A Imputed Neighborhood based Collaborative Filtering System for Web Personalization

Suresh Joseph. K
Assistant Professor
Department of Computer Science
Pondicherry University,
Puducherry, India

Ravichandran.T
Principal
Hindusthan Institute of Technology
Coimbatore,
Tamil Nadu, India

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 Imputed divisive hierarchical clustering approach to overcome this scalability issue when more number of users increases in terms of neighborhood size.

Keywords

Collaborative Filtering, Imputation Recommender Systems, Divisive Hierarchical Clustering, Web Personalization.

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 E-commerce 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. Web personalization is to help the user with customized relevant information. Recommender systems are used to improve the business intelligence in an E-commerce site. 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.

Impute means "to set propose something to other person", this model is based on INCF system for Web Recommendation. Collaborative filtering (CF) is one of the most popular technologies in recommender systems, and are widely used in many personalized recommender applications, such as E – Tourism, E-commerce, Virtual Book store, and other web based applications news sites, and so on. The underlying assumption of the CF approach is that those who agreed in the past tend to agree again in the mere future. The need for the web personalization includes Sphere Sovereignty and Discretionary Control, Saves time and cost, Better information, Addresses ongoing needs and challenges, or opportunities. INCF algorithm suggests the user using in accordance to the dynamic change. The paper is organized as follows: section1 portrays about the introduction and the Background of Filtering and recommender system. Section2 describes about the related works in collaborative filtering approach based on clustering techniques. Section3 describes about the problem definition. Section4 describes about the proposed system. Section5 describes about the divisive hierarchical clustering approach. Section6 describes about the experimental evaluation and Section7 outlines the summary.

1.1 Background

1.1.1 User-Based Collaborative Filtering

User based collaborative filtering technique main task is to find users that are similar to an active user. The users similar to a active user are called as active 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 a super user '*u',* compute the similar users from all the users data based on the similarity or their respective rating functions. 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.

1.1.2 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

1.1.3 Content Based Recommendation Approach

Content-based recommendation is an outgrowth and continuation of information filtering research. In a content-based system, the objects of interest are defined by their associated features. The content based approach profiles each user or product, allowing programs to associate users with matching products. Of course, content based strategies require gathering external information that might not be available or easy to collect including those that make use of case-based reasoning or text classification methods. A content-based recommender learns a profile of the user's interests based on the features present in objects the user has rated.

The content based recommendation approach has the following disadvantages:

- Content based can't perform in domain where there is not much content associated with items, or where the content is difficult for a computer to analyze.
- The system can only suggest items whose content match with the user's profile.

1.1.4 Hybrid Recommendation

Hybrid Recommendation approach combines collaborative filtering and content filtering methods and uses the benefits of both of them for specifying and recommending suitable items. Advanced recommender systems use hybrid systems that switch between different modules according to circumstances in order to produce better results. There are many different possibilities how to combine these modules. The most interesting examples are weighted, switching or cascade hybrid systems. A switching hybrid system works by using only one recommendation module at a time. At the beginning of the process, the system determines which module is best suited for the task and then generates recommendations using this module. This is very useful if the first recommendation module produces a very large list of possible items for suggestion.

1.1.5 Knowledge-Based Recommendation

Knowledge-based recommendation attempts to suggest objects based on inferences about a user's needs and preferences. In some sense, all recommendation techniques could be described as doing some kind of inference. Knowledge-based approaches are distinguished in that they have functional knowledge: they have knowledge about how a particular item meets a particular user need, and can therefore reason about the relationship between a need and a possible recommendation. The user profile can be any knowledge structure that supports this inference. In the simplest case, as in Google, it may simply be the query that the user has formulated. In others, it may be a more detailed representation of the user's needs. The Entree system and several other recent systems employ techniques from case-based reasoning for knowledge based recommendation.

1.1.6 Rule-based Filtering

Rule-based Filtering deliver the content to their users based on the rules. These rules are set by the system depends on what were allowed their users to do or have. More sophisticated way, the web system kept user history of their purchased. Based on this rules, web system are only allowed recommend the very similar items. The advantages of this technique were system do not have to predict out of the users need. The rules are set and users are expecting to follow those rules. It was good for surveying to obtain users profile based on demographics. Data collection will be easier to collect and analyze. The limitations of this technique are inflexibility and limiting what users can do. This technique is not really useful for E-commerce sites because users already know what type of content their will received. Ruled-based recommendation did not offer more personalize recommendation compare to others. Rule-based Filtering System are useful for insurance companies which allows them to quickly give quotes to their customer based on web or phone after filling particular information.

1.2 Collaborative Filtering Approach

Collaborative filtering (CF) is the process of filtering for information or data patterns using techniques involving collaboration. Collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting variety of information from many users (collaborating). The underlying assumption of the CF approach is that those who agreed in the past tend to agree again in the mere future. The need for the web personalization includes Sphere Sovereignty and Discretionary Control, Saves time and cost, Better information, Addresses ongoing needs and challenges, or opportunities.

In this paper, the Web recommender system uses imputed neighborhood based a collaborative filtering (INCF) algorithm which imputes the user rating data using an imputation technique, before using a traditional Pearson correlation-based algorithm on the resulting imputed data of the most similar neighbors. Collaborative filtering is used extensively on E-Tourism, E- learning , E- Commerce, social networking sites and sites which provide tools like enterprise bookmarking, in which users and promote links to sites they find interesting.

1.2.1 Imputed Nearest Neighborhood CF (INN-CF)

Imputation is the substitution of some value for a missing data point or a missing component of a data point. Imputation is not the only method available for handling missing data. Imputation is the replacement of missing values in data with estimates using techniques and assumptions. Imputed nearest neighborhood *CF* (*INN-CF*), which first finds the users most similar to the active user (for which we are making predictions), then uses the corresponding imputed data to make predictions for the active user

2. RELATED WORKS

In this section we will discuss related works in collaborative filtering approach based on clustering techniques. Songjie Gong, Hongwu Ye has proposed [10] joining user clustering and item based collaborative filtering in personalized recommendation service. Liu Hongmin et.al, Yin Zhixiet.al proposed [4] Applying multiple agents to Fuzzy collaborative filtering. They had developed a new method for recommending items using multiple agents. 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.Bi clustering Thomas George et al, Srujana Merugu et al proposed [3] 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. SongJie Gong et al, HongWu Ye et al, XiaoMing Zhu et al proposed [16] Item-Based Collaborative Filtering Recommendation using Self-Organizing Map.. Jerome Kelleher et al and Derek Bridge et al [1] proposed RecTree Centroid: An Accurate, Scalable Collaborative Recommender.

3. PROBLEM DEFINTION

The challenge in Imputed neighborhood 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 using imputation. The main advantage of using this divisive hierarchical clustering approach is that it is more suitable for making real time recommendations

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

4. PROPOSED 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.

Nearest-neighbor algorithms compute the distance between users based on their preference history which is stored. Distances vary greatly based on domain, number of users, number of recommendations, and degree of co-rating between users. Predictions of choices are computed by taking the weighted average of the opinions of a set of neighbors for that particular page. The search in nearest neighbor involves time complexity in the case of large database. This can be solved using heuristics to search for a good neighbor and may use opportunistic sampling with large data sets. Nearest Neighborhood recommender system provides high level of personalization when compared with correlation based recommenders.

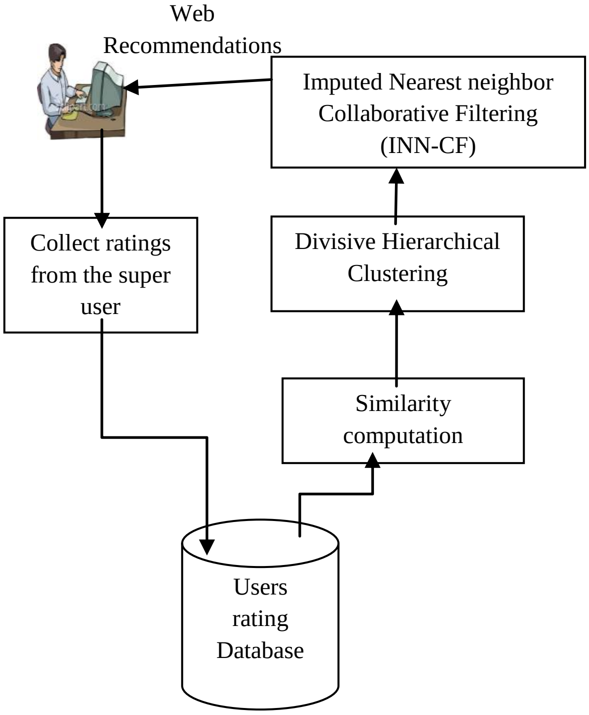
Web Recommendations
Imputed Nearest neighbor Collaborative Filtering (INN-CF)
Collect ratings from the super user
Divisive Hierarchical Clustering
Similarity computation
Users rating Database

Figure.1. Proposed Architecture for CF Web Recommendation system

The Pearson correlation-based CF (Pearson CF) algorithm is a representative memory-based CF algorithm, which calculates similarities between each pair of items rated by a user, or between each pair of users who rate the same item. This allows these systems to scale effectively with the number of users and items, as this computation involves only a small percentage of the total number of items and users – i.e., due to the sparsity. One shortcoming of Pearson CF is that its predictive performance degrades quickly as the data becomes sparser.

Neighborhood-based CF makes CF predictions using Pearson CF based on the nearest neighbors instead of the whole dataset. 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.

The similarity values are computed for the super user and then 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.

Table 2: Distance Matrix

Similarity	User			
	2	3	4	5
Super user	0.3	0.5	0.8	0.7

4.1 Predicted Rating

Now 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

5. DIVISIVE HIERERCHICAL CLUSTERING

The divisive algorithm is a top down approach. It is based on repeated cluster bi-sectioning 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.

A partition of is a list *(C1... CK)* of clusters verifying C1 $\cup$ . . . $\cup C_K$ = and $C_k \cap C_{k'} = \Omega$; 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

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

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. 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 divisions are stopped after a number L of iterations and L is given as input by the user, usually interested in few clusters partitions. 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.

6. EXPERIMENTAL EVALUATION

We used Movie lens data set for evaluating our experiment. 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. 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:

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

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

Table 3: MAE values of Neighbors

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.

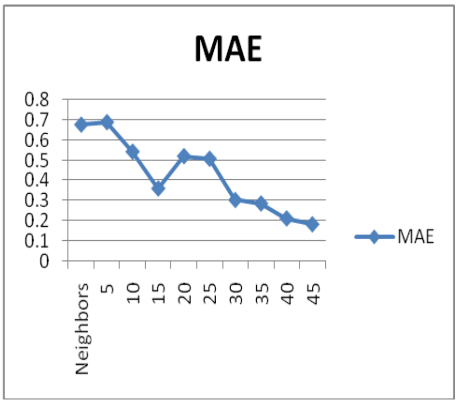

Figure.3. Sensitivity of the neighborhood size

6.1 Comparison of Imputed Divisive approach MAE with the Collaborative filtering MAE

The values of this approach is compared with the traditional CF algorithm and found that our imputed divisive approach is better in terms of prediction quality.

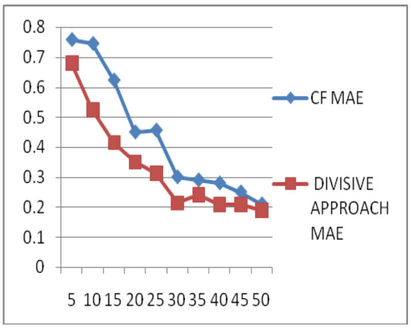

Figure.4. MAE of CF and Imputed Divisive Approach MAE

7. CONCLUSION

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

8. REFERENCES

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

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

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

[4] 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

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

[6] 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.

[7] 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, pp158 - 167

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

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

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

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

[13] Xiaoyuan Su, Taghi M. Khoshgoftaar, Russell Greiner, "Imputed Neighborhood Based Collaborative Filtering" 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[14] A novel approach for selection of the users' preferred websites, International Journal of Business Intelligence and Data Mining archive Volume 2 Issue 3, October 2007 Inderscience Publishers Inderscience Publishers.

[15] Web Personalisation through Incremental Individual Profiling and Support-based User Segmentation, IEEE/WIC/ACM International Conference on Nov. 2007.

AUTHORS BIOGRAPHY

Mr. K. Suresh Joseph received his B.E. degree from Bharathiyar University and M.E. from University of Madras, in 2002. Since 2006, he has been an Assistant Professor in the Department of Computer Science, Pondicherry University, Pondicherry, India.

Professor Dr.T.Ravichandran received the B.E degree from Bharathiar University, Tamilnadu, India and M.E degrees from Madurai Kamaraj University, Tamilnadu, India in 1994 and 1997 respectively; He received his PhD degree from the Periyar University, Salem, India. He is currently the Principal of Hindustan Institute of Technology, Coimbatore, Tamilnadu, India. His research interest includes theory and practical issues of building distributed systems, Internet computing and security. He is a member of the IEEE, CSI and ISTE.