ISSN: 2278 – 1323
International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)
Volume 6, Issue 1, Jan 2017

An Overview for Subgroup Analysis of a Recommendation System

N. S. Barley, Prof. R. R. Keole

***Abstract*-- Mostly commercial systems are based on Collaborative Filtering (CF). Collaborative Filtering (CF) is an effective and widely adopted recommendation approach. Different from content-based recommender systems which rely on the profiles of users and items for predictions, CF approaches make predictions by only utilizing the user-item interaction information such as transaction history or item satisfaction expressed in ratings, etc. In this study, we develop an efficient collaborative filtering method, called *RecTree* (which stands for RECommendation Tree) that addresses the scalability problem with a divide-and-conquer approach. A novel product recommendation method called TCRec is also studied, which takes advantage of consumer rating history record, social-trust network and product category information simultaneously. Motivated by the observation, in this paper, we studied a novel Domain-sensitive Recommendation (DsRec) algorithm, to make the rating prediction by exploring the user-item subgroup analysis simultaneously.**

Index Terms- **Collaborative Filtering, recommender systems, user-item subgroup.**

I. INTRODUCTION

Collaborative Filtering (CF) which is an effective recommendation approach with fundamental assumption that two users have similar tastes on one item if they have rated other items similarly. Due to the collaboration effects, CF only relies on users' history behaviors without collecting content information for privacy, CF systems become increasingly popular, since they do not require users to explicitly state their personal all users and items. In *RecTree* the method first performs an efficient k-means-like clustering to group data and creates neighborhood of similar users, and then performs subsequent clustering based on smaller, partitioned databases. Instead if users' interested domains are captured first, the recommender system is more likely to provide the enjoyed items while filter out those uninterested ones. Therefore, it is necessary to learn preference profiles from the correlated domains instead of the entire user-item matrix. In DsRec a user-item subgroup is deemed as a domain consisting of a subset of items with similar attributes and a subset of users who have interests in these items. The proposed framework of DsRec includes three components: a matrix factorization model for the observed rating reconstruction, a bi-clustering model for the user-item subgroup analysis, and two regularization terms to connect the above two components into a unified formulation. This short paper reports on work in progress related to applying data partitioning/clustering algorithms to ratings data in collaborative filtering. We use existing data partitioning and clustering algorithms to partition the set of items based on user rating data.

Collaborative filtering approaches can be classified into two main categories: model-based approaches and memory based approaches. Model-based approaches learn user/item rating patterns to build statistical models that provide rating estimations. Memory-based approaches, on the other hand, compute user/item similarities based on distance and correlation metrics, and use these similarities to find similar-minded people of the active user.

II. BACKGROUND

In this section, we review several major approaches for collaborative filtering.

A.. Collaborative Filtering

1) Memory-based Approaches

The memory-based approaches are among the most popular prediction techniques in collaborative filtering. The basic idea is to compute the active user's predicted vote of an item as weighted average of votes by other similar users or *K* nearest neighbors (KNN). Two commonly used memory-based algorithms are the Pearson Correlation Coefficient

Manuscript received Jan, 2017.
N. S. Barley, *P.G. Dept.of Computre Science and Information Technology, H.V.P.M's C.O.E.T. Amravati, Maharashtra, INDIA .*
Asst. Prof. R. R. Keole *,P.G. Dept.of Computre Science and Information Technology, H.V.P.M's C.O.E.T. Amravati, Maharashtra ,INDIA*

(PCC) algorithm and the Vector Space Similarity (VSS) algorithm. These two approaches differ in the computation of similarity. The PCC algorithm generally achieves higher performance than vector-space similarity method.

2) Model-based Approaches

Two popular model-based algorithms are the clustering for collaborative filtering [1] and the aspect models. *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 the degree of participation. The *aspect model* is a probabilistic latent-space model, which considers individual preferences as a convex combination of preference factors. The latent class variable is associated with each observation pair of a user and an item. The aspect model assumes that users and items are independent from each other given the latent class variable.

3) Hybrid Model

Pennock[6] et al. proposed a hybrid memory- and model-based approach. Given a user's preferences for some items, they compute the probability that a user belongs to the same "personality diagnosis" by assigning the missing rating as a uniform distribution over all possible ratings. Previous empirical studies have shown that the method is able to outperform several other approaches for collaborative filtering, including the PCC method, the VSS method and the Bayesian network approach. However, the method neither takes the whole aggregated information of the training database into account nor considers the diversity among users when rating the non-rated items. From our point of view, the clustering-based smoothing could provide more representative information for the rating

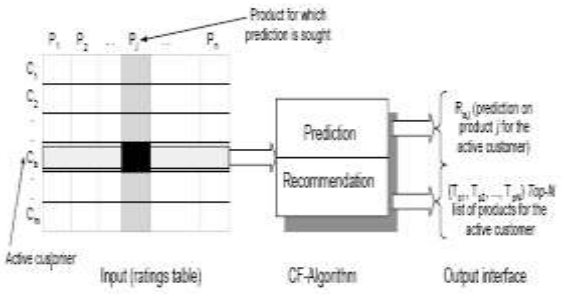

Figure 1: The Collaborative Filtering Process.

Fig.1 shows the schematic diagram of the collaborative filtering process. CF algorithms represent the entire $m \times n$ customer-product data as a ratings matrix, A. Each entry $a_{i,j}$ in A represent the preference score (ratings) of the ith customer on the jth product. Each individual rating is within a numerical scale and it can as well be 0, indicating that the customer has not yet rated that product.

III. THE *RECTREE* ALGORITHM

RecTree is the acronym for a new data structure and collaborative filtering algorithm called the RECommendation Tree. The *RecTree* algorithm [4] partitions the data into cliques of approximately similar users by recursively splitting the dataset into child clusters. Splits are chosen such that the intra-partition similarity between users is maximized while the inter-partition similarity is minimized. This yields relatively small cohesive neighborhoods that *RecTree* uses to restrict its search for advisors – which represent the bottleneck in memory-based algorithms. *RecTree* achieves its O(*n*log2(*n*)) scale-up by creating more partitions to accommodate larger datasets - essentially scaling by the number of partitions rather than the number of users. Prediction accuracy deteriorates when a large number of lowly correlated users contribute to a prediction. Herlocker et al. [8] suggest that a multitude of poor advisors can dilute the influence of good advisors on computed recommendations. The high intra-partition similarity between users makes *RecTree* less susceptible to this dilution effect – yielding a higher overall accuracy. The chain of intermediate clusters leading from the initial dataset to the final partitioning is maintained in the *RecTree* data structure, which resembles a binary tree. Within each leaf node, computing a similarity matrix between all members of that clique identifies advisors. *RecTree* then generates predictions by taking a weighted deviation from each clique's advisor ratings using.

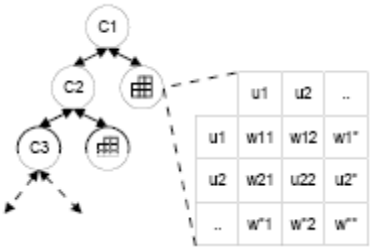

Figure 2: The *RecTree* data structure.

IV. TCREC

To achieve the ultimate target, TCRec[3] involves three components to exploit the consumer rating history record, the social-trust network and the product category information, respectively

A. Rating History Record

Normally, there are a large number of products, and a customer may only rates a small portion of the whole item set practically.

B. Social-trust Network

We assume that customer's interest to products is influenced by the other linked customers in social-trust network.

C. Product Category Information

The products in different categories should be discriminated by exploring the product-specific latent representations.

The central idea of TCRec is that the tastes of customers who trust each other are similar and the latent features of products in one category can be discriminated from others. The experimental results on two real-world datasets indicate the effectiveness of our approach.

V. DOMAIN-SENSITIVE RECOMMENDATION (DSREC) ALGORITHM

To address the problems, a novel Domain-sensitive Recommendation (DsRec) algorithm assisted with the user-item subgroup analysis, which integrates rating prediction and domain detection into a unified framework. We call the proposed algorithm DsRec for short. There are three components in the unified framework. First, we apply a matrix factorization model to best reconstruct the observed rating data with the learned latent factor representations of both users and items, with which those unobserved ratings to users and items can be predicted directly. Second, a bi-clustering model is used to learn the confidence distribution of each user and item belonging to different domains. Actually, a specific domain is a user-item subgroup, which consists of a subset of items with similar attributes and a subset of users interesting in the subset of items. In the bi-clustering formulation, we assume that a high rating score rated by a user to an item encourages the user and the item to be assigned to the same subgroups together. Additionally, two regression regularization items are imported to build a bridge between the confidence distribution of users (items) and the corresponding latent factor representations. That is, the confidence distribution over different subgroups (domains) in DsRec could be considered as soft pseudo domain labels, to guide the exploration of the latent space. Thus, connected with the regression regularizations, DsRec could learn discriminative and domain-sensitive latent spaces of users and items to perform the tasks of rating prediction and domain identification.

To the best of our knowledge, our work is the first to jointly consider the both tasks by only utilizing user-item interaction information. An alternate optimization scheme is developed to solve the unified objective function, and the experimental analysis on three real-world datasets demonstrates the effectiveness of our method.

VI. CONCLUSION

Recommender systems are rapidly becoming a crucial tool in E-commerce on the Web. A novel framework for collaborative filtering has been proposed. By integrating the advantages of memory and model-based collaborative filtering into a single framework, this approach targets two fundamental problems: data sparsity and scalability. Clusters can be used to provide smoothing operations to solve the missing-value problems. *RecTree* achieves better scale-up in comparison to other memory based collaborative filters by seeking advisors only within a clique rather than the entire database. Parallel implementation of *RecTree* will be able to realize a greater throughput, which may be the subject of future work. The central idea of TCRec is that the tastes of customers who trust each other are similar and the latent features of products in one category can be discriminated from others. The experimental results on two real-world datasets indicate the effectiveness of this approach. DsRec is a unified formulation integrating a matrix factorization model for rating prediction and a bi-clustering model for domain detection.

VII. REFERENCES

[1] Y. Zhang, B. Cao, and D.-Y. Yeung, "Multi-domain collaborative filtering," in Proc. 26th Conf. Annu. Conf. Uncertainty Artif. Intell.,
2010, pp. 725–732.

[2] X. Zhang, J. Cheng, T. Yuan, B. Niu, and H. Lu, "TopRec: Domainspecific recommendation through community topic mining in social network," in Proc. 22nd Int. Conf. World Wide Web, 2013, pp. 1501–1510.

[3] Y. Jiang, J. Liu, X. Zhang, Z. Li, and H. Lu, "TCRec: Product recommendation via exploiting social-trust network and product category information," in Proc. 22nd Int. Conf. World Wide Web, 2013, pp. 233–234.

[4] S. Han, S. Chee, J. Han, and K. Wang, "RecTree: An efficient collaborative filtering method," in Proc. 3rd Int. Conf. Data Warehousing Knowl. Discovery, 2001, pp. 141 151.

[5] B. M. Sarwar, J. Konstan, and J. Riedl, "Recommender systems for large-scale e-commerce: scalable neighborhood formation using clustering," in Proc. 5th Int. Conf. Comput. Inf. Technol., 2002, pp. 1–6.

[6] D. M. Pennock, E. Horvitz, S. Lawrence and C. L. Giles. Collaborative Filtering by Personality Diagnosis: A Hybrid Memory- and Model-Based Approach, in Proceedings of the Sixteenth Conference on Uncertainty in Artificial Intelligence (UAI), 2000.

[7] M. Fazel, H. Hindi, and S. Boyd. Log-det heuristic for matrix rank minimization with applications to hankel and Euclidean distance matrices. In Proceedings American Control
Conference, pages 2156–2162, 2003.

[8] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, An Algorithmic Framework for Performing Collaborative Filtering, In *Proc. 1999 Conf. Research and Development in Information Retrieval*, pp. 230-237, Berkeley, CA, August 1999.

[9] L. Kaufman and P. Rousseeuw, *Finding Groups in Data, An Introduction to Clustering Analysis.* John Wiley and Sons, 1989.