Journal of Combinatorial Mathematics
and Combinatorial Computing
www.combinatorialpress.com/jcmcc

Combinatorial Press

Research on job-seeking employment and career planning of college students based on Internet big data analysis

Guangli Guo[1,✉]

[1] *Shandong Vocational College of Science and Technology, Weifang, Shandong, 261500, China*

ABSTRACT

Under the background of big data era, big data mining technology is widely used, through data mining technology, deeper exploration of data, discovering the relevance of data, can provide decision support for decision makers. This paper analyzes the Internet big data of college students' employment decision-making based on big data mining technology, uses Apriori algorithm to mine the influencing factors of college students' vocational skills generation, meanwhile applies ID3 decision tree algorithm to analyze the college students' tendency of vocational choice, and explores the relevant factors affecting college students' employment through correlation analysis and clustering analysis. The results of the study show that students' personal, family and school have strong correlation with students' vocational skills generation, which affects the improvement of students' personal job-seeking ability. Meanwhile, the ID3 decision tree algorithm is applied to the employment consulting service for graduates to construct a career decision tree for individual college students, which visualizes their career choice paths under the influence of career values and helps them make more appropriate career choices. In addition, qualification certificates, social practice experience, academic performance, expected salary, ideal employment unit and other factors will affect the employment choice of college students, and there are individual differences among different students.

Keywords: Apriori algorithm, ID3 decision tree, big data mining, college students' job search and employment

1. Introduction

With the rapid development of social economy, the employment of college graduates has become the focus of social attention. The current employment of college students generally face challenges such as asymmetric employment information, unclear employer needs, and mismatch between students' own qualities and job requirements. However, the traditional employment guidance model and the current reality of information asymmetry, lack of personalized services and other problems, it is

✉ Corresponding author.
E-mail address: 13306362327@163.com (G. Guo).
Received 10 August 2024; Revised 05 October 2024; Accepted 25 January 2025; Published Online 16 April 2025.
DOI: 10.61091/jcmcc127b-042

difficult to meet the actual needs of graduates. The use of big data technology in the era of big data to optimize the employment guidance work of college graduates, and to explore new employment assistance models for college students to improve the quality and efficiency of college student employment has become one of the urgent problems to be solved at present.

In the employment market, graduates are often unable to understand the actual situation of the employment market and the needs of employers in a timely manner, which leads to an increase in the blindness and uncertainty factors of their job search [1]. On the one hand, the recruitment needs of some employers are not completely transparent, and they may cover up some information in the job advertisements, resulting in job seekers not being able to fully understand the real situation of the position [2-3]. On the other hand, college students' access to employment information mainly relies on the school's career guidance and job fairs and other channels, and the information from these channels is often limited, so that college students lack an accurate grasp of their own professional characteristics and market demand, which leads to insufficient preparation for employment [4-5]. Secondly, as the career choices of college graduates become more and more diversified, the problem of insufficient career matching corresponding to this is also becoming more and more prominent [6]. This phenomenon is not only related to the current education system is too focused on knowledge transfer and lack of practical ability and vocational skills training, but also related to the gradual decline of traditional industries and the rise of new industries [7-8]. As a result, college students are no longer satisfied with the traditional knowledge-based laborers, but are more inclined to have the spirit of innovation, teamwork ability and interdisciplinary knowledge of the composite talents [9-10].

Through the large-scale collection and analysis of data such as employment data and recruitment information in previous years, it can help predict future employment demand and provide more accurate employment guidance and career planning advice for college students, reducing the risks caused by information asymmetry. Li, L. uses TF-IDF as the feature selection method for talent information data and combines it with statistical analysis to conduct quantitative citation analysis, successfully constructing a set of industrial Talent demand characteristics research system, able to mine and analyze the characteristics of industry talent demand for the supply of talent to provide skills training reference, to achieve a dynamic balance between the supply and demand of industrial talent [11]. Saputra, A. et al. built a digital talent analytics system for corporate HR departments based on a big data analytics framework to help talent growth and enhancement through digital empowerment [12]. Wei, J. et al. analyzed the skill needs of job seekers reflected in job postings, and after occupational skill clustering analysis, classified the job postings using the LDA topic model to provide a reference for the future direction of job seekers [13]. Xiao, Q. et al. explored a text categorization method for talent demand under large-scale data and proposed a heuristic KNN text categorization algorithm based on artificial bee colony adjusting feature weights, which can automatically mine and analyze the effective information in massive recruitment data to maintain the balance of supply and demand of talents in education and society [14]. Li, L. et al. examined the application of big data technology in precise employment of college graduates, constructed a hybrid recommendation module for college students' employment based on users' real-time online behavior as well as historical information, and evaluated the current status of talent supply and the current status of talent demand of enterprises in the region from a macroscopic point of view, which provided data support for adjustment of specialization settings in the field of education [15]. Zhang, Q. et al. emphasized that quantitative modeling of dynamic changes in the hiring market is conducive to the timely adjustment of talent demand strategies based on business trends, and proposed a Talent Demand Attention Network (TDAN), which is effective in predicting talent demand at a fine-grained level, and is equally interpretable for modeling hiring trends [16]. These analyses can provide college students with the future direction of the job market and guide them to make more informed career choices.

Big data analytics can improve the success rate of a job search by analyzing an individual's career inclinations and industry preferences, and recommending positions that are more in line with a job

seeker's interests and aspirations. Nie, M. et al. proposed a data-driven career choice framework for college students, which is based on self-perception theory and plays an important role in career counseling and career guidance by capturing and analyzing students' behaviors to make predictions about their career choice options after graduation [17]. Guo, J. et al. developed a visualization system design method for college students' career planning paths and proposed an improved LSTM-Canopy algorithm to enhance the system's self-learning clustering ability, which significantly improves the analytical ability of career counselors and helps college students to choose suitable career planning paths in complex employment environments [18]. Kamal, A. et al. developed an intelligent career guidance system based on Django framework web application, which relies on machine learning techniques and algorithms to help college students to identify specific fields that suit their skills and interests with high reliability [19]. Haritha, D. D. studied a vocational course recommendation system for college students, which utilizes numerous machine learning techniques and algorithms to predict students' areas of interest and course grades from a database of student interest and aptitude information, and combines this with expert advice to recommend appropriate skill courses and certification programs for college students [20]. Yang, T. C. et al. designed a student career decision-making model that integrates institutional data and social media news, and this data-driven intelligent model provides help for college students' future career development and data support for educational administrators to make educational decisions [21]. José-García, A. et al. introduced Educational Artificial Intelligence to provide career guidance services for university students, and the proposed AI solution C3-IoC can link students' technical and non-technical skills, providing intelligent analysis and visualization conditions for self-assessment of students' competencies and future career planning [22].

This paper takes college students' job-seeking employment and career planning as the research objective, and uses big data mining technology to analyze college students' employment decision-making data, in order to provide a decision-making basis for college students' career planning and career choice. First, the correlation between the factors influencing the generation of college students' vocational skills was mined based on the Apriori algorithm. Then, the ID3 decision tree algorithm was applied to study the career choice problem of college students, and a decision tree based on college students' career values was generated. Finally, the correlation analysis of college students' employment influencing factors was carried out, and the sample college students were clustered according to the data of influencing factor indicators, and their differences in career planning and decision-making were analyzed.

2. Data mining-based model of job-seeking employment and career planning for college students

This paper utilizes big data mining technology and Apriori association rule mining algorithm to explore the influencing factors of college students' job-seeking and employment ability and employment results, and realizes the analysis of college students' career decision-making based on decision tree ID3 algorithm, which provides decision-making references for college students' career planning and career choices.

2.1. Data mining

2.1.1. Definition of data mining. Data mining originates from the knowledge discovery in the database called data mining, data mining is a hot area of current research, it is a means of information to find out the valuable, unknown, and potential connections before the data from the huge amount of real data, data containing noise, data with variables, and mixed data [23]. The most central part of data mining is the data, which is stored in databases, and it can be understood that data mining is a large database in which data is analyzed and processed.

2.1.2. Data mining process. The general process of data mining is shown in Figure 1. The process of data mining is connected with several steps, in which the main elements include data selection and acquisition, data preprocessing, data mining, and result evaluation.

1) Data selection and collection: the main task of this stage is to select relevant data from a database or data warehouse to establish a target data set, which is stored in a format that needs to be processed so that the data mining algorithm can access it directly.

2) Data pre-processing: the main purpose of this stage is to make the data incomplete, inconsistent and complex due to the large amount of data. By cleaning up the data, the data set becomes a data set from multiple sets to be stored, the transformation of data format or data planning can improve the effectiveness of the algorithm, data filling, data redundancy, data clustering and other ways to talk about the time of analysis, which aims to improve the efficiency and accuracy of data mining algorithms and reduce the time spent on data processing.

3) Data mining: this stage needs to consider the type, method and efficiency of data mining. It is necessary to decide which data mining algorithm to use, create a model that describes a set of predefined data classes or concepts, select the most appropriate data mining technique to operate by analyzing and classifying the data set, and select the appropriate software to mine the pre-processed data.

4) Evaluation of results: this stage is the main interpretation and evaluation of classification results, this step is mainly used to evaluate the results of the model and explain the value of the model. The result is to explain and evaluate the patterns mined from the data and to remove which inappropriate patterns and rules.

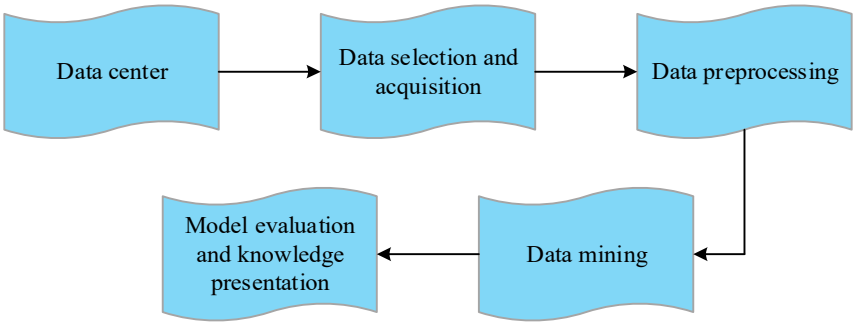

Fig. 1. General process of data mining

2.1.3. Models for data mining. In general, data mining patterns can be broadly categorized into the following four patterns:

1) Clustering mode: the main purpose of this mode is to categorize the data in the data warehouse, according to the category of the data, the similarity of the data, and the difference of the data into multiple classes. So that there is a great degree of recognition between the data of the same category and a great degree of difference between the data of different categories. Clustering mode is generally realized using statistical methods, neural network methods, etc.

2) Generalization mode: also known as data generalization, the main techniques used are data cube technology and data generalization technology. The main task of data cube technology is to store the results of data analysis in advance, so as to facilitate the use of the system. The main task of data generalization technology is to visualize the data for the convenience of users.

3) Classification model: the main task of this model is to find out a suitable feature model for the data collection. Classification mode is mainly classification rules, decision trees and so on.

4) Time series model: its main task is to characterize the values of data attributes in time from the data warehouse, including the trend characteristics of the data series, predictive features, etc.

2.1.4. Data mining techniques. Data mining techniques are mainly categorized in the following ways:

1) Clustering method: it is to follow up the data to analyze, using statistical classification methods, so that the data inside a class has a high correlation, and the data between different classes has a low correlation. There are five clustering methods such as division-based, hierarchical, density-based, grid-based, and model-based.

2) Classification methods, it is in the database of each object or attribute to find out their common characteristics, by reading the test sample data in the dataset to analyze, one or more classification rules, using this rule to build a classification model or classification function, you can use the model to make predictions, and after the establishment of the necessary evaluation. Common classification methods include: case-based reasoning, decision tree, neural network, genetic algorithm, and statistical classification.

3) Association methods, something occurs perhaps intrinsically related to other things, from these things to discover their intrinsic associations, but also to discover the rules between in the data attributes. Association rule algorithms are AIS algorithm, Apriori algorithm, SETM algorithm, DHP algorithm, PARTITION algorithm, Sampling algorithm, FP-growth algorithm and so on.

4) Prediction methods: it is by constructing a model and using the model to make predictions about the data. Methods of prediction: trend extrapolation, time series method, regression analysis, artificial neural network method, expert system method, fuzzy prediction method, wavelet analysis, preferred combination method and so on. The type of method is mostly evaluated by the accuracy of the data model, the time taken by the model to analyze the data, and the complexity of the data set.

2.2. Apriori association rule mining algorithm

2.2.1. Association rules:

1) Basic definition of association rules. Association rules are implicit formulas shaped like $X \Rightarrow Y$ and their probability of occurrence that are mined from large datasets. Association rule mining is used to discover meaningful data connections hidden in large datasets, and has now become an important method that has attracted much attention and research in the field of data mining. Association rule data mining mainly involves some of the following basic concepts:

(1) Items. Let the full set $U = \{item_1, item_2, ..., item_n\}$ be a set of items (e.g., a collection of goods in a supermarket), the elements of which are called items, e.g., instant noodles, beer, and milk in the collection of goods are three different items.

(2) Transaction. The set of items involved in the course of a single transaction. For example, if a customer buys beer, yogurt, and an electric car in the course of a single shopping trip, there exists at least one transaction such as {beer, yogurt, electric car team in the transaction set.

(3) Item set. Any subset X of the full set of items U is called an itemset in U. An itemset may or may not be a transaction. However, a transaction must be an itemset.

(4) Dimension of the itemset. The number k of items in an item set $X = \{i_1, i_2, \cdots, i_k\}$ is called the length or dimension of the item set X, and if $|X| = k$, then X is said to be a k-dimensional item set, or $k-$ item set $(k\_Itemset)$ for short.

(5) Degree of support and confidence. The support of itemset X in transaction set T is the ratio of the number of transactions containing X in T to the number of all transactions in T. Let the number of transactions in transaction set T containing item set X be $X.count(T)$, then the support $support(X)$ of item set X in transaction set T is calculated as follows:

$$support(X) = \frac{X.Count(T)}{|T|} \times 100\% \tag{1}$$

where $|T|$ is the number of transaction records in transaction set T.

For association rule $X \Rightarrow Y$ (where X and Y are both itemsets), the confidence level of the association rule is the ratio of the number of transactions in transaction set T that contain both itemsets X and Y to the number of transactions in T that contain X, i.e.

$$confidence(X \Rightarrow Y) = \frac{support(X \cup Y)}{support(X)} \times 100\% \tag{2}$$

The confidence level describes how probable it is that when X occurs in a particular transaction, Y occurs in the same transaction.

(6) Minimum support and minimum confidence level. These two attributes are set by the user according to the actual situation, the minimum support is mainly used to measure the association rule of the lowest value in the statistics, the minimum confidence measures the association rule in the reliability of the minimum value.

(7) Frequent itemset. If the support of itemset X is greater than or equal to the minimum support threshold set by the user, i.e:

$$support(X) \geq \min\_\sup \tag{3}$$

Then X is said to be a frequent term set, or frequency set for short.

(8) Strong and weak rules. Remember $support(X \Rightarrow Y) = support(X \cup Y)$, if $support(X \Rightarrow Y) \geq \min\_\sup$, and $confidence(X \Rightarrow Y) \geq \min\_conf$, then $X \Rightarrow Y$ is called a strong rule, otherwise it is called a weak rule.

2) The execution steps of the association rule mining method are as follows:

(1) Connect the database to be data mined and prepare the data theoretically.

(2) Set the minimum support degree and minimum confidence degree of the data, fully consider the data mining algorithm suitable for this paper to set the corresponding association rules.

(3) Find the frequent itemset, set the itemset frequency greater than or equal to the minimum support frequency, that is, from the transaction set T to find and discover all the frequent itemsets whose support is not less than min_sup.

(4) Based on the obtained frequent itemsets, the corresponding strong association rules are generated. By definition these rules must satisfy the minimum confidence threshold (minconf), i.e., strong association rules with confidence greater than or equal to min_conf are found from the frequent itemsets. If the frequent itemset F and its true subset S satisfy expression $\frac{support(F)}{support(F-S)} \geq \min\_conf$, then $(F-S) \Rightarrow S$ is a strong association rule.

2.2.2. Apriori algorithm. Apriori algorithm is the basic association rule mining algorithm. In order to reduce the number of candidate data item sets, the Apriori algorithm uses a layer-by-layer search to discover all frequent item sets by scanning all transactional data items of database D [24].The Apriori algorithm generates a relatively small number of candidate data item sets, and instead of having to iteratively generate the candidate data item sets based on records in the database, the candidate data item sets are found in the process of finding the k-frequent data item set in the process of finding the $k-1$-frequent data item set generated by the previous loop at once.

The Apriori property is concluded based on the following observations. Firstly, according to Definition: $P(I) < s$, this we interpret as follows: assuming that an itemset I does not satisfy the

minimum support threshold s, then the itemset I is not a frequent itemset; and $P(IUA) < s$, the interpretation is that if an item A is added to the itemset I, then the resulting new itemset IUA cannot appear in the entire transaction database more times than the number of occurrences of the original itemset I. By these two properties, it follows that IUA cannot be frequent either, i.e., if a set fails the test, all supersets of that set also fail the same test. Thus it is easy to establish that the Apriori property holds.

An information system can be denoted as $S = (U, A, V, f)$, where U is a finite non-empty set called the argument object space, A is a set of attributes, $V = \bigcup_{a} \in A V_a$, V_a denotes the value domain of attribute a, and $f : U \times A \to V$ is an information function, i.e., for $\forall x \in U, a \in A$, there is $f(x, a) \in V_a$. If the set of attributes A is divided into a set of conditional attributes C and a set of decision-making attributes D, and satisfies $A = C \cup D, C \cap D = \varnothing$. Then $S = (U, A, V, f)$ is said to be the decision-making information system or the decision table.

For a decision rule $\phi \to \psi$ in, its rule strength, confidence factor and coverage factor can be defined respectively.

Rule strength is defined as the rule strength degree $\sigma_s(\phi, \psi) = \frac{\sup p_s(\phi, \psi)}{card(U)} = \frac{card(\| \phi \wedge \psi \|)}{card(U)}$ by setting $\sup p_s(\phi, \psi) = card \| \phi \wedge \psi \|_s)$. Rule strength indicates the percentage of the decision table that can be categorized by the rule.

The confidence factor is defined as follows: if there is a probability distribution $p_U(x) = 1 / card(U)$ for $x \in U$, then the probability of any formula ϕ in S can be defined as follows: $\pi_s(\phi) = p_U \| \phi \|_s)$. Thus there exists a conditional probability for any decision rule $\phi \to \psi$:

$\pi_s(\psi | \phi) = p_U(\| \psi \|_s \| \phi \|_s) = \frac{card \| \phi \wedge \psi \|_s)}{card \| \phi \|_s)}$, where $\| \phi \|_s \neq \varnothing$, $\pi_s(\psi | \phi)$ are called the confidence factors of decision rule $\phi \to \psi$, denoted as $cer_s(\phi, \psi)$. Clearly, if $\pi_s(\psi | \phi) = 1$, $\phi \to \psi$ is called a deterministic decision rule. If $0 < \pi_s(\psi | \phi) < 1$, $\phi \to \psi$ is called an uncertainty rule. Thus $cer_s(\phi, \psi)$ reflects the confidence that rule $\phi \to \psi$ is true, i.e., the degree to which decision ψ can be believed according to the decision table.

The coverage factor is defined as follows: $\pi_s(\phi | \psi) = p_U(\| \phi \|_s \| \psi \|_s) = \frac{card \| \phi \wedge \psi \|_s}{card \| \psi \|_s}$, denoted $\mathrm{cov}_s(\phi, \psi)$. The coverage factor reflects the confidence that the inverse rule $\psi \to \phi$ of $\phi \to \psi$ is true, i.e., the degree to which the reasons leading to a decision can be believed given that decision according to the decision table.

2.3. Decision tree ID3 algorithm and its improvement

Decision tree is an effective method to generate classifiers from a large amount of data, and it is also the most widely used logical method. This paper adopts ID3 decision tree algorithm to scientifically analyze the career choices of college graduates, construct the career decision tree of individual college students, and visually present their career choice paths under the influence of career values, so as to provide reference for college students' career choice decisions.

2.3.1. ID3 algorithm:

1) Theoretical basis of ID3 algorithm. The ID3 algorithm is mainly based on information theory, and the test attributes of internal nodes are calculated by computing the information gain metric of attributes [25]. Therefore, the following first focuses on the relevant concepts of information theory and the theory of information gain calculation.

(1) Information Entropy. Let X be a discrete random variable which may take the value of X with probability $P(x)$, then, define:

$$H(X) = -\sum_{i=1}^{n} P(x) \log P(x) \tag{4}$$

Here $H(X)$ is the information entropy of the random variable X, which is used to measure the uncertainty of the value of the random variable, and for a collection of data, entropy can be used as a measure of the impurity of the collection of data.

(2) Joint entropy. The joint random variable (X,Y) is known, assuming that each possible output (x,y) corresponds to a probability of $P(x,y)$, then according to the definition of information entropy, the formula for calculating the joint entropy is:

$$H\left(X,Y\right) = -\sum_{X,Y} P(x,y) \log P(x,y) \tag{5}$$

(3) Conditional Entropy. Given random variables X and Y, the conditional entropy of X based on Y is defined as:

$$H(X \mid Y) = H(X,Y) - H(Y) \tag{6}$$

Conditional entropy is used to measure the degree of uncertainty in attribute X when another attribute Y is known.

(4) Mutual Information. Mutual information is the correlation between two random variables X and Y. Mutual information of two random variables X & Y is defined as:

$$\begin{aligned} I(X,Y) &= H(X) - H(X \mid Y) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} p(a_i, b_j) \log_2 \frac{p(a_i, b_j)}{p(a_i) p(b_j)} \\ &= H(Y) - H(Y \mid X) \end{aligned} \tag{7}$$

It can be seen that $I(Y,X) = I(X,Y)$ and $I(Y,X) = H(X) + H(Y) - H(X,Y)$, i.e., mutual information has symmetry, which can be used to reflect the strength of the interdependence between two attributes.

(5) Conditional mutual information. Under the condition that the value of random variable Z is known to be z, the conditional mutual information of random variable X and random variable Y is defined as:

$$\begin{aligned} I(X,Y \mid z) &= H(X \mid z) - H(X \mid Y,z) \\ &= -\sum_{i=1}^{n} p(a_i,z) \log_2 p(a_i,z) + \sum_{i=1}^{n} \sum_{j=1}^{m} p(a_i,b_j,z) \log_2 p(a_i \mid b_j,z) \\ &= \sum_{i=1}^{n} \sum_{j=1}^{m} p(a_i,b_j,z) \log_2 \frac{p(a_i \mid b_j,z)}{p(a_i \mid z) p(b_j \mid z)} \\ &= H(Y \mid z) - H(Y \mid X,z) \\ &= I(Y,X \mid z) \end{aligned} \tag{8}$$

It is primarily used to indicate the strength of the degree of interdependence between two random variables X and Y, given that a certain value z of a random variable Z is known. In a given certain data set, the conditional mutual information quantity can indicate the degree of interdependence between two other attributes when the value of an attribute is known.

(6) Information gain calculation theory. Let S be a collection of s data samples, the data sample set is divided into m different classes $C_i(i=1,2,...,c)$ and the number of samples in each class C_i is s_i, then the information entropy of S divided into m classes is:

$$H(S) = -\sum_{i=1}^{m} p_i \log_2 p_i \tag{9}$$

where $p_i = s_i / s$ is the probability that a sample in S belongs to class i, C_i.

Assuming that the set of all different values of attribute C is $Values(C)$ and S_v is a subset of samples in S whose value of attribute C is v, i.e., $S_v = \{s \in S \mid C(s) = v\}$, the information entropy of classifying the sample set S_v of this node at each branch node after selecting attribute C is $H(S_v)$. The expected entropy obtained by selecting C is defined as a weighted sum of information entropies of each subset S_v, with the weight being the ratio between the number of samples in sample set S_v and the number of samples in original set S, i.e., the expected entropy is the ratio, i.e., the expected entropy is:

$$H(S,C) = \sum_{v \in Vahes(C)} \frac{|S_v|}{|S|} H(S_v) \tag{10}$$

where $H(S_v)$ is the information entropy of classifying the samples in S_v into m classes.

The information gain $Gain(S,C)$ of attribute C with respect to the set of samples S is then defined:

$$Gain(S,C) = H(S) - H(S,C) \tag{11}$$

Information gain $Gain(S,C)$ is the expected compression of entropy resulting from knowing the value taken by attribute C. The larger the information gain, the more information the selection of test attribute C provides for classification.

2) Basic idea of ID3 algorithm. The basic idea of ID3 algorithm is a greedy algorithm that constructs a decision tree using top-down recursion. It starts with the set of training samples at the root node of the tree, selects an attribute with the largest information gain as a test attribute to distinguish these samples, generates a branch for each value of the test attribute, and the subset of samples corresponding to the branching attribute is shifted to the newly generated child node. The algorithm is applied recursively to each generated sub-node until all samples on a node are classified into some class, such that each path from the root node to the leaf nodes of the decision tree represents a classification rule.

2.3.2. Improvements to the ID3 algorithm. Aiming at the shortcomings of the ID3 algorithm in the decision tree that relies on the attributes with a large number of attribute values and the characteristics of the mathematical formulas used in the calculation of information gain, this paper introduces the attribute weights to improve the ID3 algorithm from the specifics of the employment information dataset. Then, using the logarithmic function is a convex function of this feature, the information gain calculation is simplified to reduce the amount of algorithmic calculations, and finally, on the basis of the original ID3 algorithm for the above improvements, the introduction of a post pruning method - Pessimistic Error Pruning (PEP) method, in order to reduce the complexity of the tree, so that the construction of decision tree models More easy to understand.

1) Introduction of attribute weights. Since the importance of decision-making attributes (non-categorized attributes) in the employment data set does not differ much, but the number of attribute

values varies greatly, this paper introduces attribute weights ω_i to reduce the impact of those attributes with more values on the classification results of the model.

The attribute weights are defined as:

$$\omega_i = \frac{q_i}{q} \tag{12}$$

where n is the total number of decision attribute sets, q_i is the number of values of a decision attribute $i=(I,2,...,n)$, q is the total number of values of all decision attributes. From the definition, the sum of attribute weights w_i on each decision attribute is 1. The information gain after the introduction of attribute weights is changed from Eq. (11) to:

$$Gain(S,C) = H(S) - w_i * H(S,C) \tag{13}$$

2) Simplification of ID3 algorithm. From the formula (9) and (13) is not difficult to find, as long as the data sample set and the number of categories are given, formula (9), that is, (13) in the $H(S)$ in the whole ID3 algorithm calculation process is always the same. In order to reduce the amount of computation, it can be taken out of Eq. (13), at which point Eq. (13) only retains the last term without affecting the final result. It can be further improved as:

$$H'(S,C) = -w_i^* \sum_{v \in Values(C)} \frac{|S_v|}{|S|}^* \sum_{i=1}^{m} p_{ij} log_2 p_{ij} \tag{14}$$

The function $\log_2 p_i$ in the above equation is a logarithmic function similar to $f(x)=\log_2 x$, where $p_{ij} \in (0,I)$ is the probability that the sample in S_j belongs to class C_i. From the fact that $f(x)=\log_2 x$ is a convex function on $x \in (0,1)$ it follows that $\log_2 p_i j$ is also convex on $p_{ij} \in (0,I)$.

According to the properties of convex functions, if $f(x)$ is a convex function, X is a non-empty set, $x_i \in X$, then $\exists \lambda \geq 0, \sum_{i=1}^{m} \lambda = 1$, making:

$$\sum_{i=1}^{m} \lambda f(x_i) \leq f\left(\sum_{i=1}^{m} \lambda x_i \right) \tag{15}$$

Similarly, $-\sum_{i=1}^{m} p_{ij} \log_2 p_{ij}$ in Eq. (14) can be converted to $-\sum_{i=1}^{m} p_{ij} \log_2 p_{ij} \geq -\log_2 \left(\sum_{i=1}^{m} p_{ij}^2 \right)$ according to the nature of convex function At this point, Eq. (14) can then be replaced by the following equation:

$$H'(S,C) = -w_i^* \sum_{veValues(C)} \frac{|S_v|}{|S|}^* log_2 (\sum_{i=l}^{m} p_{ij}^2) \tag{16}$$

The original ID3 algorithm needs to perform several operations on $\log_2 p_{ij}$, while the simplified ID3 algorithm only needs to perform one operation on $\log_2 p_{ij}$ to obtain the approximation of the information gain, which reduces the number of logarithmic operations and thus reduces the amount of computation, and improves the computational efficiency in a larger way.

3) Adding a post pruning method. Since the decision trees generated by the original ID3 algorithm are all too cumbersome and overfit the data, this paper deliberately adds a common pruning method to the improved ID3 algorithm to avoid uncontrolled growth in the height of the tree. The method uses the training samples themselves to estimate the error before and after pruning to decide whether to actually prune or not. The method uses the following formula:

$$\Pr\left[\frac{f-q}{\sqrt{q(1-q)/N}} > z\right] = c \tag{17}$$

where N is the number of instances, $f = EN$ is the observed error rate (where E is the number of classification errors in the N instances), q is the true error rate, c is the confidence level, and z is the standard deviation corresponding to the confidence c level. An upper limit of the confidence interval for the true error rate can be computed by using Equation (17), and this upper limit is used to make a pessimistic estimate for the error rate e of the node:

$$e = \frac{f + \frac{z^2}{2N} + z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}} \tag{18}$$

The size of e before and after pruning is determined by Eq. (18), which determines whether pruning is needed.

3. Model application analysis

3.1. Correlation analysis of factors influencing the formation of vocational skills of college students

Based on the data of college students' information survey, this paper uses Weka data mining tool to analyze the correlation of the three factors of individual, school and family that affect the results of college students' vocational skill formation assessment. The data collected in this paper are mostly categorical variables, and in order to explore the correlation of each factor on skill formation, the Apriori algorithm in the association rule algorithm is chosen for correlation analysis.

3.1.1. Results of correlation analysis of individual factors. Through the association rule mining experiment, the results of the correlation analysis of personal factors can be obtained as shown in Table 1. According to Table 1, it can be seen that the rules derived from data mining in this paper all satisfy the confidence level greater than 0.9, the enhancement degree greater than 1, the leverage rate greater than 0, and the certainty degree greater than 0. Therefore, all of these factors have a strong correlation with the formation of occupational skills.

Table 1. Correlation analysis results of individual factors

Coding	Association rule	Confidence	Lift	Leverage	Conviction
1	certificate=c, rank=C ==> match=c	1	1.45	0.09	6.41
2	grade=b, certificate=b ==> rank=A	0.98	1.49	0.13	4.97
3	rank=C ==> match=c	0.97	1.39	0.10	4.21
4	qizhi=a, aim=a ==>question=a	0.97	1.26	0.07	2.92
5	qizhi=a, attention=b ==> question=a	0.97	1.26	0.07	2.92
6	grade=b, aim=a ==> rank=A	0.97	1.48	0.10	3.96
7	qizhi=a, match=c ==> question=a	0.97	1.26	0.06	2.65
8	grade=a, rank=C ==> match=c	0.96	1.38	0.08	3.41
9	certificate=c, question=a ==> match=c	0.96	1.38	0.08	3.41
10	question=a, learn_num=a ==> rank=A	0.96	1.48	0.09	3.79
11	learn_num=a, e_learning=a ==> rank=A	0.96	1.48	0.09	3.79
12	grade=b, certificate=b, question=a ==> rank=A	0.96	1.48	0.09	3.79
13	qizhi=a, aim=a, rank=A ==> question=a	0.96	1.26	0.06	2.54
14	grade=b, qizhi=a ==> rank=A	0.96	1.48	0.09	3.63
15	grade=b, e_learning=a ==> rank=A	0.96	1.48	0.09	3.63
16	certificate=b, learn_numa w==> rank=A	0.96	1.48	0.09	3.63
17	question=a, rank=C ==> match=c	0.96	1.38	0.08	3.22
18	grade=b, aim=a, question=a ==> rank=A	0.96	1.48	0.09	3.63
19	qizhi=a ==> question=a	0.95	1.24	0.08	2.54
20	certificate=b, aim=a ==> question=a	0.94	1.23	0.07	2.19

From the rules derived from Table 1, the rules with the suffix “rank=A”, i.e., occupational skill formation measure with a rank of A, were filtered out and the meaning of their representation was explained. The filtered rules are as follows:

1) Grade level = junior year, number of certificates = 1 ==> skill formation assessment rank = A.

2) Grade level = Junior year, Goal = Have a clear goal and stick to it ==> Skill assessment grade = A.

3) Ways of dealing with problems = Ask teacher or classmates and discuss together Number of after-school studies = More ==> Skill assessment grade = A.

4) Number of times learning after school = more often, whether or not using electronic devices to learn = yes ==> Skill Measurement Rating = A.

5) Grade level = junior year, number of certificates = 1, way of dealing with problems = ask teacher or classmates, discuss together ==> Skill assessment grade = A.

6) Grade level = Junior year, Temperament type = Polycythemia ==> Skill assessment grade = A.

7) Grade level = junior year, whether or not you utilize electronic devices for learning = yes ==> Skill Measurement Rating = A.

8) Number of certificates = 1, Number of after-school studies = More ==> Skill assessment grade = A.

9) Grade level = junior year, goal = have a clear goal and keep working on it, way of dealing with problems = ask teacher or classmates and discuss together ==> Skill Measurement Rating = A.

In these 9 correlation rules it can be concluded that 1: personal factors of grade level, number of professional skills certificates obtained, whether there are clear goals, the number of after-school studies, the way of dealing with problems, whether they will use electronic devices for learning, and the type of temperament have a strong correlation with the formation of students' vocational skills.

3.1.2. Results of correlation analysis of school factors. The results of the correlation analysis of the school factor are shown in Table 2. According to Table 2, it can be seen that the minimum value of confidence of the association rules of school factors derived from data mining in this paper is 0.94>0.9,

the degree of enhancement is greater than 1, the leverage is greater than 0, and the degree of certainty is greater than 0, which proves the reliability of the results of this data mining.

Table 2. Correlation analysis results of school factors

Coding	Association rule	Confidence	Lift	Leverage	Conviction
1	school=C, school_att=a ==> facility=a	1	1.67	0.15	7.38
2	media=a, t_urge=a ==> relation=a	1	1.25	0.07	3.41
3	school=C ==> rank=A	0.97	1.17	0.07	2.04
4	media=b, school_att=a ==> facility=a	0.97	1.61	0.18	4.68
5	school=C, facility=a ==> rank=A	0.97	1.17	0.06	1.89
6	school=C, course=a ==> rank=A	0.96	1.17	0.06	1.78
7	t_urge=a, school_att=a, rank=A ==> relation=a	0.96	1.19	0.07	1.99
8	updata=b, t_urge=a ==> facility=a	0.96	1.58	0.12	4.08
9	media=b, school_att=a, rank=A ==> facility=a	0.96	1.58	0.12	4.08
10	facility=a, t_urge=a, school_att=a ==> course=a	0.96	1.54	0.12	3.94
11	course=a, t_urge=a, rank=A ==> relation=a	0.96	1.20	0.06	1.88
12	school=C, relation=a ==> facility=a	0.96	1.58	0.12	3.89
13	school=C, textbook=a ==> rank=A	0.96	1.16	0.06	1.63
14	school=C, relation=a ==> rank=A	0.96	1.16	0.06	1.63
15	updata=a, rank=A ==> relation=a	0.96	1.18	0.06	1.79
16	school=C, facility=a, course-a ==> rank=A	0.96	1.15	0.06	1.64
17	media=b, course=a, school_att=a ==> facility=a	0.96	1.58	0.12	3.90
18	t_urge=a, rank=A ==> relation=a	0.95	1.17	0.08	1.97
19	t_urge=a, school_att=a ==> relation=a	0.94	1.15	0.07	1.68
20	t_urge=a ==> relation=a	0.94	1.14	0.08	1.75

Specifically, from the association rules obtained from Table 2, the rules with "rank=A", i.e., occupational skill formation assessment level A, are filtered out, and the meanings of the filtered rules are as follows:

1) School=C, teaching equipment=basically functional ==>skill assessment rank=A.

2) School = C, Curriculum = Reasonable ==> Skill Measurement Rating = A.

3) Teacher-student relationship = better ==> Skill assessment grade = A.

4) Teaching equipment = basically all working, curriculum = reasonable ==> skill assessment grade = A.

In these 4 correlation rules it can be concluded 2: the formation of vocational skills has a strong correlation with the situation of school equipment, the reasonableness of curriculum and teacher-student relationship.

3.1.3. Results of correlation analysis of family factors. The results of the correlation analysis of family factors are shown in Table 3. The confidence level of the association rules for family factors derived from data mining are all greater than 0.9, the lift is greater than 1, the leverage is greater than 0, and the degree of certainty is greater than 0. This indicates that the resulting association rules have a high degree of confidence.

Table 3. Correlation analysis results of family factors

Coding	Association rule	Confidence	Lift	Leverage	Conviction
1	edu_mode=c ==> rank=A	1	1.31	0.09	5.11
2	f_type=b, edu_mode=c => rank=A	1	1.31	0.07	4.05
3	edu mode=c, plan=b ==> rank=A	1	1.31	0.07	3.86
4	f_form=a, income=a, plan=b ==> f_type=b	1	1.46	0.09	5.47
5	f_form=a, f_edu=b ==> rank=A	1	1.31	0.06	3.63
6	f_form=a, edu_mode=c => rank=A	1	1.31	0.06	3.63
7	f_form=a, communication=a ==> plan=b	1	1.52	0.09	5.11
8	m_work=c, income=a -=> f_type=b	1	1.46	0.08	4.88
9	f_edu-b communication=a ==> plan=b	1	1.52	0.09	5.11
10	m_edu=a, income=a, plan=b ==> f_type=b	1	1.46	0.08	4.88
11	f_form=a, income=a, plan=b, rank=A ==> f_type=b	1	1.46	0.08	4.88
12	f_type=b, f_work=h, rank=A==> income=a	0.97	1.46	0.10	3.36
13	m_work=c ==> f_type=b	0.96	1.38	0.08	2.75
14	m_work=c ==> rank=A	0.96	1.24	0.05	1.94
15	f_type=b, f_edu=b ==> rankA	0.96	1.24	0.05	1.94
16	f_work=h, plan=b ==> f_type=b	0.96	1.37	0.07	2.61
17	f_type=b, m_work=c ==> income=a	0.96	1.44	0.08	2.87
18	m_work=c, rank=A ==> f_type=b	0.96	1.38	0.07	2.61
19	f_type=b, m_work=c ==> rank=A	0.96	1.21	0.05	1.84
20	f_work=h, m_edu=a ==> income=a	0.96	1.43	0.08	2.85

From Table 3, we screened the rules with the post-piece of "rank=A", i.e., the rules with occupational skill formation assessment rank of A. The rules obtained from the screening and their meanings are as follows:

1) Family education style = authoritative ==> skill assessment rank = A.

2) Type of household registration = rural, family education style = authoritative ==> skill formation assessment grade = A.

3) Family education style = authoritative type, whether parents give planning and guidance = yes ==> Skill Measurement Rating = A.

4) Family type = intact family, father's level of education = Bachelor's degree ==> Skill Measurement Rating = A.

5) Family type = intact family, family education style = authoritative => skill measurement level = A.

Summarizing the above five correlation rules, it can be concluded that the correlation between family type, family education style, and parents' ability to plan and guide the students' studies are strongly related to the formation of vocational skills.

3.2. Career choice of college students based on ID3 algorithm

This paper utilizes data mining technology to obtain the ranking of various factors affecting career choice through ID3 decision tree classification algorithm, and at the same time generates an image of the decision tree, which provides a decision support role for the individual career choice of college students.

This paper is aimed at the job-seeking and employment problems of college students, which are specifically analyzed. According to the statistics of N college students' career counseling studio, 75% of the student counseling cases in the academic year 2021~2022 belong to decision-making problems. One part of the decision-making problems is emotional, which helps college students make decisions through emotional relief, and the other part needs to be helped to complete their career choices by sorting out values and repeatedly confirming career options, which is suitable for this part of college

students to analyze career values with the help of the decision tree classification algorithm.

By means of interviews or value cards, the relevant personnel will analyze the values preferences of the visiting college students, listen to 4-6 most important influencing factors in the career choice of the visiting college students, compile the options, and quantify the influencing factors according to the criteria of the college students and the viewpoints of the visitors, so as to facilitate the analysis of the algorithm.

According to the four horizontal orthogonal table generation options of "economic income", "working atmosphere", "development platform" and "work location and environment", the visitors filled in the job search willingness one by one, and obtained the simulated decision-making dataset based on occupational values, as shown in Table 4. Among them, the specific evaluation criteria for different factors and different levels are as follows:

1) Economic income level: a monthly salary higher than 8,000 is excellent, 8,000 to 500 is good, 5,000 to 3,500 is medium, and 3,500 points or less is poor.

2) Work atmosphere: loose company system and harmonious relations between subordinates is excellent, more strict company system and more harmonious relations between subordinates is good, strict company system and more serious relations between subordinates is medium, strict company system and stereotypical relations between subordinates is poor.

3) Development platform: a large company in the rising stage is excellent, a large company in the stable development stage is good, a small and medium-sized enterprise in the rising stage is medium, and a small and medium-sized enterprise is poor.

4) Workplace and environment: first and second tier big cities are excellent, third and fourth tier small and medium-sized cities are good, economically underdeveloped areas are medium, and remote mountainous rural areas are poor.

5) Job-seeking willingness grade: A for those who are very willing, B for those who are willing to try alternatives, C for those who can accept without alternatives, D for those who won't choose even if they don't work for a while, and F for those who can't make a definite choice.

According to Table 4, the information gain affecting career decision-making can be calculated, and the influencing factor with the highest information gain can be used as the test influencing factor of the set, the nodes can be generated and labeled with the test influencing factor, and for each value of the test influencing factor, a fork can be generated, so as to generate a decision tree diagram.

Table 4. Simulated decision data sets based on professional values

Serial number	Economic income	Working atmosphere	Development platform	Workplace and environment	Job-seeking intention
1	Optimal	Optimal	Optimal	Optimal	A
2	Optimal	Good	Good	Good	A
3	Optimal	Poor	Medium	Medium	B
4	Optimal	Poor	Poor	Poor	D
5	Good	Optimal	Good	Medium	A
6	Good	Good	Optimal	Poor	B
7	Good	Medium	Poor	Optimal	C
8	Good	Poor	Medium	Good	B
9	Medium	Optimal	Medium	Poor	C
10	Medium	Good	Poor	Medium	C
11	Medium	Medium	Optimal	Good	A
12	Medium	Poor	Good	Optimal	C
13	Poor	Optimal	Poor	Good	C
14	Poor	Good	Medium	Optimal	C
15	Poor	Medium	Good	Poor	C
16	Poor	Poor	Optimal	Medium	B

Taking the visitor number N2021019 as an example, the information gain of each influencing factor is calculated and the influencing factor with the highest information gain is selected as the test influencing factor for the given set, thus generating the corresponding branch nodes. The resulting node calculates the information gain obtained by dividing the sample set using the attribute “economic income”:

$$\begin{aligned} &Gain(Economic\_income) = I\left(s_1, s_2, \ldots s_m\right) \\ &= E\left(Economic\_income\right) = 0.625 \end{aligned} \tag{19}$$

where s_i is the number of samples in the different categories of influences C_i.

The same reasoning can be obtained:

$$Gain(Working\_atmosphere) = 0.541 \tag{20}$$

$$Gain(Development\_platform) = 0.824 \tag{21}$$

$$Gain(Woekplace\_and\_environment) = 0.453 \tag{22}$$

Obviously, the visitor's choice of the attribute “development platform” yields the greatest information gain and is therefore used as the root of the decision tree. The above steps are repeated for each branch, and the final decision tree is shown in Figure 2.

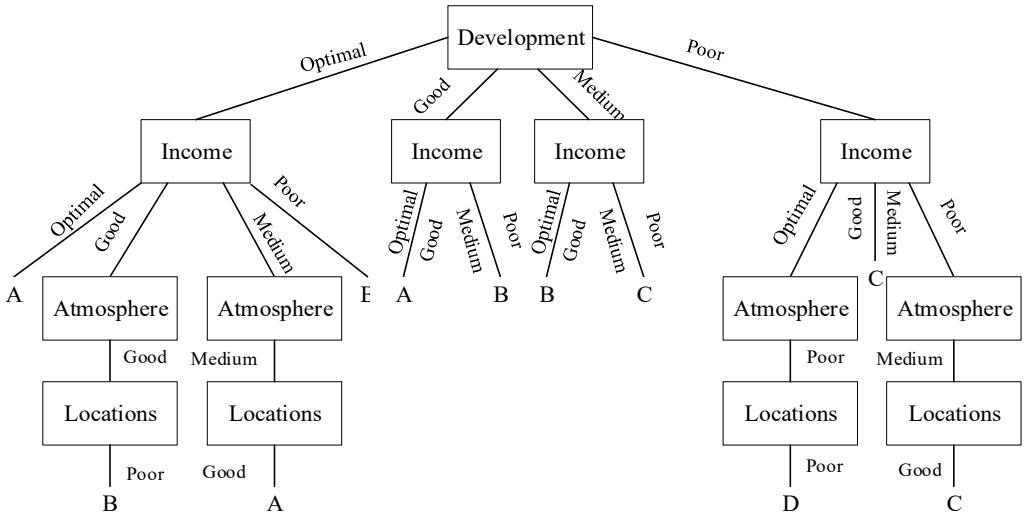

Fig. 2. Decision tree based on professional values

Through the data mining technology, the decision tree of college students' employment choices is obtained, and the following conclusion can be drawn: the visitor (No. N2021019) attaches the most importance to the prospect of employment development in employment decision-making, followed by the level of income, and the prospect and income basically determine the visitor's level of willingness to look for employment. Work atmosphere and work location are only auxiliary reference factors for the visitors.

The intuitive decision tree diagram obtained by data mining technology can guide college students to clarify the trade-offs in career choice decision-making, and then compare the results of values analysis and ranking with standard questionnaires, values classification cards and other tools, which can help college students clarify their career choices very well. At the same time, it has the advantages of easy operation and little influence by human factors, and combined with traditional career counseling techniques, it can provide support for college career guidance agencies to scientifically understand and guide students.

3.3. Analysis of Factors Influencing College Students' Career Planning

In order to better guide college students to seek employment and make career planning, it is necessary to explore the influencing factors affecting college students' employment. This section analyzes the influencing factors of college students' employment based on correlation analysis and cluster analysis.

3.3.1. Correlation analysis of several influencing factors. First of all, this paper selects college students' employment questionnaire, college students' employment cognition evaluation questionnaire, college students' employment choice questionnaire, combined with interviews with some students, to form the college students' employment influencing factors questionnaire. According to the results of the questionnaire, the college students' employment influencing factors such as image, gender, English level, computer level, qualification certificate, social practice experience, achievement, employment situation, expected salary, ideal employment unit, and efforts in school are extracted.

Then, several influencing factors among them, such as academic achievement, qualification certificate, social practice experience, expected salary, and ideal employment unit, were analyzed for correlation, and the correlation matrix of influencing factors was obtained as shown in Table 5. Where ** indicates significant correlation at 0.01 level (two-sided).

As can be seen from Table 5, there is a significant positive correlation between "academic performance" and "qualifications", "expected salary" and "ideal employment unit", and a significant negative correlation between "academic performance" and "social practice experience". There was a

significant positive correlation between "qualifications" and "social practice experience" and "ideal employment unit", and a significant negative correlation between "qualifications" and "expected salary". There was a significant negative correlation between "social practice experience" and "expected salary", but there was no significant correlation between "ideal employment unit". There is no significant correlation between "expected salary" and "ideal place of employment".

Table 5. Correlation matrix of influencing factors

	Academic performance	Qualification certificate	Social practice experience	Expected compensation	Ideal employment unit
Academic performance	1				
Qualification certificate	0.302**	1			
Social practice experience	-0.831**	0.203**	1		
Expected compensation	0.875**	-0.203**	-0.825**	1	
Ideal employment unit	0.237**	0.474**	-0.03	0.03	1

3.3.2. Cluster analysis. In order to explore the differences in college students' perceptions of employment influencing factors, this paper carries out a cluster analysis of the sample students based on the results of the questionnaire survey [26]. The results of the cluster analysis are shown in Table 6, in which clusters 1~4 indicate the four cluster centers obtained, and the views of the employment situation in the data in the table from 1~5 indicate very pessimistic, relatively pessimistic, average, relatively optimistic, and very optimistic, respectively. Expected salary from 1~5 indicates >10,000 yuan/month, 8,000~10,000 yuan/month, 5,000~8,000 yuan/month, 3,000~5,000 yuan/month, and <3,000 yuan/month, respectively. The 1~5 of other factors indicate very little to very great importance, respectively.

The first group of students think that English proficiency, computer proficiency, school effort and final employment have very great influence on employment, qualification certificates, social practice experience, grades have average influence on employment, gender and image have less influence on employment, the ideal employment unit is a foreign-funded enterprise, the employment situation is relatively optimistic, and the expected salary is 3,000-5,000 yuan/month. The second group of students think that qualification certificates and social practice experiences have a very big influence on employment, gender, image, English level, computer level and school efforts and final employment have a relatively big influence on employment, grades have an average influence on employment, the expected salary is 5,000-8,000 yuan/month, the ideal employment unit is an institution, and they are pessimistic about the employment situation. The third group of students think that English proficiency, computer proficiency, social internship experience has a very big impact on employment, gender, image and school efforts and final employment have a big impact on employment, qualification certificates and grades have a small impact on employment, the expected salary is 5000-8000 yuan/month, the ideal employment unit is state-owned enterprises, and they are pessimistic about the employment situation. The fourth group of students think that English proficiency and computer proficiency have a very big impact on employment, gender, image and school effort and final employment have a relatively big impact on employment, qualification certificates, social practice experience and grades have an average impact on employment, gender and image have a very small impact on employment, the expected salary is 3,000-5,000 yuan/month, the ideal employment unit is a governmental department, and they are more optimistic about the employment situation.

Table 6. Correlation matrix of influencing factors

Influencing factor	Cluster			
	1	2	3	4
Gender	2	4	4	2
Image	2	4	4	2
English proficiency	5	4	5	5
Computer level	5	4	5	5
Qualification certificate	3	5	2	3
Social practice experience	3	5	5	3
Academic performance	3	3	2	3
Views on employment situation	4	2	2	4
Expected compensation	4	3	3	4
Ideal employment unit	2	4	3	5
School efforts and eventual employment	5	4	4	4

4. Conclusion

In this paper, we comprehensively use Apriori algorithm, ID3 decision tree algorithm, cluster analysis, correlation analysis and other methods to mine the Internet big data of college students' job-seeking and employment and career planning, and explore its potential regularity.

1) The influencing factors of college students' professional skill generation were analyzed from three aspects: personal, school and family. Among the personal factors, grade level, the number of professional skill certificates obtained, whether there is a clear goal, the number of after-school study, the way to deal with problems, whether they will use electronic devices to study, and temperament type have a strong correlation with the formation of students' vocational skills. Among the school factors, the formation of vocational skills is strongly associated with the condition of the school equipment, the rationality of the curriculum, and the relationship between teachers and students. And among the family factors, the points of family type, family education style and whether parents can plan and guide students' studies have a stronger association with the formation of vocational skills.

2) Take the visitor with number N2015012 as an example for data mining to construct a decision tree based on vocational values. The results show that this visitor values the prospect of employment development the most in employment decision-making, followed by the income level, and the prospect and income basically determine the visitor's level of willingness to seek employment. Work atmosphere and work location are only auxiliary reference factors for the visitor. This method provides support for school career guidance organizations to understand and guide students scientifically.

3) Among the factors affecting college students' job search and employment, there was a significant positive correlation between "academic performance" and "qualification certificate", "expected salary" and "ideal employment unit", and a significant negative correlation between "social practice experience". There was a significant positive correlation between "qualification" and "social practice experience" and "ideal employment unit", and a significant negative correlation between "expected salary". There was a significant negative correlation between "social practice experience" and "expected salary", but there was no significant correlation between "ideal employment unit". There is no significant correlation between "expected salary" and "ideal place of employment". All these factors will affect the employment choices of college students, and there are great differences among different types of students.

References

[1] Zhang, J. (2021, November). The Application and Thinking of Big Data in the Career Planning Education of College Students. *In 2021 2nd International Conference on Information Science and Education* (ICISE-IE)

(pp. 868-871). IEEE.

[2] Wang, Y., Yang, L., Wu, J., Song, Z., & Shi, L. (2022). Mining campus big data: Prediction of career choice using interpretable machine learning method. *Mathematics*, 10(8), 1289.

[3] Yan, X. (2024). Research on the Strategies for Improving College Students' Career Adaptability from the Perspective of Big Data. *Frontiers in Educational Research*, 7(1).

[4] Nie, M., Xiong, Z., Zhong, R., Deng, W., & Yang, G. (2020). Career choice prediction based on campus big data—mining the potential behavior of college students. *Applied Sciences*, 10(8), 2841.

[5] Chen, L., & Jin, X. (2024). Exploring the Impact of College Students' Future Time Insights on Career Decision-making Difficulties: A Big Data Technology-based Approach. *Journal of Electrical Systems*, 20(3), 410-419.

[6] Dorsey, D. W. (2019). Big Data, Data Science, and Career Pathways. *Career Pathways: From School to Retirement*, 239.

[7] Yin, J., & Zhang, W. (2023, August). Research on Talent Demand Analysis in Big Data Related Fields Based on Text Mining. *In Proceedings of the 2023 6th International Conference on Information Management and Management Science* (pp. 33-40).

[8] Chen, J., Wu, J., & Jiang, W. (2021, March). Demand analysis of employment talents based on deep learning. *In IOP Conference Series: Earth and Environmental Science* (Vol. 692, No. 4, p. 042017). IOP Publishing.

[9] Wang, J. (2022). Optimization design of international talent training model based on big data system. *Frontiers in Psychology*, 13, 949611.

[10] Dai, J., & Su, K. (2022). Empirical Research on Data Science Talent Training and Enterprise Demand in Colleges and Universities. *Pacific International Journal*, 5(4), 188-199.

[11] Li, L. (2024). Research on the Characteristics of Industrial Talent Demand Depending on Big Data Technology. *Journal of Electrical Systems*, 20(3), 1259-1271.

[12] Saputra, A., Wang, G., Zhang, J. Z., & Behl, A. (2022). The framework of talent analytics using big data. *The TQM Journal*, 34(1), 178-198.

[13] Wei, J., & Xu, Y. (2022, August). The Application of LDA Model in the Analysis of Job Talent Demand under Big Data Technology. *In 2022 International Conference on Artificial Intelligence in Everything* (AIE) (pp. 301-305). IEEE.

[14] Xiao, Q., Zhong, X., & Zhong, C. (2020). Application research of KNN algorithm based on clustering in big data talent demand information classification. *International journal of pattern recognition and artificial intelligence*, 34(06), 2050015.

[15] Li, L., & Tsai, S. B. (2022). An Empirical Study on the Precise Employment Situation‐Oriented Analysis of Digital‐Driven Talents with Big Data Analysis. *Mathematical Problems in Engineering*, 2022(1), 8758898.

[16] Zhang, Q., Zhu, H., Sun, Y., Liu, H., Zhuang, F., & Xiong, H. (2021, August). Talent demand forecasting with attentive neural sequential model. *In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 3906-3916).

[17] Nie, M., Yang, L., Sun, J., Su, H., Xia, H., Lian, D., & Yan, K. (2018). Advanced forecasting of career choices for college students based on campus big data. *Frontiers of Computer Science*, 12, 494-503.

[18] Guo, J., & Qi, L. (2022). Visualization research of college students' career planning paths integrating deep learning and big data. *Mathematical Problems in Engineering*, 2022(1), 6006968.

[19] Kamal, A., Naushad, B., Rafiq, H., & Tahzeeb, S. (2021, November). Smart career guidance system. *In 2021 4th International Conference on Computing & Information Sciences* (ICCIS) (pp. 1-7). IEEE.

[20] Haritha, D. D. (2019). Smart career guidance and recommendation system. *International Journal of Engineering Development and Research*, 7(3), 633-638.

[21] Yang, T. C., & Chang, C. Y. (2023). Using institutional data and messages on social media to predict the career decisions of university students-A data-driven approach. *Education and Information Technologies*, 28(1), 1117-1139.

[22] José-García, A., Sneyd, A., Melro, A., Ollagnier, A., Tarling, G., Zhang, H., ... & Arthur, R. (2023). C3-IoC: A career guidance system for assessing student skills using machine learning and network visualisation. *International Journal of Artificial Intelligence in Education*, 33(4), 1092-1119.

[23] Rahul Haripriya, Nilay Khare, Manish Pandey & Sreemoyee Biswas. (2024). Decentralized big data mining: federated learning for clustering youth tobacco use in India. Journal of Big Data(1),179-179.

[24] Yang Li & Haiyu Zhang. (2024). Big data technology for teaching quality monitoring and improvement in higher education - joint K-means clustering algorithm and Apriori algorithm. Systems and Soft Computing200125-200125.

[25] Agarwal Arun, Jain Khushboo & Yadav Rakesh Kumar. (2023). A mathematical model based on modified ID3 algorithm for healthcare diagnostics model. International Journal of System Assurance Engineering and Management(6),2376-2386.

[26] Jingshuo Li, Shengmei Ma, Danyang Wu, Ziqi Zhang, Yuxian Chen, Bo Liu... & Haipeng Jia. (2025). CT-based radiomics and cluster analysis for the prediction of local progression in stage I NSCLC patients treated with microwave ablation. iScience(1),111552-111552.