International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 1, Feb 2012

AN EFFECTIVE MODEL FOR IMPROVING THE QUALITY OF RECOMMENDER SYSTEMS IN MOBILE E-TOURISM

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 major e-commerce recommendation systems, the number of users and items is very large and available data are insufficient for identifying similar users. As a result, recommender systems could not use users' opinion to make suggestions to other users and the quality of the recommendations might reduce. The main objective of our research is to provide high quality recommendations even when sufficient data are unavailable. In this article we have presented a model for this condition that combines recommendation methods (e.g., Collaborative Filtering (CF) and Content Based Filtering (CBF)) with other methods such as clustering and association rules. The model consists of four phases, at the first phase, tourists are clustered based on their location and the target tourist's cluster is sent to the next phase. In the second phase, a two level graph is made based on the similarity between the tourist interests and the similarity of the tours. According to this graph, transitive relations are discovered among the tourists and k number of items that have the highest weight of relationships and are suggested to the target tourists. According to the experiments, the standard F-measure indicates that the quality of the recommendations of this model is higher than the traditional approaches which cannot discover transitive relationships.

KEYWORDS

Recommender systems, Collaborative filtering, Content based, Association rules and Graph theory

1. INTRODUCTION

Mobile phones are becoming a primary platform for accessing information and when coupled with recommender systems technologies they can become key tools for mobile users both for leisure and business applications. Also the huge amount of data in mobile business processes and physical limitations have increased the importance of personalization process. Nowadays, expansion of mobile networks has caused the emergence of a new type of electronic business called mobile business. For this purpose the predictive and recommender systems have been developed increasingly [1], [2], [3]. Tourism is a primary application area for mobile applications to support the traveller before, during and after the travel. A mobile tourism recommender system simulates an offline travel agency. The main purpose is to help customers in travel planning because it may be so complicated and confusing to process a lot of information on the travel sites. Thus, a mobile web-based recommender system can effectively help the customers to find their trip destination according to their interests.

Recommender systems have been used to make recommendations of interesting items in a wide variety of application domains, such as web page recommendation [4], digital news [5], e-commerce [2], movie recommendation [6], travel agent [7] among others. A variety of approaches have been used to perform recommendations in these domains, including content-based, collaborative, demographic and knowledge-based [7]. Collaborative filtering approaches use user information profile and extract the users according to the similarity of their profiles. In content-based approaches, the items will be suggested to the customers that are very similar in content and character to his or her favorite items. Although, there are different ways for implementing these systems, but yet due to the increase in the precision of existing methods, recommendations are considered in various application areas.

In this research, the main objective is to increase the quality of recommendations in mobile tourism recommender systems. We provide an effective model composed of collaborative filtering and content based filtering. In this model, the tourists are clustered based on their location initially, in order to reduce the search space and making the neighbourhood of the active tourist. Clustering the tourists results in better performance of the proposed model. After clustering, tourists' profile is created based on the characteristics of the tourists and tours. The profiles are made based on their behavioural patterns, ratings and content characteristics of the tours. Then, a two-level graph is created that is composed of the tour-tour and tourism-tour similarities. Finally, recommendations are made based on the above-mentioned graph by Branch and Bound (B&B) algorithm and k recommendations are suggested.

In Section 2 of this article, we explain the common methods which could be used in recommender systems such as collaborative filtering (CF) and content based filtering (CBF). Then, we express the challenges and the most common data mining methods used in recommender systems. In Section 3, the proposed model is presented and in Section 4, the experimental results are presented. Finally, Section 5concludes the paper and presents future trends for research in this field.

2. RECOMMENDATION APPROACHES

The approaches of recommender systems are usually classified as collaborative, content-based demographic and knowledge-based. We review the most common methods in the rest of this section.

Collaborative Filtering (CF): Collaborative filtering is the most common method in the recommender systems. CF uses user profile and extracts the users according to their profiles similarity (neighborhood). This approach is based on this theory: users with common interests in the past time will have similar behaviors in the future. CF method can generate recommendations based on following information [8], [9], [10]:

a) **Weighted Recency, Frequency and Monetary (WRFM):** Customer loyalty and the importance of the membership time duration in a commercial website are very important in the process of identifying and recommending of the customer. Recommending is composed of three components: recency, frequency and monetary. Recency is an interval between the times of the last purchase date until now. Frequency is the number of purchase at a specific timeframe. Monetary is the amount of money that the customer has spent during a specific timeframe. Analytic Hierarchy Process (AHP) is used to determine the importance of RFM variables. AHP approach can identify each variable weight. First, it will be asked from the experts to compare two variables and according to the importance of the variables, put the numbers 1 to 9 beside them. Then, adaptation of the comparisons is considered, if there is not any compatibility among them, the first phase will be repeated [11], [12].

b) **Customers' ratings:** In this approach, customers' priorities are extracted from their behavioural patterns, navigations, ratings and purchasing priorities. This method has 2 phases; in the first phase, customers' behavioural patterns, navigations and their purchases are collected and in the second phase, the customer preferences in purchases are determined numerically. If an item is purchased then its priority level will be 1. If an item is selected but not purchased then it′s value will be computed via probability of its purchase [13].

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 character to his or her favourite items [7], [11].

2.1. Data Mining Techniques for Recommender Systems

Almost all recommender systems use data mining techniques to generate suggestions. The types of data mining techniques which are used in this paper are clustering and association rules mining.

Clustering: It is data division into several groups so that the data in a group should have the most similarity together and the most differences with the other groups. Among the clustering methods, self-organizing map and k-Means have been used for many decades [14].

Association Rules Mining: Association rules can discover relationships between products in a particular domain. So, they can find relations between the products in one event, this event is called transaction as a purchase transaction. We define an item set as a collection of one or more items. An association rule is an expression in the form of X→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 rule mining is to find all rules having support ≥ minsup threshold and confidence ≥ minconf threshold [15], [16], [17].

2.2. Recommender Systems Challenges

Despite their popularity and advantages, recommender systems have several shortcomings:

Cold start: It refers to the situation in which an item cannot be recommended unless it has been rated by a substantial number of users. This problem applies to new and obscure items and is particularly detrimental to users with eclectic taste. Likewise, a new user has to rate a sufficient number of items before the recommendation algorithm be able to provide reliable and accurate recommendations [18].

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 [19], [20].

2.3. Graph Based Model

In graph-based approaches, the data is represented in the form of a graph where nodes are items, users or both, and edges are the representative of interactions or similarities between the users and items. These approaches allow the nodes that are not directly connected, to influence each other by spreading information along the edges of the graph. It is shown in Figure 1, in a bipartite graph, the graph has two sets of nodes, users and items, and an edge that connects user *u* to item *i* if u rates i. The weight of each edge represents the correlation value [8], [21], [22].

Suppose the recommender system needs to recommend products for consumer c2. The standard CF algorithm will make recommendations based on the similarities between c1 and other consumers (c2 and c3). The similarity between c2 and c3 is obvious because of previous common purchases (i3). As a result, i1 is recommended to c2 because c3 has already purchased it [23], [24].

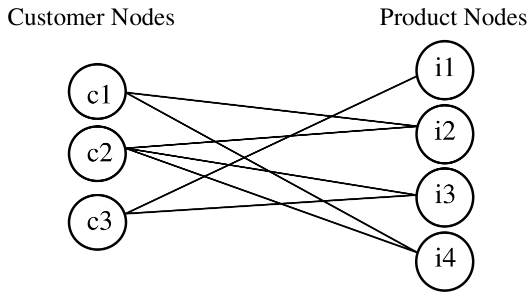

Figure 1. Representation of Users-Items in a bipartite graph [22]

3. THE PROPOSED MODEL

The proposed model consists of two phases as shown in Figure 2:

Phase 1: Almost all the recommender systems have sparsity problem. Clustering may be a way to reduce this problem. Although, this method may not solve the sparsity problem, but it can reduce the sparsity by minimizing the search space, so it would be effective in the quality of suggestions. In the tourism application, tourists' location can be the main parameter in clustering because they are scattered in the search space. So, the tourists are placed in a cluster in one particular geographic area. After this step, the cluster which includes the active user is sent to the second phase of the model as input.

Phase 2: In this phase, two types of information are used: transactional data and content of items.

a) **Similarity of the tourist-tour based on the transactional data:** At this phase, similarity between the tourists and tours is calculated by two types of input data. In the first approach, join over time and RFM parameters and in the second approach, tours ratings are used. To calculate the similarity in the first approach, we should extract the purchase transactions in the database and then RFM parameters are calculated. RFM parameters are normalized as shown in Formula 1:

$$x' = \frac{(x - x^S)}{(x^L - x^S)}. \tag{1}$$

Since the increase of the R parameter causes decreases in the tourist's loyalty, the expression shown in Formula 2 is used instead.

$$x' = \frac{(x^L - x)}{(x^L - x^S)}. \tag{2}$$

In the expression shown in Formula 1, x' is the normalized parameter and x is the original data. x_L and x_S are the maximum and minimum of RFM parameters. The weight of RFM parameters is calculated by AHP method. In this approach, tour ratings are extracted from the triple behaviours of the tourists by the web usage mining. Similarity of the tourists is calculated as shown in Formula 3:

$$\mathrm{Sim}_{WRFM}(ci,cj)=\frac{\sum_{s\in V}(WRFM_{ci}-\overline{WRFM_{ci}})(WRFM_{cj}-\overline{WRFM_{cj}})}{\sqrt{\sum_{s\in V}(WRFM_{ci}-\overline{WRFM_{ci}})^2\sum_{s\in V}(WRFM_{cj}-\overline{WRFM_{cj}})^2}}. \quad (3)$$

So, $\overline{WRFM_{ci}}$ and $\overline{WRFM_{cj}}$ are average RFM parameters for tourist *ci*, RFM parameters for tourist *cj*, respectively and *V* is a set of RFM parameters. $WRFM_{ci}$ and $WRFM_{cj}$ are normalized value of the RFM parameter ci and RFM parameter of tourist *cj*, respectively.

In the second approach, tour ratings are extracted from triple behaviours by web usage mining. Triple behaviours are composed of click on the tour link to view details, put tour in shopping basket and purchase tour. Then ratings are presented by the rating matrix, Rating=(r_{ij}), i=1,…,M and j=1,…,N (M and N are the number of tourists and tours, respectively) and each matrix element (r_{ij}) is calculated according to Formula 4. Therefore, r_{ij}^{c} is the number of user i's clicks on the tour j's link. Also, r_{ij}^{b} and r_{ij}^{p} are the number of putting tour into shopping basket and purchasing tour *j* by tourist *i*, respectively. The value of r_{ij}^{p} can be extracted by the purchase request tables in the data base.

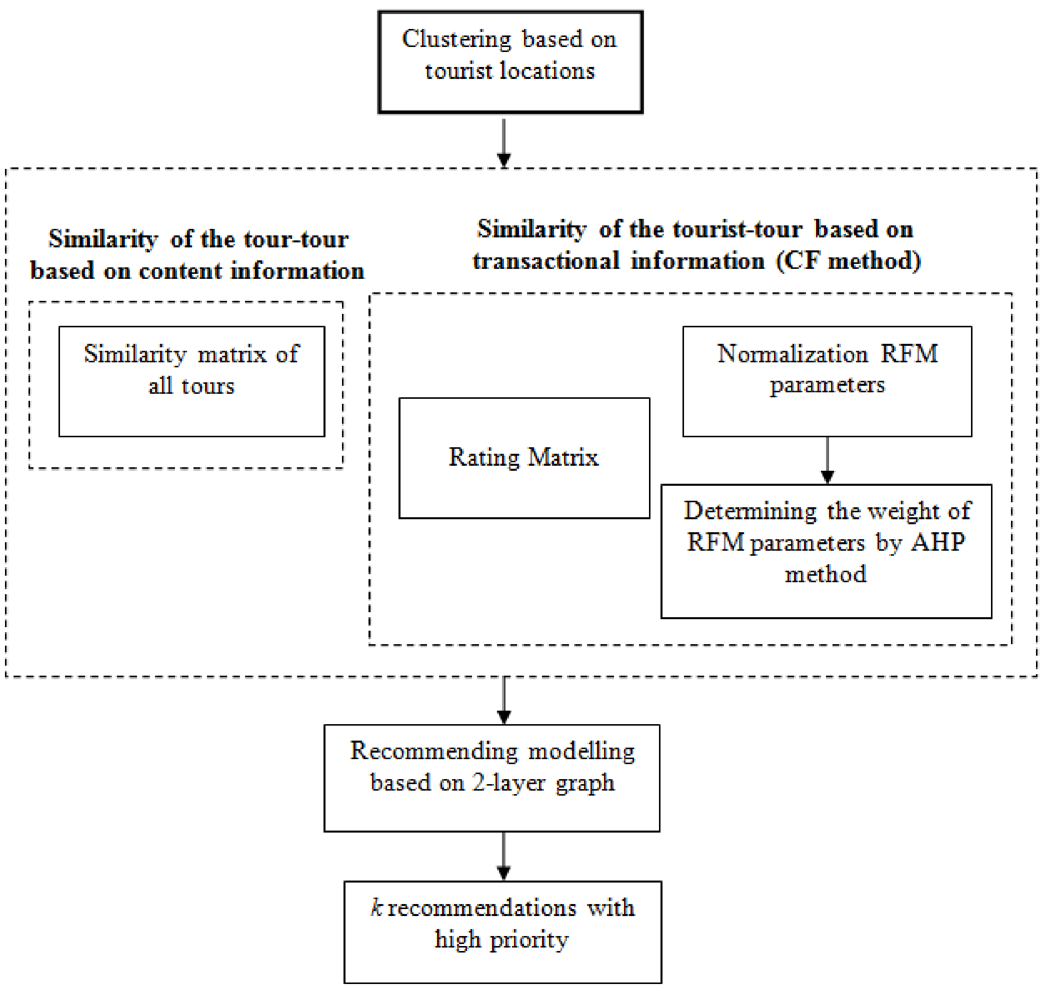

Figure 2. The Proposed Model

$$r_{ij} = r_{ij}^{c} + r_{ij}^{b} + r_{ij}^{p}. \tag{4}$$

Similarity of the tourists is computed using the above-mentioned matrix as shown in Formula 5:

$$\mathrm{Sim}_R(i,j) = \frac{\sum_{k \in S}(R_i - \overline{R_i})(R_j - \overline{R_j})}{\sqrt{\sum_{k \in S}(R_i - \overline{R_i})^2 \sum_{k \in S}(R_j - \overline{R_j})^2}}. \tag{5}$$

Therefore, $\overline{R_i}$ and $\overline{R_j}$ are the average rating of tourists *i* and *j*, respectively and *S* is a set of tourists' ratings.

Finally, the total similarity is defined as shown in Formula 6:

$$\mathrm{CSim}_{WRFM-R} = w_{WRFM} \times \mathrm{Sim}_{WRFM}(i,j) + w_R \mathrm{Sim}_R(i,j). \tag{6}$$

In the expression shown in Formula 6, w_{WRFM} and w_R are positive values in such a way that $w_{WRFM} + w_R = 1$. In this paper, w_{WRFM} and w_R are considered 0.7 and 0.3, respectively (based on our empirical evaluations).

b) **Similarity of the tour-tour based on content of the tours:** In this phase, the correlation between tours is analyzed. The tour correlations are extracted by finding relationships between tour features. For this purpose, we use association rules. There is an example of these rules in tourism recommender system:

R1: Month = January → Place = beach [sup: 0.40, conf: 0.82]

After the extraction of association rules (for example Table 1), a feature-feature matrix is generated by using the confidence parameters as shown in Table 2. Therefore, Formula 7 calculates the similarity between tours t_i and t_j as:

$$\mathrm{Sim_C}(t_i,t_j) = \frac{\sum_{t \in T} \mathrm{ASim}(t_i,t_j)}{T}. \tag{7}$$

So, *T* is the number of features in tour t_i and $\mathrm{Sim_C}(t_i,t_j)$ is the similarity between characteristics of the tours.

Table 1. An example of association rules

Rule	Support	Confidence
Month = January → Cost = Low-cost	35%	66%
Month = January → Place = Beach	40%	82%
Cost = Low-cost → Place = Beach	50%	77%

Table 2. A simple matrix of feature-feature similarity

	January	Low-cost	Beach
January	1	0.66	0.82
Low-cost	-	1	0.77
Beach	-	-	1

Phase 3: In the 2-level graph, tourists are in one layer and tours are in the other layer. There are two kind of connections between nodes; Connections that are created from similarity of the tour-tour and tourist-tour. Each Connection between tours shows their similarity. Connections between the two layers is created by transactional data that each connection shows similarity of the tourist-tour.

By selecting the various kinds of connections, we can use them to generate recommendations. If only the content of the tour is selected, it means that only the existing connections in tours layer are used, which is exactly the content based approach. If the connections between two layers (tourist and tour) become active, it means that the collaborative filtering approach has been activated. If all connections become active, it means that the hybrid approach has been selected.

Different recommendation methods can be used based on the graph model. In this paper, the nodes that have the highest correlation with the starting item are suggested (spreading activation algorithm as graph search). Spreading activation activates a number of nodes as starting nodes and then follows the links that they are connected to starting nodes. Then, it is iterated to those nodes that are already active. In spreading activation, all inactive nodes will be activated in a specific activity level. For this intention, there is Branch-and-Bound (B&B) search algorithm based on state space traversal process.

Phase 4: The output of phase 3 is the tours with various weights. k is the number of tours that are presented to the tourists. In this paper, k is 9 out of 30 tours.

4. EMPIRICAL EVALUATION

In order to evaluate the performance of the proposed model, a dialog between a tourist and a mobile agency clerk has been simulated. After the tourist registers in the system, all the places that are his favorite and he has never purchased them before, are suggested to him. To evaluate the model, the data has been collected from a set of 100 mobile tourists that rated 30 tours. The data set is divided into two parts; training set with 90% of the total ratings and the test set with 10% of the total ratings. k-means algorithm within SPSS Clementine environment has been used for clustering tourists. The optimal number of clusters has been determined into five clusters by Rapid Miner. Then, the k-means algorithm clusters the whole data set into five clusters. In this process, the records that belong to clusters 1, 2, 3, 4, 5 are 25, 16, 18, 26, 15 in order. The target tourist's cluster is the third cluster with 18 records. It is obvious that the proposed model only uses 18 tourists instead of 100 and this decrease in the number of tourists can influence the performance of the recommender system. Also, Apriori algorithm (within SPSS Clementine environment) is used for extracting association rules. To determine the similarity, only the rules that contain just one member are used. Also, for generating the recommendations, a program was written in Turbo C++ that implements the algorithm of generating k recommendations.

The Recall and Precision criteria are used to evaluate the performance of recommender systems. Recall means that how many of the purchased items that the costumers really bought have been put correctly into the recommendation list and Precision means that how many of suggested items actually have been purchased by the costumers. These two criteria can be simply calculated but they somewhat work contradictory with each other. For example, if the number of recommendations increases, Recall will increase but Precision will decrease. So, F-measure is used to combine Precision (Formula 8) and Recall (Formula 9) criteria as shown in Formula 10:

$$\text{Precision} = \frac{|\text{Test set} \cap \text{Recommended tours set}|}{|\text{Test set}|}. \tag{8}$$

$$\mathrm{Recall} = \frac{|\mathrm{Test\ set} \cap \mathrm{Recommended\ tours\ set}|}{|\mathrm{Recommended\ tours\ set}|}. \tag{9}$$

$$\mathrm{F-measure} = \frac{2 \times \mathrm{Precision} \times \mathrm{Recall}}{\mathrm{Precision} + \mathrm{Recall}}. \tag{10}$$

The following definition, as shown in formula 11, is used to measure the density degree of a graph:

$$\mathrm{Graph\ density} = \frac{\mathrm{Number\ of\ actual\ links\ present\ in\ the\ graph}}{\mathrm{Number\ of\ possible\ links\ in\ the\ graph}}. \tag{11}$$

We have compared the results of the proposed model with user-based collaborative filtering (by Pearson correlation) and content based filtering. The results are shown in Table 3.

Table 3. Recommendation Precision, Recall, F-measure for three algorithms

Algorithm	Precision	Recall	F-measure
HF-B&B based	0.29	0.73	0.42
CF-User Based (Correlation)	0.21	0.51	0.33
CBF	0.19	0.32	0.24

In Table 3, the values of Precision, Recall and F-measure criteria at one density level are presented for three types of methods, i.e., HF-B&B Based, CF-User Based (Correlation) and CBF. Also for these three methods, F-measure is considered in 11 density levels as shown in Figure 3. As shown in the figure, when the density level is low, the quality of recommendations in all the three methods is on the same level, but when the density increases, the user based CF method will have better results than the CBF method. Because the user-based algorithms are not able to utilize the associations between items, the curve for user-based (similarity function) algorithm maintains lower level of performance than HF method. Also, as can be seen in the figure, the CBF method performed poorly in our experiments. This is mainly due to the characteristics of our data set, in which the number of items is larger than the number of users, and the user-item interaction matrix is relatively sparse. However, with a different type of dataset in which the number of items is small and the number of users is large, the item-based approach should have better performance.

In high level of sparsity, the quality of the proposed model is at least 29 percent better than CF (user based) and CBF. The recommendation quality of HF (spreading activation-based) method increased faster than of the standard approaches because the transactional data accumulates during the initial deployment phase of the recommender system.

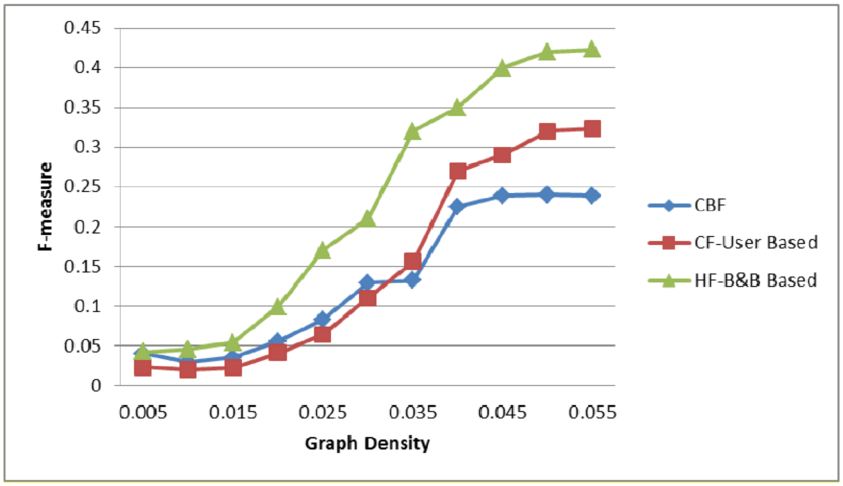

Figure 3. The amount of F-measure criteria for three methods in the eleven levels of the sparsity graph

5. Conclusions

In this paper, we aimed to improve the quality of recommendations in the proposed model. For this purpose, two effective approaches have been used in two different phases. In the first phase, clustering was used and in the second phase a graph-based model for recommendation was used. These two methods are complementary because the computational complexity of graph-based recommendation is very high and clustering reduces search space and computations. The experiments indicate that our approach improves the quality of recommendation significantly compared to standard CF and CBF. In the future research, connections of user layer with demographic data could be considered. This approach is expected to improve the precision of the recommendations.

References

[1] Fesenmaier, D. R., Werthner, H., and Woeber, K. (2006) *Destination Recommendation Systems: Behavioural Foundations and Applications*, CABI Publishing.

[2] Werthner, H. and Ricci, F. (2004) *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., 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] Ricci, F., Rokach, L., Shapira, B. (2011) *Recommender Systems Handbook*, Springer, ISBN 978-0-387-85819-7, pp. 1-184.

[9] 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.

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

[11] Banati, H., Mehta, S. (2010) *Memetic Collaborative filtering based Recommender System*, Second Vaagdevi International Conference on Information Technology for Real World Problems, Warangal, India.

[12] Soo, Y., Bong, K., Yum, J., Song, J., Myeon Kim, S. (2005) *Development of a recommender system based on navigational and behavioral patterns of customers in e-commerce sites*, Expert Systems with Applications, Vol. 28, No. 2, pp. 381-393.

[13] Liangxing, Y., Aihua, D. (2010) *Hybrid Product Recommender System for Apparel Retailing Customers*, In proceeding ICIE '10 Proceedings of the 2010 WASE International Conference on Information Engineering, Washington, DC, USA.

[14] Gupta S., Kumar, D. and Sharma, A. (2011) *Performance Analysis Of Various Data Mining Classification Techniques On Healthcare Data*, International Journal of Computer Science & Information Technology (IJCSIT), Vol.3, No. 4.

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

[16] Margahny, M.H., Mitwaly, A.A. (2005) *Fast Algorithm for Mining Association Rules*, AIML 05 Conference, Cairo, Egypt.

[17] Shaw, G., Xu, Y., Geva, S. (2009) *Investigating the use of association rules in improving recommender systems*, Proc. 14th Australasian Document Computing, Symposium, Sydney, Australia.

[18] Schein, A., Popescul, I., Ungar, A. (2002) *Methods and Metrics for Cold-Start Recommendations*, In Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Finland.

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

[20] Yujie, Z., Licai, W. (2010) *Some Challenges for Context-aware Recommender Systems*, The 5th International Conference on Computer Science & Education, Hefei, China.

[21] Huang, Z., Chen, H., Zeng D. (2004), Applying Associative Retrieval Techniques to Alleviate the Sparsity Problem in Collaborative Filtering, ACM Transactions on Information Systems, Vol. 22, No. 1, pp. 116-142.

[22] Hung, Z., Chung, W., Chen, H. A. (2004) Graph Model for E-Commerce Recommender Systems. Journal of the American Society for Information Science and Technology, Vol. 55, No. 3, pp. 259-274.

[23] Chen, H., Dhar, V. (1991) Cognitive process as a basis for intelligent retrieval systems design, Information Processing and Management, Vol. 27, No. 5, pp. 405-432.

[24] Chen, H., Lynch, K. J., Basu, K., NG, D. T. (1993) Generating, integrating, and activating thesauri for concept-based *document retrieval*, IEEE Exp., Spec. Series Artif. Intell. Text-based Inf. Systems, Vol. 8, No. 2, pp. 25-34.