International Journal for Housing Science and Its Applications
Publish August 6, 2025. Volume 46, Issue 3 Pages 8527-8539

https,//doi.org/10.70517/ijhsa463729

Research on the Integration of Data Mining-Based Intelligent Computing and Traditional Culture in the Cultivation of Innovative Talent under the New Quality Productivity Paradigm

Shuang Li[1,*], Sujie Tian[1] and Min Ding[1]
[1] Department of Automobile Engineering, Jining Polytechnic, Jining, Shandong, 272000, China
Corresponding authors: (e-mail: sdlishuang@sina.com).

Abstract This paper utilizes data mining methods to mine talent profiles. Through the Scrapy framework and text mining methods, talent tags are mined and their features are extracted. To address the limitations of the traditional FCM clustering algorithm, the CFSFDP algorithm is introduced for optimization, proposing a density peak-optimized fuzzy C-means algorithm (FDP-FCM). This algorithm is compared with other algorithms in terms of clustering performance and robustness to evaluate its effectiveness. Regression analysis is employed to explore the influencing factors and differences in graduates' job-seeking and further education decisions. Finally, a talent cultivation model for traditional cultural integration and innovation based on user profiles is proposed. The FDP-FCM algorithm achieved the best performance among all algorithms in terms of F-measure, RI, and Jaccard coefficient. Under the new productive forces, provincial types have predictive effects on the achievement of employment and further education goals for graduates. Academic performance and job preparation have more significant predictive effects on employment-oriented graduates, while academic performance has a more pronounced predictive effect on further education-oriented graduates.

Index Terms data mining, Scrapy framework, FDP-FCM algorithm, new quality productivity, talent cultivation

I. Introduction

Science and technology and culture form the two wings of modern civilization. The deep integration of the two is a key driving force for the development of new-quality productive forces, and an important practical path for building a cultural powerhouse, a scientific and technological powerhouse, and advancing the process of China's modernization [1]-[4]. Modern science and technology are the core driving force of new-quality productive forces, while traditional culture can provide inspiration and a source of innovation for modern science and technology [5], [6]. By promoting the integration of traditional culture and modern science and technology, we can uncover value concepts, ways of thinking, and practical experiences that align with the development of new-quality productive forces. This enables the creation of new products, services, and business models with unique cultural connotations and contemporary characteristics, driving breakthroughs in the concepts, technologies, and business models of new-quality productive forces, and leading industrial upgrading and economic development [7]-[10].

Furthermore, the integration of science and technology with culture has not only revolutionized and reshaped new-quality productive forces but also redefined and restructured new-quality talent cultivation models [11], [12]. The development of new-quality productive forces requires the cultivation of new-quality talent that aligns with these forces, can fully utilize new-quality production tools, and generate innovative production value [13], [14]. New-quality talent serves as the primary driving force behind the formation of new-quality productive forces. They are outstanding and exceptional individuals characterized by a sense of national pride, innovative thinking, practical capabilities, and cross-disciplinary competencies [15]. The cultivation of new-type talent in higher education institutions is an integral part of a high-quality higher education system. The integration of culture and technology provides directional guidance and a mission for talent cultivation models [16], [17]. Enhancing the quality of new-type talent cultivation through the integration of culture and technology holds significant importance for enhancing national cultural soft power, building a high-quality higher education system, and promoting high-quality industrial development.

To achieve the integration of traditional culture with a new model for cultivating talent in productive capacity, this paper first uses data mining methods to construct a talent profile, builds a standard talent profile using the Scrapy framework, and obtains occupational requirements. Text mining technology is then used to perform big data mining

and feature extraction on talent tags. The FCM clustering algorithm is selected as the base method, combined with the CFSFDP algorithm, and a fuzzy C-means algorithm (FDP-FCM) based on density peak optimization is proposed for tag clustering. To validate the effectiveness of the proposed FDP-FCM algorithm, it is compared with traditional FCM and DSFCM algorithms in terms of clustering performance and robustness. Regression analysis is employed to conduct empirical research on the factors influencing talent cultivation. Finally, a talent cultivation model based on user profiles is proposed.

II. Talent profiling based on data mining

II. A. Overview of Talent Profiles

Talent profiling is an intuitive representation of modeling and analyzing talent information data, and is a relatively common business issue. Essentially, it involves integrating individual, multi-source, heterogeneous talent information to form unique tags for each talent. These tags are derived from the analysis and mining of massive amounts of talent data, thereby forming a profile for each talent, which is then closely linked to actual business applications. Talent profiling technology involves collecting, extracting, and analyzing professional domain information about talent, then visualizing the feature data. This paper analyzes the professional domain types in which talent excels by studying their basic information, educational background, work experience, published papers and monographs, obtained patents, researched projects, and received awards and honors to analyze the types of professional fields in which talent excels. Additionally, talent models are classified by theme type, common features are extracted, and statistical methods are used for supplementary description to create talent profiles.

Talent profiling plays a crucial role in the comprehensive strength and development direction of enterprises, society, and the country. Through modeling and analysis of various aspects of talent information, talent profiling can help enterprises and institutions achieve precise talent positioning. Furthermore, based on this, combined with the currently popular “Internet+” concept, a personalized talent management system can be built, which not only helps relevant units identify talent but also helps talent recognize their current capabilities. This not only saves human resource costs for both the units and the talent but also rapidly enhances the self-awareness of the talent.

II. B. Scrapy Framework

To build a talent profile, it is necessary to obtain occupational requirements in relevant professional fields. Therefore, crawling tools are needed to extract relevant data from public academic information platforms and public talent recruitment platforms. This paper selected a web scraping tool based on the Scrapy framework. The Scrapy application framework is based on the Twisterd asynchronous processing framework [18], which can extract unstructured data from web pages and convert it into structured data during the web scraping process. The Scrapy framework is shown in Figure 1.

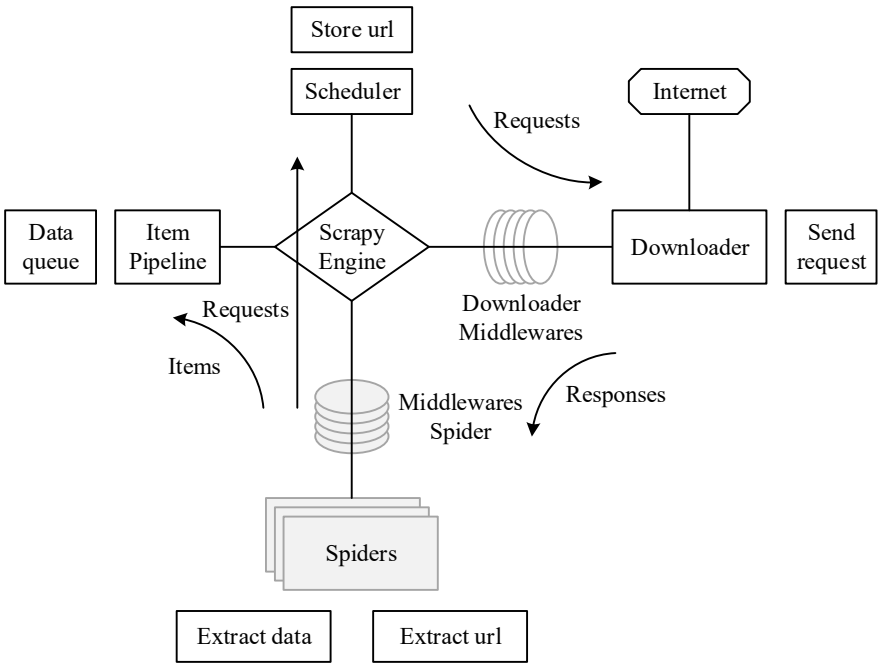

Figure 1: Scrapy framework

When starting a Scrapy project, spiders provide the initial URLs to be crawled, which are then queued in the scheduler. The scheduler maintains a queue of URLs and dispatches them in a first-in, first-out order. Each URL removed from the queue is downloaded from the web or retrieved as data by the downloader, which then returns the downloaded data to the spiders. The data is then analyzed and sent to the pipeline for further processing. If there are other URLs to download, the spiders extract them and send them to the scheduler, where they are queued and wait. When the URL queue in the scheduler is empty, Scrapy completes its task.

The Scrapy framework has a clear module architecture, low coupling, high scalability, stable crawling speed, and can be flexibly adjusted according to requirements, making it very suitable for data mining, data processing, and other tasks in practical applications.

II. C. Text mining technology

II. C. 1) Part-of-speech tagging technology

Word segmentation is an important component of natural language processing and serves as the foundation for many tasks.

Jieba word segmentation is a representative tool based on statistical word segmentation. Jieba word segmentation technology supports three types of segmentation modes: precise mode, full mode, and search engine mode. In precise mode, Chinese sentences can be segmented with the highest accuracy, making it suitable for text analysis. Full mode can quickly scan Chinese sentences to identify words that can form terms, but this method does not effectively address semantic ambiguity issues. The search engine model is based on the precise mode, re-segmenting long words to improve retrieval efficiency, making it suitable for search engine word segmentation. Additionally, jieba word segmentation supports traditional Chinese characters and custom dictionaries to accommodate newly added words. While jieba can recognize new words, adding new words manually can further improve its accuracy.

For unregistered words, jieba word segmentation uses a hidden Markov model (HMM) based on the word formation ability of Chinese characters. This model uses the Viterbi algorithm [19], which generally takes text sequence data as input and outputs the corresponding hidden sequence. The joint probability distribution formula of the HMM model can be expressed as:

$$P(x_1, x_2, ..., x_n) = p(x_1)p(y_1 \mid x_1)\prod_{t=2}^{n} p(y_t \mid y_{t-1})p(x_t \mid y_t) \tag{1}$$

II. C. 2) Feature Selection

The feature selection process mainly consists of four steps: generation process, evaluation function, termination conditions, and verification. The generation process produces a subset of candidate features, the evaluation function evaluates each feature subset, the termination conditions determine when feature selection should end, and the verification step confirms whether the final feature subset is valid.

Currently, commonly used feature selection methods include mutual information, expected cross-entropy, information gain, square root fitting, and TF-IDF. The jieba word segmentation tool used in the preceding text can be combined with the TF-IDF algorithm to extract keywords. TF-IDF is a common weight-based method [20] widely applied in information retrieval and text mining. TF represents the frequency of a word, indicating how many times a keyword appears in a document. IDF stands for inverse document frequency, calculated by dividing the total number of documents by the number of documents containing the word, then taking the logarithm of the result. The formula used is as follows:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{2}$$

Among them, $n_{i,j}$ indicates the frequency of word i appearing in document d_j, and $\sum_k n_{k,j}$ indicates the total number of words in document d_j.

$$IDF_i = log\frac{|D|}{|\{j : n_i \in d_j\}| + 1} \tag{3}$$

In this formula, the numerator represents the total number of documents, and the denominator represents the number of documents containing word n_i. Adding 1 to the denominator is mainly to prevent the denominator from becoming 0.

The TF-IDF formula is expressed as follows:

$$TFIDF = TF * IDF \tag{4}$$

The TF-IDF method can be used to build talent profiles, which not only improves the accuracy of talent tag matching, but also reduces the complexity of data computation due to the reduction in dimensions.

II. C. 3) Text Classification

Text classification is essentially the process of using computer learning classification methods to automatically categorize data texts according to text classification standards. In this paper, the KNN text classification algorithm is used to match talent information data with domain labels. The algorithm works by statistically analyzing training samples to identify the K samples most similar to the sample to be classified. The classification type of the sample to be classified is determined based on the weighting of similar sample types and the similarity of the sample to be classified with the text.

II. C. 4) Clustering Algorithms

Cluster analysis is one of the more important algorithms in text mining. Cluster analysis does not require prior knowledge of the categories of the samples, nor does it require knowledge of the number of categories, making it a typical unsupervised learning algorithm. Cluster analysis can group data together based on their similarity, with high similarity within the same category and significant differences between different categories. It is commonly used in statistical data analysis techniques.

Currently, there are numerous types of clustering analysis algorithms, which can primarily be categorized into the following groups: model-based methods, partition-based methods, density-based methods, hierarchical methods, and grid-based methods.

II. D.Big Data Mining and Feature Extraction of Talent Tags

II. D. 1) Big Data Mining of Talent Tags

To obtain the edge contour features of talent user profiles in big data, a comprehensive input model combining internal and external factors is adopted to establish a big data statistical analysis model for talent tags. Fuzzy decision-making methods are employed, where $A=\{(a_1,a_2,\cdots,a_n)\}$ represents the data sample set of talent tags in business scenarios, with n denoting the number of samples, $\{A_1,A_2,\cdots,A_c\}$ denotes the label category corresponding to the sample, and the membership degree of sample a_i for category A_k is denoted as $u_k(a_i)$, abbreviated as u_{ik}. In the talent behavior space, the fuzzy clustering statistical feature information of business scenario talent labels is obtained based on the associative characteristics of A, leading to the following two-dimensional statistical information distribution for business scenario talent label identification:

$$H_b=\sum_{i=1}^{n}\sum_{k=1}^{c}(u_{ik})^b(d_{ik})^2 \tag{5}$$

Among these, d_{ik} represents the Euclidean distance, used to measure the distance between the i nd sample a_i and the k th cluster center point; b is the category attribute parameter of the talent label at the decision-making moment. If $b=0$, it indicates that the k th category of business rejects the label attribute of this type of talent. If $b=1$, it indicates the k th category of talent mobility and information access. Through the comprehensive input of internal and external factors, the a th category corresponds to the talent label attribute set of the A th category. In the talent label state behavior data distribution space A_x, the results of human resource basic data mining can be represented as:

$$\mu_{ik}=\sqrt{\sum_{i=1}^{n}(a_i-d_{ik})^2} \tag{6}$$

Under the constraint of maximizing talent efficiency, based on various human resource business application scenarios such as job-talent matching, selection, cultivation, utilization, and retention, the return value under specific state $a\in A$ and specific human resource business application behavior A_k is obtained as the return function $r(a_i,A_k)$. For the evolutionary distribution of talent tag attributes, the maximum return function is obtained as:

$$r(a_i,A_k)=\sum_{k=1}^{c}\times\sum_{i=1}^{n}\left(\frac{\mu_{ik}}{w_{ik}}\right)^{\frac{b-1}{2}}/1 \tag{7}$$

Among them, w_{ik} are used to describe the benefit weights of application tags in specific business scenarios. Based on the above analysis, a talent tag big data mining model is obtained, which combines fuzzy state evolution clustering and feature recognition to achieve talent tag big data mining and user profiling. Based on the results of edge contour feature extraction, the accurate identification ability of talent tags is improved.

II. D. 2) Extraction of talent tag attribute characteristics

To optimize the identification of talent tags, based on the results of big data mining of talent tags, combined with user profile feature analysis and big data evolutionary clustering analysis methods, a comprehensive feature analysis model for talent tags in specific business scenarios is constructed to extract the relevant feature information of talent tags in specific business scenarios. Assuming the maximum interference of the main user is T_n, under the condition of maximizing talent benefits, the feature state distribution $x \in X$ of talent tags in business scenarios is obtained. In specific business scenarios, the behavioral state distribution feature components of talent tag recognition can be expressed as:

$$v_{ik} = \frac{\sum_{i=1}^{n} (u_{ik})^b x_i}{\sum_{i=1}^{n} (\mu_{ik})^b T_n}, k = 1, 2, \cdots, c \tag{8}$$

Among them, x_i is the talent tag attribute status of the i nd sample. Then, the minimum conditional probability density function for transferring from talent tag attribute status x to status y can be expressed as:

$$y = \exp\left[-\frac{(x - x_i)^T (x - x_i)}{2\sigma^2} \right] \tag{9}$$

Among these, σ represents the decision-making model for each type of talent attribute. Based on professional human resources and psychological methodologies, a label fusion feature clustering analysis method is employed. Let $P_n^*, P_{vh}^*, P_{hh}^*$ represent the maximum allowable values for talent demand under market constraints in various human resources business application scenarios. Through fuzzy optimization, the correlation statistical feature values for various human resources business application scenarios are obtained as O_n, O_{vh} and O_{hh}, respectively representing the priority of talent label feature scheduling under three different label attributes. Based on the priority of talent label scheduling in various human resources business application scenarios, we perform feature extraction and association analysis of talent labels in different application scenarios, obtaining the size relationship of association distribution items as $w_{hh} \geq w_{vh} \geq w_n$. Under these circumstances, the probability density feature distributions are $P_{C_1}^* \leq P_{C_2}^* \leq P_{C_K}^*$ and $w_{C_1} \geq w_{C_2} \geq w_{C_k}$. By combining specific business scenario requirements with basic label identification, we obtain feature weights η_{ik} that satisfy an exponential distribution. Thus, the maximum utility of talent labels can be approximately expressed as:

$$S_i = \sum_{k=1}^{c} \sum_{i=1}^{n} y \eta_{ik} \tag{10}$$

Under balanced game control, the output of talent tag attribute feature mining is:

$$f(x, y) = \frac{1}{v_{ik} (2\pi)^{\frac{\eta_{ik}+1}{2}} \sigma^{\eta_{ik}+1}} \exp\left[-\frac{(x - x_i)^T (x - x_i)}{2\sigma^2} \right] \tag{11}$$

Based on the above analysis, a three-level tagging system for talent resources was established, including basic tags, feature tags, and high-level tags. Edge contour feature extraction and pixel fusion methods were used to generate tags and cluster attributes.

II. E. Talent Profile Tag Clustering Method

In order to cluster talent profile tags, this paper introduces the density peak clustering algorithm based on the FCM clustering algorithm and proposes the FDP-FCM algorithm to achieve the best results in talent profile tag clustering.

II. E. 1) FCM Clustering Algorithm

The core idea of the FCM algorithm is an improvement on traditional hard clustering [21], which is based on finding the minimum value of the objective function J:

$$J = \sum_{i=1}^{N} \sum_{j=1}^{C} (u_{ij})^m x_i - c_j^2, 1 \leq m < \infty \tag{12}$$

where m is the fuzzy coefficient, c_j is the j cluster center point, and $x_i - c_j$ represents the distance between x_i and c_j, which is generally calculated using Euclidean distance.

The specific FCM algorithm is as follows.

Step 1: Initialize parameters, given the number of clusters c, initialize the fuzzy coefficient m, and allowable error ε.

Step 2: Initialize the membership matrix using a random method, but it must comply with the normalization rule, i.e., the sum of the memberships of a data set is 1:

$$\sum_{i=1}^{c} u_{ij} = 1, \forall j = 1, 2, \cdots, n \tag{13}$$

Step 3: Calculate the fuzzy cluster centers using formula (14):

$$c_j = \frac{\sum_{i=1}^{n} u_{ij}^m x_i}{\sum_{i=1}^{n} u_{ij}^m} \tag{14}$$

Step 4 Update the membership matrix using formula (15):

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\| x_i - c_j \|}{\| x_i - c_j \|} \right)^{2/(m-1)}} \tag{15}$$

Step 5: If the objective function $J(U,V) < \varepsilon$ or the maximum number of generations is reached, stop; otherwise, jump to step 3.

When the FCM clustering algorithm reaches the iteration termination condition, it is considered to have converged. During each iteration of the algorithm, the process proceeds in the direction of minimizing the objective function value. However, the objective function may have multiple extrema, and the initial clustering centers may only be near one of these extrema. This can lead the algorithm to converge to a local optimum, making the FCM clustering algorithm sensitive to the initial clustering centers. Additionally, the FCM clustering algorithm requires a specified number of cluster centers, which is often difficult to determine precisely because the data to be clustered is typically unordered, making it challenging to determine the number of cluster centers. Furthermore, while the FCM clustering algorithm considers the aggregation relationship between data and cluster centers, it ignores the associations between data points, specifically the density around data points and the distances between them. Based on the above principles, this paper proposes an improved FCM algorithm—the FDP-FCM algorithm.

II. E. 2) Density Peak Clustering Algorithm

The CFSFDP clustering algorithm is a density-based clustering method [22]. The density peak clustering algorithm is based on the assumption that, for a given dataset, cluster centers are surrounded by neighboring points with lower local density, and these points with lower local density are relatively distant from other points with higher local density. For each data point i, the density peak clustering algorithm must calculate two key parameters: local density ρ_i and the distance δ_i between high-density points.

where the local density ρ_i of data point i is defined as follows:

$$\rho_i = \sum_j \chi(d_{ij} - d_c) \tag{16}$$

Among them, $\chi(x) = \begin{cases} 1, x < 0 \\ 0, x \geq 0 \end{cases}$ and d_c are the cutoff distances. For large amounts of data, the local density is essentially the relative density between data points, so the choice of d_c is robust. The δ_i of data point i is the minimum distance from the point to any point with a density greater than it, defined as follows:

$$\delta_i = \min_{j:\rho_j > \rho_i} (d_{ij}) \tag{17}$$

The point with the highest local density requires special treatment, and its value is generally set to $\delta_i = \max_j (d_{ij})$.

The specific algorithm for CFSFDP clustering is as follows.

Step 1: Initialize and calculate the distance matrix $D = \{d_{ij}\}, i, j = 1, 2, \cdots, n$ between each data point, and determine the cutoff distance d_c, which provides the basis for subsequent calculations of local density and high-density distance.

Step 2: Use formulas (16) and (17) to calculate the local density ρ_i and high density distance δ_i for each data point i.

Step 3: Construct a decision graph with ρ as the horizontal axis and δ as the vertical axis. Based on the decision graph, manually select the density peak points in the upper right corner of the graph that have high local density ρ and high density distance δ and are clearly far away from most of the samples as the cluster centers.

Step 4: Calculate the minimum distance between each point in the dataset and each cluster center, and finally assign each data point to the cluster center closest to it.

The CFSFDP clustering algorithm can effectively determine the density-based clustering centers and the number of clusters, but it has certain limitations. The CFSFDP clustering algorithm employs a decision tree-based heuristic method that manually selects clusters based on the expected high local density ρ and high density distance δ. However, in some cases, a single cluster may contain multiple density peaks, and the density peak clustering algorithm treats different density peaks as potential cluster centers, making it difficult to precisely determine the number of clusters when selecting cluster centers. To effectively select cluster centers on decision graphs, users should be experts in the domain of the underlying dataset. However, this clearly imposes certain limitations on the application of the algorithm.

II. E. 3) Fuzzy C-means algorithm based on density peak optimization

To address the shortcomings of the FCM clustering algorithm mentioned above, such as the need for manual specification of the number of clusters, slow convergence speed, and susceptibility to local optima, this paper combines the density peak clustering algorithm to propose an improved fuzzy C-means clustering algorithm—FDP-FCM.

(1) Initial selection of cluster centers

The principle for selecting cluster centers is to ensure they accurately reflect the overall density of the dataset, thereby avoiding the selection of noise data points as cluster centers, which could lead to iterative adjustments and local convergence. FDP-FCM utilizes the local density and high-density distance proposed by the density peak clustering algorithm, with improvements to effectively determine cluster centers and demonstrate robust performance on large datasets.

Therefore, to better select initial cluster centers and accurately determine the number of clusters, the Fuzzy-CFSFDP clustering algorithm is optimized to obtain initial cluster centers and determine the number of clusters. The optimization of the Fuzzy-CFSFDP clustering algorithm over the CFSFDP algorithm is based on the following formula:

$$EC_i = (\delta_i) \geq 2\sigma(\delta_i) \tag{18}$$

Among these, EC_i represents the expected cluster center, while $\sigma(\delta_i)$ is the standard deviation of all high-density distances calculated using formula (17). According to the density peak clustering algorithm, the cluster center has a significant clustering relationship with other cluster centers; therefore, the high-density distances of other points within the cluster should be less than $2\sigma(\delta_i)$. However, the density peak clustering algorithm may also produce larger δ values and noise cluster centers with low density. Therefore, this noise should be separated from the desired cluster centers using the following formula:

$$LC_i = EC_i \geq \mu(\rho_i) \tag{19}$$

Among these, LC_i is the local cluster center after removing the noise cluster center, while $\mu(\rho_i)$ is the mean of ρ_i. In this way, the local cluster center has higher density and greater high-density distance compared to neighboring data points.

After determining the local cluster centers using the above two formulas, the local cluster centers are then merged. If the minimum distance between any two local cluster centers is less than the cutoff distance d_c, they are merged into a single cluster center. After the final local cluster centers are merged, the global cluster centers are constructed, and the number of initial cluster centers determined is the number of clusters.

(2) Algorithm acceleration iteration

After determining the cluster centers and the number of clusters, the algorithm iteration begins. The traditional FCM algorithm iterates continuously without considering the influence of local density during the iteration process. The proposed FDP-FCM algorithm takes into account the impact of local density on clustering during iteration, introducing a density-weighted coefficient to accelerate the algorithm's convergence speed. Additionally, to further enhance convergence speed, an oscillation factor λ is incorporated into the iteration process. The final algorithm iteration formula is as follows:

$$u_{ij} = \frac{1+\tau}{\sum_{k=1}^{c}\left(\frac{x_i - c_j}{x_i - c_k}\right)^{2/(m-1)}} \times \lambda + (1-\lambda) \times \tau \tag{20}$$

Among them, τ is the density weighting coefficient, whose value is calculated as follows:

$$\tau = \begin{cases} e^{-(x_i - c_j / d_c)}, x_i - c_j < d_c \\ 0, x_i - c_j \geq d_c \end{cases} \tag{21}$$

After extensive experimentation, the oscillation factor λ has a range of $\lambda \in (0.6,1]$, with a typical value of 0.9, which can be adjusted in specific algorithms.

III. Analysis of research results

III. A. Experimental Analysis

III. A. 1) Initialization of Cluster Center Comparison Analysis

To validate the effectiveness of the CFSFDP algorithm, the CFSFDP algorithm was used to select the initial cluster centers for the Iris dataset, and the results were compared with the actual cluster centers of the Iris dataset. The results are shown in Table 1.

As shown in Table 1, the results of the CFSFDP algorithm for initializing the cluster centers of the Iris dataset are very close to the actual cluster centers of the Iris dataset, with errors at the percentage level. This demonstrates that the CFSFDP algorithm can optimize the results of initial cluster center selection, laying a solid foundation for subsequent FCM algorithm clustering and validating the effectiveness of the FDP-FCM algorithm.

Table 1: Initial clustering center comparison

CFSFDP initial Iris clustering center result	Iris real clustering center
(6.2016 4.2328 1.7793 0.7476)	(6.20 4.23 1.78 0.75)
(6.8406 3.5609 5.1544 2.0296)	(6.84 3.56 5.15 2.03)
(7.1218 3.2378 5.9536 2.8641)	(7.12 3.24 5.95 2.86)

III. A. 2) Clustering Effect and Robustness Comparison Analysis

To validate the clustering performance and robustness of the FCM, DSFCM, and FDP-FCM algorithms, this paper conducts simulation comparison experiments using the Wine dataset from the UCI database. Like the Iris dataset, the Wine dataset is a commonly used dataset in clustering algorithm experiments. The F-measure metric, Rand Index (RI), and Jaccard coefficient were used to evaluate the clustering performance of the three algorithms on the Wine dataset. Higher values for these metrics indicate better clustering performance.

The FCM algorithm, DSFCM algorithm, and the FDP-FCM algorithm proposed in this paper were used to perform clustering on the Wine dataset. The clustering results of the three algorithms were evaluated using the aforementioned clustering performance metrics, and the results are shown in Table 2. As can be seen from the table: Under the F-measure metric, the FCM algorithm has a slightly higher value than the DSFCM algorithm in the Class 2 category, while in all other categories, FDP-FCM > DSFCM > FCM. Under the RI index, the FCM algorithm has a higher value than the DSFCM and KDPC-FCM algorithms in the Class 3 category, while in all other categories, FDP-FCM > DSFCM > FCM. Under the Jaccard coefficient, all three categories follow the order FDP-FCM > DSFCM > FCM. Overall, under the Wine dataset, all metrics follow the order FDP-FCM > DSFCM > FCM, indicating that the FDP-FCM algorithm outperforms the DSFCM and FCM algorithms in terms of clustering performance on the Wine dataset.

Table 2: Clustering evaluation index comparison of clustering algorithms in Wine dataset (%)

		Class 1	Class 2	Class 3	Mean
F-measure	FCM	83.65	72.15	55.21	70.34
	DSFCM	87.45	69.44	57.86	71.58
	FDP-FCM	94.23	78.13	63.17	78.51
RI	FCM	85.47	60.34	66.42	70.74
	DSFCM	88.49	66.18	62.73	72.47
	FDP-FCM	94.26	74.92	64.86	78.01
Jaccard	FCM	82.74	68.45	55.63	68.94
	DSFCM	84.27	71.39	58.41	71.36
	FDP-FCM	90.48	74.64	62.32	75.81

To evaluate the robustness of the three algorithms, the following method was employed: each algorithm was run 100 times on the Wine dataset, and the standard deviation of each algorithm was calculated for the three clustering

evaluation metrics. A smaller standard deviation indicates lower variability, greater stability, and stronger robustness. After conducting the experiments and calculating the standard deviation for each algorithm, the standard deviation results for the three comparison algorithms across the three different clustering evaluation metrics are shown in Table 3.

As shown in Table 3, the standard deviation of the FDP-FCM algorithm is smaller than that of the other two algorithms across all three evaluation metrics.

Table 3: Standard error comparison in evaluation index

Algorithm	F-measure	RI	Jaccard
FCM	0.1325	0.1065	0.1005
DSFCM	0.1106	0.0824	0.0568
FDP-FCM	0.0824	0.0743	0.0389

In summary, we can draw the following preliminary conclusions: Among the three comparison algorithms, the FCM algorithm has the poorest clustering performance and robustness, followed by the DSFCM algorithm, while the FDP-FCM algorithm developed in this paper has the strongest clustering performance and robustness.

III. B. Empirical Analysis

III. B. 1) Common method bias test

Considering that this study draws from the same data source, this paper employs Harman's single-factor analysis to test for common method bias. Through unrotated principal component analysis, the results show that the variance explained by the first factor is 14.52%, which is far below the critical threshold of 40%. Therefore, this study does not exhibit severe common method bias.

III. B. 2) Descriptive statistical analysis

The descriptive statistical results for each research variable are shown in Table 4. Observing Table 4, it can be seen that social support is significantly positively correlated with job-seeking mindset, academic performance is significantly positively correlated with job-seeking preparation, and job-seeking mindset is significantly positively correlated with reading type, social support, job-seeking preparation, and job-seeking flexibility.

Table 4: Descriptive statistical results of each variable

Variable	M	SD	Social support	Academic grade	Extracurricular performance	Job-hunting mentality	Job-hunting preparation	Flexibility of job-hunting
Province	1.62	0.75	-0.035	-0.425**	0.312**	0.046	-0.156*	0.077
Political appearance	2.13	0.65	-0.015	-0.362**	-0.232**	-0.123*	-0.174*	0.008
Family rank	2.62	1.28	-0.079	-0.077	0.105	0.096	0.006	0.144*
Funding object or not	1.94	0.52	0.163**	0.063	-0.175*	0.012	0.003	-0.093
Living expense	2.08	0.56	0.121*	0.032	-0.038	0.125	0.102	-0.052
Student cadre or not	1.72	0.46	-0.053	0.122	0.264**	0.079	0.132*	-0.071
Industry	1.41	0.48	-0.092	-0.178	-0.054	-0.182	-0.112	-0.079
Considering factor	1.65	0.79	-0.081	-0.179	--0.061	-0.068	-0.188*	-0.128*
Interest type	3.16	1.86	0.054	0.044	0.003	-0.054	-0.046	-0.041
Commonly used APP	1.99	0.77	-0.072	-0.082	0.033	-0.123*	-0.127*	-0.002
Reading type	6.97	1.45	1.000**	0.079	0.076	0.142*	0.045	0.096
Social support	6.95	1.49	1	0.079	0.076	0.142*	0.045	0.096
Academic grade	7.88	1.39		1	-0.026	0.36	0.138*	-0.062
Extracurricular performance	6.11	2.62			1	0.103	-0.002	0.071
Job-hunting mentality	8.74	1.06				1	0.368**	0.267**
Job-hunting preparation	6.91	1.04					1	0.186**
Job-hunting flexibility	9.27	1.62						1

III. B. 3) Difference test

Following t-tests and analysis of variance, significant differences were found in certain variables. In terms of family economic circumstances, students who were recipients of financial aid scored higher on extracurricular performance (t=4.854*) and job-seeking flexibility (t=7.058**) than those who were not recipients. Regarding whether students held student leadership positions, those who did scored higher on extracurricular performance (t=35.062***) than those who did not. In terms of intended industry, students planning to work in their current field had higher job-seeking mindset scores (t=5.847*) and job-seeking flexibility scores (t=7.934**) than those seeking jobs in other industries. In terms of provincial type, students from eastern and central provinces had significantly higher academic performance scores (F=31.862***) than those from western provinces. Students from eastern provinces had significantly higher job preparation scores (F=3.546*) than those from western provinces, while students from central and western provinces had higher extracurricular performance scores (F=15.425***) than those from eastern provinces. In terms of family ranking, students from multi-child families who are in the middle of the family hierarchy have higher extracurricular performance scores (F=4.857**) than students in other family rankings. Only children have lower job flexibility scores (F=3.662*) than students from multi-child families. In terms of primary considerations for job-seeking, students who prioritize job location have higher academic performance (F=3.256*) than those who prioritize benefits and compensation, and are better prepared for job hunting (F=3.885*). In terms of commonly used apps, students who frequently use learning interest-related apps are the most prepared for job hunting (F=6.849**). In terms of extracurricular reading types, students who read books receive more social support (F=5.745**) than those who do not read books. Students with a broad range of reading interests have the most optimistic job-seeking mindset (F=3.766*) and are better prepared for job-seeking (F=5.941**).

III. B. 4) Regression Analysis

In further regression analysis, the sample data were divided into two categories based on students' primary goals at graduation: employment-oriented (N=182) and further education-oriented (N=150). The achievement assessment values of their goals one month after graduation were used as the dependent variable for stratified regression analysis.

For graduates whose primary goal was employment, the results of the stratified regression analysis indicated that provincial type had a predictive effect on goal achievement, while academic performance and job search preparation had more pronounced predictive effects. The regression analysis of goal achievement for graduates whose primary goal was employment is shown in Table 5.

Table 5: Regression analysis result of employment goal achievement

Variable		Step 1 (β)	Step 2 (β)	Step 3 (β)	Step 4 (β)
Step 1	Province	-0.425*	-0.415	-0.263	-0.216
	Family rank	-0.048	-0.033	-0.016	0.018
	Funding object or not	-0.705	-0.704	-0.612	-0.674
	Student cadre or not	0.384	0.365	0.168	0.093
	Industry	-0.306	-0.386	-0.293	-0.294
	Considering factor	-0.294	-0.312	-0.241	-0.175
	Commonly used APP	0.295	0.194	0.188	0.175
	Reading type	-0.284	-0.226	-0.184	-0.288
Step 2	Social support		-0.052	-0.048	-0.021
	Job-hunting mentality		-0.158	-0.206	-0.316
	Job-hunting flexibility		-0.096	-0.098	-0.132
Step 3	Academic grade			0.406**	0.376**
	Extracurricular performance			0.144	0.136
Step 4	Job-hunting preparation				0.411**
$R^2(\Delta R^2)$		0.098 (0.042)	0.132 (0.044)	0.182 (0.093)*	0.211 (0.126)**

The primary target group is graduates aiming for further education. The results of the hierarchical regression analysis indicate that provincial type has a predictive effect on goal attainment, while academic performance has a more pronounced predictive effect. The regression analysis of goal attainment for graduates whose primary objective is further education is shown in Table 6.

From the research results, it can be seen that for college graduates, in terms of key indicators influencing the achievement of employment and further education goals, students from eastern and central provinces perform

better than those from western provinces in terms of academic performance. Students who prioritize work location perform better than those who prioritize benefits and compensation. This is related to the students' place of origin. Regarding differences in preferred work location, further analysis reveals that many students tend to return to their hometowns for work, prioritizing family and stability, and thus are more likely to perform better academically to meet family expectations. Regarding job preparation, students who frequently use learning interest-related apps and read more extracurricular books are better prepared. This is related to students' daily behaviors. Allocating time spent online and during extracurricular hours to learning and cultivating interests is significantly more beneficial for job hunting and college preparation than purely for entertainment.

Table 6: Regression analysis result of further-study goal achievement

Variable		Step 1 (β)	Step 2 (β)	Step 3 (β)	Step 4 (β)
Step 1	Province	-0.989**	-0.893*	-0.586	-0.586
	Family rank	0.326	0.312	0.307	0.307
	Funding object or not	-0.496	-0.274	0.078	0.078
	Student cadre or not	0.402	0.495	0.182	0.182
	Industry	0.142	0.052	0.672	0.672
	Considering factor	-0.089	-0.198	0.069	0.069
	Commonly used APP	-0.225	-0.178	-0.046	-0.046
	Reading type	-0.284	-0.043	-0.021	-0.021
Step 2	Social support		-0.285	-0.082	-0.082
	Job-hunting mentality		0.274	-0.362	-0.362
	Job-hunting flexibility			0.274	0.274
Step 3	Academic grade			0.625**	0.625**
	Extracurricular performance			0.008	0.008
Step 4	Job-hunting preparation				0.000
$R^2(\Delta R^2)$		0.098 (0.038)	0.144 (0.044)	0.188 (0.089)*	0.188 (0.083)

IV. Building a talent cultivation model based on traditional cultural innovation

From the perspective of traditional cultural innovation and talent cultivation, the value of user profiling lies in its ability to generate personalized profiles (portraits) based on the academic circumstances of different talents. These profiles accurately reflect an individual's knowledge, skill levels, learning experiences, internship experiences, practical experience, and other data. For enterprises, talent user profiles can serve as a standard for evaluating the comprehensive capabilities of talent, enabling accurate matching between talent and job positions. For educational institutions, talent user profiles can serve as the basis for precise talent cultivation, enabling tailored education, dynamic adjustment of talent cultivation programs, and adaptation to the new requirements posed by changes in new occupations. For talent themselves, talent user profiles can reflect their strengths and weaknesses, clarifying their development direction. By innovating traditional culture, we connect enterprises, institutions, and talent, integrate enterprise job talent demand data and talent growth data, construct talent cultivation models, generate talent user profiles, and achieve perfect alignment between schools, talent, and enterprises.

(1) Connecting enterprises and constructing job competency models

Companies are at one end of the industrial chain and are the employers, with needs for flexible staffing and precise talent. Therefore, talent development must be based on the competency standards derived from company job roles. To obtain these competency standards, it is necessary to conduct a detailed analysis of e-commerce company job roles, identify typical tasks within those roles, and then establish competency models around those roles. After constructing the models, competency assessments are used to evaluate talent levels, match them with job role requirements, and provide precise feedback on whether the talent meets the company's needs.

(2) Connecting institutions and establishing intelligent adaptive learning paths

Institutions are at one end of the education chain, with talent cultivation as their primary objective. However, institutions find it difficult to obtain job talent standards and internal training resources from e-commerce companies. Additionally, companies are profit-driven and unwilling to invest significant resources in assisting institutions with talent cultivation, severely limiting institutions' ability to develop talent. Therefore, a third-party platform is needed to connect institutions with companies, integrating corporate hiring standards and job competency models into talent

development. This allows for flexible adjustments to the training direction of different types of talent, precise education, and the adoption of optimal learning methods to enhance the efficiency of talent cultivation.

(3) Connecting talent and creating digital talent profiles

The talent chain records talent's learning trajectories and growth data, connecting e-commerce companies and educational institutions through data, and ultimately providing feedback through digital talent profiles. Digital profiles record and consolidate all content included in corporate job competency models and student learning paths, and are continuously updated based on talent development directions and capabilities, forming user profiles specific to individual talents, i.e., talent data profiles.

V. Conclusion

This paper builds a Scrapy framework to perform data mining and feature extraction on talent tags, proposes the FDP-FCM algorithm to address the shortcomings of the FCM clustering algorithm, verifies the effectiveness of the FDP-FCM algorithm, and conducts empirical research to classify talent attributes and explore factors related to talent cultivation.

The F-measure, RI, and Jaccard values of the FDP-FCM algorithm in this paper are all superior to those of the FCM algorithm and DSFCM algorithm, demonstrating better clustering performance and robustness. Students receiving financial aid exhibit higher levels of extracurricular performance and job-seeking flexibility compared to those not receiving aid. Students preparing to enter the same industry demonstrate higher job-seeking mindset and flexibility than those seeking employment in different industries. Students from eastern provinces demonstrate significantly higher job preparation levels than those from western provinces. Only children exhibit lower job flexibility than students from families with multiple children. Students with a broad reading background exhibit the most optimistic job-seeking mindset and more thorough job preparation. Provincial type has a predictive effect on graduates' goal achievement. Academic performance and job preparation have a more pronounced predictive effect on graduates whose primary goal is employment. Academic performance has a more pronounced predictive effect on graduates whose primary goal is further education.

References

[1] Henriksen, D., Mishra, P., & Fisser, P. (2016). Infusing creativity and technology in 21st century education: A systemic view for change. Journal of Educational Technology & Society, 19(3), 27-37.

[2] Raja, R., & Nagasubramani, P. C. (2018). Impact of modern technology in education. Journal of applied and advanced research, 3(1), 33-35.

[3] Hamidi, H., & Chavoshi, A. (2018). Analysis of the essential factors for the adoption of mobile learning in higher education: A case study of students of the University of Technology. Telematics and Informatics, 35(4), 1053-1070.

[4] Serdyukov, P. (2017). Innovation in education: what works, what doesn't, and what to do about it?. Journal of research in innovative teaching & learning, 10(1), 4-33.

[5] Subotnik, R. F., Olszewski-Kubilius, P., & Worrell, F. C. (2021). Unlocking Creative Productivity: A Talent Development Approach. Journal of Modern Foreign Psychology, 10(4), 17-32.

[6] Chen, S., & Lin, N. (2021). Culture, productivity and competitiveness: disentangling the concepts. Cross Cultural & Strategic Management, 28(1), 52-75.

[7] Han, K., & Zhang, L. (2021, December). Exploration on the path of cultivating innovative talents under the background of intelligent era. In 2021 International Conference on Forthcoming Networks and Sustainability in AIoT Era (FoNeS-AIoT) (pp. 270-274). IEEE.

[8] Hidayat, M. C., & Arifin, S. (2020, May). Integration Science Technology with Islamic Values: Empowering Education Model. In 1st Borobudur International Symposium on Humanities, Economics and Social Sciences (BIS-HESS 2019) (pp. 966-970). Atlantis Press.

[9] Sanusi, A. M., & Puteh, S. (2017). An Approach of Excellence Talent in Engineering Education Programme of Enhancing the Quality of Students. Advanced Science Letters, 23(2), 1109-1112.

[10] Djiraro Mangue, C. L., & Gonondo, J. (2021). Academic Culture and Talent Cultivation: The Chinese Experience. Journal of Comparative & International Higher Education, 13.

[11] Xie, F., Jiang, N., & Kuang, X. (2025). Towards an accurate understanding of 'new quality productive forces'. Economic and Political Studies, 13(1), 1-15.

[12] Gukalenko, O., Borisenkov, V., Kuznetsov, V., Panova, L., & Tkach, L. (2021). Technological effectiveness of modern education: features, traditions, innovations. In E3S Web of Conferences (Vol. 273, p. 12074). EDP Sciences.

[13] Qiu, Y., Lu, W., Guo, X., & Yen, T. T. (2025). Cultivation of Innovative Marketing Talents in Universities under the New Quality Productivity: A Review. Asian Journal of Education and Social Studies, 51(1), 253-261.

[14] Cheng, X., Tang, Z., & Yao, Y. (2024, September). Research on Digitally Empowering Vocational Education with the Integration of Labor and Creativity in the Perspective of New Quality Productivity. In 2024 14th International Conference on Information Technology in Medicine and Education (ITME) (pp. 686-688). IEEE.

[15] Liu, L., Si, S., & Li, J. (2023). Research on the effect of regional talent allocation on high-quality economic development—Based on the perspective of innovation-driven growth. Sustainability, 15(7), 6315.

[16] Jing'ai, L., & Weiqing, L. (2019). Research on talent training model of new applied undergraduate colleges. International Journal of Information and Education Technology, 9(9), 652-660.

[17] Frankiewicz, B., & Chamorro-Premuzic, T. (2020). Digital transformation is about talent, not technology. Harvard business review, 6(3), 1-6.

[18] Deng Kaiying,Chen Senpeng & Deng Jingwei. (2020). On optimisation of web crawler system on Scrapy framework. International Journal of Wireless and Mobile Computing,18(4).
[19] Xiaofeng Jiang,Xiude Guo,Hui Feng & Fangling Ren. (2024). Discrete Dynamic Modeling Analysis of Badminton Games Based on Viterbi Algorithm in College Badminton Physical Education. International Journal of High Speed Electronics and Systems,34(02).
[20] Ali Alammary & Saeed Masoud. (2025). Towards Smarter Assessments: Enhancing Bloom's Taxonomy Classification with a Bayesian-Optimized Ensemble Model Using Deep Learning and TF-IDF Features. Electronics,14(12),2312-2312.
[21] Fanbo Zeng,Xiaojun Rao,Jianhua Lei,Xujie Huo,Yuan Shi & Deng Ai. (2025). Behavioral Correlation-Based Residential Space Modularization Using Design Structure Matrix and Fuzzy C-Means Clustering Algorithm. Buildings,15(4),647-647.
[22] Mengnan Cai,Siye Wangpp,Qinxuan Wu,Yijia Jin & Xinling Shen. (2019). DTCluster: A CFSFDP Improved Algorithm for RFID Trajectory Clustering Under Digital-twin Driven. IEICE Proceeding Series,59,TS1-1.