Hybrid collaborative filtering model using hierarchical clustering and PCA

[1]Ashutosh Lokhande, [2]Pooja Jain

[1]Research Scholar, [2]Assistant Professor
Computer Science,
Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India

***Abstract*: Recommendation system uses different types of algorithms to make any type of recommendations to user. Collaborative filtering recommendation algorithm is most popular algorithm, which uses the similar types of user with similar likings, but somewhere it is not that much efficient while working on big data. As the size of dataset becomes larger then some improvements in this algorithm must be made.**

Here in our proposed approach we are applying an additional hierarchical clustering technique with the collaborative filtering recommendation algorithm also the Principle Component Analysis (PCA) method is applied for reducing the dimensions of data to get more accuracy in the results. The hierarchical clustering will provide additional benefits of the clustering technique over the dataset and the PCA will help to redefine the dataset by decreasing the dimensionality of the dataset as required. By implementing the major features of these two techniques on the traditional collaborative filtering recommendation algorithm the major components used for recommendations can be improved.

The proposed approach will surely enhance the accuracy of the results obtained from the traditional CFRA and will enhance the efficiency of the recommendation system in an extreme manner. The overall results will be carried out on the combined dataset of TMDB and Movielens, which is used for making recommendations of the movies to the user according to the ratings patterns created by the particular user.

***Index Terms*: Recommender system; Collaborative filtering recommendation algorithm; Hierarchical Clustering; Principle Component Analysis; Big Data.**

I. INTRODUCTION

As the data is being generated over the internet by the users and is called as big data, is an important topic to make research. The day by day evolution in the technology and the services provided to users makes it important to do more and more research and provide the solutions of the problems in the existing systems. Nearly each and every organization belonging to any of the industry requires an efficient system for their users to make recommendations and it is a difficult task to make proper recommendations. Recommendation system is a way to provide the expected likings of the users as a suggestion to user at a very complex information platform. Three Filtering algorithms i.e. Content based filtering, Collaborative filtering and Hybrid filtering are the filtering algorithms which are used to make any recommendation system. Among all these three techniques the mostly used algorithm is the collaborative filtering recommendation algorithm (CFRA) which is more popular and has been used by major providers and consumers of big data such as: eBay, Amazon and Facebook. It is also seen that the collaborative filtering works efficiently irrespective of the size of data and therefore it is suitable for using it under the scenario of big data. Collaborative filtering makes recommendations on the basis of the neighbor user's likings and ratings which are similar to the user for whom the recommendations are being made.

Many researchers have proved that on the traditional CFRA if we add the features of K means clustering technique then this will improve the working efficiency of the CFRA and to improve it more we are applying Hierarchical clustering by replacing the traditional K means clustering. It is also proved that Hierarchical clustering works more efficiently than K means as the number of clusters need not to be provided at the beginning in the hierarchical clustering. Also it is proved that the additional PCA applied on this will make the accuracy of the recommendation higher. It is to be noted that the PCA should be applied before applying the clustering technique as the dimensions will be reduced the clustering will become easier.

II. LITERATURE SURVEY

Authors have proposed an algorithm which is more efficient as compared to traditional collaborative filtering recommendation algorithm. They used two additional techniques over CFRA i.e. K means clustering and PCA to improve the traditional CFRA. They proposed two algorithms which are combinations of the K means clustering and another is the combination of K means clustering along with PCA. Both the algorithm worked well than the traditional CFRA. The Netflix dataset was used for the experiment [1].

The authors have firstly introduced the Frechet distance to calculate the similarity of AIS trajectories. Secondly, Principal component analysis (PCA) is used for decomposing the distance matrix obtained in the very initial step to state the number of clusters in last level of clustering. Finally, the above distance matrix and clustering number are fused with the traditional hierarchical clustering algorithm [2].

In this paper the author proposed new methods to automatically mention the number of clusters in agglomerative hierarchical clustering. Here they showed two approaches for this purpose, one is to use a variation of cluster validity measure, and another is to use statistical model selection method like BIC. Through this the authors tried to deal with the problem of the selecting appropriate number of clusters [3].

Authors have stated a formal method to determine the number of clusters, which is based on an agglomerative hierarchical clustering (AHC) algorithm. The new index and the method can calculate the clustering results, which are produced from the AHC and determine the optimal number of clusters which may be used for different types of datasets, such as linear, manifold, annular, and convex structures [4].

In this paper the authors examines the K-means approach of clustering and how the selection of primary seeding affects the result. Hierarchical algorithms are used as a base line and it is compared with a data set. The authors have evaluated k-means clustering and hierarchical clustering algorithm for pure numeric synthetic data set [5].

Authors have explained the types of clustering that are being used frequently over the internet. Here the Hierarchical clustering is explained in detail and the different types of the hierarchical clustering is also explained. And the comparison between clustering techniques is done. The authors measured the quality of clusters with the help of three parameters: Cohesion measurement, Silhouette index and Elapsed time [6].

All the authors in this paper stated six methods that collaborative filtering recommender systems may use to learn about some new users. In these techniques a sequence of items is selected for the collaborative filtering system to present it to every next new user for the purpose of rating. They tried to state that not only filtering should be applied but also the user’s interest should be kept in mind. They have studied the techniques thru offline experiments with a large preexisting user data set, and thru a live experiment with over 300 users [7].

An approach proposed by authors to remove the sparsity and the scalability problems of the collaborative filtering method. They tried to address both the problem at the once by using a new CF model, which is based on the Artificial Immune Network Algorithm (aiNet). The reason to choose this is because aiNet is capable of reducing sparsity and offering the feature of scalability of the dataset by describing the data structure, by including their spatial distribution and cluster interrelations. They have also reduced the sparsity rate and with aiNET they have applied k means clustering technique [8].

Authors presented their research on the profit factor of the selling over the websites. Normally the recommendations made for the interests and likings of the user but the profit of the seller is not considered. For this they proposed two profitability-based recommender systems called CPPRS (Convenience plus Profitability Perspective Recommender System) and HPRS (Hybrid Perspective Recommender System). Successful implementation of the proposed approach is carried out [9].

They all proposed an approach to provide accurate online predictions, a recommendation system is developed by authors which is called WebPUM, an online prediction which uses the Web usage mining system and propose an approach for classifying the user’s navigation patterns to predict future intentions of that user. They studied the system on CTI and MSNBC datasets and carried out a successful implementation [10].

Authors has given simple article of the study of the amazon website. They have discussed about the growth of the amazons recommendation system. Also shown that collaborative filtering is used by amazon for making proper recommendations. They studied the two decades of the amazon for making recommendations [11].

Authors presented an approach to remove the two major problem of collaborative filtering i.e. the sparsity and scalability. For removing these problems they have proposed a new method that is finding of a neighbor user method which has been implemented from the subspace clustering approach. This method works as, the authors finds out different subspaces of the items rated under the categories some of the category like Interested, Neither Interested, Nor Uninterested, and Uninterested. The proposed method tested across the Movielens 100K, Movielens 1M and Jester datasets to make some comparison with the traditional techniques. The results express that the proposed method can increase the working capability of the Recommender Systems [12].

They have projected a compression scheme for large-scale data sets which leads to both computational time and memory benefits in unsupervised learning tasks such as PCA and K-means clustering. The main factor of the approach given by them is that it requires just one pass over the data and for this they used randomized preconditioning transformation, which makes it possible to apply to streaming and distributed data settings [13].

The authors have presented a paper that provides an overview of many different machine-learning algorithms which is being used in context of big data analytics. They tried to remove the gap between the researchers that many machine algorithms are provided, so to make it clear to researchers they have presented such an overview of some different machine learning algorithms. The paper presents a simple study of big data analytics, with keeping a special focus over data-intensive distributed machine-learning algorithms for big data. This review is theoretical in nature [14].

In this the author presented a paper which deals with the similarity measure technique. Here he used different similarity measures for validating their performance. In this paper the various similarity measures that are being used in collaborative filtering are analyzed for evaluating their performances. For doing this they have used Apache Mahout which helped to make proper analysis in an efficient manner [15].

Author’s has made their analysis on the data present over internet by using hadoop framework. They used the data like ratings, likings, reviews of the users to make recommendations to users, for this they used hadoop framework. They first filtered the data and then applied this data under mahout interface. This overall work was done for a movie recommendation using movielens dataset [16].

In this the authors has analyzed the MyMediaLite library/API, which provides support for the algorithms. This library address collaborative filtering for the two scenarios of data that are prediction of ratings (e.g. on a scale of 1 to 5 stars) and prediction of items from positive-only implicit feedback. Also they addressed some future work on this library [17].

The authors presented a keyword awareness recommendation which uses the keywords to make recommendations. Collaborative filtering was used for the analysis and is implemented on hadoop. The real world data is used to make this experiment and by KSAR the accuracy of recommendation is improved in an extreme [18].

III. PROBLEM DOMAIN

The traditional collaborative filtering recommendation algorithm is having lack of accuracy and efficiency as this uses formal method of filtering which makes it inefficient to use at alone. In terms of recommendation made by the collaborative filtering algorithm it may be concluded that the algorithm needs many more improvements. By implementing traditional collaborative filtering recommendation algorithm we get less accuracy which makes it typical to use and inefficient to apply on huge datasets i.e. Big Data. Dealing with big data the less accuracy makes it inappropriate and less accurate. As applying this algorithm on huge amount of data in real world applications the less accuracy will not be efficient for making recommendations to users. The numbers of attributes which are available are totally considered for extracting information to recommend items to users which makes the collaborative filtering recommendation algorithm inefficient. Also the higher the number of attributes used to make recommendations, results in higher computing time and higher number of comparisons to be made. The overall dimensions included for making recommendation should be removed as per the requirement.

Apart from this the k-means clustering applied previously with the collaborative filtering algorithm can be replaced by different clustering technique. There are some drawbacks that can be seen in the k-means clustering technique which may be overcome by replacing this clustering technique with the newer one. In the k-means clustering the numbers of the clusters that should be made need to be defined at the start of the algorithm which makes it inefficient to use if the numbers of the clusters are not properly defined.

One more thing to be noted, that is the dimensionality of the given dataset should be less in number to lower the comparisons that will be made at the time of execution. The more the number of the dimensions to evaluate the results, makes the accuracy lesser and requires more time to make recommendations to the user. Hence to reduce the number of attribute or the dimensionality of the dataset is major task.

IV. PROPOSED APPROACH

The problem observed in the previous algorithm can be removed by replacing the existing techniques by newer techniques. As in the previous, the algorithm combines the K-means clustering technique with the PCA as dimensionality reduction technique. Combining both this techniques in the collaborative filtering algorithm was a solution proposed earlier by the authors.

Here we have proposed a better clustering technique as compared to the k-means clustering, while keeping the PCA as earlier it was used. The k-means clustering can be replaced by the hierarchical clustering as it is better clustering technique to work on. The PCA will be used as the dimensionality reduction technique to decrease the dimensionality of the data.

The Hierarchical clustering will provide better results in comparison to the k-means clustering, as stated that in hierarchical clustering there is no need to define the number of clusters at the beginning of the clustering. Defining the required number of clusters after applying the hierarchical clustering will make it feasible to break the clusters as per the dataset. But before applying the clustering technique on the dataset the dataset should be improved. If the Input to the algorithm will be accurate then the obtained output will be more efficient. So, to improve the input dataset the dimensionality reduction should be done and to do this the PCA have to be applied on the dataset.

In final words we are going to apply the PCA on the dataset before giving it as input and after getting the principal components this are given as input to the hierarchical clustering. The collaborative filtering algorithm will firstly perform the PCA and after that the hierarchical clustering is applied and the final recommendations are made. Hence in this way the collaborative filtering algorithm can be improved and the recommendations can be made accurate.

IV.1 Algorithm For Proposed Approach

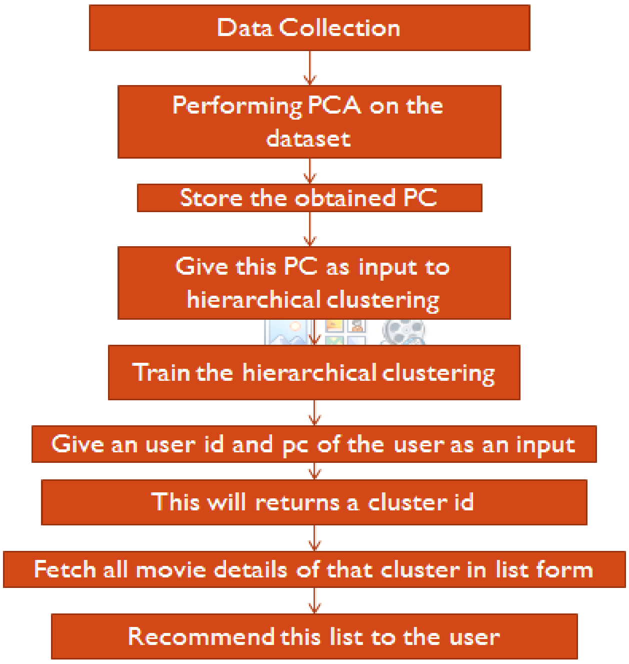
The proposed algorithm using both the techniques, the first one is the PCA which will help in reducing the dimensions of the given dataset and the second one is the clustering technique which is the hierarchical clustering. Here in the our algorithm we are applying the PCA at first because it will reduce the dimensions of data and after that the hierarchical clustering will be performed on the obtained principal components. The working algorithm is as follows:

Step 1 : Data collection - collect the movie related data like name, rating etc. in the form of csv file.
Step 2 : Data pre-processing - perform manual data analysis and eliminate the feature which is less correlate to other feature.
Step 3 : Perform PCA (principal component analysis) on the data and save the data in to csv file.
Step 4 : Define hierarchical clustering (agglomerative) model.
Step 5: Train the hierarchical clustering (agglomerative) model on the data.
Step 6: Take the one user input and apply PCA on that.
Step 7: Perform the prediction in the input it give the cluster id.
Step 8: Fetch all the movie detail which belong to this cluster id and make the list of it.
(This list is recommended movie list)

IV.2 Flowchart of the Proposed Approach

Below we have given the flowchart for the proposed approach which will help in understanding the flow of the steps performed:

Fig.1 Flowchart for the proposed system


Data Collection
Performing PCA on the dataset
Store the obtained PC
Give this PC as input to hierarchical clustering
Train the hierarchical clustering
Give an user id and pc of the user as an input
This will returns a cluster id
Fetch all movie details of that cluster in list form
Recommend this list to the user

V. EXPERIMENTAL RESULTS AND EVALUATION

For making the analysis of the proposed approach we have used the TMDB dataset along with the Movielens Dataset. The data about movies is taken from the TMDB dataset and the ratings pattern and user details are combined from the Movielens dataset. The combination of both the datasets are used for the analysis purpose. The ratings are provided from 1 to 5.

The experiment was carried out to evaluate the accuracy of the recommendations produced by the algorithm we have proposed in our paper. The accuracy term is calculated in this experiment by which the comparison between the proposed and the existing algorithm can be made.

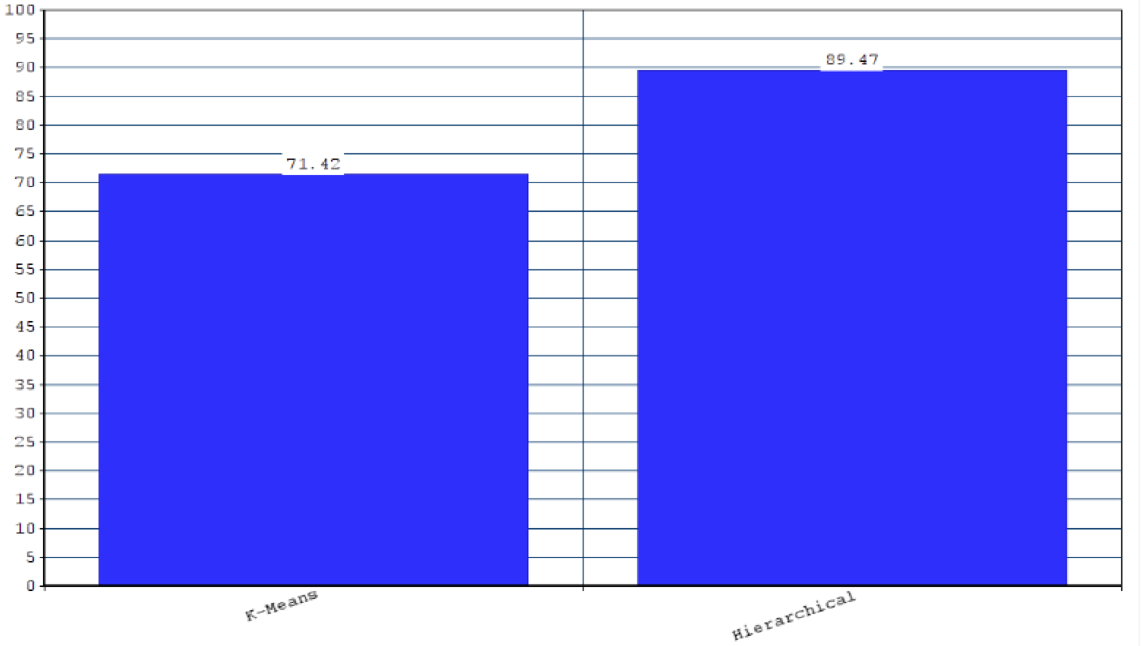
We are applying this data on the previous collaborative algorithm with pca and kmeans and the results are obtained, so the accuracy of the previous algorithm is calculated.

Accuracy = ({Relevant Document} intersection {Retrieved Document} / {Relevant Document}) *100

Now the proposed algorithm with hierarchical clustering is taken for analysis. The collaborative filtering algorithm along with pca and hierarchical clustering is analyzed over the same data. This algorithm's accuracy is compared with the existing algorithm.

The experiment clearly results in an increase in the accuracy of the recommendations made by our proposed algorithm. The results are compared between both the algorithms using k-means clustering with pca and hierarchical clustering with pca in terms of accuracy are shown in the following graph:

Fig. 2. Accuracy results for both the algorithms



The Fig.2 clearly concludes that the proposed hierarchical clustering works much better as compared to the previously used k-means clustering. The results in terms of accuracy of the proposed algorithm is higher than the earlier clustering technique. So it is better to use the Hierarchical clustering with pca on the collaborative filtering algorithm as compared to earlier one.

VI. CONCLUSIION AND FUTURE WORK

The proposed research work observes the recommendations made by the system to the user. The entire work is done by the hierarchical clustering technique along with the pca, by which the accuracy of the system is evaluated. The accuracy of the system is evaluated by the intersection of the recommended movies with the ratings made by the user for the movies earlier. The experiment shows better results from the earlier algorithms.

In future we can use other datasets to carry out the experiment. The other parameters apart from the accuracy can be tested. Different clustering technique may be applied to improve the algorithm.

REFERENCES

[1] Zarzour, H. , Maazouzi, F. , Soltani, M. , and Chemam, C., 2018: An Improved Collaborative Filtering Recommendation Algorithm for Big Data. © IFIP2018 Springer International Publishing AG 2018, CIIA 2018, 660–668.
[2] Jing Cao, Maohan Liang, Yan Li, Jinwei Chen, Huanhuan Li, Ryan Wen Liu and Jingxian Liu, 2018: PCA-Based Hierarchical Clustering of AIS Trajectories with Automatic Extraction of Clusters. IEEE 3rd International Conference on Big Data Analysis. 978-1-5386-4794-3/18/$31.00 ©2018 IEEE.
[3] Ryosuke Abe, Sadaaki Miyamoto and Yukihiro Hamasuna, 2017: Hierarchical clustering algorithms with automatic estimation of the number of clusters. IFSA-SCIS 2017, Otsu, Shiga, Japan, June 27-30, 2017
[4] Shibing Zhou, Zhenyuan Xu, and Fei Liu, 2016: Method for Determining the Optimal Number of Clusters Based on Agglomerative Hierarchical Clustering. IEEE Transactions On Neural Networks and Learning Systems. 2162-237X © 2016 IEEE.
[5] P. Praveen and B. Rama, 2016: An Empirical comparison of Clustering using Hierarchical methods and K-means. International Conference on Advances in Electrical, Electronics, Information, Communication and Bio-Informatics (AEEICB16). 978-1-4673-9745-2 ©2016 IEEE.
[6] Nisha and Puneet Jai Kaurt 2015: Cluster Quality Based Performance Evaluation of Hierarchical Clustering Method. 1st International Conference on Next Generation Computing Technologies (NGCT-2015). 978-1-4673-6809-4/15/$31.00 ©2015 IEEE.
[7] Rashid, A.M., Albert, I., Cosley, D., Lam, S.K., McNee, S.M., Konstan, J.A., Riedl, J., 2002: Getting to know you. In: Proceedings of the 7th International Conference on Intelligent User Interfaces - IUI.
[8] Acilar, M.A., Arslan, A,. 2009 : A collaborative filtering method based on artificial immune network. Expert Syst. Appl. 36(4), 8324–8332 .
[9] Chen, L., Hsu, F., Chen, M., Hsu, Y., 2008 : Developing recommender systems with the consideration of product profitability for sellers. Inf. Sci. 178(4), 1032–1048.
[10] Jalali, M., Mustapha, N., Sulaiman, M.N., Mamat, A., 2010 : WebPUM: a web-based recommendation system to predict user future movements. Expert Syst. Appl. 37(9), 6201–6212.
[11] Smith, B., Linden, G., 2017 : Two decades of recommender systems at amazon.com. IEEE Internet Comput. 21(3), 12–18 .
[12] Koohi, H., Kiani, K., 2017 : A new method to find neighbor users that improves the performance of collaborative filtering. Expert Syst. Appl. 83, 30–39.

[13] Pourkamali-Anaraki, F., Becker, S., 2017 : Preconditioned data sparsification for big data with applications to PCA and k-means. IEEE Trans. Inf. Theory 63(5), 1.
[14] Gupta, P., Sharma, A., Jindal, R., 2016 : Scalable machine-learning algorithms for big data analytics: a comprehensive review. Wiley Interdisc. Rev.: Data Min. Knowl. Disc. 6(6), 194–214.
[15] Bagchi, S., 2015 : Performance and quality assessment of similarity measures in collaborative filtering using mahout. Proced. Comput. Sci. 50, 229–234.
[16] Verma, J.P., Patel, B., Patel, A., 2015 : Big data analysis: recommendation system with Hadoop framework. In: 2015 IEEE International Conference on Computational Intelligence and Communication Technology.
[17] Gantner, Z., Rendle, S., Freudenthaler, C., Schmidt-Thieme, L., 2011 : MyMediaLite. In: Proceedings Of The Fifth Acm Conference On Recommender Systems - Recsys.
[18] Meng, S., Dou, W., Zhang, X., Chen, J., 2014 : KASR: a keyword-aware service recommendation method on mapreduce for big data applications. IEEE Trans. Parallel Distrib. Syst. 25(12), 3221–3231.