Development of the Employment Recommendation System based on K-Means Improved Collaborative Filtering Algorithm

Pengying Wan

School of Economics and Management, Beijing Jiaotong University, Beijing, China
pennywan2606@163.com

ABSTRACT

With the massive expansion of higher education, the employment pressure of college graduates has increased dramatically. Meanwhile, graduates are unable to find their preferred positions from the massive and heterogeneous data, and are extremely disturbed by many irrelevant information while searching. To address the above problems, this paper proposes the development of an employment recommendation system based on K-means improved collaborative filtering recommendation algorithm. First, the job seeker and employment job data are collected and preprocessed to understand the characteristics of job seekers and job resources, then the job seeker behavior matrix is established, and similar users are clustered by K-means clustering algorithm, and in the clustering process, the distance between data is calculated by using Euclidean formula, and then the set of neighboring items is selected by improved similarity calculation to predict the job seeker rating and realize recommendation. The experiments show that the method has improved in precision, recall and F-score ratio to some extent.

Keywords: K-means clustering algorithm; collaborative filtering; Employment recommendation system

1. INTRODUCTION

The employment situation of college graduates directly affects the healthy development of social economy, especially in the context of massive expansion of higher education. According to the data of China Business Industry Research Institute, from 2013 to 2021, the number of college graduates grows year by year. There were 8.34 million college graduates in 2019, 8.74 million in 2020, while 2021 broke through 9 million [1]. The expansion of higher education has resulted in the growth of graduates' employment demand and little change in job supply, thus intensifying the contradiction between employment and talent market supply, while the expansion has led to the uneven quality of graduates and the surge of employment pressure.

In response to the employment problem of college students, the government and guidance agencies have put forward a series of policies and methods to promote employment, relieve employment pressure, and improve employment and entrepreneurship rates, but they are generally not detailed enough and do not reflect professional differences, while favoring theory and neglecting practice, without considering the gap between comprehensive market and professional demand [2]. At the same time, there is a polarization of college students' understanding of employment information. Part of the graduates lack understanding of employment information and have high expectations leading to a mismatch between personal quality and work ability; the other part, with the development of technology network, job seekers are unable to find their preferred positions from the massive and complicated data, and are extremely disturbed by a lot of irrelevant information while searching [3].

Therefore, based on the above problems, this paper proposes the development of an employment recommendation system based on K-means improved collaborative filtering recommendation algorithm.

2. LITERATURE REVIEW

With the advent of the era of big data, recommendation systems can effectively solve the problem of information overload.

S. A. R. Khan et al. (Eds.): ICMSSE 2022, AHCS 12, pp. 489–494, 2023.
https://doi.org/10.2991/978-94-6463-056-5_72

Up to now, there have been many successful applications of recommendation systems at home and abroad. A study of Netflix indicated that 60% of users thought the movie genre recommended by the movie recommendation system matched their preferences [4]. Meanwhile, Amazon had increased its annual revenue by about 20% by analyzing users' behavioral historical data as a way to provide personalized products. In addition, domestic Douban and Today's Headlines are also examples of recommendation system applications.

Hengkai Li et al. proposed an employment recommendation model for college graduates based on big data analysis, which can obtain relevant recruitment information through web crawler technology and can meet the employment needs of college students more accurately [5]. Tsaic et al. combined collaborative filtering algorithm with clustering algorithm, which effectively improved the accuracy of the recommendation system, but only considered the different rating criteria among users' differences [7]; Shunyong Li et al. applied the K-means clustering algorithm to the collaborative filtering algorithm and used an improved similarity formula to find the set of users' neighbors, which improved the accuracy of recommendation results to some extent [8].

3. RESEARCH METHOD

3.1 Recommendation Algorithm

Recommendation algorithm can be divided into content-based recommendation, collaborative filtering algorithm and hybrid recommendation. Among them, collaborative filtering recommendation is one of the most widely used algorithms. The main collaborative filtering algorithms are memory-based collaborative filtering and clustering-based collaborative filtering.

3.1.1 Memory-based Collaborative Filtering

The nearest neighbor technique is generally used to calculate the distance between users using their historical information, and then assign the product ratings of the target user's neighbor users to predict the target user's preference for a specific product to achieve the recommendation result. The memory-based collaborative filtering can be divided into User-based CF and Item-based CF.

User-based collaborative filtering is to predict the ratings of similar target users based on the rating behavior among users, which is suitable for scenarios with a small number of users, more real-time and more social.

Item-based collaborative filtering is to predict user ratings based on the similarity between the predicted item and the actual item selected by the user The new behavior of users leads to real-time changes in the recommendation results, and is suitable for areas with strong personalization needs.

3.1.2 Collaborative Filtering Algorithm Based on K-means

Clustering is an unsupervised machine learning algorithm that constructs data based on a defined model. By clustering users with similar interests or similar items into one cluster. The K-means clustering algorithm is a widely used algorithm with high recommendation accuracy and relatively simple algorithm implementation. The goal of this algorithm is to classify all data into k clustering according to the input parameter k (the number of clustering targets). The basic idea is to assign each data point to the cluster class where the nearest cluster centroid is located. The specific steps are as follows.

First, randomly select K data points among the data points as the initial clustering centroids.

Second, for each data point, use the Euclidean distance formula to calculate the distance from the point to each cluster center and assign it to the nearest cluster. The Euclidean distance formula is

$$dis(X,Y)=\sqrt{\sum_{i}^{m}(x_i-y_i)^2} \tag{1}$$

where: X, Y are user data, n is the number of items, and x_i and y_i denote the rating of item i by users x, y.

Third, recalculate the mean value of the data points in each category by using equation (2), and the mean point is used as the new cluster centroid.

$$c_i=\frac{1}{m}\sum_{x\in C_i}x \tag{2}$$

If no data point is the same as the mean point, the distance from the point to the mean point is calculated (2), and the data point with the closest distance is used as the new cluster center.

Fourth, if the change of the cluster center does not exceed the preset threshold, then converge; otherwise, go back to step 2.

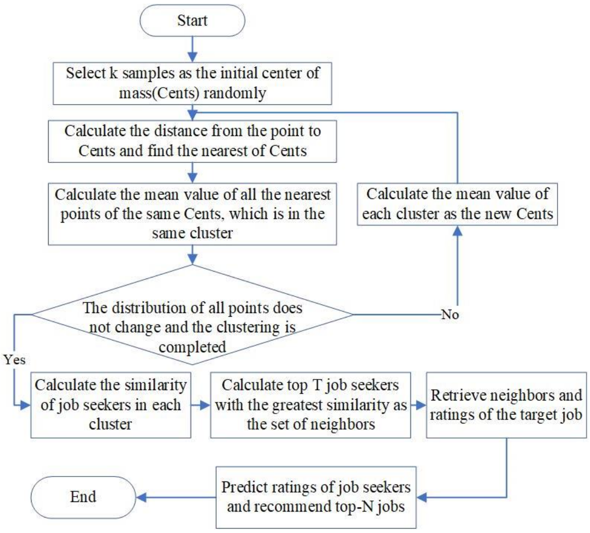
Start
Select k samples as the initial center of mass(Cents) randomly
Calculate the distance from the point to Cents and find the nearest of Cents
Calculate the mean value of all the nearest points of the same Cents, which is in the same cluster
Calculate the mean value of each cluster as the new Cents
The distribution of all points does not change and the clustering is completed
No
Yes
Calculate the similarity of job seekers in each cluster
Calculate top T job seekers with the greatest similarity as the set of neighbors
Retrieve neighbors and ratings of the target job
Predict ratings of job seekers and recommend top-N jobs
End

Figure 1: Improved Flow Chart based on k-means clustering

3.2 Similarity Computation

3.2 1 Pearson's Correlation Coefficient

Pearson's correlation coefficient measures the degree of linear correlation between any two variables, and a larger value of the coefficient means a high degree of similarity. Its value ranges from [- 1,1], and positive values indicate that the two variables are positively correlated, while negative values indicate that they are negatively correlated.

$$sim(x,y)=p(x,y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}\frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})}^2} \tag{3}$$

3.2.2 Cosine Similarity Coefficient

The similarity of two vectors is evaluated by calculating the cosine of their angle. It takes values between [- 1,1] and has the same meaning as Pearson coefficient.

$$sim(x,y)=\cos\theta=\frac{\sum_{i=1}^{m}x_iy_i}{\sqrt{\sum_{i=1}^{m}x_i^2}\sqrt{\sum_{i=1}^{m}y_i^2}} \tag{4}$$

3.2.3 Jaccard's Similarity Coefficient

The ratio between the number of items jointly rated by user x and user y and the number of total items rated by them is called the Jaccard similarity coefficient. Because this metric only considers the number of items rated jointly between users and does not reflect information about users' specific rating preferences, this metric is often used to assess whether users would rate items.

$$sim(x,y)=\frac{|N(x)\cap N(y)|}{|N(x)\cup N(y)|} \tag{5}$$

Where $N(x)$, $N(y)$ denote the set of items rated by user X, Y, respectively.

3.3 Feature Selection

3.3.1 Features of Job Seekers

3.3.1.1 Education

The highest education level of job seekers is classified as high school, college, bachelor's degree, master's degree and doctoral degree. The education level of job seekers largely affects the job search. The similarity of education is calculated by setting the same education level as 1 and different as 0.

$$E(x,y)=\begin{cases}0, & \textit{the education background of x and y are different}\\ 1, & \textit{the education background of x and y are the same}\end{cases} \tag{6}$$

3.3.1.2 Gender

Job seekers are influenced by the nature of some of the jobs they are looking for, and gender can also affect the results of recommended positions. The gender similarity is calculated as

$$G(x,y)=\begin{cases}0, & \textit{the gender of x and y are different}\\ 1, & \textit{the gender of x and y are the same}\end{cases} \tag{7}$$

3.3.1.3 Expected Salary

The expected salary of job seekers is an important factor that affects the expected job position and will influence the recommendation result. According to the amount of expected monthly salary is divided into four categories.

$$I(u)=\begin{cases}0, & \textit{expected income} < ¥5{,}000\\ 1, & ¥5{,}000<\textit{expected income}< ¥10.000\\ 2, & ¥10{,}000<\textit{expected income}< ¥15.000\\ 3, & \textit{expected income} > ¥15.000\end{cases} \tag{8}$$

On this basis, the formula for calculating the similarity of expected job salary is

$$I(x,y)=\begin{cases}0, & x,y \textit{ are in the different level}\\ 1, & x,y \textit{ are in the same level}\end{cases} \tag{9}$$

3.3.1.4 Work Experience

Many companies have certain requirements for job seekers' work experience when posting job offers.

Different lengths of work experience can affect the accuracy of recommended employment information.

$$S(x,y) = \begin{cases} 0, \mathit{the\ time\ of\ working\ experience\ are\ different} \\ 1, \mathit{the\ time\ of\ working\ experience\ are\ the\ same} \end{cases} \tag{10}$$

Therefore, the calculation formula for the similarity of the total characteristics between job applicants is as follows:

$$P(x,y) = aE(x,y) + bG(x,y) + cI(x,y) + dS(x,y) \tag{11}$$

where: $a,b,c,d \in [0,1]$ and $a+b+c+d=1$.

3.3.2 Features of Jobs

3.3.2.1 Academic Requirements

When a company posts a job, it will have a basic requirement for the highest level of education of the applicant.

$$R(x,y) = \begin{cases} 0, & \mathit{the\ education\ requirement\ of\ x\ and\ y\ are\ different} \\ 1, & \mathit{the\ education\ requirement\ of\ x\ and\ y\ are\ the\ same} \end{cases} \tag{12}$$

3.3.2.2 Work Experience

Companies prefer to recruit employees with a certain level of experience, all things being equal, which means less training costs and quicker delivery of the same benefits. The formula is the same as equation (10).

3.3.2.3 Work Salary

Salaries vary by geography and company position. The salary offered will greatly influence the similarity of the position. The formula is the same as equation (9).

3.3.2.4 Professional Similarity

Some companies consider the degree to which a candidate's major and related skills are relevant to the position. To a large extent, the similarity of the profession will also affect the similarity of the job.

$$T(x,y) = \begin{cases} 0, & \mathit{the\ major\ and\ skills\ requirements\ of\ x\ and\ y\ are\ different} \\ 1, & \mathit{the\ the\ major\ and\ skills\ requirements\ of\ x\ and\ y\ are\ the\ same} \end{cases} \tag{13}$$

Therefore, the formula for calculating the similarity of the total characteristics between the job seekers is obtained through a comprehensive weighting calculation

$$J(x,y) = \alpha R(x,y) + \beta S(x,y) + \mu I(x,y) + kT(x,y) \tag{14}$$

where: $\alpha, \beta, \mu, k \in [0,1]$ and $\alpha + \beta + \mu + k = 1$.

Also due to the influence of the time difference factor on the students' employment environment, as the practice interval increases, the similarity of job seekers tends to be smaller. The calculation formula at this point is

$$sim(x,y) = P(x,y) * e^{\alpha(t_x - t_y)} \tag{15}$$

where t denotes the graduation time of students, α denotes the influence coefficient of time factor, is a constant value.

The information between job seekers can be fitted to the degree of matching between the employer and the job seeker and the degree of compliance with the attributes of special evidence, so as to achieve the recommendation results.

$$sim(x,S) = \frac{\sum_{i=1,j=1}^{n} (x_i - \bar{x})(s_j - \bar{s})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \sqrt{\sum_{j=1}^{n} (y_j - \bar{y})}^{2}} \tag{16}$$

where x denotes the set of characteristic values of job seekers, S denotes the set of characteristic values of enterprise positions.

3.4 Evaluation Metrics

The evaluation metrics used for different recommendation tasks are often different. Scoring prediction tasks usually require prediction accuracy.

The recommendation performance is evaluated in terms of accuracy, recall, and F-score of the recommendation algorithm evaluation metrics by varying the size of k (the number of nearest similar users). $R(u)$ denotes the list of recommendation items calculated by the model, and $T(u)$ denotes the actual list of users' favorites on the test set.

3.4.1 Precision Ratio

Precision is used to reflect how many items in the predicted recommendation list are of interest to the user in the list. The formula is as follows.

$$\mathrm{Pr}\,ecision = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |R(u)|} \tag{17}$$

3.4.2 Recall Ratio

This ratio is used to represent how many items in the user's true favorite list are predicted by the recommendation algorithm. The formula is as follows

$$\mathrm{Re}\,call = \frac{\sum_{u \in U} |R(u) \cap T(u)|}{\sum_{u \in U} |T(u)|} \tag{18}$$

3.4.3 F- score

F-score can combine both accuracy and recall, It can be regarded as a weighted average of model accuracy and recall. The formula is as follows:

$$F = 2 \bullet \frac{\Pr ecision \bullet \operatorname{Re} call}{\Pr ecision + \operatorname{Re} call} \tag{19}$$

4. RESEARCH RESULT

The research first constructs the scoring matrix between job seekers and jobs, then calculates the similarity of job seekers and similarity of jobs respectively, and uses the user-based collaborative filtering algorithm, the project-based collaborative filtering algorithm and the K-means clustering-based collaborative filtering algorithm to calculate the recommended relevant jobs respectively. By changing the size of the number of nearest similar users k, the performance is evaluated in terms of three indexes: accuracy, find-back bar and F-score, respectively.

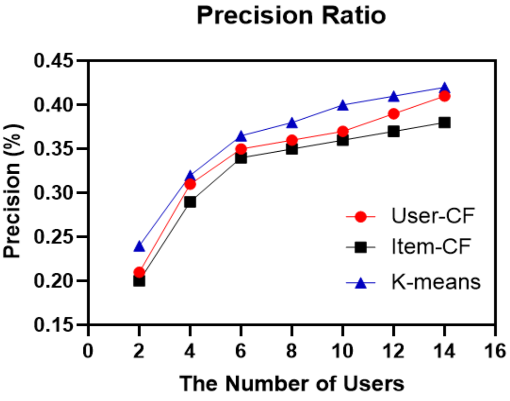

Figure 2: Comparison of the Precision Ratio of the Recommended Results

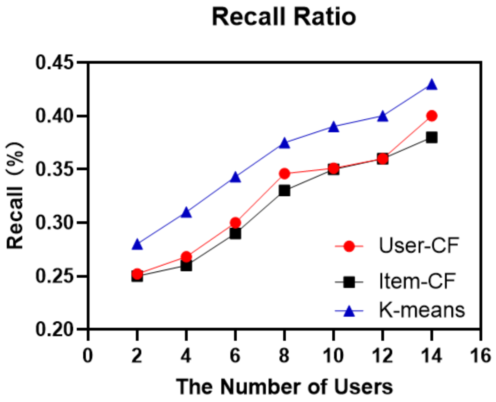

Figure 3: Comparison of the Recall Ratio of the Recommended Results

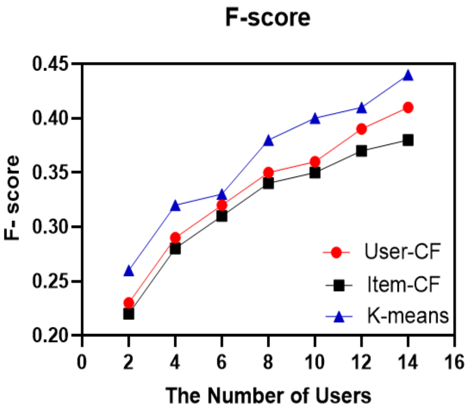

Figure 4: Comparison of the F-score of the Recommended Results

The results illustrate that the accuracy, recall and F-score increase with the growth of the number of users in a certain range. From Figure 2, it is noted that the accuracy of the k-means based collaborative filtering algorithm is significantly higher than the other two traditional recommendation algorithms for a certain number of users. From Figure 3, it indicates that the recall ratio of the improved collaborative filtering algorithm is about 3% higher than that of the traditional algorithm, and from Figure 4, the F-score of the k-means-based collaborative filtering algorithm is about 0.05 higher.

5. CONCLUSIONS

This paper combines the clustering algorithm and the collaborative filtering algorithm, and improves the similarity calculation of job seekers and corresponding jobs, and proposes an employment recommendation system based on the improved collaborative filtering algorithm of K-means.

However, when the actual recommendation system is applied, it is still a challenge to ensure the accuracy of the recommendation system while maintaining the efficiency as the data increases day by day. At the same time, when collecting the basic registration information and the implicit information of users, it involves the problem of users' personal privacy, and it is worth further research and improvement to provide user services while effectively protecting users' privacy.

REFERENCES

[1] Zeng K., Wu X., Huang Z., (2022) Research on Employment Status and influencing Factors of College Students based on SWOT Analysis --Taking Business, Logistics and Economics and Management Majors as Examples. China Storage and Transportation, 02: 94-95.

[2] Liu Y., (2020) Study on current Situation, Problems and Countermeasures of College Students' Employment Guidance, Fujian Tea, 04: 361-362.

[3] Zhang S., (2021). On the Difficulty of Employment of College Students from the Perspective of Public Economics. Science and Technology Information, vol. 17, 164-166.

[4] Zhao J., Zhuang F., Ao X., He Q. Jiang H. (2021) Survey of Collaborative Filtering Recommender Systems. Journal of Cyber Security, vol. 6, No. 5:17-34

[5] Li H., Cao H., Wang X., Liu Y. (2019). Research on Employment Recommendation Model of College Graduates based on big Data Analysis[J]. Chinese Metallurgical Education 03:93-97.

[6] Tsaic F., Hung C. (2012). Cluster ensembles in collaborative filtering recommendation[J]. Applied soft computing, 12(4): 1417-1425.

[7] Li S., Zhang Y., Zhang H. (2020) “Collaborative Filtering Recommendation Algorithm Based on NKL and K-means Clustering. J. Henan science,38(01): 6-12

[8] Huang L., Jiang B., Lyu S., Liu Y., Li D., (2018) Survey on Deep Learning Based Recommender Systems. Chinese Journal of Computers (07),1619-1647.