Study on Tourism Recommendation Model Based on Neogeneity and Diversity Identification

Xiaoxue Zhao [1, a], Yucheng Zhang [1,b] Zhangli Zhou[1,c] and Lili Wang[1,d]

[1]Xijing University, Xi'an 710123 , China

[a]55100246@qq.com,[b]58911533@qq.com,[c]124649226@qq.com,[d]1736656241 @qq.com

Keyword: Recommendation system; Novelty; Diversity

Abstract. With the development of Internet technology, people often choose to make travel plans through the recommendation system of travel websites, but many travel recommendation systems can not meet the real needs of tourists. Based on the novelty and diversity of tourism recommendation system, this paper introduces two common recommendation and calculation methods of diversified algorithms, and then compares the diversity and novelty of the two methods through the tourism data in Xi'an and carries out experimental simulation .

The Necessity of Using Tourism Route Recommendation System

Accurately Recommend the Needs of Traveling Consumers. If list out the tourist attractions that visitors are interested in, then you will find the list of attractions recommended by the system accounts for about 20% of popular attractions. Such recommendations do not have a personalized feature, which is to recommend some popular attractions to tourists. So it's important to study a recommendation system that can provide visitors with scenic spots that they're interested in but difficult to find. In the field of travel recommendation system, the direction of attention is for one tourist. How to make the recommended scenic spots more personalized, diversified and novelty, how to improve the satisfaction of users, and the precise recommendation of tourists is what we need to solve.

Big Data Era to Promote the Development of Tourism Market. With the development of the Internet and big data technologies, people no longer only know about scenic spots through travel agencies, but more information on scenic spots and purchases of tourist products through the Internet. As can be seen, big data has had a positive impact on the development of the travel market. Therefore, some scholars at home and abroad have started to study how to forecast tourists' needs by using tourism-related big data generated by tourists on the Internet[1-2].

Common Diversification Algorithm Recommendations and Calculation Methods

Collaborative filtering algorithm is a commonly used personalized recommendation technology at present stage[3-12], following introduces two commonly used collaborative filtering algorithms.

Collaborative Filtering Algorithm Based on the Influence of User Behavior. Influence is the extent to which one user affects another user's intention to travel.

Original user service interaction matrix

First define a matrix I_{mn} , the rows of this matrix represent services $S\{s_1, s_2, \cdots, s_n\}$, columns represent users $U\{u_1, u_2, \cdots u_m\}$, and the size of this matrix represents the number of services used by the user D_{mn} .

Assuming the user (X,Y) and the service (a,b,c,d,e,f) are regarded as different nodes and the behavior between the user and the service is regarded as the edge, a multi-node infinite network can be constructed, and the weight of the edge is the interaction between the user and the service.
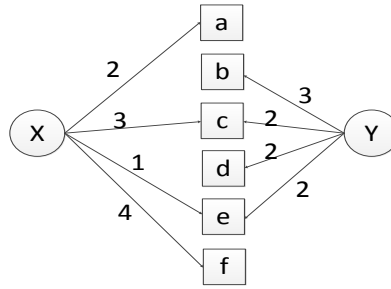
Figure 1. User - service original interaction diagram

For users, they are more interested in choosing more services. For example, user X and user Y each have an interaction with service a、c、e、f and service b、c、d、e, then user X's interest in service f is large compared e to. Therefore, when considering the user's interaction behavior, we must consider the impact of different user interaction on the number of services.

User influence

Normally, in a complex network, there will be a power law distribution and small world features, so no matter which two users are able to use a certain number of unicom, the number of steps to form a unicom path. The path between users can be divided into two types: normal path and short path. In the entire path, the proportion of short path affects the size of users' mutual influence. The larger the short path, the greater the impact on each other between users, the proportion of short path is measured by network medium.

Through the above figure, we can draw the user X and user Y have the same service c、e. So the bridge established by these two services can communicate user x and user y only in two steps, and it can be concluded that the roles of both sides of c.e will be greater than those of other sides. Here, the same edges that exist between users are considered as a short path, the other path as a normal path. Above the short path, the number and weight of its stripes have a direct impact on other users, so the short path is more important than the normal path, although the weights on the edges are different in size. Therefore, the degree of importance of the short path can be measured by the mediation of the complex network, which can be said to be the impact on other users. The short path ratio between user and service is user-service edgemedium. The greater the number of edgemediums, the greater the importance of the user for the selection of some special services, and also Can also be considered the more obvious impact on the target user.

$$C_{xy} = \sum_{X,K=(I_X \cap I_Y)} D_{xk} / \sum_{j} D_{xj}$$

(1)

In this formula, I_x is the set of services selected by user X, I_y is the set of services selected by the user Y, $I_x \cap I_y$ is the duplicate of the set of services used by both users. In this way, we can know the shortest path of the molecule in the formula that connects the two users at the same time. j is a service used by user X, the denominator in the formula is the size of the degree of connection that user X in the service space on the edge.

C_{XY} is the size of user X's impact on user Y, the same reason can calculate C_{YX}, and the two values are not the same.

Score forecast

Through the influence among the users, the following formula is used to calculate the prediction score p_{yj} of the user y for the specific item j.

$$p_{yj} = \sum_{x \in N_y} C_{xy} * A_{xj}$$

(2)

In this formula, A_{xj} is the average rating of user x for item j, and N_y is the set of neighbors for which the number of targeted y is k.

User Behavior Similarity Measure. An important part of the collaborative filtering algorithm is the similarity between users. The most basic way to calculate the similarity between users is cosine similarity.

In the cosine similarity method, all users' scores for each item are regarded as multi-dimensional vectors. If an item does not exist, the default is 0, and the similarity between users can be equivalent to the corresponding vector Cosine angle:

$$sim(u',v')^{\cos} = \cos(u,v) = \frac{u \cdot v}{\|u\| \times \|v\|}$$

(3)

In this formula, u' u and v' represent two users; the vector u,v represents the vectors of these two users in the multidimensional project space.

Collaborative Filtering Algorithm Based on User Behavior. The collaborative filtering algorithm based on user behavior is effective in the search of the nearest neighbor of users. This paper introduces a method to find the nearest neighbor of target users based on large interval. This method can obtain the neighbor with more similar target users, but also adaptive to find the target user's neighbors.

First, a definition is made, assuming that the same group as the target user U_a is s and d is different from the target user U_a. Then, the similarity between the target user U_a and other users U_b can be calculated according to the following formula.

If $\quad \max sim(U_a,U_b)_s > \max sim(U_a,U_b)_d,\quad$ then $\quad sim(U_a,U_b) = 0.5|\max sim(U_a,U_b)_s -$
$\max sim(U_a,U_b)_d \ | + \max sim(U_a,U_b)_s$

$$sim(U_a,U_b) = 0.5 \left| \sum_{k \in s} sim(U_a,U_b) - \right.$$

If $\quad \max sim(U_a,U_b)_s \le \max sim(U_a,U_b)_d \quad$, $\quad$ then

$\sum_{k \in d} sim(U_a,U_b) \ | + \min sim(U_a,U_b)_s$

In this formula, $sim(U_a,U_b)$ is the similarity between user U_a and user U_b; U_a and U_b are the target customers; $\max sim(U_a,U_b)_s$ is the most similar value between users U_a and s; $\max sim(U_a,U_b)_d$ is the most similar value between users U_a and d; $\sum_{k \in s} sim(U_a,U_b)$ is the sum of similarity of U_a and k-nearest neighbor in s; $\min sim(U_a,U_b)_s$ is the smallest similarity between user U_a and s, the purpose of this value is to ensure that U_a can at least find an approximate neighbor.

The main idea of this method is: to design a new measurement method with different degrees of similarity, so that the nearest neighbors in user space which have the same preferences as the target users are put in one category, and the users with different preferences are different from the target users. In this way, potential preferences of the user can be reflected, so that the distance between users who have the same preference can be shortened, and users with different preferences can increase their distance. This method can automatically determine the number of neighbors, so that the recommended effect is improved.

Then a definition is made to define the similarity $sim(U_a,U_b)$ between the target user U_a and other users U_b. Then a new similarity calculation formula is as follows:

$$Newsim(U_a,U_b) = \begin{cases} sim(U_a,U_b), sim(U_a,U_b) < S(U_a,U_b); \\ 0 \end{cases} \quad (4)$$

The approximate neighbor of the target user can be obtained through the new definition, and forecast the item that the target user does not score, which is calculated as follows:

$$P_{t,i} = \overline{R_t} + \frac{\left[\sum_{x_k \in v} Newsim(U_t,U_{x_k})\right](R_{x_k,i} - \overline{R_{x_k}})}{\sum_{x_k \in v}\left|Newsim(U_t,U_{x_k})\right|} \quad (5)$$

In this formula, $P_{t,i}$ is the score of the target user U_t for item I_i; v is the neighbor set of the target user; x_k is the k th neighbor of the target user U_t, and k is determined by the large number of neighbors; $\overline{R_t}$ is the target user The average rating.

Travel Recommendation System Diversity and Novelty

Novelty. The novelty of a travel system refers to the ability of a travel recommendation system to recommend to a user a tourist attraction that has not previously been covered.

The measure of novelty used in this article is measured by distance. The measure of this approach is expressed by the weighted average e of the product in an environment with all other content dissimilarity. The formula is as follows:

$$N(i|R) = \sum_{j \in V} p(j|R)d(i,j) \quad (6)$$

In this formula, $d(i,j)$ is a function of measuring distance, $d(i,j) = 1 - sim(i,j)$, it is possible to calculate the distance by the similarity, R is the sum of the items in which user interaction exists, and the novelty measurement is carried out by EPD.

Diversity. A recommendation system can be expressed in the following two ways of diversity.
Individual diversity

The approach to individual diversity used in this article is Internal List Distance (ILD). This method defines the diversity of the user's individual system recommendation list as the average dissimilarity of

the P items in the system recommendation list, that is, the average similarity between all the recommended items in the system recommendation list. The calculation process is as follows:

$$Div(I_1, I_2, \cdots, I_n) = ILD(R) = \frac{\sum\limits_{i=1\cdots n} \sum\limits_{j=i\cdots n} sim(I_i, I_j)}{\frac{n}{2} \times (n-1)}$$

(7)

In this article, the ild method is used as the diversity criterion for this article.

Overall diversity

The overall diversity and individual diversity is not the same, compared to the individual diversity, the overall diversity is more for different users. So it's for all users, so increasing overall diversity can make the system's recommendations more focused on items of high popularity. In this article, more is reflected in the user's personality experience, the goal is independent travelers, individuals with higher diversity does not lead to overall diversity as high. For example, when a travel recommendation system treats several popular products that are unrelated to all users, the recommended products of the travel system are diverse for all customers, but the diversity has not been raised for the travel system. Therefore, the content of this article does not put the overall diversity in the scope of the discussion.

Experimental Data Comparison Conclusion. The simulation experiment in this paper is based on the tourism data of a travel site in Xi'an. Data includes more than 3000 passengers' route data, attractions data, time data, comment rating data. Because of the sparsity of these data, preprocessing is required. The first step is to take the statistics of the scenic spots according to the response volume, and get the statistical tables of scenic spots in xi 'an. This form can reflect the relative popularity of the scenic spots in xi 'an city, as shown in the following table:

Table 1 attractions popularity table

Number	Attractions	Post	Number	Attractions	Post
1	Terracotta Warriors	2160	30	Guan Zhong College	25
2	Da-Yan Tower	1848	31	Xi'an Botanical Garden	23
3	Muslim Street	1839	32	Xi'an Daxing Temple	21
4	Ancient City Wall	1459	33	Xi'an Guangren Temple	21
5	Shaanxi History Museum	1280	34	TV tower	17
6	Bell and Drum Tower	1275	35	Zhaoling	17
7	Famen Temple	1250	36	Datang West City	15
8	Huaqing pool	794	37	Wetland Park	13
9	Huashan	786	38	Eight immortals	11
10	Datang Furong Garden	597	39	Xi'an Incident Memorial	9
11	Xi'an Museum	506	40	Qiao Family Courtyard	9
12	Crown Falls	500	41	Yang Hucheng old home	9
13	Forest of steles	398	42	Opera Grand View Garden	9
14	Safari	290	43	Catholic Church	9
15	Lishan	271	44	Wolong Temple	9
16	Qianling	210	45	Shaanxi Provincial Library	9
17	College door	208	46	Xi'an Du Ling	8
18	Daming Palace site	172	47	Zhongnanshan	7
19	Dragon Temple	107	48	Xi'an Art Museum	7
20	Qujiang pool site	98	49	Seoul Lake Park	6
21	Banpo site	68	50	Bailuyuan	6
22	Expo	63	51	Hong plot Temple	6
23	Xingqing Gong Park	62	52	Museum of Art	5
24	Tsui Wah Shan	50	53	Suzaku Forest Park	5
25	Qujiang ocean world	38	54	Wang Shun Hill	4
26	A house palace ruins	33	55	Peasant Painting Exhibition Hall	4
27	Great Mosque	32	56	Xiang Yu Forest Park	4
28	Feng Meiyukou	33	57	Xi'an Caotang Temple	3

The second step is to denoise these data, and then randomly selected 30 representative spots as experimental data, the number is $J1 - J30$. The data of the 30 spots basically contains two categories:

human landscape and natural landscape. Due to the sparseness of the data, the data of all the tourists who have scored more than 5 scenic spots are screened. Finally, 1000 visitors, 30 spots, 9885 scoring, 7394 comments and some time data are selected from the users who have at least two lines interactive records and their interaction record. Finally, the scoring matrix is obtained from these data.

Then start the experiment simulation. Through the diversity and novelty, this article compares the performance of two collaborative filtering algorithms that based on user behavior influence and user behavior habit.

Table 2   Performance Comparison Table

	Novelty	Diversity
Collaborative Filtering Algorithm Based on the Influence of User Behavior	0.635	0.745
Collaborative Filtering Algorithm Based on the Habit of User Behavior	0.876	0.847

From the table above, it can be concluded that the collaborative filtering algorithm based on user behavior has better diversity and novelty. This is because that collaborative filtering algorithm adopts a design that is different from the previous similarity measurement. In user space, the nearest neighbors that have the same preference as the target user are placed in a category, and the users with different preferences are distinguished from the target users. In this way, the potential users' preferences can be reflected, the distance between users who have the same preferences can be shortened, and the users with different preferences can be increased in distance. More tourists have access to attractions they are interested in but have not seen before.

Conclusion

Based on the novelty and diversity of tourism recommendation system, this paper introduces two common recommendation and calculation methods of diversified algorithms, and then compares and simulates the diversity and novelty of the two methods through the tourism data in Xi'an.

Acknowledgements

References

[1] Boim R,Milo T,Novgorodov S.Diversification and refinement in collaborative filtering recommender[C]//CIKM 2011,2011: 739-744.

[2] Adomavicius G,Kwon Y O.Maximizing aggregate recommendation diversity:a graph-theoretic approach[C]//Proceedings of the Workshop on Novelty and Diversity inRecommender Systems.New York:ACM, 2011:3-10.

[3] AN Wei, LIU Qihua, ZHANG Liyi. Research Progress on the Diversity of Personalized Recommendation System [J]. Library and Information Service, 2013,57 (20): 127-135.

[4] Zheng Cuicui, Li Lin.Study on similarity measurement in collaborative filtering [J] .Computer Engineering and Applications, 2014,50 (8): 147-149.

[5] Deng Ailin, left cotyledons, Zhu Yang Yong. Collaborative Filtering Recommendation Algorithm Based on Item Clustering [J] .Microsoft Microcomputer Systems, 2004,25 (9): 1665-1670.

[6] Shen Jian, Yang Yupu.Cooperative filtering recommendation algorithm based on two-stage similarity learning [J] .Application Research of Computers, 2013, 30 (3): 715-719.

[7] Deng Ailin, Zhu Yangyong, Shi Bole. Collaborative Filtering Recommendation Algorithm Based on Item Grading Prediction [J] .Journal of Software, 2003,14 (9): 1621-1628.

[8] Huang Chongguang, Yin Jian, Wang Jing, et al.Research on collaborative filtering recommendation algorithm for uncertain nearest neighbors [J] .Acta Ch Computers, 2010,33 (8): 1369-1377.

[9] Zhang Guangwei, Li Deyi, Li Peng, et al.Collaborative filtering recommendation algorithm based on cloud model [J] .Journal of Software, 2007, 18 (10): 2403-2411.

[10] Peng Shi, Zhou Zhibin, Wang Guojun. Collaborative Filtering Algorithm Based on Scoring Matrix Prefilling [J]. Computer Engineering, 2013,39 (1): 175-178,182.

[11] Tian Chao, Zhu Qing, Tan Zuo-yan, et al.Requiring query service recommendation ranking based on comment analysis [J] .Microsoft Microcomputer Systems, 2011,32 (9): 1740-1746.

[12] PENG Shi, ZHOU Zhibin, WANG Guojun. Collaborative Filtering Algorithm Based on Scoring Matrix Pre-Filtering [J]. Computer Engineering, 2013,39 (1): 175-178.