which have higher confidence and stronger association.

Table 1  Association rules for job information in the printing industry

| Association rules                                      | Confidence |
|--------------------------------------------------------|------------|
| Dongguan, junior high school education or above-> Production technology | 76%        |
| Prepress-> junior high school education or above        | 76%        |
| Sales and customer service-> junior high school education or above | 75%        |
| Production technology, two years working experience-> junior high school education or above | 75%        |
| Management-> college student                            | 71%        |

The association rules are obtained based on Apriori algorithm, and then the relationship network graphs are constructed according to the confidence level, the antecedent and the consequent items of the frequent item set to show the inner connection between the job information. As shown in Table 3.1, the association confidence level of low-technology jobs such as pre-press, sales and customer service, and production technology with education requirements above junior high school reaches more than 70%, and the education requirements of these jobs are not high. And the confidence level of the jobs of operation and management requiring college degree or above reaches 71%. It can be found that although the educational threshold for printing-related low-end jobs is low, it is still quite important to have a college degree or above to get a more desirable position.

3.2. Job and requirement visualization design

By comparing the matching relationship between job requirements and their own abilities, it not only
enables job seekers to quickly choose the jobs they can do, but also promotes job seekers to understand which requirements the industry needs talents to meet and helps job seekers to clarify the direction of subsequent efforts.

In order to quickly extract the key information in large paragraphs of text about recruitment requirements, this paper extracts the keywords in each job posting based on word frequency. The WordCloud[4] library is used to adjust the word cloud parameters, construct the word cloud according to the word frequency, change the word cloud layout, color and style, and use the word cloud map to display the printing industry recruitment requirements. The job data is similar to the demand data, which has the characteristics of clear keywords and obvious word frequency gaps, and is also displayed in the form of word cloud to achieve the purpose of fast, clear and intuitive understanding of industry demand.

3.3. Geographical distribution visualization design

Through the crawling data found that the printing industry recruitment geographical distribution is obvious, so we use Echart to draw a heat map based on the number of jobs in each city, and show the demand for printing talents in the region through the change of regional color shades.

4. Visualization results and analysis

After the analysis and processing of the recruitment information and the visualization design, this paper will show the essence of the recruitment information in four forms: network diagram, word cloud, heat map and area map, and analyze the current situation of talent demand in the printing industry based on the visualization results.

4.1. Network diagram display

Based on R language and Apriori algorithm to obtain association rules between recruitment data, when the support threshold is set to 0.01 and confidence threshold is set to 0.7, association rules with strong association are obtained, but their number is limited because the total amount of data is not large and the data that meet the support threshold and confidence threshold are few and far between. Therefore, in order to better show the possible hidden connections between the data, the support threshold is set to 0.005 and the confidence threshold is set to 0.3, which can maximize the possible correlations in the limited data. Finally, the network diagram about 100 association rules is drawn by the arulesViz package, as shown in Figure 2.

![Graph for 100 rules](image-url)

Figure 2  Printing industry recruitment information association rules network diagram
According to the rules of correlation, it can be found that positions with low experience requirements are often not demanding on education requirements, while jobs with higher education requirements for candidates, the probability is that there is also an emphasis on work experience. For the current printing industry, highly educated people need to have not only sufficient knowledge but also certain work experience to stand out in the job application.

4.2. Word cloud display
Job categories, competency requirements and other data in the job information have obvious keywords, and the word cloud display according to the hotness of the keywords can achieve a clear effect at a glance. Through Echart, the word cloud of popular positions is drawn according to the word frequency of keywords in job requirements, as shown in Figure 3.

"These types of positions cover different positions from technology to management, indicating that the demand for talents in the printing industry is comprehensive and not concentrated in a certain field. The word cloud of ability demand mainly revolves around "junior high school education or above", "high school education or above", "college education or above", "more than 2 years of experience " and "more than 3 years of experience", it can be seen that the printing industry recruiters are more inclined to candidates with sufficient work experience than education.

4.3. Heat map display
The use of geographical heat map to show the distribution of jobs can not only show the difference in the number of jobs, but also have a clear concept of the geographical location, from the upper perspective to understand the nationwide demand for each position. Figure 4 shows that the printing industry jobs are mainly concentrated in the central and southern coastal areas, including Dongguan and the main.
4.4. Area map display
Salary is often linked to work experience and education, and the relationship between them is also the focus of attention of job seekers. Echart draws an area chart for salary and experience, as shown in Figure 5, to show the relationship between salary distribution and work experience requirements.

The area diagram for salary and education is shown in Figure 6, which shows that the requirements for education in the printing industry are not high, and the minimum education required for each salary level is not very different.

Experience is a more important indicator in printing industry recruitment than education. For jobs with salaries above 5k, work experience is a knock on the door of recruitment, and as salaries rise, the requirements for experience become more stringent.

5. Conclusion
In this paper, we collected job information of printing industry through Python crawler, cleaned and processed the information, and then visualized different types of data to maximize the value behind the data. Finally, through the visualization results, the current recruitment situation of printing industry is displayed concisely and intuitively. The visualization method adopted in this paper has certain reference significance for recruitment information, which is helpful to express the recruitment demand more intuitively, concretely, simply and clearly, and is helpful for job seekers to make a choice.
Acknowledgments
This paper is supported by the scientific research program of Beijing Municipal Education Commission (No. KZ202010015021, KM201810015010, KM201610015001) and BIGC project (Ec202002, Eb202103).

References
[1] Chen, L., Ma J. (2020) The Dynamic Monitoring Algorithm of Agricultural Network Spatial Information Based on Scrapy. Journal of Shandong Agricultural University (Natural Science Edition), 51(02):253-258.
[2] Hao, M., Xu X., Yu H., Zhao X., Xu Z. (2018) Emotion classification and prediction algorithm based on Chinese microblog. Journal of Computer Applications., 38(S2):89-96.
[3] Yang M., Fan X., Xu H. (2020) Research on application of user interest model and Apriori algorithm in employment recommendation of college students. Modern Electronics Technique., 43(13):119-122.
[4] Bao C., Wang Y. (2021) A Survey of Word Cloud Visualization. Journal of Computer-Aided Design & Computer Graphics., 33(04):532-544.