Combination of a Cluster-Based and Content-Based Collaborative Filtering Approach for Recommender System

Ms. Ashwini A. Chirde
ME-Computer Networks
GHRCEM, Wagholi,
Pune, India
Ashwinichirde1@gmail.com

Prof. Ms. Urmila K. Biradar
Computer Engineering
GHRCEM, Wagholi
Pune, India
urmilakb@gmail.com

***Abstract*—** with the development in technology in the field of e-commerce, the problem with information overload has been at its peak. Oftentimes the user is overwhelmed by the huge amount of options he/she is provided with while searching for an item. This is when recommender system comes in handy, which is an information filtering technique aimed at presenting the user with the most possible options based on certain reference characteristics. However, the problem with many recommender systems is that they are associated with a high cost of learning customer preferences. The current agricultural web application uses recommendation system along with the collaborative filtering concept which introduces the Agricultural Informative System (AIS) that uses pseudo feedback, which provides a method for automatic local analysis about the user preferences with the help of clustering in collaborative filtering. The AIS uses pseudo feedback to capture the preferences which are stored in the users profile for future personalized recommendations to address the problem.

Keywords- *Collaborative Filter, Clustering, E-Commerce, Recommendation Systems*

*****

I. Introduction

The advancement in technology in the field of Electronic commerce has enabled businesses to open up their products and services to a massive client base. As the competition ‘between businesses becomes increasingly fierce, consumers are faced with a multitude of choices and hence information overload. In any online business, the problem of information overload occurs when the user is provided with too many options to choose from, most of which may not be what the user is looking for. Hence an effective system would capture the user preferences from his/her previous purchase history and use the same in the process of personalized recommendation in the future.

The user preferences are stored every time he makes a purchase. Explicit feedback from the user could be one way of collecting these preferences. But, explicit feedback often times may not be reasonable, as it may result in boring the user and hence increasing the user’s frustration level. [1] One way of working around this problem would be to collect implicit feedback about the user preferences by mining for information from the data stored in the user profile. The aim of this project is to use the Java framework in effectively creating an application that provides the farmers a friendly interface to sell their products and buy the essentials. The key aspect of the system being the ability to recommend the user, the locations that best fit his/her selections and nearest to his/her proximity, based on purchase history.

The system will take in the user input and will generate a list of locations that the user is free to select from depending on his personal preference. The input will include the user requirement, his/her current location, and the approximate date when he/she is going to buy the product. The inputs required are strongly imposed by validating the required fields, which ensures the user selections are correct and prompt. A sorted list of recommended results will then be displayed dynamically by searching the database for the entries that meet the input criterion. The system is developed using JSP with the database MySQL (WampServer2.0i).

This paper is organized as follows. In section II we discuss the works related to the collaborative filtering systems. In section III the proposed system is discussed. In section IV the cluster based collaborative filtering is introduced. In section V the system architecture is discussed. In section VI implementation results are summarized. In section VII our conclusions are proposed.

II. Related Work

The biggest challenge in collaborative filtering recommender system is scalability. The system should provide accurate recommendations for the super user as the more number of users is increasing in the site. The imputed divisive hierarchical clustering approach is used by Suresh Joseph K and Ravichandran T to overcome the scalability issue when more number of users increases in terms of neighborhood size.

A literature survey on cluster based collaborative filter and an approach to construct is givenby R. Venu Babu and K. Srinivas. A coverage metric that uncovers and compensates for the incompleteness of performance evaluations based only on precision is proposed by Alejandro Bellogin, Ivan Cantador, Fernando Diez, Pablo Castells and Enrique Cavarriaga. They use this metric together with precision metrics in an empirical comparison of several social, collaborative filtering, and hybrid recommenders.

F. R. Sayyed, R. V. Argiddi, S. S. Apte proposed a Collaborative Filtering Recommender System which can be used for financial markets such as stock exchanges for future predictions.[4] F. Darvishi - mirshekarlou, SH. Akbarpour and M. Feizi-Derakhshi, by reviewing some recent approaches in which clustering has been used and applied to improve scalability, the effects of various kinds of clustering algorithms (partitional clustering such as hard and fuzzy, evolutionary based clustering such as genetic, memetic, ant colony and also hybrid methods) on increasing the quality.

The idea of object typicality from cognitive thinking is borrowed by Yi Cai, Ho-fung Leung, Qing Li, Senior Member, IEEE, Huaqing Min, Jie Tang and Juanzi Li and they[5] suggested a novel typicality-based collaborative filtering recommendation method termed TyCo A distinct feature of typicality-based CF is that it finds "neighbors" of users based on user typicality degrees in user groups (instead of the co-rated items of users, or common users of items, as in traditional CF). TyCo has lower time cost than other CF methods and it outstrips many CF recommendation methods. Further, it can obtain more exact predictions with less number of big-error predictions.

Xindong Wu, Fellow, IEEE, Xingquan Zhu, Senior Member, IEEE, Gong-Qing Wu, and Wei Ding, Senior Member, IEEE presents a HACE theorem that describes the features of the Big Data revolution, and suggests a Big Data processing model, from the data mining perspective. [1] The demand-driven, mining and analysis, user interest modelling, and security and privacy considerations are involved in the demand-driven aggregation of data sources.

Junhao WEN and Wei ZHOU presents improved collaborative filtering recommendation algorithm based on dynamic item clustering method was proposed in the paper.
Item-based collaborative filtering recommendation algorithm is one of the most widely used recommendation algorithm, which is widely used in many recommendation systems. But there are some drawbacks when used in large e-business systems. The existing traditional algorithms can't perform well when the item space changes; on the other side, the performance of the recommendation system will go down as the items increase into a large amount. The improved collaborative filtering recommendation algorithm performs relatively good recommendation with less resource consumption.

III. EXISTING SYSTEM

Amazon and YouTube use Item Clustering Collaborative Filtering technique. Last.fm and Reddit use Collaborative Filtering technique. Pandora uses Content based approach. Facebook, MySpace, LinkedIn use 'Collaborative Filtering technique' to make friend suggestions, groups and other social connections by observing the network of connections between a user and people present in their connections.

Twitter makes use of several signals and in-memory calculations for suggesting whom to follow. Netflix is a hybrid system. "They make recommendations by comparing the watching and searching habits of similar users (Collaborative Filtering) as well as by offering movies that share characteristics with films that a user has rated highly (Content based Filtering)".

Pandora utilizes the features of a song or artist for tuning into a station which plays music with similar features. Feedback given by the users is used to tune the station; not considering some features like dislikes and gives more importance to the other features like liking a particular song. This is a 'Content Based approach'. With little information Pandora can get started and has limited scope. If the song is alike to the original seed, only then it is suggested.

In Last.fm, a station is created in which songs are suggested to the users based on history of the user and compares the taste of current user against other users. Last.fm plays songs which the users with similar taste listen frequently. This is an example of 'User based Collaborative Filtering' because suggestions are made by considering other users choice.Last.fm needs huge data related to a particular user to make reliable suggestions. This experiences cold start problem.

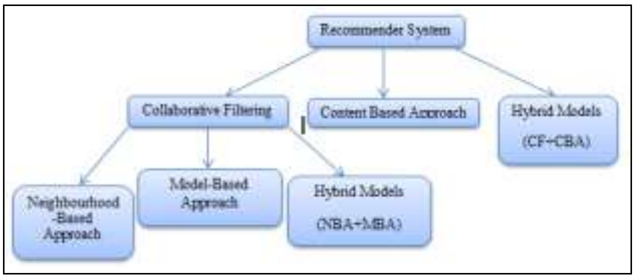

Figure 1: Collaborative filtering for recommender system

By considering all the pros n cons of the existing systems we are going to develop a new collaborative filtering-recommender system, which can be explained as follows. Now-a-days everything can be sold and bought online using commercial websites. But still there is very less work done in agricultural domain. If we can sell anything online then, why not the agricultural products? The bigger part of our economy is dependent on agriculture. The way of farming and the quality and quantity of the products are improving day by day. The more scientific approach is being applied to farming, so that the better outputs can be obtained. So, if the way of farming and the output obtained are getting better with the passing time, then why can't we use advanced techniques for selling the products? We can develop an application which will be useful for selling the products as well as will be used as an informative system by the farmers. The application will give whole information about the inputs to be put in the farming like seed, fertilizers, instruments, machines needed, etc.

There are some websites owned by government which will be useful to the farmers working as an information center and some by wealthier farmers who have developed their own sites to sell their products. Each and every farmer can't have their own website for selling their products. So, we are going to develop an application which will be useful to the common farmers, by using the concepts like collaborative filtering, Big Data and Semantic Analysis.

IV. SYSTEM ARCHITECTURE

The work can be done by dividing the application in three major sections:

1. The Farmer
2. The Seed and Fertilizer vendor
3. The Customer

A. The Farmer:

- The farmers will be clustered together. And there will be a cluster head for each group of farmers.
- After registration at the Cluster Head, the farmers will be able to update the information about the products at the CH, like the available products, upcoming products, products not in stock, etc.

B. The Seed and Fertilizer Vendor:

- Along with the seed and fertilizers there will be an additional feature for the famers that he will get information about the instruments and machines required for farming.

C. The customer:

- The customer will send the requirements through the website.
- The admin will then send this requirement notification to all CH's.
- Every CH will then check to see whether their farmers can fulfill the requirements.
- If yes, then will revert the admin informing that they can satisfy the requirements. And if they can't then they will inform the same.
- After successful processing admin will then inform the customer regarding whether the requirement can be satisfied or not.
- The locations will be suggested to the customer from where he/she can buy the products.

The farmers/customers will also get recommendations for the similar products and the nearby locations based on the other similar user ratings to the products and the past interest of the farmer/customer who want to buy the things.

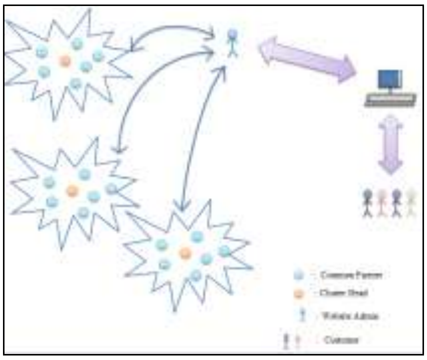

Figure 2: System Architecture

Algorithm 1: Algorithm for similarity computation

When given enough and clear information, the traditional algorithm usually shows good performance. But with the increment of users and items of the rating matrix, item-based collaborative filtering algorithm gradually exposes some shortcomings. With items and users added in the rating matrix every day, we need an algorithm that adjusts to the dynamic change of the rating matrix.

There are x clusters in the cluster set, each cluster has many items which have similar ratings. The cluster center of each cluster is the mean rating value of all the items in the cluster. When a new item is added in, the similitude between the item and other cluster centers will be calculated.

If the max value of similitude is bigger than the threshold, the item will be added to the cluster which has the biggest similitude with the item. The cluster center will be recalculated. Else a new cluster center will be built with the rating score of the item be the cluster center of the new cluster.

Definition:

Collection of all items I = $\{i_1,i_2,\ldots,i_m,\ldots,i_n\}$, while i_m is an item of the collection I or collection of products.

Collection of all items $U = \{u_1,u_2,\ldots,u_a,\ldots,u_b\}$, while u_m is an item of the collection U called users.

Collection of cluster $C = \{c_1,c_2,\ldots,c_p,\ldots,c_x\}$, while c_p is a cluster of the cluster collection C.

Initialize $C = Ø, x = 0$.

Collection of cluster centers $CC = \{cc_1,cc_2,..,cc_p,\ldots, cc_x\}$, while cc_p is a cluster center of the collection CC. Initialize, $CC = Ø, x = 0$.

Input: User Rating Database (URDB) and similitude threshold threshold.

Output: x clusters

1. Retrieve all n items from the database URDB, assigned as a collection $I = \{i_1,i_2,\ldots,i_m,\ldots,i_n\}$, while i_m is an item in I.
2. The first item i_1 is retrieved from URDB; initialize an empty cluster, assign the cluster center of the cluster with the score of the item in URDB. $I = I-i_1$. $x = 1$.
3. for each item i_m in collection I
4. for each cluster center cc_p
5. Calculate the similitude of item i_m and cluster center cc_p, $sim(i_m,cc_p)$
6. endfor;
7.$max(i_m,CC)=sim(i_m,cc_p)=max(sim(i_m,cc_1);sim(i_m,cc),\ldots,sim(i_m,cc_x))$
8. if $max(i_m;CC) < threshold$
9. $x = x + l$, add a new cluster cc_x, assign the cluster center of cc_x with the score of the item in URDB. $I = I-i_m$.
10. else $c_p = c_p + i_m$, $I = I-i_m$
11. for each user u_a in U
12. Recalculate the mean score of the items in the cluster by user u_a. Generate the new cluster center.
13. endfor;
14. end;

A cluster center is the typically score of the cluster. The similitude score of items in the same cluster as high as possible, and similitude score of items as low as possible between different clusters. In the next step, choose the highest similitude clusters with the target item as the search space, finding the nearest neighbor in the search space.

Algorithm 2: Finding the nearest neighbors

Definition:

TargetClusters: Collection of the clusters that fulfill the condition of similitude threshold. Initialize NULL.

NearestNeighbors: Output collection of the algorithm. Initialize NULL.

Input: the target item; numbers of nearest neighbors k; URDB, Collection of clusters C; collection of cluster centers CC; the threshold of similitude simthre.

Output: k nearest neighbors of the target item.

1. TargetClusters = NULL,
2. for each cluster c_p in collection C
3. Calculate the similitude of item i_m and cluster of c_p
4. if $sim(i_m,cc_p) > simthre$
5. TargetClusters = TargetClusters + c_p, $C = C-c_p$.
6. Endif;
7. Endfor;
8. Sort items in TargetClusters by the similitude with item i_m, find the top-k neighbors.
9. End;

A rating matrix is constructed according to the generated cluster centers.

V. IMPLEMENTATION

The system is developed using JSP and MySQL (WampServer2.0i) database has been used as the back end. Eclipse JAVA framework will be needed for development of JAVA Server Programs. The system makes personal recommendations to users based on their purchase history. All the algorithms are written in java. We will conduct a number of experiments to verify effectiveness of the proposed methods. All the experiments will be conducted on PC with an Intel Core i3 and 4GB RAM.

VI. BEHAVIOURS OF OUR METHOD

We make 2 experiments to test the algorithm. The first one the new algorithm was compared with other algorithm. In the next experiment, we compare our new method with the existing algorithms. In figure 3, the time required for clustering the users in particular locality is compared. Larger the cluster size, the time required will also increase.

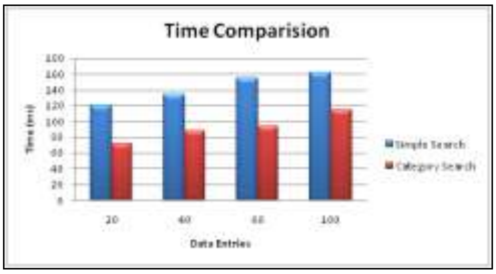

Figure 3: Graphical representation of time consumption for clustering

We can see from Fig 4 that, the predicted ratings given by the users to products will be compared using item based algorithm and the proposed algorithm. The results from proposed algorithm are expected to be better as they are location based. As the number of users increases the predicted ratings will become more and more accurate.

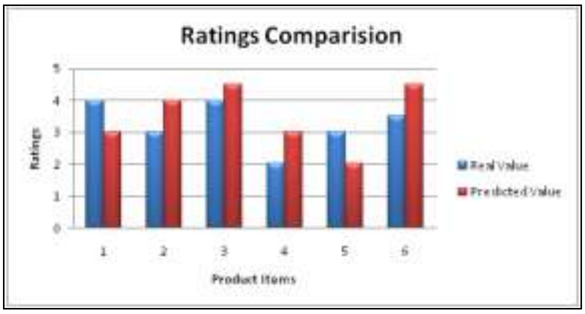

Figure 4: Comparison between real values and predicted values

VII. CONCLUSION

Recommender systems are considered as a filtering and retrieval technique developed to alleviate the problem of information and products overload. Collaborative filtering is the most popular and successful method that recommends the item to the target user. In this paper, we have proposed a new collaborative filtering approach for recommender system, which have been studied in dynamic environment. We are developing an improved collaborative filtering recommender system based on proposed algorithms, which will be useful in the agricultural field. Pseudo feedback concept and the missed least rating item concept can be added as a future work.

ACKNOWLEDGMENT

I am grateful to number of individuals, faculty members, whose professional guidance along their encouragement have made it very pleasant endeavour to undertake this work. I have a great pleasure in presenting the work under the guidance of Asst. Prof. Ms. Urmila K. Biradar for giving me the opportunity to work on this topic. I would like to express our gratitude to Prof. Dr. Tanuja Dhope, Head of Computer Engineering Department and Dr. Shaha sir (Principal, GHRCEM, Pune) for their support and guidance also all the teaching and non-teaching staff of Computer Engineering Department for their encouragement, support and untiring cooperation Finally I express my sincere thanks to our parents, friends and all those who helped us directly or indirectly in many ways in completion of this work.

REFERENCES

[1] Qing Li, Byeong Man Kim, “An Approach for Combining Content-based and Collaborative Filters”, (This work was supported by Korea Research Foundation Grant (KRF-2002-041-D00459)).

[2] X. Wu, X. Zhu, G. Q. Wu, “Data mining with big data,” IEEE Trans. on Knowledge and Data Engineering, vol. 26, no. 1, pp. 97-107, January 2014.

[3] Suresh Joseph. K, Ravichandran. T, “A Imputed Neighborhood based Collaborative Filtering System for Web Personalization”, International Journal of Computer Applications (0975 – 8887) Volume 19– No.8, April 2011.

[4] Manh Cuong Pham, Yiwei Cao, Ralf Klamma, Matthias Jarke, “A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis”, Journal of Universal Computer Science, vol. 17, no. 4 (2011), 583-604 submitted: 30/10/10, accepted: 15/2/11, appeared: 28/2/11 © J.UCS.

[5] F.R.Sayyed, R.V.Argiddi, S.S.Apte, “Collaborative Filtering Recommender System for Financial Market”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-6, August 2013.

[6] Yi Cai, Ho-fung Leung, Qing Li, Senior Member, IEEE, Huaqing Min, Jie Tang, and Juanzi Li, “Typicality-Based Collaborative Filtering Recommendation,” IEEE Transactions on Knowledge and Data Engineering, Vol. 26, No. X, XXXXXXX 2014

[7] F. Darvishi - mirshekarlou, S. H. Akbarpour, M. Feizi - Derakhshi, “Reviewing Cluster Based Collaborative Filtering Approaches”, International Journal of Computer Applications Technology and Research Volume 2– Issue 6, 650 - 659, 2013.

[8] Fay Chang, Jeffrey Dean, Sanjay Ghemawat, Wilson C. Hsieh, Deborah A. Wallach Mike Burrows, Tushar Chandra, Andrew Fikes, Robert E. Gruber, “Bigtable: A Distributed Storage System for Structured Data”, Google, Inc, OSDI2006.

[9] Junhao WEN∗, Wei ZHOU, “An Improved Item-based Collaborative Filtering Algorithm Based on Clustering Method”, Journal of Computational Information Systems 8: 2 (2012) 571–578.

[10] Alejandro Bellogin, Ivan Cantador, Fernando Diez, Pablo, Castells and Enrique Chavarriaga, “An Empirical Comparison of Social, Collaborative Filtering, and Hybrid Recommenders”, © 2011 ACM 1073-0516/01/0300-0034.

[11] Rong Hu, Member, IEEE, Wanchun Dou*, Member, IEEE, Jianxun Liu, Member, IEEE,” ClubCF: A Clustering-based Collaborative Filtering Approach for Big Data Application”, IEEE Transactions On Emerging Topics in Computing.