DOI: http://dx.doi.org/10.26483/ijarcs.v11i3.6616

Volume 11, No. 3, May-June 2020

ISSN No. 0976-5697

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

CLUSTERING AND TECHNIQUES USED IN COLLABORATIVE FILTERING – AN OVERVIEW

Sanjana Roy
Student, BCA
The Heritage Academy
Kolkata

Riddhima Shome
Student, BCA
The Heritage Academy
Kolkata

Snigdha Roy
Student, BCA
The Heritage Academy
Kolkata

Asmita Majumder
Student, BCA
The Heritage Academy
Kolkata

Madhurima Banerjee
Asst. Prof., BCA
The Heritage Academy
Kolkata

Abstract: The purpose of this paper is give an overview of the concept of clustering used in recommendation systems, to study the different kind of clustering approaches and techniques getting used in recommendation systems and how implementation of the clusters vary. The paper highlights the pros and cons of using clustering technique in collaborative filtering.
Keywords: Collaborative filtering, Clustering, DBSCAN, Hierarchical Clustering

I. INTRODUCTION

Recommender Systems are created using Content Based Techniques, Knowledge Based Techniques or using Collaborative Filtering (CF).

Techniques in RS can be divided into two categories: memory-based and model-based algorithms. Memory-based algorithms operate on the entire user-item rating matrix, while model-based techniques use the rating data to train a model and then the model will be used to derive the recommendations.

CF is implemented by finding users similar to the active user. These similar users are called neighbors of the active user. The item ratings of the neighbors are calculated and then the active user is recommended with the items that he/she has not rated but is most likely to rate high as his neighbors have liked the items. The similarity between users is often derived from the intersection of the items rated by the users.

Sparsity is a concern for Collaborative Filtering because users rate a very small subset of the items, therefore, finding accurate neighbors can be difficult at times.

Efficiency is another issue with Collaborative Filtering. CF has to compute the similarity between every pair of users to determine their neighbourhoods, which is the traditional approach for collaborative filtering. However, it becomes computationally expensive in a system with huge database. Moreover, efficiency further decreases when new user enters a system.

To overcome the above isuue of Collaborative Filtering, clustering is applied. When clustering technique is applied, it reduces the sparsity and improves the scalability of the systems. With clustering, efficiency of the system also enhances.

Different clustering techniques have been proposed by various researchers in a quest to improve the quality of recommendation.

II. COLLABORATIVE FILTERING

Recommender system is used to suggest items or products to users after measuring users' tastes and needs.

Collaborative Filtering is a well-known technique used in recommender systems. In collaborative filtering, the characteristics of the items or products are not required to process a recommendation. The recommendation is made on the basis of users' previous choices and choices of his peers. The peers are those users in the list those whose choices matches with the user's the most.

There are three steps involved in the collaborative filtering process. First is to calculate similarity between every pair of users and second, is to calculate the target user's rating for items those the user has not rated taking into consideration the ratings for such items given by similar users found in the first step. Third step is to recommend Top-N items to the target user.

Similarity between items or users can be calculated using following methods:

- Cosine Similarity
- Euclidean Distance
- Jaccard Similarity

- Pearson's Correlation

Challenges of Collaborative filtering can be identified as

- Sparsity
- Scalability

Calculating similarity between users employ the idea of finding similarity quotient between every pair of users. This process can be very expensive which is why clustering techniques are being proposed to make the recommendation process less costly.

Clustering is the process of grouping objects in such a manner that the objects belonging to same group are closer to each other with respect to similarity pertaining to some criteria.

Using clustering, means that the user list is grouped on basis of criteria which leads to reduction of the user list, or in other words, since the user list gets fragmented, the number of users per cluster is much less than the total number of users in the dataset. The user-item rating matrix gets divided into sub matrices, which reduces the number of comparisons required to produce result, thus accelerating the process of recommendation.

Employing clustering techniques, however, can have the disadvantage of taking a toll on the accuracy of the recommender system. Therefore the challenge is to use a clustering technique that does not affect the accuracy considerably. The parameters that are used to detect the accuracy of recommendation are

Precision: Ratio of relevant items recommended to the number of items recommended.

Recall: Ratio of relevant items recommended to the total number of relevant items available.

III. RESEARCH METHODOLOGY

This study is an explanatory study that leads to understandings of role of clusters and techniques used in recommender system. Recommender system is a popular area of research and researchers are proposing various methods and techniques to optimise recommendation and to make the result of their recommendation more accurate. In this paper, four papers have been taken at random where researchers have proposed different approaches for recommendation.

IV. CLUSTERING AND TECHNIQUES USED IN VARIOUS RECOMMENDER SYSTEMS

A. Case I

[1]Phum M. C., et al proposed a clustering technique based on social networking. Their method is based on the concept that user behaviour is impacted by their social relationships.

In the algorithm, they have represented the user-rating matrix as a network G= (V, E). Where V is the set of users and E is the social relationship between the users. This graph represents the social network.

The clustering is applied on G using the rule that:

If users are clustered such the v1,v2,v3…..vN represents all the clusters, then

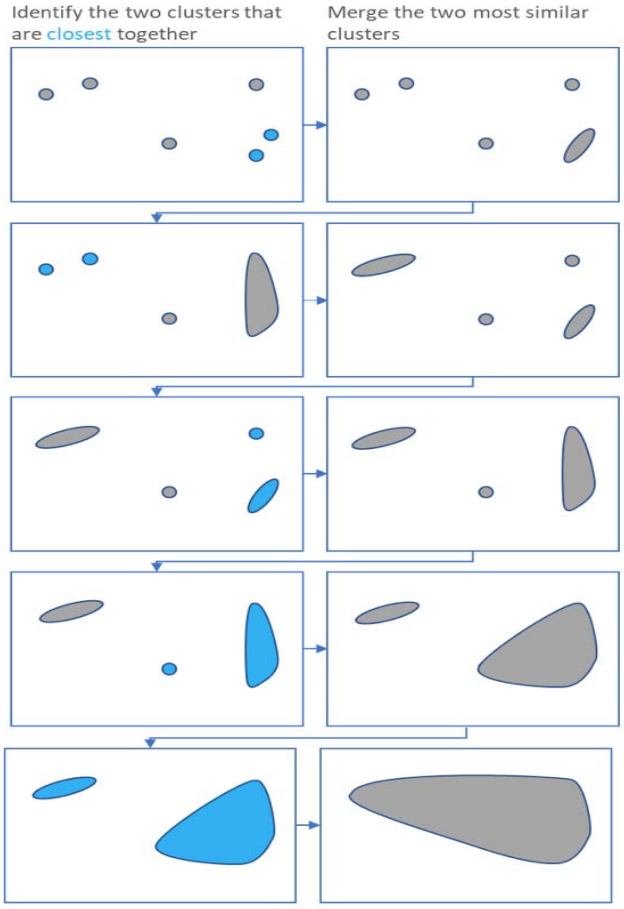
i) $v_i \cap v_j = \phi$, i.e., the clusters should be disjoint and should not be having any overlapping members

ii) $v1 \cup v2 \cup v3 \cup \ldots \cup vN = V$, i.e, union of all the clusters should give back the set all the users in the system.

With the clustering, the collaborative filtering is applied on each of the clusters.

There are two ways to cluster networks, graphical clustering and hierarchical clustering. The algorithm uses hierarchical clustering since that avoids declaring any predefined structure of the network.

Hierarchical clustering starts with nodes where each node is a cluster and then at each step closest clusters are agglomerated together to create a hierarchical structure. Depending on agreed criteria, hierarchical clustering can stop at any level. Measure used in this case is modularity. The algorithm clusters nodes which would result in greatest increase in modularity, the measure of quality of the cluster. The time complexity of the method used in O (Nlog2N). In [6], Book T has explained hierarchical clustering using the following diagram.



In their work, Phum M.C. et al, used the above clustering system to recommend events to researchers for publication of research papers, or events they would like to attend or would be relevant for them. The clusters they have used uses the social network of a researcher. Out of various social networks available, the authors have used social network of co-authors for clustering.

The measure used to create the user matrix is the rating of the researchers on the venues and events they have already attended.

Now, a researcher rates only those events or venues where they have participated. Reason for non-participation in an event might be wide spread. However, when the researchers are clustered on the basis of their social networks that gives an

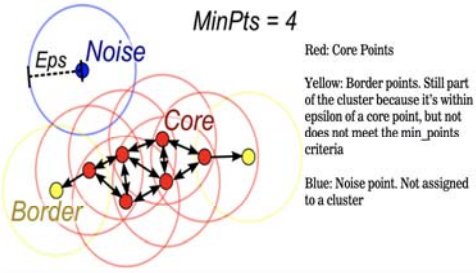
insight about what are the events a particular community of researchers appreciate and would like to attend.
In the above method of clustering, the precision and recall has improved compared to the traditional system, but the improvement is not very significant. Authors have pointed out that the ratings are based on venue, which might not show the true picture, if criteria of rating is changed, may be the results would show a considerable improvement. However, never the less the clustering process has shown improvement over traditional CF method.
From point of view of the patients as well, the recommender system can help the service seekers to get recommendations on more drugs, tests, and treatments.

B. Case II

[2]Das. J, et al has proposed a clustering based recommender system using principals of voting theory. They have used the above algorithm to build a recommender system for movies. The user is asked to give four genre preferences, and the algorithm uses DBSCAN method to cluster the users.
DBSCAN clusters the objects depending on two criteria [3]:

1. Eps: Epsilon Neighborhood where Epsilon € is measure of the radius of the cluster. In other words, all objects lying within the eps radius is considered to be inside a cluster
2. Density of a point, i.e., how many objects would belong to a cluster.

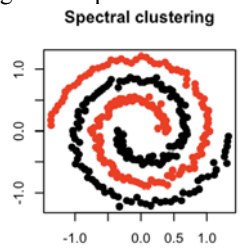
In [7] DBSACN has been pictorially represented as follows. The image is self-explanatory and nicely explains DBSCAN clustering system.



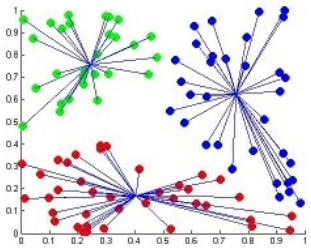
For the recommender system, the users are clustered on the basis of the genres of the movies they prefer.
For each user, genres are given preference weights ranging from 0 to 1, and weights have been assigned to each genre randomly according to order of preference, in other words, more preferred genres have got higher weight than less preferred genres. After assigning weight, they have been normalized.
For creating the clusters, each user is represented as a vector having k dimensions, which is a vector of his genre preferences and k is the number of movie genres present in the dataset.
The values in the vector are normalized weights of the corresponding genres for that user, so that added sum of the preferences is 1.
Euclidean distance between the vectors are calculated to put the users in specific clusters.
After forming the clusters, voting theory has been applied to individual clusters to rank the genres in the clusters.
Out of the various voting techniques, Borda count voting rule has been used in this algorithm. Borda count is a technique where, points are assigned to each of the candidates based on their ranking and then the points in the ballots are added to choose the candidate having the highest rank [4].
For implementing the Borda Count, the authors have created a ranking vector for every cluster. The ranking vector has dimension equal to number of genres, i.e. k which is same as the dimension of the user vector. Borda Count for each of the genres corresponding to the cluster is put in the ranking vector. The genre getting highest ranking is the winner in the cluster. Now when a user belongs to a cluster say A, the movies recommended to the user are movies belonging to the winning genre of the cluster A.
This clustering technique has also considered the cold start problem.
When a new user enters the system, he specifies four preferred genres, this helps to form the preference vector of the user, which is used to put the user into a particular cluster, now, top-10 movies of the cluster is recommended to the user.
The results of the clustering system as depicted in the paper shows there is considerable reduction in the runtime of the algorithm, as the number of clusters increases, the recommendation time reduces and precision and recall also shows high value as the dataset is clustered. Thus the clustering system reduces execution time without much effecting the quality of recommendation.

C. CASE III

[5]Trans C. et al has proposed a recommender system based on concept of incentives and penalty.
They have used the above algorithm to build a recommender system for movies.
The clustering technique they have used is spectral clustering and FCM.
Following is a diagram of spectral clustering [8]:



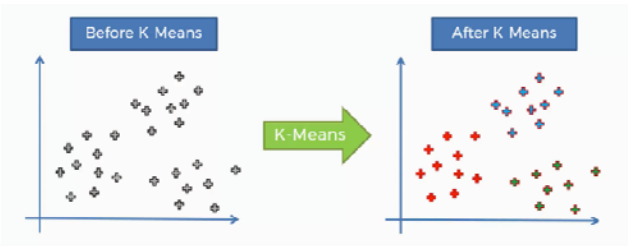
We are aware of the fact that if points lie close to each other, they would be clustered together. In case of spectral clustering, even if the distance between two points is less, if they are not connected, they are not clustered together.
The authors have also used FCM clustering technique in their research. FCM stands for Fuzzy C-Mean clustering algorithm [9] [10].
FCM algorithm measures distance of each data point from the cluster centre and assigns membership of each data point to clusters depending on the distance measured. Closer the data point to the centre of the cluster, more strong is its membership in that cluster. One data point can belong to more than one cluster and summation of membership of each data point should be equal to one.



The authors have shaped their research depending on item model. The essence of the research is to increase the recall and F1 score.
It is known that a recommender system suggests an item to a target user depending on a pre-decided threshold value of predicted rating. If the predicted rating of the target user for an item is greater than or equal to threshold, the item is recommended to the user, if the predicted rating goes below the threshold, it is not recommended to the user, but in actual scenario, the user may like the product, and if the user does like the product, then recall and F1, which are measures of prediction quality score decreases.
To enhance the prediction quality, the authors in the research have suggested a system of penalty or incentive which is set according to preference tendency of the users.
User clusters are created for each item I, and incentive and penalty is calculated to decide whether the item is to be recommended to the user or not.
γ is a positive number.
There are 3 positive system parameters considered in this process, which have been found with intensive testing: α β γ. $\alpha >= \beta$. Let Average preference of an item in a cluster is X and predicted rating of the item for a target user is UR.
If $X > \gamma$ and $UR > \beta$ then an incentive is added to the item i and it is recommended to the user.
If $X < \gamma$ and $UR > \alpha$ then also an incentive is added to the item i and it is recommended to the user.
For any other case, a penalty is added to the item i and it is not recommended to the user.
The authors have applied this method both on user based and item based Collaborative Filtering and the above method has shown a good improvement in recall and F1 score over baseline method.
In the results in both cases, user based and item based, it is seen that when FCM has been used as the clustering technique, a better performance result was obtained than when spectral clustering is used.

D. CASE IV

[11]Wasid, M. and Ali R. in their research have proposed a multi criteria clustering approach. Multi criteria clustering approach, known as MCRS allows to represent the preferences of users on several aspects of the items. Thus multi criteria helps us to exploit more specialized or detailed preferences of the users. Now this approach brings with it two fold challenges:
First, how MCRS can be incorporated in a traditional collaborative filtering
Second, how similarity between users can be calculated with multiple criteria rating.
K-Means clustering technique has been used by the authors to incorporate MCRS, since K means is one of the algorithms suited for multidimensional dataset.



The above diagram shows that how K-Means takes a scattered dataset and clusters the dataset into predefined number of clusters [14].
[12][13]K Means clustering is an example of unsupervised algorithm where datasets are provided as input and the dataset is clustered into K clusters each having a centroid. The algorithm first assigns random centroids and then optimizes the position of the centroid. The position of the centroid has to be determined wisely, as the output changes with position of the centroid.
In this approach, the authors have used Mahalanobis distance to find distance between users in clusters. According to them, since Mahalanobis distance considers variance and co variance along with commonly rated items by the users it can produce more similar neighbors than other distance calculating methods.

V. CONCLUSION

Here we have reviewed four research works, each using different clustering systems to implement collaborative filtering: Hierarchical clustering, DBSCAN, Spiral clustering, Fuzzy C Means and K-means.
Clustering has been implemented on overall rating as well as on multi criteria rating.
The common analysis of all the researchers is that, since clustering reduces the user dataset, it helps in reducing the time it takes to recommend items to users, but at the same time, since the dataset has been reduces it at times takes toll on the accuracy of the recommendation.
Research is also being done on the various methods that can be implemented on the clusters such that the recommendation accuracy can also be enhanced. We have come across the concept of voting theory and incorporation of penalty and incentive on clusters to work on the accuracy of recommendation.
Recommendation is influencing commerce from every dimension and research is being done to recommend faster and with more accuracy, and clustering has proved a welcome concept and researchers are making clustering the back bone of research on collaborative filtering by applying various techniques on clustered data.

VI. REFERENCES

[1] Pham, M.C., et al. 15.02.2011, "A Clustering Approach for Collaborative Filtering Recommendation Using Social Network Analysis" in Journal of Universal Computer Science, vol 17 no. 4, pp 584 - 604

[2] Das, J., et al., 2014, "Clustering-Based Recommender System Using Principles of Voting Theory" in 2014 International Conference on Contemporary Computing and Informatics, pp 231 - 235

[3] Prado, K., 02.04.2017, How DBSCAN works and why should we use it? , viewed 27th May, 2020 from https://towardsdatascience.com/how-dbscan-works-and-why-should-i-use-it-443b4a191c80.

[4] "Module 7: Voting Theory", lumen, Mathematics for Liberal arts, viewed 27th May, 2020 from https://courses.lumenlearning.com/waymakermath4libarts/chapter/borda-count/

[5] C. Tran, J. Kim, W. Shin and S. Kim, 02.05.2019, "Clustering-Based Collaborative Filtering Using an Incentivized/Penalized User Model," in IEEE Access, vol. 7, pp. 62115-62125

[6] Book, T., n.d., "What is Hierarchical Clustering", viewed 27th May, 2020 from https://www.displayr.com/what-is-hierarchical-clustering/
[7] Lutins, E., 06.09.2017, "DBSCAN: What is it? When to Use it? How to use it.", viewed 27th May, 2020 from https://medium.com/@elutins/dbscan-what-is-it-when-to-use-it-how-to-use-it-8bd506293818
[8] Doshi, N, 05.02.2019, "Spectral clustering. The intuition and math behind how it works!", viewed 27th May, 2020 from https://towardsdatascience.com/spectral-clustering-82d3cff3d3b7
[9] "Fuzzy c-means clustering algorithm", viewed on 27th May, 2020 from https://sites.google.com/site/dataclusteringalgorithms/fuzzy-c-means-clustering-algorithm
[10] " A tutorial on Clustering Algorithms", viewed on 27th May, 2020 from https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/cmeans.html
[11] Wasid M., Ali R., 2018, "An Improved Recommender System based on Multi-criteria Clustering Approach" in 8th International Congress of Information and Communication Technology (ICICT- 2018), pp 93 – 101
[12] Garbade, M.J., 13.09.2018, "Understanding K-means Clustering in Machine Learning", viewed on 27th May, 2020 from Understanding K-means Clustering in Machine Learning https://towardsdatascience.com/understanding-k-means-clustering-in-machine-learning-6a6e67336aa1
[13] " A tutorial on Clustering Algorithms", viewed on 27th May, 2020 from https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html
[14] SuperDataScience Team, 29.09.2018, "Self Organizing Maps (SOM's) - K-Means Clustering (Refresher)", viewed on 28th May, 2020 from https://www.superdatascience.com/blogs/self-organizing-maps-soms-k-means-clustering-refresher

