SCALABLE COLLABORATIVE FILTERING RECOMMENDATIONS USING DIVISIVE HIERARCHICAL CLUSTERING APPROACH

S. Saint Jesudoss*

Abstract: *Recommender system is the most important technology in e-commerce .It is used to suggest valuable products for the customer and improve their business intelligence. Collaborative filtering is a technique which is used to suggest information from similar kinds of users. Scalability is the biggest challenge in collaborative filtering recommender system. When more number of users is increasing in the site the system should provide accurate recommendations for the super user. We use divisive hierarchical clustering approach to overcome this scalability issue when more number of users increases in terms of neighborhood size.*

Keywords: *Collaborative Filtering, Recommender Systems, Divisive Hierarchical Clustering, Web Personalization*

*Department of Information Technology, Dr SJS Paul Memorial college of Engineering and technology, Pondicherry, India

1. INTRODUCTION

In the present scenario, e-commerce sites are used to provide customers with products like books, movies etc. The main purpose of an ecommerce site is used to provide valuable products for the customers and improve their website. Therefore the site has to introduce new techniques to improve their business intelligence. The items in the e-commerce site are very large so that the user may be spending much time in searching his or her favorite items. Web personalization is the solution to these kinds of information overload problems. The goal of web personalization is to help the user with customized relevant information. Recommender systems are used to improve the business intelligence in a e-commerce site. The main use of a recommender system is that it suggests products or items that a user may be interested. Recommender systems are based on the concept of information filtering. The goal of an information filtering system is to sort through large volumes of dynamically generated information and it provides those information to users which are likely to satisfy his or her requirement. It reduces the problem of information overload and gives precise information to the customer. Therefore the user does not need to search for his favorite products

Collaborative filtering technique is the most successful information filtering technique used in ecommerce. Collaborative recommender is the process of filtering and evaluating items through the opinions from other people. Collaborative filtering techniques identify the likely preferences of a user based on the known preferences of other user’s is used for generating very effective quality recommendations. Collaborative filtering recommendations are considered better than the content based approaches and it does not require users to explicitly state their preferences.

Clustering techniques have been used in collaborative filtering recommender systems. Clustering techniques work by identifying groups of users who appear to have similar preferences. Once the clusters are created, predictions for an individual can be made by averaging the opinions of the other users in that cluster. Some clustering techniques represent each user with partial participation in several clusters. The prediction is then an average across the clusters, weighted by degree of participation.

In this paper we propose a user based collaborative filtering approach using the divisive hierarchical clustering algorithm. Our proposed system is used to cluster similar users and to recommend items from the top k neighbors.

The paper is organized as follows: section 1 portrays about the introduction about the recommender system. Section 2 narrates about the user based collaborative filtering and item based collaborative filtering .Section 3 describes about the related works in collaborative filtering approach based on clustering techniques .section 4 describes about the research literature in collaborative filtering recommender system. Section 5 describes about the problem definition. Section6 describes about the proposed system .section 7 describes about the divisive hierarchical clustering approach. Section 8 describes about the experimental evaluation and Section 9 outlines the summary.

2. BACKGROUND

User-based collaborative filtering

User based collaborative filtering technique main task is to find users that are similar to a super user. The users similar to a active user are called as super users neighbor. These neighbors for the active user are used as recommenders. Generally the user-based collaborative filtering working process can be broken down into two major steps:

Neighborhood formation: It is the application of the selected similarity metric leads to the construction of the active user's neighborhood Given an super user ' *u'* , compute the similar users from all the users data based on the similarity or their respective choices. Pearson correlation and cosine distance are popular functions for the similarity computing. The top-N most similar users become members of *u's* neighborhood.

Rating prediction**:** It is based on these neighborhood predictions for items rated by the active user are produced. Once the top-N closest neighbors have been selected, for each item predicted, these highest ranking neighbors that have rated tie item in question are used to compute a prediction.

Item-based Collaborative Filtering

The main task of item based collaborative filtering is to find items that are similar to an active item. The populations rating on the items are used to determine item similarity. Item-based collaborative filtering analyzes the user-item matrix to identify relations between the different items, and then use these relations to compute the list of top-N recommendations are made by considering the active users rating on items similar to the active item, usually referred to as the active item neighbors.The basic idea of Item-based collaborative filtering algorithm is choosing K most similar items and getting the corresponding similarity according to the similarity of rated item and target items. Then we can compute the rating of predictions through the formula with the ratings of the target user to the best several similar neighbors and their similarity

3. RELATED WORKS

In this section we will discuss related works in collaborative filtering approach based on clustering techniques

K-means clustering technique

Songjie Gong, Hongwu Ye has proposed[17] joining user clustering and item based collaborative filtering in personalized recommendation service. K-means clustering technique has been used in collaborative filtering recommender systems to cluster the users and items. They have used K-means clustering algorithm to cluster users and items. The idea is to divide the users of a collaborative filtering system using K-means clustering algorithm and use the divide as neighborhoods. The k-means cluster the user into groups called as cluster centers.

Fuzzy C-Means (FCM) clustering

Liu Hongmin et.al, Yin Zhixiet.al proposed [11] Applying multiple agents to Fuzzy collaborative filtering. They had developed a new method for recommending items using multiple agents. The agents were established by employing the fuzzy C-means clustering technique. FCM clustering is an iterative technique that starts with a set of cluster centers and generates membership grades, used to induce new cluster centers. The resulting cluster centers will be taken as multiple agents. The multiple agents are constructed based on all users in the database. Their ratings do not actually exist. They use the multiple agents to take the place of actual users as neighbors. This will greatly increase the co-rated items between user and its neighbors.

Bi Clustering

Pablo A. D. de Castro e. al, Fabricio O. de Franca Hamilton M. Ferreira and Fernando J. Von Zuben et al[11] have proposed Applying Biclustering to Perform Collaborative Filtering. Biclustering Technique has been used in collaborative filtering recommender systems to cluster similarities between users and items. They cluster rows and columns of the user item matrix at the same time.

Co Clustering

Thomas George et al, Srujana Merugu et al proposed [5] A Scalable Collaborative Filtering Framework based on Co-clustering which solves the recommendation problem in terms of a weighted matrix approximation and motivate the co-clustering approach for solving it. Co Clustering simultaneously obtain user and item neighborhoods via co-clustering and generate predictions based on the average ratings of the co-clusters while taking into account the individual biases of the users and items.

Self Organizing Map (SOM) clustering

SongJie Gong et al , HongWu Ye et al, XiaoMing Zhu et al proposed[16] Item-Based Collaborative Filtering Recommendation using Self-Organizing Map .They used SOM to cluster items. Firstly, it employs clustering function of self-organizing map to form nearest neighbors of the target item. Then, it produces prediction of the target user to the target item using item-based collaborative filtering.

Genetic clustering

Feng Zhang et al, Hui-you Chang et al [3] proposed a collaborative filtering algorithm employing genetic clustering to ameliorate the scalability issue. They have considered issues like coding scheme, the fitness function, the genetic operator, running parameters and the decoding method. Their research designs and implements an off-line running genetic clustering algorithm embedded in a memory-based collaborative filtering algorithm

RecTree Clustering

Jerome Kelleher et al and Derek Bridge et al [1] proposed RecTree Centroid: An Accurate, Scalable Collaborative Recommender It builds a binary tree of clusters; it was previously used to cluster users. RecTree recursively invokes the k-means clustering algorithm. In k-means, elements are repeatedly assigned to clusters on the basis of their similarity to the cluster centre.

4 .RESEARCH LITERATURE

In this section we brief present some of the research literature related to collaborative filtering recommender systems.

Tapestry is one of the earliest implementations of collaborative filtering based recommender systems [2]. This system relied on the explicit opinions of people from a close community, such as an office workgroup. Tapestry is an experimental mail system which was developed at Xerox Palo Alto Research Center in the year 1992 .This system was used as a mail filtering information system utilizing the collaborative sense of a small group of participating users.

The Group Lens research system in 1994 provides a collaborative Filtering solution for Usenet news and movies implemented a Collaborative Filtering algorithm based on common user’s preferences. Nowadays, it is known as user-based CF algorithm [12]. Then the concept of real automated collaborative filtering first appeared in the Group Lens Research Project system using neighborhood-based algorithm for providing personalized predictions for Usenet news articles. Since then, the research on collaborative filtering has drawn significant attention and a number of collaborative filtering systems in broad variety of application domains have appeared.

Amazon.com introduced the item to item recommendation collaborative filtering technique [13]. It was considered as the most publicly aware system. It is used to personalize the online store for each customer. The store radically changes based on customer interests.

5. PROBLEM DEFINTION

The challenge in user based collaborative filtering recommender system is scalability. We propose a divisive hierarchical clustering approach to cluster similar kind of users and propose a solution to the scalability problem. The main advantage of using this divisive hierarchical clustering approach is that it is more suitable for making real time recommendations

6. PROPOSED SYSTEM

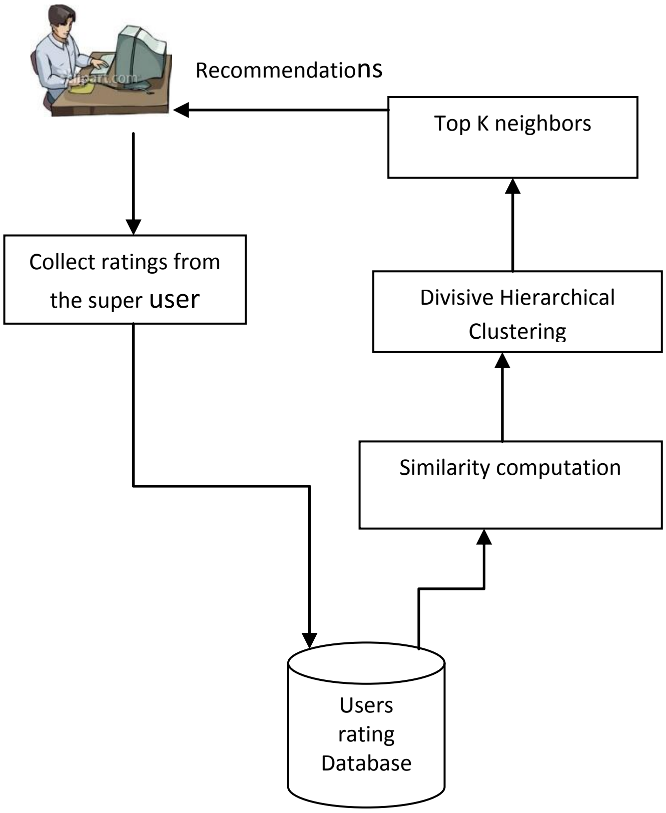

Figure.1. Proposed Architecture for collaborative filtering recommendation system

The ratings are collected from the super user. The ratings are collected and are represented in the form of user-item matrix shown in table-1.The ratings of the super user are compared with other users in the rating database and their similarity are computed using Pearson's correlation coefficient. Using the similarity values we cluster the users based on divisive hierarchical clustering approach and find the top k neighbors for producing recommendations. The recommendations from the top k neighbors are the products that the super user has not accessed yet that are given high ratings by their top k neighbors.

Table1: User-Item Matrix

Ratings	Item1	Item2	Item3	Item4
User1	5	5	1	1
User2	4	5	1	2
User3	1	1	5	5
User4	2	1	5	4
User5	1	1	1	3

Similarity Computation:

The similarities of the super user and the other users are calculated using Pearson's correlation coefficient

$$w(a,i)=\frac{\sum_j (v_{a,j}-\bar{v}_a)(v_{i,j}-\bar{v}_i)}{\sqrt{\sum_j (v_{a,j}-\bar{v}_a)^2 \sum_i (v_{i,j}-\bar{v}_i)^2}} \qquad (1)$$

here *w(a,i)* represents the Pearson's coefficient and $v_{i,j}$ is the rating that user i gave to item j and $\bar{v}_i$ represents the average rating of user i.

Table 2: Distance matrix

Similarity	User 2	User3	User 4	User 5
Super user	0.3	0.5	0.8	0.7

The similarity values are computed for the super user and the divisive hierarchical clustering algorithm is applied. Using that algorithm we are clustering the most similar users for the super users and selecting the top k neighbors for the super user and recommending the items from the them that the super user has not accessed yet.

Predicted rating:

After calculating the similarities for the super user and the other users we have to calculate the predicted ratings for the super user. The predicted ratings can be calculated using

$$P_{a,j} = \bar{v}_a + \sum_{i=1}^{n} \frac{v_{i,j}-\bar{v}_i * w(a,i)}{\sum_{i=1}^{n} w(a,i)} \qquad (2)$$

here $P_{a,j}$ represents the predicted rating of the user a on the unknown item j, *w(a,i)* represents the Pearson's coefficient $v_{i,j}$ the rating that user i gave to item j and $\bar{v}_i$ represents the average rating of user i. The predicted ratings can be used for calculating the ratings for the super user who doesn't rate a particular item. Predicted ratings are useful for finding the mean absolute error of a collaborative filtering approach. When the user is new and the item is an existing one, the predicted value is the item average. Similarly, when the item is new and the user is an existing one, the predicted value the user average. When both the user and item are new, there is no specific information and we just return the global average of all the known ratings

7. DIVISIVE HIERCHICAL CLUSTERING

The divisive algorithm is a top down approach. It is based on repeated cluster bisectioning approach [2]. Initially all the user similarity values in the dataset are assigned to a single cluster. Then the Users similarity in the cluster is further divided into two based on bi sectioning approach. Replace the chosen cluster with the sub-clusters. The process continues until n-1 times and it leads to n leaf cluster

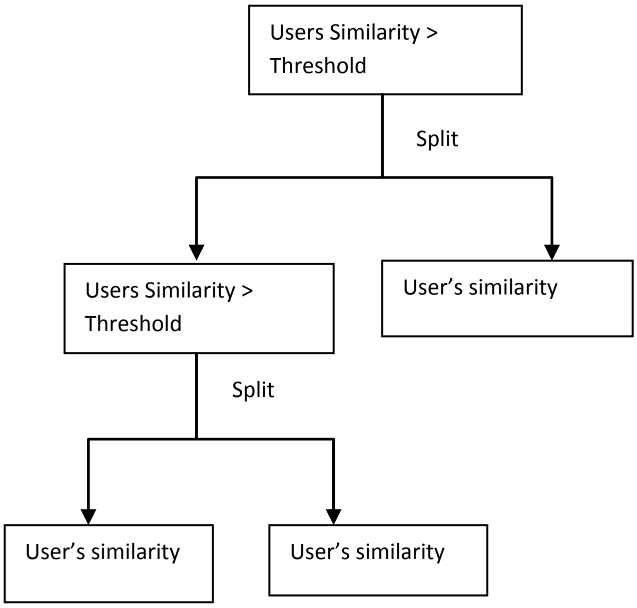

Figure.2.Formation of user clusters using divisive hierarchical clustering approach

Algorithm

A partition of is a list *(C1, . . . ,*C_K*)* of clusters verifying

C1 ∪ . . . ∪ C_K = and $C_k \cap C_{k'}$ =Ω ; for all k ≠ k'.Let N be the number of users in Ω. Each user is described on p real variables $y_1 \ldots\ldots y_p$

The inertia I of a cluster C_k is a homogeneity measure equal to

$$I(C_k) = \sum_{X_i \in C_k} p_i d^2{}_M(X_i, \bar{X}_k) \qquad (3)$$

Bipartitioning a cluster

Let C be a set of n objects. We want to find a bipartition *(C1,C2)* of C such that the within cluster inertia is minimum. Optimal bipartition *(C1,C2)* among the $2^{n-1}-1$ possible bipartitions.

Choice of the cluster

Let $P_K = (C_1, \ldots, C_K)$ be a K-clusters-partition of Ω . At each stage, a new (K+1)-clusters partition is obtained by dividing a cluster $C_k \in P_K$ into two new user clusters C_1k and C_2k. The purpose is to choose the cluster $C_k \in P_K$ so that the new partition,$P_{K+1} = P_K [\{C^1k, C^2k\} - \{C_k\}$ has minimum within-cluster inertia.

The criterion used to determine the cluster that will be divided is then equal to:

$$\Delta (C_k) = I(Ck) - I(C^1k) - I(C^2k) \qquad (4)$$

It means that the bipartitions of all the clusters of the partition P_K have been defined previously. At each stage, the bipartitions of the two new clusters C^1k and C^2k are defined and used in the next stage.

The stopping rule and the output

The divisions are stopped after a number L of iterations and L is given as input by the user, usually interested in few clusters partitions. Indeed, the last partition obtained in the last iteration is a L+1-clusters-partition. The output of this divisive clustering method is a hierarchy H which singletons are the L+1clusters of the partition obtained in the last iteration of the algorithm.

8. EXPERIMENTAL EVALUATION

Dataset

We use Movie lens data set for evaluating our experiment (www.grouplens.org). The data set contained 100,000 ratings from 943 users and 1682 movies (items), with each user rating at least 20 items. We have taken a sample of 100 users for evaluating our divisive hierarchical clustering approach. The ratings provided by the users are in the range 1 to 5.

Evaluation Metric

Mean Absolute Error (MAE) is the evaluation metric for our collaborative filtering. It evaluates the accuracy of a system by compare the numerical recommendation scores against the actual user ratings for the user-item pairs in the test dataset. We assume that *{p1,p2….pm}*is the predicted ratings for the super user and {q1,q2…qm}is the actual ratings of the super user and the MAE metrics is formulated as:

$$MAE = \frac{\sum_{i=1}^{m} |p_i - q_i|}{M} \qquad (5)$$

Experimentation of sensitivity of neighborhood size

We performed an experimentation to determine the sensitivity of the neighborhood size. We can see that in fig(3) the sensitivity of the neighborhood size has a big impact of the quality of the prediction. We can see that after neighborhood size is greater than 35, the accuracy of the prediction has been increased.

Table 3: MAE values of Neighbors

No of neighbors	MAE
5	0.675
10	0.687
15	0.541
20	0.359
25	0.519
30	0.506
35	0.302
40	0.285
45	0.211
50	0.182

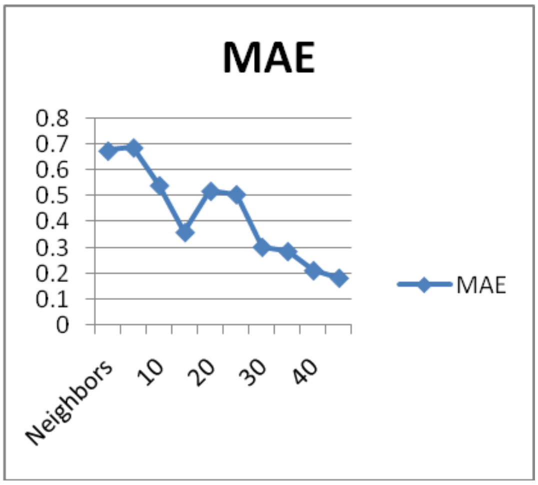

Figure.3. Sensitivity of the neighborhood size

9. CONCLUSION

In this paper, we have proposed a framework for collaborative filtering using divisive hierarchical clustering approach. We partitioned the users based on their neighborhood similarity and the size of the cluster. Experimental results show that the proposed framework can significantly improve the accuracy of prediction as well as solve the scalability problem. The future enhancement will concentrate on agglomerative hiererchical clustering in collaborative filtering approach

REFERENCES

[1]Chee_J Han_K. Wang "_Rectree: An efficient collaborative filtering method."Lecture Notes in Computer Science, volume no:2114, 2001 pages: 141-145

[2]Chris Ding and Xiaofeng He" Cluster merging and splitting in hierarchical clustering algorithms ", Second IEEE International Conference on Data Mining (ICDM'02), 2002 pages:139-146

[3]Feng Zhang, Hui-you Chang" A Collaborative Filtering Algorithm Employing Genetic Clustering To Ameliorate The Scalability Issue" Proceedings of the IEEE International Conference on e-Business Engineering 2006 pages:231-236

[4] Gabor Takacs, Istvan Pilaszy, Bottyan Nemeth" Scalable Collaborative Filtering approaches for Large Recommender Systems" journal of Machine Learning Research 2009vol no:10 pages. 623-656

[5]George, T., & Merugu, S." A scalable collaborative filtering framework based on co-clustering." Proceedings of the Fifth IEEE International Conference on Data Mining 2005 pages:625 - 628

[6] Goldberg, D. Nichols, B. M. Oki, and D. Terry." Using collaborative filtering to weave an information tapestry".Communications of the ACM, 1992 volno35, issue:12 pages:61–70

[7] Kelleher.J and D. Bridge.:" Rectree :An accurate, scalable collaborative recommender". In Procs. of the Fourteenth Irish conference on Artificial Intelligence and Cognitive Science, 2003 pages: 89–94,

[8] Liang ZhanG, bo Xiao, Jun guo, Chen zhu " A scalable collaborative filtering algorithm based on localized preference" Proceedings of the Seventh International Conference on Machine Learning and Cybernetics, Kunming, July 2008 pages:12-15

[9] Liu Hongmin Yin Zhixi "Applying Multiple Agents To Fuzzy Collaborative Filtering" International Conference On E-Business And Information System Security, 2009. Ebiss '09 pages1-5

[10] Marie Chavent" A monothetic clustering method" *Pattern Recognition Letters*, Volume 19, Issue 11,1998, pages 989-996

[11] Panagiotis Symeonidis, Alexandros Nanopoulos, Apostolos N. Papadopoulos, Yannis Manolopoulos. "Nearest-biclusters collaborative filtering based on constant and coherent values" information retrieval 2008 Volume 11 , Issue 1 Pages: 51 - 75

[12]Resnick, N. Iacovou, M. Suchak, P. Bergstorm, and J. Riedl. “Grouplens: An open architecture for collaborative filtering of netnews”. From Proceedings of ACM 1994 Conference on Computer Supported Cooperative Work, Chapel Hill, NC: pages 175-186.

[13] Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. Analysis of Recommender Algorithms for E-Commerce. Proceedings of the 2nd ACM conference on Electronic commerce Pages: 158 - 167

[14] Sarwar.B, G. Karypis, J. Konstan and J. Riedl,” Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering”, Proceedings of the Fifth International Conference on Computer and Information Technology, 2002

[15]Sergio M. Savaresi,, Daniel L. Boley, Sergio Bittanti and Giovanna Gazzaniga” Cluster selection in divisive clustering algorithms”, SIAM Internation Conference on Data Mining 2002 pages:299-314

[16] SongJie Gong, HongWu Ye, XiaoMing Zhu” Item-Based Collaborative Filtering Recommendation using Self-Organizing Map” in proceedings of the 21st annual international conference on chinese control and decision conference 2009 pages:4065-4067

[17] Songjie Gong, Hongwu Ye” Joining User Clustering And Item Based Collaborative Filtering In Personalized Recommendation Service” in proceeding of international conference on industrial and information systems 2009 pages: 149-151

[18] Xue, G., Lin, C., Yang, Q,”Scalable collaborative filtering using cluster-based smoothing” In Proceedings of the ACM SIGIR Conference 2005 pages:114–121

[19]www.grouplens.org