Data Mining System and Construction Analysis of Skilled Talents based on Big Data

Yibo Guan¹,*
⁰Guangdong Business and Technology University, Zhaoqing, Guangdong, 526060
*Corresponding author e-mail: dxyhaohao1183@gzgs.edu.cn

Abstract: The cultivation of high-quality skilled talents is an important subject that needs to be improved in the current vocational college education. Through the construction of the training mechanism corresponding to the specialty, the students' practical ability can be promoted. Driven by big data, many technical enterprises gradually strengthen the recruitment of technical talents. However, due to the wide distribution of talents, data mining is more difficult. Therefore, it introduces the data mining system and construction analysis of skilled talents based on big data. This paper first analyzes the demand of skilled talents; secondly, the popular data mining technology is applied to this big data system, and the reasons for the low efficiency of data analysis and processing of skilled talents are obtained; then, the Bayesian algorithm in data mining technology is used to classify the technical talents, so that the enterprises in need can select and locate the skilled talents Data positioning is also carried out. Finally, it is concluded that the system can effectively promote the recruitment of technical talents in enterprises, and the ability level of these people after entering the company also meets the post requirements.

Keywords: Big Data, Bayesian Algorithm, Skilled Talents, Data Mining Technology

1. Introduction

From the current talent training direction of higher vocational colleges, the positioning of training objectives is mainly reflected in the training of skilled and applied talents. In today's economic society, we need not only skilled talents, but also innovative skilled talents. It can be seen that innovative skilled talents are the new force of the current social development. Therefore, to promote the reform of Higher Vocational Education [1-3] and strengthen the cultivation of innovative skilled talents is not only an important measure to improve higher vocational education, but also an inevitable requirement to adapt to the society. While the society cultivates skilled talents, major enterprises need to tap skilled talents to fill the vacancy problem [4]. However, due to the huge amount of resume investment, it is difficult to accurately find the right candidates, which leads to the low efficiency of human resources department in recruiting employees. In the face of massive big data, we have encountered
unprecedented barrenness.

In the management of modern technical talents [5-6], the level of informatization has been improved year by year. With the accumulation of recruitment system data of major enterprises over the years, the big data environment for technical talents [7] has been formed. It is mainly reflected in the characteristics of large-scale, multi-type, high-speed and low-density of technical talent data. How to effectively mine the data of different types of technical talents has become an important content to improve the level of talent information management. In this paper, through the establishment of data mining system [8], using Bayesian algorithm [9-10] to mine the post data of technical talents. Through the establishment of enterprises to analyze the characteristics of technical talent demand, and applied to the talent recruitment network, and then through the feasibility of a large number of data mining and the scientificity of Bayesian algorithm, the inherent talent recruitment mode is changed, and the enterprise's goal of information-based recruitment of technical talents is implemented, so that those who have the ability to choose their own occupation through the right way and give full play to it Important role.

The main purpose of this paper is to integrate the historical data of technical talents from all over the world, select suitable candidates and deal with qualified personnel in combination with enterprise recruitment information. Bayesian algorithm is used to classify a large number of data, and the classification model of technical talents is updated in real time. Finally, the vacancy of each enterprise post recruitment is implemented, and the data classification efficiency of the whole data mining system is strengthened. Finally, the system is applied to the enterprise talent recruitment network, and the test results show that the system can effectively classify skilled talents and fill the vacancies of enterprises. The majority of the people who are qualified for the post are selected, which provides a new technology-based talent acquisition system for major enterprises.

2. Establishment of Data Mining System Model for Skilled Talents based on Big Data

2.1 Data Mining Technology

Data mining technology is to search data in the system. In other words, it automatically selects the most valuable information from a large database. If you apply it to a business model, you can interpret it like this. The company locks in the company's project objectives, then searches in the company's large amount of data, finds the undiscovered or discovered rules in the data, so as to get the method to achieve the company's goals.

2.2 The Establishment Process of Bayesian Algorithm Model

Bayesian algorithm model realizes the integration of prior data and sample data. Namely:

\[ (I_t) \text{Posterior data} = (I_0) \text{Prior data} = (S) \text{Sample data} \]  

(1)

The data mining system of skilled talents will use Bayesian algorithm to complete the index extraction. NB algorithm has the advantages of high precision and high efficiency and has been widely used. Through the algorithm research, we can see that NB algorithm can even compare with neural network algorithm and decision tree algorithm. NB algorithm has been widely used in medical diagnosis, text processing and performance management because of its high efficiency, simplicity and accuracy. However, developers will choose naive Bayesian algorithm to complete the construction of large database, and in NB algorithm, if the category is clear, the attributes will remain independent of each other. Then the formula of NB algorithm is as follows:

\[ P(C_i|X) = \frac{P(X|C_i)P(C_i)}{P(X)} \]  

(2)

The system can query the data \( P(x_1|C_i), P(x_2|C_2), \ldots, P(x_n|C_i) \) probability value to evaluate the sample value. This is because there is too much data. If we don't find a shortcut to set up the independent relationship between various skilled talents, it is difficult to complete the accurate classification of data.
The system can also solve the $C_i$ of $P(X|C_i)P(C_i)$, and calculate the value of $X$. If the system needs to use $C_i$ for classification processing, at this time:

$$P(X|C_i)P(C_i) > P(X|C_j)P(C_j), 1 \leq j \leq m, j \neq i$$

(3)

Thus, in $P(X|C_i)P(C_i)$, the system uses $C_i$ to implement the classification of each skilled talent sample $X$. Based on the classification of skilled talents, major enterprises can comprehensively grasp the position direction of skilled talents, provide scientific guidelines for enterprise decision-making, ensure the accuracy of professional counterparts, and promote the expansion and recommendation of enterprise recruitment.

3. Experiment Idea and Parameter Setting

3.1 Experimental Ideas

In this paper, data mining technology is used to analyze the amount of recruitment information published by Internet recruitment websites. After preprocessing, unstructured text data such as filtering, copying, deleting empty data, Chinese word segmentation and stop word filtering are used to preprocess the recruitment data mining association rules. The emerging work of internal association rules of each index is mining in the field of big data and in-depth analysis of industry distribution, industry skill demand and demand growth trend in the field of big data. This system looks for the recruitment information and demand of skilled talents from the Internet, collects and classifies the recruitment information on the recruitment platform by using Bayesian algorithm, and allocates the corresponding talents to different recruitment companies. Such as the basic information of the enterprise, workplace, education requirements, industry field, professional title, etc. will be displayed. We apply the system to the recruitment network, which is reflected by visiting the enterprises' satisfaction with the skilled talents and the post competency.

3.2 Experimental Design

This paper studies the data mining system and construction analysis of skilled talents, which involves many components of Hadoop ecological family. HDFS has high fault tolerance and high throughput, which lays a solid foundation for big data processing. As a non-relational database, HBase is different from the basic data type of previous relational database. The information stored in the original database is stored in a two-dimensional table in the form of key value pairs. The system is divided into resource layer, perception layer, service layer, portal layer and application layer. The system architecture can cover all concepts and applications involving big data and build a service-oriented system architecture supporting the whole life cycle activities of big data engineering. Through data mining, statistical analysis, machine learning and related computing technology, the analysis results are obtained. The parameter values used in this paper are shown as Table 1.

| Parameter setting | Numerical value |
|-------------------|-----------------|
| Computing open source framework | Hadoop |
| distributed file system | HDFS |
| Large scale data processing | MapReduce |
| Database | Hbase |
| Management tool | Zookeeper |
| Data extraction tools | Hive |
4. Discussion

4.1 Evaluation of the System by Enterprise Personnel

![Figure 1: Survey on the satisfaction of enterprise personnel to the system](image)

Shown as Figure 1, we applied the established system to the recruitment network. The time span of this paper to collect recruitment information is three months, and the amount of data collected is 300000, including the data related to the classification data of skilled talents. Among them, there are basic information of the enterprise and personal information of candidates, which are distributed and stored in different storage structures. The processed information bar is more effective, and the quality of recruiters received by enterprises has been improved as a whole. We conducted a questionnaire survey on many enterprise staff and found that they agreed with the data collection efficiency of the system, and the overall satisfaction increased to 85%. The collected and classified information is not only true and effective, but also greatly speeds up the recruitment plan of skilled talents, screening and deleting many unqualified candidates, or putting them into other qualified recruitment information.

4.2 Position Information Collection Results of Energy Saving Talent Mining System based on Big Data
Figure 2 The system summarizes the relevant position information

Shown as Figure 2, the first-tier cities, the second-tier cities, the third-tier cities and the fourth-tier cities are respectively represented by a, B, C and D. In addition, the years of working experience are also included. We mainly classify these four types of cities and make statistics on the demand for skilled talents in different level cities, as well as the number of years of work experience needed for recruitment. The results show that the demand for skilled talents is different in cities with different levels, and the number of people in the first-tier cities is the largest. Second, the second-tier cities, by analogy, the demand of backward cities is the least. Different cities have different demand for skilled talents, and the demand for working years is basically the same. All major enterprises need candidates with working experience of at least one year. It can be clearly seen from the data that each year, the demand for skilled talents in different cities makes many job seekers have a clear direction and the cognition of post information, which plays a very important role in grasping the direction of their own development.

4.3 System Construction Analysis

The overall architecture of the system is divided into six layers: data support layer, data preprocessing layer, data storage layer, data analysis layer, data display layer and enterprise management. The design of the overall architecture is based on the recruitment needs of enterprises, the method of big data analysis and processing, and the improvement of enterprise development. The details are as follows:

1) Enterprise management is mainly carried out in three directions of personnel training, mainly to strengthen the three aspects of moral, intellectual and physical training of skilled talents. Among them, we should focus on the cultivation of professional quality, life management and political ideology, and vigorously develop the skills of skilled talents in other aspects, which can quickly make benefits for the company and eliminate all adverse effects.

2) The data display layer shows the results of data analysis in the form of website. The website of student behavior big data analysis system is written in PHP language, and the analysis results are displayed in the form of webpage, which is convenient to establish contact with enterprise recruitment business system.

3) The data analysis layer is mainly used for the underlying data mining and data analysis, such as MAPE data mining and business analysis.

4) Data storage layer is mainly divided into structured data and unstructured data. Structured data
is stored in MySQL database of NTFS file system, and unstructured data is stored in HBase non-relational database of HDFS file system. Data is extracted by hive tool.

5) The main task of data preprocessing layer is to clean up "dirty" data, including duplicate data, abnormal data and default data. The method to clean up abnormal data is box chart.

6) Data support layer is the basis of big data acquisition. This layer is mainly responsible for the interface connection with the major recruitment business systems of the enterprise and collecting the data of skilled talents into the big data analysis system.

Through the above-mentioned six-tier architecture, big data mining is carried out on skilled talents data, and big data analysis system is used to process big data of resume of skilled talents. The following difficulties exist in the processing process:

1) The sources of data collection include many web pages. The data information of some web pages includes the personal privacy of skilled talents. Therefore, it may be the communication between relevant departments and skilled talents to ensure that the required data are collected on a legal basis to ensure the personal privacy and security of skilled talents.

2) The data collected are semi-structured data and unstructured data, and their structures are quite different. A large number of data greatly tested the bearing pressure of the equipment.

3) Whether it is necessary to analyze and monitor the individual's resume information, so that everyone can dynamically monitor their own resume information, and can change the information at any time, but it is necessary to ensure the authenticity of the information and bear the consequences of untrue information.

5.Conclusions
This paper uses advanced big data technology and data mining technology to build a skilled talent identification system and classifies and stores the input resume information by Bayesian algorithm. At the same time, the information of skilled talents needed in the recruitment information is matched with the stored resume information, which makes the recruitment channels of enterprises more extensive and the recruitment efficiency more effective. Through the experimental test, the system has a strong reliability, which can accurately detect the demand for different talents in different cities of different levels and the demand conditions, so as to realize the accurate classification and release of resumes. According to the questionnaire survey of enterprise personnel, 85% of the staff feel that the system accelerates their recruitment task and the information authenticity of personnel meeting the requirements is extremely high. Let skilled talents play their instincts in the right positions and speed up the development of enterprises.

Acknowledgements
Department of Education, Guangdong Province Youth Innovation Talent Project (Philosophy, social sciences) (2020WQNCX094)

References
[1] Wierik, M. L. J. T. , Beishuizen, J. , & Os, W. V. . (2015). Career guidance and student success in dutch higher vocational education. Studies in Higher Education, 40(10), 1947-1961.
[2] Jr, D. B. E. , & Loucel, C. . (2016). The educo program, impact evaluations, and the political economy of global education reform. Education Policy Analysis Archives, 24(92), 1-50.
[3] Fu, T. T. , Sung, A. D. , See, K. F. , & Chou, K. W. . (2019). Do optimal scale and efficiency matter in taiwan's higher education reform? a stochastic cost frontier approach. Socio Economic Planning Sciences, 67(SEP.), 111-119.
[4] Team, E. E. . (2016). Job vacancy at the european centre for disease prevention and control. Eurosurveillance: bulletin europeen sur les maladies transmissibles = European communicable disease bulletin, 23(1), 2–6.
[5] Cascio, W. F. , & Boudreau, J. W. . (2016). The search for global competence: from
international hr to talent management. Journal of World Business, 51(1), 103-114.

[6] Sparrow, P. R. , & Makram, H. . (2015). What is the value of talent management? building value-driven processes within a talent management architecture. Human Resource Management Review, 25(3), 249-263.

[7] Hashem, I. A. T. , Yaqoob, I. , Anuar, N. B. , Mokhtar, S. , Gani, A. , & Khan, S. U. . (2015). The rise of "big data" on cloud computing: review and open research issues. Information Systems, 47(jan.), 98-115.

[8] Buczak, A. , & Guven, E. . (2017). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153-1176.

[9] Jing, S. , Pan, T. , & Li, Z. . (2016). Recursive bayesian algorithm with covariance resetting for identification of oear models with non-uniformly sampled input data. Circuits Systems & Signal Processing, 35(3), 919-932.

[10] Xunzhang, G. , Xiang, L. I. , Filos, J. , & Wei, D. . (2015). A sequential bayesian algorithm for doa tracking in time-varying environments. Chinese Journal of Electronics, 24(001), 140-145.