A review: Recommender System using Collaborative Filtering and Gray Sheep Problem

Rajatha Prabhu[1], Pooja Shetty[2], Shilpa[3], Shwetha D R[4], Ramakrishna Hegde[5]
[1-4]Engineering Graduate Students, [5]Assistant Professor
Dept. of Computer science and Engineering, SDM Institute of technology, Ujire

Abstract - Recommender system is a powerful process that plays an important role in the online marketing system. The recommendation technique used has become one of the important tools of personal service in website. The recommender systems are classified into three main categories: content-based, collaborative and hybrid recommendation approaches. Recommender system act as the bridge gap between the customers and the applications or websites by providing many options from which the customers make their choice of interest. In designing such recommenders' designers face several problems. This review paper provides different methods to solve gray sheep problems. This can be used for the further research of fine-tuning and designing high quality recommender system.

Keywords - recommender system, content-based, collaborative, hybrid recommendation, clustering algorithms.

I.INTRODUCTION

Recommendation is a subclass of information filtering system that seeks to predict the "rating" or "preference" a user would give to an item. Recommender system have become more popular in recent years and are utilized in many areas like movies, news, books, music, search queries and products in general.

Recommendation System Techniques

This is categorized into three classes: Collaborative Filtering, Content Filtering, Hybrid based approach and Collaborative Filtering is classified into two categories that are Memory based and Model based Recommendation[1]. This paper explains about limitation of these approaches.

1. **Collaborative Filtering**

Collaborative filtering is method of making automatic prediction about the interest of user by calculating user's ratings on item. This approach focusing on relationship between items and user's. Based on review of other user Collaborative filtering recommend items to users. For example, in book recommendation system it tries to search like-minded people and recommend books that are mostly liked by them. The Workflow of Collaborative filtering is by giving rating to items a user convey his/her preferences and these ratings can be taken as representation of the User interest in certain domains. System tries to find people with similar taste by matching certain user ratings against others and system recommend items with similar user have rated highly but not yet being rated by this user.

- **Memory based Collaborative filtering -** To compute similarity between user's and item Memory based approach uses user data ratings. Memory based Collaborative filtering use data like likes, votes, clicks to establish link between either user or items to recommend items to user. This approach access database directly so it requires large computational time according to data size. Whereas Model based approach has constant computational time. Memory based approach uses entire user-item data make prediction. Advantage of Memory based approach is easy to implement and effective. It have limited scalability for large dataset. Nearest neighbor-is usage of statistical techniques to find neighbor. Minkowski distance, Pearson correlation and cosine metrics are used to calculate the similarity or distance or weight between users or items.
- **Model based Collaborative filtering -** By developing model from user ratings on item, model based approach recommend item to user. The advantage of model based approach is scalability, speed and it is easier to try to avoid over fitting and disadvantage of this approach is inflexibility and quality of prediction. Bayesian networks, clustering and latent semantic models are examples of this approach. Sometime it takes complete dataset for training data or sometime it divide dataset into ratio to train model.

2. Content-based Filtering

Using rating history of current user related to previously used items this method finds the preferences of the current user about new item. By measuring the similarity in their properties similarity of items is determined. So, there is no dependency on rating records of other users in order to generate preferences for current user in this type of filtering method. Content-Based systems focus on properties of items. For example, if the user has purchased a book on amazon.com which uses recommender system then the user starts getting additional preferences for buying books from online book store which includes same keywords information for books.

3.Hybrid Filtering

A hybrid is one that joins multiple technique together to achieve some synergy between them. No single recommender system approach is to be efficient enough to generate relevant and accurate recommendation preferences, so, a hybrid recommender systems came into existence to overcome the limitations of traditional recommendation approaches mentioned above. These systems are based upon combining advantages of more than one traditional approach for recommendation generation for example; collaborative filtering approach with content-based approach or collaborative filtering etc.

II.RECOMMENDER SYSTEM CHALLENGES

Recommender systems have been used for providing recommendations for various application domains and this has brought to notice many challenges. Research areas are emphasizing on solving the issues mentioned below[2]:

- **Data Sparsity**
  Problem appears if the user-item matrix containing ratings details is extremely sparse and this situation further leads to inefficient recommender systems which are based upon **nearest-neighbor algorithms** for calculating user similarity. This problem is further classified as **reduced coverage problem and neighbor transitivity problem.**
- **Scalability**
  Scalability problem occurs when numbers of items and users giving ratings those items increased to very great extent and it becomes difficult for a recommender system to handle such a big data due to computational complexity and constrained resources. So, it goes beyond the limit of acceptability of recommender systems.
- **Synonymy**
  It arises when recommender system fails to recognize the similarity among two items when some similar items have different names and the recommender system treats them as if they are different items. This leads to problem of recommending similar items, called as synonymy problem.
- **Gray sheep**
  Problem arises because the user's choice does not match with any other user or group of users in agreement or disagreement consistently.
- **Cold-start problem**
  With the progress of e-commerce, product recommendation has been recognized as critical tool for improving sales in online store. We can increase income through persuading customer to buy something additional or more expensive. There have been more numerous way of product recommendation methods one of the most successful method is collaborative filtering method. Collaborative filtering is method of making automatic prediction about the interest of user by calculating user's ratings on item. Here we focusing on improving the performance of recommendation systems under cold start problem. Cold start problem occurs when it is not possible to make recommendation due to lack of user's rating on item. This kind of situation occurs when new user or items enters having no rating records or no one has given ratings till now.

III.RELATED WORKS

The Recommender Systems are applied in different domains such as recommendations of books, film, webpages, news, songs, products. But the Recommender Systems pay less attention to the existence of grey-sheep users. These users will have negative effects on individual user recommendations. This study focuses on clustering the users and separating these grey-sheep users from normal users to provide better performance.

Clustering method using classification algorithm

Problem in Recommender Systems is the cold-start problem. This problem is related to recommendations for new users and new items. When the new users arrives, Recommender Systems do not have sufficient information for creating recommendations. Lika[7] method which is also known as classification algorithm, is combined with similarity methods and recommendation methods to provide instruments required for producing recommendations. Classification methods are mixed with the collaborative filtering system in order to identify users with similar behaviors. The system uses a three-stage method for creating recommendations for new users. This method is adopts a mechanism based on which similarity methods find neighboring users. The users with the highest level of similarity with new users are called neighbors. Therefore, each new user in a group is classified and is the rating prediction method and responsible for producing rates for items. The advantage of this system is that produces lower mean absolute error values and increases the precision of rating prediction (Lika et al., 2014)[7].

Hierarchical Cluster Method

This method is used for recommendation. Hierarchical clustering is a cluster analysis method to build hierarchy of clusters. It consists of two types[4] such as agglomerative and divisive. Users preferences is not limited only to literature review on users, but it acts according to group literature of users with similar preferences and with this hypothesis that each group of users have unique preferences .In addition, users personal pages is hierarchically indicated with a mixture of several topics via combining hierarchy related to users groups in which users use adaptive hierarchical clustering[5]. In this part, users personal profiles become enriched regarding similar users profiles. In this users are divided into some groups in terms of similarity and then, users personal page in each group is created using a weighting method in such a way that users having more interest to a topic share more pieces of information in group profiles (Zheng et al., 2013)[5].

Neighborhood-Based Collaborative

Filtering - This approach is also referred as k-Nearest Neighbor or kNN. This method is used in order to solve cold start problem. This method identifies pairs of like-minded users with similar history of rating. The three major components in kNN

approach[6] are: (1) data normalization (2) neighbor selection (3) determination of interpolation weights. In this method, ratings of neighborhoods are combined with each other in order provide complete preferences of active users and illustrate them. Prediction of the items depends on calculation of mean rates obtained from the similar users.

Centroid-based Clustering Method Based

Centroid-based clustering algorithm developed by Shinde[8] are used for Personalized Recommender Systems. This method consists of two phases for recommendation. In the first phase, users comments are collected in the form of the User-Item Ranking Matrix and then they are clustered offline and stored in databases for future recommendations. In the second phase recommendations for active users are done using similarity criteria and are produced by selecting clusters with the best rating quality. The efficiency and quality of Recommender Systems for active users are improved. This helps in the exploration of other clusters which have closer similarities with active users and provide better recommendations for them. This method reduces problems such as cold start, the first rater and gray sheep.

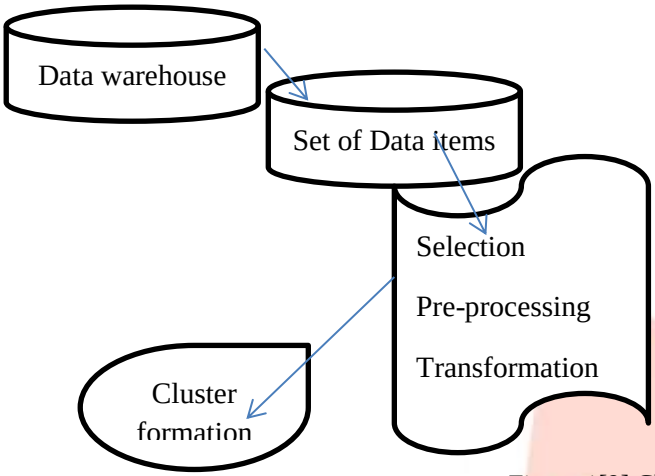

Figure 1[9].Cluster formation mechanism

Fuzzy C-means clustering to user-based CF

At the first step of the system[10], dataset is divided to fivefold cross-validation sub sets.

We will have five different results based on the five different testing subsets. We will consider average of these results. In User Clustering process, we select some clustering methods such as: K-means, SOM and fuzzy C-means, in order to group together users in clusters, minimizing the dissimilarity between elements assigned to the same cluster. Each clustering attempt will run individually with different number of clusters set in order to find the clustering models which can provide the highest rate of recommendation accuracy. Fuzzy C-means is used in order to provide a different membership degree of every user that belongs to different clusters. In order to define what users to consider in the prediction step we apply a defuzzification method. In another word, every user is assigned to k clusters with different degrees of membership, so a defuzzification process is necessary.

After finding clusters, the neighbor finding step with defining similarity measure between two users elicits the information that correlate users. So the most similar k users are selected as neighbors to contribute to the prediction process. In which, users whose similarity weights satisfy a pre-defined threshold can be chosen as neighbors. While the next step predicts ratings for items similar to rated ones by neighbors.

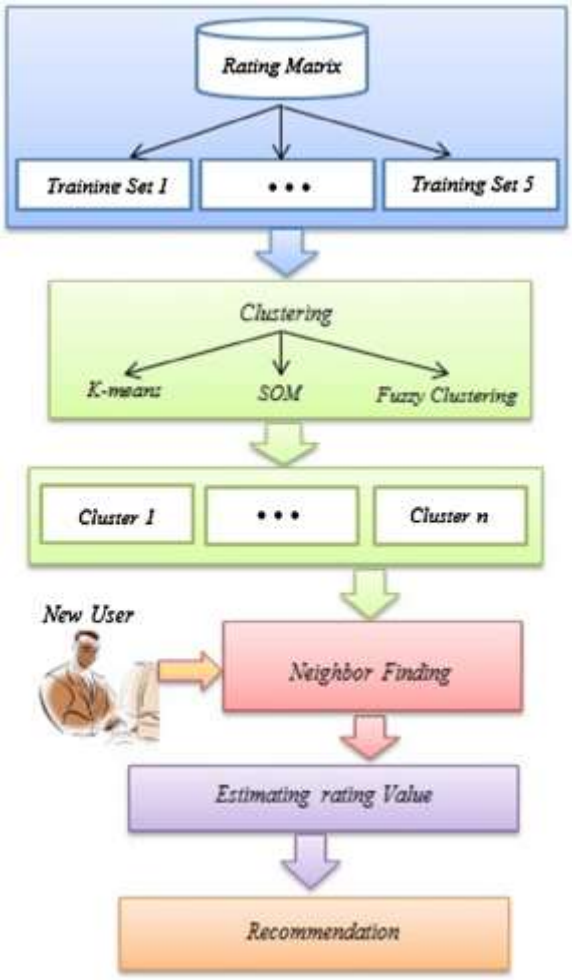

Figure 2[10]. Experimented model

IV. CONCLUSION

Machine learning algorithm is used to predict the outcome of data processing. Nowadays, recommender systems are widely used in e-commerce and other domains. It helps to solve the problem of data overload and enables user to have access to products which are not available to users. In this paper we have discussed the three recommendation techniques with their challenges. Here we highlighted Grey sheep challenge and strategies used to improve performances. Various machine learning algorithms used in measuring quality and performance of recommendation algorithm.

V. REFERENCES

[1] Gediminas Adomavicius ,Member, IEEE, and Alexander Tuzhilin, Member, IEEE." Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extension", 2005.
[2] Shano Solanki, Dr.Shalini Batra."Recommender System using Collaborative Filtering and Demographic Characteristics of Users".
[3] Qing Li, Byeong Man Kim. "Clustering Approach for Hybrid Recommender System",2003.
[4] Rokach, Lior and Oded Maimon "Clustering methods." Data mining and knowledge discovery handbook. Springer US, 2005.
[5] Zhang, et al. "agglomerative clustering via maximum incremental path integral." Pattern Recognition(2013).
[6] Robert M Bell and Yehuda Koren "Improved Neighborhood-based Collaborative Filtering" .
[7] Lika B., Kolomvatsos K., Hadjiefth-ymiades S. (2014). Facing the cold start problem in recommender systems. Expert Systems with Applications.41, 2065–207
[8] Hybrid personalized recommender system using centering-bunching based clustering algorithm by SK Shinde, U Kulkarni- Expert system with Applications,2012
[9] Santhosh Kumar Uppada "Centroid based clustering algorithms".
[10] Hamidreza Koohi , Kourosh Kiani "User based Collaborative Filtering using fuzzy C-means", 2016.