A Survey on Collaborative Filtering in Accordance with the Agricultural Application

Ashwini A. Chirde
M.E. Computer Networks , Student
GHRCEM, Wagholi
Savitribai Phule Pune University, India

Umila K. Biradar
Asst. Professor in Computer Dept
GHRCEM, Wagholi
Savitribai Phule Pune University, India

ABSTRACT
In modern E-Commerce it is not easy for the customers to find the best goods of their interest as there are millions of products available online. Recommender systems, one of these systems, are one of information filtering systems forecasting the items that may be additional interest for user within a big set of items on the basis of user's interests. This system uses the Collaborative filtering, which offers some recommendations to users on the basis of matches in behavioral and functional patterns of users and also shows similar fondness and behavioral patterns with those users. We are going to develop an agriculture based application by using the Collaborative Filtering, Semantic Analysis and Big Data concepts. This system will be useful to farmers for selling their products and getting information regarding the required material in farming.

Keywords
Collaborative Filter, Clustering, LSA-Latent Semantic Analysis, Recommendation Systems.

1. INTRODUCTION
The recent rapid expansions in computer use and a massive increase in the popularity of the Internet have inspired lots of companies to enter the online market, which results in a giant expansion in the number of E-commerce sites. The companies started to consider how to provide users with information about the most desirable items among their many products to be an outstanding E-commerce site. Resulting in the development of recommender systems. Recommender systems are used to progress the business intelligence in an E-commerce site. By using Collaborative filtering (CF) we can describe a number of processes involving the recommendation of items to the users based upon the opinions of a neighborhood of human advisors.[8] The main idea behind CF approach is that those who granted in the past tend to agree again in the future. It is usually relevant to e-Business, e-Learning, and so on. Presently, many e-business sites are previously using the recommended system, such as Amazon, CDNow and Drugstore etc. The correspondence computation is an important issue and the definition of similarity may vary as per the application. The degree of match between two objects can be decided from the final forecasts to be done. Some old-style CF algorithms have quadratic complexity. The social recommenders' are another and/or stability to Collaborative Filtering (CF) methods. The social recommenders are represented as links among users in the community to suggest interesting items.[9] The social recommenders naturally require the availability of a social context and they perform very well in some situations. This requirement introduces a vital limitation of real applications, namely, the incapability to produce endorsements for users, when insufficient social links are known to the system. This limitation may go somewhat ignored in studies focusing on overall recommendation accuracy, in which a failure to produce recommendations gets blurred when averaging the obtained results or, even worse, is just not accounted for, as users with no recommendations are typically excluded from the performance calculations. The collaborative filtering can also be applied to big data so that the data filtering can be done effectively through the enormous amount of data.

Big data arose as a widely accepted trend, attracting attentions from government, industry and academia. In general, the Big Data concerns large-volume, complex, growing data sets with multiple, autonomous sources.[7] Big Data applications is on the rise, where the data gathering has grown extremely and is past the capacity of normally used software tools to capture, manage, and process within a "tolerable elapsed time". The most basic challenge for the Big Data applications is to explore the large volumes of data and extract useful information or knowledge for future actions. The BigTable resembles a database: it shares many implementation strategies with databases. The parallel databases and main-memory databases have achieved scalability and high performance, but BigTable provides a different interface than such systems. BigTable does not support a full relational data model; instead, it provides clients with a simple data model that supports dynamic control over data layout and format. Data is indexed using row and column. BigTable also treats data as uninterpreted strings, although clients often serialize various forms of structured and semi-structured data into these strings. Clients can control the locality of their data through careful choices in their schemas. Finally, BigTable schema parameters let clients dynamically control whether to serve data out of memory or from disk.

2. 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 given by 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.[6]

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.

The innovation in LSA method for improving its restrictions is given by Soniya Patil, Ashish T. Bhole. LSA is a fully involuntary mathematical/statistical technique for extracting and concluding relations of expected contextual usage of words in passages of discourse. It uses no humanly created dictionaries, knowledge bases, semantic networks, grammars, syntactic parsers, or morphologies, or the like, and takes as its input only raw text parsed into words defined as unique character strings and separated into meaningful passages or samples such as sentences or paragraphs also it is not a traditional natural language processing or artificial intelligence program. The enactment will be raised with the improvement in LSA method. The restrictions of existing statistical approach can be overcome and improved output can be attained using LSA method.

3. EXISTING SYSTEM

The recommended concept of a Clustering-based Collaborative Filtering approach (ClubCF), consists of two stages: Clustering and Collaborative Filtering. The pre-processing step to separate big data into controllable parts is a clustering. A cluster contains some similar services just like a club contains some like-minded users. As the cluster comprises less number of services as compared to the total number of services, the computation time of CF algorithm can be reduced considerably. Besides, since the ratings of similar services within a cluster are more relevant than that of dissimilar service, the endorsement accuracy based on users' ratings may be enhanced.

Although these recommendation methods are broadly used in E-Commerce, a number of shortages have been recognized, including:

3.1 Data Sparsity:

The data sparsity problem is the problem of having too few ratings, and hence, it is hard to find out connections between users and items. It happens when the existing data are unsatisfactory for identifying similar users or items. It is a major issue that limits the quality of CF recommendations.

3.2 Recommendation Accuracy:

People need recommender systems to estimate users' likings or ratings as accurately as possible. However, some predictions provided by current systems may be very dissimilar from the actual preferences or ratings given by users. These inaccurate predictions, especially the big-error predictions, may decrease the trust of users on the recommender system.

With the above stated issues, [5]it is clear that a good mechanism to find "neighbors" of users is very important. An improved way to select "neighbors" of users or items for collaborative filtering can facilitate well handling of the challenges.

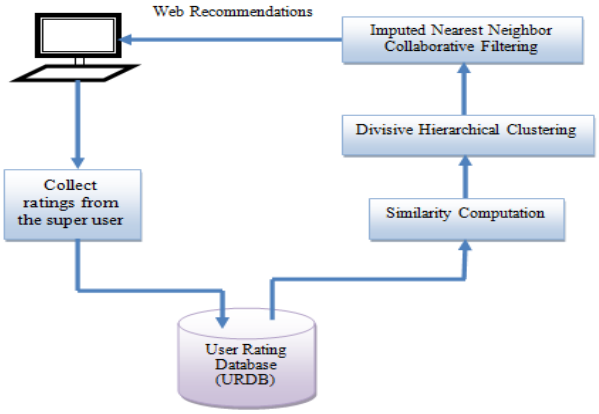

Figure 1: Proposed Architecture for CF Web Recommendation System

Stage 1: Clustering Stage

1. The words will be stemmed in D_t and D_j using Porter Stemmer. The stemmed words in D_t are set into $D_{t'}$ and the stemmed words in D_j are put into $D_{j'}$.
2. Compute $D\_(s_{t,}s_j)$, and $F\_sim(s_t, s_j)$ using Jaccard similarity coefficient individually.
3. Calculate $C\_(s_t, s_j)$ by weighted sum of $D\_sim(s_t,s_j)$ and $F\_sim(s_t,s_j)$. Construct a matrix D each record of which is a characteristic similarity.[3]
4. Cluster the services rendering to their characteristic similarities in D using an agglomerative hierarchical clustering algorithm.

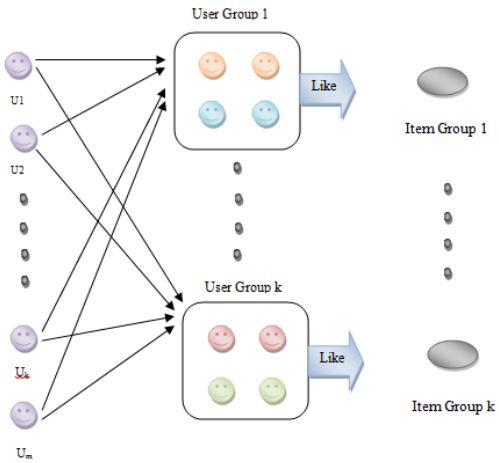

Figure 2: The relations among users, user groups, and item groups

Stage 2: Collaborative filtering Stage

1. Calculate the $R\_sim(s_t,s_j)$ [3]using Pearson correlation coefficient if st and sj belong to the similar cluster, and calculate $R\_sim'(s_t,s_j)$ by weighting $R\_sim(s_t,s_j)$.
2. Select services whose boosted rating similarity with s_t exceed a rating similarity threshold γ, and put them into a neighbours set.
3. Compute the expected rating of s_t for an active user. If the expected rating exceeds a recommending threshold, it will be suggested to the active user.

The idea of a ClubCF method for big data applications related to service recommendation is explained.[10] Services are merged into some clusters via an AHC algorithm before applying any CF technique. After that the rating likenesses between services within the same cluster will be calculated. Moreover, as the ratings of services in the same cluster are more related with each other than with the ones in other clusters. The prediction based on the ratings of the services in the same cluster will be more correct than that of based on the ratings of all similar or dissimilar services in all clusters.

4. PROPOSED SYSTEM

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.

4.1 Work Flow

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

This can be explained as:

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.

5. CONCLUSION AND FUTURE WORK

A ClubCF method for big data applications is related to service recommendation. Big Data are now rapidly expanding in all science and engineering areas, including physical, biological and biomedical sciences. So we can develop number of applications to filter the data. The future work can be done in two areas. First, by introducing the actual buying and selling concept in the website. So that the farmers can be able to buy the essentials and sell the products through the

application. For this, the banking concept can be introduced. Second, the SMS alert can be activated after the farmer registration at the cluster head. The transaction information and also the information regarding product requirement and the product delivery will be provided to the farmers through this SMS alert. Also by considering the point of ignorance of least rated items we can modify the system. We can also use the concept of pseudo feedback, which provides a method for automatic local analysis. It automates the manual part of relevance feedback, so that the user gets improved retrieval performance without an interaction.

6. ACKNOWLEDGEMENT

I express my sincere thanks to Asst. Prof. Ms. Urmila K. Biradar, whose supervision, inspiration and valuable discussion has helped me tremendously to complete my work. Her guidance proved to be the most valuable to overcome all the hurdles in the fulfillment of this paper.

7. REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.