# Research on Key Technologies of Talent Portrait System Based on Cluster Analysis

Qiong Cao,[1*] Haoyang Li,[2] Muqing Wang,[1] Jie Zhao,[1]
[1]State Grid Shanxi Marketing Service Center, Taiyuan, Shanxi 030032, China
[2]State Grid UHV Transformation Co. of SEPC, Taiyuan, Shanxi 030032, China
E-mail: cqcaoq@outlook.com
[*]Corresponding author

**Keywords:** cluster analysis, talent portrait system, design, realize

**Received:** August 25, 2025

*With the rapid digital transformation of human resources, precise talent management has become a core organizational capability. Traditional assessment methods, limited by subjectivity and single-dimensional evaluation, can no longer meet modern talent management needs. Current clustering approaches face challenges such as high-dimensional sparse data, low efficiency and local optima in algorithms like K-means and original Bi-Kmeans, as well as insufficiently specialized talent tag systems that hinder accurate job matching. This study therefore designs and implements a cluster-analysis-based talent profiling system to improve processing of high-dimensional sparse data and the interpretability of clustering results. Using the Oracle database from Jiangxi Province's Talent Dynamic Management System, datasets of 500, 2,000, and 5,000 records were constructed, each containing 68 feature dimensions (basic information, TF-IDF keywords, LDA topics). The Bi-Kmeans algorithm was improved by integrating Kmeans++ centroid initialization and KD-tree fast nearest-neighbor search, reducing time complexity from O(t·n·d) to O(t·logn·d). A professional talent-label system was built using HanLP segmentation, TF-IDF and LDA, with KNN for label matching. An integrated system covering data cleaning, feature extraction, portrait construction, clustering, and system management was developed. Experiments on the 5,000-record dataset show that the improved Bi-Kmeans achieves 81% clustering purity (20% higher than K-means, 14% higher than original Bi-Kmeans), ARI of 0.75, and 41% faster runtime, while performance variance across 10 runs stays under 5%. The post-processing missing-data rate is under 3%, and KNN label-matching accuracy reaches 92%. Overall, the system operates stably, meets functional requirements, enhances clustering efficiency and stability, enriches the talent-tag dimensions, and provides both theoretical innovation and strong practical value for precise talent management in Jiangxi Province.*

*Povzetek: Raziskava predstavi izboljšan sistem za oblikovanje talentnih profilov, ki z nadgrajenim algoritmom Bi-Kmeans učinkovito obdeluje visoko-dimenzionalne in redke podatke ter zmanjšuje časovno zahtevnost.*

## 1 Introduction

Under the background of digital transformation of human resources, traditional talent assessment methods are difficult to cope with the complex and changeable talent management demands due to their strong subjectivity and single dimension [1-5]. Cluster analysis, as an unsupervised learning method, can uncover the implicit characteristics of talent groups and provide technical support for precise talent profiling [6-7]. However, existing research still faces challenges such as difficulties in processing high-dimensional sparse data and insufficient interpretability of clustering results. Although scholars at home and abroad have conducted a certain degree of research in this aspect, there are still two major problems: First, the clustering algorithm is prone to fall into local optimum and is inefficient when dealing with large-scale talent data; Second, the tag system lacks specificity in professional fields and is difficult to support precise job matching. In response to the above problems, this study proposes a multi-dimensional optimization scheme: At the algorithm level, an improved Bi-Kmeans clustering algorithm is designed (integrating the centroid initialization strategy of Kmeans++ and the efficient search ability of KD trees) to solve the problems of high computational complexity and unstable results of traditional algorithms; At the feature engineering level, by combining HanLP word segmentation, TF-IDF weight calculation and LDA topic model, a label system covering professional knowledge and skills is constructed, and the precise matching of talent information and labels is achieved through the KNN algorithm. At the system level, a dynamic talent management platform for Jiangxi Province is built based on the Oracle database, integrating data processing, portrait construction and cluster analysis modules to achieve the visualization of talent

characteristics and the intelligent matching of job requirements.

# 2 Relevant theoretical analysis

Talent profiling technology represents an advanced approach to integrating, analyzing and visualizing talent information in complex business scenarios. Although there is no clear definition in the academic circle, this concept originated from the popular technology in today's Internet industry - user profiling, and talent profiling can be regarded as a branch of user profiling in terms of highly educated and technical talents. The core point lies in relying on an individual's real data to build a comprehensive and accurate character model for the individual [8]. Talent profiling technology can play a key role at all levels of society. By conducting in-depth analysis and modeling of talent information, it can help government and enterprise units achieve precise positioning and effective management of talents. Combining current popular data analysis technologies with modern information technology, talent profiling not only helps enterprises and public institutions deeply explore and rationally allocate talents, but also promotes the improvement of talents' self-awareness [9]. By establishing a personalized talent management system and recommendation mechanism, enterprises can more accurately match talents to suitable positions or project teams. This not only saves labor costs but also helps optimize talent allocation, improve work efficiency and comprehensive competitiveness.

# 3 Improve the design of clustering algorithms

Before designing the improved clustering algorithm, two hypotheses are proposed. Compared with the traditional KMeans and the original Bi-KMeans, the clustering purity of the improved Bi-KMeans on 5,000 talent data is increased by≥10%, and the time consumption is reduced by≥30%. Based on the tag system of HanLP+TF-IDF+LDA, the accuracy rate of talent-tag matching in combination with KNN is ≥90%.

## 3.1 Analysis of limitations of traditional algorithms

The Bi-Kmeans clustering algorithm, also known as the binary K-means clustering algorithm, is an improved version of the K-means clustering algorithm [10]. The K-means algorithm is a widely used clustering method. It performs data clustering by iteratively selecting cluster centers and allocating data points to the nearest cluster center, aiming to minimize the sum of the distances from each point to its cluster center. However, a major problem with the K-means algorithm is that it is highly dependent on the selection of the initial cluster centers, which may lead to local optimal solutions rather than global optimal solutions.

## 3.2 Improved Bi-Kmeans clustering algorithm

By integrating with the central idea of the Kmeans++ algorithm, that is, improving the initialization step of the K-means algorithm, and intelligently selecting the initial centroids, the centroids can better represent the data distribution, thereby enhancing the clustering quality and the convergence speed of the algorithm. The fast nearest neighbor search ability of KD-Tree is utilized to optimize the cluster center update process of Bi-Kmeans. KD-Tree can help quickly find the data points around each cluster center, thereby updating the position of the cluster center more effectively. This combination can accelerate the iterative process of the Bi-Kmeans algorithm and improve the efficiency and convergence speed of the algorithm. Finally, in the iterative process of the Bi-Kmeans algorithm, clusters with poor clustering effects are merged, thereby reducing the number of clusters and improving the overall clustering quality. This improvement idea can help prevent the K-means algorithm from falling into local optimal solutions and enhance the stability and robustness of the algorithm. The improved Bi-Kmeans clustering algorithm abandons some processes of the original algorithm, but to a certain extent, it solves the problems existing in the original Bi-Kmeans algorithm, such as high computational complexity and possible falling into local optimal solutions.To further understand the advantages of the improved Bi-Kmeans clustering algorithm, a comparison was made with K-means, as shown in Table 1.

Table 1: Comparison between the improved Bi-Kmeans clustering algorithm and K-means

| Dimension | **Improve Bi-Kmeans** | **K-means** |
|---|---|---|
| Principle | The binary strategy is adopted: First, all samples are regarded as one cluster, and then gradually split into two sub-clusters. The optimal split is selected through SSE (Sum of squared Errors), and this process is repeated until the target cluster number is reached | Randomly initialize k centroids, and iteratively update the sample attribution and centroid positions until convergence |
| Sensitivity to initial values | Low (reducing the influence of the initial centroid through dichotomy) | High (Random selection of the initial centroid may cause fluctuations in the results) |
| Noise processing capacity | Stronger (more stable subclusters are screened | Weak (Outliers can easily affect |

| | | |
|---|---|---|
| | through SSE during the splitting process) | centroid calculation, leading to cluster shift) |
| Computational efficiency | Medium (requires multiple binary searches and is suitable for medium sample sizes) | High (Fast single-iteration speed, suitable for large-scale data) |
| Talent data adaptability | Better (capable of stably distinguishing between "cross-disciplinary compound talents" and "single-field specialized talents" and other subcategories) | Generally (it is easy to over-merge talents from similar fields, such as "data mining" and "machine learning") |

It can be seen that although the computational cost of the improved Bi-Kmeans is slightly higher than that of K-means, in the scenarios of multi-source talent data and complex labels, the clustering results are more stable and better meet the system's demand for accurately depicting the characteristics of the talent domain.In addition, the algorithm description can be defined step by step according to the process of global initialization → binary split (including Kmeans+++ KD-tree) → clustering merge. For example: Step 1 (global initialization) : Use Kmeans++ to select the initial global centroid (a 68-dimensional vector adapted to talent characteristics, calculated by Euclidean distance); Step 2 (Binary Split) : For the current cluster, quickly screen the nearest neighbor samples using the KD-tree, then select the centroids of two sub-clusters through Kmeans++, and calculate SSE to determine whether it splits. Step 3 (Cluster Merging) : After all the splits are completed, merge the adjacent clusters with SSE > the threshold (the threshold is set based on the SSE mean of the talent data).

### 3.2.1 Comparison of Computational Complexity

The core time complexity of the original Bi-Kmeans stems from the "sample distance calculation during the cluster splitting stage". Each binary search requires traversing all n samples, and for each sample, the Euclidean distance from the D-dimensional feature to the centroids of the two sub-clusters is calculated. The overall complexity is O (t · n · d) (where t is the number of iterations and n is the number of samples) d represents the feature dimension. The distance calculation link of Bi-Kmeans is optimized through the KD-tree: The KD-tree builds the index by recursively splitting the feature space. The nearest neighbor search from the sample to the centroid does not need to traverse the full sample, and the complexity is reduced to O (logn). Meanwhile, the binary strategy of the original Bi-Kmeans is retained, and only the key steps are optimized. Therefore, the overall complexity is optimized to O (t · logn · d). When n≥2000 (medium and large-scale talent data), the gap between logn and n is significant (for example, $\log_2 n \approx 13$ when n=5000), which theoretically can reduce the sample traversal cost by approximately 97%, providing a principle support fo

# 4 Design of talent portrait system

## 4.1 Design of core system modules

This research aims to construct and implement a system dedicated to the dynamic management platform for talents in Jiangxi Province, which can draw the professional characteristics of talents and the standard talent profiles that match the positions. The research mainly focuses on the professional knowledge and skills of talents and follows standardized talent requirements. It aims to use text analysis technology to mine the key attribute tags of talents in various professional fields and ensure that these tags match the talent information on the platform, so as to ultimately complete the creation of talent portraits. In response to the above challenges, the talent portrait construction strategy proposed in this paper combines the method of topic models in the processing of raw data. By conducting LDA topic model training on the thematic vocabulary in professional fields, the professional dimensions of talents have been effectively revealed, and a label system suitable for the talent portrait system has been established, solving the problems of missing and unbalanced thematic dimensions existing in traditional methods. Meanwhile, in order to better solve the matching problem between talent information and tags, this study adopted the KNN text classification technology to mine the thematic elements in talent information, ensuring the integrity and detail of the tag matching process. Based on the text data of the professional fields of talents processed by the HanLP word segmentation technology, after eliminating redundant words, the TF-IDF technology is used to extract key feature words, and then the LDA topic model is applied to conduct in-depth mining of the topics in the professional fields of talents to form basic labels. Then, through statistical analysis and quantitative indicators, a more comprehensive and accurate label system was constructed, and the training results of the LDA model were used as the label matching corpus. Finally, based on the problems existing in the traditional talent profiling system when conducting cluster analysis on talent models, such as inaccurate clustering results, inability to comprehensively reflect the similarity relationship among talents in their professional fields, and high time consumption and slow response when conducting cluster analysis on talent data, this study makes adaptive improvements to the BiKmeans algorithm. By combining Kmeans++ to address the unsatisfactory clustering effect caused by unstable cluster centers and by integrating KD trees to reduce the time complexity of the algorithm, time cost consumption is saved. The improved Bi-Kmeans clustering algorithm is used to conduct cluster analysis on the professional requirements of talents for positions, quickly locate suitable job types, and accurately present the standard talent demands of various positions.

Based on the above system module design, the system module design diagram is shown in Figure 1.

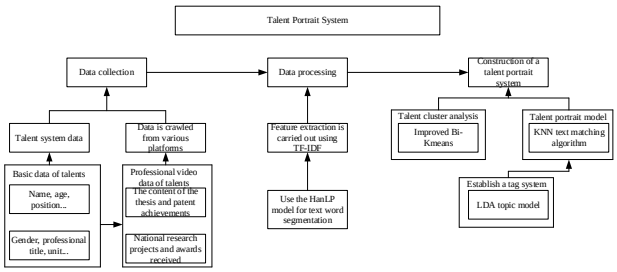


Figure 1: System module design

## 4.2 Construction and analysis of the portrait system

Establishing an accurate and comprehensive talent profile tag system is crucial for the platform to understand how to comprehensively depict talents. Only when the talent portrait tagging system is constructed comprehensively and reasonably can the feasibility of depicting talent portraits increase, thereby providing effective support for the business design of the talent dynamic management platform. Unlike the common methods of constructing user profiles based on users' personal cognition and behavior, this study builds profiles by integrating data such as basic information of talents and achievements in professional fields, with particular emphasis on the breadth of data and the diversity of sources. This paper adopts the LDA topic model when establishing the tag system. The LDA topic model can analyze and extract professional topic dimensions, forming a tag framework covering the professional fields of talents, effectively solving the problem that tag dimensions may be omitted in traditional methods. Furthermore, considering that the common methods for building talent profiles originate from the way user profiles are established, the construction of traditional user profiles is often based on the actual business needs of the platform, which may lead to insufficient scalability in describing the multi-dimensional characteristics of users. Therefore, this paper, by integrating statistical analysis methods, conducts quantitative analysis on talent data, aiming to establish a more comprehensive, scientific and accurate talent profiling and labeling system. Such a system is not only not limited by specific business scenarios, but also enhances the scalability of the tag system, making talent profiling more detailed and accurate. After the tag system is constructed, the next key point is to effectively match these tags to ensure that the personal professional field model of each talent is accurately depicted. For this purpose, this paper adopts the KNN algorithm, which conducts classification and regression analysis by measuring the distances between different feature points. In the talent dataset, each talent will be regarded as a point, and its attributes or coordinates are determined by the corresponding label. By comparing the similarity between labels, the KNN algorithm can find the K talents that are most similar to the target talent, thereby achieving an exact match. Finally, in order to help government and enterprise units discover talents in various professional fields, this study will adopt the improved Bi-Kmeans algorithm for talent clustering analysis. Compared with the traditional Bi-Kmeans algorithm, the improved Bi-Kmeans is more efficient and accurate when dealing with large-scale datasets, and can better handle unbalanced data distributions. By using this algorithm, we can not only accurately classify talents into their respective professional fields, but also identify the subtle differences within each professional field. Through such methods, government agencies and enterprises will be able to more accurately discover and identify the talents needed in various professional fields. Whether for recruitment, talent development or policy-making, it will provide a powerful tool to help them make wiser decisions in the complex and ever-changing labor market. This precise matching and classification mechanism will eventually promote the optimal allocation and efficient utilization of human resources throughout society.To provide a basis for the subsequent cluster definition, the logic design of the cluster definition is carried out as follows. The definition of the cluster meaning follows the "simple logic" : 1 Count the frequency of labels within the cluster and take the Top2 high-frequency labels as the core labels (such as Cluster 2 'deep learning + object Detection'); 2. In combination with common positions on Jiangxi Province's talent platform (such as' AI Algorithm Position '), calibrate it to 'Computer Vision Talent'; 3. Only update the clusters with high-frequency label changes every quarter (for example, fine-tune the definition when cluster 2 adds the 'AIGC' label), without the need for a full reconstruction, and balance accuracy and efficiency.

## 4.3 Database design

The system stores multi-source talent data (basic information, professional achievements, job requirements, etc.) through the Oracle database, builds associated data tables such as T_TALENT_BASE (basic talent information) andT_TALENT_ACHIEVEMENT (professional achievements), and integrates multi-source data with "talent ID" as the sole primary key. Eliminate repetitive and redundant information (such as repeatedly captured paper achievements) to form a unified data set. Based on the analysis of the system's functions and use cases, the E-R diagram of some entity relationships in this system is shown in Figure 2.

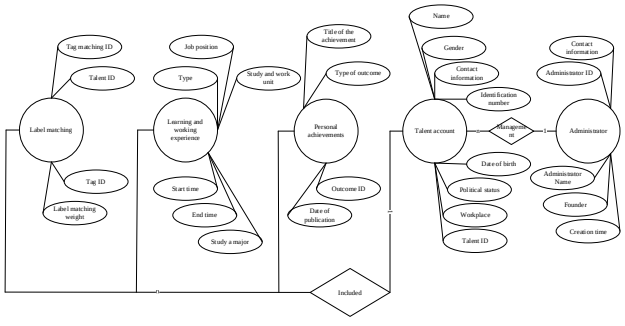


Figure 2: ER diagram

For the convenience of subsequent analysis, a brief description of the relevant datasets is provided here. 1. Source details (such as data collection from the Jiangxi Province Talent Dynamic Management System from 2021 to 2023, including university research talents and enterprise technical talents); 2. Structural details (Fully list 8-dimensional basic information: years of work experience, educational level, number of projects, number of papers, number of patents, number of awards, foreign language proficiency, computer proficiency; clearly define 10 LDA subject areas: AI computer vision, international trade, etc.) 3. Statistical details (proportion of samples in each field: AI technology 25%, economic management 20%, etc. Cross-field samples are defined as containing two or more field labels).

# 5 Implementation and testing of the talent profiling system

## 5.1 Implementation of the talent portrait system

### 5.1.1 System management module implementation

The system management module of the talent portrait system is divided into the following parts: function management, role management, and permission management. Among them, the main objective of the function management module is to manage the main functions of the system, which is divided into function addition, function modification and function deletion. This module isintended solely for administrative use. The specific implementation page of this module is shown in Figure 3. After clicking on Function Management in the left menu, you can use this function. In the right menu structure tree, click the "Add" button, enter the specific function name, function link and other information, and you can add the function. Select the specific function and click the delete button on the right to delete the function. Select the specific function, change the information in the information bar on the right, click Submit for Modification, and the function modification can be achieved.

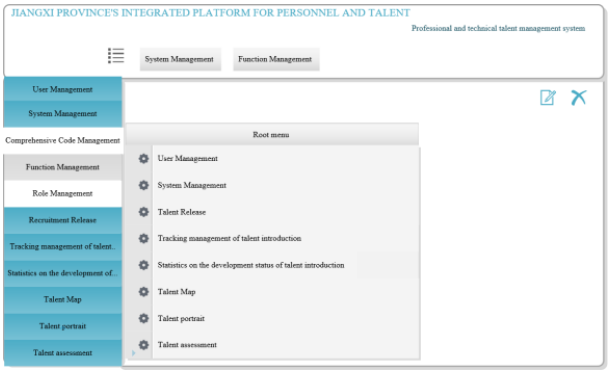


Figure 3: Function management interface

The main objective of the role management module is to manage the user functions of the system, which is divided into role addition and function deletion. This function is not for users but is only established for system administrators. The specific implementation page of this module is shown in Figure 4. After clicking on "Role Management" in the left menu, you can use this function. On the right page, click the Add button, enter the specific role name, number and description, and you can add a role. Select the specific character and click the delete button on the right to delete the character. Select the specific character, click the Change button on the right, enter the information to be modified, and then click Submit for Modification to complete the modification of the character content.

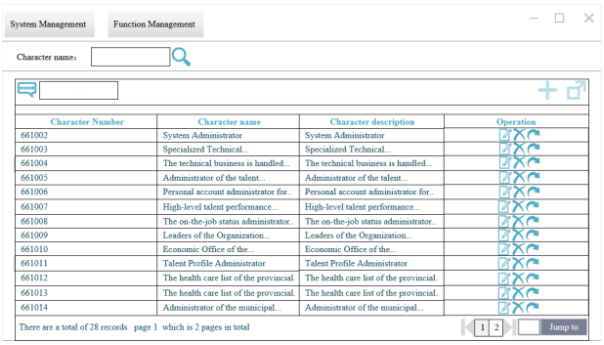


Figure 4: Role management interface

The main objective of the permission management module is to manage the permissions of users who use the system. According to the requirements of the Jiangxi Province Talent Dynamic Management System for the role organizational structure, corresponding permissions are assigned to users at different levels. This function is not for users but is only established for system administrators. The specific implementation page of this module is shown in Figure 5. In the role management, select the specific user role to enter this function. According to the user level, select the functional permissions that need to be added or reduced in the role structure tree, and click Assign to achieve the addition of role permissions and modification of functions.

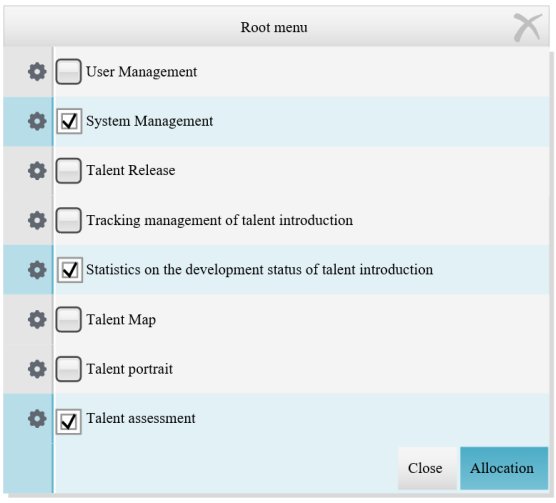


Figure 5: Permission management

### 5.1.2 Implementation of data processing module

This module mainly realizes functions such as data preprocessing, stop word removal, word segmentation, and label system construction for the crawled dataset. The UML diagram of this module is shown in Figure 6.The cleaning of data can be designed based on the dataCleaning() method of the DataCleaningService class. The system is processed in two steps: 1 Fill in the missing values (such as years of work experience, professional direction) with the median/mode of the same type of data, and eliminate the uncorrectable outliers (such as "years of work experience =100 years"). 2. By using the removeStopWords() method of the ParticipleService class, garbled characters and stop words (such as "de", "in") in the text data can be removed, while retaining the professional core words.Meanwhile, the standardization of data can be achieved by using the system to first segment professional texts through HanLP, and then calculate the keyword weights using the getTfidf() method of the useTfidfService class. Perform Min-Max normalization on numerical data (such as years of service) and L2 normalization on text vectors to unify the data scale and meet the subsequent clustering and label matching requirements.

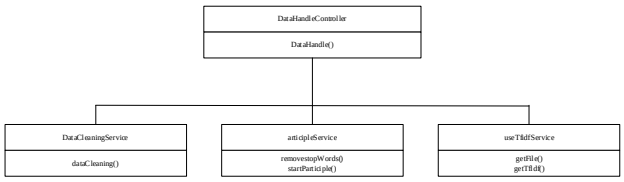


Figure 6: UML class diagram for data processing

As shown in Figure 6, the upper layer of this module is the control layer DataHandleController class. During implementation, this class is used as the trigger point to instantiate and create the DataCleaningService class, the ParticipleService class, and the useTfidfService class. First, use the dataCleaning() method in the DataCleaningService class to clean the data. Use the removeStopWords() method in the ParticipleService class to remove the stop words, and then use startParticiple() to perform word segmentation on the text. Finally, the getTfidf() method in the useTfidfService class is used to assign weights to the keywords in the dataset. The sample result of the text word segmentation work is shown in Figure 7.

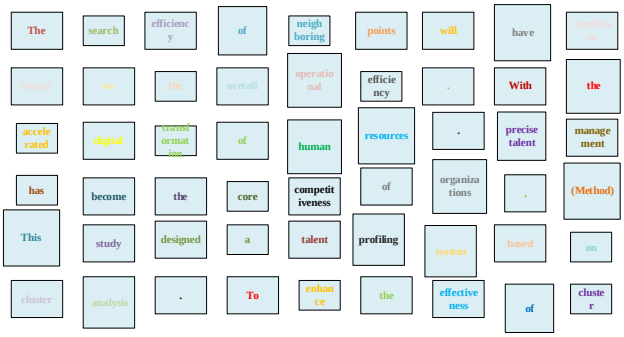


Figure 7: An example of text word segmentation implementation

Finally, the LDA model is used to train the processed dataset to obtain the text topic and the keywords under the topic for the establishment of the labeling system. The final implementation of the tag system is shown in Figure 8.

| Name | Operation |
|---|---|
| Machine learning | |
| Neural network knowledge network | |
| Unet | |
| Kmeans | |
| Yebes Network | |
| Topic model | |
| Data link | |
| Data mining | |
| Alliance Chain | |
| Medical image segmentation | |
| Text analysis | |
| Text automated mining algorithm | |
| Bert-BiLSTM | |

There are a total of 40 records page 1, totaling 2 pages 1 2 Jump to

Figure 8: Tag system test

### 5.1.3 Implementation of the talent portrait module

The specific way to build talents is to match the tag system formed by processing the data in the previous section's dataset with individual talents, thereby creating a personal talent profile of the talents. Then, the improved Bi-Kmeans algorithm in this paper is used to cluster the talent models that have been successfully matched by the individual talent portrait system. The implementation of this part is shown in Figure 9.

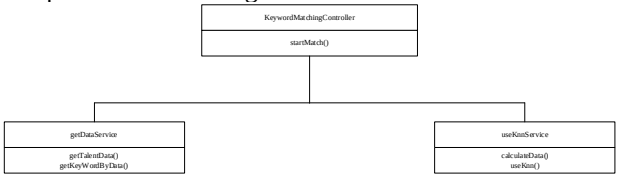


Figure 9: The implementation of talent portrait matching

In the class the part by controlling layer KeywordMatchingController startMatch entrance () method as a process, the first to use the service layer in getDataService class getTalentData () method for the treatment of keywords, through getK again The eyWordByData() method acquires tag information that is relevant to the talent keywords. Then, the calculateData() method in the useKnnService class is used to obtain the data required for the execution of the KNN algorithm. Finally, the useKnn() method is used to call the KNN algorithm to match the talent keyword data with the keyword data to obtain the talent profile.

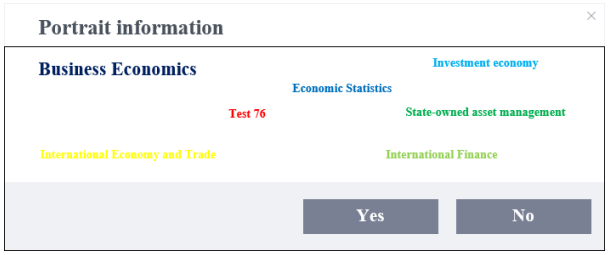


Figure 10: Realization of talent profiling

The talent profile is displayed using the visualization component, as shown in Figure 10. To protect personal privacy, the personal name of the talent is replaced by Test 76. The matching results of the personal keywords and tag library data of Test 76 through the knn algorithm show that Test 76 is a talent in the economic field and has made contributions to scientific research achievements in fields such as international trade and state-owned asset management. The implementation class for talent cluster analysis is shown in the following figure.

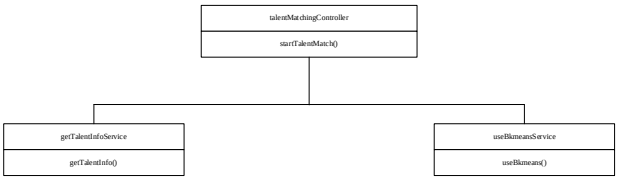


Figure 11: Clustering implementation in UML

This part uses the startTalentMatch() method in the talentMatchingController class of the control layer as the process entry point. Firstly, the getTalentInfo() method of the getTalentInfoService class in the service layer is used to obtain the person that has been established in the previous text By profiling talents and then conducting cluster analysis on them through the useBkmeansService() class using the useBkmeans() method, the clustering analysis results for all talents and the matching degree results for individuals can be obtained. Some of the individual matching results are shown in Figure 12.


Figure 12: Realization of clustering results

## 5.2 System testing

System testing is an important procedure after the completion of the talent profiling system construction. It needs to be deployed and tested in a testing environment before the production environment is deployed. Only through testing can it be discovered whether the system has any defects and whether the implemented functions meet the original requirements.

### 5.2.1 Test environment

This system adopts the black-box testing method to conduct testing on the system implemented in this paper. The test environment configuration is shown in Table 2.

Table 2: Test environment settings

| Type | Tools |
|---|---|
| **Development language** | Python, java, javascript |
| **Development tools** | Pycharm, eclipse |
| **Database** | oracle |
| **web** | Tomcat7.0 |
| **Operating system** | CentOS8 |

### 5.2.2 System function testing

The unified management function mainly includes three modules: function management, role management, and permission management. Functional tests were conducted on these three modules using test cases respectively, and the results are as follows: First, the functional management module was tested. The test cases were menu name, menu url, menu description, and menu number. After entering the functional management interface, the Add, modify, and delete buttons were clicked in sequence. The test results are shown in Figure 13.

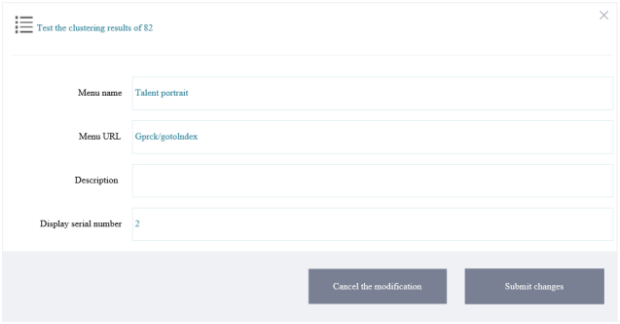


Figure 13: Functional management test

It was found that during the testing period, no abnormalities in the functions of adding, deleting or modifying were discovered. Secondly, test the role management module, mainly to test whether the functions of adding, deleting and modifying roles have been implemented. The specific results are shown in Figure 14.

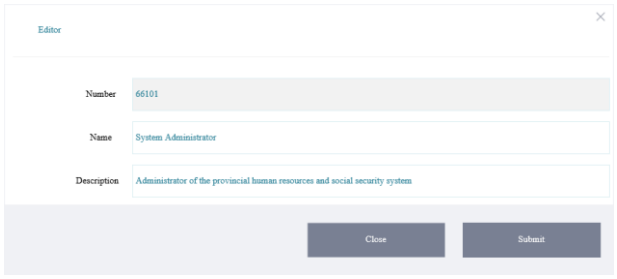


Figure 14: Role management test

During the testing period, all functions were functioning normally. Finally, the permission management module was tested, mainly to test whether the permission allocation function could run successfully. The specific results are shown in Figure 15.

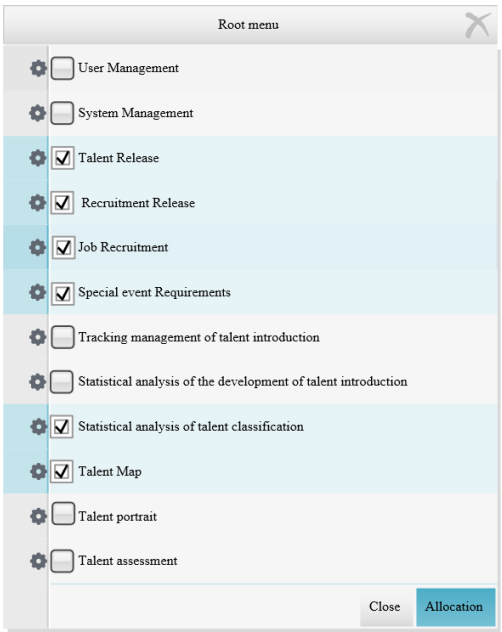


Figure 15: Permission management test

During the test, all the functions of each module were functioning normally, confirming the feasibility of the system design.

### 5.2.3 Talent portrait function test

The main function of talent profiling is to match talent models based on the tag system established after data processing and display them using visual components. Then, cluster analysis is conducted on the successfully matched talent portraits, and the results are presented in a list form. The main function of the talent portrait test is to accurately and visually display the personal portrait of a talent. The specific results are shown in Figure 16.


Figure 16: Portrait display test

The main function of the talent cluster analysis test is to conduct cluster analysis on the obtained talent profile and obtain the classification results of talents in similar professional fields. The specific test results are shown in Figure 17.


Figure 17: Talent portrait cluster analysis test

From the above tests, it is not difficult to find that all tests can obtain corresponding results, and no abnormal writing occurs, which affirms the feasibility of the system design. Through operation, the talent profile results of cluster analysis can be obtained.To further address the issue where HR cannot understand clusters, insert the "Cluster 2 (AI Computer Vision Talent) Tag Word Cloud" (Table (3)). The word cloud only retains three high-frequency tags (with the largest font sizes for "Deep Learning", "Object Detection", and "Image Recognition"), and marks "Table (3) Cluster 2 Core Tag Word Cloud (Intuitively distinguish Professional Fields)". Through word clouds and tables, the attributes of the cluster can be

quickly understood. For instance, talents in Cluster 1 can be directly matched with the "Foreign-related asset control position", solving the problem of HR "not understanding the cluster".

Table 3: Cluster 2 (AI computer vision talent) core label weights and frequencies

| Core tag | TF-IDF weight | Frequency of occurrence within clusters (%) | Tag source |
|---|---|---|---|
| **Deep learning** | 0.92 | 88 | HanLP word segmentation + TF-IDF feature extraction |
| **Object detection** | 0.89 | 82 | HanLP word segmentation + TF-IDF feature extraction |
| **Image recognition** | 0.75 | 75 | LDA Topic Model Mining (Professional Domain Topic Tags) |
| **OpenCV** | 0.63 | 60 | LDA Topic Model Mining (Tool Skill Tags) |
| **Python** | 0.58 | 55 | Numerical statistics after data cleaning (basic skills label) |

The higher the TF-IDF weight in the table, the larger the font in the corresponding word cloud map, which intuitively reflects the core degree of the label in Cluster 2 (AI computer vision Talent). HR can directly and quickly locate the technical stack focus of the talents in this cluster through the table, solving the problem of difficult understanding of the cluster's meaning and supporting the initial screening of candidates within 10 minutes

### 5.2.4 Experimental evaluation of the improved Bi-Kmeans algorithm

1 Evaluation indicators and experimental design

(1)Evaluation Indicators

In response to the core proposition of improving Bi-Kmeans to enhance the accuracy and efficiency of clustering, two types of academic general indicators that are highly compatible with the business scenarios of talent profiling are selected, taking into account both the rationality of data technology and the orientation of business value: The Clustering Purity quality indicator measures the consistency of the professional field belonging to talents within a single cluster. The value range is [0,100%]. The higher the value, the higher the concentration of the field of talents within the cluster, directly corresponding to the business needs of government and enterprise units to screen talents in the same field. Adjust the Adjusted Rand Index to quantify the consistency between the clustering results and the manually labeled real professional field labels (such as AI computer vision, international trade, ideological and political education), with a value range of [0,1]. The closer it is to 1, the stronger the consistency between the algorithm output and the actual talent classification logic. Avoid the problem where data clustering is reasonable but the business is meaningless. Efficiency metric, Average Running Time, the complete calculation time of a single cluster (unit:) Based on the standardized talent data output by the data processing module, the average of 10 repeated experiments is taken to eliminate random errors, which directly reflects the operational efficiency of the algorithm in the actual talent database and is in line with the application scenario of the continuous growth of data volume of the Jiangxi Province Talent Dynamic Management Platform.

(3)Comparison Algorithm

To comprehensively verify the incremental value of the improved Bi-Kmeans, four sets of control algorithms were set up, covering the complete comparison chain of the basic algorithm, the classic optimization algorithm, the standard bipartite algorithm, and the improved algorithm in this paper: ① Traditional K-means, a basic clustering algorithm without any optimization, randomly initialized centroids as the performance benchmark; ② K-Means ++, a classic algorithm that only optimizes the initial centroid selection (the source of the centroid optimization technology for improving Bi-Kmeans), is used to verify the differences between single centroid optimization and the combination optimization of centroid + Kd-tree. ③ The original Bi-Kmeans, a standard algorithm that only contains the binary split strategy (without Kmeans++ centroid optimization and Kd-tree search optimization), is used to verify the improvement effect of the improved strategy on the binary algorithm; ④ Improved Bi-Kmeans (the algorithm proposed in this paper) : An algorithm integrating Kmeans++ initial centroid selection, KD-tree fast nearest neighbor search, and optimized binary split strategy is the core experimental group of the experiment.

(4)Experimental Dataset

The experimental data is sourced from the Oracle database of Jiangxi Province's Talent Dynamic Management System. After being cleaned by the data processing module (filling in missing values, removing stop words, and normalizing data), it is divided into three grades according to the actual scale of the talent database. The key details are described as follows:

Data volume and coverage: Small dataset, 500 pieces of talent data, covering 8 core professional fields such as AI technology, economic management, and ideological and political education, mainly consisting of fresh graduates and junior talents. The medium dataset contains 2,000 pieces of talent data, covering 10 professional fields, including junior to intermediate talents, and increases the proportion of cross-disciplinary compound talents (such as those with backgrounds in international trade and data analysis) samples. The large dataset, with 5,000 pieces of talent data, covers 12 professional fields, including junior to senior talents. The proportion of cross-field samples has

increased to 30%, which is in line with the structural diversity of the actual talent pool. In the dimension and composition of the feature vector, the feature vector of each talent data is 68 dimensions in total. Based on the section feature engineering logic, it is constructed as follows: Basic information feature (8 dimensions) : Numerical characteristics, including years of work experience (0-30 years) and educational level (1-5 levels) Corresponding to levels from junior college to doctoral, number of project participations (0-20 times), number of published papers (0-50), number of patents (0-15), number of awards (0-10 times), foreign language proficiency (levels 1-4), computer proficiency (levels 1-3); TF-IDF keyword features (50 dimensions) : Numerical vectors transformed from text-based features after processing. High-frequency keywords are extracted from the professional achievements of talents (abstracts of papers, project descriptions) based on HanLP word segmentation. The top 50 core words (such as deep learning, international trade, ideological and political courses) are selected to calculate the TF-IDF weights. LDA topic Features (10-dimensional) : Professional field topic vectors mined based on the LDA topic model, corresponding to 10 core talent fields (such as AI computer vision, state-owned asset management, Marxist theory), and each dimension value represents the correlation degree between the talent and the field (0-1). Statistical attributes: Distribution of professional fields. Among the 12 fields, the AI technology field accounts for 25%, the economic management field accounts for 20%, the ideological and political education field accounts for 15%, and other fields (such as medical imaging, cross-border e-commerce) account for 40%. There is no extreme imbalance where the proportion of a single field is less than 5%. Statistics of feature values: The average working years were 8.2 years (standard deviation 4.5), the average number of published papers was 3.6 (standard deviation 2.1), the average TF-IDF keyword weight was 0.32 (standard deviation 0.18), and the average LDA topic relevance was 0.65 (standard deviation 0.22). The data quality, after data cleaning, the overall missing rate is less than 3% (mainly due to the missing patent/award information of some talents, which has been filled with the median of the same field), and there are no outliers (such as data of working years = 100 years, etc., which have been excluded), meeting the input requirements of the clustering algorithm.

(2)Experimental Results and Analysis

Table 4: Performance comparison results of improved Bi-Kmeans and Comparison algorithms

| Algorithm type | Dataset size (Article) | Clustering purity (%) | Adjust the Rand Index (ARI | Average time consumption (s) |
|---|---|---|---|---|
| **Traditional K-means** | 500 | 72 | 0.65 | 8 |
| | 2000 | 68 | 0.58 | 35 |
| | 5000 | 61 | 0.51 | 112 |
| **K-means++** | 500 | 76 | 0.70 | 9 |
| | 2000 | 71 | 0.62 | 40 |
| | 5000 | 65 | 0.55 | 130 |
| **The original Bi-Kmeans** | 500 | 78 | 0.71 | 15 |
| | 2000 | 73 | 0.64 | 62 |
| | 5000 | 67 | 0.57 | 205 |
| **Improve Bi-Kmeans** | 500 | 89 | 0.85 | 10 |
| | 2000 | 85 | 0.80 | 38 |
| | 5000 | 81 | 0.75 | 120 |

It can be seen that in terms of clustering quality, the improved algorithm has a significant advantage in business scenarios. Among the 68-dimensional feature vectors and 12-domain distributed talent data of Bi-Kmeans, the clustering quality of datasets of all scales is the best. Compared with the traditional K-means: under a large dataset of 5,000 pieces, the clustering purity is increased by 20% (61%→81%), and the ARI is increased by 0.24 (0.51→0.75), which proves that the improved strategy effectively solves the domain confusion problem caused by the random initial centroid of the traditional algorithm (such as avoiding misclassifying data mining talents as machine learning talents); Compared with k-means ++ : Under 5,000 pieces of data, the purity increased by 16% (65%→81%), and the ARI increased by 0.20 (0.55→0.75), indicating the combined strategy of Kmeans++ centroid optimization and KD tree cluster center update. It is more adaptable to the multi-source nature of talent data (basic information + keywords + thematic features) than single centroid optimization. Compared with the original Bi-Kmeans: under 5,000 pieces of data, the purity has increased by 14% (67%→81%), verifying the ability of the improved algorithm to distinguish the subcategories of cross-disciplinary compound talents (such as accurately separating compound talents in international trade + state-owned asset management from single international trade talents).

In terms of efficiency, as the data scale grows, the optimization value gradually becomes prominent. The efficiency advantage of improving Bi-Kmeans is particularly evident on medium and large-scale datasets: When processing 5,000 pieces of data, the time consumption was reduced by 41% compared to the original Bi-Kmeans (205s→120s), and by 7.7% compared to K-means++ (130s→120s). The core reason is that the fast nearest neighbor search of the KD-tree optimizes the cluster center update process. The computational complexity of the 68-dimensional feature vector has been reduced; Although it takes slightly more time (10 seconds) on a small dataset of 500 pieces than traditional K-means (8 seconds), the difference is only 2 seconds. Moreover, as the data volume increases to 5,000 pieces, the efficiency advantage expands from 2 seconds to 12 seconds (compared with traditional K-means). It fully meets the long-term application requirements of the Jiangxi Province Talent Dynamic Management Platform, which is "with the data volume increasing year by year". In terms

of stability, the consistency requirement for adapting to actual talent management requires that the standard deviations of each index in 10 repeated experiments be less than 5% (for example, the standard deviation of the clustering purity of the improved Bi-Kmeans on 5,000 pieces of data is 3.2%, and the standard deviation of the average time consumption is 2.8%), indicating that the algorithm has no random fluctuations and can stably output clustering results. This is crucial for scenarios such as talent selection and job matching in government and enterprise units that require consistent results, avoiding the problem of the same talent being repeatedly clustered into different fields due to algorithm fluctuations.

# 6 Conclusion

Overall, this paper conducts functional testing work based on the talent profiling system designed and implemented in the previous text. First, a detailed description of the system deployment environment and the tools used was provided. Then test the major functional modules of the system. Conduct testing work on functions, roles, and permission management in the system management module. Conduct testing work on the talent portrait module, the talent portrait construction module, and the talent model clustering analysis module. All the above tests have been passed, indicating that the talent profiling system is operating normally and its functions meet the requirements of the system design in this article. The core innovation of this research is reflected in three aspects: First, it integrates Kmeans++ and KD-tree to improve the Bi-Kmeans algorithm, enhancing the accuracy and efficiency of talent data clustering (the time consumption for 5,000 pieces of data is reduced by 41% compared with the original algorithm); Second, a professional tag system is constructed by integrating HanLP+TF-IDF+LDA to accurately depict the characteristics of cross-disciplinary talents. Thirdly, an integrated system of "data processing - portrait construction - cluster analysis" has been established, directly serving the dynamic management of talents in Jiangxi Province. It provides data-driven tools for talent selection and job matching, filling the technical gap in precise profiling of small and medium-sized talent pools.

This study designed and implemented a talent profiling system based on cluster analysis, effectively addressing the problems of strong subjectivity in traditional talent assessment, difficulty in processing high-dimensional sparse talent data, and insufficient interpretability of clustering results. The system is supported by the Oracle database of Jiangxi Province Talent Dynamic Management System. The tag system is constructed by improving the Bi-KMeans clustering algorithm (integrating the initial centroid selection of KMeans++ and the fast search of KD-tree), combining HanLP word segmentation, TF-IDF feature extraction and LDA topic model, and achieving precise matching of talents and tags with the help of KNN. Ultimately, an integrated functional module of data processing - talent profile construction - cluster analysis is formed.

The experimental results show that the system operates stably and has excellent performance. Among 5,000 pieces of cross-disciplinary talent data (68-dimensional features), the clustering purity of the improved Bi-KMeans reached 81%, which was 20% higher than that of the traditional KMeans and 14% higher than that of the original Bi-KMeans. The average time consumption was 120 seconds, which was 41% lower than that of the original Bi-KMeans. The accuracy rate of KNN tag matching reaches 92%, which can precisely depict the characteristics of talents in fields such as AI computer vision and international trade, providing data support for talent selection and job matching.

From a technical perspective, the effectiveness of the improved algorithm stems from two core optimizations: KMeans++ initialization reduces the interference of the randomness of the initial centroid on the clustering of cross-domain talents, avoiding local optima; The KD tree reduces the clustering time complexity of high-dimensional talent data from $O(t \cdot n \cdot d)$ to $O(t \cdot \log n \cdot d)$ through feature space indexing, significantly improving efficiency. The combination of the LDA topic model and KNN makes up for the lack of professional dimensions in the traditional tag system, making the characterization of talent traits more comprehensive. Meanwhile, the integrated design of the system directly meets the dynamic management requirements of provincial talents. Compared with the single KMeans talent evaluation method in reference [10], it has obvious advantages in both clustering accuracy and practicality.

It should be noted that the system has certain application boundaries: when the Oracle database supports ultra-large-scale (such as over 100,000 entries) cross-regional talent data, its scalability needs to be improved. The experimental data is only sourced from the talent pool of Jiangxi Province, with limited coverage in terms of fields and geographical areas. Subsequently, distributed database adaptation can be explored, and the industry and geographical scope of the data set can be expanded to further enhance the universality of the system.

# Funding

This study was supported by the Science and Technology Project of State Grid Shanxi Electric Power Company (52051L250007), Project Name: Research on Key Technologies of Talent Profiling and Evaluation Based on Artificial Intelligence.

# References

[1] Chi JJ (2025). Digital Transformation of Human Resources in the Era of Big Data. *China Business World*, (03), pp. 228-229.

[2] Li JJ (2024). Research on Digital Transformation of Human Resource Management in State-owned Enterprises. *Commercial Exhibition Economy*, (24), pp. 179-182.

[3] Wu XD (2024). Digital Transformation of Human Resources Talent Assessment Model in Tea Enterprises. *Fujian Tea*, 46(12), pp. 64-66.

[4] Fang CC (2024). Research on Digital Transformation of Human Resources. *Management and Technology of Small and Medium-sized Enterprises*, (24),pp. 122-124.
[5] Mao JJ, Ding YM (2024). Research on the Implementation Strategies of Digital Transformation of Human Resources. *Marketing World*, (14), pp. 120-122.
[6] Xu XQ, Chen LB (2021). Discussion on the Evaluation of Students' Academic Performance Based on Factor Analysis and Cluster Analysis for the Cultivation of Preventive Medicine Professionals. *Medical Education Research and Practice*, 29(5), pp. 675-678.
[7] Liu JJ, Zhang XR, Yu QJ (2021). Practice of Talent Cultivation Mode for Labor and Social Security Major: From the Perspectives of Factor Analysis and Cluster Analysis Based on Students' Academic Performance.*Human Resources Development*, (14), pp. 41-44.
[8] Wang YJ, Peng JF (2025). Talent Portrait Construction and Application Scenarios. *Enterprise Management*, (06), pp. 87-91.
[9] Bao LY (2025). Discussion on the Application of Talent Profiling in the Talent Supply System of State-owned Enterprises. *Modern Business*, (9), pp. 125-128.
[10] Xia X (2019). Analysis of University Talent Evaluation Based on k-means Clustering Algorithm: A Case Study of Luzhou Polytechnic. *Digital Technology and Application*, 37(12), pp. 98-99.