CCF Transactions on Pervasive Computing and Interaction (2019) 1:260–274
https://doi.org/10.1007/s42486-019-00022-1

REGULAR PAPER

Job recommendation algorithm for graduates based on personalized preference

Qing Zhou[1] · Fenglu Liao[1] · Chao Chen[1] · Liang Ge[1]

Received: 14 June 2019 / Accepted: 31 October 2019 / Published online: 28 November 2019

Abstract
It is challenging for graduates to find a proper job. Unlike those with occupational history, graduates generally are short of work experience and the support from social network, so they have to face hundreds of recruitment companies. The process of applying for a job is time-consuming, especially in preparing and attending tests and interviews. Not knowing which companies are most proper for them, graduates need to devote their energy and time to preparing for each potential recruitment. This job-hunting strategy can easily lead to employment dissatisfaction or failure. Therefore, it is very helpful to recommend a few most suitable jobs to graduates. Collaborative filtering (CF) method is currently the most frequently adopted and effective recommendation algorithm, but it cannot be directly applied to job recommendation for graduates because graduates generally have no historical records on employment. Besides, job recommendation should take into account graduate preferences for jobs, such as enterprise types and company locations, which are crucial to job choices. To address these challenges, we first analyze the pattern of job choices of graduates. Based on this, we propose a personalized preference collaborative filtering recommendation algorithm (P2CF), which can not only recommend jobs for graduates through massive campus records, but also identify graduate personal preferences for jobs. Graduates are first clustered into different groups according to their academic performances and family economic conditions. Then Bayesian personalized ranking (BPR) method is introduced to calculate the scores of graduate groups to jobs. Finally the scores and graduate personalized preferences are combined to recommend a few potential jobs. P2CF is a recommendation algorithm with hierarchical structure, which takes account of both the group records of job choices and the individual preferences for jobs. Experimental results show that P2CF on job recommendation outperforms state-of-the-art CF methods and identifies graduate personalized preference for jobs accurately.

Keywords Job recommendation · Collaborative filtering · Graduate groups · Personalized preferences

1 Introduction

The number of graduates exceeds 8,000,000 in 2018 in China. Through all over the world, it is always a challenge for graduates to find a satisfactory job due to the keen competition among a legion of candidates (Liu et al. 2017; Paparrizos et al. 2011). Moreover, unlike people with occupational history, graduates generally are short of work experience and the support from social network, so they have to screen a few jobs from hundreds of companies and organizations (Liu et al. 2017; Peterson 2014). The process of applying for a job, especially in preparing and attending tests and interviews, is quite time-consuming. If graduates cannot discern what jobs are proper for them, they have to devote their energy and time to preparing for each potential recruitment. This job-hunting strategy can easily lead to employment dissatisfaction or failure. Therefore, it is valuable to recommend a few suitable jobs to graduates, so that they can focus on the applications of these jobs suitable for them, increasing both the efficiency and success rate of job applications.

To address this problem, some studies recommend jobs based on the similarity between the skills extracted from a resume and the skills required by a job position (Al-Otaibi and Ykhlef 2012; Patel and Kakuste 2017; Razak et al. 2014). The accuracy of this method relies on the objectivity of skills described in the resume, but it is difficult to achieve

✉ Qing Zhou
tzhou@cqu.edu.cn

Fenglu Liao
Fenglu_liao@cqu.edu.cn

[1] School of Computer Science, Chongqing University, Chongqing, China

because a resume is essentially a self-report. Collaborative filtering (CF) method is the most frequently used and effective recommendation technique currently (Chen et al. 2017; He et al. 2016; Li et al. 2017), such as the recommendation of movies, music, shopping and tourism places (Chen et al. 2017; Linden et al. 2003; Nilashi et al. 2015). An item-based CF method has been proposed for online job-hunting based on users' job application records (Zhang et al. 2014). It proves that CF method has remarkable effect on job recommendation. While CF technique predicts which items users will choose according to a large amount of data collected from their past behaviors and records (Wei et al. 2017; Su and Khoshgoftaar 2009). As most graduates apply for jobs the first time in their lives, they have no occupational records and their behaviors in looking for jobs are known for little. Therefore, CF method cannot be directly applied to recommend jobs for graduates. In addition, CF method does not take account of user preferences for items, while these preferences are crucial for the choice of an item (Chen et al. 2017; Paparrizos et al. 2011; Wang et al. 2017). Similarly, graduate preferences for jobs will affect the accuracy in job recommendation.

To sum up, the construction of a job recommendation system for graduates faces three challenges:

1. how to measure the ability and conditions of a graduate in a relatively objective way;
2. how to apply CF method to recommend jobs for graduates with no historical job records;
3. how to extract graduate preferences for jobs accurately and integrate them with CF method.

To solve the first problem, we draw support from massive campus records stored in information systems, or the "campus big data" as named in Nie et al. (2016). Campus big data can reflect a wide range of student characteristics and behaviors on campus. Those information are more objective than a self-reported resume because they are records of actual behaviors and events. Therefore, the job recommendation system based on "campus big data" is promising to achieve a higher accuracy than those based on resumes or questionnaires.

To solve the second and the third problems, we first analyze the pattern of job choices of graduates. We find that different types of graduates have different considerations about job attributes and locations when they choose jobs. Based on this, we propose a personalized preference collaborative filtering recommendation algorithm (P2CF), which is a job recommendation framework with hierarchical structure. First, graduates are divided into different groups with reference to their academic performances and family economic conditions. Then a latent factor model is used to calculate the scores of graduate groups on jobs. Finally, these scores are combined with graduate personalized preferences, such as their preferences for positions, employer types and job locations, to recommend a number of potential jobs for graduates. Experimental results show that P2CF outperforms competing methods ranging from CF-based method and content-based method.

The main contributions of the paper can be summarized as follows:

- To apply CF method to recommend jobs for graduates without historical job records, we propose a recommendation algorithm with hierarchical structure, which takes account of both the group records of job choices and the individual preferences for jobs.
- We analyze the pattern of job choices of graduates and propose a job recommendation method (P2CF), which can not only recommend jobs for graduates, but also identify graduate preferences for jobs.
- We introduced massive campus records so as to evaluate the academic performances and family economic conditions of a graduate in a relatively objective way.

2 Related work

2.1 Collaborative filtering

CF is one of the most frequently adopted methods in recommendation systems. Instead of analyzing the contents of items, it employs users' ratings on items to make predictions and recommendations (Wei et al. 2017; Su and Khoshgoftaar 2009). Early CF method is based on neighborhood, as item-based CF and user-based CF. Nevertheless, user-item interaction matrix is sparse, which is not suitable to recommend items by CF based on neighborhood. To address this problem, CF method based on latent factor model is employed, such as matrix factorization (MF) technique (He et al. 2016; Koren et al. 2009). MF factorizes the user-item interaction matrix into two d-dimension matrices, i.e., an user latent factor matrix and an item latent factor matrix. The multiplication of two matrices is close to the original matrix, thus preserving user-item ratings. Unlike standard MF models, Bayesian personalized ranking (BPR) tries to preserve users' preference between two items (Rendle et al. 2009). BPR is reported to perform better than many other MF models such as singular value decomposition (SVD) (Almalis et al. 2015; Rendle et al. 2009), and has been successfully applied in many applications, such as media recommendation (Chen et al. 2017) and friend recommendation in social network (Ding et al. 2017). BPR is also adopted in this paper for job recommendation.

An item-based collaborative recommender system has been proposed for online job-hunting based on users' job

application records (Zhang et al. 2014). It proves that CF method has remarkable effect on job recommendation. CF technique, however, cannot be directly applied to recommend jobs for graduates, because they generally have no job records. This is one of the challenges facing us if we apply BPR to recommend jobs for graduates.

2.2 Graduate job recommendation

At present, career centers of institutions and colleges mainly provide manual job recommendation to graduates based on information from questionnaire surveys (Al-Otaibi and Ykhlef 2012; Nguyen et al. 2013). Other graduate job recommendation systems are implemented based on the similarity calculation (Almalis et al. 2015; Liu et al. 2017; Patel and Kakuste 2017; Razak et al. 2014). For example, Patel and Kakuste (2017), Razak et al. (2014) have proposed a job recommendation system, which match the similarity between the skills extracted from a resume and the skills required by a job. Both of the above strategies are unreliable because the questionnaire surveys and resumes are self-report results.

To address this problem, a few job recommendation systems based on objective records have been proposed (Liu et al. 2017; Nie et al. 2016). For example, Liu et al. (2017) proposed a job recommendation model for graduates based on their similarities among various campus records. Although the performance of this model is not so good as BPR, as shown in Sect. 5, the thought of "similar students may have similar preferences for jobs" motivates us to cope with the problem facing CF and BPR by clustering similar students into the same group.

2.3 Campus big data

Campus big data refers to the massive collection of records and data sets produced by information systems on campus (Nie et al. 2016). These data can not only reveal student behaviors such as the browse of websites and consumption in the campus canteens, but also reflect students' characteristics such as their academic performances and life styles. Compared with questionnaires data, campus big data is a more objective information source, because it records students' behaviors in real situations. For example, university students who need financial aid are identified through analyzing their campus card usages (Guan et al. 2015), because students with poor family economic conditions tend to eat more often in the canteen and have a smaller average cost per meal. This inspires us to construct a probability model to learn graduate family economic conditions by using their consumption records on campus.

Nie et al. (2016) have employed students campus records to predict their job choices after graduation by using a number of supervised learning models such as Logistic regression and random forest. It shows that the academic performances and family economic conditions affect student job choices to a large extent, which motivates us to cluster students according to these two characteristics. The predicted results in Nie et al. (2016) are a number of coarse categories such as "abroad further study", "domestic further study", and "seeking jobs", while our method tries to recommend specific jobs for graduates.

3 Pattern analysis of graduate job choices

3.1 Datasets

We investigate 1389 graduates from the college of computer science in one university. All of them were admitted to enterprises, institutions or governmental agencies in China or abroad before the summer of 2017. The datasets include following information and records:

- Personal information records, including basic information of graduates such as gender, age, hometown, and whether the graduate is non-poor student.
- Academic performances records. There are total 106,514 academic performances records for 1389 graduates throughout their four-year studies in the university. Features related to graduate academic performances (GPI) are extracted, including the number of failed courses, the average scores of all courses and the average scores of mathematics courses and computer science major courses.
- Consumption records of campus card. There are total 144,871 consumption records which are produced when 1389 graduates use their campus cards for eating or shopping in campus in a month. Statistical features related to graduate family economic conditions (GEI) are extracted, including the average times of canteen consumption in a month and the average cost per meal.
- Job records. There are 1389 job records for all graduates, who found 414 different jobs. A job record includes employer name, employer type (e.g., state-owned enterprise, government agency), job location and the job position (e.g., technical post, non-technical post). 33 job attributes are extracted including engineer technicians, further studying, civil servant, going aboard, finance, education, self-employment, manufacturing, services and so on.
- Geographical location records, including latitude and longitude of 34 provincial administrative regions in China.
- Regional demographic and economic records, including the total population, the urban–rural population ratio and

the average amount of consumption and income of 34 provincial administrative regions in China.

3.2 Extracting of features

3.2.1 Extracting of regional features

The employment situation of graduates in past years shows that the locations of jobs are major consideration when graduates choose jobs. Figure 1a, b present the distribution of graduates' hometown locations and job locations respectively. We can find that a large proportion of graduates come from Sichuan and Chongqing, and job locations are concentrated in five regions: Sichuan, Chongqing, Beijing, Shanghai and Guangdong. This distribution of job locations shows that graduates prefer economically developed cities, such as Beijing, Shanghai and Guangdong or cities they are most familiar with, such as Sichuan and Chongqing. So, regional economic index and familiarity index are two crucial features of job locations, as defined as follows respectively.

The regional economic index (*REI*) indicates the degree of economic development in a region. It is generally believed that the more developed a region is, the higher its consumption and income will be, and so will be its total population and urban–rural population ratio. Hence, these characteristics can be used to estimate *REI* of a region:

$$REI = \sum_{i=1}^{n} a_i C_i, \tag{1}$$

where C_i denotes various economic characteristics, including the total population, the urban–rural population ratio, the average amount of consumption and income of a region, and the weight a_i is estimated by Pearson correlation coefficient between the feature C_i and the gross domestic product (GDP) of the region.

The regional familiarity index (*RFI*) indicates people's familiarity with a region. Generally, people are relatively familiar with the areas where they have lived for a long time, such as the location of their hometown. Therefore, people's familiarity with such long-term living area is considered as 1. When the distance between two regions increases, people's customs and living habits will be different. This difference will lead to a decrease in familiarity. So the distance from area where familiarity equals 1 is used to measure graduate familiarity index. In this paper, there are two areas where familiarity equals 1, namely, the location of hometown and graduate school. *RFI* is calculated as follows:

$$RFI = \max(HFI, SFI), \tag{2}$$

where *HFI* and *SFI* represents the regional familiarity index over hometown and graduate school respectively, as calculated as follows:

$$HFI = \mathrm{a}^{-\mathrm{D(h,l)}}, \tag{3}$$

$$SFI = \mathrm{a}^{-\mathrm{D(s,l)}}, \tag{4}$$

where $D(h, l)$ is the distance between job location (l) and hometown (h) and $D(s, l)$ is the distance between job location (l) and the location of graduate school (s). a is used to adjust the shape of function, and is set to 1.15 in this paper.

3.2.2 Extracting of graduate features

Previous studies show that graduates with different academic performance and family economic conditions often have different job choices (Liu et al. 2017; Nie et al. 2016; Nie et al. 2018). For example, graduates with good academic performance are more likely to enter well-known enterprises, while students with good family economic conditions are more likely to go abroad for further study.

Academic performances of a graduate can be estimated from their performances in the courses they have learned (Liu et al. 2017; Nie et al. 2016). We define the graduate academic performances index (*GPI*) as follows:

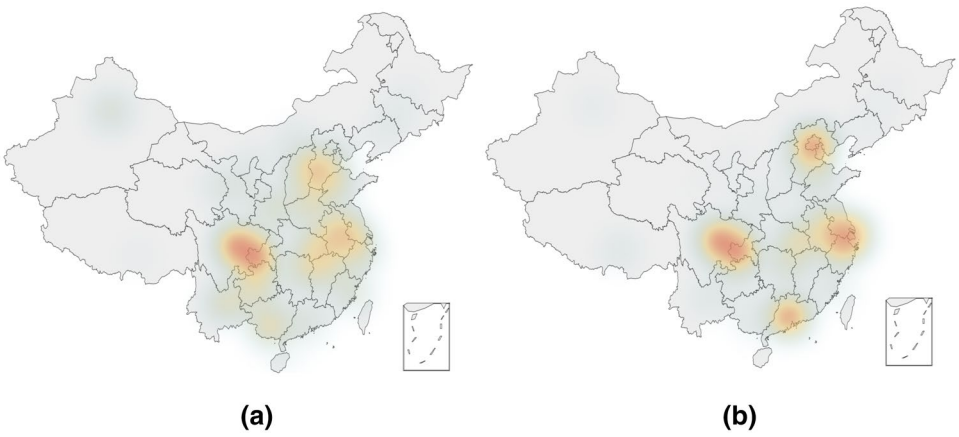


Fig. 1 **a**, **b** Distribution of the location of graduate's hometowns and jobs respectively

$$GPI = \frac{\sum_{i=1}^{n} PI_i}{n}, \tag{5}$$

where PI_i is the normalized value of performance features, including the number of failed courses, the average scores of all courses, and the average scores of mathematics courses and computer science major courses.

Graduate family economic conditions index (*GEI*) is used to measure the conditions of graduate family economic. About one fourth of students accept financial aids every year, because of the poor family economic conditions. So *GEI* can be modeled as the probability that the graduate is a non-poor student:

$$GEI = P\{Y = 0|X\}, \tag{6}$$

where $Y = 0$ denotes that the graduate is a non-poor student, and X denotes the features related to graduate economic conditions, including the average times of canteen consumption in a month, the average cost per meal, the economic index of hometown (*REI*), and whether or not the graduate comes from rural areas. The probability in Eq. (6) is modeled by logistics function:

$$GEI = \frac{1}{1 + e^{-(w^T X + b)}}, \tag{7}$$

where w and b are estimated from sample data.

All above four features (REI, RFI, GPI and GEI) are normalized to [0,1] by Min–Max normalization.

3.3 Pattern analysis of job choices

The choice of job is influenced by many factors, such as gender, family background, academic performance and so on. We first analyze the pattern of job choices with regards to these three factors.

3.3.1 Analysis of job attributes

We collected 33 attributes for each job, such as the types of jobs, the working posts and the industries of companies. Tables 1 and 2 show the analysis of different types of graduates for part of job attributes. From Tables 1 and 2, we can find that:

1. Male and female graduates have different tendency in choosing jobs with various attributes. The proportion of males in computer posts is about 10 percentage points higher than that of females. Among them, about two-third males are engaged in engineer posts, while only about half of females are engaged in engineer posts. On the contrary, the proportion of females in civil service positions is about 3 percentage points higher than that of males.
2. Graduates with different academic performance have significant difference in the types of jobs. About 96% of graduates with good academic performances choose to continue their studies, while about 93% of graduates with poor academic performances choose to work in companies or other organizations. The proportion of graduates with poor academic performances choosing self-employment is as high as 21%, possibly because they cannot find satisfactory job due to their poor academic performances.
3. The proportion of graduates who choose further study in China is almost the same for graduates with good family economic conditions and for those with poor conditions. However, the proportion of graduates with good family conditions going abroad is about 9% higher than those with poor conditions. And about 36% of graduates with poor family economic conditions choose to work in private enterprises, which is about 10% higher than those with good family economic conditions.

Table 1 Analysis of male and female graduates for part of job attributes

Working posts	Males (%)	Females (%)
Civil servant	1.69	4.96
Other technical personnel	15.04	22.07
Office staff	6.96	9.91
Business and service personnel	1.13	1.80
Engineer	66.54	50
Teaching staff	2.82	3.60
Legal professionals	0	0.45
Scientific researcher	3.38	2.70
Financial personnel	2.44	4.51

3.3.2 Analysis of job locations

As shown in Sect. 3.2, the economic condition and graduates' familiarity with the job locations are important factors for job choices. Figure 4 shows the distribution of regional economic index and familiarity index for different types of graduates. From Fig. 2, we see that:

1. Graduates with good academic performances generally have high regional economic index, with the median value approaching 0.6. Compared with male graduate, females graduates have a lower regional economic index. And graduates with better economic conditions have a higher regional economic index.
2. Graduates with poor academic performance and better economic conditions generally have higher familiarity

Table 2 Analysis of graduates with good or poor GPI and graduates with wealthy or poor GEI for part of job attributes

Job types	GPI (good) (%)	GPI (poor) (%)	GEI (wealthy) (%)	GEI (poor) (%)
Foreign-funded enterprises	0	4.94	4.81	6.09
Government-affiliated institutions	0	2.47	2.94	1.79
Private enterprise	3.30	34.57	25.58	35.48
Going abroad	6.59	2.47	8.84	0.36
Further study	89.01	4.94	32.87	32.62
State-owned enterprise	1.10	20.99	14.88	16.49
Self-employment	0	20.98	6.67	4.66
Army	0	8.64	3.41	2.51

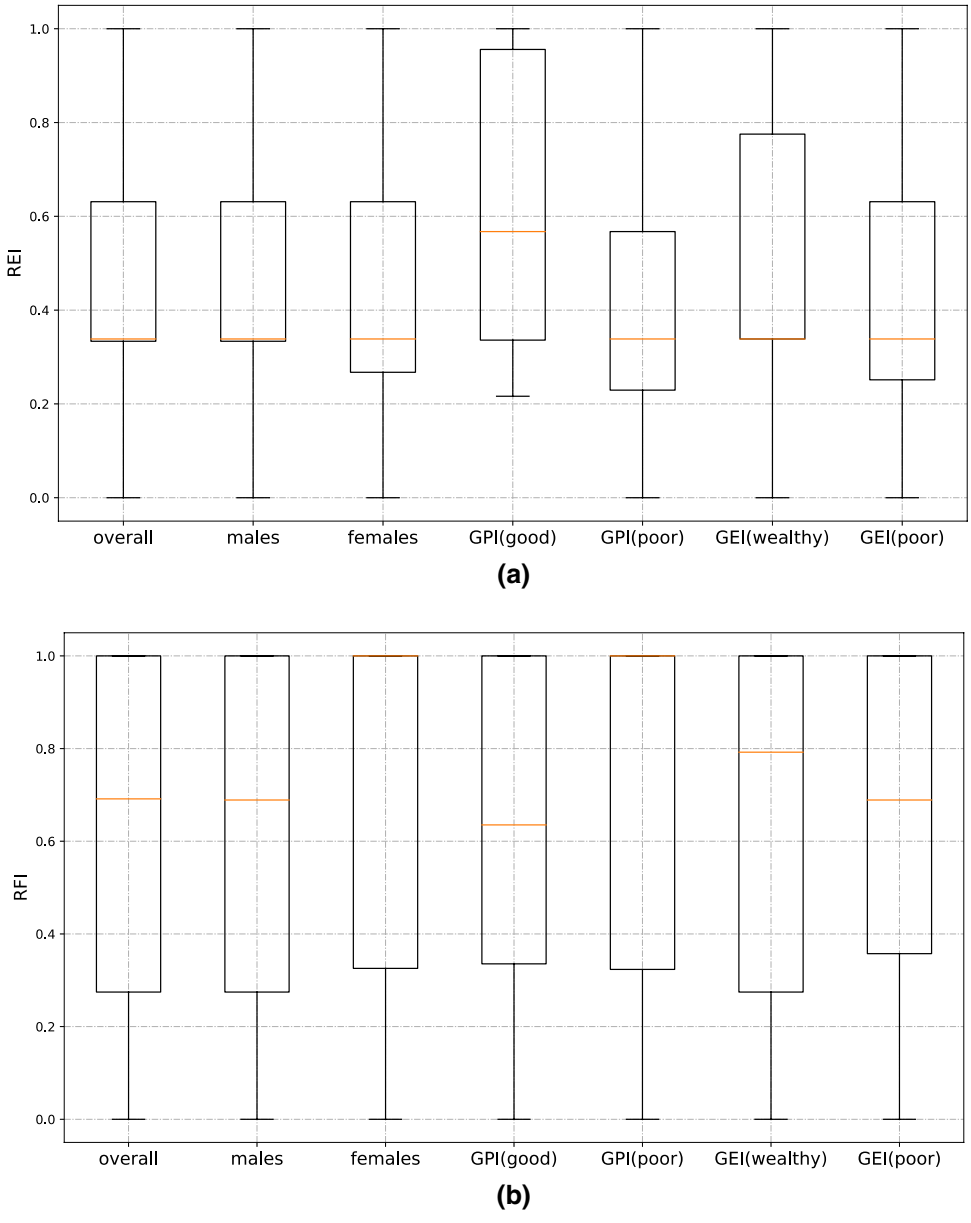


Fig. 2 Distribution of regional economic index and familiarity index of different types of graduates. **a**, **b** Distribution of regional economic index and regional familiarity index respectively

index, which indicates that these graduates take more account of the familiarity of the job locations. The familiarity index of graduates with good academic performance, on the contrary, are lower than the average, indicating that these graduates consider the familiarity of job location to less extent.

4 Job recommendation algorithm for graduates

4.1 General framework

Figure 3 shows the framework of P2CF recommendation algorithm, which consists of two phases, i.e., graduate group identification and graduate job recommendation.

Graduate group identification: in this phase, graduates are clustered into groups according to their academic performances, family economic conditions.

Graduate job recommendation: in this phase, group records of job choices, personalized preferences for job attributes and job locations are combined to score jobs for graduates, and a few jobs with higher scores are recommended.

4.2 Graduate group identification

Since academic performances and family economic conditions influence student job choices to a large extent (Nie et al. 2016), we can cluster graduates into different groups according to these two factors, and then recommend jobs for each group. There are three categories of commonly used clustering algorithms: partition-based clustering method, hierarchical clustering method and density-based clustering method. In view of the similarity of the results of the same type of clustering algorithm, the K-means algorithm (Jian and Wang 2010), AGNES algorithm (Davidson and Ravi 2005) and DBSCAN algorithm (Ester et al. 1996) are selected to classify the graduates.

To determine the best clustering algorithm for group identification, the performance of three clustering algorithms are tested when students are divided into eight groups, as shown in Table 3. K-means clustering algorithm is the best algorithm to identify graduate groups which achieves the highest Silhouette coefficient of 0.415. Silhouette coefficient is used to measure the similarity of an object with objects in its own clusters relative to the distance with objects in other clusters (Rousseeuw 1987). The larger the value, the better the matching degree with its own cluster. Therefore, we choose k-means algorithm to cluster graduates into groups.

4.3 Job recommendation algorithm for graduates

P2CF recommends jobs for a graduate considering three aspects: group records of job choices ($\mathbf{UV}^T$), preferences for job attributes ($\mathbf{AF}^T$) and preferences for job locations ($\mathbf{P}$).

Table 3 Performance comparison of three clustering algorithms

Clustering algorithm	Silhouette coefficient	Execution time (s)
K-means	0.415	0.085
AGNES	0.360	0.069
DBSCAN	0.029	0.013

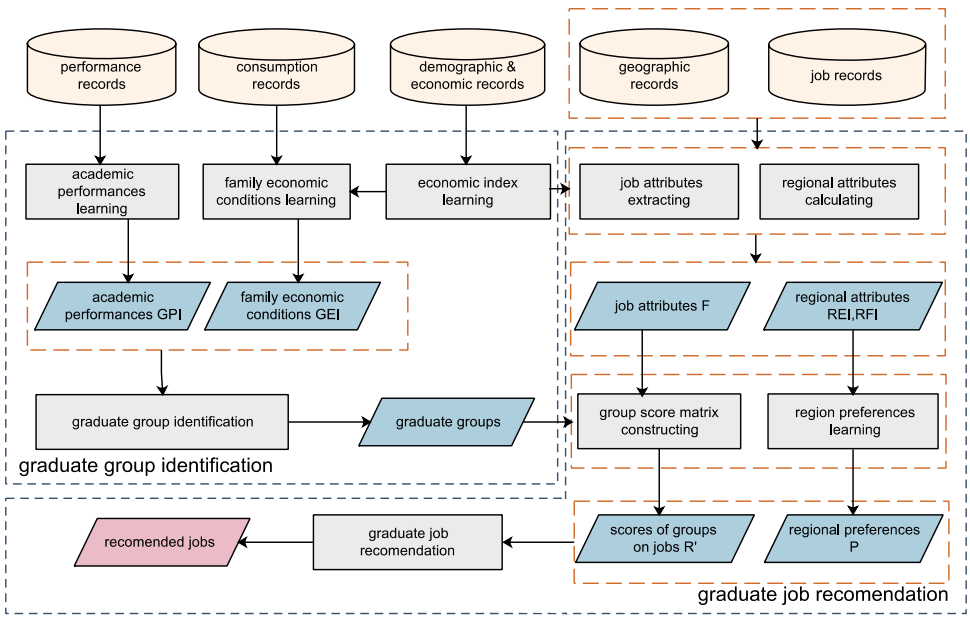


Fig. 3 Framework of our job recommendation algorithm

Group records of job choices reveal the job choices of different groups. Preferences for job attributes reflect the preferences for specific types of jobs. For example, some graduates prefer stable jobs such as teachers and civil servants, while others prefer challenging jobs such as working in competitive enterprises. Finally, the preferences for job locations reflect graduate' geographic consideration when choosing a job. For example, some graduates prefer to work in economically developed areas, while others prefer to work in the place they are familiar with. We first model above three aspects before presenting the details of the job recommendation algorithm.

4.3.1 Group records of job choices

Let $\mathbf{R} \in \mathbb{R}^{M\times N}$ denotes the job choice matrix, where M is the number of graduate groups and N is the number of jobs, and R_{ui} represents the number of times the graduates in group u have chosen job i. Let $\mathbf{U} = \left[\mathbf{U}_1, \ldots, \mathbf{U}_D\right] \in \mathbb{R}^{M\times D}$ and $\mathbf{V} = \left[\mathbf{V}_1, \ldots, \mathbf{V}_D\right] \in \mathbb{R}^{N\times D}$ be two latent factor matrices, a new matrix $\hat{\mathbf{R}}$ approximate to $\mathbf{R}$ can be estimated as follows:

$$\hat{R}_{ui} = \mathbf{U}_u \mathbf{V}_i^T, \tag{8}$$

where $\hat{R}_{ui}$ is the element of $\hat{R}$ in uth row and ith column and then $\mathbf{U}$, $\mathbf{V}$ are chosen to minimize the difference between R and $\hat{R}$:

$$\arg\min_{\mathrm{U},V} \sum_{(u,i)\in\mathcal{R}} \left(R_{ui} - \hat{R}_{ui}\right)^2 + \lambda\left(||\mathbf{U}||^2 + ||\mathbf{V}||^2\right). \tag{9}$$

Bayesian personalized ranking (BPR) is a popular method to solve the recommendation problem (Chen et al. 2017; Rendle et al. 2009). Instead of point-wise learning in traditional matrix decomposition, BPR models a triplet $\mathcal{R}_B$ with one user and two items to indicate the user preferences ranking to items. $\mathcal{R}_B$ is generated as:

$$\mathcal{R}_B = \left\{(u,i,j)|i \in \mathcal{R}(u) \; and \; j \in \mathcal{R}(u)_{<i}\right\}, \tag{10}$$

where $\mathcal{R}(u)$ denotes the set of items that user u interacted with, and $\mathcal{R}(u)_{<i}$ denotes the set of items that user u scores lower than item i. So a triplet (u,i,j) indicates that user u prefers item i to item j.

BPR enlarges the difference of a user's score between frequently chosen items and seldom chosen items via the maximization of following objective function:

$$\arg\max_{\mathrm{U,V}} \sum_{(u,i,j)\in\mathcal{R}_B} ln\sigma\left(\hat{R}_{uij}\right) - \lambda\left(||\mathbf{U}||^2 + ||\mathbf{V}||^2\right), \tag{11}$$

where σ is the function *sigmoid*, and

$$\hat{R}_{uij} = \hat{R}_{ui} - \hat{R}_{uj}, \tag{12}$$

where $\hat{R}_{ui}$ is the latent factor model as Eq. (8).

4.3.2 Preferences for job attributes

Different groups may have distinct preferences for jobs. These preferences $\mathbf{A} = \left[\mathbf{A}_1, \ldots, \mathbf{A}_L\right] \in \mathbb{R}^{M\times L}$, where M and L are the number of user groups and job attributes. $\mathbf{F}$ denote the job attribute matrix, $\mathbf{F} = \left[\mathbf{F}_1, \ldots, \mathbf{F}_L\right] \in \mathbb{R}^{N\times L}$, where N is the number of jobs. By integrating a group's job choices with its preferences for job attributes, the score of group u to job i ($\hat{R}_{ui}$) can be recalculated as:

$$\hat{R}_{ui} = \mathbf{U}_u \mathbf{V}_i^T + \mathbf{A}_u \mathbf{F}_i^T. \tag{13}$$

4.3.3 Preferences for job locations

Graduates may have different preferences for job locations. Let $\mathbf{P}$ denote the regional preference matrix and P_{gi} represents the regional preference of graduate g to job i, calculated as:

$$P_{gi} = f\left(RFI_{gi}, REI_i; \theta\right), \tag{14}$$

where f is the function of probability density, and it is fitted by the multivariate Gauss distribution:

$$f(\boldsymbol{x};\theta) = \frac{1}{\sqrt{(2\pi)^2|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right), \tag{15}$$

where $= \left(RFI_{gi}, REI_i\right)^{\mathrm{T}}$, and $\theta = (\mu(\boldsymbol{x}), \Sigma(\boldsymbol{x}))$ denote the mean and covariance matrix for $\boldsymbol{x}$. So different function f will be fitted by data from different groups. REI_i is the regional economic index of job i as calculated by Eq. (1), and RFI_{gi} is the regional familiarity index of graduate g to job i, calculated by Eq. (2).

4.3.4 Job recommendation algorithm

The proposed job recommendation algorithm takes account of group records of job choices, preferences for job attributes and preferences for job locations. The score of graduate g to job i is calculated as follows:

$$\hat{R}_{gi} = \hat{R}_{ui} + \eta \cdot P_{gi}, \tag{16}$$

where η is the adjustment parameter, P_{gi} represents graduate preferences for job locations, and is calculated as Eq. (14); while $\hat{R}_{ui}$, the group's score on job i, is calculated as Eq. (13). In order to determine the parameters $\mathbf{U}$, $\mathbf{V}$ and $\mathbf{A}$ in Eq. (13), the following objective function is minimized:

$$\begin{aligned}\arg\min_{U,\mathrm{V,A}} \sum_{(u,i,j)\in\mathcal{R}_B} -ln\sigma\left\{\left(\mathbf{U}_u\mathbf{V}_i^T + \mathbf{A}_u\mathbf{F}_i^T\right) - \left(\mathbf{U}_u\mathbf{V}_j^T + \mathbf{A}_u\mathbf{F}_j^T\right)\right\} \\ + \lambda\left(||\mathbf{U}||^2 + ||\mathbf{V}||^2 + ||\mathbf{A}||^2\right).\end{aligned} \tag{17}$$

We apply the stochastic gradient descent method to solve this optimization problem.

Algorithm 1 describes the detailed process of P2CF. In step 1, the latent factor matrix **U**, **V** and the job attribute matrix **A** are randomly initialized. The triple relationship $\mathcal{R}_B$ is generated from the training set in step 2. From step 3 to step 10, a stochastic gradient descent algorithm is employed to determine **U**, **V** and **A**. When the cost of loss is minimal or its fluctuation is not obvious, the iteration is terminated. From step 12 to step 17, the preference P_{gi} is calculated and combined with $\hat{R}_{ui}$ to obtain $\hat{R}_{gi}$.

Algorithm 1 Personalized preference collaborative filtering

Input: $\mathbf{data}, \mathbf{F}, \mathbf{RFI}, \mathbf{REI}, \mathbf{D}, \lambda, \alpha, \eta$
Output: $\hat{\mathbf{R}}$

1: initialize $\mathbf{U}, \mathbf{V}, \mathbf{A}$ randomly
2: generate the triplet $\mathcal{R}_B$ from training data
3: **repeat**
4: draw (u, i, j) from $\mathcal{R}_B$
5: $\hat{R}_{ui} = \mathbf{U}_u \mathbf{V}_i^T + \mathbf{A}_u \mathbf{F}_i^T$
6: $\hat{R}_{uj} = \mathbf{U}_u \mathbf{V}_j^T + \mathbf{A}_u \mathbf{F}_j^T$
7: $\hat{R}_{uij} = \hat{R}_{ui} - \hat{R}_{uj}$
8: **for** θ in $[\mathbf{U}, \mathbf{V}, \mathbf{A}]$ **do**
9: $\theta = \theta - \alpha(-\frac{e^{-\hat{R}_{uij}}}{1+e^{-\hat{R}_{uij}}} \cdot \frac{\partial \hat{R}_{uij}}{\partial \theta} + \lambda\theta)$
10: **end for**
11: **until** convergence
12: compute $\hat{R}_{ui}$ according to (13)
13: learn personalized preferences function f according to (15)
14: **for** graduate g in testing data **do**
15: calculate regional preferences P_{gi} according to (14)
16: compute $\hat{R}_{gi}$ according to (16)
17: **end for**
18: return $\hat{\mathbf{R}}$

5 Experiment

5.1 Experimental setting

5.1.1 Baselines

The baseline methods include item-based collaborative filtering (ICF), user-based collaborative filtering (UCF), singular value decomposition matrix decomposition (SVD), Bayesian personalized ranking (BPR) and content-based filtering (CBF). Since graduates have no historical employment records, all baseline methods based on collaborative filtering are implemented on groups. Only CBF, which is not a collaborative filtering method, can be implemented based on individuals.

5.1.2 Parameter settings

There are five hyper parameters in P2CF, including the number of clusters *kn* in k-means, the feature dimension *D* of the latent factor model, the learning rate α of stochastic gradient descent algorithm, the regularize parameter λ in Eq. (17) and the adjustment parameter η in Eq. (16). These parameters are tested in candidate sets of {4, 8, 12, 16, 20}, {8, 16, 32, 64, 128}, {0.001, 0.01, 0.05, 0.1, 0.5}, {0.0001, 0.0005, 0.001, 0.005, 0.01, 0.1}, and {0.2, 0.4, 0.6, 0.8, 1.0} separately. By comparing the performance on validation set, these parameters are chosen as $kn = 20$, $D = 64$, $\alpha = 0.05$, $\lambda = 0.001$, $\eta = 0.4$.

5.1.3 Evaluation

10-fold cross validation evaluation method is adopted to evaluate the performance of P2CF. To assess the ranked list with the ground truth job, we adopt hit ratio (*HR*) and mean reciprocal rank (*MRR*), where *HR* measures whether the actual job is contained in the recommended list and *MRR* accounts for the position of hit.

5.2 Result of group identification

Figure 4 shows the performance of HR@50 and MRR@50 with respect to *kn*, the number of clusters, by using k-means algorithm. With the increase of the number of clusters, P2CF demonstrates consistent improvement over other models. When *kn* is set to 20, P2CF achieves the best performance than all other models. But the recommendation performance of P2CF is still increasing, so $kn = 25$ and $kn = 30$ are introduced to observe the performance of clustering. as shown in Fig. 5. It is found that Silhouette Coefficient decreases with the increase of *kn*. Considering both the recommendation performance and clustering effects, *kn* is chosen as 20 in this study.

Figure 6 shows the clustering results for 20 groups of graduates. Each group differs from the others in either *GPI* or *GEI*. As shown in Fig. 6, axis *GEI* is roughly divided into five regions, so we use letter from "A" to "E" to indicate graduate family economic conditions (*GEI*) from small to large; and axis GPI is divided into six regions referring group "E", so we use number from 1 to 6 to indicate graduate academic performances (*GPI*) from small to large. For example, group "E6" represents someone with good family economic conditions and outstanding academic performances and group "A1" represents someone with poor family economic conditions and poor academic performances.

Moreover, previous studies show that men and women have obvious differences in cognition pattern and personal preference. For example, men generally prefer competitive and risky activities than women (Croson and Gneezy 2009). So we use 'F' or 'M' to label these groups so as to distinguish male and female. For example, 'B1.M' denotes all male graduates in group "B1", and 'E1.F' denotes all female graduates in group "E1".

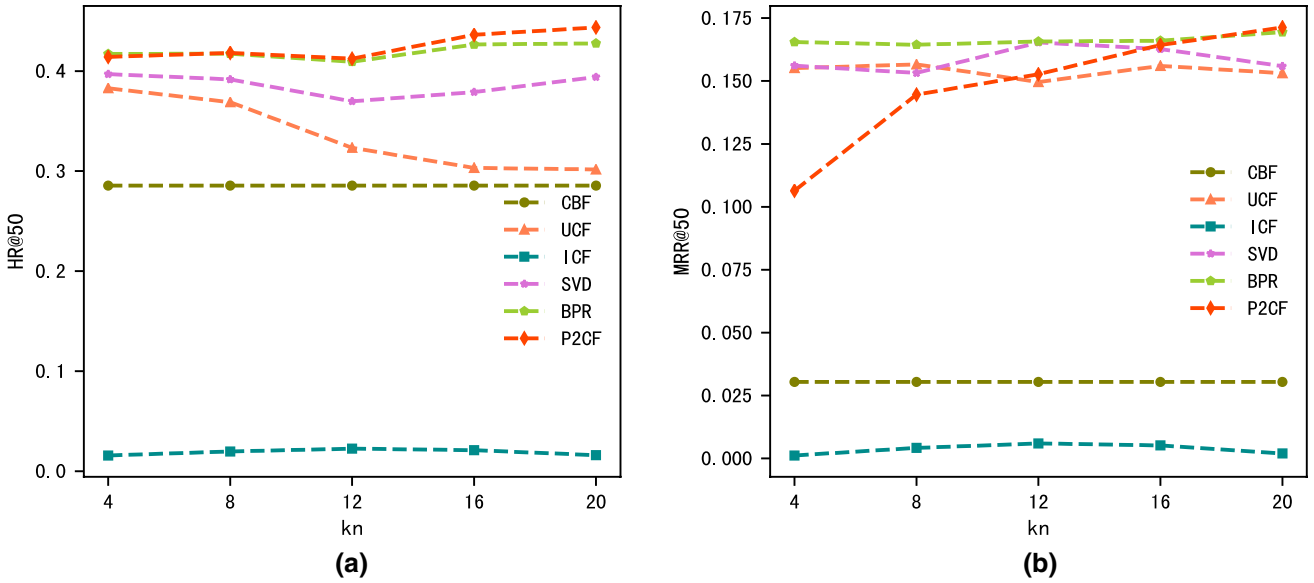


Fig. 4 Performance of HR@50 and MRR@50 w.r.t. the number of clusters in k-means

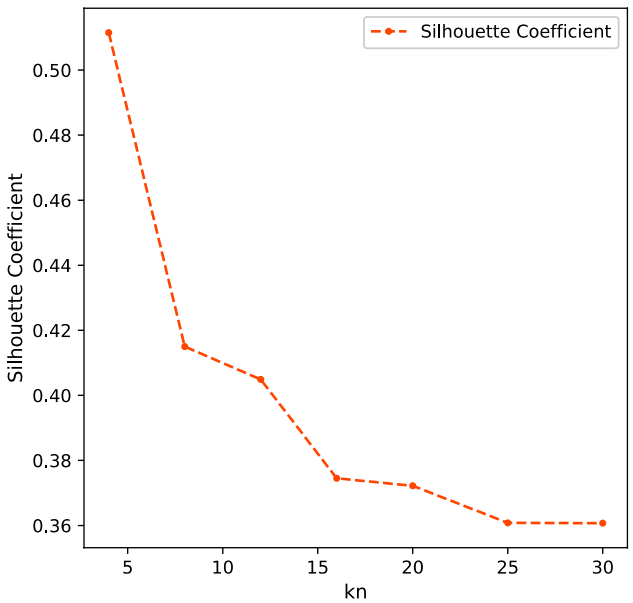


Fig. 5 Performance of clustering w.r.t. the number of clusters in k-means

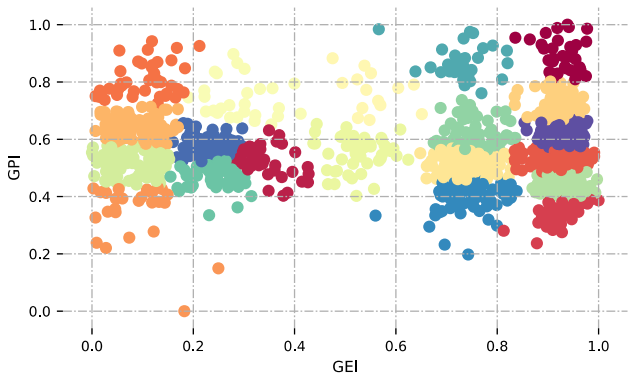


Fig. 6 Results of clustering graduates into 20 groups

5.3 Performance of job recommendation models

Figure 7 shows the performance of Top-K job recommendation, where K ranges from 1 to 50. All methods are implemented on groups except "CBF", which is implemented on individual graduate. To better observe the performance among different recommendation methods, "CBF" and "ICF" with poor performance are omitted in Fig. 7c, d. From Fig. 7, we observe that:

1. Compared with most CF method, CBF method, which relies on the similarity of users, performs poorly. It is about 20% lower in *HR* and 15% lower in MRR than P2CF.
2. The performance of UCF and ICF based on neighborhood is worse than CF method based on latent factor model, probably because they score items based on historical interaction between users and items only, ignoring users' hidden interests.
3. Of all baseline methods, SVD and BPR, which are based on latent factor model, outperform other methods. Furthermore, the proposed P2CF algorithm, however, achieves the best performance. When $K=50$, *HR* reaches 44.37%, about twice as much as CBF method; *MRR* reaches 17.14%, about 7 times as much as CBF method.

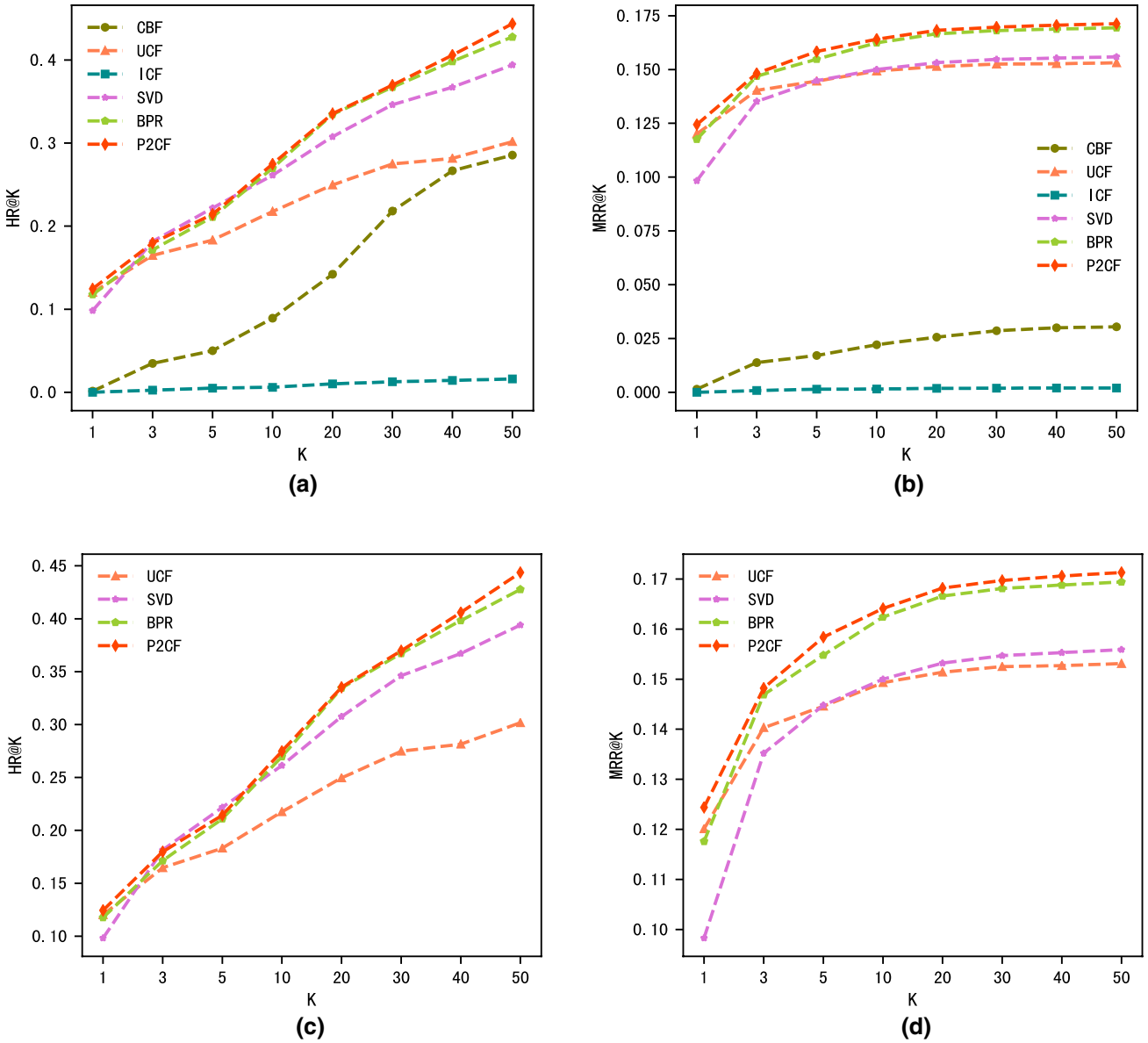


Fig. 7 Performance of Top-K job recommendation where K ranges from 1 to 50

Figure 8 shows the performance of HR@50 and MRR@50 with respect to the number of latent factors. For BPR algorithm, its performance fluctuates with the increase of dimension, but the fluctuation is not obvious from 8 to 64. When $D = 128$, it decreases rapidly. However, for P2CF algorithm, with the increase of dimension D, its performance first increases and then decreases. When $D = 64$, it reaches the maximum ($HR = 44.37\%$, $MRR = 17.14\%$). Therefore, it can be considered that the best value of hidden factor D is 64.

5.4 Effect of introducing the preferences for job attributes and locations

P2CF achieves better performance than BPR by introducing the preferences for job attributes (**A**) and locations (**P**). This effect is demonstrated in Table 4. From the table, we observe that:

1. BPR algorithm achieves the worst performance because it does not consider any graduate preferences for jobs.

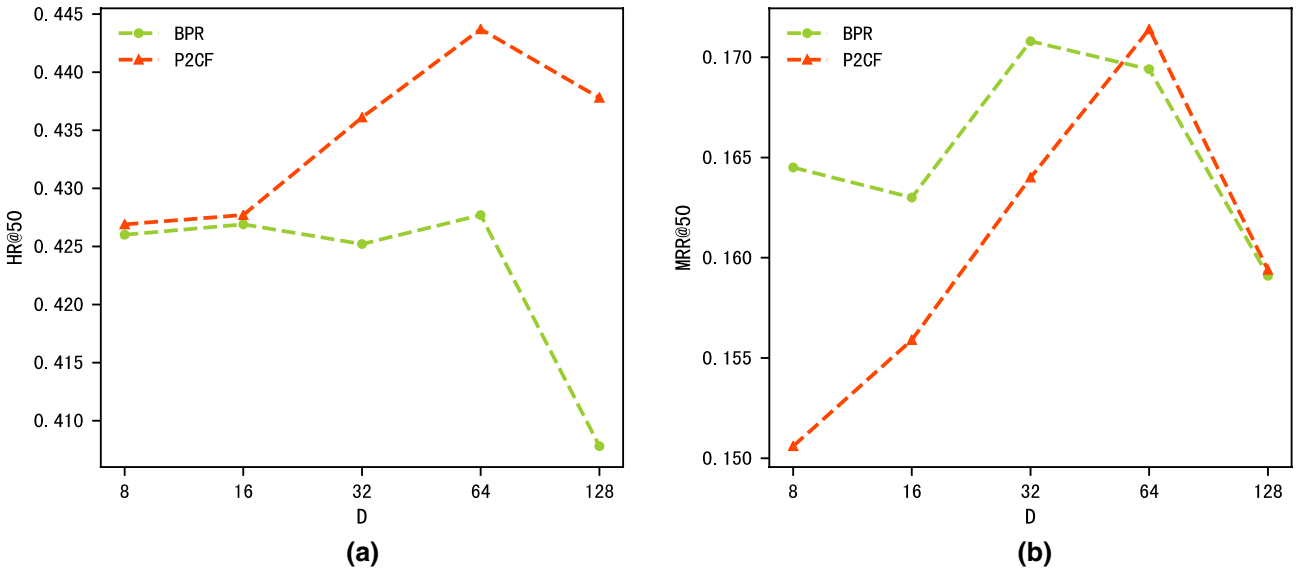


Fig. 8 Performance of HR@50 and MRR@50 w.r.t. the number of latent factors

Table 4 Effect of introducing the preferences for job attributes (**A**) and locations (**P**)

model	HR@50 (%)	MRR@50 (%)
BPR	42.77	16.94
BPR + **A**	43.11	17.30
BPR + **A**+**P** (P2CF)	44.37	17.14

2. The preferences for both job attributes (**A**) and job locations (**P**) contribute to improve the Hit Ratio. But the preferences for job locations are more effective to improve Hit Ratio than job attributes.
3. When introducing the preferences for job attributes, *MRR* increases by 0.36 percentage point. There is a negative effect on *MRR*, however, by introducing the preferences for job locations.

5.5 Display of personalized preferences

5.5.1 Preferences for job attributes

Matrix **A**, as estimated by Eq. (17), can reveal group preferences for different job attributes. Figure 9 shows the display of preferences of some groups for five typical job attributes, including engineering technician, civil servant, going abroad, non-technical post and self-employment. A larger circle indicates that the group has a higher degree of preferences for this job attribute, and circles with small radius are filtered for the sake of simplicity. From Fig. 9, we observe that:

1. Graduate groups "E1" and "E6" have a higher preferences for going abroad than other groups. We can conclude that graduates with better family economic conditions tend to work or study abroad.
2. Female groups have higher preferences for civil servants and non-technical positions than male groups. They choose the non-technical positions might due to the lack of interests or confidence in technical positions. In contrast, male graduates prefer technical jobs such as engineers than females.
3. Most of the graduates who choose self-employment are from group "A1", "E1" and "E2". These graduates are all with worse academic performances. It's likely that their academic performances are too bad to find a job.

5.5.2 Preferences for job locations

Figure 10 shows the distribution of regional economic index (*REI*) and regional familiarity index (*RFI*) in the choices of employment for several groups. The distribution for group "D3.F" and "E2.M" is depicted in Fig. 10a. A darker dot indicates that there are more graduates choosing jobs in the region with similar *REI* and *RFI*. Comparing the dots in the top with those in the bottom in Fig. 10a, we see that regional familiarity index is an important factor for graduates to choose jobs. Female graduates seldom work in regions they are not familiar with, but this is not the case for male graduates. Actually, quite a lot of male graduates would like to work in a developed region although they are not familiar.

Figure 10b shows the function f of group "D3.F" and "E2.M" fitted by multivariate Gauss distribution. The center of function f in group "D3.F" is close to (0.4, 1.0) and its

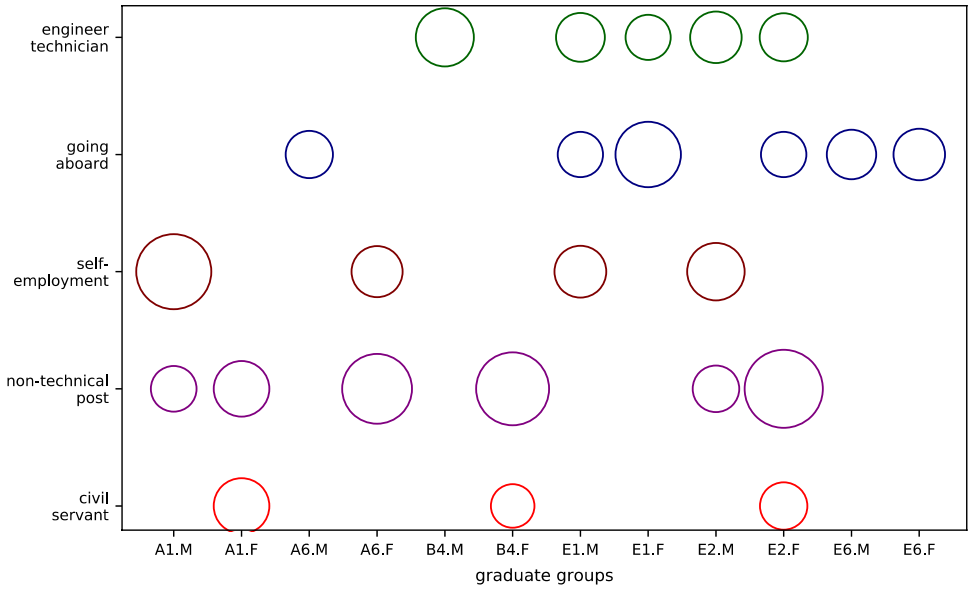


Fig. 9 Display of personalized preferences for job attributes among different groups

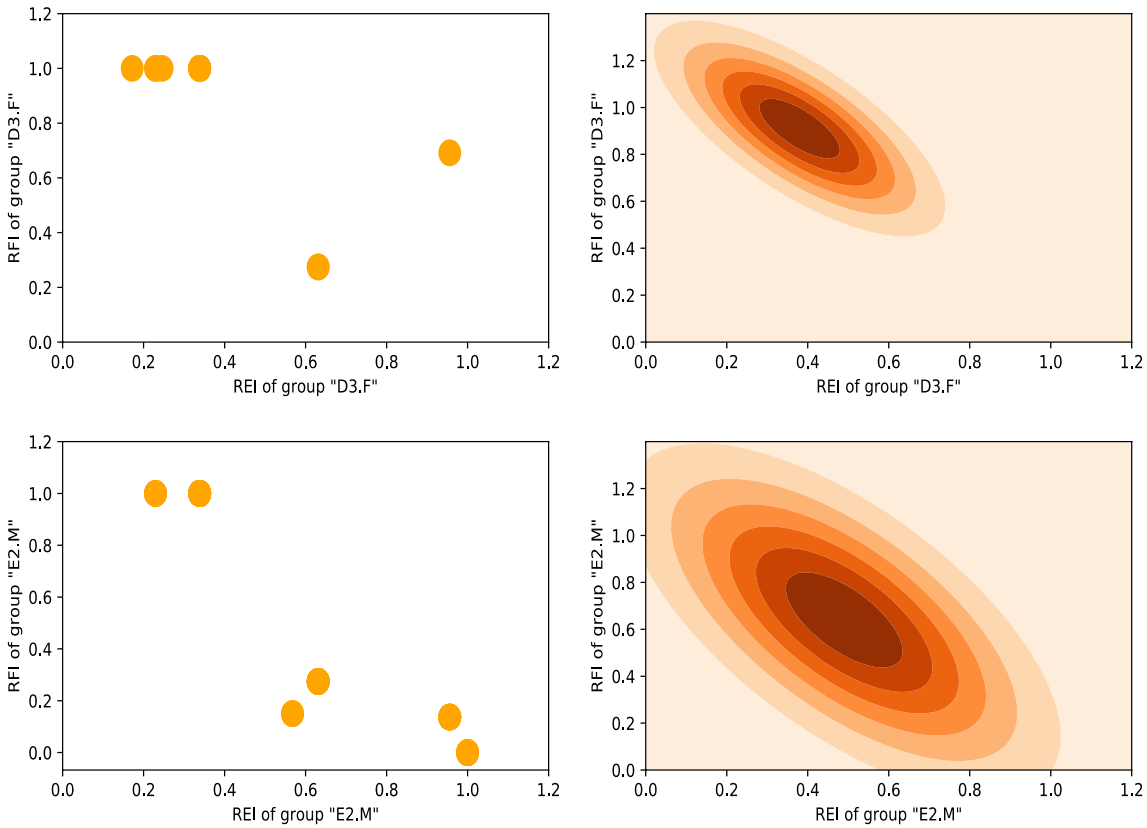


Fig. 10 **a** Distribution of regional economic index (REI) and regional familiarity index (RFI) for group "D3.F" and "E2.M"; **b** Function f of group "D3.F" and "E2.M" fitted by Multivariate Gauss distribution

variance is small, which indicates that female graduates in group "D3" are likely to choose jobs in familiar regions. While in the bottom figure, the function f of group "E2.M" is centered around (0.5, 0.6) with a large variance, which indicates that male graduates in group "E2" have diverse choice for job locations.

6 Conclusion

We proposed a personalized preference collaborative filtering algorithm, P2CF, to recommend jobs for graduates. Unlike the existing job recommendation methods, P2CF first divides graduates into groups, and then recommends jobs for graduates based on both group records of job choices and graduate preferences for jobs. By introducing the personalized preferences to collaboration filtering, the hit rate of P2CF in recommending jobs can be twice higher than traditional collaborative filtering algorithms. Hence, P2CF is a promising algorithm to increase the accuracy and efficiency in finding a proper job for graduates. Besides this, P2CF can also identify graduate preferences for jobs, helping graduates to adjust their job-hunting strategies. We expect the proposed method can be applied to graduates from more colleges and universities and adopted by career centers of these institutions.

Funding This work was funded by Fundamental Research Funds for the Central Universities Grant number 2018CDXYJSJ0026.

References

Almalis, N.D., Tsihrintzis, G.A., Karagiannis, N., Strati, A.D.: FoDRA—a new content-based job recommendation algorithm for job seeking and recruiting. In: 2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA), pp. 1–7. IEEE (2015)

Al-Otaibi, S.T., Ykhlef, M.: A survey of job recommender systems. Int. J. Phys. Sci. **7**(29), 5127–5142 (2012)

Chen, J., Zhang, H., He, X., Nie, L., Liu, W., Chua, T.S.: Attentive collaborative filtering: Multimedia recommendation with item-and component-level attention. In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 335–344. ACM (2017)

Croson, R., Gneezy, U.: Gender differences in preferences. J. Econ. Lit. **47**(2), 448–474 (2009)

Davidson, I., Ravi, S.S.: Agglomerative hierarchical clustering with constraints: theoretical and empirical results. PKDD **3721**, 59–70 (2005)

Ding, D., Zhang, M., Li, S.Y., Tang, J., Chen, X., Zhou, Z.H.:. Baydnn: friend recommendation with bayesian personalized ranking deep neural network. In: Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pp. 1479–1488. ACM (2017)

Ester, M., Kriegel, H.P., Xu, X.: A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In: International Conference on Knowledge Discovery & Data Mining (1996)

Guan, C., Lu, X., Li, X., Chen, E., Zhou, W., Xiong, H.: Discovery of college students in financial hardship. In: 2015 IEEE International Conference on Data Mining, pp. 141–150. IEEE (2015)

He, X., Zhang, H., Kan, M.Y., Chua, T.S.: Fast matrix factorization for online recommendation with implicit feedback. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 549–558. ACM (2016)

Jian, Z., Wang, H.: An improved K-means clustering algorithm. In: IEEE International Conference on Information Management & Engineering (2010)

Koren, Y., Bell, R., Volinsky, C.: Matrix factorization techniques for recommender systems. Computer **8**, 30–37 (2009)

Li, J., Wang, J., Sun, Q., Zhou, A.: Temporal influences-aware collaborative filtering for QoS-based service recommendation. In: 2017 IEEE International Conference on Services Computing (SCC), pp. 471–474. IEEE (2017)

Linden, G., Smith, B., York, J.: Amazon.com recommendations: item-to-item collaborative filtering. IEEE Internet Comput. **1**, 76–80 (2003)

Liu, R., Rong, W., Ouyang, Y., Xiong, Z.: A hierarchical similarity based job recommendation service framework for university students. Front. Comput. Sci. **11**(5), 912–922 (2017)

Nguyen, C.D., Vo, K.D., Nguyen, D.T.: Supporting career counseling with user modeling and job matching. In: Advanced Computational Methods for Knowledge Engineering, pp. 281–292. Springer, Heidelberg (2013)

Nie, M., Yang, L., Ding, B., Xia, H., Xu, H., Lian, D.: Forecasting career choice for college students based on campus big data. In: Asia-Pacific Web Conference, pp. 359–370. Springer, Cham (2016)

Nie, M., Yang, L., Sun, J., Su, H., Xia, H., Lian, D., Yan, K.: Advanced forecasting of career choices for college students based on campus big data. Front. Comput. Sci. **12**(3), 494–503 (2018)

Nilashi, M., bin Ibrahim, O., Ithnin, N., Sarmin, N.H.: A multi-criteria collaborative filtering recommender system for the tourism domain using expectation maximization (EM) and PCA–ANFIS. Electron. Commerce Res. Appl. **14**(6), 542–562 (2015)

Paparrizos, I., Cambazoglu, B.B., Gionis, A.: Machine learned job recommendation. In: Proceedings of the 5th ACM Conference on Recommender Systems, pp. 325–328. ACM (2011)

Patel, B., Kakuste, V., Eirinaki, M.: CaPaR: a career path recommendation framework. In: 2017 IEEE 3rd International Conference on Big Data Computing Service and Applications (BigDataService), pp. 23–30. IEEE (2017)

Peterson, A.: On the prowl: how to hunt and score your first job. Educ. Horiz. **92**(3), 13–15 (2014)

Razak, T.R., Hashim, M.A., Noor, N.M., Halim, I.H.A., Shamsul, N.F.F.: Career path recommendation system for UiTM Perlis students using fuzzy logic. In: 2014 5th International Conference on Intelligent and Advanced Systems (ICIAS), pp. 1–5. IEEE (2014)

Rendle, S., Freudenthaler, C., Gantner, Z., Schmidt-Thieme, L.: BPR: Bayesian personalized ranking from implicit feedback. In: Proceedings of the 25th conference on uncertainty in artificial intelligence, pp. 452–461. AUAI Press (2009)

Rousseeuw, P.J.: Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. **20**, 53–65 (1987)

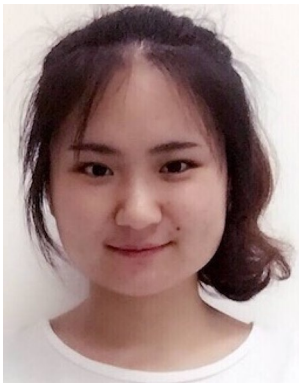
Su, X., Khoshgoftaar, T.M.: A survey of collaborative filtering techniques. In: Advances in Artificial Intelligence (2009)

Wang, S., Gong, M., Qin, C., Yang, J.: A multi-objective framework for location recommendation based on user preference. In: 2017 13th International Conference on Computational Intelligence and Security (CIS), pp. 39–43. IEEE (2017)

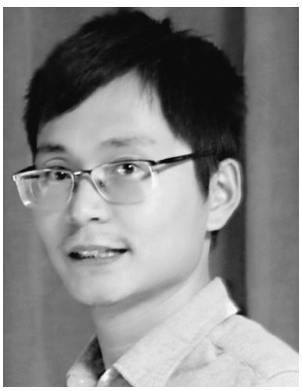
Wei, J., He, J., Chen, K., Zhou, Y., Tang, Z.: Collaborative filtering and deep learning based recommendation system for cold start items. Expert Syst. Appl. **69**, 29–39 (2017)

Zhang, Y., Yang, C., Niu, Z.: A research of job recommendation system based on collaborative filtering. In: 2014 Seventh International Symposium on Computational Intelligence and Design, vol. 1, pp. 533–538. IEEE (2014)

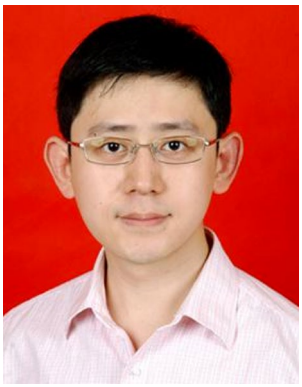
Qing Zhou is a full professor at the College of Computer Science, Chongqing University. He got his PhD in Computer Science from Chongqing University in 2008. His current research interests include machine learning and educational data mining.



Fenglu Liao is a software engineer in China Merchants Bank, Chengdu. She got her Master Degree in Computer Science from Chongqing University in 2019. Her current research interests include machine learning and data mining.



Chao Chen is a full professor at College of Computer Science, Chongqing University, Chongqing, China. He obtained his Ph.D. degree from Pierre and Marie Curie University and Institut Mines-Télécom/Télécom SudParis, France in 2014. He received the B.Sc. and M.Sc. degrees in control science and control engineering from Northwestern Polytechnical University, Xi'an, China, in 2007 and 2010, respectively. Dr. Chen got published over 80 papers including 20 ACM/IEEE Transactions. His work on taxi trajectory data mining was featured by IEEE Spectrum twice, in 2011 and 2016 respectively. He was the winner of the Best Paper Runner-Up Award at MobiQuitous 2011. In 2009, he worked as a Research Assistant with Hong Kong Polytechnic University, Kowloon, Hong Kong. His research interests include pervasive computing, mobile computing, urban logistics, data mining from large-scale GPS trajectory data, and big data analytics for smart cities.



Liang Ge is an associate professor at the College of Computer Science, Chongqing University. He got his PhD in Computer Science from Chongqing University in 2009. His current research interests include machine learning and educational data mining.