Acta Technica 62 (2017), No. 7A, 1073–1086

Personalized recommendation method for innovation and entrepreneurship resource base of undergraduates based on low-rank and sparse matrix factorization

Hai Li[1], Yingping Li[1]

Abstract. To improve effectiveness of personalized recommendation for innovation and entrepreneurship resource base of undergraduates, a personalized recommendation method for innovation and entrepreneurship resource base of undergraduates based on low-rank and sparse matrix factorization was proposed. Firstly, personalized recommendation model form for innovation and entrepreneurship resource base of undergraduates was given and vector space model was used to calculate the similarity between user configuration files and resource description. Secondly, low-rank and sparse matrix factorization method was introduced to solve personalized recommendation method for innovation and entrepreneurship resource base of undergraduates; finally, simulation experiment verifies the effectiveness of the proposed method.

Key words. Sparse matrix, Innovation and entrepreneurship, Resource base, Personalized recommendation.

1. Introduction

Innovation and entrepreneurship of undergraduates have developed considerably in the recent decade and competition between innovation and entrepreneurship of undergraduates also gradually needs undergraduates in need of innovation and entrepreneurship to master more accurate user demands and preferences, thus providing services or others for customers pertinently. Hence, a recommendation filtration algorithm with high accuracy and high performance becomes very important[1]. Collaborative filtration recommendation can make full use of relationship between

[1]School of Economics and Management, JiuJiang University, JiuJiang City, JiangXi, China

information, and be of high execution efficiency to get good recommendation result, becoming current research hotspot[2].

As for collaborative filtration recommendation, domestic and foreign scholars and experts have conducted a lot of in-depth researches. Up to now, there are many collaborative filtration recommendation algorithms[3-6] and each collaborative filtration recommendation algorithm has different operation thoughts. In actual application, these collaborative filtration recommendation algorithms have their own advantages and their defects are also very obvious: such as sparse data, cold starting, poor extendibility [7-9]. To solve these defects, some scholars have proposed adopting association rule data mining, Bayesian network, neural network, support vector machine and other technologies [10-12] to improve recommendation accuracy of recommendation system and obtain good recommendation effect. However, user interests are subject to comprehensive function and influence of several factors and the similarity value of current collaborative filtration recommendation algorithm is not scientific in accounting, lacks rationality, and ignores users' interest information. The recommendation precision remains to be further improved. [13]. Recently, researchers have shifted focus to research in data probability distribution and it supports data sampling and drawing in submanifold environment space. To detect potential manifold structure, manifold learning algorithm has been proposed, such as local linear imbedding algorithm[7], Laplacian characteristic mapping algorithm [8] etc. All these algorithms are built upon local invariance of data. According to the latest research progress of matrix factorization and manifold learning, a new non-negative matrix factorization algorithm for image normalization was proposed in the Thesis and it is also built upon local invariance. Through building a nearest neighbor, the geometrical information of data space was encoded. The objective is to find a module-based representation space. If two nodes in the graph are connective, the two corresponding data should be similar.

To realize the objective, we designed a new objective function for matrix factorization and included it into its graphical structure. At the same time, a simple optimization scheme was developed with gradient descent algorithm to realize iteration update and optimization of personalized recommendation objective function for innovation and entrepreneurship resource base of undergraduates.

2. Recommendation for innovation and entrepreneurship resource of undergraduates

In this section, active users are provided with detailed methods to recommend innovation and entrepreneurship resources of undergraduates. Firstly, a basic method suitable for vector space model was introduced in Section 2.1 and then detailing for text vector of semantic similarity based on innovation and entrepreneurship resources was shown in 2.2.

2.1. Content-based vector space filtering

Vector space model (VSM) is a kind of calculation method for similarity in resource space commonly used in information retrieval field. Each innovation and entrepreneurship resource is deemed as a dimension. Each file and character can constitute $n-$dimension space vector. Vector element of each innovation and entrepreneurship resource in the resource base can be assigned with corresponding weight, thus reflecting their importance in the resource base. These weights can be calculated through term frequency (TF) and inverse document frequency (IDF). specifically, a resource base d_i can be expressed as a vector [13].

$$\overrightarrow{d_i} = \{w_{i1}, w_{i2}, \cdots, w_{in}\} . \tag{1}$$

Where: n is total resource quantity in the resource base, w_{ik} $\left(k = \overline{1, \cdots, n}\right)$ is weight of element k in the vector, then w_{ik} can be calculated as:

$$w_{i,k} = TF(i,k) \times IDF(k) = \frac{|w_k|}{|d_i|} \times 1og\frac{n}{|D_k|} . \tag{2}$$

Where: $|w_k|$ is occurrence times of resource w_k in the resource base d_i, $|d_i|$ is total resource quantity in the resource base, $|D_k|$ is the quantity of including w_k document.

Then, document similarity can be calculated with cosine angle of vector in the resource base. For example, the similarity between documents d_i and d_j can be calculated as:

$$sim(d_i, d_j) = \cos ine(\overrightarrow{d_i}, \overrightarrow{d_j}) = \frac{\overrightarrow{d_i} \cdot \overrightarrow{d_j}}{|\overrightarrow{d_i}| \times |\overrightarrow{d_j}|} . \tag{3}$$

In the proposed method, vector space model is used to calculate the similarity between user configuration file and resource description. User configuration file is defined as a group of keywords for resources browsed by him recently and resource description is all texts used to describe resources.

Specifically, a resource base of innovation and entrepreneurship resources for m undergraduates is considered, with active users u_a viewing resources r_a. Suppose m resource contains n different resources, , let d_i as text description of resource r_i $\left(i = \overline{1, \cdots, m}\right)$, then it can be expressed as $\overrightarrow{d_i} = \{w_{i1}, w_{i2}, \cdots, w_{in}\}$, with w_{ik} $\left(k = \overline{1, \cdots, n}\right)$ as TF-IDF weight of corresponding resource element k in the description of resource text. It can be calculated as per Equation (2).

Let $\{k_1,\ k_2,\ .\ .\ .\ ,\ k_t\}$ as t key resources of resource r_a recently browsed by user h. Considering that these historic keywords are taken as an query q_a. Such query can be expressed as vector in the same resource description space.

$$\overrightarrow{q_a} = \{w_{a1}, w_{a2}, \cdots, w_{an}\} . \tag{4}$$

Where $\overrightarrow{q_a}$ is weight of corresponding resource semantics in query.

Then, according to Equation (3), the calculation form of similarity between query

q_a and resource base d_i is:

$$sim(q_a, d_i) = \frac{\overrightarrow{d_i} \cdot \overrightarrow{q_a}}{|\overrightarrow{d_i}| \times |\overrightarrow{q_a}|} = \frac{\sum\limits_{k=1}^{n} w_{ak} w_{ik}}{\sqrt{\sum\limits_{k=1}^{n} w_{ak}^2} \times \sqrt{\sum\limits_{k=1}^{n} w_{ik}^2}}. \tag{5}$$

Equation (5) is applied to all $d_i \in \{1, 2, \cdots, N\}$. Then, resources are sorted in descending order as per calculation value of similarity with q_a. Finally, top-k resource selection is conducted for active users in historic key resources of related resources.

As resources can appear in different forms (such as singular or plural forms), tense (such as present, past or future tenses), before similarity calculation, it is required to treat resources correspondingly. In addition, in terms of key resources which may be singular or plural resources, resource description can be pre-treated through plural resource recognition matching historic key resources. Through this identification, plural resources can be deemed as singular resources.

On the basis of above pretreatment process, Equation (2) (TF-IDF) and Equation (5) (VSM) can be used to calculate the similarity between resource base d_i and query q_a.

2.2. Query of resource similarity matching

Polysemy resource and synonymy resource are a common problem faced by text processing. if only syntax matching of resource is processed without consideration of semantic similarity, it is easy to miss potential matching with different resources and the same meaning. In this section, similarity integration method of resource semantics is proposed to more accurately identify innovation and entrepreneurship resources of undergraduates. Different from previous literature, the Thesis mainly focuses on resource matching instead of semantic similarity. In the experiment in the Thesis, the form of literature [14] is adopted for similarity matching and this method is a high-precision calculation method for similarity of resource semantics based on Wikipedia data set.

Considering that the key resource list of resource h browsed by active user u_a recently is: $q_a = \{k_1,\ k_2,\ \ldots,\ k_t\}$ and the list is taken as query. For each resource r_i considered to match query q_a, it is suggested that resource description should be used to substitute each resource. For example, d_i can be used to substitute similar semantics in query, although it does not appear in key resources queried. Resource semantics weight of resource description is updated according to calculation value for similarity between queried semantics and selected resource semantics. Finally, resource and weight resource under query are represented with TF-IDF and VSM method is used to calculate their similarity.

Specifically, considering that the occurrence times of resource v_x in resource description d_i is o_x. Suppose v_x is the resource which is the most similar to key resource k_y in query q_a in semantics and the similarity value is $s\,(v_x,\ k_y) \in (0,\ 1)$. k_y is used to substitute resource v_x in resource base d_i and its weight value is updated to $w_{xy} = o_x s\left(v_x,\ k_y\right)$. It means that if resource v_x in resource base d_i occurs for o_x times, it can be deemed that resource k_y in resource base d_i occurs for $o_x s\left(v_x,\ k_y\right)$ times. Above substitution process is executed repeatedly for all resources in resource base d_i.

In consideration of calculation example of similarity between resource base d_i and query q_a as stated in Section 2.1, resource in query q_a of resource base d_i is used to substitute the most similar resource semantics, specifically as shown in Table 1. For example, resource "domain" in the resource d_i is substituted with resource "resource" in query q_a and its weight is updated to be $1 \times 0.015 = 0.015$.

Suppose n_1 resources in resource base d_i are substituted with k_1 and their weights are updated to be$\{w_{11},\ w_{21}, \cdots,\ w_{n_1 1}\}$. $n_2$resources in resource base d_i are substituted with $k_2$and their weights are updated to be $\{w_{12},\ w_{22}, \cdots,\ w_{n_1 2}\}$. n_0 resources in resource base d_i are not substituted with any resource in query. Resource description d_i is changed to:

$$d_i' = \{k_1,\ k_2, \cdots,\ k_t,\ k_{t+1}, \cdots,\ k_{t+n_0}\}. \tag{6}$$

Corresponding weight is:

$$w = \left\{ \sum_{j=1}^{n_1} \omega_{j1}, \sum_{j=2}^{n_2} \omega_{j2}, \cdots, \sum_{j=1}^{n_2} \omega_{j2}, 0, \cdots, 0 \right\}. \tag{7}$$

The similarity between l_a and d_i can be obtained through calculating the similarity between q_a and d_i'. For corresponding weights of $k_{t+1}, \cdots,\ k_{t+n_0}$ in Equation (12) are 0, these elements can be eliminated from d_i'. Spatial vector dimension of d_i' is changed to be consistent with q_a. Then TF-IDF of resource in resource base d_i and query q_a is calculated according to Equation (2) and the similarity is calculated with VSM in Equation (5). Finally, similar resources of top-k are selected according to the similarity between resource base d_i and query q_a.

3. Recommendation algorithm for semi-supervised innovation and entrepreneurship resource base of undergraduates with gradient descent

In this section, on the basis of discussion about non-negative matrix factorization for recommendation of innovation and entrepreneurship resource base of undergraduates in last section, firstly, the recommendation framework for regularized and semi-supervised innovation and entrepreneurship resource base of undergraduates based on potential space is proposed and how to apply it to NMF is discussed (see Section 3.1). Then, some specific solutions are proposed (see Section 3.2). Finally, calculation complexity calculation of the algorithm is provided (see Section 3.3).

3.1. Non-negative matrix factorization for recommendation of innovation and entrepreneurship resource base of undergraduates

During network generation, a_{ij} is an observational variable and it indicates the probability of interaction between resource i and resource j. Suppose the probability of connection between existing resource i and resource j is determined by whether the generated edge and edge belong to the same community. Two talent variables are defined to be $W = [w_{ik}] \in R_+^{N\times K}$ and $H = [v_{jk}] \in R_+^{N\times K}$ and their elements w_{ik} and v_{jk} respectively indicate the probability that inner edge and outer edge generated by node i belong to community k. These latent variables also mean the probability that node i belongs to inner or outer community k. Each row of W or H can be deemed as resource membership distribution, as shown in Fig. 1. Hence, the probability of mutual connection between resource i and resource j can be expressed as:

$$\hat{a}_{ij} = \sum_{k=1}^{K} w_{ik} v_{jk} \,. \tag{8}$$

Hence, the recommendation of innovation and entrepreneurship resource base of undergraduates can be transformed to NMF problem, namely $\hat{A} = WH^T$. In each row of W or H, the element with the largest grade is community. If network is directed, adjacent matrix is asymmetrical. If network is undirected, adjacent matrix is symmetrical. There is a gap of constant multiplication factor for factorization of W or H. In this Thesis, focus is put on researching un-weighted network and H is used to determine node membership. From the perspective of cluster, the factorization process can be projected to K-dimension potential space as N-dimension in adjacent matrix.

As for quality quantization of factorization result, there are two commonly used objective functions. The first is square loss function, which is equivalent to Frobenius square norm of difference between two matrixes:

$$\mathcal{L}_{LSE}(A, WH^T) = ||A - WH^T||_F^2 \,. \tag{9}$$

Kullback-leibleer(KL)The second objective function is defined according to Kullback-leibleer (KL) divergence of two matrixes:

$$\mathcal{L}_{KL}(A, WH^T) = KL(A||WH^T) \,. \tag{10}$$

Different types of objective functions shall be selected for different application programs. A modified form of NMF algorithm is SNMF which introduces symmetrical constraint framework to NMF algorithm. If network is connected and undirected, adjacent matrix is symmetrical. Hence, the factorization process should be symmetrical. Then the modified form of square loss function is;

$$\mathcal{L}_{SYM}(A, HH^T) = ||A - HH^T||_F^2 \,. \tag{11}$$

As $\mathcal{L}_{LSE}$, $\mathcal{L}_{KL}$ and $\mathcal{L}_{SYM}$ are non-convex for W and H, it is difficult to adopt heuristic optimization algorithm to find global minimum value of these loss functions with W and H are all existing. However, if $\mathcal{L}_{LSE}$, $\mathcal{L}_{KL}$ and $\mathcal{L}_{SYM}$ are convex for W or H, $\mathcal{L}_{LSE}$ can be optimized through iterative optimization algorithm:

$$\begin{cases} w_{ik} \leftarrow w_{ik} \dfrac{(AH)_{ik}}{(WH^T H)_{ik}} \\ v_{jk} \leftarrow v_{jk} \dfrac{(A^T H)_{jk}}{(HW^T W)_{jk}} \end{cases} \tag{12}$$

Similarly, $\mathcal{L}_{KL}$ and $\mathcal{L}_{SYM}$ can be respectively solved through the following two update schemes:

$$\begin{cases} w_{ik} \leftarrow w_{ik} \dfrac{\sum_j (a_{ij} v_{jk} / \sum_k w_{ik} v_{jk})}{\sum_j v_{jk}} \\ v_{jk} \leftarrow v_{jk} \dfrac{\sum_j (a_{ij} w_{ik} / \sum_k w_{ik} v_{jk})}{\sum_i w_{ik}} \end{cases} \tag{13}$$

$$\begin{cases} w_{ik} \leftarrow w_{ik} \dfrac{\sum_j (a_{ij} v_{jk} / \sum_k w_{ik} v_{jk})}{\sum_j v_{jk}} \\ v_{ik} \leftarrow v_{ik} (\dfrac{1}{2} + \dfrac{(AH)_{ik}}{(2HH^T H)_{ik}}) \end{cases} \tag{14}$$

Input of recommendation algorithm for innovation and entrepreneurship resource base of undergraduates is network topology information and it can be expressed as adjacent matrix. Most recommendation algorithms for innovation and entrepreneurship resource base of undergraduates firstly obtain a new matrix of adjacent matrix through minimization of objective function. New matrix can be expressed in latent variable space. Then, the algorithm clusters row in the new matrix and k mean value cluster or other cluster algorithms are used to classify nodes. Hence, it is similar to the cluster process in latent variable space.

On the basis of definition of above index, regularization item $\mathcal{R}_\beta(O, H)$, $\beta \in \{LSE, KL\}$, the objective function of its topology information can be defined as:

$$\mathcal{F}_{\alpha,\beta}(H|A, O) = \mathcal{L}_\alpha(A, H) + \lambda \mathcal{R}_\beta(O, H) \ . \tag{15}$$

Where $\alpha \in \{LSE, KL, SYM, MOD, ADJ, LAP, NLAP\}$, λ is balance parameter of topology information and prior information. Under most cases, the value of symbol λ is positive. As for the first term in $\mathcal{L}_{ADJ}$ or $\mathcal{L}_{MOD}$, the value of symbol $\lambda$is negative, for these terms need to be maximized and the second term needs to be minimized. For convenience of calculation, the same distance function is selected for two parts in Equation (14). When $\alpha \in \{LSE, SYM, MOD, ADJ, LAP, NLAP\}$, $\beta = LSE$; when $\alpha = KL$, $\beta = KL$. Hence, the following objective function can be obtained:

$$\mathcal{F}_{LSE}(H|A, O) = ||A - WH^T||_F^2 + \lambda Tr(H^T LH) \,. \tag{16}$$

$$\mathcal{F}_{SYM}(H|A, O) = ||A - HH^T||_F^2 + \lambda Tr(H^T LH) \,. \tag{17}$$

$$\begin{aligned}\mathcal{F}_{KL}\left(H\,|A,O\right) = &\sum_{i=1}^{N}\sum_{j=1}^{N}(a_{ij}\log\left(\frac{a_{ij}}{\sum_{k=1}^{K}w_{ik}v_{jk}}\right) + \sum_{k=1}^{K}w_{ik}v_{jk} \\ &- a_{ij} + \frac{\lambda}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{K}\left(v_{ik}\log\left(\frac{v_{ik}}{v_{jk}}\right)\right) + v_{jk}\log\left(\frac{v_{jk}}{v_{ik}}\right)\end{aligned} \tag{18}$$

$$\mathcal{F}_{MOD}\left(H\,|A,O\right) = -Tr(H^TBH) + \lambda Tr(H^TLH)\,. \tag{19}$$

$$\mathcal{F}_{LAP}\left(H\,|A,O\right) = Tr(H^T(D-A)H) + \lambda Tr(H^TLH)\,. \tag{20}$$

Focus will be put on introducing how to solve these optimized objective functions in next section.

3.2. Optimization rule based on matrix trace and Frobenius norm

As objective functions $\mathcal{F}_{LSE}$, $\mathcal{F}_{SYM}$ and $\mathcal{F}_{KL}$ in the proposed framework are non-convex relative to H and W, it is difficult to obtain global minimum value point with simple iteration algorithm. In this section, three iteration update rules are developed for three forms of objective functions in this section to obtain local minimum value of objection function.

With $\mathcal{F}_{LSE}$ objective function in Equation (15) as an example, based on matrix trace and Frobenius norm characteristics [13∼14], such as $Tr(A) = Tr(A^T)$, $Tr(AB) = Tr(BA)$ and $||A||_F^2 = Tr\left(AA^T\right)$, then the modified form of $\mathcal{F}_{LSE}$ can be:

$$\begin{aligned}\mathcal{F}_{LSE}\left(H\,|A,O\right) =&Tr((A-WH^T)(A-WH^T)^T) \\ &+ \lambda Tr(H^TLH) = Tr(AA^T) + Tr(WH^THW^T) \\ &- 2Tr(AHW^T) + \lambda Tr(H^TLH)\,.\end{aligned} \tag{21}$$

Through respectively introducing Lagrangian multipliers $\psi = [\psi_{ij}] \in R^{N\times K}$ and $\Phi = [\varphi_{ij}] \in R^{N\times K}$ for $W = [w_{ij}] \geq 0$ and $H = [h_{ij}] \geq 0$, $\mathcal{L}_{LSE}$ form of Lagrangian operator is

$$\begin{aligned}\mathcal{L}_{LSE} =&Tr(AA^T) + Tr(WH^THW^T) - 2Tr(AHW^T) \\ &+ \lambda Tr(H^TLH) + Tr(\psi W^T) + Tr(\Phi H^T)\,.\end{aligned} \tag{22}$$

To calculate minimum value of $\mathcal{L}_{LSE}$ under H and W parameters, the following can be obtained through solving partial derivative form:

$$\frac{\partial\mathcal{L}_{LSE}}{\partial W} = -2AH + 2WH^TH + \psi\,. \tag{23}$$

$$\frac{\partial \mathcal{L}_{LSE}}{\partial H} = -2A^T W + 2HW^T W + 2\lambda LH + \Phi \,. \tag{24}$$

In combination with partial derivative equaling to $\psi_{ik} w_{ik} = 0$ and Karush–Kuhn–Tucker condition equation $\varphi_{jk} h_{jk} = 0$, the solution to w_{ik} and h_{jk} can be:

$$-(AH)_{ik} w_{ik} + (WH^T H)_{ik} w_{ik} = 0 \,. \tag{25}$$

$$-(A^T W)_{jk} v_{jk} + (HW^T W)_{jk} v_{jk} + \lambda (LH)_{jk} v_{jk} = 0 \,. \tag{26}$$

Finally, the following objective update rules are obtained:

$$\begin{cases} w_{ik} \leftarrow w_{ik} \dfrac{(AH)_{ik}}{(WH^T H)_{ik}} \\ v_{jk} \leftarrow v_{jk} \dfrac{\left(A^T W + \lambda OH\right)_{jk}}{\left(HW^T W + \lambda DH\right)_{jk}} \end{cases} \tag{27}$$

When $\lambda = 0$, update rule in Equation (26) is degraded to that as shown in Equation (5). Such rule is based on update rule of standard NMF of Euclidean distance.

Similarly, the objective update rule of minimized $\mathcal{F}_{SYM}$ can be:

$$v_{ik} \leftarrow v_{ik} \frac{\left(AH + \lambda'' OH\right)_{ik}}{\left(HH^T H + \lambda' DH\right)_{ik}} \,. \tag{28}$$

To ensure equation consistency, $\lambda'' = 2\lambda'$.

Then, $\mathcal{F}_{KL}$ is minimized with the same method and the form of update rule is:

$$\begin{cases} w_{ik} \leftarrow w_{ik} \dfrac{\sum_j \left(a_{ij} v_{jk} / \sum_k w_{ik} v_{jk}\right)}{\sum_j v_{jk}} \\ v_k \leftarrow (\sum\limits_i w_{ik} I + \lambda L)^{-1} \hat{v}_k \end{cases} \tag{29}$$

Where, h_k is line k in H and I is $N \times N$ unit matrix and:

$$\hat{h}_k = \begin{bmatrix} h_{1k} \sum_i \left(a_{i1} w_{ik} / \sum_k w_{ik} h_{1k}\right) \\ h_{2k} \sum_i \left(a_{i2} w_{ik} / \sum_k w_{ik} h_{2k}\right) \\ \vdots \\ h_{Nk} \sum_i \left(a_{iN} w_{ik} / \sum_k w_{ik} h_{Nk}\right) \end{bmatrix} . \tag{30}$$

3.3. Solution with gradient descent method

Gradient descent method is also called the steepest descent method and its theoretical basis is the concept of gradient. The relation between gradient and directional derivative is that: the gradient direction is consistent with the direction of maximum directional derivative and the model of the gradient is maximum value for directional derivative of function at the point[15]. The advantage of solution to optimization with gradient descent method lies that the realization of the algorithm is simple. With gradient update rule, the convergence rate is fast. It is of high practical value for recommendation for innovation and entrepreneurship resource base of undergraduates, especially recommendation of complex resources. On the basis of update iteration rule of above objective function, gradient descent method is used for objective optimization and calculation pseudocode is shown in algorithm 1.

Algorithm 1: recommendation for innovation and entrepreneurship resource base of undergraduates based on gradient descent method

1. Initiative: randomly select any number in the value scope
2. Operate repeatedly:

Calculate objective value:

$$\mathcal{F}_{LSE}\left(H\left|A,O\right.\right) = ||A - WH^T||_F^2 + \lambda Tr(H^T LH)$$

$$\mathcal{F}_{SYM}\left(H\left|A,O\right.\right) = ||A - HH^T||_F^2 + \lambda Tr(H^T LH)$$

$$\mathcal{F}_{KL}\left(H\left|A,O\right.\right) = \sum_{i=1}^{N}\sum_{j=1}^{N}(a_{ij}\log\left(\frac{a_{ij}}{\sum_{k=1}^{K} w_{ik}v_{jk}}\right) + \sum_{k=1}^{K} w_{ik}v_{jk}$$

$$- a_{ij} + \frac{\lambda}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{k=1}^{K}\left(v_{ik}\log\left(\frac{v_{ik}}{v_{jk}}\right)\right) + h_{jk}\log\left(\frac{v_{jk}}{v_{ik}}\right)$$

Judge whether it reaches the ending point: if the absolute value of difference between the previous and the present function values is less than the threshold value, jump out of the circulation; otherwise, it will continue;

Update community division:

$$\begin{cases} w_{ik} \leftarrow w_{ik}\frac{(AH)_{ik}}{(WH^TH)_{ik}} \\ v_{jk} \leftarrow v_{jk}\frac{\left(A^TW+\lambda OH\right)_{jk}}{\left(HW^TW+\lambda DH\right)_{jk}} \end{cases}$$

$$v_{ik} \leftarrow v_{ik}\frac{(AH + \lambda'' OH)_{ik}}{(HH^TH + \lambda' DH)_{ik}}$$

$$\begin{cases} w_{ik} \leftarrow w_{ik}\frac{\sum_j \left(a_{ij}v_{jk}/\sum_k w_{ik}v_{jk}\right)}{\sum_j v_{jk}} \\ v_k \leftarrow (\sum_i w_{ik}I + \lambda L)^{-1}\hat{v}_k \end{cases}$$

3. Output final result

4. Experiment analysis

Visual C++ programming is adopted for simulation test in the computers of Intel(R) Core i5-3337U 3.0GHz CPU, 4GB RAM, Windows XP operating systems. Data come from public data set *MovieLens* and its description is shown in Literature [11].

4.1. Comparison of recommendation precision

When the nearest neighbor is 35, recommendation algorithm is adopted to solve the problem. Specific result is shown in Fig. 1. It can be seen clearly from Fig. 1 that the MAE value of collaborative filtering recommendation algorithm in the Thesis is lower than the comparison algorithm, effectively improving recommendation precision and obtaining ideal recommendation result.

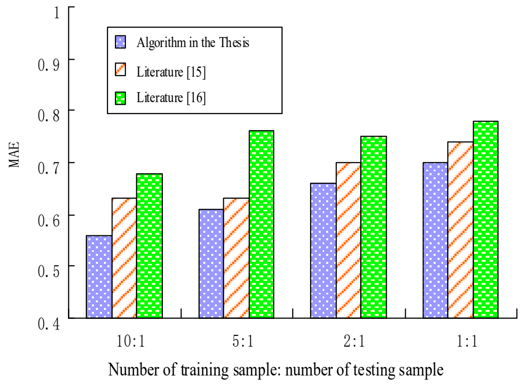

Fig. 1. Comparison of recommendation precision in different algorithms

4.2. Result analysis under cold starting conditions

To simulate cold starting conditions, 10 users are selected and their evaluation information is deleted, with result shown in Fig. 2. Through detailed analysis for Fig. 2, the collaborative filtering recommendation algorithm integrating user's scoring and attribute similarity in the Thesis can solve the current existing problem that the recommendation algorithm cannot be implemented under cold starting conditions, improving recommendation precision and obtaining more superior recommendation result.

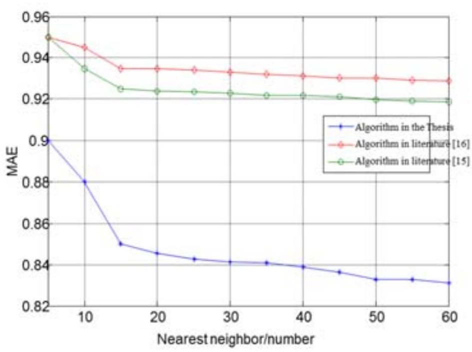

Fig. 2. Comparison of algorithm performance under cold starting conditions

4.3. Performance comparison under different spareness

The recommendation error of different data spareness is shown in Fig. 3. There is approximately linear change relation between data spareness and MAE. However, under equal conditions, compared with comparison results in Literatures [15] [16], it is found that MAE value of recommendation result by algorithm in the Thesis is smaller. Hence, the recommendation precision of algorithm in the Thesis is superior to algorithm in Literatures [15] [16] under equal conditions.

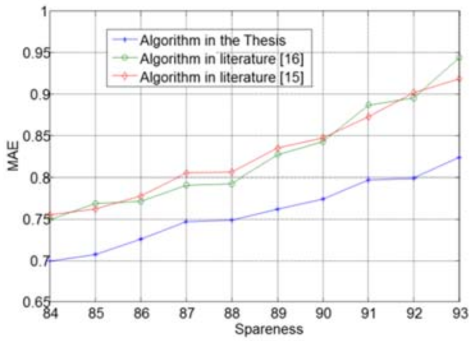

Fig. 3. Performance comparison of three algorithms under three different spareness

4.4. Generality test

To verify the generality of collaborative filtering recommendation algorithm integrating user's scoring and attribute similarity in the Thesis, we select Book-Crossing data set for simulation test, with information of 287558 users and 1491807 scoring data for 231797 E-books. We adopt scoring system for evaluation modeling of data in range [0, 10], with 1 for the highest score and 0 for the lowest score. The experiment results of different algorithms are shown in Fig. 4. It can be known from Fig. 4 that compared with other collaborative filtering recommendation algorithm, the MAE of collaborative filtering recommendation algorithm in the Thesis is also the minimum and the recommendation precision is higher. It further proves superiority of algorithm in the Thesis and good generality.

5. Conclusion

A gradient descent algorithm for recommendation of semi-supervised innovation and entrepreneurship resource base of undergraduates based on low-rank and sparse matrix factorization is proposed in the Thesis and recommendation framework of semi-supervised innovation and entrepreneurship resource base of undergraduates based on low-rank and sparse matrix factorization is built on the basis of expression form of latent space. In addition, optimization rule of gradient descent algorithm for recommendation of semi-supervised innovation and entrepreneurship resource base

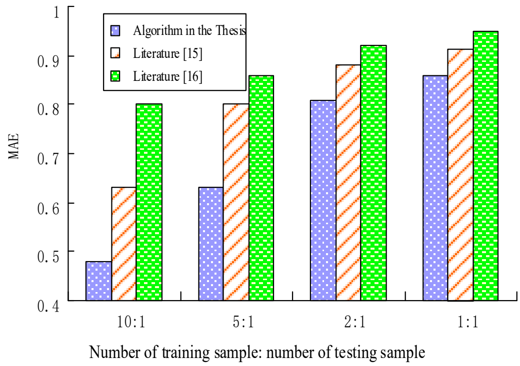

Fig. 4. Performance comparison with classic algorithm

of undergraduates is built on the basis of matrix trace and Frobenius norm. Experiment results prove effectiveness of the proposed method. As the calculation efficiency of recommendation of large online innovation and entrepreneurship resource base of undergraduates is considered too much in the design of the algorithm, with gradient descent algorithm adopted, we seek to obtain local extreme point for recommendation of innovation and entrepreneurship resource base of undergraduates. How to better improve approximation accuracy of local extreme point is next research emphasis.

References

[1] Y. Y. Zhang, A. Algburi, N. Wang, V. Kholodovych, D. O. Oh, M. Chikindas, and K. E. Uhrich: *Self-assembled Cationic Amphiphiles as Antimicrobial Peptides Mimics: Role of Hydrophobicity, Linkage Type, and Assembly State*, Nanomedicine: Nanotechnology, Biology and Medicine, *13* (2017), No. 2, 343–352.

[2] N. Arunkumar, K. R. Kumar, V. Venkataraman: *Automatic detection of epileptic seizures using new entropy measures*, Journal of Medical Imaging and Health Informatics, *6* (2016), No. 3, 724–730.

[3] R. Hamza, K. Muhammad, N. Arunkumar , G. R. González: *Hash based Encryption for Keyframes of Diagnostic Hysteroscopy*, IEEE Access (2017), https://doi.org/10.1109/ACCESS.2017.2762405

[4] D. S. Abdelhamid, Y. Y. Zhang, D. R. Lewis, P. V. Moghe, W. J. Welsh, and K. E. Uhrich: *Tartaric Acid-based Amphiphilic Macromolecules with Ether Linkages Exhibit Enhanced Repression of Oxidized Low Density Lipoprotein Uptake*, Biomaterials, *53* (2015), 32–39.

[5] W. S. Pan, S. Z. Chen, Z. Y. Feng: *Automatic Clustering of Social Tag using Community Detection.* Applied Mathematics & Information Sciences, *7* (2013), No. 2, 675–681.

[6] Y. Y. Zhang, E. Mintzer, and K. E. Uhrich: *Synthesis and Characterization of PEGylated Bolaamphiphiles with Enhanced Retention in Liposomes*, Journal of Colloid and Interface Science, *482* (2016), 19–26.

[7] N. Arunkumar, K. M. Mohamed Sirajudeen: *Approximate Entropy based ayurvedic pulse diagnosis for diabetics - A case study*, TISC 2011 - Proceedings of

the 3rd International Conference on Trendz in Information Sciences and Computing, (2011), art. No. 6169099, 133–135.

[8] N. Arunkumar, K. Ramkumar, S. Hema, A. Nithya, P. Prakash, V. Kirthika: *Fuzzy Lyapunov exponent based onset detection of the epileptic seizures*,2013 IEEE Conference on Information and Communication Technologies, ICT 2013, (2013), art. No. 6558185, 701–706.

[9] J. J. Faig, A. Moretti, L. B. Joseph, Y. Y. Zhang, M. J. Nova, K. Smith, and K. E. Uhrich: *Biodegradable Kojic Acid-Based Polymers: Controlled Delivery of Bioactives for Melanogenesis Inhibition*, Biomacromolecules, *18* (2017), No. 2, 363–373.

[10] N. Arunkumar, V. Venkataraman, T. Lavanya: *A moving window approximate entropy based neural network for detecting the onset of epileptic seizures*, International Journal of Applied Engineering Research, *8* (2013), No. 15, 1841–1847.

Received May 7, 2017