Recommendation on Bundling the Items using Item based Collaborative Filtering TechniquE

S.Saranya
Post Graduate Student - CSE
Dr.Mahalingam College of Engineering and Technology, Coimbatore.

N.Gobi
Assistant Professor (SS) - CSE
Dr.Mahalingam College of Engineering and Technology, Coimbatore.

ABSTRACT

Recommender System (RS) is a personalized information filtering technique used to provide personalized recommendations of products or services to the users. The goal of the RS is to obtain ratings for items(such as music, books, or movies) from the users and based on the result, the system will predict ratings for each item and suggests interesting items to the users. Collaborative filtering technique is widely used recommendation algorithm that predicts item ratings by considering the users with similar preferences (i.e., "neighbors") who liked in the past. Recommending the highly rated items can improve the accuracy but it does not provide more diverse recommendations. The previous works was modeled using optimization algorithms for recommending bundled items. i.e., set of items in packages but does not provide more efficient recommendations. In the proposed system, a ranking algorithm along with collaborative filtering technique is used to provide different set of composite package of items to improve both the accuracy and diversity. The empirical evaluation of this proposed technique will be implemented based on real-world datasets to providedifferent set of composite packages to all theusers.

I. INTRODUCTION

Recommendation system helps user to find and select items from the huge number of available resources on the web or in other electronic information sources. Given a large set of items and its description given by the user, the RS will present a small set of items to the end user that are well suited to their description.

Commonly, Recommender Systems are classified into four categories: collaborative filtering, content-based filtering, knowledge based systems and hybrid recommendation approaches. Recommender systems became an important research area after the appearance of the collaborative filtering since the mid-1990s and also plays a major role in current recent years. The interest in this area still remains high because it constitutes a problem rich research area and because of the abundance of practical applications the recommender system will help users to deal with information overloads and provide personalized recommendations, content and services to them. Examples of such applications include recommending books, CDs and other products from the websiteslike Amazon.com, movielens.com, Netflix.com etc.

Theclassical Recommendation System provides recommendations consisting of single items, e.g., books or DVDs. These techniques help to improve the accuracy of the system. However, accuracy alone is not enough to measure the quality of the system. The importance of diverse measure has to be considered. This is the major challenge focused in the recommender system. To obtain this challenge various re-ranking approaches has been proposed but still there is a trade-off between accuracy and diversityi.e., the system must able to recommend both the popular items (highly rated items) and long tail items (minimum rated items) to the users. Some recent studies showed that several applications can benefit from a system capable of recommending packages of items, in the form of sets. This helps to improvethe quality ofthe recommender system. In the existing system algorithms like Approximation, Clustering, Greedy, Histograms are used along with filtering techniques for packing the items.

II. RELATED WORK

There are various techniques available in recommender system. The most widely used technique is collaborative filtering technique. In general, collaborative filtering is the process of filtering for information or patterns using techniques involving collaboration among multiple agents, viewpoints, data sources, etc. Applications of collaborative filtering typically involve very large data sets. Collaborative filtering methods have been applied to many different kinds of data including: sensing and monitoring data, financial data, electronic commerce and web applications etc where the focus is on user data.

In the narrower sense, collaborative filtering is a method of making automatic predictions (filtering) about the interests of a user by collecting preferences or taste information from many users (collaborating). The underlying assumption of the collaborative filtering approach is that if a person A has the same opinion as a person B on an issue, A is more likely to have

B's opinion on a different issue x than to have the opinion on x of a person chosen randomly.

The Collaborative filtering approach can be divided in two categories:

1. Memory-based
2. Model-based

A Memory-based Collaborative filtering algorithms use the entire or a sample of theuser-item rating matrix to generate a prediction. Every user is part of a group of people with similar interests. By identifying those neighbors of an active user, a prediction of preferences on items for the active user can be produced. A memory-based Collaborative filtering algorithm includes user–based and item-based collaborative filtering methods.

Model-based recommendation systems involve building a model based on the dataset of ratings. In other words, information is extracted from the dataset and can use a "model" to make recommendations without having to use the complete dataset every time. Model-basedcollaborative filtering algorithms include Bayesian models (probabilistic), Clustering Models and Matrix Factorization methods.

Zhang and Hurley (2008) [1] proposed a method called Quadratic Programming (QP) which is a special type of mathematical optimizationproblem. QP is the problem of optimizing (minimizing or maximizing) a quadratic function of several variables subject to linear constraints on these variables. The optimization function of an objective function under certain constraints provides feasible solutions and is represented in terms of binary vectors. The solution was the choice of quadratic relaxation otherwise the problem itself becomes a hard problem. The spectral relaxation method was proposed to solve the problem.

Park, Joo and Tuzhilin (2008) [2] discussed the problem of recommending long tail items to the users because many long tail items had only few ratings. This becomes hard for the recommender system to recommend those items. To solve this problem, they proposed Clustered Tail (CT) method to split the whole item set into head and tail parts and do clustering only on the tail items because it consists of only long tail items. Then the tail part of items is recommended based on the ratings of the clusters and the head items based on the ratings of individual items.

Nguyen and Tho (2009) [4] worked on large dataset of Netflix's movies to improve recommender systems by incorporating confidence interval and genres of movies. This new approach enhances the performance and quality of service of recommender systems and gives better result than Netflix commercial recommender system, Cinematch.

The content-based algorithms use user's profile to find matching items for the user.For example, a 23 year old user based on the algorithm will select all items which are interested by their age. It can also use item's profile to recommend item to user. The various set of user and item's profiles are difficult to collect and it should need an external source to collect these profiles.

The user-based and item based collaborative filtering techniques are used for calculating the similarity between the users and items in recommender systems. There are various algorithms used in this technique and one of most popular algorithm used in CF technique is Pearson correlation algorithm.

G. Karypis (2001) [6] proposed item based recommendation algorithm is used to determine similarity between items and to combine these similarities to recommend a basket of items. The cosine-based similarity is used to find the similarity between two items is computed by treating any two items in terms of vectors in the space of users or customers and use cosine function between these vectors as a measure of similarity. The conditional probability-based similarity is an alternative way of computing the similarity between pairs of items u and v is based on conditional probability where assuming that purchasing one of the items given that all other items have been already purchased.But these two methods have some limitations where each item v should have high conditional probability.

Khabbaz, Xie, and Laks(2011) [3] proposed an efficient and scalable itemrecommendation engine which recommends top-k interesting items, and an efficient package recommendation engine which recommends top-k interesting packages which satisfy all the user specified constraints. The two phase algorithm is used for finding top-k interesting items by maintaining optimal threshold. The instance optimal and greedy algorithm is used for packing the items in the form of sets.

Adomavicius and Kwon (2012)[7][12] introduced a number of item re-ranking approaches to generate more diverse recommendations and also to maintain comparable level of accuracy. So the authors proposed various re-ranking approaches working along with filtering techniques will improve the diversity and they are Item popularity approach, reverse predicted rating value, Item average rating, Item absolute likeability, Item relative likeability, Item rating variance, Neighbor's rating variance. These techniques have been worked for individual items.

In summary, the goal of the proposed ranking approachesis to improve the diversity of recommendations. However, there is a potentialtradeoffbetween recommendation accuracy and diversity.Thus, in this paper, we aim to find techniques that canimprove both accuracy and diversity of recommendations.

III. EXISTING WORK

The main objective of the existing system[5] is to recommend different set ofcomposite packages to the users. Each item is associated with a value (rating or score) and a cost, then the user specifies a maximum total cost (budget) for recommending the set of items.

A. SIMILARITY MEASURE:

The user-based collaborative filtering (CF) technique was applied to predict the unknown ratings for each item by finding similarity between two users. The algorithm used for the prediction is Pearson correlation algorithm which finds the users who have similar tastes.

The similarity can be calculated by

$$sim(x,y)=\frac{\sum_{s\varepsilon S}(r_{x,s}-\overline{r_x})(r_{y,s}-\overline{r_y})}{\sqrt{\sum_{s\varepsilon S}(r_{x,s}-\bar{r}_x)^2}\sqrt{\sum_{s\varepsilon S}(r_{y,s}-\bar{r}_y)^2}}$$

Where $r_{x,s,}$ $r_{y,s}$ - An active user x & y who rated an item s.

$\bar{r}_x$ - The average rating of the co-rating of user x.

$\bar{r}_y$ - The average rating of the co-rating of user y.

After this process, the prediction for the active user for an item is implemented by the weighted average of all the ratings.

The main drawback of User –based CF technique is scalability problem because the computation will become complex when the number of items grows bigger and the performance of the technique is slow.

B. APPROXIMATION ALGORITHMS:

The authors proposed different types of approximation algorithm for recommending top-k packages to the user. Generally, the approximation algorithms are mainly used to find approximate solution to theoptimization problem. The three different approximation algorithms used in the existing system are Instance optimal algorithm (Inst-Opt-CR) which is also called as 2-approximation algorithm, Greedy algorithm and Histogram optimization algorithm.

The Instance optimal algorithm is used to access the items only in non-increasing order of their value because of the huge size of the sets of items and provides high cost for retrieving item information from the source. So it is crucial
for an algorithm to find high-quality solutions while minimizing the number of items accessed. Even the instance optimal algorithm provides optimal solutions but it has some issues. The computational cost for accessing the items was
high and it does not provide the quality of the recommendation.

In Histogram-based optimization method, the costs of items are represented in terms of histograms and it will be divided into non-overlapping buckets. Each bucket will store the number of items whose cost falls inside the specific cost range. These two methods can able to provide optimal solution but the quality of the system is not feasible and also provides high computational cost.

Greedy algorithm is a simple and efficient algorithm that generates high qualitypackages from the list of items. The items are sorted based on value/cost ratio.The cost budget is assigned by the users and based on the cost budget the system will check for the best quality items i.e., the selected package value should be greater than or equal to threshold value and must be less than or equal to cost budget.If the package satisfies the condition then the package is recommended or else the system has to look for next set of items. This iteration is repeated with different set of threshold values.

For each dataset, the quality of the top-5 composite recommendations returned by theapproximation algorithms was measured by the aggregated value of each package andthe average item value of each package.The techniques used in existing system improve the accuracy and diversity but still thediversification of the system has to be improved with more efficient algorithms.

IV. PROPOSED SYSTEM

In the proposed system, additionally a re- ranking approach is proposed which will help to improve the diversity. Greedy algorithm is incorporated from the existing system to increment the quality of the recommender system.A generic design has been developed so that various types of datasets can be used for recommending the products on web.

A. SIMILARITY MEASURE:

A RS normally focuses on a specific type of item (e.g., CDs, or news) and also its design, graphical user interface and the core recommendation techniques will generate the customized recommendations to provide useful and effective suggestions for the specific type of item.

Instead of performing computation between two users, the item-basedcollaborative filtering technique is proposed to predict similarity between two items. The item-based collaborative filtering technique is proposed to predict similarity between two items. The advantage of item-based algorithm is the computation is much simpler and more scalability than user-based algorithm. The Pearson correlation algorithm is used which finds the no. of users who have rated the items i and j and also both. The similarity is calculated by

$$sim(i,j)=\frac{\sum_{u\varepsilon U}(R_{u,i}-\overline{R_i})(R_{u,j}-\overline{R_j})}{\sqrt{\sum_{u\varepsilon U}(R_{u,i}-\overline{R_i})^2}\sqrt{\sum_{u\varepsilon U}(R_{u,j}-\overline{R_j})^2}}$$

Where, $R_{u,i}$ - An active user u who rated an item j.

$\bar{R}_i$ - The average rating of the co-rating users on the i-th item.

$\bar{R}_j$ - The average rating of the co-rating users on the j-th item.

After this process, the prediction for the active user for an item is implemented by the weighted average of all the ratings.

The formula is expressed by

$$P_{a,x}=\bar{r}_b+\frac{\sum_{i\varepsilon I}(r_{a,i}-\bar{r}_i).w_{x,i}}{\sum_{i\varepsilon I}|\,w_{x,i}\,|}$$

Where $r_x$and r_i are the average ratings for item b and item ithat all other users have rated.
$W_{x,i}$is the Pearson correlation between items b and i.
i € I - summations over all the items that the user a has rated and an item i should belongs to list of items i.

For example, to predict rating for an average user A for an item 'x' is calculated by summing the average ratings for item 'x' and item 'i' that all other users who rated those items
and the Pearson correlation between items 'x' and 'i', This process was repeated until the ratings for each item is predicted.

B. ITEM AVERAGE RATING FUNCTION

With the list of predicted items, a threshold value is assigned. For example, consider the threshold rating value to be 3.5. The items which falls above or equal to 3.5 is said to be highly rated items and the items which falls below or equal to 3.5 is said to be long tail items. In order to include long tail items in recommendation, an average item ranking approach is used to re–ranking the items according to average ratings of all known users for each item. So that the item ratings in the list is re-ranked and sorted in descending order. The formula used for calculating average item rating is

$$rank_{AvgRating}(i)=\overline{R(i)},$$
$$where\,\overline{R(i)}=\frac{1}{|U(i)|}\sum_{u\in U(i)}R(u,i)$$

Where | U(i)| - No.of users who rated an item i.

∑ R (u,i) – summing the ratings of each users who rated for an item i.

This method helps the recommender system to include long tails items into popular items.

C.GREEDY ALGORITHM

The working of the greedy algorithm is it sorts the items in descending order based on value/cost ratio. A threshold value is fixed by using minimum value of the item in the item set and minimum cost of the items in the item set (v_{min} /c_{min}).

The user will fix the cost budget to the system. The system will find the set of items in terms of subset and checks whether the subset value is greater than or equal to the threshold and less than or equal to cost budget.

If this condition is satisfied, the system will recommended the subset to the user otherwise the system will move to next subset of items to satisfy the condition. If the upper bound cost of the item i.e., the maximum cost from the list of items satisfies the condition then that item
is recommended or else the system will increment to next item with maximum cost and condition is checked. Based on this process the packages of items are done.

V. PERFORMANCE EVALUATION

The performance evaluation of the proposed system is between accuracy and diversity. The recommendation accuracy is measured based on the percentage of truly "highly ranked" ratings. i.e., items that were predicted to be the N most relevant "highly ranked" items for each user[7].

The recommendation diversityis measured using the total number of distinct items recommended across all users as an aggregate diversity which is referred as diversity-in-top-N. By using these techniques
and measures the performance of the recommended system will be improved.

VI. CONCLUSION

In this paper we have proposed Item average ranking approach combining with filtering techniques and greedy algorithm which is expected to improve the quality of the recommendation system.

REFERENCES

[1] Zhang and Hurley (2008),"Avoiding monotony: improving the diversity of recommendation lists", Proceedings of the 2008 ACM conference on Recommender systems, pp.123-130.

[2] Park, Joo and Tuzhilin (2008),"The Long tail of recommender systems and how to leverage it", Proceedings of the 2008 ACM conference on Recommender systems, pp.11-18.

[3] Khabbaz, Xie, and Laks (2011),"TopRecs: Pushing the Envelope on Recommender Systems", Data Engineering, pp.61.

[4] Nguyen and Tho (2009), "Web based Recommender Systems and Rating Prediction".

[5]Xie, Laks and T.Wood (2012), "Composite recommendations: from items to packages”, Frontiers of Computer Science, pp.264-277.

[6] Sarwar and Badrul, et al (2001), "Item-based collaborative filtering recommendation algorithms", Proceedings of the 10th international conference on World Wide Web , ACM, pp.285-295.

[7] Adomavicius. G and Y. Kwon (2012),“Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques”, IEEE Transactions on Knowledge and Data Engineering , vol.24, pp.896-911.

[8] Schafer .J, et al. (2007), "Collaborative filtering recommender systems", The adaptive web, ACM, pp.291-324.

[9] Su, M. Khoshgoftaar and Russell Greiner (2008),"Imputed neighborhood based collaborative filtering" , Web Intelligence and Intelligent Agent Technology, WI-IAT'08. IEEE/WIC/ACM International Conference on Vol.1.

[10] Anderson, Chris and ManishaHiralall (2011), "Recommender systems for e-shops".

[11] Adomavicius and Y. Kwon. (2009), "Toward more diverse recommendations: Item re-ranking methods for recommender systems”, Workshop on Information Technologies and Systems.

[12] Adomavicius and Kwon (2007), "New recommendation techniques for multi criteria rating systems", Intelligent Systems, pp.48-55.