7-17-2019

Novel Recommendation System for Tourist Spots Based on Hierarchical Sampling Statistics and SVD++

Guangli Li

Jin Hua

Tian Yuan

Jinpeng Wu

Hongbin Zhang

See next page for additional authors

Authors

Guangli Li, Jin Hua, Tian Yuan, Jinpeng Wu, Hongbin Zhang, and Tao Li

Research Article

Novel Recommendation System for Tourist Spots Based on Hierarchical Sampling Statistics and SVD++

Guangli Li,[1] **Jin Hua** ⓘ **,**[1] **Tian Yuan,**[1] **Jinpeng Wu,**[2] **Ziliang Jiang,**[2]
Hongbin Zhang ⓘ **,**[2] **and Tao Li**[3]†

[1]*School of Information Engineering, East China Jiaotong University, China*
[2]*Software School, East China Jiaotong University, China*
[3]*School of Computing and Information Science, Florida International University, USA*
†*Deceased*

Correspondence should be addressed to Hongbin Zhang; zhanghongbin@whu.edu.cn

Recommendation system for tourist spots has very high potential value including social and economic benefits. The traditional clustering algorithms were usually used to build a recommendation system. However, clustering algorithms have the risk on falling into local minimums, which may decrease the final recommendation performance heavily. Few works focused their research on tourist spots recommendation and few recommendation systems consider the population attributes information for fitting the user implicit preference. To address the problem, we focused our research work on designing a novel recommendation system for tourist spots. First a new dataset named "Smart Travel" is created for the following experiments. Then hierarchical sampling statistics (HSS) model is used to acquire the user preference for different population attributes. A new recommendation list named L_A is generated in turn by fitting the excavated the user preference. More importantly, SVD++ algorithm rather than those traditional clustering algorithms is used to predict the user ratings. And a new recommendation list named L_B is generated in turn on the basis of rating predictions. Finally, the two lists L_A and L_B are fused together to boost the final recommendation performance. Experimental results demonstrate that the mean precision, mean recall, and mean F1 values of the proposed recommendation system improve about 7.5%, 6.2%, and 6.5%, respectively, compared to the best competitor. The novel recommendation system is especially better at recommending a group of tourist spots, which means it has higher practical value.

1. Introduction

The rapid development of the Web: people are now stepping into an "information overload" era. Not only the information consumers (online websites users) but the information producers (online websites managers) should face a big challenge: website managers strive to generate valid information of their products and recommending them to their users, which may bring them lots of economic benefits, whereas website users want to know whether the recommended information can satisfy their preference. Recently, a tourism wave has emerged, and tourists (travelers) are smart to retrieve and use valid tourism information to appropriate plan their travel. Obviously, lots of traveling tips are helpful. Hence, statistics show that approximately three-quarters of tourists

prefer retrieve travel reviews or ratings of their destinations. As previously mentioned, tourists are often given excessive information due to the dawn of the "information overload" era. Addressing the problem has become increasingly important. As it is known, recommendation system is an effective method which can resolve the "information overload" problem to some extent. However, few works focused their research on tourist spots recommendation and few recommendation systems consider the population attributes information for fitting the user implicit preference. To address the problem, we focus our research work on designing a novel recommendation system for tourist spots: a new dataset named "Smart Travel" is created initially for the following experiments. Then, a hierarchical sampling statistics (HSS) model is utilized to acquire the user implicit preference

for different population attributes. Thirdly a novel SVD++ algorithm is designed to complete the final recommendation. More importantly, the excavated the user preference is fused into the collaborative filtering-based recommendation framework to more effectively generate recommendation results.

The rest of this paper is organized as follows. Section 2 introduces the preliminaries and related works. Section 3 presents the architecture and some details of the proposed recommendation system. Section 4 provides relevant experimental results and discussions. Finally, Section 5 presents the concluding remarks and future works.

2. Related Works

In general, recommendation results can be generated on the basis of the user preference, the item features, and other environmental factors such as time, season, and location. In recent recommendation literatures [1, 2], recommendation systems are usually categorized into three categories, namely, content-based, collaborative filtering-based, and hybrid recommendation systems. A content-based recommendation system focuses on selecting those objects with similar characteristics according to the objects that users have browsed. A collaborative filtering-based recommendation system has been widely used in some real websites like Amazon and YouTube. It can generate lots of personalized recommendation results for users. As it is known, collaborative filtering theory was proposed by Goldberg [3] and was first used in "Tapestry," which is a famous recommendation system. However, "Tapestry" only offered recommendation services for fewer users. The rating-based collaborative filtering recommendation system was presented by Resnick [4], who designed another famous recommendation system called "GroupLens." "GroupLens" predicted user implicit preference on the basis of all user ratings. Traditional clustering algorithms are usually used to mine users who have same or similar preference. Finally, both news and films are recommended to those target users according to semantic similarity computing. On the basis of "GroupLens", Konstan [5] and Miller [6] designed a novel recommendation system on an open computing structure. Yu [7] proposed a multilinear interactive matrix factorization algorithm (MLIMF) to model the interactions between the users and each event associated with their final decisions; the model considered not only the user-item ratings but also the pairwise interactions based on some empirically supported factors. In our real world, items will suffer from the sparsity problem more severely than users since items are usually observed by fewer features to support a feature-based or content-based algorithm. To tackle this problem, Yu [8] proposed a new method to study the collaborative retrieval (CR) task from the users' perspective, which aimed to sufficiently explore the sophisticated relationship of each triple as the form is query × user × item.

Recently, researchers paid their attentions on the recommendation system for tourist spots. Li [9] proposed a hybrid recommendation system for tourist spots based on hierarchical sampling statistics and Bayesian personalized ranking (BPR), and the recommendation performance was improved mainly by the BPR algorithm. Fenza [10] proposed a context-aware recommendation system based on the collaborative filtering theory. First, both users and tourist spots are clustered using the traditional fuzzy clustering algorithm, and context information is analyzed by the situational awareness technology. Then, the similarities of the tourist the spots that users prefer to be were calculated and tourist spots were recommended to users in order. The present work provides an adaptive environment in the process of dynamical user clustering and alleviates the cold start problem to some extent. Hsu [11] adopted the traditional collaborative filtering method to construct a novel recommendation system for tourist spots. Bayesian network was used to calculate the user preference for spots, to improve the final predication accuracy. Nilash [12] designed a new recommendation system for tourist spots based on the multistandard collaborative filtering algorithm. Multiobjective collaborative filtering technology was adopted to mine much more valuable user preference. Finally, recommendation accuracy was improved by a Gaussian mixture-based clustering algorithm.

In summary, we can observe: (1) Different kinds of recommendation system [3–12] have entered into people's daily life and they have played very important roles in changing people's working, living and studying. To the best of our knowledge, few works focused their research on recommendation system for tourist spots. Moreover, the above-mentioned works only used the traditional collaborative filtering algorithm. And they did not consider the population attributes, which may help to represent the deep-level semantic information of the user preference. Recently, tourism wave has emerged, which indicates recommendation system for tourist spots has very large potential value. It will give us much more economic benefits as well as social benefits. (2) The traditional clustering algorithms have been used in most recommendation systems [3–5]. But they have several limitations: first they need a good initial value; second, they usually fall into local minimums; third outliers will interfere with the final clustering results. So the traditional clustering algorithms are not suitable for designing our novel recommendation system. (3) Additional information such as visual content and user personal information [13] could help to improve the final recommendation performance. However, most recommendation systems only used user ratings rather than other additional information to build their models. The paper is an extended version of the reference [9]. Our contributions in the paper are two folds:

(1) A novel dataset named "Smart Travel" is created by combining a well-designed questionnaire survey and automatic data crawling on the Web. On the basis of the new dataset, the hierarchical sampling statistics (HSS) model [14] is utilized to acquire the user preference for different population attributes, namely, travel season, travel interest, and travel method are used to describe users, which are a good supplement to the traditional collaborative filtering-based recommendation model.

(2) A new recommendation list called L_A is generated on the basis of the user preference. Then a novel collaborative
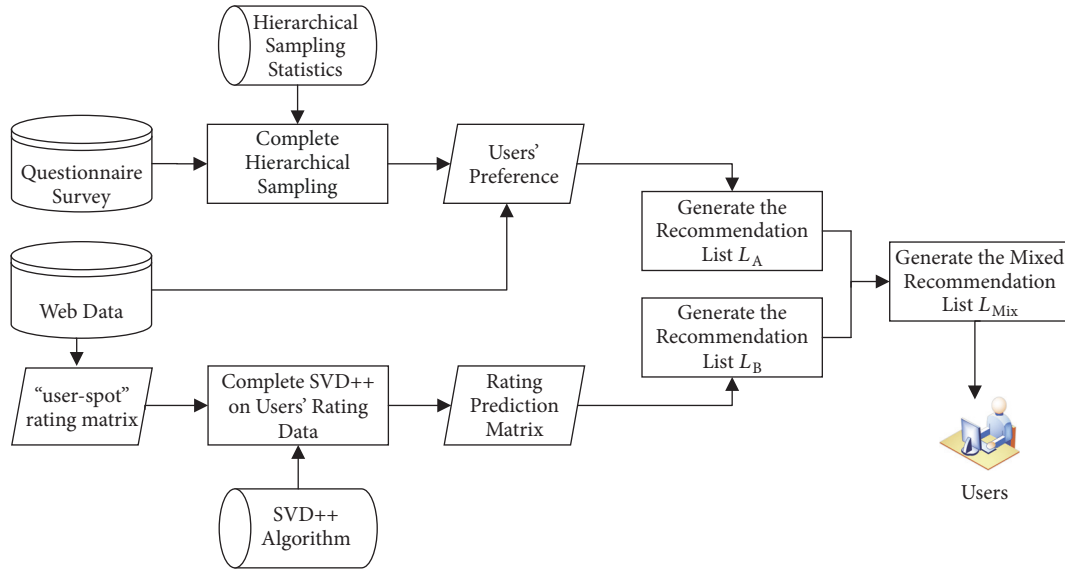
Figure 1: Proposed recommendation system framework.

filtering-based model called SVD++ algorithm [15, 16] is designed to generate a new recommendation list called L_B. L_A and L_B were fused together to obtain the hybrid recommendation system. The hybrid recommendation system could better fit the user preference. More importantly, it was especially better at recommending a group of tourist spots, which means it has higher practical value.

3. The Proposed Recommendation System

3.1. Fundamental Theory. The above-mentioned collaborative filtering theory has been widely used in recommendation system. As it is known, a collaborative filtering-based recommendation system has several apparent advantages: it can process unstructured objects, and it does not need any domain knowledge to discover new user preference. It can generate lots of personalized recommendation results for users. Hence, we adopt the popular collaborative filtering theory into the proposed recommendation system. It first analyzes the tourist spots other users have rated, then it predicts the ratings of the target users for all spots. Therefore, the matrix of users' ratings is factorized into two matrices (user matrix and spot matrix). They all contain a group of latent factors. The user vector is associated with the user matrix while the spot vector is associated with the spot matrix. Dot product between these two vectors is computed to complete the final rating prediction. Meanwhile, the hierarchical sampling statistics model is used to mine users' preference.

In general, the proposed recommendation system for tourist spots has several components, user data collection, hierarchical sampling statistics, SVD++ algorithm, collaborative filtering model, and hybrid recommendation list. Figure 1 illustrates the recommendation framework.

In Figure 1, travel preference is collected through a well-designed questionnaire survey because we cannot grab

travelers' personal information from web. Users' ratings of different tourist spots are crawled from a tourist website (its URL link is http://www.ctrip.com) automatically. Then a data preprocessing strategy is designed to discretize all users' ratings: "0~5" is used to represent users' satisfaction. "5" means the highest positive rating while "0" means the lowest negative rating. Then, the user preference is acquired by using the hierarchical sampling statistics model. The crawled rating data (rated tourist spots) and users' travel preference are matched automatically to create the "Smart Travel" dataset. Based on the users' preference, a new recommendation list called L_A is generated from the statistical perspective. Meanwhile, based on the preprocessed users' ratings, SVD++ algorithm is introduced to complete matrix factorization and both a user matrix and a spot matrix are obtained in turn. The user matrix is also called users' embedding, which represents the latent semantic information of users. The spot matrix is also called spots' embeddings, which represents the latent semantic information of spots. With the help of these two embeddings, it is easy to predict the ratings of the target users for all spots: dot product between these them is computed to complete rating prediction. And a novel recommendation list called L_B is generated on the basis of the rating prediction. Finally, a hybrid recommendation result called L_{Mix} is acquired by fusing L_A and L_B.

3.2. Hierarchical Sampling Statistics Model. The HSS model completes the sampling procedure from different hierarchies randomly on the basis of different proportions. First, the target samples (users) are divided into *num* disjoint subsets in proportion. Second, the sampling procedure is completed independently in each subset. Each subset (e.g., $E_1, E_2, \ldots, E_{num}$) is called a "hierarchy." Finally, *num* subsets are merged into an overall distribution of the target samples. The proposed sampling procedure is described in detail as follows.

Step 1 (select target random variables). Target random variables are introduced to reflect the differences of tourism preference among different kinds of users. They are the key factors for hierarchical sampling statistics. Several target random variables, such as travel season, travel interest, and travel method, are utilized to instruct the following sampling procedure.

Step 2 (divide the target samples into different hierarchies). On the basis of the previously selected target random variables, all target samples are divided into *num* disjoint hierarchies (subsets). If E_i users exist in the i-th ($i = 1, \ldots, num$) hierarchy, then the overall distribution of the target samples is obtained by merging all *num* hierarchies, as shown as formula (1).

$$E = \sum_{i=1}^{num} E_i \tag{1}$$

Step 3 (determine the sampling number of each hierarchy). M is defined as the overall sample size, and *num* is defined as the total number of hierarchies. E_i is defined as the total number of the target users in the i-th hierarchy. On the basis of the definitions, X_i is defined as the number of sampled users in the i-th hierarchy by the proposed HSS model.

$$X_i = M * \frac{E_i}{E} \tag{2}$$

$$s.t. \quad i = 1, \ldots, num \tag{3}$$

Now the target samples are divided into several disjoint hierarchies by the HSS model, which helps decrease the differences within a homogeneous hierarchy and increase the differences between heterogeneous hierarchies. Hence the HSS model strives to sample a certain number of target samples (E_i) to describe the feature space of the current hierarchy and construct the overall distribution of all target samples (E). All classified results are determined through a well-designed questionnaire survey, which can represent the actual tourism preference of users.

On the basis of the subjective weighting method [14], the weight of each population attribute is tuned by the analytic hierarchy process, which means the relative importance of each population attribute should be compared pairwise and a discriminant matrix is established. The weight of each population attribute is illustrated in the discriminant matrix. According to experts' suggestion, six attributes, namely, gender (C_1), district (C_2), age (C_3), education (C_4), job (C_5), and wage (C_6), are chosen for creating the discriminant matrix. These attributes depict users' preference from diverse perspectives. So appropriate weight for each attribute can fit the proposed sampling results. The importance scale of each indicator A relative to index B is as follows: very important = 6, important = 4, slightly important = 2, equally important = 1, slightly minor = 1/2, minor = 1/4, and very minor = 1 / 6. Then a discriminant matrix called G can be constructed as follows. (Note: other values have ever been

used, but relative worse recommendation performance was obtained.)

$$G = \begin{bmatrix} & C_1 & C_2 & C_3 & C_4 & C_5 & C_6 \\ C_1 & 1 & \frac{1}{2} & \frac{1}{4} & \frac{1}{4} & \frac{1}{2} & \frac{1}{6} \\ C_2 & 2 & 1 & \frac{1}{2} & 4 & 4 & \frac{1}{4} \\ C_3 & 4 & 2 & 1 & 1 & \frac{1}{2} & \frac{1}{4} \\ C_4 & 4 & \frac{1}{4} & 1 & 1 & \frac{1}{2} & \frac{1}{2} \\ C_5 & 2 & \frac{1}{4} & 2 & 2 & 1 & \frac{1}{2} \\ C_6 & 6 & 4 & 4 & 4 & 2 & 1 \end{bmatrix} \tag{4}$$

The weight of the k-th population attribute ($k = 1 \sim 6$) is calculated on the basis of the matrix G. Thus each target random variable introduced above can be described by six weighted population attributes. The weight of each population attribute is expressed as follows.

$$W_k = \frac{\prod_{j=1}^{6} G_{kj}}{\sum_{i=1}^{6} \prod_{j=1}^{6} G_{ij}} \tag{5}$$

Finally, all population attributes are ranked by their weights and the corresponding prerecommendation results are generated according to the ranked attributes. In summary, the proposed HSS model for recommendation is described as in Algorithm 1.

3.3. SVD++ Algorithm. SVD++ [15, 16] is an improved algorithm based on the traditional SVD [17] (Singular Value Decomposition) algorithm. It regards the users' rating matrix R as a product of two matrices P and Q. And it also maps both all users and all tourist spots into a K dimensional latent semantic space. The semantic space is made up of a group of latent factors. As described above, the users' rating matrix R is factorized as follows:

$$R_{U \times Z} = P_{U \times K} * Q_{K \times Z}$$

$$= \begin{bmatrix} p_{11}, \ldots, p_{1K} \\ p_{21}, \ldots, p_{2K} \\ \cdots \\ p_{U1}, \ldots, p_{UK} \end{bmatrix} \begin{bmatrix} q_{11}, \ldots, q_{1Z} \\ q_{21}, \ldots, q_{2Z} \\ \cdots \\ q_{K1}, \ldots, q_{KZ} \end{bmatrix} \tag{6}$$

$U = (u_1, u_2, \ldots, u_m)$ represents a user set while $Z = (z_1, z_2, \ldots, z_n)$ represents a tourist spot set. p_{ik} represents the preference degree of the user i for the k-th latent factor of the tourist spots. q_{kj} represents the distribution of the k-th latent factor among the tourist spot j. So each user is associated with a user vector $p_u \in R^K$ which is a row of the matrix P. And each tourist spot is associated with a tourist spot vector $q_z \in R^K$ which is a column of the matrix Q. p_u describes the preference of a user. q_z describes the feature space of a tourist spot. They are now located in a homogeneous data space. Hence, the traditional collaborative filtering theory can

> **Input**: Results of questionnaire survey (T).
> **Output**: Recommendation list (L_A)
> 1. Target random variables including travel season, travel interest, and travel method are used to depict the user preference. Each variable is described by six population attributes namely gender, district, age, education, job, and wage.
> 2. On the basis of the T, the sampling dataset is obtained by the HSS model. and the sampling number of the i-th hierarchy is obtained. which is expressed as Equation (2).
> 3. The proportional value of each attribute hierarchy is calculated by X_W/N, which can depict the actual distribution of the corresponding attribute.
> 4. On the basis of the preceding proportional values, the relative importance of each attribute is determined by the subjective weighting method, and a discriminant matrix (G) is obtained, which is expressed as Equation (4).
> 5. The weight of each attribute hierarchy is calculated on the basis of the matrix G and Equation (5), and each target random variable is described by the six weighted previously population attributes.
> 6. Population attributes are ranked by their weights and recommendation results are generated according to the ranked attributes.
> 7. Recommendation list L_A is generated by matching the above recommendation results and users' population attributes collected from the survey.

ALGORITHM 1: Hierarchical sampling statistics model for recommendation.

be applied here: the dot product between these two vectors is utilized to get the prediction rating $\hat{r}_{uz}$ which is rated by the user u for the tourist spot z:

$$\hat{r}_{uz} = \boldsymbol{p}_u^T \boldsymbol{q}_z = \boldsymbol{q}_z^T \boldsymbol{p}_u \tag{7}$$

By analyzing the users' ratings matrix $\boldsymbol{R}$, we find some users always give high ratings (or low ratings) compared with others. This means the corresponding ratings are biased. However, (7) does not consider the bias. Hence several key bias factors should be taken into account, which contributes to getting more objective ratings. The modified rating equation is shown as follows:

$$\hat{r}_{uz} = b_{uz} + \boldsymbol{q}_z^T \boldsymbol{p}_u \tag{8}$$

$$b_{uz} = \mu + b_z + b_u \tag{9}$$

$$\hat{r}_{uz} = \mu + b_z + b_u + \boldsymbol{q}_z^T \boldsymbol{p}_u \tag{10}$$

$\boldsymbol{b}_{uz}$ represents the overall "bias information" of the user u for the spot z. μ is a rating mean. $\boldsymbol{b}_z$ represents the "bias information" of the tourist spot z, which is an item offset to the rating mean. $\boldsymbol{b}_u$ represents the "bias information" of the user u, which is a user offset to the rating mean. In addition to the "bias information", many implicit parameters are added to the SVD++ algorithm to better reflect users'

latent preference for tourist spots. Generally, users' ratings are usually called "explicit information" while users' behaviors are usually called "implicit information". Finally, a novel perspective of users' preference is acquired by combining the above-mentioned "explicit information", "bias information", and "implicit information" together. The final rating equation is shown as follows:

$$\hat{r}_{uz} = \mu + b_z + b_u + \boldsymbol{q}_z^T \left(\boldsymbol{p}_u + |N_u|^{-1/2} \sum_{j \in N_u} \boldsymbol{y}_j \right) \tag{11}$$

N_u represents the behavior data of the user u. This means the user u has rated the corresponding tourist spots. $|N_u|$ is the size of the behavior data. $-(1/2)$ is the contraction factor which is an empirical value. $\boldsymbol{y}_j$ represents the implicit parameters used for describing the "implicit information" in recommendation, which indicates that the user u has rated the tourist spot j. Finally, besides "explicit information", the SVD++ algorithm takes both the above-mentioned "bias information" and "implicit information" into account. It firstly analyzes users' preference degree for the latent semantic factors of the tourist spots. Then it obtains the distribution of the latent semantic factors among all tourist spots. Lastly, it takes both the above-mentioned "bias information" and "implicit information" into account. Based on the above analysis, the cost function J of the SVD++ algorithm is shown as follows:

$$J = \min_{\boldsymbol{b}_z, \boldsymbol{b}_u, \boldsymbol{q}_z, \boldsymbol{p}_u} \sum_{(u,z) \in K} \left(r_{uz} - \mu - \boldsymbol{b}_u - \boldsymbol{b}_z - \boldsymbol{q}_z^T \left(\boldsymbol{p}_u + |N_u|^{-1/2} \sum_{j \in N_u} \boldsymbol{y}_j \right) \right)^2$$

$$+ \lambda \left\{ \sum_u \left(\boldsymbol{b}_u^2 + \|\boldsymbol{p}_u\|^2 \right) + \sum_z \left(\boldsymbol{b}_z^2 + \|\boldsymbol{q}_z\|^2 + \|\boldsymbol{y}_z\|^2 \right) \right\} \tag{12}$$

Input: Users' rating matrix (R)
Output: Rating prediction matrix (D) and Recommendation list (L_B)
1. Compute the mean rating based on the matrix R.
2. Initialize the "bias information" b_u and b_z. Initialize the user vector p_u. Initialize the tourist spot vector p_u. Initialize the implicit parameters y_j.
3. User ratings are grouped by user ids.
4. Compute the inner product of the user vector p_u and the tourist spot vector q_z by Equation (11). User's ratings are predicted.
5. Compute the prediction error e_{uz} based on the real rating and the predicted rating. As shown in Equation (13)~(18), the SGD method is utilized to complete optimization.
6. Repeat the third step and fourth step to get the prediction rating r_{uz}. Update the rating prediction matrix D.
7. Generate the recommendation list L_B based on the matrix D.

ALGORITHM 2: Proposed SVD++ based collaborative filtering model.

Input: Questionnaire results T, Users' rating matrix R
Output: Hybrid recommendation list L_{Mix}
1. Generate the recommendation list L_A using Algorithm 1.
2. Generate the recommendation list L_B using Algorithm 2.
3. The mix merge method is used to generate the recommendation list L_{Mix}.
4. Compute the precision, recall, and F1 values (Equations (21)–(24), respectively) of the hybrid recommendation system.

ALGORITHM 3: A novel recommendation system for tourist spots.

The first part of J is the loss based on the least square method. The second part of J is the regularization term. The proposed SVD++ algorithm is optimized by the stochastic gradient descent method:

$$b_u \longleftarrow b_u + \gamma \left(e_{uz} - \lambda b_u \right) \tag{13}$$

$$b_z \longleftarrow b_z + \gamma \left(e_{uz} - \lambda b_z \right) \tag{14}$$

$$p_u \longleftarrow p_u + \gamma \left(e_{uz} q_z - \lambda p_u \right) \tag{15}$$

$$q_z \longleftarrow q_z + \gamma \left(e_{uz} \left(p_u + |N_u|^{-1/2} \sum_{j \in N_u} y_j \right) - \lambda q_z \right) \tag{16}$$

$$y_j \longleftarrow y_j + \gamma \left(e_{uz} |N_u|^{-1/2} q_z - \lambda q_z \right) \tag{17}$$

$$e_{uz} = r_{uz} - \hat{r}_{uz} \tag{18}$$

e_{uz} is the prediction error, γ is the learning rate, and λ is the regularization parameter.

3.4. The Proposed Collaborative Filtering Model. As described above, the SVD++ algorithm decomposes the users' rating matrix R into two matrices P and Q that contain a group of latent semantic factors. Each user is associated with a user vector that is a row of the matrix P. Each tourist spot is associated with a tourist spot vector that is a column of the matrix Q. By computing the inner product of these two vectors and taking both the above-mentioned "bias

information" and "implicit information" into account, the corresponding ratings of users can be predicted. Algorithm 2 shows the details about the proposed SVD++ based collaborative filtering model.

3.5. Novel Recommendation System for Tourist Spots. As described in Section 3.2, the HSS model is used to generate the recommendation list L_A from the statistical perspective. As described in Section 3.4, the proposed SVD++ algorithm is used to generate the recommendation list L_B from the matrix factorization perspective. On the basis of the two lists, a novel hybrid recommendation system for tourist spots is achieved, which is shown in Algorithm 3.

4. Experimental Results and Discussions

4.1. The Dataset. All user ratings in the Smart Travel dataset is crawled from http://www.Ctrip.com. approximately 5000 users' ratings of 60 tourist spots are crawled from the website. On the basis of experts' suggestion, we classify the 60 tourist spots into 8 categories: Seashore Island (SI), World Heritage (WH), Blessing Buddha (BB), Cruise Trip (CT), Ancient Town (AT), Family Travel (FT), Health Leisure (HL), and Folk Experience (FE). About 4000 ratings (80%) are select randomly for training while the rest 1000 ratings (20%) are selected for testing. Then a well-designed questionnaire survey is used to collect users' preference from diverse population attributes, namely, gender, age, district, and wage. It is a very important complementary part of user ratings,
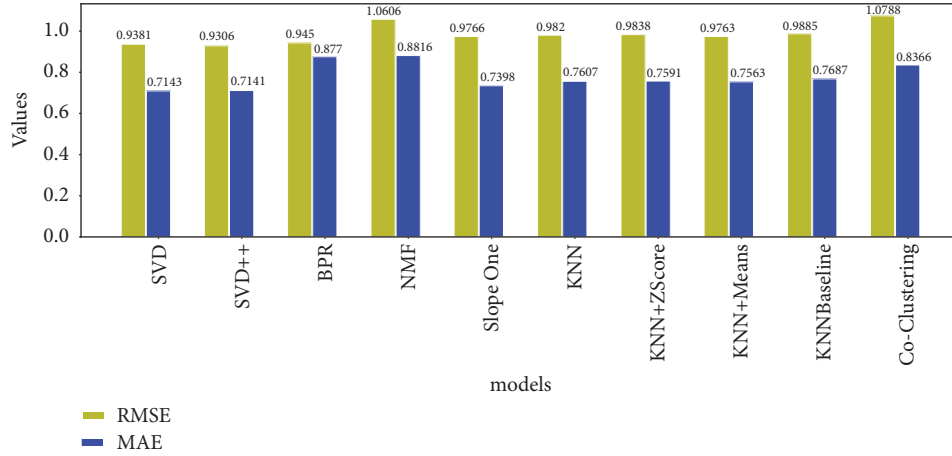
Figure 2: The prediction rating comparisons with baselines.

which can assist the proposed model in recommending better performance. Finally, all user ratings are discretized: "0~5" represent users' satisfaction degree, where "5" means the highest positive rating whereas "0" represents the lowest negative rating. Users' population attributes such as gender, age, district, and wage are discretized into integer values too.

Generally, some detailed information about the "Smart Travel" dataset and the well-designed questionnaire survey can be found in https://github.com/CVNLP/SmartTravel.

4.2. Results of Hierarchical Sampling Statistics Model. We randomly sent the survey table to 2170 visitors and completed the corresponding statistical analysis by the proposed HSS model. This survey can better cover the key population attributes of the interviewed visitors. In each population attribute, we took 1,000 visitors as sampling objects and got the statistics results of the corresponding preference. Three angles, namely, travel season, travel interest, and travel method, are chosen as the tourist preference. On the basis of this setting, the final HSS statistics can be computed. And several valuable conclusions can be obtained from the HSS statistics [9]. Most of them are close to people's cognition: (1) Most interviewees prefer to travel in spring and fall. Tourist spots in spring (or fall) are more attractive than those in other seasons, given the suitable temperature and beautiful scenes; thus, people are willing to embrace nature. Meanwhile, most students (age ≤ 20) prefer to travel in summer during their vacation, which is a good opportunity for them to travel freely and extend their knowledge. (2) Most interviewees prefer to travel with their families, which is especially for the aged people (age >=40). We guess it is due to Chinese traditional culture. Chinese enjoy themselves in a family trip. Surprisingly, males prefer to travel by themselves. (3) Most middle-aged or the elder (age >=50) prefer to choose the tourist spots of "HL" since these tourist spots require less physical strength. (4) Most interviewees from the east district (or the central district) in China prefer to choose the tourist spots of "SI", while most interviewees from the west district (or the north district) in China prefer to choose the tourist spots of "FE" and "AT". We guess this is mainly due to the

living habits of the interviewees. (5) Most interviewees with a master's degree or above tend to travel by themselves.

4.3. RMSE and MAE Evaluations. We use both Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) to compute the corresponding rating prediction accuracy. Small RMSE and MAE indicate good recommendation performance. RMSE and MAE are expressed as (19) and (20), respectively, where *Test* is the testing dataset; $|Test|$ is the number of testing data; (u, z) is a user-spot combination in the user rating matrix, where z and u denote a spot and a user, respectively; r_{uz} represents the actual rating value user u rated spot z in the Wisdom Tourist dataset; and $\hat{r}_{uz}$ is the predicted rating generated by the recommendation system.

$$RMSE = \sqrt{\frac{1}{|Test|} \sum_{(u,z)\in Test} \left(\hat{r}_{uz} - r_{uz}\right)^2} \qquad (19)$$

$$MAE = \sqrt{\frac{1}{|Test|} \sum_{(u,z)\in Test} \left\|\hat{r}_{uz} - r_{uz}\right\|} \qquad (20)$$

The proposed SVD++ based recommendation system is compared with a group of traditional recommendation systems namely SVD [17], BPR [9], NMF [18], Slope One [19], KNN [20], "KNN + Zscore" [21], "KNN + means" [22], KNN (Baseline) [22], and Co-Clustering [22]. Experimental results are shown in Figure 2.

The RMSE values of the proposed SVD++ algorithm decrease about 0.8%, 1.55%, 13.97%, 5.05%, 5.52%, 5.72%, 4.91%, 6.22%, and 15.92%, respectively, compared with SVD, BPR, NMF, Slope One, KNN, "KNN + Zscore", "KNN + means", KNN (Baseline), and Co-Clustering. The MAE values of the proposed SVD++ algorithm decrease about 0.03%, 22.81%, 23.46%, 3.6%, 6.53%, 6.3%, 5.91%, 7.65%, and 16.73%, respectively. These indicate the proposed SVD++ algorithm can improve the rating prediction accuracy, which helps to generate a better recommendation list (L_B) for users. The implicit reasons are as follows: (1) Domain knowledge of tourist spots plays an important role; all tourist spots are classified into eight categories. The idea is very close to
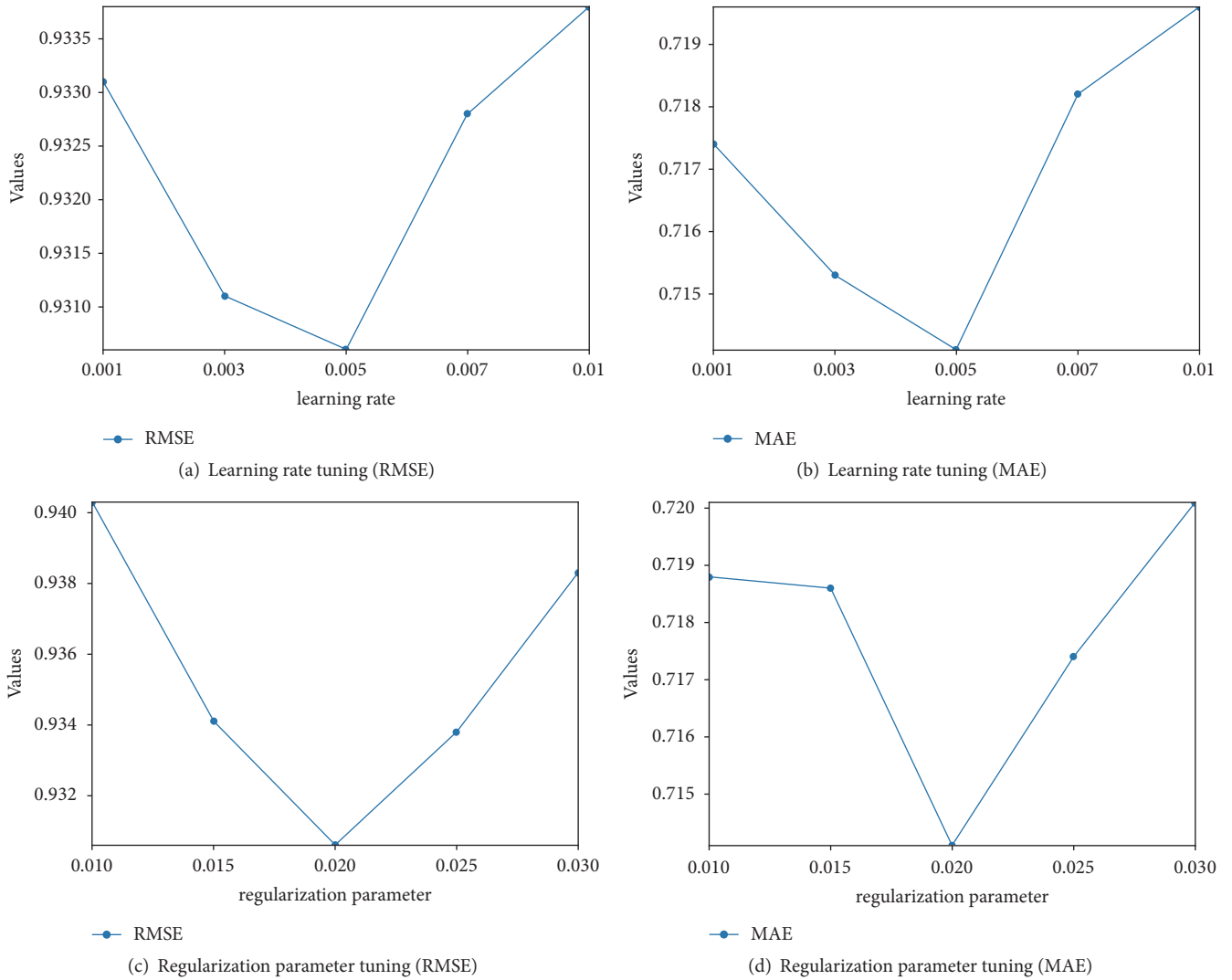
(a) Learning rate tuning (RMSE)

(b) Learning rate tuning (MAE)

(c) Regularization parameter tuning (RMSE)

(d) Regularization parameter tuning (MAE)

FIGURE 3: The parameter tuning procedure of the proposed SVD++ algorithm.

people's cognition. (2) The proposed SVD++ algorithm finds out the latent semantic space which is suitable for depicting the user preference. (3) The proposed SVD++ algorithm also finds out the latent semantic space which is suitable for depicting different tourist spots. (4) Both the above-mentioned "implicit information" and "bias information" also play very important roles in the recommendation procedure.

As described in Section 3.2, the proposed SVD++ algorithm plays the most important role in the hybrid recommendation system. So it should be tuned carefully. The SVD++ algorithm has several key parameters such as the regularization parameter (λ), the learning rate parameter (γ), and different split ratio. These parameters are tuned by cross-validation. The corresponding results are shown in Figures 3-4:

In Figures 3(a)–3(d), when $\gamma = 0.005$, $\lambda = 0.02$ the proposed SVD++ algorithm achieves the best recommendation performance (its MAE = 0.7141 and its RMSE = 0.9306 in Figure 2). Second, in all subfigures, the variation trends of

MAE are similar to that of RMSE, which demonstrates that the proposed SVD++ algorithm is more robust.

In Figure 4, different split ratios are utilized to validate the best split method (other parameters are fixed, for example γ =0.005 and λ =0.02).Obviously,when the train set accounts for 80%, the proposed SVD++ algorithm achieves the best recommendation performance.

4.4. Precision, Recall, and F1 Evaluations. As shown in HSS statistics result, different population attributes determine the corresponding travel preference. Based on the fact, a new model called Our_HS is created: as described in Section 3.2, the weight of each attribute is set by the subjective weight evaluation method [14]: "gender" is 0.2, "district" is 0.05, "age" is 0.3, "education" is 0.05, "job" is 0.1, and "wage" is 0.3. Moreover, a new model called Our_Mix is created too: the recommendation list L_A generated by the Our_HS model is fused with the recommendation list L_B generated
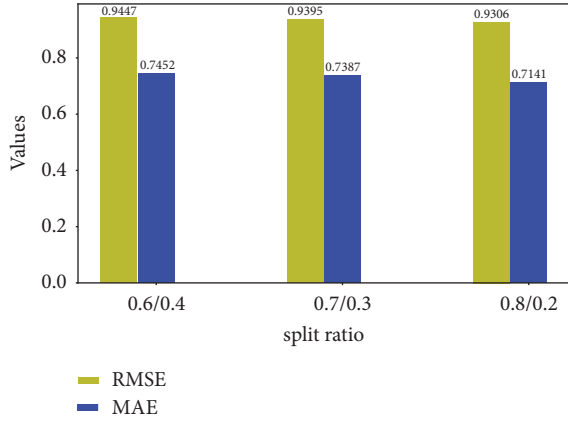
FIGURE 4: Different recommendation performance by setting different split ratios.

TABLE 1: The classification method of the recommended tourist spots.

User Preference	Recommended	Not Recommended
Prefer	True-Positive (tp)	False-Negative (fn)
Not Prefer	False –Positive (fp)	True- Negative (tn)

by the proposed SVD++ algorithm. Finally, a new list L_{Mix} is obtained.

In addition to MAE and RMSE, Precision, Recall, and F-Score are often used to evaluate recommendation system. The type that the recommended tourist spots belong to should be determined before computing the corresponding Precision, Recall, and F-Score values. In Table 1, "True-Positive(tp)" indicates the recommended tourist spots are preferred by users, "False-Positive(fp)" indicates the recommended tourist spots are not preferred by users, "False-Negative (fn)" indicates the tourist spots which are not recommended are preferred by users, and "True-Negative(tn)" indicates the tourist spots which are not recommended are not preferred by users.

Based on Table 1, Precision, Recall, and F-Score are computed as follows:

$$\text{Precision} = \frac{\#tp}{\#tp + \#fp} \tag{21}$$

$$\text{Recall} = \frac{\#tp}{\#tp + \#fn} \tag{22}$$

$$\text{F-Score} = \frac{\left(\alpha^2 + 1\right) * \text{Precision} * \text{Recall}}{\alpha^2 \left(\text{Precision} + \text{Recall}\right)} \tag{23}$$

means the number. The F1 (α=1) metric (see (24)) is used to evaluate the overall recommendation performance, when N' tourist spots are recommended to users. A novel metric named P' (see (25)) is used to evaluate the mean performance.

$$\text{F1} = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}} \tag{24}$$

$$P' = \frac{\sum_{w=1}^{N'} index\,(w)}{N'} \tag{25}$$

The *index (w)* represents the evaluation metric described above such as Precision, Recall, and F1. N'=1~10. Precisions are listed in Table 2, Recalls are listed in Table 3, and F1 values are listed in Table 4.

In Table 2, the "KNN + means" algorithm acquires the best Precision when N'=1; it is better for recommending only one tourist spot. However, this cannot meet users' personalized requirement. Except the SVD, SVD ++, NMF, and KNN algorithms, the Precision cannot obtain significant improvement when more tourist spots are recommended to users because the rating matrix $\boldsymbol{R}$ is sparse. The proposed SVD++ algorithm can obtain higher Precision and beat other baselines when $N'\geq3$. This indicates the proposed SVD++ algorithm is better at recommending more tourist spots. It is very close to people's cognition and it can better meet users' personalized requirement. However, the superiority of the SVD++ algorithm is a margin compared with the best competitor. It still has a large improvement space.

As expected, the proposed Our_Mix model wins nine first places among ten results. Figure 5(a) describes the precision improvement between different models with respect to N'. For example, the precision improvement of the proposed Our_Mix model is improved about 2.49% when N'=2. Meanwhile, the precision improvements of the proposed Our_Mix model are always positive, and it has an increasing trend. On the contrary, the SVD ++ algorithm has a tortuous trend which twists and turns. More importantly, the overall improvement of the SVD++ algorithm is far less than that of the Our_Mix model. Hence, when more and more tourist spots are recommended to users, the proposed Our_Mix model is more stable than any other baseline. We conclude the proposed SVD++ algorithm plays a major role while the HSS model plays a minor role. They complement each other to obtain the best recommendation performance.

In summary, based on the P' values, we find, compared with the best competitor, the mean precision of the proposed Our_Mix model is improved about (65.18%-60.66%)/ 60.66% ≈7.5%. It is a larger margin, which indicates the proposed hybrid recommendation model is more practical than any other baseline. Moreover, the mean precision descend order of all models is shown as Our_Mix>SVD++>"KNN + means"> "KNN + Zscore"> SVD > "KNN Baseline">Slope One>Co-Clustering>NMF>KNN>BPR>Our_HS.

In Table 3, several baselines, i.e., SVD++, Co-Clustering and KNN, obtain better recall values. Among ten results, the proposed SVD++ algorithm wins one first place and seven second places. It is better at recommending a group of tourist spots (especially for $N'\geq5$). As expected, the proposed Our_Mix model gets the best recall values. When more spots are recommended to users, the proposed Our_Mix model is more stable and superior to any other baseline. Figure 5(b) indicates the recall improvement between different models with respect to N'. For example, the recall value of the proposed Our_Mix model is improved about 2.49% when N'=2. The recall improvements of the Our_Mix model are always positive and it has an increasing trend.

In summary, based on the P' values, compared with the best competitor, the mean recall of the proposed Our_Mix model is improved about (18.67%-17.57%)/17.57% ≈6.2%, It

TABLE 2: Precisions of different algorithms (the best value of each row is shown as **60.52**, the suboptimal value of each row is shown as **59.84**, and unit is %).

N'	SVD	SVD++	NMF	Slope One	KNN	KNN + Zscore	KNN + means	KNN Baseline	Co-Clustering	BPR	Our_HS	Our_Mix
1	59.80	59.29	58.60	59.05	57.31	58.42	**60.52**	58.08	58.60	37.77	5.33	**60.03**
2	58.91	58.82	57.86	58.24	56.43	59.20	**59.80**	58.96	57.66	37.58	8.62	**62.52**
3	60.27	**60.76**	57.43	58.86	57.32	59.32	60.62	58.98	58.41	39.32	7.81	**63.07**
4	59.51	**61.75**	57.96	58.93	57.68	60.35	60.60	59.36	58.80	39.77	8.62	**65.86**
5	59.07	**60.69**	57.50	59.06	57.83	60.39	59.84	60.29	58.87	40.05	8.38	**65.75**
6	60.38	59.21	58.59	59.73	56.10	60.28	**60.39**	58.67	58.65	40.49	8.77	**65.99**
7	58.74	**61.15**	58.98	59.98	57.56	60.48	60.47	60.60	59.26	40.56	9.08	**66.87**
8	60.07	**61.15**	59.06	60.02	58.15	60.67	60.51	60.69	59.57	41.28	9.01	**66.98**
9	61.43	**61.49**	59.36	60.25	58.79	60.76	60.75	60.84	59.69	41.04	8.51	**67.08**
10	62.16	**62.33**	59.40	60.32	59.42	61.15	60.83	59.98	60.03	41.67	9.08	**67.69**
$\bar{P}$	60.04	**60.66**	58.47	59.44	57.66	60.10	60.43	59.65	58.95	39.95	8.32	**65.18**

TABLE 3: Recall values of different algorithms (the best value of each row is shown as **16.13**, the suboptimal value of each row is shown as _17.89_, and unit is %).

N'	SVD	SVD++	NMF	Slope One	KNN	KNN + Zscore	KNN + means	KNN Baseline	Co-Clustering	BPR	Our_HS	Our_Mix
1	13.31	15.94	14.92	13.39	15.93	15.36	15.04	14.57	**16.13**	9.38	0.33	_15.94_
2	15.36	_17.89_	15.55	15.40	16.41	14.08	14.36	14.71	16.73	10.25	0.80	_17.96_
3	13.40	_17.93_	12.65	13.95	15.87	13.57	14.39	15.23	17.63	9.91	1.12	_17.83_
4	14.81	16.54	14.30	14.01	16.27	15.48	14.61	15.63	**16.77**	10.42	1.80	_17.84_
5	14.99	_18.29_	15.76	14.88	16.32	15.51	14.86	15.42	17.16	10.62	2.11	_18.69_
6	16.21	14.65	15.78	15.80	16.61	16.62	14.97	16.31	**17.47**	9.89	2.34	_17.66_
7	17.97	_18.86_	16.13	16.39	16.85	16.73	16.54	16.47	18.85	11.64	3.14	_19.20_
8	17.20	_18.65_	16.38	16.45	16.94	16.74	16.81	16.58	16.13	12.08	3.43	_19.68_
9	17.29	_18.75_	16.83	17.17	16.96	16.82	16.23	16.74	17.49	13.07	4.81	_20.20_
10	18.16	_18.25_	16.28	17.39	17.56	16.70	16.92	17.08	17.64	13.12	5.56	_21.70_
$\bar{P}$	15.87	17.57	15.46	15.48	16.57	15.76	15.47	15.88	17.20	11.04	2.54	_18.67_

TABLE 4: F1 values of different algorithms (the best value of each row is shown as **25.30**, the suboptimal value of each row is shown as **27.44**, and unit is %).

N'	SVD	SVD++	NMF	Slope One	KNN	KNN + Zscore	KNN + means	KNN Baseline	Co-Clustering	BPR	Our_HS	Our_Mix
1	21.77	25.13	23.78	21.83	24.93	24.32	24.09	23.30	**25.30**	15.03	0.62	**25.19**
2	24.37	**27.44**	24.51	24.36	25.43	22.75	23.16	23.55	25.93	16.11	1.46	**27.90**
3	21.93	**27.69**	20.73	22.55	24.86	22.09	23.26	24.21	27.08	15.84	1.96	**27.80**
4	23.72	26.09	22.94	22.64	25.38	24.64	23.54	24.74	**26.10**	16.51	2.98	**28.08**
5	23.91	**28.11**	24.74	23.77	25.46	24.68	23.81	24.56	26.57	16.79	3.37	**29.11**
6	25.56	23.49	24.86	24.99	25.63	26.06	23.99	25.52	**26.92**	15.89	3.69	**27.86**
7	27.52	**28.83**	25.33	25.74	26.07	26.21	25.98	25.90	28.60	18.08	4.67	**29.83**
8	26.74	**28.58**	25.65	25.82	26.24	26.24	26.31	26.04	25.39	18.69	4.97	**30.42**
9	26.98	**28.74**	26.22	26.72	26.33	26.35	25.62	26.26	27.05	19.83	6.15	**31.05**
10	28.11	**28.23**	25.56	27.00	27.11	26.24	26.48	26.59	27.27	19.96	6.90	**32.86**
$\overline{P}$	25.06	27.23	24.43	24.54	25.74	24.96	24.62	25.07	26.62	17.30	3.68	**29.01**

(a) The precision improvements
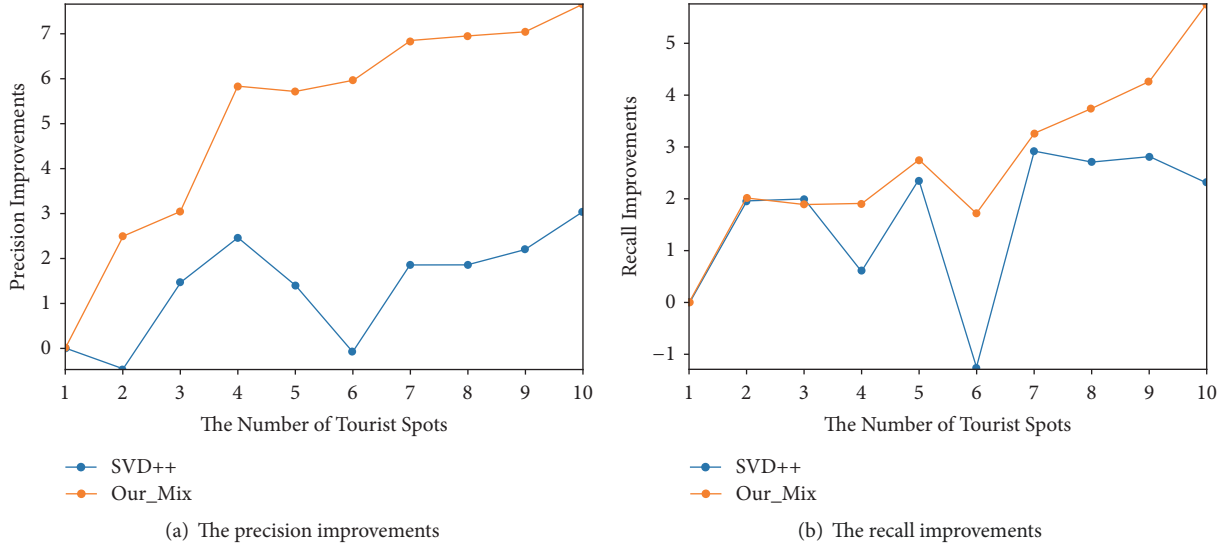

(b) The recall improvements

FIGURE 5: The performance comparisons.

TABLE 5: Case study results.

	Ground-truth	SVD++	Co-Clustering	KNN	Our_Mix
User id: 9	Jiuhua Mountain	Sri Lanka	Gulangyu islet	Phuket	**Jiuhua Mountain** √
	Huanglong	**Jiuhua Mountain** √	Kyoto	Nara	Fuji
	Bali	**Lijiang** √	**Huanglong** √	**Bali** √	**Bali** √
	Lijiang	Shangri-la	Sanya	**Lijiang** √	Cairo
		Nara	Guilin	Wuhan	**Lijiang** √
		Hong Kong	Nusa Lembongan	Indonesia	Kyoto
User id: 53	Terra Cotta Warriors	Wutai Mountain	Bali	Sanya	**Nara** √
	Nara	**Kyoto** √	Wuyuan	Lijiang	**Kyoto** √
	Kyoto	Cairo	**Shangri-la** √	Tokyo	Cairo
		Disney	Jiuzhangou	**Terra Cotta Warriors** √	Sri Lanka
		Nara √	Munich	Haikou	**Terra Cotta Warriors** √
		Phuket	Dali	Disney	Phuket

is also a larger margin, which indicates the Our_Mix model is more practical than other models. Moreover, the mean recall descend order of all models is shown as Our_Mix >SVD++ > Co-Clustering > KNN > "KNN Baseline" > SVD > "KNN + Zscore" > Slope One > "KNN + means" > NMF > "BPR" > Our_HS.

Table 4 is more valuable because the F1 metric focuses on evaluating the overall recommendation performance. In Table 4, the proposed Our_Mix model acquires the best F1 values. Among ten results, it wins nine first place and one second place. When more tourist spots are recommended to users, the Our_Mix model is more stable and it has an increasing trend. Secondly, based on the P' values, compared with the best competitor, the mean F1 of the proposed Our_Mix model is improved about (29.01%-27.23%)/27.23% ≈6.5%. The value is between 6.2% (the mean recall improvement of the Our_Mix model) and 7.5% (the mean precision improvement of the Our_Mix model) because F1 is an overall evaluation matric. Moreover, the value is a larger margin compared with

the best competitor, which indicates the Our_Mix model is more practical. Finally, the mean F1 descend order of all models is shown as: Our_Mix > SVD++ > Co-Clustering > KNN > SVD > "KNN + Zscore" > "KNN + means" > Slope One > NMF > "BPR" > Our_HS.

All in all, the proposed Our_Mix model obtains the best recommendation performance. It also is better at recommending a group of tourist spots, which really helps to improve users' interactive experience and enhance the influence of tourism websites.

4.5. Case Study. In this section, we will show the advantages of the proposed model through some quintessential examples. We provide the recommendation lists in Table 5, which is built by three baselines (SVD++, Co-Clustering, and KNN. They get better performance in Table 4) and the proposed Our_Mix model. Two users are randomly selected to complete case study. The first column presents the ground-truth

of the two users, and the following columns are the results of different recommendation systems.

As shown in Table 5, the proposed hybrid model predicts more positive tourist spots and more accurate tourist spots. For example, the tourist spots that the user "53" prefers to, i.e. "Terra Cotta Warriors", "Nara" and "Kyoto", are predicted accurately by the proposed hybrid model. We conclude that our model benefits from the hybrid mode which consists of the HSS model and the SVD++ algorithm. The proposed SVD++ algorithm finds out more valuable latent semantic space which can better describe users' preference. Moreover, the HSS model is a good complementary to the SVD++ algorithm.

5. Conclusions and Future Works

Recommendation system is an effective method to resolve the "information overload" problem and it can further promote the value of information. As a hot research topic, recommendation system for tourist spots is chosen as our research object: a new dataset named "Smart Travel" is created firstly. After that, a novel recommendation system is designed by fusing the HSS model and the SVD ++ algorithm. Experimental results demonstrated that the mean precision, mean recall, and mean F1 of the proposed hybrid model improve about 7.5%, 6.2%, and 6.5%, respectively, compared with the best competitor. The novel recommendation system is especially better at recommending a group of tourist spots. Our future works are shown as follows:

(1) The user preference will be modeled by spots' images. We plan to introduce the state-of-the-art relative model [23] to describe the extent variation of users' preference, which may contribute to characterizing users' preference more accurately.

(2) Several state-of-the-art deep learning-based models such as CNN [24] and DSSM [25] will be introduced to better mine the nonlinear relationship between users and tourist spots, which also help to improve the final recommendation performance.

(3) The state-of-the-art IRGAN [26] model will be used to train a more robust recommendation system. It will give us some new intuitions about designing a novel recommendation system.

Data Availability

The data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

References

[1] L. Lü, M. Medo, C. H. Yeung, Y. Zhang, Z. Zhang, and T. Zhou, "Recommender systems," *Physics Reports*, vol. 519, no. 1, pp. 1–49, 2012.

[2] J. Bobadilla, F. Ortega, A. Hernando, and A. Gutiérrez, "Recommender systems survey," *Knowledge-Based Systems*, vol. 46, pp. 109–132, 2013.

[3] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry, "Using collaborative filtering to weave an information tapestry," *Communications of the ACM*, vol. 35, no. 12, pp. 61–70, 1992.

[4] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "GroupLens: An open architecture for collaborative filtering of netnews," in *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pp. 175–186, CSCW, 1994.

[5] J. A. Konstan, B. N. Miller, D. Maltz, J. L. Herlocker, L. R. Gordon, and J. Riedl, "Applying collaborative filtering to usenet news," *Communications of the ACM*, vol. 40, no. 3, pp. 77–87, 1997.

[6] B. Miller N, J. Riedl T, and J. Konstan A, "Experiences with grouplens: making usenet useful again," in *Proceedings of the USENIX Winter Technical Conference*, pp. 219–231, Anaheim, US, 1997.

[7] L. Yu, C. Liu, and Z. Zhang, "Multi-linear interactive matrix factorization," *Knowledge-Based Systems*, vol. 85, pp. 307–315, 2015.

[8] L. Yu, J. Huang, G. Zhou, C. Liu, and Z. Zhang, "TIIREC: A tensor approach for tag-driven item recommendation with sparse user generated content," *Information Sciences*, vol. 411, pp. 122–135, 2017.

[9] Li. G-L, T. Zhu, J. Hua et al., "Hybrid recommendation system for tourist spots based on hierarchical sampling statistics and Bayesian personalized ranking," *Journal of Central China Normal University (Natural Sciences)*, vol. 53, no. 2, pp. 214–221, 2019.

[10] G. Fenza, E. Fischetti, D. Furno, and V. Loia, "A hybrid context aware system for tourist guidance based on collaborative filtering," in *Proceedings of the IEEE International Conference on Fuzzy Systems (FUZZ '11)*, pp. 131–138, 2011.

[11] F.-M. Hsu, Y.-T. Lin, and T.-K. Ho, "Design and implementation of an intelligent recommendation system for tourist attractions: the integration of EBM model, Bayesian network and Google Maps," *Expert Systems with Applications*, vol. 39, no. 3, pp. 3257–3264, 2012.

[12] M. Nilashi, O. bin Ibrahim, N. Ithnin, and N. H. Sarmin, "A multi-criteria collaborative filtering recommender system for

the tourism domain using expectation maximization (EM) and PCA–ANFIS," *Electronic Commerce Research and Applications*, vol. 14, no. 6, pp. 542–562, 2015.

[13] L. Zheng, F. X. Zhu, and X. Yao, "Recommendation rating prediction based on attribute boosting with partial sampling," *Chinese Journal of Computers. Jisuanji Xuebao*, vol. 39, no. 8, pp. 1501–1514, 2016.

[14] T. Li, *Application and Practice of Data Mining*, Xiamen University press, 2013.

[15] Y. Koren, "Factorization meets the neighborhood: a multifaceted collaborative filtering model," in *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 426–434, ACM, 2008.

[16] J. Luo, C. Jiang, and P. Tian, "A hybrid recommendation algorithm with LDA and SVD++ considering the news timeliness," *Revista de La Facultad de Ingenieria*, vol. 31, no. 12, pp. 1–7, 2017.

[17] X. Zhou, J. He, G. Huang, and Y. Zhang, "SVD-based incremental approaches for recommender systems," *Journal of Computer and System Sciences*, vol. 81, no. 4, pp. 717–733, 2015.

[18] C.-J. Lin, "Projected gradient methods for nonnegative matrix factorization," *Neural Computation*, vol. 19, no. 10, pp. 2756–2779, 2007.

[19] D. Lemire and A. Maclachlan, "Slope one predictors for online rating-based collaborative filtering," in *Proceedings of SIAM Data Mining*, 2005.

[20] Z. Zhang, "Introduction to machine learning: K-nearest neighbors," *Annals of Translational Medicine*, vol. 4, no. 11, 2016.

[21] P. Bruce and A. Bruce, "Practical statistics for data scientists," *O'Reilly Media*, pp. 216–218, 2017.

[22] H. Nicolas, *Surprise, A Python Library for Recommender Systems*, 2017, http://surpriselib.com.

[23] D. Jayaraman, F. Sha, and K. Grauman, "Decorrelating semantic visual attributes by resisting the urge to share," in *Proceedings of Computer Vision and Pattern Recognition*, pp. 1629–1636, 2014.

[24] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

[25] H. Palangi, L. Deng, Y. Shen et al., *Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval*, Computer Science, 2015.

[26] J. Wang, L. Yu, W. Zhang et al., "IRGAN: A minimax game for unifying generative and discriminative information retrieval models," in *Proceeding of the 40th Int ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 515–524, ACM, 2017.