MATEC Web of Conferences 22, 01013 (2015)
DOI: 10.1051/matecconf/20152201013
© Owned by the authors, published by EDP Sciences, 2015

Algorithm Research of Individualized Travelling Route Recommendation Based on Similarity

Shan Xue
Qingdao Vocational and Technical College of Hotel Management, Qingdao, Shandong, China

Song Liu
Qingdao Technological University, Qingdao, Shandong, China

ABSTRACT: Although commercial recommendation system has made certain achievement in travelling route development, the recommendation system is facing a series of challenges because of people's increasing interest in travelling. It is obvious that the core content of the recommendation system is recommendation algorithm. The advantages of recommendation algorithm can bring great effect to the recommendation system. Based on this, this paper applies traditional collaborative filtering algorithm for analysis. Besides, illustrating the deficiencies of the algorithm, such as the rating unicity and rating matrix sparsity, this paper proposes an improved algorithm combing the multi-similarity algorithm based on user and the element similarity algorithm based on user, so as to compensate for the deficiencies that traditional algorithm has within a controllable range. Experimental results have shown that the improved algorithm has obvious advantages in comparison with the traditional one. The improved algorithm has obvious effect on remedying the rating matrix sparsity and rating unicity.

Keywords: collaborative filtering; multi-similarity; element similarity; trust degree; fraction of coverage; accumulated gain in normalization depreciation

1 INTRODUCTION

Tourist industry is the pillar industry of the tertiary industry. In contemporary society, tourist industry is experiencing a higher accelerated speed in development. Hongliang Lv et al. (2012) once pointed that the contribution that tourism made to GDP rose from 9.1% in 2011 to 9.6% in 2012. Meanwhile, with the development of internet, more and more tourism enterprises start to post their service information online [1]. The explosive growth of online tourism information has created a big challenge for users to choose appropriate tourism service. Jianguo Liu et al. (2009) called this challenge as an information overload problem in tourism field [2]. An important method to solve information overload is the application of individualized recommendation technology [3]. This technology has been successfully used in e-commerce [4], movie [5] and music [6]. In order to provide recommended routes for individualized tourism and design high-quality travelling routes for all tourism customers, this paper proposes a collaborative filtering system based on similarity algorithm.

Among the existing recommendation algorithms, collaborative filtering recommendation algorithm is a sophisticated one and has been widely applied in practical use. Yifan Shi et al. (2014) have pointed that collaborative filtering algorithm is the most basic algorithm of recommendation system. It can be divided into collaborative filtering recommendation based on items and collaborative filtering recommendation based on users [7]. Xu Pengyuan et al. (2011) have proposed that the core concept of collaborative filtering recommendation algorithm is to calculate the similarity between the target item and other items or the similarity between the target user and other users according to user's rating of the target item; and then calculate the target user's rating of the target item based on the resulting similarity [8]. There are certain differences existing between the tourism recommendation system and traditional recommendation systems. First, compared with the data of traditional fields, travelling data are sparser. Second, traditional recommendations will generally recommend a single item [3] whereas travelling routes always include more than one scenic spots. Last, traditional recommendation systems basically reply on users' ratings. However, in the data of tourism system, only implicit feedback data such as browsing and reservation need to be normalized for use. In order to further improve the recommendation quality of the collaborative filtering recommendation algorithm based on user, scholars have proposed some new ways to improve the algorithm and solve the existing deficiencies in it, such as new customer problem, new item problem [9], data sparsity [9], and data expandability [10]. For example, Sarwar B. et al (2001) have proposed to implement relatively stable observation on the similarity among items. They have recommended item-based collaborative filtering recommendation algorithm to improve the expandability of the system [10]. Breese J, et al(1998)have suggested to solve data sparsity problem by default voting and back scheduling of user frequency [11]. In the study of the individualized recommendation for travelling route, Lingwei Zeng (2013) has pointed that the travelling route recommendation model of Apriori algorithm based on association rules can be applied [12]. On the basis of traditional collaborative filtering algorithm research, this paper combines the multi-similarity algorithm based on user and the element similarity algorithm based on user, so as to improve the traditional algorithm; make up the deficiencies in the individualized travelling route recommendation generated by similarity algorithm; and make certain contribution to the tourist industry of our country.

2 COLLABORATIVE FILTERING RECOMMENDATION ALGORITHM BASED ON IMPROVED USER'S SIMILARITY

The traditional collaborative filtering recommendation algorithm based on user can be divided into three phases shown as below:

1) Calculate the similarity among users' ratings according to their rating record history.

2) Select several users from the current ones sharing the highest similarity as the nearest neighbors. Forecast the current users' ratings on one item based on the nearest neighbors' actual ratings on the item.

3) Select several items with the highest forecast rating as the recommendation results for the current users.

It can be concluded from the three phases mentioned above that the traditional algorithm only relies on single rating similarity to forecast user's rating on a to-be-evaluated item without any consideration about item type. As a result, this similarity can only reflect the similarity between the historical rating records of two users. It cannot reflect the similarity between the users' preference on one item. Besides, in the traditional algorithm, there's serious sparsity problem in the user-product rating matrix which makes it impossible to calculate the similarity between two users who do not give ratings on the same product. Based on the two reasons mentioned earlier, this chapter proposes to apply the collaborative filtering recommendation algorithm based on user-element similarity and multi-similarity with analysis of the traditional collaborative filtering algorithm, so as to lay a theoretical foundation for individualized travelling route.

2.1 *Traditional Collaborative Filtering Algorithm*

In order to calculate user's forecast ratings on a to-be-evaluated item, the similarity between users' historical rating records should be calculated at first. This similarity can reflect the similarity between users' preference on each item. Apply user information, item information and users' rating information on items to build the rating matrix R as shown in formula (1) of which p refers to the number of users; q refers to the number of items; r_{ij} refers to the rating that user i has on item j; and No. i row of the matrix refers to the rating vector quantity of user i. If the user hasn't given any rating to the item, the rating value should be assigned as 0.

$$R_{p,q}=\begin{bmatrix} r_{11} & r_{12} & \cdots & r_{1q} \\ r_{21} & r_{22} & \cdots & r_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ r_{p1} & r_{p2} & \cdots & r_{pq} \end{bmatrix} \tag{1}$$

During the calculation of relative similarity, the Pearson correlation coefficient [13] can be calculated in accordance with all items that any two users have given ratings to, and the coefficient can be regarded as the similarity degree between those two users. If i and j are used to represent the two users, use $I_{i,j}$ to express the item collection that has been given ratings to by the two users. Besides, p refers to the number of users; q refers to the number of items; $r_{i,\alpha}, r_{j,\alpha}$ respectively refer to the ratings that user i and user j have given to item α ; and $\bar{r}_i, \bar{r}_j$ respectively refer to the average rating score that user i and user j have given to all items. Therefore, the calculation for the relative similarity between user i and user j can be completed by quotation (2) as shown below:

$$sim(i,j)=\frac{\sum\limits_{\alpha\in I_{i,j}}\left(r_{i,\alpha}-\bar{r}_i\right)\left(r_{j,\alpha}-\bar{r}_j\right)}{\sqrt{\sum\limits_{\alpha\in I_{i,j}}\left(r_{i,\alpha}-\bar{r}_i\right)^2}\cdot\sqrt{\sum\limits_{\alpha\in I_{i,j}}\left(r_{j,\alpha}-\bar{r}_j\right)^2}} \tag{2}$$

The calculation of forecast rating is based on the actual ratings of the users who are similar to the current users on the item. Forecast rating can manifest the current users' preference on the item. The forecast rating is the main basis for most recommendation systems to generate recommendation results. In the recommendation algorithms which are mostly based on similarity, forecast is accomplished by completing weighted average process on the voting values generated according to similarity. See quotation (3) for the calculation method:

$$R_{i,\alpha}=\frac{\sum\limits_{\beta} sim(\alpha,\beta)\cdot r_{i,\beta}}{\sum\limits_{\beta}\left|sim(\alpha,\beta)\right|} \tag{3}$$

It can be seen from formula (2) and formula (3) that when similarity sim (i,j) approaches 1, the similarity between two users is higher; on the contrary, when similarity sim (i, j) is much bigger or smaller than 1, there's distinct difference between the two users' interest. When the values of $sim(\alpha,\beta)$ and $r_{i,\beta}$ turn higher, the evaluating value will become higher correspondingly. Therefore, when any user sharing great similarity with the user is fond of any item, the forecast value of the user's preference on the item will be higher; or will turn lower in the contrary situation.

2.2 *Improved Algorithm*

People tend to have different degrees of preference on various item types in real life. In order to make up the unicity, the deficiency of collaborative filtering algorithm should be considered and the first step for improvement can be accomplished on the basis of user multi-similarity theory.

The multi-similarity degree between two users refers to the several independent rating similarity degrees in various item types between the users. The similarity degree of one item type between two users is based on the calculation of the users' rating records in item type. As a result, the similarity degree can reflect the similarity degree between two users' preference on item type more accurately. By calculating users' independent rating similarity degrees in items of various types, the differences between users' interests in different item types can be more accurately manifested through the similarity degrees, and thus more precise forecast ratings can be obtained accordingly. Set a collection $T=\{t_1,t_2,\mathrm{L}\ ,t_q\}$ to represent all types, and each item belongs to one type or several types of the collection T. Users' similarity degrees based on all q item types need to be calculated in order to describe users' different degrees of preference on those q item types. The q similarity degrees of any two users—user i and user j—in the collection T can be expressed as

$$sim(i,j,t_1),sim(i,j,t_2),\mathrm{L}\ ,sim(i,j,t_q)\ .$$

Set $I_{i,j}^{t_k}$ as the item collection for the items which belong to type t_k $t_k\in T,1\le k\le q$ and have been rated by both user i and user j. Use $\overline{r_i^{t_k}}$, $\overline{r_j^{t_k}}$ to respectively represent the average ratings that user i and user j give to the items belonging to type t_k . User i and user j's relative similarity degrees in type t_k can be obtained through formula (2). See formula (4) for related calculation formula:

$$sim(i,j,t_k)=\frac{\sum\limits_{\alpha\in I_{i,j}^{t_k}}\left(r_{i,\alpha}-\overline{r_i^{t_k}}\right)\left(r_{j,\alpha}-\overline{r_j^{t_k}}\right)}{\sqrt{\sum\limits_{\alpha\in I_{i,j}^{t_k}}\left(r_{i,\alpha}-\overline{r_i^{t_k}}\right)^2}\cdot\sqrt{\sum\limits_{\alpha\in I_{i,j}^{t_k}}\left(r_{j,\alpha}-\overline{r_j^{t_k}}\right)^2}} \tag{4}$$

In the traditional collaborative filtering algorithm, there's serious problem of sparsity existing in the user-item rating matrix which makes it impossible to calculate the similarity degree between two users who haven't given any rating to the same product. In order to solve this problem, this paper introduces the concept of calculating the indirect element similarity and the overall element similarity in similarity relations based on the first step of improvement as the second step of improvement. The second step of improvement can be divided into three phases. The first phase is to calculate the original similarity $Isim(i,j,t_k)$ according to formula (4). The second phase is to calculate the degree of element similarity according to formula (5). And the third phase is to make weighting integration of $Isim(i,j,t_k)$ that represents local/direct similarity relations and $metasim(i,j,t_k)$ that represents indirect/overall similarity relations.

$$metasim(i,j,t_k)=\frac{\sum\limits_{\alpha\in I_{i,j}^{t_k}} sim'_{i,\alpha}\cdot sim'_{j,\alpha}}{\sqrt{\sum\limits_{\alpha\in I_{i,j}^{t_k}}\left(sim'_{i,\alpha}\right)^2}\cdot\sqrt{\sum\limits_{\alpha\in I_{i,j}^{t_k}}\left(sim'_{j,\alpha}\right)^2}} \tag{5}$$

The $Isim_i$ of $sim'_{i,\alpha}=Isim_{i,\alpha}-\langle Isim_i\rangle$ in formula (5) refers to the average value of the similarity degree between user i and other users. This formula uses the longest spreading path of the constructed social network and the shortest path between users. The trust degree between users can be calculated by formula (6).

$$\begin{cases} trust(i,j)=\dfrac{d_{\max}-n+1}{d_{\max}} \\ d_{\max}=\left\lceil\dfrac{\ln(m)}{\ln(k)}\right\rceil \end{cases} \tag{6}$$

The m in formula (6) refers to the total number of social network users while k refers to the average degree of the users. If user i cannot reach user j, it means user j has no trust value to user i. However, if user i can reach user j within the longest spreading distance, it means that user j has certain trust degree to user i.

The **Isim** in formula (7) is the original matrix of similarity degree. **Metasim** refers to the matrix of element similarity degree and λ refers to weighting coefficient.

$$\mathbf{newsim}=\lambda\mathbf{Isim}+(1-\lambda)\mathbf{metasim} \tag{7}$$

After calculating the similarity degree, user's rating value $R'_{i,\alpha}$ on the item can be obtained by formula (8):

$$R'_{i,\alpha}=\frac{\sum\limits_j newsim(i,j,t_k)\cdot r_{j,\alpha}}{\sum\limits_j\left|newsim(i,j,t_k)\right|} \tag{8}$$

The $metasim(i,j,t_k)$ in formula (8) represents the element of matrix **newsim** defined in formula (7). When $\lambda=0$, it means only the situation of direct and original similarities should be considered which can also regarded as the algorithm stopped in the first phase of algorithm improvement. When $\lambda=1$, it means only the situation of the overall and indirect similarities should be considered. When $0<\lambda<1$, it means only the situation of the multi-similarity existing in the overall information and local information should be considered. So far, the collaborative filtering algorithm of user's similarity improved on basis of the foundation proposed in this paper has been completely elaborated. This algorithm can not only avoid user's single rating on one item, but can also solve the serious sparsity problem existing in user's item-rating matrix.

3 DESIGN OF EXPERIMENTAL SCHEME

3.1 *Tripartite Graph of Travelling Route*

The objective of offering recommendation to users of individualized traveling route recommendation is to point out the optimal scenic spot traveling routes. Therefore, users can participate in the travelling routes consisted of scenic spots. In order to indicate the relations among users, travelling routes and scenic spots, this paper has designed the user-scenic spot-route tripartite graph. In their process of selecting travelling routes, users take the internet as the platform. The web pages can form a circle through super interlinking as shown in Figure 1:

In user-tourism relations, there're three different elements which are users, traveling routes and scenic spots. This paper defines three top points in Figure 2: top point of user, top point of travelling route and top point of scenic spot. If any known user (User 1) visits the travelling route (Route 3), and the scenic spot Spot2 is included in Route 3, sides should be drawn between User 1 and Route 3, Route 3 and Spot 2, and User 1 and Spot 2. If there are sides already drawn between the two top points, the weights of the sides should be added.

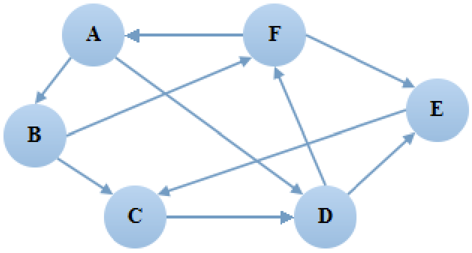

Figure 1. Web Page Linking Graph

After constructing the tripartite graph of user-scenic spot-travelling route according to the relations among users' behaviors, travelling routes and scenic spots, the problem of recommending travelling routes to user U_i becomes the relevancy sequencing problem of corresponding all item nodes in the computer graph and users' nodes. The relevancy is the similarity degree studied in this paper.

3.2 *Experimental Data Set*

This paper applies crawling technology to obtain the web page date of some tourism service website [14]. The required experimental data can be obtained from content extraction. There're five collected tourism spots: Guilin, Shanghai, Xiamen, Hangzhou and Hong Kong. See Table 1 for the basic situation of the obtained data:

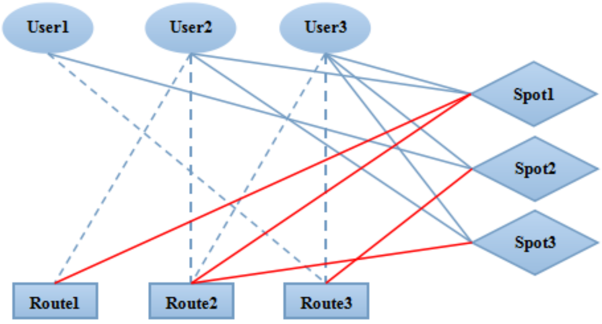

Figure 2. Tripartite Graph of User-Scenic Spot-Travelling Route

Table 1. List of the Basic Situation of the Obtained Experimental Data

Number of Scenic Spots	Rating Users	Rating Records	Type Tags	Social Network Number of Users	Number of Sides	Number of Sides Formed among Rating Users
998	9202	22502	132	15375	26931	10471

Table 2. List of the Basic Situation of the Data after Pre-treatment

Number of Scenic Spots	Rating Users	Rating Records	Degree of Data Sparsity	Social Network	
				Number of Users	Number of Sides
998	1015	12339	1.23%	15375	26931

In the obtained tag information of the scenic spots, each record represents the tag information of each scenic spot, among which each element is divided by "#". For example, "A#B#C#D" means that the tag of A includes B, C and D.

For the convenience of conducting the experiment, pre-treatment needs to be made for the data. The processing steps are as follows:

STEP1. Extract and integrate the data required in this experiment from the original data.

STEP2. Delete repeated rating records and select the recent ratings of one user on the scenic spot as the ratings on the scenic spot of that user.

STEP3. Delete the users who only have one or two rating records.

STEP4. Delete the scenic spots with no user rating.

After data pre-treatment, the basic situation of the final experimental data is shown in Table 2:

3.3 *Evaluation Metrics and Experimental Design*

This paper applies the coverage [15] and normalized discounted cumulative gain (NDCG) [16] to measure the accuracy of recommendations.

The coverage can calculate the proportion of the forecast rating sample quantity in the overall sample quantity. In this paper, if a location rating cannot be forecasted, it means the forecast rating cannot be 0. If a location cannot be forecasted, the forecast rating should be set as 0. If the number of the samples of which the forecast ratings are not 0 is k, the number of the total samples is M, the calculating formula of coverage can be set as formula (9) shown below:

$$coverage=\frac{k}{M}\times 100\% \qquad (9)$$

NDCG is mainly used in information indexing domain to measure the quality of the sequencing results. NDCG is based on an assumption which is a travelling route and will have a higher rank in the recommendation list if it can obtain more attention. NDCG only evaluates the top k routes in the travelling route recommendation list. The calculation of the NDCG value in the recommendation list with a quantity of k can be completed as follows:

$$\begin{cases} NDCG@k=\dfrac{RL@k}{IRL@k} \\ RL@k=\left(utility\left(P_1\right)+\sum\limits_{i=2}^{k}\dfrac{utility\left(P_i\right)}{\log_2(i)} \right) \end{cases} \qquad (10)$$

The RL in formula (10) is the sorted list offered by the system. However, IRL is the ideal sorted list which is completely arranged by the users' actual preference. P_i refers to the No. i travelling route in the recommendation list while $utility\left(P_i\right)$ refers to the effect that route P_i offers to the given user. This paper applies normalized interactive information between users and routes for measurement as shown in formula (11) which means the K times of web browsing done by the given user equals to the highest rating score in one time. Take K as 10 in this experiment.

$$utility\left(P_i\right)=N_i^{book}+\frac{N_i^{view}}{K} \qquad (11)$$

The value range of NDCG is [0, 1]. The higher the value is, the more accurate the sequencing forecast is.

This paper uses contrast experiment to verify the advantages of the individualized travelling route recommendation algorithm based on improved similarity. It applies the leap-one-out cross validation [17] used in most social recommendation system researches to conduct experimental verification on algorithm. Besides, it completes cyclic utilization on one datum of the whole data set as the test set and takes the other data as the training set. The forecast result of each cycling integration process can be used to calculate the evaluation metrics: the coverage and NDCG. In the meanwhile, in order to verify the effectiveness of the method, this paper makes comparison between the experimental results of the C1-tradtional similarity algorithm, C2-element similarity algorithm and C3-multi-user similarity algorithm used in the algorithm improvement process and the evaluation metrics obtained form the improved algorithm. According to the social network information collected in experimental data, the longest spreading path length obtained through formula (6) is 18.

4 ANALYSIS OF EXPERIMENTAL RESULTS

See Table 3 for the data obtained from the experimental results:

Trends of the four individualized recommendation

Table 3. Results of the Four Algorithms on NDCG

Algorithm Type		Evaluation Indicator	
Name of Algorithm	Symbol	Coverage	NDCG
Traditional Algorithm	C1	78.32%	56.97%
Algorithm Based on User Multi-similarity	C2	78.56%	75.31%
Algorithm Based on User-element Similarity	C3	92.46%	76.67%
Improved Algorithm	C	93.92%	81.52%

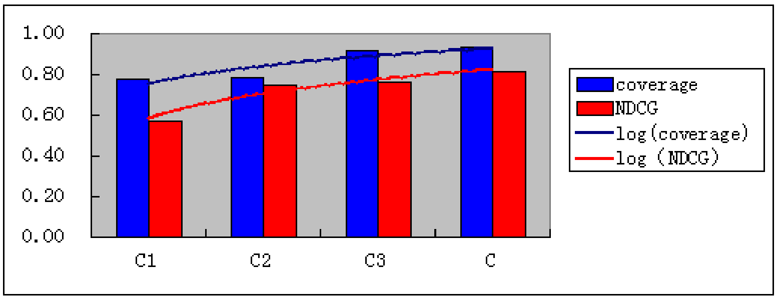

Figure 3. Trends of the Four Individualized Recommendation Algorithms on the Evaluation Indicators

algorithms on the evaluation indicators shown as in Figure 3:

From Figure 3 and Table 3, it can be seen that in the traditional algorithm, both evaluating indicators show their minimum values; while in the improved algorithm, the same indicators show their maximum values. The NDCG values of the algorithm based on user multi-similarity and the algorithm based on user-element similarity share some similar features; however, the value of the latter one is higher than that of the former one. It shows that these two algorithms have certain superiority over the traditional algorithm on NDCG. In the case of Coverage, the traditional algorithm and the algorithm based on user multi-similarity have similar features. The author thinks that the reason for the generation of the similarity is because of the rating matrix sparsity. The algorithm based on element similarity can make up this deficiency and bring a higher value in the evaluating indicator of Coverage. The improved algorithm combines this advantage of the element similarity algorithm and it is the reason for why there's similarity between the improved algorithm and the algorithm based on element similarity in Coverage.

To sum up, after integrating the advantages of the algorithm based on user multi-similarity and the algorithm based on user-element similarity, the improved algorithm has the best evaluation results. It has obvious advantages in comparison with the traditional algorithm, and thus can serve for the individualized travelling route recommendation in a better way.

5 CONCLUSION

The optimization of individualized travelling route recommendation system can help improve user's travelling enthusiasm. To some extent, the traditional collaborative filtering algorithm can no longer satisfy user's demands. This paper has made improvement on the traditional recommendation algorithm based on the analysis of the unicity and rating matrix sparsity in the traditional algorithm. From related experiment, it has been verified that the improved algorithm has obvious advantages in comparison with the traditional one. In conclusion, the improved algorithm can serve for the individualized travelling route recommendation system in a more efficient way.

REFERENCES

[1] Lv H.L., Wang J.L. & Deng F, 2012. A Recommendation Algorithm for Individualized Travelling Route. *Network New Media Technology*. 1(3): 42-48.

[2] Liu J.G., Zhou T. & Wang B.H. 2009. Development Progress of Individualized Recommendation System. *Natural Science Progress*. 19(1): 1-15.

[3] Adomavicius G. & Tuzhilin. A. 2005. Toward the next generation of recommender systems: A survey of the state-of-the art and possible extensions. *IEEE transactions on knowledge and data engineering*. pp: 734-749.

[4] Linden G, Smith B. & York J. 2003. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet computing*. 7(1): 76-80.

[5] Said A, Berkovsky S. & De Luca E W. 2010. Putting things in context: Challenge on context-aware movie recommendation. *ACM*.

[6] Su J H, Yeh H H. & Yu P S, et al. 2010. Music recommendation using content and context information mining. Intelligent Systems. *IEEE*. 25(1): 16-26.

[7] Shi Y.F., Wen Y.M., Cai G.Y. & Miu Y.Q. 2014. Collaborative Filtering Recommendation Based on Scenic Spot Label. Computer Application. 34(10): 2854-2858.

[8] Xu P.Y. & Dang Y.Z. 2011. Recommendation Algorithm Based on Element Similarity. *Computer Application Research*. 28(10): 3646-3659.

[9] MASSA P. & AVESANI P. 2007. Trust-aware recommender systems. Proceedings of the 2007 ACM Conference on Recommender Systems. New York: *ACM Press*.

[10] Sarwar B, Karypis G. & Konstan J, et al. 2001. Item-Based Collaborative Filtering Recommendation Algorithms. *Proceedings of the 10th International World Wide Web Conference*. New York.

[11] Breese J, Hecherman D. & Kadie C. 1998. Empirical Analysis of Forecast Algorithms for Collaborative Filtering. *Proceedings of the 14th Conference on Uncertainty in Artifical Intelligence* (UAI 98).

[12] Zeng L.W., Wang D. & Wu J. 2013. Analysis of Travelling Route Recommendation Model Practice Based on Apiori Algorithm. *Computer Knowledge and Technology*. pp: 1906-1908.

[13] Shardanand U. & Maes P. 1995. Social Information Filtering: Algorithms for Automating ‘World of Mouth’. Proceeding of the Conference on Human Factors in Computing Systems.

[14] Ctrip. Ctrip.com [EB/OL]. [2014-02-12]. http://www.ctrip.com.

[15] VICTOR P, COCK M D. & CORNELLS C. 2011. Trust and recommendations. // KANTOR P B, ROKACH L, RICCI F, et al. Recommender Systems Handbook. *Berlin: Springer.*

[16] Jarvelin K. & Kekalainen J. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* (TOIS). 20(4): 422-446.

[17] JAMALI M. & ESTER M. 2009. Trustwalker: a random walk model for combining trust-based and item-based recommendation.// Proceedings of 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: *ACM Press*.