iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** VII **Month of publication:** July 2018

DOI: http://doi.org/10.22214/ijraset.2018.7053

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VII, July 2018- Available at www.ijraset.com

Hybrid Similarity and Clustering in Recommendation System

Twinkle[1], Dr. R.K. Chauhan[2]
[1]*M.Tech Scholar, Dept. Of Computer Science and Applications, Kurukshetra University, Haryana, India.*
[2]*Professor, Dept. Of Computer Science and Applications, Kurukshetra University, Haryana, India*

Abstract: Recommender System are subclass of Information Filtering System (IFS) that seeks to predict the rating or preference that a user would give to an item. Recommender system basically provide recommendations in one of two ways – through collaborative and content-based filtering. Collaborative filtering that bases its prediction and recommendation on the rating or behavior of other user in system. So, the main work is to find best neighbor users for target users and then recommend items to target user. Finding similarity among the users is the most difficult task because the accuracy and the quality of the recommendations depend majorly on them. This paper is bases upon collaborative filtering and find best similar user or neighbors by following ways- Hybrid Similarity. And another way is by making clusters, in this paper the improved k-means algorithm is proposed. Proposed algorithm is applied on Movie-Lens dataset , and graphical output by comparing RMSE show that these algorithm give better result for finding neighbors.
Keywords: Recommendation System, RMSE, Collaborative Filtering, Clustering.

I. INTRODUCTION

Recommender systems (RSs) are playing a significant role since 1990s as it provides relevant, personalized information to the users over the internet. It is a subclass of information filtering system that search to foretell the "rating" or "preference" that a user would give to an item. It is popular in a various areas including movies, music, news, books, research articles, search queries, social tags, and products in general.

Types Of Recommender System

A. Content Based Recommendation System

Content-based filtering methods are based on utilize the characteristics of item i.e find out similar item to that which is liked in past by user. In another way it tells about items that are similar to those that a user liked in the past (or is examining in the present).

B. Collaborative Recommendation System

[1]Collaborative filtering (CF) is a popular recommendation algorithm that bases its predictions and recommendations on the ratings or behavior of other users in the system. The fundamental assumption behind this method is that other users' opinions can be selected and aggregated in such a way as to provide a reasonable prediction of the active user's preference. Mostly in all fields collaborative filtering algorithms in service today. Collaborative Filtering done in following way:

1) Computing Prediction[2]
2) Calculating Similarity

II. PROPOSED ALGORITHM

In traditional collaborative filtering - The process of the recommendation system can be divided into three steps: data initialization; find the nearest neighbors; produce recommendation data set[2]. The key step in this algorithm is accurately calculating similarity of the users or the items to search the nearest neighbors of them. In this paper our focus is to find the algorithm which give best similar users . So to provide best recommendation to users. We achieve this by algorithm called Hybrid similarity and improved k-means clustering in collaborative filtering on dataset of movie lens. We focus on User rating matrix and Calculate the user similarity by Hybrid approach and improved-K mean clustering to and then find RMSE to show which is better. We implemented this on [3] 1M Movie Lens Dataset.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VII, July 2018- Available at www.ijraset.com

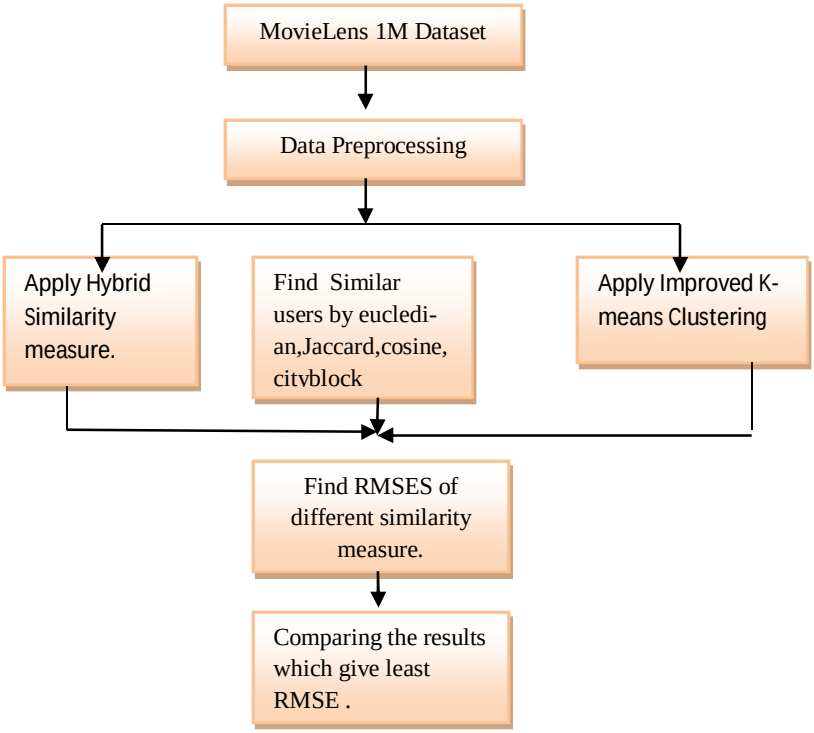


Fig 1 Flowchart of this paper

A. Step By Step Procedure of Hybrid Similarity Recommendation System:

1) From 1M movie lens dataset take only user rating data.

2) Find the nearest neighbor in whole user rating data ,by selecting several users that have highest similarity as neighbors: N(N1,N2,N3…Nk) need to meet the following condition : Sim(u,N1)>Sim(u,N2)>Sim(u,N3)…Sim(u,Nk).

Where u is user and N1,N2…Nk are neighbors.

3) Calculate the Similarity between users by [4] cosine as

For i=1 to n

$Sim(Cos\Theta) = (\sum Ai\ Bi) \div (\sqrt{\sum} Ai^2 * \sqrt{\sum} Bi^2)$

Here summation($\sum$) is also run from 1 to n

Where n = number of items in metrics, Ai & Bi are rating given by user A and user B to same item i .

Similarly, we Calculate Similarity by [5]Jaccard Similarity as-

$SIMj\ (ta,tb) = (ta \cap tb) \div (|ta|^2 + |tb|^2 - ta.tb)$

Here ta ,tb are rating given by user ta and tb .

4) Now, Fuse these above formulas into one-

$SIMh = alpha * Sim(Cos\Theta) + beta * SIMj$

Where SIMh is similarity in Hybrid approach, SIMj is Similarity in Jaccard, $Sim(Cos\Theta)$ is similarity in Cosine.

and alpha=0.6, beta (1-alpha).

5) Now Calcualte [6]Predicate Rating of user u to item m as-

P u,m=

$Uavgrating + \sum(Simh(u,nb) * (selrating - nbavgrating)) / \sum(Simh(u,nb))$

Where Uavgrating =average of ratings given by user u,

Simh =Hybrid similarity, selrating =rating given by neighbor to item m

nbavgrating=average rating of neighbors.

6) Now Calculate RMSE

For i= 1 to n

Rsum=Rsum+(actualrating- predicaterating)

Here n= number of users
RMSE=√(Rsum ÷ n)

B. Step by Step Procedure of Improved-K mean Clustering in Recommendation System

1) From 1M movie lens dataset take only user rating data.
2) Sort out all rating column in user rating matrix.
3) Divide the rows into K-cluster rows , here we calculated 'K' by (k=logn).
4) Now, Calculate the column average in different – different rows and by this we calculated centroid of cluster.
5) Calculate the distance between each data point and cluster centroid by [7] Eucledian distance as-

$$J(V) = \sum_{i=1}^{c_i} \sum_{j=1}^{c_i} (\|x_i - v_j\|)^2$$

Where $\|x_i - v_j\|$ is the Euclidean distance between x_i and v_j.
'c_i' is the number of data points in i^{th} cluster.
'c' is the number of cluster centers.

6) Allocate the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
Here summation limit from i to Ci
Vi= (1/Ci)∑ Xi
where, 'c_i' represents the number of data centers in i^{th} cluster.

7) Repeat Step 4.
8) If no data point was reassigned and same center came then stop, otherwise repeat from step 4.
9) Similarily in this predicated rating calculated
P u,m=
Uavgrating+∑(Sim(u,nb)*(selrating –nbavgrating) / ∑(Simh(u,nb)
Where Uavgrating =average of ratings given by user u,
Sim =Imp k-mean Cluster similarity, selrating =rating given by neighbor to item m
nbavgrating=average rating of neighbors.

10) Now Calculate RMSE
For i= 1 to n
Rsum=Rsum+(actualrating- predicaterating)
Here n= number of users
RMSE=√(Rsum ÷ n)

III. RESULTS

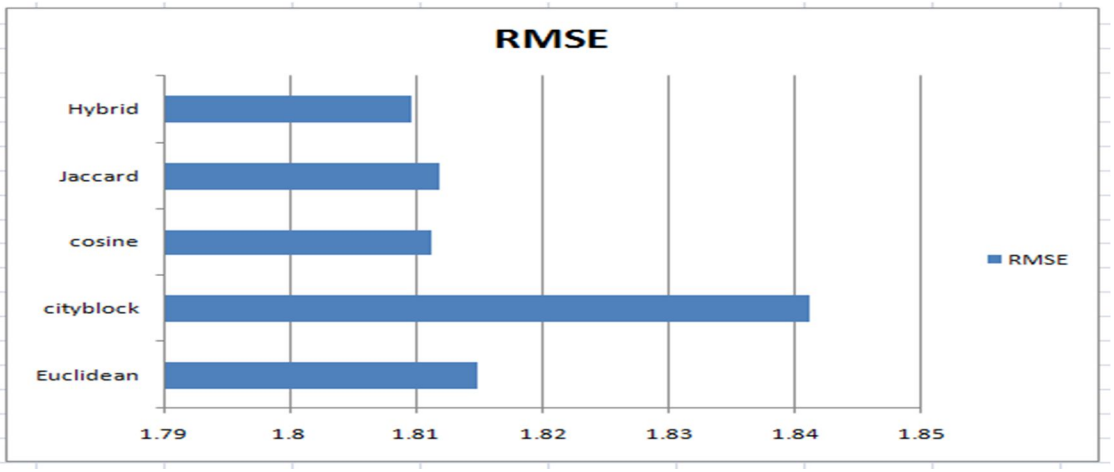


Fig 2Graphical Presentation of Comparison of Similarity measure in Movielens dataset.

In figure2 We graphically show the different similarities measure like hybrid ,jaccard , Eucledian, city block on rating data ,and we get RMSE of Hybrid is least

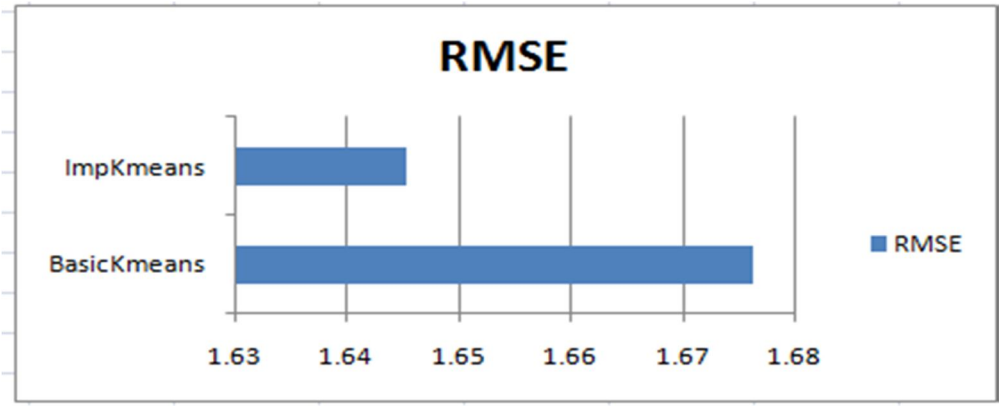


Figure 3 Compare RMSE between basic k-means and improved k-means

In this figure 3 we compare k-means and improved k means by RMSE, by graph we show this, and we see improved k-means perform better. As we see in graph value of RMSE in imp-K-means is 1.6454 which is least among all other algorithm i.e Hybrid and k-means.

Hence, improved k-means give most similar users. Also, it recommend best suited items in our case movie to users .

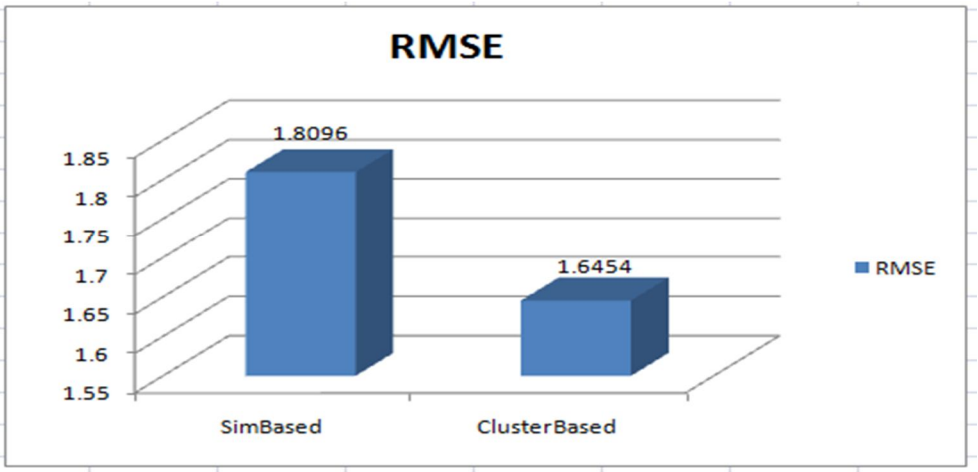


Figure 4 Compare RMSE between Similarity and Cluster based.

In this figure 4 ,this graph of overall conclusion of this paper i.e Comparing RMSE between Similarity measure and Cluster based approach. By graph we see improved k-means gives better result. Therefore, We used this for similarity among users, because of providing by this best recommended item we suggest to users.

In table 1 ,we see RMSE of different similarity measure and clustering and as you see least is imp-k means. Clustering

Table 1 Comparison of different similarity measure

S.No.	Similarity Type	RMSE
1.	Eucledian	1.8149
2.	City Block	1.8412
3.	Cosine	1.8112
4.	Jaccard	1.8119
5.	Hybrid	1.8096
6.	K-means Clustering	1.6629
7.	Improved K-means Clustering	1.6454

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VII, July 2018- Available at www.ijraset.com

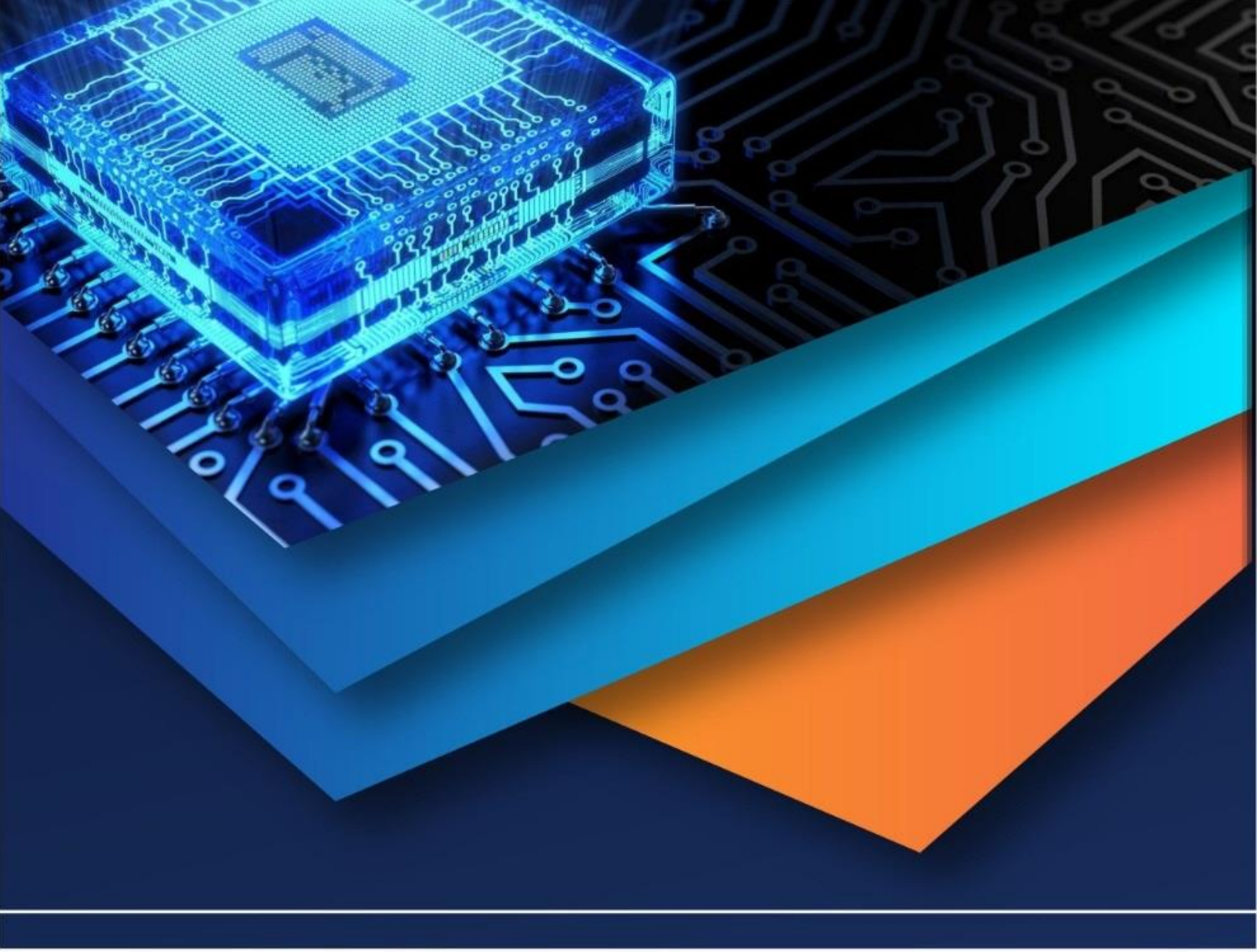
IV. CONCLUSION

As you see in this paper we find neighbor users because on the basis of similarity we recommend items to users. We proposed algorithm to find best similar users with target users. As various method have in past for similarity among users i.e [8], Jaccard ,Cosine, Eucledian etc. But when we check their result by RMSE then our algorithm Hybrid Similarity show less error i.e RMSE. In Input dataset we take 1M movie lens ,user rating file .

By Improving the K-means clustering ,we proposed other algorithm i.e improved K-means. In it we sort the rating data column wise, then divide into K-clusters row wise here k is not it is calculated as Log n, where n is size of matrix or users in dataset, then calculating average and other formulas we applied ,we see etter result i.e, reduced RMSE than Hybrid .So, finally we get an algorithm which gives best similar users or neighbors ,by this we provide best liked item to user .

REFERENCES

[1] "Recommender An Introduction System" Dietmar Jannach, Markus Zanker, Alexander Felfering, Gerhard Friedrich (2011) .
[2] Sridhar Dilip Sondur , Amit P Chigadani and Shanthran Nayak ,"Similarity Measures For Recommender Systems : A Comparative Study ", Journal for Research ,Volume 02 ,May 2016.
[3] https://grouplens.org/datasets/movielens
[4] Jun Ye," Cosine similarity measures for intuitionistic fuzzy sets and their applications", : www.elsevier.com,2011.
[5] Suphakit Niwattanakul , Jatsada Singthongchai, Ekkachai Naenudorn and Supachanun Wanapu, "Using of Jaccard Coefficient for Keywords Similarity", International MultiConference of Engineers and Computer Scientists 2013.
[6] Pooyan Adibi and Behrouz Tork Ladani ," A Collaborative Filtering Recommender System Based on User's Time Pattern Activity", 5th Conference on Information and Knowledge Technology, 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY