Research Article

Visualization Research of College Students’ Career Planning Paths Integrating Deep Learning and Big Data

Jing Guo and Lei Qi

1Students’ Affairs Division, Hebei Agricultural University, Baoding 071001, Hebei, China
2Faculty of Built Environment, University of Malaya, Kuala Lumpur 50603, Malaysia

Correspondence should be addressed to Jing Guo; nxguojing@hebau.edu.cn

Received 27 February 2022; Revised 16 March 2022; Accepted 19 March 2022; Published 12 April 2022

Abstract

As China’s education enters a high-level stage, more and more students graduate from Chinese colleges and universities. In particular, the current employment environment is flexible and multilateral, and there are more and more opportunities to choose from. In view of this situation, this article aims to visualize the career planning (CP) path of college students, so as to help college students adapt to the environment of flexible employment. For deep learning and big data (DLBA) technology, this article proposes the LSTM-Canopy algorithm, which is added to the traditional Canopy algorithm to enhance the self-learning clustering ability of the algorithm. Also, this study applies this algorithm to the visualization system of college students’ CP path, which can effectively improve the analysis and judgment of experts on career. The experiments in this article have proved that the system can meet the normal use of 400–500 users, and the system server has successfully passed 40 load tests, and the running time is also less than 2.5s, which proves the reliability of the system.

1. Introduction

The current era is the era of the Internet. In recent years, computer technology and network technology have developed rapidly, and social informatization has also developed into a new stage. Some people say that the current era is the era of information, but it is more accurate to say that it is the era of data.

The number of students continues to rise, and new types of information continue to emerge, and the raw student data on college campuses has grown exponentially. These raw data hold enormous value. In the face of massive data, if we can use scientific methods to process it? "Take the essence and get rid of the dross," extracting the information we are interested in hidden in the data can make the data a real resource that can be used by us.

Of course, as more and more information accumulates without knowing how to use it. Especially, in the CP of college students, there is a lot of information, but they do not know how to use it, and they do not know their direction. Therefore, it is necessary to design a visualization system of college students’ CP paths for this problem.

For the analysis of college CP path, this article has the following two innovations: (1) For the problem of data classification, based on DLBA technology, this article proposes the LSTM-Canopy algorithm, which is a data clustering algorithm based on the LSTM algorithm. It adds LSTM self-learning factor to the traditional Canopy algorithm to make the clustering effect better. (2) This article designs a visualization system of college students’ CP path and designs an expert system for college students’ CP. Experienced teachers and entrepreneurs analyze the career reports submitted by students and present their own analysis reports to facilitate data visualization.

2. Related Work

College students are the main reserve force for the construction of the motherland, and there are many studies on the careers of college students. Tavabie and Simms explored common abilities and characteristics of nonclinical roles. They organized this characteristic information into a generic job description at four key levels. These form the basis of a career path [1]. Jackson believed that in the case of intense competition in the graduate
labor market and underemployment of graduates, effective CP for college students is becoming more and more important. His research examined the impact of work-study integration on students’ CP. Also, his research deepens our understanding of how work-study integration shapes college students’ career goals and improves current weak levels of student engagement in CP, discussing the implications for future career counseling [2]. Nordin and Hong aimed to explore the impact of career coaching activities on children’s CP. Their study was conducted on 12-year olds who were having problems with the CP process. The findings showed that four topics were successfully identified, namely, occupational understanding, sources of occupational information disclosure, occupational choice, and employment understanding through parental occupations [3]. Hung and others investigated the motivations that led Vietnamese students to choose to study in Taiwan. Quantitative research results showed that there is a significant correlation between student motivation and CP, and both directly affect decision making [4]. Ying et al. provided a comprehensive overview of research using deep learning methods to process clinical data. They believed that despite the challenges of applying DL technology to clinical data, the prospect of DL application in clinical big data in precision medicine is still worth looking forward to [5]. Deep learning (DL) is a branch of machine learning techniques based on algorithms that learn multilevel representations. Big data analytics (BDA) is the process of examining large-scale data and various data types. Hordri et al. identified existing features of DL methods used in BDA and identified key features that affect the effectiveness of DL methods. It is experimentally demonstrated that DL for BDA is an active research area [6]. Duncan et al. aimed to define and highlight some of the “hot” new perspectives in the field of biomedical imaging and analysis. They aimed to shed light on where the field is headed over the next few decades and highlighted areas where electrical engineers are already involved and are likely to have the greatest impact. They have a good discussion about medical imaging being “big data.” Also, they believed that the development of biomedical imaging and analysis is very good [7]. However, after relevant research, it is found that the CP for college students is reflected in the course design and teaching experience, and there is little research on the visualization of the system.

3. DL Big Data Algorithms

3.1. Neural Network Algorithm. The overall structure of a long short-term memory network (LSTM) is similar to that of a recurrent neural network (RNN). The “gate” mechanism and the structural concept of cell state are introduced into the LSTM hidden layer calculation. The “gate” mechanism determines how much information is input at each time step, how much state information is saved and forgotten, and the cell state is responsible for recording the state information at the current time step. Through these two structures, the earlier information of the sequence can be preserved, and the learning of long-distance dependence of sequence features can be realized [8, 9]. Also, due to the existence of the cell state, the gradient can be saved during training to solve the problem of gradient disappearance [10]. The hidden layer computing structure of LSTM is shown in Figure 1.

The calculation formula of the hidden layer at the time \( t \) of LSTM is as follows, including the forgetting gate \( f_t \), the input gate \( i_t \), the output gate \( o_t \), the cell state update \( c_t \), and the calculation of the hidden layer output \( h_t \).

\[
\begin{align*}
    f_t &= \sigma(W_f [h_{t-1}, X_t] + b_f), \\
    i_t &= \sigma(W_i [h_{t-1}, X_t] + b_i), \\
    o_t &= \sigma(W_o [h_{t-1}, X_t] + b_o), \\
    c_t &= \tanh(W_c [h_{t-1}, X_t] + b_c), \\
    h_t &= o_t \cdot \tanh(c_t),
\end{align*}
\]

where

\[
\sigma(x) = \frac{1}{1 + \exp(-x)}
\]

Formula (2) is the sigmoid function, \( \cdot \) is the element-wise multiplication of vectors, and \( W \) and \( b \) are the weights and biases.

Position embedding is a matrix that represents time-step information in the same shape as the input feature. It can use trainable variables as position embedding, or it can be a custom matrix that can represent the difference between time steps. A common method is shown in Formulas (3) and (4).

\[
\begin{align*}
    \text{PE}_{\text{pos},2i} &= \sin \left( \frac{\text{pos}}{10000^{2i/n}} \right), \\
    \text{PE}_{\text{pos},2i+1} &= \cos \left( \frac{\text{pos}}{10000^{2i/n}} \right),
\end{align*}
\]

where \( \text{PE} \) represents the position-embedding matrix, \( \text{pos} \) represents the index of the time step dimension of the PE matrix, \( i \) represents the index of the feature dimension of the PE matrix, and \( n \) represents the total number of features.

Batch normalization is an optimization algorithm that normalizes the outputs of neurons in the same position in a training batch. The algorithm makes the distribution of the output of each layer of the neural network more stable and reduces the coupling between the layers, thereby accelerating the convergence of the neural network algorithm. Assuming that the value of a neuron on a certain minibatch is \( \{x_1, x_2, \ldots, x_m\} \), the batch-normalization calculation method is shown in the following formula.

\[
\begin{align*}
    \mu &= \frac{1}{m} \sum_{i=1}^{m} x_i, \\
    \sigma^2 &= \frac{1}{m} \sum_{i=1}^{m} (x_i - \mu)^2, \\
    \bar{x}_i &= \frac{x_i - \mu}{\sqrt{\sigma^2 + \epsilon}}, \\
    \text{BN}_{\gamma, \beta}(x_i) &= \gamma \bar{x}_i + \beta,
\end{align*}
\]
where $y$ and $\beta$ are trainable variables and $\epsilon$ is a positive epsilon.

Dropout randomly discards some neurons on the output of a layer of neurons and then scales up the training method of the signal, so that the weight does not completely depend on certain features. It achieves the purpose of reducing the degree of network overfitting and improving the generalization ability of the algorithm. The calculation method of dropout is as follows:

$$
\begin{align*}
    r_j^{(l-1)} & \sim \text{Bernoulli}(p), \\
    \bar{y}_j^{(l-1)} & = x^{(l-1)} \circ y_j^{(l-1)}, \\
    z_j^{(l)} & = \frac{W_j^{(l)} \bar{y}_j^{(l-1)} + b_j^{(l)}}{\epsilon}, \\
    y_j^{(l)} & = \sigma(z_j^{(l)}),
\end{align*}
$$

where $z_j^{(l)}$ represents the unactivated feature output of the $l$th layer of the fully connected neural network and $y_j^{(l)}$ represents the activated feature output of the $l$th layer of the fully connected neural network. $r(l)$ is the vector of the Bernoulli distribution with probability $p$ for the elements of the neural network in the $l$th layer, and $W^{(l)}$ represents the weight matrix of the $l$th layer of the fully connected neural network. $b(l)$ represents the bias vector of the $l$th layer of the fully connected neural network. $\sigma(x) = 1/(1 + \exp(-x))$ is the sigmoid function, and $\circ$ is the elementwise multiplication of vectors [11]. The Canopy algorithm performs rough clustering without preselecting the number of clusters $k$, and at the same time, it greatly optimizes the method for determining the initial cluster center.

### 3.2. Canopy Algorithm

Before introducing the Canopy algorithm, first define several concepts of the Canopy algorithm:

Suppose $D = \{d_1, d_2, \ldots, d_n\}$ is a data set with $n$ data elements, for any $d_i \in D$ if

$$
\{C_k \mid \|C_k - d_i\| \leq T_2, T_2 < T_1, C_k \in D, i \neq k\},
$$

then the set $C_k$ is called the non-Canopy candidate center point set.

If any

$$
\{C_j \mid \|C_j - d_i\| \leq T_1, C_j \in D, i \neq j\},
$$

then $d_i$ is defined as a Canopy, $T_1$ is called the set radius of Canopy, and $C_j$ is the center point of Canopy.

The basic idea of the Canopy algorithm can be divided into two stages: The first step is to use the roughly calculated distance as a metric. Efficiently divide the data in the data set into several subsets. Different subsets can intersect but cannot completely overlap, and each divided subset is called a Canopy. In the second step, a more rigorous clustering algorithm is selected to perform clustering operations on the data in Canopy obtained in the first step. The Canopy algorithm is a clustering strategy of thick and thin sets, and it is very suitable for preanalysis of high-dimensional data [12, 13]. The Canopy clustering process is shown in Figure 2.

The first step of the algorithm generates several Canopy, and each Canopy is a collection of sample data. The algorithm presets a threshold for the measure of difference [14].

Because the first step of the algorithm allows two Canopy to intersect, a data object may belong to more than one Canopy, but each data object should exist in at least one Canopy. The second step of the algorithm can be done using conventional clustering algorithms like $k$-means, but it is worth noting. The clustering work in the second step is performed on the data in the same Canopy, and there is no need to calculate the data in different Canopy. It is generally considered that the distance between the data between different Canopies is infinite. Consider an extreme case, if all data objects are divided into the same Canopy in the second
step of the algorithm, then the work of the second step degenerates into a traditional clustering algorithm.

The Canopy algorithm does not need to preset the $k$ value for the number of clusters but uses two similarity measures $T_1$ and $T_2$ to indirectly determine the number of Canopy subsets after clustering and the number of data objects in each Canopy [15, 16]. $T_1$ is called a loose threshold (loose distance), $T_2$ is called a tight threshold (tight distance), and $T_1 > T_2$. If the distance between a data object and a Canopy center is less than $T_1$, then the data object is added to the Canopy; if the distance is less than $T_2$, the data object will not become a Canopy center. The relationship between the discrimination of data objects and $T_1$ and $T_2$ is shown in Figure 3.

The Canopy algorithm does not need to select the number of clusters, which saves the cost of the algorithm. Bringing the results of the Canopy algorithm into the next precise clustering algorithm can greatly reduce the number of iterations of the clustering algorithm, thereby improving the efficiency and accuracy of the clustering. It can also reduce the probability of local optimal solutions to a certain extent [17, 18]. Canopy’s algorithm mechanism also determines that it has better performance for high-dimensional data processing [10].

However, the Canopy algorithm is not perfect. It is obvious that the selection of $T_1$ and $T_2$ when executing the algorithm has a great impact on the performance of the algorithm. If the selection of the $k$ value is the core of the $k$-means algorithm, then the determination of the $T_1$ and $T_2$ thresholds is the key to the Canopy algorithm. If the selected loose threshold $T_1$ is too large, then a data object may be included in many different Canopy, and the computational cost of the algorithm in the second step will be greatly increased. If the selected tight threshold $T_2$ is too large, the number of clusters obtained will be very small. The accuracy of the algorithm in the initial Canopy center selection will be greatly reduced, and it may even fall into the trap of local optimality, thereby reducing the accuracy of the clustering results.

3.3. Canopy Algorithm Based on Neural Network. Through the analysis and summary of the Canopy algorithm, this section proposes an idea to improve the Canopy algorithm. The Canopy algorithm can be used to obtain relatively rough clusters. The results of the Canopy algorithm are added to the LSTM algorithm for improvement, and then, the $k$-means algorithm iteration is performed to complete the final clustering of the data [19, 20]. From the analysis of the algorithm principle, the improved new algorithm first avoids the manual selection of the number of clusters $k$, and then, the preprocessing of Canopy will greatly reduce the burden of the $k$-means algorithm, thereby improving the accuracy and operating efficiency of the algorithm. The next section will test the performance of the improved algorithm through specific experiments.

In this article, the improved algorithm is named LSTM-Canopy algorithm, and the specific process of the algorithm is as follows:
3.4. Algorithm Performance Analysis. In this section, the Canopy algorithm and the improved LSTM_Canopy algorithm are used to compare the two algorithms with multiple data sets, focusing on the improvement of the performance of the LSTM_Canopy algorithm to the traditional clustering algorithm.

The University of California has created a database dedicated to scientific research, called the Machine Learning Standards Database (UCIMachineLearningRepository). The data in this database is widely used in the field of artificial intelligence research. This test selected three data sets in the UCI database for testing.

The Seeds data set contains 210 pieces of data, the data dimension is 7, and the standard number of clusters is 3, which is recorded as data set DS1 in this test. The Glass data set contains 214 pieces of data, the data dimension is 10, and the standard number of clusters is 7, which is recorded as data set DS2. The CarEvolution data set has 1728 data, the data dimension is 6, and the standard number of clusters is 4, which is recorded as data set DS3. The parameters of each data set are summarized in Table 1.

Download these data sets from the UCI database website, perform data processing (for example, the Seeds data set needs to be normalized), and then import them into MATLAB software for simulation. The simulation of two-dimensional data clustering is shown in Figure 4.

There are two standards to measure the accuracy of the algorithm in this experiment: the accuracy rate $P$ and the minimum error squared sum $E$.

Assuming that the data set contains $n$ data objects, after the clustering algorithm ends, the data objects in the data set are clustered into $k$ classes. $d_i$ is the number of data objects that are correctly assigned to the $i$th cluster (compared with the standard clustering of the UCI data set to judge whether it is correct), then the accuracy is defined as

$$P = \frac{\sum_{i=1}^{k} d_i}{n} \times 100\%.$$  \hfill (9)

Let $s$ be the number of clusters and $n_k$ the number of data objects in the $k$th cluster. $c_k$ represents the cluster center of the $k$th cluster, $d^k_i$ represents the $i$th data object in the $k$th cluster, and the error sum of squares is defined as

$$E = \sum_{k=1}^{s} \sum_{i=1}^{n_k} |d^k_i - c_k|^2.$$  \hfill (10)

The smallest sum of squares of errors $E_{\text{min}}$ is the smallest of the sums of squares of errors after the algorithm ends. Three sets of parameters with different values \{ $k = 3, T1 = 9, T2 = 6+$ \}, \{ $k = 3, T1 = 10, T2 = 7$ \}, and \{ $k = 4, T1 = 8, T2 = 6$ \} were selected to perform ten clustering simulation experiments on DS1, DS2, and DS3 data sets using $k$-means algorithm and LSTM_Canopy algorithm. Then, calculate the average of the accuracy of ten clustering results and the minimum error sum of squares, the results are shown in Table 2 and Figures 5 and 6.

It can be seen from Table 2 and Figure 5 that the LSTM_Canopy algorithm has different degrees of improvement in accuracy compared with the traditional $k$-means algorithm. When the amount of data in the data set is large, the dimension is high, and there are many types; the accuracy of the $k$-means algorithm is greatly reduced. However, the LSTM_Canopy algorithm can greatly improve the clustering accuracy in the clustering of high-dimensional...
Input: dataset $D$ to be clustered
Algorithm flow:
1. Set the thresholds $T_1$ and $T_2$ as the area radius ($T_1 > T_2$)
2. Calculate the distance $l$ between other data units in the dataset $D$ and the object $d$;
3. If $l < T_1$, this data is marked with weak correlation, and this data is divided into Canopy; If $l < T_2$, this data mark is strongly associated, and this data object is deleted from the data set $D$;
4. Repeat steps (2) (3) (4) until the dataset $D$ is empty;
5. For the obtained series of Canopy centers and the Canopy centered on them, each Canopy is regarded as a cluster, and the center of Canopy is regarded as the center of the cluster, and the number of clusters is recorded as $k$.
6. Calculate the distances from other data objects to the $k$ centers, and assign these objects to the clusters closest to a certain cluster center;
7. According to the data allocated to $k$ clusters, update the center point of the cluster;
8. Repeat steps (6) and (7) until the clusters no longer change.
Output: set of $k$ clusters.

**Algorithm 1: LSTM-Canopy algorithm.**

**Table 1: Test data set parameter table.**

| Dataset name | Number of objects | Data dimension | Quantity classification |
|--------------|-------------------|----------------|------------------------|
| DS1          | 210               | 7              | 3                      |
| DS2          | 214               | 10             | 7                      |
| DS3          | 1728              | 6              | 4                      |

![Simulation diagram of two-dimensional data clustering.](image)

**Table 2: The clustering accuracy results of the two algorithms.**

| Data set name | $k$-means algorithm | LSTM_Canopy algorithm |
|---------------|----------------------|-----------------------|
| DS1           | 80.70%               | 83.20%                |
| DS2           | 27.30%               | 48.70%                |
| DS3           | 78.50%               | 84.30%                |
data and massive data. The minimum sum of squared errors of the LSTM_Canopy algorithm is also much lower than that of the k-means algorithm, and its clustering accuracy and stability are also improved, which proves that the data similarity is higher in the clusters obtained using the improved algorithm.

It can be seen from Figure 6 that the accuracy of the LSTM_Canopy algorithm is always higher than that of the k-means algorithm, and its curve is significantly smoother, which proves that the improved algorithm has higher clustering accuracy and algorithm stability, which further verifies the previous results.

4. Visualization System Design and Realization of College Students’ CP Path

The core task of the college CP expert system based on data mining is to complete the CP work for college students with the idea of an expert system. The college CP of this subject is divided into academic planning and employment guidance.
for college students, among which employment guidance is the core content of the study. In order to provide correct employment guidance for college students, it is necessary to mine the data of college students and design the reasoning mechanism of expert system based on the knowledge excavated. Therefore, the system should not only have the knowledge acquisition means and reasoning mechanism of the expert system but also have the functions of data processing and management, which can complete the work of the simple management information system.

The functions of the college CP expert system based on data mining mainly include two parts: the foreground function and the background function: the foreground function is the business function of the system, that is, the service function that the system can provide when the user uses the system. The background function is the management function, that is, the system administrators at all levels manage and maintain the system according to their own authority. This section will analyze the requirements of the system from the perspective of the system’s foreground business functions and background management functions.

4.1. Overall System Design. This chapter designs the system workflow with the idea of modularization according to the requirement analysis. The workflow takes into account the functional requirements and performance requirements of the system and provides the basis for the realization of the system. The overall framework of the system is shown in Figure 7.

4.2. System Function Design. The core of the expert system is the knowledge base and the inference engine. The knowledge base stores a large amount of professional knowledge and experience in a certain field provided by domain experts. The reasoning engine uses computer thinking to simulate the thinking process of human experts based on the knowledge in the system and the demands of expert system users and answers the user’s questions instead of them. The design of the knowledge base and inference engine determines the performance of the expert system. The construction of the knowledge base and inference engine determines the performance of the expert system. The selection of the algorithm are the keys to the design of the expert system. The framework of the expert system for college CP based on data mining is shown in Figure 8.

The core function of the college career planning system based on data mining is the student’s academic planning and employment guidance. Click the Career Planning tab at the top of the page to switch to the career planning page. Select the academic planning function on the page to view the professional courses recommended by the expert system for students, which can be a reference for students when choosing courses. When the student user clicks to enter the academic planning page, the front end of the web page obtains the student’s id and sends a post request to the server. The server first obtains the institute and major fields in the student table according to the id, so as to determine the student’s college information and major information. Then, the server obtains the course selection information of the student’s seniors in the same major from the system database and selects a course with a relatively high degree of selection. It renders the course data to the Career_Plan.html web page in templates and returns it, so as to realize the academic planning function for the students in school. The realization of the function of the college CP system is shown in Figure 9.

4.3. System Data Layer Design. The data layer of the system stores all the data and information in the system, and the data layer is invisible to ordinary users of the system, only visible to those who design and maintain the system, which ensures the security of the system. The expert system of college CP is based on data mining, and the data layer includes original data, knowledge data, and system data.

Raw data are raw data without data preprocessing. System data are the data generated and used by the users of the system in the system and are stored using the MySQL relational database. Knowledge data are the key data for the operation of the expert system, including the knowledge stored in the knowledge base and the reasoning logic of the inference engine. The knowledge data are formed by the preprocessing and data mining of the original data of the students. It is huge in quantity and has high value, so it needs to be stored in the backup database in the form of text documents. Such data isolation operation not only ensures the security of the data but also facilitates the research of the data by domain experts and system knowledge engineers. Its data types are shown in Table 3.

4.4. System Test. After completing the design and implementation of the system, this chapter tests the system. The system test includes the test of each functional module of the system and the performance test of the system to verify whether the function of the system is complete and whether the performance can meet the needs of multiple students using at the same time.

4.4.1. Server Load Test. Load testing is an important component of performance testing. The purpose is to test whether the system can operate under overload conditions that are close to the upper limit or even exceed the upper limit. Load testing can help testers examine the stability of the system, determine the volume of system users, and make targeted improvements to the system.

When load testing the server of this system, increase the number of user terminal computers connected to the server in batches. From 10 to 40, the testers use various functions of the system on the terminal computer and record the average response time of the system when the testers use it as the performance index of the server load test. In this test, the ratio of students and managers to test is 4:1. When testers use various front-end functions of the system, the response time of the system (accurate to 0.1 seconds) is shown in Figure 10.
As shown in Figure 10, the system showed relatively good performance in the load test. When the load reached 40 units, the response time of the system remained below 5 s.

4.4.2. Server Stress Test. Stress testing is a common means of system testing, and it is a destructive performance test. Through the stress test of the server, the stability of the system under high load for a long time can be tested, and it can also be applied to the system by applying an excessive load. It crashes the system and exposes hidden problems early and then makes targeted improvements to the system.

Apache JMeter is a very popular testing tool. It is an open-source software with a graphical interface. It is widely used in the performance testing of simulated network systems, servers, and other systems. This test uses the latest JMeter 5.1 version. The pressure test results are shown in Table 4.

As shown in Table 4, the performance of the system varies with the number of users. Specifically, as the number of users increases, the error rate also increases.

Throughput refers to the number of all requests processed by the server in one test, while throughput is the ratio of throughput to time, reflecting the ability of the server to process requests per unit time. As the number of simulated
Table 3: Message information data table.

| Data field   | Data type | Data length | Data information       |
|--------------|-----------|-------------|------------------------|
| Id           | Int       | 16          | Principal linkage      |
| Title        | Varchar   | 64          | Information title      |
| Content      | Text      | 500         | Message text           |
| Pubtime      | Datetime  | 8           | Release time           |
| Publisher    | Char      | 32          | Promulgator            |
| Note         | Varchar   | 128         | Remarks                |

Figure 9: Functional diagram of college CP.

Figure 10: The average response time of each function of the system function.
users increases, the server throughput continues to decrease. When the number of simulated users reaches 500, the decreasing trend of throughput slows down and finally stabilizes at around 800 req/s. Combined with the previous error rate analysis, it can be concluded that the system can meet the normal use of 400–500 users when executing 1000 cycles of server request operations. As far as the network environment used in the current test is concerned, the server of this system has successfully passed the test task.

5. Conclusion

This article studies the related technologies of college CP expert system based on DLBA and designs and implements the system. The whole system includes the front-end business system used by students, the back-end management system used by managers at all levels, and the expert system that provides students with CP. This article uses a large number of real student data and adopts the data mining algorithm combining cluster analysis and classification analysis to complete the establishment of the knowledge module and reasoning mechanism of the expert system, thus realizing the core function of the system. Finally, this article tests the system by professional means. The test results show that the system has achieved the expected requirements in function and performance. However, due to the limitation of the author’s time and personal ability, the front-end web page of the system only realizes the basic functions required by the system, and the web page design is not beautiful. It does not implement the system on the mobile terminal, and the user can only access the system through the web page, which is not convenient enough to use. Therefore, follow-up research will be carried out in this way.

Data Availability

The data used to support the findings of this study are included in the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was funded in part by the Ministry of Education university-industry collaboration education Project number: 202102647006.

Table 4: Server stress test.

| Number of threads | Total request count | Average sound should do (ms) | median (ms) | 90% line (ms) | Lowest response (ms) | Maximum response (ms) | Error rate | Throughput (RPS) | Throughput test time (ms) |
|-------------------|---------------------|-------------------------------|------------|--------------|----------------------|-----------------------|------------|-----------------|--------------------------|
| 100               | 100000              | 102                           | 74         | 219          | 1                    | 672                   | 0.00%      | 1639            | 61000                    |
| 200               | 200000              | 227                           | 135        | 364          | 0                    | 1055                  | 0.00%      | 1163            | 172000                   |
| 300               | 300000              | 352                           | 202        | 588          | 0                    | 1677                  | 0.02%      | 1003            | 299000                   |
| 400               | 400000              | 478                           | 244        | 976          | 1                    | 1969                  | 0.17%      | 913             | 438000                   |
| 500               | 500000              | 618                           | 295        | 1362         | 1                    | 2204                  | 1.42%      | 848             | 590000                   |
| 600               | 600000              | 740                           | 436        | 1584         | 2                    | 2380                  | 4.35%      | 829             | 724000                   |
| 700               | 700000              | 867                           | 560        | 1765         | 2                    | 2491                  | 8.78%      | 818             | 856000                   |
| 800               | 800000              | 989                           | 729        | 1848         | 1                    | 2588                  | 13.81%     | 814             | 983000                   |

References

[1] J. A. Tavabie and J. M. Simms, "Career planning for the non-clinical workforce - an opportunity to develop a sustainable workforce in primary care," *Education for Primary Care*, vol. 28, no. 2, pp. 94–101, 2017.
[2] D. A. Jackson, "Using work-integrated learning to enhance career planning among business undergraduates," *Australian Journal of Career Development*, vol. 26, no. 3, pp. 153–164, 2017.
[3] M. Nordin and C. S. Hong, "Exploring children's CP through career guidance activities: a case study," *International Journal of Academic Research in Progressive Education and Development*, vol. 10, no. 2, pp. 754–765, 2021.
[4] N. T. Hung, "The role of motivation and CP in students’ decision-making process for studying abroad: a mixed-methods study," *Revista Argentina de Clínica Psicologica*, vol. 29, no. 4, pp. 252–264, 2020.
[5] Y. Ying, M. Li, L. Liu, Y. Li, and J. Wang, "Clinical big data and DL: Applications, Challenges, and future outlooks," *Big Data Mining and Analytics*, vol. 2, no. 04, pp. 288–305, 2019.
[6] N. F. Hordri, A. Samar, S. S. Yuhnaz, and S. M. Shamsuddin, "A systematic literature review on features of DL in big data analytics," *International Journal of Advances in Soft Computing and Its Applications*, vol. 9, no. 1, pp. 32–49, 2017.
[7] J. S. Duncan, M. F. Insana, and N. Ayache, "Biomedical imaging and analysis in the age of big data and DL [scanning the issue]," *Proceedings of the IEEE*, vol. 108, no. 1, pp. 3–10, 2020.
[8] Y. Yu, M. Li, and L. Liu, "Clinical big data and deep learning: a," *Big Data Mining and Analytics*, vol. 2, no. 4, pp. 288–305, 2019.
[9] W. Zhong, N. Yu, and C. Ai, "Applying big data based DL system to intrusion detection," *Big Data Mining and Analytics*, vol. 3, no. 3, pp. 181–195, 2020.
[10] T. T. Ziniz and C. W. Lin, *Big Data Analysis and Deep Learning Applications Proceedings of the First International Conference on Big Data Analysis and Deep Learning: Proceedings of the First International Conference on Big Data Analysis and Deep Learning*, Springer, NY, USA, pp. 48–57, 2019.
[11] P. Li, Z. Chen, L. T. Yang, J. Gao, Q. Zhang, and M. J. Deen, "An incremental deep convolutional computation model for feature learning on industrial big data," *IEEE Transactions on Industrial Informatics*, vol. 15, no. 3, pp. 1341–1349, 2019.
[12] L. Gi-In and K. Hang-Bong, "A DL-based streetscapes safety score prediction model using environmental context from big data," *Journal of Korea Multimedia Society*, vol. 20, no. 8, pp. 1282–1290, 2017.
[13] T. Yang, G. Yuan, and J. Yan, “Health analysis of footballer using big data and deep learning,” *Scientific Programming*, vol. 2021, Article ID 9608147, 8 pages, 2021.

[14] L. Yin, Y. Zhang, Z. Zhang, Y. Peng, and P. Zhao, “ParaX,” *Proceedings of the VLDB Endowment*, vol. 14, no. 6, pp. 864–877, 2021.

[15] B. Liu, “Text sentiment analysis based on CBOW model and deep learning in big data environment,” *Journal of Ambient Intelligence and Humanized Computing*, vol. 11, no. 2, pp. 451–458, 2020.

[16] C. NiNg and F. You, “Optimization under uncertainty in the era of big data and deep learning: when machine learning meets mathematical programming,” *Computers & Chemical Engineering*, vol. 125, no. JUN.9, pp. 434–448, 2019.

[17] Y. Chen, Z. Lin, Z. Xing, and G. Wang, “DL-based classification of hyperspectral data,” *Ieee Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 7, no. 6, pp. 2094–2107, 2017.

[18] S. Lee, H. Ko, and S. Oh, “Multisensor fusion and integration in the wake of big data, deep learning and cyber physical system,” *Lecture Notes in Electrical Engineering*, Springer, NY, USA, pp. 268–283, 2018.

[19] M. Park and J. Cho, “Effects of the realization of career goals, career planning, and mentoring on career satisfaction of public corporations’ office workers,” *JOURNAL OF SECRETARIAL STUDIES*, vol. 29, no. 2, pp. 31–59, 2020.

[20] T. Catanzano, J. Robbins, and P. Slanetz, “OK boomer: are we oversupporting junior faculty and neglecting career planning for mid and senior rank?” *Journal of the American College of Radiology*, vol. 18, no. 1, pp. 214–218, 2021.