The Tourism Recommendation of Jingdezhen Based on Unifying User-based and Item-based Collaborative filtering

Tao Liu[1, a], Mingang Wu[1, b] and Donglan Ying[2, c]

[1]School of Information Engineering, Jingdezhen ceramic University, Jingdezhen 333403, China.

[2]Department of Foreign Language, Jingdezhen ceramic University, Jingdezhen 333403, China.

[a]10922475@qq.com, [b]2383127286@qq.com, [c]ltjdz@126.com

Abstract

The vigorous development of tourism and high mass tourism enthusiasm makes huge amounts of information overloaded provided by the current mainstream tourism information service platform. An important way to the problem of travel information overloaded is a personalized recommendation technology. This paper tries to address the issue how recommend the attractions of Jingdezhen to the tourist based on unifying user-base and item-base collaborative filtering. Otherwise, realize the existing hybrid model and propose the strategy to meet the recommend of specific tourism area with very few network information. The information outside the network was utilized to computer the similarity of attractions and the tourist similarity takes into account of the social information. The results demonstrate the hybrid algorithm that we realized can competent to recommend.

Keywords

Tourism recommendation, co-rating, collaborative filtering, hybrid algorithm, Jingdezhen.

1. Introduction

The rapid growth of web and tourism has become source of large amount of available online information. One hand, this information may be helpful for tourist to select the more perfect attraction to tour. On the other hand is led to information overloaded problem. In order to solve the problem of "information overloaded", many large e-commerce sites, such as Amazon, Alibaba, Dangdang, Ebay, etc. use various forms of recommendation systems. Recommender system plays the role of generating suggestions by collecting user information such as preferences, interests, and locations [1]. The research on recommender systems gained importance after the emergence of collaborative filtering [2, 3]. Current collaborative filtering recommendation algorithm can be mainly divided into two categories, which are memory-based algorithm and model-based algorithm. The memory-based algorithm generates the recommendation according to similar neighbors. The model-based algorithm produces recommendations by establishing grading model. Existing studies show the memory-based algorithm can obtain better accuracy, but they cannot solve scalability problems caused by increased data [4], model-based algorithm has a better scalability, but its quality of recommendation is not as good as the memory-based algorithm because the model cannot show the diversity of users' interests and hobbies [5]. Memory-based algorithm can be divided into user-based and item-based algorithm. User-based or item-based algorithm only considers the influence of users or item unilaterally, ignoring the connection between the user and the item. On the other hand, user-based or item-based algorithm considers the effect of users or items from all rating data, ignoring that users or items may have similar data in some circumstances. Doing research on these issues, some researchers had put forward the unifying user-based and item-based collaborative filtering [6]. The article computes the similarity of users by making use of users' co-rating data and calculates items' similarity by using users' co-rating data on items. The experimental results show that the proposed algorithm greatly improves the prediction accuracy and can get better recommendation quality.

Nowadays, there are many research fruits about tourism recommendation by utilizing the collaborative filtering. For example, Ahas R, such as Estonia makes a questionnaire survey of the tourism activities of tourists, and analyzes their individualized tourism destination information [7]; Fengmei Ma gets users' travel demands by means of online interview, and recommends scenic spots by calculating comprehensive interest degree [8]; Xian-fei wang inspires users' requirements by means of talks, and recommends attractions and hotels by constraining tourism knowledge base [9]; HAO Qiang recommends tourism destination to users by finding out users' information by means of topic modeling [10]; Yin H produces trip planning system through geo-tagged photos [11]. Although many existing researches can recommend travel information for users effectively, in certain tourist area explicit information is less on the web and model developers are unfamiliar with the tourist area, it will not role to play. The information outside the network will gradually become the main source of getting users' information, and the traditional grading recommendations will face a bigger challenge. Hence, how to effectively mine users' implicit preference information becomes an urgent problem in the field of the travel recommendation.

The main aim of this paper is to generate attractions of Jingdezhen based on the hybrid user-based and item-based collaborative filtering. The information outside the network that model developer had mastered is considered to form neighbor attractions. The tourist similarity takes into account of the social information. This paper focuses on the unifying collaborative filtering and its application in tourism of Jingdezhen. Since many scholars are trying to study how to improve the efficiency and accuracy of recommender algorithms, this paper will modify the existing model to meet the recommend of specific tourism area with very few network information.

2. Related Research

2.1 User-based Collaborative Filtering

The user-based algorithm finds several most similar neighbors by analyzing rating of different users. For the items current users have not yet rated, the algorithm makes the prediction according to users rating for it. This algorithm mainly concerns the connection of different users' interests, so it's called user-based collaborative filtering recommendation algorithm. The user-based algorithm was first proposed by Goldberg when building the email filtering system Tapeslry [12], and was later used by GroupLens to recommend news [13]. Herlocker J first proposed users' similarity parameters and neighbor users' selected threshold, and proved through the experiment the accuracy of recommendation were improved after these parameters' introduction [14]. Lee H C puts forward the method of using users' evaluation to predict users' rating of unevaluated items. The experimental results show that the improved algorithm obtains a higher recommendation quality [15].

2.2 Item-based Collaborative Filtering

Item-based algorithm was first put forward by Sarwar [16]. The algorithm looks for the most similar items to items the current users have not yet rated, then predicts the current users' rating by using their rating on similar items. Amazon website uses the algorithm to recommend commodities for users, achieving better real-time recommendation effects [17]. Deshpande M proves through experiments that item-based algorithm has higher recommendation quality than user-based algorithm [18].

2.3 Hybrid User-based and Item-based Collaborative Filtering

Xue G R proposes cluster-based collaborative filtering recommendation algorithm, using clustering technology to fill vacancy value and to select users' neighbors and items' neighbors [19]. Wang J establishes probability model to predict users' vacant rating data, and puts forward the algorithms of predictive value which combines user-based algorithm and item-based algorithm [20]. Ma H proposes a user-based and item-based weighting algorithm to fill vacancy value, and then predicts the collaborative filtering recommendation algorithm. Obviously, there is maybe impact the authentic prediction results utilizing fill the vacancy value before predicting. Through experiments, he proves its recommendation quality is better than other algorithms' [21]. These algorithms take into account of the connection between users and items rely on mining social information or data set, and are based on weight combining user-based algorithm and item-based algorithm, but it is hard to give a suitable weight without full understanding of the data set in advance. In this paper not only considers the situation in which users or items may have similar data, but also solves the problem that some algorithms are sensitive to the weight. Furthermore, the offline data is considered to calculate the weight in special tourism area due to the network data of attractions is very few.

3. The Unifying Model

The contribution in this paper is developed to recommend for the tourists for travelling in Jingdezhen. It will mainly consider the following two aspects: (1) the building of the unifying algorithm; (2) the application strategy of the hybrid algorithm on the attraction of Jingdezhen.

3.1 The Building of Unifying Algorithm

Relying on user-based collaborative filtering or item-based collaborative filtering only is undesirable, especially when the ratings from these two sources are quite often not available. Consequently, predictions are often made by averaging ratings from 'not-so-similar' users or items. Many scholars had pointed out that the accuracy of prediction can be improved by fusing user-based collaborative filtering and item-based collaborative filtering.

In general, many factors may influence the efficiency and accuracy of recommendation in some way according to relevant research. There are three main factors, which are similarity based on co-rating data, neighbor threshold and dynamic weighting of prediction. Similarity computation is the most critical step in collaborative filtering recommendation algorithm. Documents [14,15] show through experiments that Pearson correlation coefficient can better indicate users' or items' similarity degree. This paper also uses the Pearson correlation coefficient.

3.1.1. Similarity Based on Co-rating Data

When choosing users' co-rating data to compute users' similarity degree, such circumstance may appear: the number of items two users have commonly rated is very rare, but similarity is very high because their ratings are very close. Actually this circumstance may overestimate the similarity of users, because when the number of items the users are both interested in is small, their interests and hobbies are probably different. So, besides selecting co-rating data, the number of items users have commonly rated should also be considered. Only in this way can users' similarity degree in interests and hobbies be shown more accurately. So, some scholars use the number of items users have commonly rated to adjust users' similarity, said as the following formula 1.

$$Sim^{'}(a,u)=\frac{Min(k,\gamma)}{\gamma}\bullet Sim(a,u) \tag{1}$$

Among them, k is the number of items of user a and user u have commonly rated; γ is the preset parameter to adjust users' similarity. $Min(k,\gamma)$ is to take the smaller number between k and γ. If the number of items users have commonly rated is greater than or equal to γ, the similarity of users doesn't need to be adjusted. On the other hand, the similarity of users is adjusted according to the ratio of k and γ.

Similarly, items' similarity can be adjusted by using the number of users items have commonly rated. The formula shows below.

$$Sim^{'}(i,j)=\frac{Min(S,\delta)}{\delta}\bullet Sim(i,j) \tag{2}$$

Among them, S is the number of users of item i and item j have commonly rated; δ is the preset parameter to adjust items' similarity.

The paper won't explain the similarity formula $Sim(a,u)$ or $Sim(i,j)$ in detail because many scholars have described in the thesis.

3.1.2. Neighbor Threshold

Though the similarity of users as well as items after adjusting may be small, if such users or items are used as the similar neighbors to predict ratings, the accuracy of prediction may be reduced, which will affect the final recommendation quality. So this paper uses threshold θ and η to respectively restrict the selection of users' and items' similar neighbors. The formula 6 and 7 reflect respectively the relationship between threshold and similarity.

$$S(u) = \{a \mid Sim'(a,u) > \eta, a \neq u\} \tag{3}$$

$$S(i) = \{j \mid Sim'(i,j) > \theta, j \neq i\} \tag{4}$$

Among them, S(u) means the users set of similar neighbors, η is the selection threshold of similar neighbors, the value of η determines the number the users sets of similar neighbors. Only when the similarity to user u is greater than η, can other users be selected as similar neighbors. S(i) expresses the items set of similar neighbors, θ is the selection threshold of similar neighbors, the value of θ determines the number the items sets of similar neighbors. Only when the similarity to item j is greater than θ, can other items be selected as similar neighbors.

3.1.3. Dynamic Weighting of Prediction

To combine user-based algorithm and item-based algorithm, this paper uses weighting factor λ_i and λ_u, both of which are in the [0, 1] value. λ_u shows the dynamic impact of user-based algorithm on prediction results; λ_i shows the dynamic impact of item-based algorithm on prediction results. Choosing appropriate weighting factors automatically can combine dynamically advantages of user-based algorithm and item-based algorithm, making the prediction results achieve a good balance, making recommendation more accurate and more stable. If the more similar neighbors the user has, the stronger their similarity is, and the more influence the user- based algorithm has on the prediction results, then the greater the value of λ_u should be; On the other hand, the value of λ_i should be bigger. Because the users' or items' similarity reflects the characteristics of dynamic change, as described in formulas 8 and 9, this article sets λ_i and λ_u dynamically by using the ratio of sum of squares of users' similarity to sum of squares of items' similarity. Sum of squares is used to enlarge similarity's influence, making prediction results more inclined to be obtained by the party with stronger similarity, expressed in formula 5 to 9 as follows:

$$\lambda_i = \frac{\sum_{j \in S(i)} Sim'^2(i,j)}{\sum_{a \in S(u)} Sim'^2(a,u) + \sum_{j \in S(i)} Sim'^2(i,j)} \tag{5}$$

$$\lambda_u = \frac{\sum_{a \in S(u)} Sim'^2(a,u)}{\sum_{a \in S(u)} Sim'^2(a,u) + \sum_{j \in S(i)} Sim'^2(i,j)} \tag{6}$$

$$\lambda_u + \lambda_i = 1 \tag{7}$$

$$P(R_{u,i}) = \lambda_u \times P_{user}(R_{u,i}) + \lambda_i \times P_{item}(R_{u,i}) \tag{8}$$

3.1.4. Experimental process and Discussion of results

The data set used in the paper's experiment is provided by MovieLens website (http: / / movieLens. umn. edu) [21]. It is created and maintained by GroupLe Research Team of the University of Minnesota in the United States. This experiment in this paper uses MAE (Mean Absolute Error) as the standard of measuring an algorithm. MAE measures prediction accuracy by calculating the error between prediction scores and actual scores. MAE is the most commonly-used method in collaborative filtering recommendation algorithm. The smaller the value of MAE is, the more accurate the prediction is, and the higher quality the recommendation has.

$$MAE = \frac{\sum_{u,i} |R_{u,i} - \hat{R}_{u,i}|}{N} \tag{9}$$

Where N denotes the number of test ratings.

In this paper, 500 users are extracted from the data set to do this experiment. To increase the density of the rating data, the 500 users' ratings on the unevaluated items are filled 2 and are divided into two groups. In one group, 300 users are selected as the training set; in the other group, remaining 200 users become the testing set. At last, calculate the error between prediction score and actual score, using MAE to measure the accuracy of prediction. η and θ respectively represents similar neighbors selecting threshold of users and items, and these two parameters directly determine the number of similar neighbors collection. In this paper, to analyze the impact of η or θ, parameter $\delta = \gamma = 40$ is set, and then let η or θ change from 0.045 to 0.075, adding 0.005 each time, the performance of the algorithm is shown in figure 1. If 100 users are selected as the testing set, the value of other parameters still is same as the original, the result of predict is shown in figure2.

From the results of two figures, the MAE of is obviously lower than that of other value when the value of η and θ is 0.06 or 0.065, showing that the prediction score of the algorithm is closer to the actual score. η and θ are preset parameter to adjust the similarity of users or items, which makes its calculation more reasonable. The two parameters can be used to adjust the circumstance in which users' co-rating data are few but their similarity is very high.

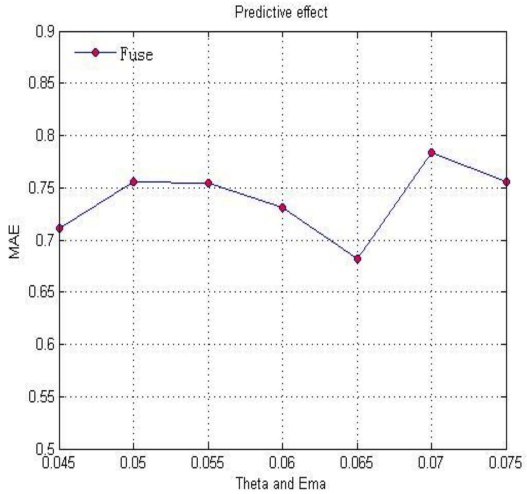

Fig1. The impact of similarity testing set=200

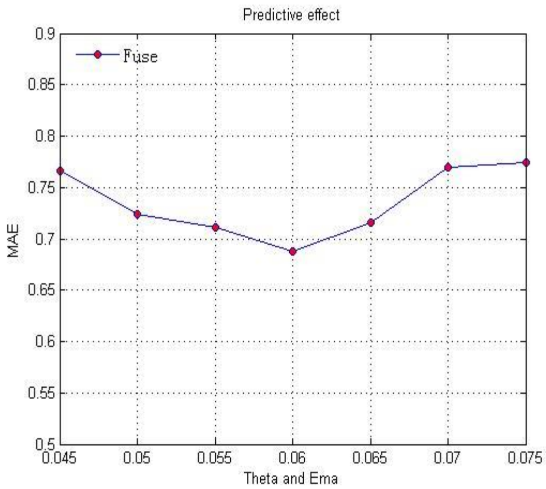

Fig2. The impact of similarity testing set=100

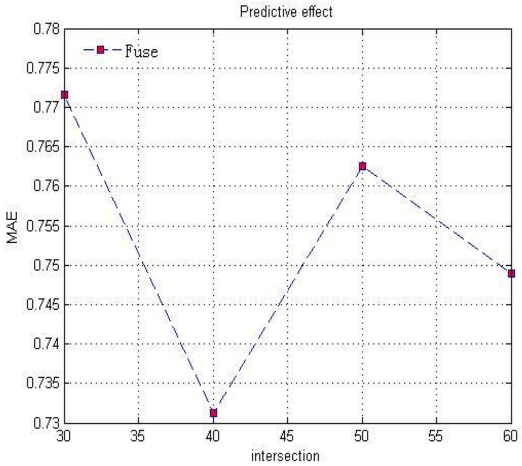

Fig3. The impact of intersection testing set=200

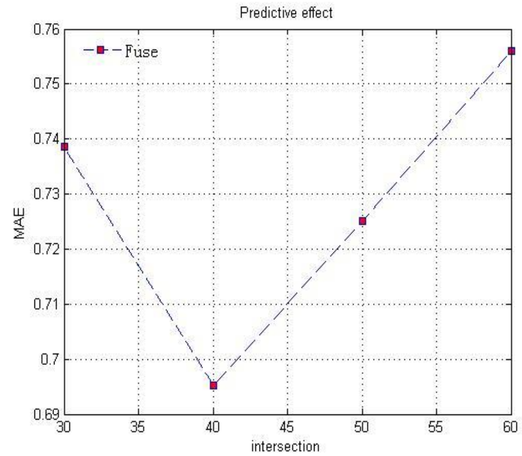

Fig4. The impact of intersection testing set=100

Figure 3 and figure 4 show the effect of δ and γ on prediction rate. Moreover, no matter what circumstance, the MAE is lowest when δ and γ is equal to 40. When δ and γ is more than 40, the MAE is increased rapidly. So, δ and γ being 40 can make the algorithm more effective in the prediction accuracy and range.

The experimental results On MovieLens dataset show that the fused method proposed in this paper can obtain better recommendation quality. Next, we will propose the recommend strategy on the attraction of Jingdezhen based on it.

3.2 The application Strategy

In addition to rating data, users and items have their own properties. How to use these properties to predict current user's interests and classify these items and improve the quality of the recommendation system is an important part of the future research. We divide the attractions into three categories according to offline knowledge including ancient books and the accumulation of previous research, because the specific tourism area of Jingdezhen with very few network information. The first class is shopping sites, such as Guomao square, Ceramics city and Antique market. The second class is kiln sites with natural landscape including Yaoli, Hutian, Ancient kiln, Kaolin Mountain and imperial kiln. The last class is ceramic art sites, such as Taoxichuan, Jingdezhen Ceramics Museum. In addition, we can select users' neighbor according to social information. The strategy will increase similarity precision compared to the method by presetting threshold. With the similarity increases, the result will more accurate because the similarity plays an important role in the unifying model.

4. Conclusion

This paper puts forward the application strategy based on the unifying collaborative filtering recommendation algorithm. The main contributions of this paper are: (1) The situation is considered in which users or items have similar data, and the similarity of users and items is calculated by making use of co-rating data. (2) The connection between users and items is considered. By using the ratio of sum of squares of users' similarity to sum of squares of items' similarity as the weight, and by combining dynamically the user-based and item-based Collaborative Filtering Recommendation Algorithm, the problem that data are sensitive to fixed weight is solved, making the combination of the two algorithms achieve a good balance. (3) We use the offline knowledge to classify the tourist attractions of jingdezhen. How to formulaic represent the similarity by the classified attractions is also an important research in the future research.

Acknowledgments

This research was supported by the Science and Technology Research Foundation of Education Bureau of Jiangxi Province, China (No. GJJ150922).

References

[1] Ravi, L. and Vairavasundaram, S., 2016. A Collaborative Location Based Travel Recommendation System through Enhanced Rating Prediction for the Group of Users. Computational intelligence and neuroscience, 2016, p.7.

[2] Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P. and Riedl, J., 1994, October. GroupLens: an open architecture for collaborative filtering of netnews. In Proceedings of the 1994 ACM conference on Computer supported cooperative work (pp. 175-186). ACM.

[3] Shardanand, U. and Maes, P., 1995, May. Social information filtering: algorithms for automating "word of mouth". In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 210-217). ACM Press/Addison-Wesley Publishing Co.

[4] Sarwar, B., Karypis, G., Konstan, J. and Riedl, J., 2001, April. Item-based collaborative filtering recommendation algorithms. In Proceedings of the 10th international conference on World Wide Web (pp. 285-295). ACM.

[5] Xue, G.R., Lin, C., Yang, Q., Xi, W., Zeng, H.J., Yu, Y. and Chen, Z., 2005, August. Scalable collaborative filtering using cluster-based smoothing. In Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval (pp. 114-121). ACM.

[6] Wang J, Yin J, Zheng L R, et al. Collaborative filtering recommendation algorithm based on co-ratings and similarity weight [J]. Computer Science, 2010, 37(2): 99-104.

[7] Ahas R, Aasa A, Roose A, et al. Evaluating passive mobile positioning data for tourism surveys: An Estonian case study [J]. Tourism Management, 2008, 29(3): 469-486.

[8] Fangmei Ma. The Recommendation of Travel Destinations Based on Comprehensive Interest Measure [J]. Geomatics ＆ Spatial Information Technology, 2014, 37(3): 55-56.

[9] Xianfei Wang, Mei Chen, Xiaotian Li. Research and Design of Constraint-Based Travel Recommendation System [J]. Computer Technology and Development, 2012, 22(2): 141-145.

[10] Hao Q, Cai R, Wang C, et al. Equip tourists with knowledge mined from travelogues[C]//Proceedings of the 19th international conference on World wide web. ACM, 2010: 401-410.

[11] Yin H, Lu X, Wang C, et al. Photo2trip: an interactive trip planning system based on geo-tagged photos[C]//Proceedings of the 18th ACM international conference on Multimedia. ACM, 2010: 1579-1582.

[12] Goldberg D, Nichols D, Oki B M, et al. Using collaborative filtering to weave an information tapestry [J]. Communications of the ACM, 1992, 35(12): 61-70.

[13] Resnick P, Iacovou N, Suchak M, et al. GroupLens: an open architecture for collaborative filtering of netnews[C]//Proceedings of the 1994 ACM conference on Computer supported cooperative work. ACM, 1994: 175-186.
[14] Herlocker J, Konstan J A, Riedl J. An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms[J]. Information retrieval, 2002, 5(4): 287-310.
[15] Lee H C, Lee S J, Chung Y J. A study on the improved collaborative filtering algorithm for recommender system[C]//Software Engineering Research, Management & Applications, 2007. SERA 2007. 5th ACIS International Conference on. IEEE, 2007: 297-304.
[16] Sarwar B, Karypis G, Konstan J, et al. Item-based collaborative filtering recommendation algorithms[C]//Proceedings of the 10th international conference on World Wide Web. ACM, 2001: 285-295.
[17] Linden G, Smith B, York J. Amazon. com recommendations: Item-to-item collaborative filtering[J]. IEEE Internet computing, 2003, 7(1): 76-80.
[18] Deshpande M, Karypis G. Item-based top-n recommendation algorithms[J]. ACM Transactions on Information Systems (TOIS), 2004, 22(1): 143-177.
[19] Xue G R, Lin C, Yang Q, et al. Scalable collaborative filtering using cluster-based smoothing[C]//Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2005: 114-121.
[20] Wang J, De Vries A P, Reinders M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006: 501-508.
[21] Ma H, King I, Lyu M R. Effective missing data prediction for collaborative filtering[C]//Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007: 39-46.
[22] Wang J, De Vries A P, Reinders M J T. Unifying user-based and item-based collaborative filtering approaches by similarity fusion[C]//Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2006: 501-508.