Effective Hybrid Recommender Approach using Improved K-means And Similarity

Prerana Khurana[#1], Shabnam Parveen[#2]
[#1]*M.tech, Computer Science Department, Kurukshetra University*
[#2]*Assistant Professor in Computer Science Department, Kurukshetra University*
JMIT Radaur, Haryana, India

Abstract- *In this age of information load, it become a herculean task for user to get the relevant information. Recommender system plays an important role in suggesting relevant information that is likely to be preferred by the user. Different type of clustering is used for recommender system like K-means, fuzzy C-mean, chameleon hierarchical etc. This papers aims at proposing a recommender system that uses hybrid approach using improved K-means clustering with Spearman's rank correlation similarity to reduce the RMSE and time complexity and results are compared with basic K-means clustering.*

Keywords- *Recommender System, Hybrid Recommender, clustering, k-means, similarity, RMSE, Spearman's rank correlation*

I. INTRODUCTION

In today's world where internet has become an important part of human life, there is exponential increase in the amount of digital information and on-line services available and we are drowning in it that has led to the problem of information overloading - where people face difficulties in finding their required information from an overwhelming set of choices. Today the problem is not about how to get correct information to make decision, rather, how to make a right decision out of the enormous amount of information. Looking for a motel to looking for good investment options, there is abundant much information available [3][12][13].

The abundance of data has increased interest in data analysis and knowledge discovery. To extract knowledge or filter useful and interesting information from overwhelming collection of information has led to the development of various information retrieval technologies. Moreover, users' increased expectation has made it very difficult to please them by searching based on the few keywords or some criteria only as it may not capture all their actual interests. To fulfil the demand of intelligent data analysis, various tools and applications have been discovered and recommender system is one of them. Recommender system is an Information Retrieval tool that helps the users discover items or products of their interest from a large collection of items [19].

Recommender System use the opinions of a group of users to help individuals in that group more effectively identify content of interest from potentially large search space.

Recommender System recommends everything from movies, news, books, songs and Web sites to more complicated suggestions for electronic gadgets, matrimonial matches, financial services, etc.

A number of such online recommendation systems implemented and used are the recommendation system for books at Amazon.com , for movies at MovieLens.org etc. [12][14][3]

Recommendation algorithms mainly follow collaborative filtering, content-based filtering, demographics-based filtering and hybrid approaches.

Collaborative filtering:-It recommends items based on the similarity measures between users and items. The system recommends those items that are preferred by similar category of users. Collaborative filtering has many advantages

1. It is content-independent
2. In CF people makes explicit ratings so real quality assessment of items is done.
3. It provides effective recommendations because it is based on user's similarity rather than item's similarity.

Content based filtering:-It is based on profile of the user's preference and the item's description. In CBF, to describe items we use keywords apart from user's profile to indicate users preferred likes or dislikes. In other words CBF algorithm recommend items or similar to those items that were liked in past. It examines previously rated items and recommends best matching item.

Demographic: It provides recommendation based on the demographic profile of the user. Recommended products can be produced for different demographic niches, by combining ratings of users in those niches.

Knowledge-based: It suggests products based on inferences about user's needs and preferences. This knowledge will sometimes contain product features that meet user needs.

Hybrid recommender: Hybrid recommender system is the one that combines multiple recommendation techniques together to produce the output [16].

If one compares hybrid recommender systems with collaborative or content-based systems, the recommendation accuracy is usually higher in hybrid systems. The reason is the lack of information about the domain dependencies in collaborative filtering, and about the people's preferences in content-based system. The combination of both leads to common knowledge increase, which contributes to better recommendations. The knowledge increase makes it especially promising to explore new ways to extend underlying collaborative filtering algorithms with content data and content-based algorithms with the user behaviour data [15].

II. RELATED WORK

A. Collaborative filtering

Sanjeev Dhawan et al. [2] build a high rating recent preferences based recommendation system in which classification is done in WEKA data mining tool and similarity index is calculated by using Pearson correlation, Cosine based similarity and Euclidean distance based similarity. **D.K. Yadav et al. [3]** introduced a movie recommendation system named MOVREC that allows a user to select the choices from a given set of attributes and then recommend a movie list based on the cumulative weight of different attributes and using K-means algorithm. **Utkarsh Gupta et al. [5]** proposed efficient technique based on Hierarchical clustering. The user or item specific information is grouped into a set of cluster using Chameleon Hierarchical clustering algorithm. Further voting system is used to predict the rating of particular item. This produces lower Mean Absolute Error. **Hirdesh Shivhare et al. [6]** Used combinatorial approach by combining fuzzy c-means clustering and genetic algorithm based weighted similarity measure and provide optimal similarity measures and similarity metrics. **Zebin Wuet al. [8]** introduced personalised recommendation algorithm by improving similarity calculation method to solve problem of similarity.

Then algorithm use user's similarity fuzzy clustering. After that matrix is used to produce recommendation to improve accuracy and real time response speed. **Hideyuki Mase et al. [9]** resolves the problem in collaborative filtering of data smoothing in entire user database by incorporating hybrid clustering.

B. Hybrid Recommender

Harpreet Kaur Virk et al. [1] proposed a hybrid recommender system based on content and collaborative filtering and also using context that provide better movie recommendation based on user feedback using simple GUI. **Manisha Chandak et al. [4]** presented an effective hybrid (Collaborative Filtering and Content-based technique along with demographic attributes) technique for book recommendation with use of Ontology for user profiling to increase the system efficiency. **Jyoti Gupta et al. [7]** proposed hybrid system that combine prediction using item based collaborative filtering and demographic based user cluster in weighted scheme and item similarity and user cluster are computed offline. This achieves lower MAE and higher coverage than the traditional collaborative filtering algorithm. **Li Chao et al. [10]** proposed hybrid approach that combines recommendation based-on content with collaborative filtering .The concept of user intimacy is also used. Model applied to Social Networking Services. **L. Martínez et al. [11]** introduced hybrid (collaborative and a knowledge-based to avoid the cold start problem) recommender system for restaurants, called REJA that introduces a new facility for its users that consists in a geographic information referred by Google Maps of the recommended restaurants.

III. PROPOSED WORK

In this paper we are proposing a recommender system that uses a hybrid approach using clustering and similarity to recommend things to users.

Clustering is used to make user clusters. An **improved K-means** clustering is used.

Similarity is used to calculate similar users. **Spearman's rank correlation** similarity is used.

A. Overall recommendation process

In our model dataset is splitted into two parts: Training and test.

Training:

i. In this data is trained to improve the knowledge of algorithm about the user to user rating.

ii. Then user rating matrix is created.

Iii. After that user similarity is calculated of this rating matrix. **Spearman's rank correlation coefficient** is used for calculating similarity of rating matrix in our model.

iv. Then user clustering is used. In our model we used **Improved k-means** clustering.

Test:

i. In this we test our records. In this actual ratings are known already.

ii. Then cluster is predicted using Euclidean distance.

iii. After that Neighbour members can found out.

iv. Then Top N neighbours are selected using similarity.

v. After that calculate A_i and A_{nb} , where A_i is average rating of item and A_{nb} is average rating of neighbours.

vi. Then prediction formula is used to calculate prediction.

Recommender Prediction Formula:

$$\mathbf{P_{ut} = A_u + [\sum^{c}_{i=1}(R_{it} - A_i) \times sim(u,i) \div \sum^{c}_{i=1} sim(u,i)]}$$

where P_{ut} is the prediction for the active user, A_u is the average user rating , R_{it} is the rating given to item i ,A_i the item average rating and sim(u,i) is the similarity function between the user and item.

vii. Then errors are calculated using **RMSE**

The Root Mean Square Error (**RMSE**) is a frequently used measure of the difference between values predicted by a model and the values actually observed from the environment that is being modelled. These individual differences are also called residuals, and RMSE serves to aggregate them into single measure of predictive power. The RMSE of a model prediction with respect to the estimated variable X_{model} is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(X_{obs,i} - X_{model,i})^2}{n}}$$

Where X_{obs} is observed values and X_{model} is modelled values at time/place *i*.

Spearman's rank correlation coefficient

In our proposed approach Spearman's rank correlation is used for calculating similarity.

Spearman's rank correlation coefficient or **Spearman's rho**, is a nonparametric measure of rank correlation. It assesses how well the relationship between two variables can be described using monotonic function. The Spearman correlation between two variables will be high when observations have a similar rank and low when observations have a dissimilar rank between the two variables [21].

It is calculated using formula:

$$\rho = 1 - \frac{6\sum d_i^2}{n(n^2-1)}$$

where d_i = difference in paired ranks and n = number of cases.

B. Improved k-means: Our approach

The main idea is to classify a given set of data into K number of disjoint clusters, where the value of K is fixed already. The algorithm consists of two distinct phases: the first phase is to calculate the K centroids, one for each cluster. Next phase is to take each point belonging to given data set and associate it to the nearest centroid. Euclidean distance is used to compute the distance between data points and the centroids. When all the points are included in some clusters, the first step is complete and early grouping is done. At this point we need to recalculate the new centroids, as the inclusion of new points may lead to change in the cluster centroids. Once we find the K new centroids, new binding is to be created between same data points and the nearest new centroid, thus generating a loop. As a result of this loop, K centroids may change their position in step by step manner. Eventually, a situation will arrive where the centroids do not move longer. This signifies the convergence criteria for clustering. Pseudo code for the improved k-means clustering algorithm is listed as Algorithm 1 [17][18].

Algorithm 1: The k-means clustering algorithm

Input:

D = {d1, d2,.....,di} //set of *i* data items.

K// Number of desired clusters

n // the total number of rows

N// size of each cluster

Output:

A set of K clusters.

Steps:

1. For predicting initial centroids data is first sorted. Then data is divided into K clusters. Size of each cluster can be calculated using N=row (n/K)
2. Take the mean of each column of a cluster to calculate the center. In this way a matrix is created having rows equal to the size of cluster
3. Repeat

Assign each item dj to the cluster which has the closest centroid;

Calculate new mean for each cluster;

Until convergence criteria is met

IV. EXPERIMENT SETUP AND RESULTS

To demonstrate the effectiveness of our proposed approach we have conducted experiment on the most popular **MovieLens** dataset. Experiment is performed on MovieLens 1M dataset using MATLAB tool.

In MovieLens Dataset rating data files have at least three columns: the user ID, the item ID, and the rating value.

MATLAB stands for **MATtrix LABoratory** allows matrix manipulations, plotting of functions and data, implementation of algorithms, creation of the user interfaces, and interfacing with

programs written in other languages, including C, C++, Java, Fortran and Python.

System details: We have validated our results on the machine with configuration of installed memory (RAM 3.00 GB) ,32-bit operating system, Intel(R) Core (TM)i3 CPU M380 @ 2.53 GHz and MATLAB 7.0 tool and Table is shown below with their corresponding readings.

Table1: Comparison of existing and proposed approach

	RMSE	Time Complexity(sec)
Recomm –Kmeans	1.7220	3.847169
Recomm- iKmeans-sim	1.6850	15.011400

In table1 recommender sytem using the existing basic k-means and recommender with Hybrid approach (Improved K-means with **Spearman's rank correlation** similarity) are compared to show RMSE and time complexity. RMSE and Time complexity are reduced in our proposed work than the existing model which will enhance the performance of the system and produce better and effective recommendations in less time.

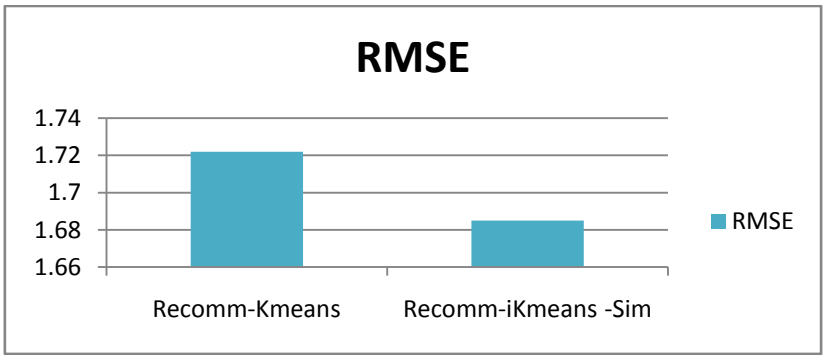

Figure 1: RMSE comparison of existing and proposed approach

Figure1 shows the RMSE graph of existing and proposed recommender system. RMSE of existing model is 1.7220 and of proposed is 1.6850 which reduces the error and give better recommendations.

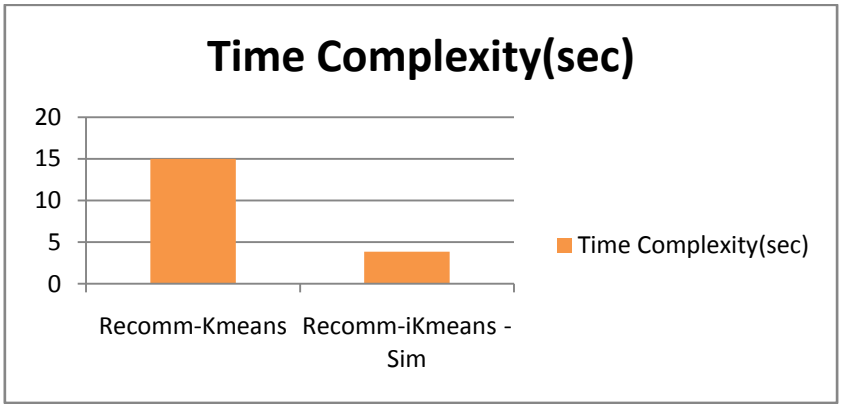

Figure 2: Time Complexity (sec) comparison of existing and proposed approach

Figure 2 shows the time complexity (sec) graph of existing and proposed model. Time complexity of existing model is 15.011400 seconds and that of proposed is 3.847169 seconds which is much less than the existing approach.

The following below are the snapshots of the recommender system using existing and proposed approach.

.

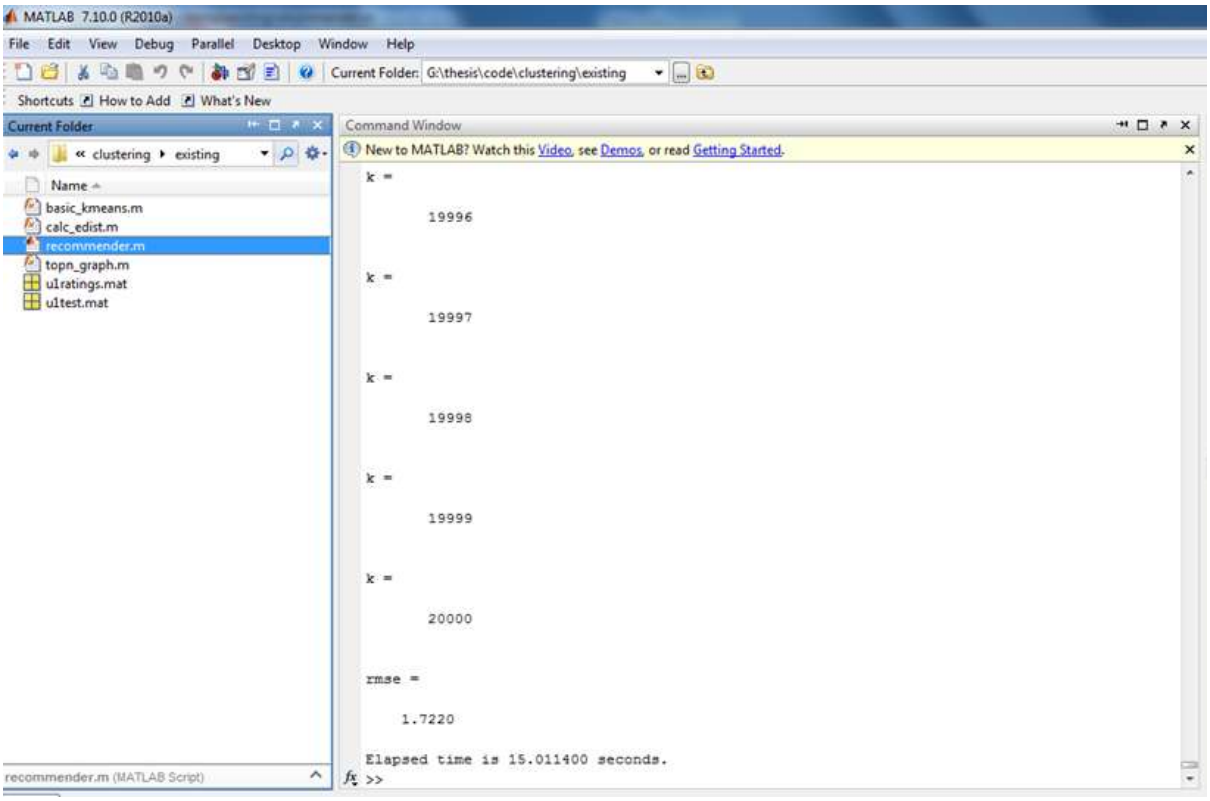

Figure 3: Snapshot of recommender system using existing approach

Figure 3 depicts the snapshot of the existing recommender system when implemented on MATLAB tool calculating the RMSE and elapsed time.

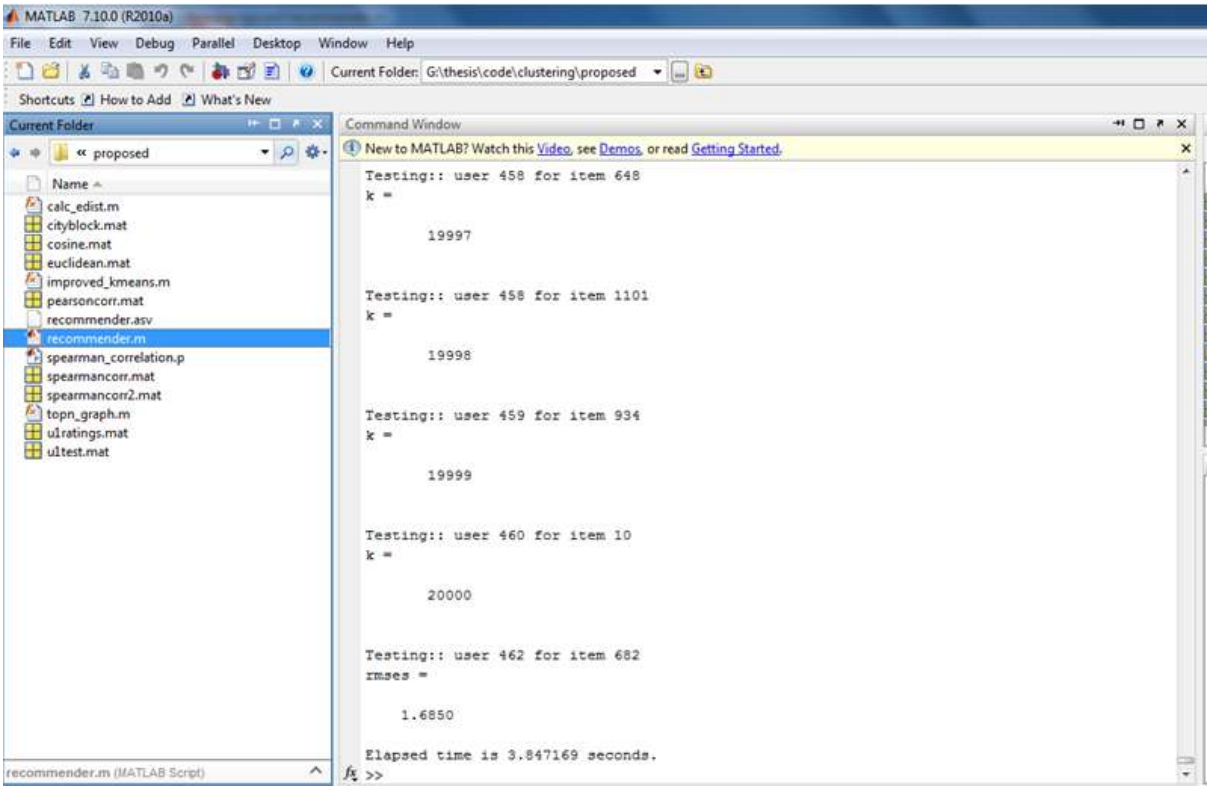

Figure 4: Snapshot of recommender system using proposed approach

Figure 4 shows the snapshot of the proposed approach when implemented on MATLAB tool showing the RMSE and elapsed time which are better than the existing approach. Hence, it will provide better recommendations.

V. CONCLUSION AND FUTURE SCOPE

In this paper we proposed a hybrid approach(improved K-means clustering with Spearman's rank correlation similarity) which reduces the RMSE and time complexity and produces better recommendation than the existing basic K-mean algorithm in which quality of final clusters highly depends on selection of initial centroids.

In future various clustering algorithms can be combined with different similarity to produce effective results.

REFERENCES

[1] Harpreet Kaur Virk, Er.Maninder Singh," Analysis and Design of Hybrid Online Movie Recommender System" International Journal of Innovations in Engineering and Technology (IJIET)Volume 5 Issue 2,April 2015.

[2] Sanjeev Dhawan, Kulvinder Singh, Jyoti," High Rating Recent Preferences Based Recommendation System" 4thInternational Conference on Eco-friendly Computing and Communication Systems,ICECCS 2015

[3]Manoj Kumar,D.K Yadav,Ankur Singh,Vijay Kr. Gupta," A Movie Recommender System: MOVREC" International Journal of Computer Applications (0975 – 8887)Volume 124 – No.3, August 2015

[4] Manisha Chandak,Sheetal Girase, Debajyoti Mukhopadhyay," Introducing Hybrid Technique for Optimization of Book Recommender System" International Conference on Advanced Computing Technologies and Applications(ICACTA-2015)

[5] Utkarsh Gupta1 and Dr Nagamma Patil2," Recommender System Based on Hierarchical Clustering Algorithm Chameleon" 2015 IEEE International Advance Computing Conference (IACC)

[6] Hirdesh Shivhare, Anshul Gupta, Shalki Sharma," Recommender system using fuzzy c means clustering and genetic algorithm based weighted similarity measure" IEEE International Conference on Computer, Communication and Control (IC4-2015)

[7] Jyoti Gupta, Jayant Gadge," Performance Analysis ofRecommendation System Based On Collaborative Filtering and Demographics"2015 International Conference on Communication, Information & Computing Technology (ICCICT), Jan. 16-17, Mumbai, India

[8] Zebin Wu,Yan Chen,Taoying Li,"Personalized Recommendation Based On The Improved Similarity and Fuzzy Clustering" The National Natural Science Foundation of China(No.71271034)2014 IEEE

[9] Hideyuki Mase, Hayato Ohwada," A Collaborative Filtering Incorporating Hybrid-Clustering Technology" 2012 International Conference on Systems and Informatics (ICSAI 2012)

[10] Li Chao, Yu Jian, Li Xiang, Chen Jia Hui," A Social Network System Oriented Hybrid Recommendation Model" 2012 2nd International Conference on Computer Science and Network Technology

[11] L. Martínez, R.M. Rodríguez, M. Espinilla," *REJA*: A GEOREFERENCED HYBRID RECOMMENDER SYSTEM FOR RESTAURANTS" 2009 IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology

[12]Joydeep Das , Shreya Dugar, Harsh Gupta, Subhashis Majumder and Prosenjit Gupta," An Adaptive Approach To Collaborative Filtering Using Attribute Autocorrelation"2015 IEEE

[13]Ying Liu, Jiajun Yang," Improving Ranking-based Recommendation by Social Information and Negative Similarity" Procedia Computer Science 55 (2015) 732 – 740

[14]Mohammed Wasid and Vibhor Kant," A Particle Swarm Approach to Collaborative Filtering based Recommender Systems through Fuzzy Features" Eleventh International Multi-Conference on Information Processing-2015 (IMCIP-2015), Procedia Computer Science 54 (2015) 440 – 448

[15] Oleksandr Krasnoshchok, Yngve Lamo," Extended content-boosted matrix factorization algorithm for recommender systems" 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014, Procedia Computer Science 35 (2014) 417 – 426

[16] Prerana Khurana, Shabnam Parveen, "Approaches of Recommender System: A Survey", International of Computer Trends and Technology (IJCTT) – Volume 34 Number 3 - April 2016

[17]]Margaret H. Dunham, Data Mining- Introductory and Advanced Concepts, Pearson Education, 2006

[18] K. A. Abdul Nazeer, M. P. Sebastian," Improving accuracy of recommendation system by means of Item-based Fuzzy Clustering Collaborative Filtering"2011 IEEE

[19] Rajani Chulyadyo, Philippe Leray ," A personalized recommender system from probabilistic relational model and users' preferences" 18th International Conference on Knowledge-Based and Intelligent Information & Engineering Systems - KES2014, Procedia Computer Science 35 (2014) 1063 – 1072

[20] Dr. Sarika Jain, Anjali Grover, Praveen Singh Thakur, Sourabh Kumar Choudhary," Trends, Problems And Solutions of Recommender System" International Conference on Computing, Communication and Automation (ICCCA2015)

[21] Saikat Bagchi," Performance and Quality Assessment of Similarity Measures in Collaborative Filtering Using Mahout" 2nd International Symposium on Big Data and Cloud Computing (ISBCC'15), Procedia Computer Science 50 (2015) 229 – 234