Parameter boosted approach to Collaborative Filtering based Recommender System

Sarthak Tandon, Deepak Jain1, Himanshi Agrawal1, Swasti Singhal1

***Abstract-*Because of the changing nature of human interests and preferences ,it is very difficult to make accurate recommendations to a user .This has been one of the major challenges in developing an accurate recommender system .This work proposes a collaborative filtering based method to make better recommendations by overcoming this challenge .Initially we will add some extra parameters, like genre of the movie or the mood of the user, to our. After that a clustering algorithm like K means clustering will be used to group similar movies together .Then ,on the basis of correlation between a set movies ,we can make recommendations to a user.**

Keywords— Recommender system, K-Means clustering, Collaborative filtering

**I. INTRODUCTION**

Recommendation system is a technology which is used to predict and thereby suggest most appropriate preferences to the users on the basis of their past activities and reviews. Basically, the data is collected that can be implicit in the form of user behavior or explicit in the form of user’s previous ratings. Movies and E-commerce websites are widely using the concept of recommendation system. It increases the revenue of the company and provide a great user experience.

**II. LITERATURESURVEY**

With the advent of Recommender systems in the 90’s the service providers were in huge profit.. Hence, variety of solutions were proposed to enhance the performance of Recommender System. Some solutions involved hybrid of algorithms, other introduced new parameters to be considered.

Most common used technologies were discrete matrix and explicit responses. So, in IEEE Conference 2019 [2]“*Movie Recommender System Based on Percentage of View*” a solution was proposed which took implicit opinion major (IOM) in account. Which consider percentage of view as its crucial parameters to increase the efficiency by 5 folds. So, it discussed the limitations of Content Based filtering,

Collaborative based filtering and Hybrid Based filtering and how percentage of view parameter could be accountable for it. A positive correlation between like and percentage of view establish to prove the parameter[[1]](#footnote-1) where the percentage of view varies linearly with the like probability.

Different algorithm was used such as Random Forest and Linear of evaluation through RMSE and turned out to be the solution to the problem of movie recommender system at an extent.

But every solution comes with another problem as the leader chose for the clusters sometimes may not be Regression. The Role of Percentage of view in memory based collaborative filtering were using the user-item-rate matrix and the KNN algorithm. Finally, such model could be evaluated as evaluation matric by finding mean absolute error given by:

***Mean Absolute Error = absolute difference of real & predicted value.***

This parameter proved to be the suitable for collaborative filtering by achieving minimum MAE. Furthermore, this concept could be improved by taking data on the large scale thus increasing accuracy of the system.

*According to IEEE Conference 2019,“Movie recommender system using K-means clustering and K nearest neighbor* “[6]the technology used are k means clustering, collaborative filtering and KNN algorithm. Basically, both the algorithms are used to get the best optimized result. In this technique we form the clusters of the similar data like user id, user info etc. Previously clusters formed were huge but the proposed technique is such that, the parameters are taken in a way to minimize the no. of clusters. To calculate the average rating, we form the clustered matrix. For implementation we use the different steps like collecting and preparing the data and then creating and training the model. NumPy and pandas library are used to pre-process the data. Within clustered sum of squared is used to find the number of clusters. RMSE is calculated which increases with the number of clusters. Lesser the error more precise the prediction.

*(Personalized real time Movie Recommendation system: Practical Prototype and Evaluation- 2020)[4].*Subsequently, more researches talked about the collaborative filtering being the mostly used algorithm for movie recommendation system. But collaborative filtering suffers from the issue of time complexity which in turn led to poor performance in real time scenarios. It was known that the problem with movie recommendation system was accounting the dynamic data and feedback. In order to meet the above problems, an algorithm based on K means and

Virtual User concept was introduced which clustered the user’s attributes and assigned a virtual opinion leader to each, representing each item/user in the cluster. Rather than using the whole user-item-rate matrix, algorithm helped in reducing the matrix. Such implication reduced the time complexity of the recommender system for predicting or suggesting the

movies to users. The result was compared to general concept.

(An efficient collaborative movie recommender system with cuckoo search 2016, Elsevier)[7] Clustering technique in recommending movies has been an indispensable approach for an efficient and precise result. But the challenges with the most clustering based algorithm is the computation in limited time and is conducted for large data matrix. To handle such problem to some extent, a hybrid model was proposed which was the fusion of K- means and bio-inspired optimization, cuckoo search algorithm. By using cuckoo search optimization method there was an appreciable improvement in time complexity and found efficient compared to other algorithms. Subsequently, the evaluation metrics such as root mean square error (RMSE), mean absolute error (MAE)[1,3,4], standard deviation (SD) was calculated and the method found to be the efficient among all the cluster-based methods for movie recommender system as the value of evaluation metrics were less as compared other popular algorithms. The limitation with this approach is that in case the first step of cluster formation is not done efficiently then the result could be inaccurate in nature. So future work may include some other bio-inspired optimization algorithm which is reliable in terms of performance and predictions.

(Solving Sparsity Problem in Rating-Based Movie Recommendation System, Springer 2017) Dealing with the problem of searching the most efficient movie recommendation system, most researchers suggest the usage of Clustering algorithms to generate the similarity matrix which would be further used to suggest movies to users of same taste. But the most concerned limitation of using this method is the sparsity of data as user merely acknowledge to rate the movies or services. This decreases the ratio of number of users rated to the total number of users, serving to problem of sparsity in the data. Sparsity not only makes it difficult to suggest a preferred movie but also decreases the efficiency of whole recommendation system. Afterwards, a solution to such problem was conceived which solved the issue with great extent. The approach utilized K-medoid clustering technique in place of K-means. The advantage of K- medoid over K-means is that the method deals with the absolute difference between the distances of points in the same cluster. Whereas in K-means clustering, the difference of squared distance is minimized. This approach helped in decreasing the sparsity of data and hence a better recommender system would be.

In this, Krista Rizman Zalik discusses about the K means clustering algorithm and its limitations[8]. The biggest drawback of this algorithm is that it requires the number of clusters to be pre-assigned. He also proposes a different version of the K-means algorithm which would not require the exact number of clusters beforehand. He was able to achieve these results by minimizing the cost function. The optimum number of clusters is found when the cost function is at its global minimum.

In this, Xin Luo, Mengchu Zhou, Fellow, IEEE, Yunni Xia, Member, IEEE, and Qingsheng Zhu developed an efficient matrix factorization method to develop a recommender system. They used Non-negative matrix factorization based collaborative filtering to make recommendations. They sufficient to describe each This filtering algorithm is used to find similarities between items from interactions. It uses an interaction

And every attribute of the user. This led to slightly varied accuracy as compared to other algorithms. Moreover, this algorithm didn’t keep any track of suggesting the latest watch list to the user. And focused on each feature separately rather than the feature matrix. This made them achieve high accuracy and less computational power.

# Problem Statement

**PS: 1**

To provide a balance between efficiently and accuracy to produce better recommendations. Basically, the main aim is to improve the recommendation accuracy by removing the complexity of the system.

**PS: 2**

In this recommendation system we are trying to maximize the speed of the system. The system should be able to deal with large amount of dataset.

# Study of Various Algorithm

**1. K- Nearest Neighbor:**

A supervised machine learning algorithm which is easy to implement and can be used to solve both classification as well as regression problems. This algorithms assumes that the closer the things, the more related they are to each other. But it has a major drawback of becoming slow as the size of the data increases.

Basically, it stores the dataset and learn from it only at the time of real time prediction. In order to make the recommendation fast , it really helped in improving the speed. It is quite easy to be implemented, thereby reduces the complexity. Moreover, we can add the new data set quite easily it does not even affect the accuracy of the system. It has a variety of distance criteria so it provides more flexibility to the system.

# 2. K - Means Clustering:

It is an unsupervised learning algorithm which takes inference only from the input vectors without knowing about the known outcomes. This algorithm starts by randomly selecting a first group of centroids. Then it repeats this process until the most optimal position of the centroid is found.

When it comes to large datasets, some of the algorithms are not able to maintain their accuracy, while K-means clustering helps in dealing with large dataset. In this recommender system, sometimes dataset is too large so it helps in dealing with such scenarios. It even provider somehow tight cluster in comparison with other algorithms.

Basically, maintaining the accuracy and efficiency in case of large dataset is something we need from K- means clustering.

**3. Collaborative Filtering:**

This algorithm uses the historical data of a user to

provide a similar item in the future.

When we have quite massive and diverse content then it is difficult to apply attributes manually. It follows the approach that if people have a common interests in a particular thing then they will also have common interests in other thing. It is based on the real life approach that is simply based on making connections to recommend better and this something that helped in this approach. It may be possible that this approach may be not producing so accurate result but definitely it produces some of very interesting recommendations.

# 4.Matrix Factorization:

This algorithm is used to reduce the size of the interaction matrix. It helps in discovering latent features between 2 entities. These are one of the most widely used algorithms in the field of machine learning.

It is more effective than collaborative filtering algorithm as it allows to discover more latent features between items and users. It helped in improving the accuracy of the system.

# V. Result

After studying various algorithms, we evaluated them. We converted our data to a pandas data frame and formatted it as per our requirements. We had two different datasets, one is the user data which has their user id and the other is movie data which consists of ratings made by every user for a particular item. These two datasets are linked together with the help of user id, which uniquely identifies every user.

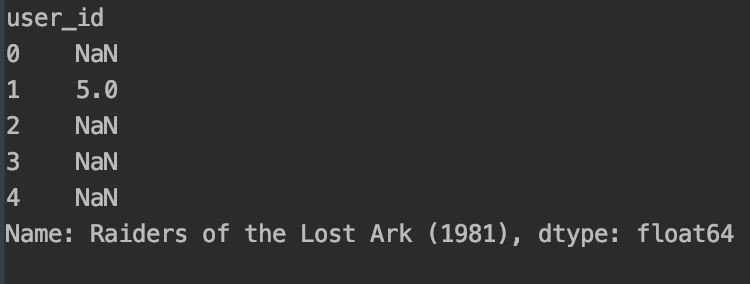


Fig [1]

Fig 1 is a matrix which consist of user id and their ratings for every movie in the dataset. If a user has not rated the movie, it will show NaN otherwise rating will be shown.

# In this matrix ,only one user has rated only the movie “Raiders of the Lost Ark(1981)”. Negatively affecting factors contributing in decreasing the optimality of approach.

We merged these datasets into one for a better data representation. After loading the data, the first step was to analyze the data. Our data had an average rating of 3.53(maximum being 5) and consisted of 100003 records. After data analysis, we created a new data frame consisting of average rating for each item(movie in our case) and number of ratings. This will help us in analyzing correlation between movies.

We used the correlation between the ratings of a movie as the similarity metric. For finding the correlation between the ratings of the movie, we needed to create a matrix where each column is a movie name and each row contains the rating assigned by a specific user to that movie. It should be noted that this matrix will have a lot of null values since every movie is not rated by every user.

We knew that each column contains all the user ratings for a particular movie. So we started to find all the user ratings for the movie "Forrest Gump (1994)" and find the movies similar to it. We chose this movie since it had the highest number of ratings and we wanted to find the correlation between movies that have a higher number of ratings.

But, correlation alone is not the best option to find similarity between items. So we decided to put a threshold on minimum number of ratings for a movie which will be considered. We used 50 as the threshold value for the minimum number of ratings.

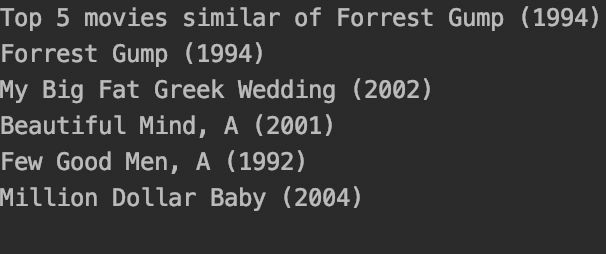


Fig [2]

In Fig[2] it can be seen from the output that the movies that are highly correlated with "Forrest Gump (1994)” are shown. The movies in the list are some of the most famous Hollywood movies, and since "Forest Gump (1994)" is also a very famous movie, there is a high chance that these movies are correlated.

# VI. Conclusion And Future Work

In this paper, we proposed the parameter boosted approach to collaborative filtering based recommender system.

Experimental approach has proved that considering the supplementary parameter, the efficiency has been improved as compared to conventional approach of collaborative filtering for recommender system. Increasing the number of parameters taken in account has compensated the problem emerging due to never ending choices of user as well as data sets in real time scenario.

We also witnessed that the parameter boosting approach

outshines other algorithm in many cases. Integrating this

idea and filling the vacuum created by earlier proposed

methods helped to achieve a remarkable results. K-means

has an edge to uniquely distinguish between different

types of groups. Here the empirical analysis has proved

that accounting the clustered data, has drastically reduce

the

In future, the work could be don’t in gathering the

Feedback and reviews from the user. We haven’t considered this in our work and probability of getting improved results are much. The feedback and the reviews after utilizing the service, given by the user can thereby be used as an explicit factor for generating improved results. The clustered groups could be allowed to calculate over root mean square distance and the external parameters will finely divide the clusters and distinguish them evenly.

**REFERENCES**

1. G. Eason, B. Noble, and I. N. Sneddon, “On certain integrals of Lipschitz-Hankel type involving products of Bessel functions,” Phil. Trans. Roy. Soc. London, vol. A247, pp. 529–551,April19**5.**
2. *(references)*
3. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon,1892,pp.68–73.
4. I. S. Jacobs and C. P. Bean, “Fine particles, thin films and exchange anisotropy,” in Magnetism, vol. III, G. T. Rado and H. Suhl, Eds. New York: Academic, 1963,pp.271–350.
5. K. Elissa, “Title of paper ifknown,”unpublished.
6. R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev.,inpress.
7. Y. Yorozu, M. Hirano, K. Oka, and Y. Tagawa, “Electron spectroscopy studies on magneto-optical media and plastic substrate interface,” IEEE Transl. J. Magn. Japan, vol. 2, pp. 740–741, August1987[Digests9thAnnualConf.MagneticsJapan,p.301,1982].
8. M. Young, The Technical Writer’s Handbook. Mill Valley, CA: University Science,1989.

1. *Galgotias College of Engineering and Technology, Greater Noida, India* [↑](#footnote-ref-1)