Research on Data Analysis and Visualization of Recruitment Positions Based on Text Mining

Yuanyuan Chen and Ruijie Pan
School of Information Technology, Luoyang Normal University, Luoyang 471934, China
Correspondence should be addressed to Yuanyuan Chen; chenyuanyuan@lynu.edu.cn

Received 23 May 2022; Accepted 1 July 2022; Published 20 July 2022

Copyright © 2022 Yuanyuan Chen and Ruijie Pan. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

With the rapid development of the Internet and the impact of COVID-19, online recruitment has gradually become the mainstream form of recruitment. However, existing online recruitment platforms fail to fully combine the job seekers’ demands for salary, region, benefits, and other aspects, which cannot display the information related to recruitment positions in a multidimensional way. To solve this problem, this paper firstly uses a web crawler to collect job information from recruitment websites based on keywords retrieved by users, then extracts job information using regular expressions, and cleans and processes the extracted job information using third-party libraries such as Pandas and NumPy. Finally, through the probabilistic theme model of text mining, the topic model of job description content in the recruitment information is modeled. Combining with the django development framework and related visualization technology, the relationship among education requirement, experience requirement, job location, salary, and other aspects in the recruitment information is visually displayed in a multidimensional way. At the same time, the GM model is used to realize the gray prediction of the number of employment personnel in related industries, which provides employment reference for the majority of job seekers and enterprises.

1. Introduction

In recent years, under the dual influence of the strong development of the Internet and the COVID-19 pandemic, enterprises and job seekers are no longer limited to traditional offline recruitment but gradually tend to online recruitment. With the rapid development of the Internet, online recruitment has become the main channel, which has obvious advantages compared with the traditional channels. Moreover, the form of online recruitment has been recognized by more and more job seekers and enterprises in the past ten years and has been gradually accepted by people. At present, recruitment websites can be roughly divided into three categories: comprehensive, vertical, and social. Among them, vertical recruitment websites are highly praised with high recruitment efficiency and good user experience, such as online recruitment platforms such as pull hook and push. According to statistics, about 20 million pieces of employment information are released every day around the world, and about 30 million people send job resumes on the Internet. Online recruitment not only provides convenience for job seekers but also brings problems to users. The question of how to extract and analyze effective information from the massive amount of information released by various enterprises so as to better realize the interaction between talents and enterprises and improve the matching degree has become particularly important.

As for the research of recruitment information, some foreign scholars used the text mining method to analyze the talent market demand in the early stage. Todd et al. [1] analyzed the contents of advertisements published in newspapers based on the frequency of keywords, studied the changes of knowledge and skill requirements of relevant positions, and provided reference and guidance for education and recruitment. Lee and Lee [2] collected and analyzed the recruitment advertisements issued by Fortune 500 companies, constructed a skill requirement classification list, and showed the overall trend of job skill requirements by...
counting the number. Sodhi and Son [3] proposed a computer-based content analysis method based on the employment information related to operation research as the data source to construct a skill dictionary and keyword dictionary, which is very useful for regularly analyzing recruitment advertisements to monitor changes. Smith and Ali [4] analyzed and collected the data of employment requirements in the programming field based on data mining technology, analyzed several popular programming languages in recent years, provided guidance for the arrangement of computer-related courses in colleges and universities, and made an in-depth analysis of the market trend of programming work. Compared with foreign countries, the research on recruitment information in China is a little slow, which is related to the development stage of online recruitment. However, with the in-depth development of the Internet in China, the forms of online recruitment show a diversified trend, and domestic scholars' research on recruitment information is becoming more and more in-depth. For example, Zhang and Ruibin [5] built a data job recruitment dictionary based on Chinese word segmentation and natural language processing technology to analyze and mine the talent demand characteristics of domestic data jobs. Yan et al. [6] proposed a three-level curriculum knowledge model of "post-curriculum-knowledge point" and combined it with natural language text mining technology to realize the automatic construction of the curriculum knowledge point model so as to provide teaching and learning reference for colleges and students. Ling and Gao [7] analyzed multisource, heterogeneous, and unstructured online recruitment information based on text mining technology to help college professional managers and builders quickly and accurately understand the needs of enterprises for professional talents and provide decision-making guidance for them to formulate professional talent training programs that meet the needs of enterprises. On the whole, the above research is relatively simple in the mining of recruitment information without digging deeper information, which is of limited help to job seekers. Moreover, a complete visualization system has not been built, so users cannot have a clear and intuitive in-depth understanding of recruitment information for a certain industry.

This paper, taking the research of the above problems, focuses on the mining of the hidden information of the recruitment. Combined with the relevant needs of job seekers, based on a recruitment website and web crawler technology, the data collection of recruitment information is collected. Further data processing and data analysis are carried out through Pandas, NumPy, Python, and other third-party function libraries, and the probabilistic topic model is used to model the content of job description in the recruitment information. GM (1, 1) algorithm is used to predict the number of employees employed in information transmission, computer services, and the software industry in the next ten years. Combined with the django development framework and PyEcharts visualization technology, this paper makes a visual multidimensional display of the relationship among education, experience, job location, and salary in the recruitment information. For graduates or former students, they can more clearly understand the relevant skill requirements of artificial intelligence positions, including salary distribution, geographical distribution, academic requirements, experience, and other information so as to cultivate relevant skills more pertinently, improve their own strength, and then face it calmly in the job search process. For enterprises, it can help companies analyze data in a shorter time, promote business growth, assist enterprises in data integration, and express the inherent value of data more vividly through the intuitionistic and interactive nature of charts, thus speeding up data analysis and decision-making speed. For colleges and universities, it is convenient for them to adjust relevant professional courses timely, better connect with relevant enterprises, and cultivate more high-quality talents to meet social needs and enterprise recruitment needs.

2. Related Work

2.1. Web Crawler. Web Crawler, also known as the web information collector, is a computer program or automatic script that automatically downloads web pages and is an important part of the search engine [8]. The schematic diagram of the general process of the Web Crawler is shown in Figure 1. With the diversity and complexity of information, the web crawler has attracted more and more attention and has been widely applied in various fields. For example, Peng et al. [9] based on the Python open source Scrapy framework not only helped researchers to carry out subsequent data mining analysis by collecting shipping job-hunting information but also provided data support for subsequent shipping job-hunting information databases. Chen et al. [10] used web crawler technology to mine and analyze the post information of online recruitment data and mine the information in massive network data so as to realize the accurate connection between the demand and supply of professional jobs. Long [11] realized scientific and technological literature retrieval based on web crawler technology, which greatly improved the efficiency and accuracy of scientific and technological literature retrieval and better served scientific research. Cong [12] designed an intelligent advertising delivery system based on web crawler technology to deliver advertisements accurately according to the needs of users, which significantly improved the product conversion rate of advertising.

According to the usage scenarios, web crawlers are mainly divided into general crawlers and focused crawlers. The general crawler is an important part of the search engine capture system (Baidu, Google, and many others). Its main purpose is to download web pages on the Internet to the local area and form a mirror backup of Internet content. The focused crawler, also known as the topic crawler, refers to web crawlers that selectively crawl pages related to predefined topics. Compared with the general crawler, it only needs to crawl the theme-related pages, which greatly saves hardware and network resources. The saved pages are also updated quickly due to the small number, which can better meet the needs of some specific people for information in specific fields. Taking Douban book information acquisition
2.2. Data Mining and Analysis. The mathematical basis of data mining and analysis was established in the early 20th century, but it was not until the emergence of computers that practical operation became possible and popularized. It mainly refers to the use of appropriate statistical analysis methods to extract valuable and meaningful information from a large number of messy and fuzzy data, conduct detailed research and summary, and find out the internal law of the research object. The data analysis process can be roughly divided into five stages, as shown in Figure 1. First of all, we need to clarify the purpose and ideas of data analysis, understand the data objects to be analyzed and the business problems to be solved, sort out the framework and ideas of analysis, and determine the analysis means and tools to be used. Data collection is the basis of data analysis, and there are many ways to collect data, such as some open source data sets published by many universities or government departments, data sets of major competitions, and data sources based on web crawlers mentioned above. At the same time, the quality of data analysis largely depends on the effect of data processing, which mainly refers to cleaning, processing, and sorting the collected data to lay a foundation for data analysis. Data analysis refers to the exploration and analysis of processed data through analytical means, methods, and techniques and the discovery of causal relationships and internal connections. The results of data analysis are often presented through visualizations such as the chart to clearly convey information to users, increase the spirituality of the data, and help users to quickly and easily extract the meaning of the data to a large extent, thereby reducing the time cost of users.

This research mainly uses the Pandas library for data analysis, which is a NumPy-based python library specially created to solve data analysis tasks. It not only includes a large number of libraries and some standard data models but also provides tools for efficient operation of large data sets. It is widely used in academic and commercial fields such as economics, statistics, and analysis. The focus of data analysis is to observe, process, and analyze the collected data to extract valuable information and play the role of data. Different from data analysis, data mining mainly refers to the process of mining unknown and valuable information and knowledge from a large amount of data through statistics, artificial intelligence, machine learning, and other methods. CRISP provides an open, freely available standard process for data mining that makes it suitable for problem-solving strategies in business or research units. As shown in Figure 2, this process is defined as six phases: business understanding, data understanding, data preparation, model building, model evaluation, and model deployment. Data mining is similar to the first three stages of data analysis. The main difference is that data mining processes data by constructing models, allowing models to learn the rules of data, and producing models for subsequent work. The purpose of model evaluation is to select the best model from many models so as to better reflect the authenticity of data. For example, for the prediction or classification model, even though it performs well in the training set, the results in the test set are mediocre, indicating that there is overfitting in the model.

According to incomplete statistics, 80% of the time in the data mining process is data preparation, and then appropriate models are considered for modeling. The task of data mining is to discover patterns hidden in data. The patterns that can be divided into descriptive patterns and predictive patterns. A descriptive pattern is a normative description of the facts existing in the current data, describing the general characteristics of the current data. The predictive model takes time as the key parameter and predicts the future value of time series data based on its history and current value [14]. In this study, probabilistic topic modeling based on the Gensim algorithm is used to mine the job description information in the recruitment positions. The gray prediction algorithm GM (1,1) is used to predict the gray level of the employment personnel in other units of information transmission, computer service, and software in China, which provides a comprehensive reference for job seekers.
2.3. Data Visualization. Data visualization is a technology that is widely applied in the data field and plays an important role. According to its application in different fields and tasks, different researchers have different understandings of it [15]. Waskom [16] pointed out that it is an integral part of the scientific process, and effective visualization will enable scientists to understand their own data and communicate their insights to others. Azzam et al. [17] proposed that it is a process that generates images representing original data based on qualitative or quantitative data, which can be read by observers and support data retrieval, inspection, and communication. Unwin [18] proposed that it refers to the use of graphic display to display data, and its main goal is to visualize data and statistical information and interpret the display to obtain information. Cheng et al. [15] believed that data visualization is a method of data modeling and expression, which aims to show some characteristics and internal laws of data through the model so that observers can more easily discover and understand these characteristics and laws of data. On the whole, data visualization is to visually convey the data displayed in texts or numerical values to users in a graphical way so that users can observe data from different dimensions and discover the inherent rules implied in data information, which is convenient for more in-depth observation and analysis of data.

The whole data visualization process can be divided into three steps: analysis, processing, and generation. The analysis stage can be similar to the first three stages of data analysis. The processing stage can be subdivided into two parts, data processing and visual coding processing, and the generation stage is mainly to put the previous analysis and design into practice. As early as 1990, Haber and McNabb [19] have proposed the basic process of data visualization. The whole process of this model is linear and very advanced, and the nested model and cyclic model proposed later are derived from this model. As shown in Figure 3, the model divides the data into five stages, which go through four processes, respectively, and the input of each process is the output of the previous process. In short, data visualization is the mapping of data space to graphic space.

A classic visualization implementation process is to process and filter the data, transform it into a visually expressive form, and then render it into a user visible view, as shown in Figure 4 [20]. In contrast to the previous linear model, the process adds a user interaction component at the end and keeps each step circular. At present, the visualization process model is widely used in almost all well-known information visualization systems and tools. It can be seen that no matter how the model changes, its essence needs to go through three stages of analysis, processing, and generation.

Data visualization perfectly combines art and technology and uses graphical methods to display massive data visually and intuitively. Python provides a variety of third-party libraries for visualization, such as Matplotlib, PyEcharts, Plotly, and so on. This study is mainly based on PyEcharts to achieve relevant visual display, which is produced by the combination of Baidu open source ECharts and Python. Compared with foreign HighEcharts, all the documents in this library are written in Chinese, which is very friendly to developers who are not good at English and has richer content.

3. Data Acquisition and Processing

3.1. Data Sources and Data Collection. Before data collection, this study comprehensively compares the authority, timeliness, and difficulty of data collection of recruitment websites, such as Zhaopin.com, 51 job, Lagou, Liepin, and other mainstream recruitment websites. This paper takes the Lagou.com platform as the experimental object and uses the keywords “artificial intelligence,” “AI,” and “algorithm” as search criteria to collect data on artificial intelligence jobs, and it obtains 1932 recruitment information including post name, salary distribution, city distribution, skill requirements, benefits, company name, company scale, financing, job description, post release time, and so on.

Data collection is mainly divided into data fetching and data parsing, and rich third-party libraries in Python are provided for them, such as urllib and requests libraries for data fetching and XPath, lxml, Beautiful Soup, or regular expressions for data parsing. This study is mainly based on a more user-friendly request library for data fetching. Compared with the URLlib library, it is not only convenient to
use but also saves a lot of work. Most importantly, it inherits all the features of urllib and supports some other features such as the use of cookies to maintain sessions and automatically determine the encoding of the response content. In addition, this study analyzes the data of the captured web pages based on regular expression, extracts the required information, and converts it into dictionary data through the JSON module so as to convert the unstructured data in the web page into structured data and store it in CSV file, which lays the foundation for subsequent data processing and data mining.

3.2. Data Processing

3.2.1. Data Preprocessing. The quality of data will greatly affect the results of data analysis. Some of the data collected in the early stage may be incomplete, such as data missing, outliers, duplicate values, and other problems as shown in Figure 5. True represents missing values, while False represents the presence of values. Therefore, data need to be preprocessed before data analysis and data mining, including data cleaning, merging, reshaping, and transformation. Among them, data cleaning is the primary and core link, and its purpose is to improve data quality, clean dirty data, and make the original data more complete, consistent, and unique. This research is mainly based on the Python third-party function Pandas library for data preprocessing. Data cleaning operations in Pandas include the processing of null and missing values, duplicate values, and outliers. A null value means that the data is unknown, inapplicable, or will be added later, and a missing value means an incomplete attribute in a piece of data. For null and missing values, one can generally choose to delete or fill them in. For duplicate values, in most cases, the duplicate entries are deleted, and only one valid piece of data is retained. Outliers refer to certain values in the data that deviate significantly from other observations in the sample to which they belong, and these values are unreasonable or wrong. For example, the keyword “overseas” appears in the statistics of artificial intelligence positions in various provinces, which is obviously inconsistent with the name of China’s provinces. After data cleaning, 1928 pieces of valid data are obtained, which lay a foundation for subsequent related operations.

3.2.2. Chinese Word Segmentation. Chinese word segmentation refers to adding boundary markers between words in a Chinese sentence. Compared with English, there is no space boundary between words in Chinese sentences, which makes them blurry. Guoju [21] pointed out that Chinese word segmentation, which is the basis of Chinese information processing and a subset of natural language processing, is the automatic addition of dividing lines between words in Chinese text by the machine, and its essence is demarcation. Wang and Liang [22] mentioned that word segmentation, as the first step of natural language processing, plays an indispensable role. Chinese word segmentation has become a research hotspot due to the complexity of language. It can be seen that it plays a crucial role in many fields, including the exploration of job description information or skill requirements in the recruitment positions based on probabilistic topic modeling and word cloud technology described in this paper, which requires Chinese word segmentation of text information, making it easier for computers to understand the text.

The representative methods of Chinese word segmentation include shortest path word segmentation, n-gram word segmentation, word segmentation by word structure, recurrent neural network word segmentation, transformer word segmentation, and so on. This study is mainly based on the Jieba word segmentation tool for Chinese word segmentation, as shown in Table 1; the data is mainly from GitHub. As of April 2022, it can be seen that Jieba’s number of stars ranks first. Jieba adopts the word segmentation method based on word formation and the Viterbi algorithm based on HMM, supporting four-word segmentation modes, such as the accurate mode, full mode, search engine mode, and paddle mode. Among them, the accurate mode tries to cut sentences in the most accurate way, which is very suitable for text analysis.

4. Data Mining

4.1. Probabilistic Topic Modeling. A topic model is a modeling method that can effectively extract the hidden topics of large-scale text [23]. It is mainly used for document modeling to convert documents into numerical vectors, and after conversion, each dimension of numerical vectors corresponds to a topic, thus the essence of topic model is to realize the structure of text data. As structured documents could be queried and compared with each other so as to realize traditional machine learning tasks. In addition, a topic model is a generalization concept, which generally refers to the classic topic model of Latent Dirichlet Allocation (LDA). It is the simplest probability topic model among topic models, which was proposed by Blei et al. [24] in 2003, and it is used to predict the topic distribution of documents, which could also present the topic of each document in the form of probability distribution. The model mainly solves the problems of document clustering and word aggregation, realizes the abstract analysis of text information, and helps analysts explore the implied semantic content.

This research is mainly based on the Gensim algorithm to carry out LDA probability subject modeling for job description text information of recruitment posts. The modeling algorithm is shown in Algorithm 1. First, load the corresponding data, preprocess the data, including data cleaning, Chinese word segmentation, and other operations, then conduct text vectorization, including generating corpus dictionary and sparse vector set, then conduct model training, input the vectorized text into the LDA model, set the corresponding number of topics, and finally obtain the distribution results of subject words.

4.2. Grey Prediction. Prediction refers to people making predictions about the development trend of human society, science, and technology based on available historical and
realistic data by using certain scientific methods and means so as to guide the direction of future actions. Prediction usually includes white prediction and black prediction; white prediction means that the internal characteristics of the system are completely known, and the system information is completely sufficient; black prediction means that the internal characteristics of the system are unknown and can only be correlated by observing its relationship with the outside world. The gray forecast is between the two, some information is known, the other is unknown, and there is an uncertain relationship between the system factors. Gray prediction is a method to predict the system with uncertain factors, by identifying the degree of difference between the development trends of system factors, that is, by association analysis and generating and processing the original data, the law of system changes can be found to generate data series with strong regularity. Then, a differential equation model is established to predict the future development trend of things [25]. The core system of gray prediction is the gray model (GM), which is a method of accumulating (accumulating and mapping) the original data to generate an approximate exponential law and then modeling. The results of the gray model prediction are relatively stable; it is not only suitable for prediction of a large amount of data, but also the prediction result is still more accurate when the amount of data is small (more than 4). The gray model mainly includes the GM (1, 1) model, GM (2, 1) model, DGM model, and Verhulst model, and this study is mainly based on the single-

**Algorithm 1: Gensim algorithm.**

```
LDA modeling using the Gensim algorithm
Input: job description text set (texts)
Output: topic inference
(1) function Gensim(texts)
(2) create part of speech table flags, stop word table stop words
(3) use the Jieba library to segment and filter
(4) words_ls ← []
(5) for text in texts:
(6) words ← remove_top words([w.word for w in jp.cut(text)])
(7) words_ls.append(words)
(8) end for
(9) dictionary ← corpora.Dictionary(words_ls)
(10) corpus ← [dictionary.doc2bow(words) for words in words_ls]
(11) LDA ← models.ldamodel.LdaModel(corpus, dictionary, num_topics=1)
(12) show the top 30 words in each topic
(13) for topic in lda.print_topics(num_words=30):
(14) print topic
(15) end for
(16) end function
```

### Table 1: Overview of mainstream word splitting tools.

| Short name | Full name | Development Time | Star |
|------------|-----------|------------------|------|
| Jieba      | Jieba participle | fsjsy          | 2012.09 | 28.5k |
| HanLP      | Chinese language processing package | hankcs        | 2014.10 | 25.8k |
| snownlp    | Chinese class library | Isnowfy     | 2013.11 | 5.8k  |
| FoolNLTK   | Chinese processing toolkit | rockyzhengwu | 2017.12 | 1.6k  |
sequence first-order linear differential equation model GM (1, 1) model in the gray system to make a gray prediction of the number of employees in information transmission, computer services, and software in the next decade so as to provide a reference for relevant job seekers.

The GM (1, 1) model represents a gray prediction model based on a first-order differential equation and one variable. Let the time series \( X^{(0)} \) have \( n \) observations shown as the following equation.

\[
X^0 = \{X^{(0)}(1), X^{(0)}(2), X^{(0)}(3), \ldots, X^{(0)}(n)\}. \tag{1}
\]

Take an accumulation generation on \( X^{(0)} \) to obtain an accumulation sequence \( X^{(1)}(i) \) shown as the following equation.

\[
X^{(1)}(i) = \sum_{m=1}^{i} X^{(0)}(m) (i = 1, 2, \ldots, n). \tag{2}
\]

The sequence obtained by one accumulation is shown as follows (\( m = 1 \)) as the following equation.

\[
X^{(1)} = \{X^{(1)}(1), X^{(1)}(2), X^{(1)}(3), \ldots, X^{(1)}(3), \ldots, X^{(1)}(n)\}. \tag{3}
\]

Using the first-order univariate linear dynamic model GM (1,1), the first-order differential equation of \( X^{(0)}(t) \) is shown as the following equation.

\[
\frac{dX^{(1)}(t)}{dt} + aX^{(1)}(t) = b. \tag{4}
\]

Then, use the least square method to find the values of \( a \) and \( b \), which are shown as equations

\[
A = (B^T B)^{-1} B^T y_n, \tag{5}
\]

\[
B = \begin{bmatrix}
-\frac{1}{2} [X^{(1)}(1) + X^{(1)}(2)] & 1 \\
-\frac{1}{2} [X^{(1)}(2) + X^{(1)}(3)] & 1 \\
\vdots & \vdots \\
-\frac{1}{2} [X^{(1)}(n-1) + X^{(1)}(n)] & 1
\end{bmatrix}. \tag{6}
\]

The prediction model is obtained by solving the differential equation shown as the following equation.

\[
X^{(1)}(k+1) = \left( X^{(1)}(1) - \frac{b}{a} \right) e^{(-ak)} + \frac{b}{a} (k = 1, 2, \ldots, n). \tag{7}
\]

The selection of the prediction model should be based on sufficient qualitative analysis conclusions, and it is necessary to go through a variety of tests to determine whether it is reasonable and effective. The prediction model is mainly judged by the following aspects: the accuracy test, relative stagger residual test, variance ratio test, and small error probability test. As shown in Table 2, the smaller the variance ratio \( C \) and the larger the small error probability \( P \), the higher the prediction accuracy. Generally, it is necessary to ensure that the \( C \) value is small enough. Even if the law of the original data is not obvious, it can ensure that the error range of the predicted value will not be large.

5. Visual Case Analysis

For the collected massive recruitment position information, information visualization and visual analysis methods are adopted so that users could have a clear and intuitive in-depth understanding of relevant industries. The main interface of the recruitment information visualization platform system is shown in Figure 6, which mainly includes four modules: home console, data management, job description, and chart analysis. The homepage console module is the overview display interface of recruitment information, which mainly includes the regional distribution of recruitment positions, the proportion of educational requirements, the dynamic sliding display of recruitment information (job name, salary distribution, regional distribution, and job posting time), and the function options bar. The data management module displays the artificial intelligence job information in detail, including skills requirements, company size, job description, company name, and other information in addition to the fields already displayed on the home page. Meanwhile, it supports global search, which is convenient for users to query relevant information. The job description module mainly displays the visualization results of probability topic modeling. Chart analysis is a visual analysis of the collected field information in a graphical way, including salary analysis, company size analysis, work experience analysis, word cloud map, and gray prediction.

5.1. Examples of Data Analysis

5.1.1. Visual Analysis of Salary and Other Factors. Salary is a direct reflection of the value of employees and an important factor for job seekers to choose an employer [26]. Figures 7 to 10 show the relationship between the average annual salary of AI positions and the province where the positions are located, educational background requirements, experience requirements, and company size. As can be seen from Figure 7, the highest average annual salary in this industry is about 400,000 yuan, and the high salary is mostly distributed in the eastern region and southeast region. It can be seen from Figure 8 that educational background is directly proportional to the average annual salary; that is, the higher the educational background is, the higher the salary is. In addition, the salaries offered by the industry for undergraduate graduates are still very impressive, which can greatly ease the concerns of job seekers. Figure 9 shows that work experience and the average annual salary have an
increasing trend. The average annual salary after working for more than one year can reach more than 300,000 yuan, and the salary of college students or fresh graduates can also reach about 160,000 yuan. Compared with other industries, the income is quite considerable. As can be seen from Figure 10, the average annual salary of an enterprise with more than 50 employees is more than 300,000 yuan, while that of an enterprise with less than 15 employees is nearly 300,000 yuan. Generally speaking, for job seekers, especially fresh graduates, when applying for relevant positions in this industry, priority can be given to relevant large-scale enterprises in southeast China.

5.1.2. Visual Analysis of Company Size and Work Experience. The number of small and medium-sized enterprises in China is increasing year by year, and each company requires more work experience for job seekers. It can be seen from Figures 9 to 10 that the longer the work experience and the larger the company scale, the higher the average annual salary. For job seekers, especially fresh graduates, whether there is an advantage when applying for a job, the current proportion of small and medium-sized enterprises is very concerned about the problem. Figure 11 shows the proportion distribution of company size in the 1928 effective recruitment messages, in which 32.69% of companies with more than 2000 employees take up the largest proportion, 21.71% of companies with 500–2000 employees, and 19.78%
of companies with 150–500 employees. The proportion of 50–150 is 15.04%, and the proportion of 15–50 is 9.79%. At the same time, Figure 12 shows the distribution of work experience required by each company in the recruitment process. The quantity distribution from high to low is 3–5 years of work experience, 1–3 years of work experience, college/fresh graduate, 5–10 years of work experience, more than 10 years of work experience, and less than one year of work experience. Figure 11 shows that the average annual salary is about 300,000 RMB in the company with a scale of more than 50 people, and the size of the company with more than 50 people accounted for 89.22%.

5.1.3. Visual Analysis of Skill Requirements and Benefits. While paying attention to salary, job seekers will also pay corresponding attention to the skill requirements and benefits provided by the company. Our results show keyword extraction of skill requirements and welfare benefits using word clouds, and it can be concluded that job seekers focus on learning python, deep learning, natural language, and image processing. At the same time, the company provides employees with benefits such as performance bonuses, five insurances and one housing fund, paid vacations, and flexible systems.

5.2. Data Mining Examples

5.2.1. Visual Analysis of Topic Modeling. The LDA probabilistic topic model based on the Gensim algorithm was visualized by PyLDAvis as shown in Figure 13. After Chinese word segmentation of the job description text, the topic model was established by Gensim and divided into 6 topics. The bubbles on the left are different topics, and
the top 30 feature words within the topic range are on the right. Light blue shows how often the word appears throughout the document, and dark red shows the weight of the word in the topic. In the upper right corner is an adjustable parameter $\lambda$. When $\lambda$ approaches 1, words frequently related to the topic will be displayed, indicating close relationship with the topic. When $\lambda$ approaches 0, words that are special and unique to the topic will be displayed. As can be seen from Figure 13, when $\lambda$ is 0.5, the topics in all job descriptions at this time mainly emphasize the learning ability, algorithm experience, technical proficiency, etc.

5.2.2. Grey Prediction. Figure 14 shows the gray prediction results of the number of employed persons in information transmission, computer services, and the software industry in the past decade. The abscissa axis is the year, and the ordinate axis is the number of employed persons in information transmission, computer services, and the software industry every year. In the Figure, from the year 2011 to 2020, the number of employed persons is the actual employed persons, and from the year 2021 to 2030, the number of employed persons is the estimated value by gray prediction. As shown in Figure 14, it is estimated that 5.75
6. Conclusion

This paper takes an artificial intelligence position on a recruitment website as an example and builds a recruitment position information visualization platform based on the web crawler, text data mining, mining data analysis, and other related technologies. Through a pie chart, histogram, funnel chart, word cloud chart, probability theme modeling, gray prediction, and other methods, this paper makes a visual analysis on the relationship between the salary and various factors, company size, and work experience proportion that users are concerned about and makes a gray prediction on the number of employees in other units of information transmission, computer service, and the software industry in the next decade so as to help job seekers understand the recruitment information of relevant industries in a more intuitive way and provide a comprehensive reference.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] P. A. Todd, J. D. McKeen, and R. B. Gallupe, “The evolution of IS job skills: A content analysis of IS job advertisements from 1970 to 1990,” MIS Quarterly, vol. 19, no. 1, pp. 1–27, 1995.

[2] S. M. Lee and C. K. Lee, “IT managers’ requisite skills,” Communications of the ACM, vol. 49, no. 4, pp. 111–114, 2006.

[3] M. S. Sodhi and B. G. Son, “Content analysis of OR job advertisements to infer required skills,” Journal of the Operational Research Society, vol. 61, no. 9, pp. 1315–1327, 2010.

[4] D. T. Smith and A. Ali, “Analyzing computer programming job trend using web data mining,” Issues in Informing Science and Information Technology, vol. 11, no. 1, pp. 203–214, 2014.

[5] J. Zhang and W. Rubin, “Mining the demand characteristics of data posts on domestic recruitment websites,” Journal of Intelligence, vol. 37, pp. 176–182, 2018.

[6] Y. Yan, C. Lei, and N. Zhao, “Research on automatic construction of curriculum knowledge model based on online recruitment text mining,” Library and information work, vol. 10, Article ID 10848, 2019.

[7] Li Ling and M. Gao, “Analysis on the skill needs of professionals in the era of online recruitment -- Taking the specialty of information management and information system as an example,” Intelligence exploration, vol. 11, pp. 53–57, 2018.

[8] L. Sun, G. He, and L. Wu, “Research on web crawler technology,” Computer knowledge and technology, vol. 15, pp. 4112–4115, 2010.

[9] D. Peng, T. Li, Y. Wang, and C. L. Philip Chen, “Research on information collection method of shipping job hunting based on web crawler,” in Proceedings of the Eighth international conference on information science and technology (ICIST), p. 132, July 2018.

[10] J. Chen, K. Li, Z. Liu et al., “Data Analysis and Knowledge Discovery in Web Recruitment-Based on Big Data Related Jobs,” in Proceedings of the International Conference on Machine Learning, Big Data and Business Intelligence (MLBDII), p. 1274, November 2019.

[11] X. Long, “Application of web crawler in scientific and technological literature retrieval,” Modern information technology, vol. 5, no. 24, pp. 150–152, 2021.

[12] C. Cong, “Hong Minmin Intelligent advertising recommendation based on web crawler technology,” Information technology and informatization, vol. 7, pp. 239–242, 2021.

[13] C. Du, “Preliminary design and implementation of focus crawler based on Python,” Modern manufacturing technology and equipment, vol. 12, no. 289, pp. 30–31, 2020.

[14] G. Wang and P. Jiang, “Overview of data mining,” Journal of Tongji University, vol. 2, pp. 112–118, 2004.

[15] J. Cheng, H. you, and S. Tang, “Application of data visualization technology in military data analysis,” Information theory and practice, vol. 43, no. 9, pp. 171–175, 2020.

[16] M. L. Waskom, “Seaborn: statistical data visualization,” Journal of Open Source Software, vol. 6, no. 60, p. 2021, 2021.

[17] T. Azzam, S. Evergreen, A. A. Germuth, and S. J. Kistler, “Data visualization and evaluation,” New Directions for Evaluation, vol. 2013, no. 139, pp. 7–32, 2013.

[18] A. Unwin, “Why is data visualization important? what is important in data visualization?” New Directions for Evaluation, vol. 2, no. 1, pp. 1–7, 2020.

[19] R. B. Haber and D. A. McNabb, “Visualization idioms: A conceptual model for scientific visualization systems,” Visualization in scientific computing, vol. 74, p. 93, 1990.

[20] L. Chou, “Data-visualization-tools-that-you-cannot-miss-in-2019,” 2019, https://towardsdatascience.com/9-data-visualization-tools-that-you-cannot-miss-in-2019-3f23222a927.

[21] S. Guojun, “Research on Chinese word segmentation technology based on Python,” Wireless Internet technology, vol. 18, no. 23, pp. 110-111, 2021.

[22] J. Wang and Y. Liang, “A review of Chinese word segmentation Software guide,” Journal of Biomedical Informatics, vol. 20, no. 4, pp. 247–252, 2021.

[23] B. Wang, S. Liu, and K. Ding, “Patent content analysis method based on LDA subject model,” Scientific research management, vol. 36, no. 3, p. 7, 2015.

[24] D. M. Blei, A. Y. Ng, and M. I. Jordan, “Latent dirichlet allocation,” Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[25] X. Zhou, J. Xiang, and X. Guo, “Analysis and Mining of Network Recruitment Information,” Statistics and Applications, vol. 5, no. 4, pp. 389–396, 2016.

[26] L. Chang, “Research on job requirements of data analysis based on Web text mining,” China management informatization, vol. 21, no. 10, p. 4, 2018.