PROVIDING A MODEL FOR PREDICTING TOUR SALE IN MOBILE E-TOURISM RECOMMENDER SYSTEMS

Masoumeh Mohammadnezhad[1] and Mehregan Mahdavi[2]

[1]Department of Computer Science, Islamic Azad University, Shabestar branch, Shabestar, Iran
mohamadnezhadma@google.com

[2]Department of Computer Science and Engineering, University of Guilan, Guilan, Iran
MAHDAVI@GUILAN.AC.IR

Abstract

In this article, a new model is proposed for tourism recommender systems. This model recommends tours to tourists using data-mining techniques such as clustering and association rules. According to the proposed model, tourists are initially clustered. Self Organize Map (SOM) algorithm is used for determining the number of clusters and the clusters are created by K-means algorithm. Then, the clusters are analyzed and validated considering Quantization error, Topographic error and Davies-Bouldin error parameters. This model is implemented using two methods; according to the first method, recommendation is made based on tourists' location, and in the second method this is done based on tourists' behavioural patterns in the past. The results from evaluating the model using Pearson Correlation show that recommendations based on the behavioural patterns are closer to tourists' interests.

Keywords

Recommender systems, Clustering, Association rules mining, Cold start, Sparsity

1. INTRODUCTION

Nowadays, mobile phone is a platform for obtaining information, therefore, the characteristics and challenges provided by this device are increasing. Besides, great amount of information in mobile business processes and also physical limitations of mobile devices have resulted in an increasing attention to personalization process, so recommender systems have been developed increasingly as well [1], [2], [3]. A mobile tourism recommender system tries to simulate an off-line travel agent. The main objective of such a system is to help customers during the planning phase of their travel, because processing a huge amount of information in web sites that provide tour information/services could be complicated and confusing. In this regard, a web-based mobile recommender system can help customers to find their travel destinations considering their preferences and interests.

Recommender systems are used in a wide range of applications such as web page recommendation [4], digital news [5], e-commerce [2], movie recommendation [6], travel agent [7] and many others. In this regard, different approaches including collaborative, content based, knowledge based and demographic based filtering are applied. Collaborative filtering approaches use user information profile and extract the users according to the similarity of their

profiles. One of the disadvantages of this approach is that the recommender system cannot recommend an item until several users have rated it. This is known as the cold start problem. In content-based approaches, the items will be suggested to the customers that are very similar in content and characteristics to his or her favorite items. The problem with this approach is that the recommender system only recommends items closely related to those items that the user has shown interest in the past. However, no new items are suggested. Demographic recommenders aim at categorizing users based on their personal attributes that belong to stereotypical classes [7]. Although, there are different ways for implementation these systems but yet most of such approaches suffer from cold start and sparsity problems. The sparsity problem occurs when available data are insufficient for identifying similar users and it is a major issue that limits the quality of recommendations and the applicability of collaborative filtering in general. This is why in these systems, besides eliminating cold start problems and sparsity, the goal is to provide good suggestions, i.e., suggestions close to costumers' interests and preferences. Our proposed model addresses these issues. The proposed model initially clusters tourists, and then the created clusters are evaluated and validated using quantization, topographic and Davies-Bouldin error parameters. There are two methods to perform the model: in the first method, recommendation is performed only based on tourists' location and in the second it is done based on the behavioural patterns of tourists in the past.

In Section 2 of this article, we explain the common methods used in the recommender systems such as collaborative filtering (CF) and content based filtering (CBF). Then, we express challenges and the most common data mining methods used in recommender systems. In Section 3, the proposed model is presented and in the Section 4, it is evaluated. Finally, we conclude the paper in Section 5 and present future research.

2   RECOMMENDATION APPROACHES

Recommender systems are information processing systems which gather different kinds of data to perform suggestion. Considering the data they filter, very often, these systems are divided using collaborative, content based, knowledge based and demographic based filtering methods. The information references for these systems could be different. Table 1 presents different kinds of information used in recommender systems [8].

Table 1. A Taxonomy of input data

Data type	Explanation
Demographic data	name, age, gender, birth date, telephone, address, profession, hobbies, salary and so on.
Rating data	comments such as best, good, bad, worse; and rating such as discrete ratings and so on.
Behavioural pattern Data	click times, duration of browsing, the links of webs; open, close, save, delete of the web and even download of the web content and so on.
Transaction Data	number of purchase, purchasing date and so on.
Production data	for travel tour, it means destination, price, during of travel, hotel and so on.

Collaborative Filtering (CF): Collaborative filtering is the most common method in the recommender systems. CF uses user profile and extracts the users according to the similarity of their profiles (neighbourhood). This approach is based on this theory: users with common interests in the past will have similar behaviours in the future. CF method can generate recommendations based on WRFM parameter and customer ratings [9], [10], [11].

Content Based Filtering (CBF): Content-based filtering makes suggestions according to customer′s past interests. Therefore, the items will be suggested to the customers that are very similar in content and characters to his or her favourite items [7], [12]. One of the disadvantages of this approach is that the recommender system only recommends items closely related to those items the user has shown interest in the past. As a result, no new items are suggested [13].

Demographic Based Filtering (DF): Demographic-base filtering uses user profile information such as age, gender, marital status, postcode, occupation, and so forth. One problem with DF method is that it is time consuming, and if users do not voluntarily provide personal information, it is not possible to build any profile for them. Although, the users might provide some profile information, this could be inconsistent or not properly updated causing the profile become inconsistent over time [14], [15].

2.1. Data mining methods

Most of the recommender systems have a core algorithm that in most cases is a special king of a data mining method. The methods used in this research are clustering and Association rules mining.

Clustering: It is data division into several groups so that each group or cluster is homogeneous; the items in each group should be similar to each other and the items in each group or cluster should be different from the items in other clusters. Among the clustering methods, Self-Organizing Map (SOM) and k-Means have been used for many decades. K-means is the simplest clustering algorithm; this algorithm uses a predefined number of clusters (k) as input, which is included its name. Mean stands for an average (an average location of all the members of a particular cluster). When dealing with clustering techniques, a notion of a high dimensional space must be adopted, or space in which orthogonal dimensions are all attributes from the tables of analyzed data. The value of each attribute of an example represents a distance of the example from the origin along the attribute axes.

Association rules: These rules have the power of finding relationships among merchandises in a special scope. Therefore, they can find relations between the products in one event. This event is called a transaction such as a purchase transaction. An item set is defined as a collection of one or more items. An association rule is an expression of the form $X \rightarrow Y$, where X and Y are item sets. In this case the support of the association rule is the fraction of transactions that have both X and Y. On the other hand, the confidence of the rule is how often items in Y appear in transactions that contain X. Given a set of transactions T, the goal of association rules mining is to find all rules having support $\geq$ minsup threshold and confidence $\geq$ minconf threshold [8].

2.2. Challenges of Recommender Systems

Each kind of recommender systems has its own strengths or weaknesses. The most principle challenges related to these systems are cold start (new user / new item) and sparsity.

Cold start: Cold start is one of the problems that the collaborative filtering approach suffers from. When a system is created based on collaborative filtering, there might exist lots of products that the system is not able to offer a recommendation until some users rate them; such items are called cold start items. The notion of cold start user means that when a new user enters the system, in order to provide a recommendation, the system must know about his rating history and preferences and this is impossible unless such users enter their preferences and interests to the system [14], [15].

Sparsity: In many commercial recommender systems, both the number of items and the number of consumers are large. In such cases, even very active users may have purchased less than 1% of the items. So, the consumer-product interaction matrix can be extremely sparse. This problem is commonly named as the sparsity [16], [17].

3. THE PROPOSED MODEL

The proposed model includes four stages: tourist clustering, cluster validation, association rules mining and recommending; Figure 1 shows our proposed model. Throughout this section we exemplify different parts of the proposed method using a datasat used for the evaluation which contains 1069 tourists and 3058 purchase transactions.

3.1. Clustering

In addition to the problem of cold start, recommender systems always suffer sparsity. Clustering is a possible solution for alleviating this problem. Although, the method does not completely eliminate the problem, it alleviates the influence of sparsity on the quality of suggestions by decreasing search space.
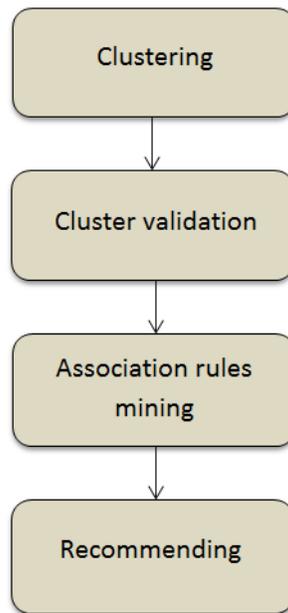

Figure 1. The proposed four-stage model

Location based method: Because tourists are scattered in different locations so their location can be applied as a principle parameter of clustering. Therefore, tourists in the same geographical area are put in one cluster. An example is given in Table 2.

Behavior based method: In this method, clustering is performed based on the behavioural similarities and shopping items among tourists. Therefore, tourists having the most similarities in their shopping items are put in the same cluster (as shown in Table 3). In this method, a relationship between customers' behavioural patterns, navigations and their shopping preferences is established. Different stages of the method work as follows: In the first stage, the data related to customers' shopping and behavioural patterns and navigation is gathered. In the second stage, costumers' priorities, when buying a special purchase, are determined numerically. When something is purchased, the level of its priority for that costumer is 1 and if that item is selected but not purchased, the number is calculated by determining the probability of entering the shopping stage by the customer which is computed considering the data obtained in the first stage.

Table 2. Clusters created based on location

Clusters	Tourists
Cluster-1	110
Cluster-2	230
Cluster-3	90
Cluster-4	42
Cluster-5	59
Cluster-6	215
Cluster-7	69
Cluster-8	51
Cluster-9	203

Table 3. Clusters created based on shopping behaviour of the tourists

Clusters	Tourists
Cluster-1	605
Cluster-2	230
Cluster-3	1330
Cluster-4	57
Cluster-5	503
Cluster-6	333

In both methods, we determine the optimum number of clusters by SOM algorithm and K-means algorithm is used for clustering.

3.2. Cluster Validation

Validating Clusters is very important for their evaluation. In different clustering algorithms, the number of clusters is quantified by certain parameters. There are lots of methods to find the optimum number of clusters. In the current research, the method used for validating clusters is Davies- Bouldin. The method is a function of total ratio in clusters' sparsity to the separation among them, as shown in Formula 1.

$$DB = \frac{1}{n}\sum_{i=1}^{n} \max\left\{\frac{S_n(Q_i) + S_n(Q_j)}{S_n(Q_i, Q_j)}\right\} \tag{1}$$

Where n is the number of clusters and S_n is the average distance of all the items included in cluster i compared to the center point and $S(Q_i, Q_j)$ is the distance between center points of clusters i and j. The less the ratio is, the denser the clusters are and the farther they get from each other.

Using numerical and topographic parameters and also Davies-Bouldin method, validating clusters for the first method is determined according to Table 4 and for the second one according to Table 5. The results are extracted after considering the above-mentioned parameters and also after validating the number of clusters:

Table 4. Evaluation of clusters in the location based method

Davies-Bouldin error	Topographic error	Quantization error
0.9312	0.024	0.516

Table 5. Evaluation of clusters in the behaviour based method

Davies-Bouldin error	Topographic error	Quantization error
0.9021	0.013	0.515

3.3. Association Rules Mining

After clusters were created, transactions of each costumer should be determined in each cluster in order for SPSS Clementine software to extract association rules from training data set. We use A-priori algorithm, which is one of the most efficient methods in association rules mining, for extracting association rules. The results of the rules found in each cluster are shown in Table 6 for the first method and in Table 7 for the second one.

Table 6. Finding association rules in each cluster in the location based method

Tours	T1	T2	T3	T4	T5	T6	T7	T8	T9	T10
Cluster-1	T8	T8,T10					T4		T1	
Cluster-2		T1,T6,T10	T8				T5			
Cluster-3		T6,T8,T10		T8		T9	T2,T4,T8		T1,T5	
Cluster-4	T7	T7	T9				T2	T6	T3	
Cluster-5	T6,T9,T10	T1,T8,T9	T7,T8				T1,T2,T5,T9	T2	T3	
Cluster-6		T7,T8,T10	T10	T7			T3,T4	T3,T6	T8	
Cluster-7	T5	T9,T10		T10			T5,T15	T5,T6	T1	
Cluster-8	T3	T1,T3,T10	T4		T7		T1,T6,T8	T1,T2	T8	
Cluster-9	T6	T8,T10					T5,T9			T3

According to Table 6, in cluster number 1, if a tourist has purchased tour T2, the system will suggest tours T8 and T10. This suggestion will be based on the abundance of the tours purchased and also un-clicked tours. Also if a tourist makes a purchase for the first time, the suggestion will be made only based on the shopping abundance in the related cluster.

Table 7. Finding association rules in each cluster in the behaviour based method

Tours	T1	T2	T3	T4	T5	T6	T7
Cluster-1	T3	T3,T8	T2				
Cluster-2	T2,T10	T6,T8,T10			T3		T5
Cluster-3		T6,T8,T10		T8		T9	T2,T4,T8
Cluster-4	T7	T7	T9				T2
Cluster-5	T6,T9,T10	T1,T8,T9	T7,T8				T1,T2,T5,T9
Cluster-6		T7,T8,T10	T10	T7			T3,T4
Cluster-7	T5	T9,T10		T10			T5,T15
Cluster-8	T3	T1,T3,T10	T4		T7		T1,T6,T8
Cluster-9	T6	T8,T10					T5,T9

3.4. Recommending

After clusters were created and existing rules were found, suggestions are made based on the cluster in which the targeted tourist is and also his past shopping history.

4. EVALUATION

In order to evaluate the proposed model, the correlation of the suggestions is obtained by Pearson method. In fact, Pearson correlation is a correlation between the actual suggestions that are made and the suggestions which are made based on the test data. Therefore, we consider two random variables with σ_x and σ_y standard deviations. Their correlation coefficient is shown using ρ and it is defined as shown in Formula 2 [18]:

$$\rho_{xy} = \frac{\mathrm{cov}(x, y)}{\sqrt{(\mathrm{var}\ x)(\mathrm{var}\ y)}} = \frac{\sigma_{xy}}{\sigma_x \sigma_y} \tag{2}$$

Different values are considered for correlation coefficient by different researchers. Some researchers assume 65% as the correlation coefficient and some others assume 35%. The amount of correlation coefficient using Pearson method for both methods is according to Table 8. As shown in the table, the suggestion based on the shopping taste is closer to tourist's interests.

Table 8. Correlation coefficient using Pearson method

Number of Recommendations	Correlation Coefficient Location Based Clustering	Correlation Coefficient Behaviour Based Clustering
K=5	68.3152	86.2943
K=10	67.23157	85.36147
K=15	66.4012	84.2511
K=20	65.0016	83.5548
K=25	63.6382	82.3175
K=30	61.9373	80.5284
K=35	59.7486	78.4152

5. CONCLUSION

This paper proposes a model for recommender systems. The model is according to clustering models and profiling customers in tourism recommender systems with a compound point of view. Considering the application, the proposed model provides an effective solution, as it provides suggestions close to tourists' interests and preferences. Besides, it effectively addresses cold start user and sparsity issues in recommender systems.

This model provides suggestions based on two criteria. If tourist has purchased tours before, the system recommends tourist's favorite tours. This recommendation is based on the abundance of related purchased tours and un-clicked tours as well. If the customer makes a purchase for the first time, the recommendation is carried out only based on the shopping abundance in the related cluster. After evaluating correlation coefficient it was found that in the second method which is based on tourist's shopping behaviour, this coefficient performs better tha the location based method. Therefore, a suggestion closer to tourist's interest is presentable. In this research, RFM parameters were not used for making suggestion. In future research these parameters may be used for increasing the accuracy and precision of the recommendations.

REFERENCES

[1]   Fesenmaier, D. R., Werthner, H., and Woeber, W. (2006) *Destination Recommendation Systems: Behavioural Foundations and Applications*, CABI Publishing, ISBN-13 978-0-85199-023-1, pp. 67-93, United Kingdom.

[2]   Werthner, H., and Ricci, F. (2003) *Electronic commerce and tourism*, Communication of ACM, Vol 47, No 12, pp. 101-105.

[3]   Werthner, H. (2003) *Intelligent systems in travel and tourism*, In Proceeding of the 18th International Joint Conference on Artificial Intelligence. IJCAI-03, Acapulco, Mexico.

[4]   Lieberman, H.., Fry, C., and Weitzman, L., (2001) *Exploring the web with reconnaissance agents*, Communications of the ACM, Vol. 44, No. 8, pp. 69-75,

[5]   Resnick, P., Iacoyou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994) *GroupLens: An open architecture for collaborative filtering on netnews*, In Proceedings of the ACM Conference on computer supported cooperative work, USA.

[6]   Salter, J., Antonopoulus, N. (2006) *CinemaScreen recommender agent: Combining collaborative filtering and content-based filtering*, IEEE Intelligent Systems, Vol. 21, No. 1, pp. 35-41.

[7]   Schiaffino, S., Amandi, A. (2009), *Building an expert travel agent as a software agent*, Expert Systems with Applications, Vol 36, No. 2, pp. 1291-1299.

[8]   Kangning, W., Jinghua, H., Shaohong, F. (2007) *Survey of E-Commerce Recommender Systems*, International Conference on In Service Systems and Services Management,China.

[9]   Ricci, F., Rokach, L., Shapira, B. (2011) *Recommender Systems Handbook*, Springer, pp. 1-184.

[10]   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, Vol. 17, No. 6, pp. 734-749.

[11]   Adomavicius, G. Sankaranarayanan, R., Sen, S., Tuzhilin, A. (2005) *Incorporating Contextual information in Recommender Systems*, ACM Transactions on Information Systems, Vol. 23, No. 1, pp. 104-112.

[12]   Agrawal, R., Srikant, R. (1994) *Fast algorithms for mining association rules*, In Proceeding of the 20th VLDB conferece, San Mateo.

[13]   Sathya Bama, B., Mohana Valli, S., Aysha Mariam M., Raju S. and Abhaikumar V. (2011) *Content Based Leaf Image Retrieval (CBLIR) Intended for e-commerce*, International Journal of Information Technology Convergence and Services (IJITCS), Vol.1, No.2. pp. 1-5.

[14]   Krulwich, B. (1997) *Lifestyle finder: intelligent user profiling using large-scale demographic data*, Artificial Intelligence Magazine, vol. 18, no. 2, pp. 37–45.

[15]   Pazzani, M. J. (1999) *A framework for collaborative, content-based and demographic filtering*, Artificial Intelligence Review, vol. 13, no. 5-6, pp. 393–408.

[16]   Manos, P., Dimitris, P., Themistoklis, K. (2005) *Alleviating the Sparsity Problem of Collaborative Filtering Using Trust Inferences*, In Proceedings of the 3rd International Conference on Trust Management (I-Trust), Paris, France.

[17]   Chen, Y. (2011) *Solving the Sparsity Problem in Recommender Systems Using Association Retrieval*, Journal of computers, Vol. 6, No. 9, pp. 1896-1902.

[18]   Topchy, A., Punch, W. (2003) *Combining multiple weak Clustering*, In Proc. Of IEEE Conf. on Data Mining (ICDM), Orlando.