A Hybrid Recommendation Method For Movies Using The Bisect K-Means And The Parallel SGD Algorithm

Pushpa Mohan, Shanthi M. B
Department of CSE
CMR Institute of Technology,Bengaluru
E-mail: *pushpa.m@cmrit.ac.in*,E-mail: *shanthi.mb@cmrit.ac.in*

Abstract

A Recommender System is an approach for providing adequate information to the customer which speeds up the decision making process related to a particular task of interest. It is common, the taste of interest varies from person to person, and hence it is a challenge to design a recommender system which provides the suggestions based on individuals interest. Most of the current researches have given focus on either implementing the recommender system either by using content based or collaborative approaches in their research. As the effectiveness of the recommendation system depends on the accuracy and the speedy results, there is an ample space for the improvisation through upgrades in such systems. In this paper, we have proposed a hybrid movie recommender system for the user by combining the better features of Bisect K-Means and Parallel Stochastic Gradient Descent (SGD) algorithms. The experimental results have shown the considerable improvement over the approaches currently are in use

Key words: *Recommender, adequate, particular, customer, challenge, suggestions, bisect, collaborative, effectiveness, improvisation, K-means*

I. INTRODUCTION

In general, any recommendation system is made up of two basic elements called ‘User’ and ‘Item’. It takes a database having users and the items of interest and gives recommendations as the output. In order to generate recommendation, the system uses the user inputs or the sources of browse history in past by the user who wants to get the recommendation or the user ratings for the requested item, etc... [3]. Movie Recommender System (MRS) is a collection of software and technical tools that will provide best suited suggestions as per the user choice and displays a list of movies based on user interest. Generally a Recommendation applies one among five different filtering approaches [5]. 1. Filtering based on the content: In this filtering of the item has been done based on the content. 2. Collaborative filtering: Acquires the knowledge for filtering by observing the past behavior of the user. 3. Demographic filtering: Recommendation system uses the demographic parameters of the users like age, gender, location information, employment status etc for recommendation purpose. 4. Knowledge-based filtering: Acquiring the knowledge about the user interest based on the specific domain. User preferences are considered as the external inputs and the recommendation criteria has been supplied to the algorithm as external input for generating the recommendation. 5. Hybrid filtering: This type of approach is built by combining the effective features of the filtering approaches and bringing under a single framework of data filtration for recommendation system.

The approach presented in the paper has a focus on hybrid filtering by combining the best features of collaborative filtering and content based filtering mechanisms. The experimental results have shown that the accuracy in the recommendation based on the user interest is better than the currently existing movie recommender systems. It has also been observed that the computational time consumed by the proposed approach is less compared to the current approaches.

Goksu Tyysuzoglu and Zerrin Isik [1] have introduced a content based movie recommendation system based on collaborative filtering. They have applied a graph based approach to identify the user preferences based on their past preferences. Used demographic information to update the recommendation

ISSN: 2005-4297 IJCA

list. Combination of two different mechanisms have shown more precise recommendations concerning movies. Jeffrey Lund and Yiu-Kai Ng [2], have proposed a deep learning based approach by using auto-encoders to generate a collaborative filtering system working based on the user ratings given by the users to produce movie recommendation. They have used K-Nearest-Neighbor and Matrix factorization.. Raymond Li et al. [3] have introduced RADIAL a new high quality real world data set, a human generated conversations around the theme for providing the movie recommendations. They have used general purpose sentence representations and hierarchical encoder and decoder architectures extended with dynamically instantiated RNN models to generate auto-encoder based recommendation system. Nagamanjula R, A. Pethalakshmi [4], have proposed a movie recommendation system based on new user similarity metric and opinion mining. Main focus was in finding type of the user opinion about the movies and to generate the top-K recommendation list for users. They have extracted aspect-based specific ratings from reviews and also recommended reviews to users depending on user similarity and their rating patterns.. Bei-Bei CUI [5] has proposed a movie recommendation system using KNN algorithm and collaborative filtering algorithm. They have applied a detailed principle and architecture of JAVAEE system relational database model. Finally, the test results showed that the system has a good recommendation effect.. Lili Zhao, Zhongqi Lu et al. [6], have investigated both high level and low level visual features from

movie posters for improvising the accuracy in movie recommendation. They have used real world data sets for the experimentation and shown the significant improvement in the results found. Zehra C Ataltepe et al. [7] have proposed a hybrid approach for movie recommendation by combining content-based and collaborative recommendation methods for a Turkish movie recommendation system. They have used user behavior and different types of content features for the prediction of movie rating. Experimental results on a data set with hundreds of users and movies have shown that the users who have watched a small number of movies in the past, feature selection can increase recommendation success. OJoun Lee, Jason J. Jung [8] have introduced a new approach based on following two features. 1. Composition of movie characters and 2. Interactions among the characters. Story-based features of the movies that are extracted from character networks. They have anticipated that the proposed method could improve the reasonability of the recommend systems for movies. Sajal Halder, A.M. Jehad Sarkar et.al. [9], have used Movie Lens database and proposed a data mining tool that gathers all important information which is required in a movie recommendation system. Generated movie swarms which would help the producer to plan a new movie and also useful for movie recommendation. Experimental studies on the real data reveal the efficiency and effectiveness of the proposed system. Wei Yang et al. [10] have proposed IMRHN (interest-based movie recommendation in heterogeneous network) with user's information and users' influence on each other to implement personalized movie recommendation successfully. It also investigates the user's impact of interest in movies with others. Furthermore, the approach reduced time utilization efficiently to the scale of data set to expand.

II. PROPOSED WORK

Proposed approach uses benchmark data set of Movie Lens consisting of 1 to 256 million ratings for generating the recommendation. This data set MovieLens describes user details and movie ratings from Movies. The Fig.1 shows the flow chart illustrating different steps followed in implementation.

The overall work has been carried out in 3 different phases. In the first phase, the selected data set has been loaded to Apache Hive and required features have been extracted out. The processed data set is supplied as training data set for generating recommendations. The second phase is to apply matrix factorization used to decompose the user-item interaction into product of two lower dimension rectangular matrix. Stochastic Gradient Descent algorithm is used to factorize the matrix and reduce error. The training data generated by applying matrix factorization has been supplied to generate recommendation based on the combined approach of Bisect K-Means and the SGD algorithms.

A. Matrix Factorization

Matrix factorization algorithms work by decomposing the user-item interaction matrix into the product of two lower dimensional rectangular matrices. The idea behind matrix factorization is to represent users and items in a lower dimensional latent space. Solution to sparse data problem in matrix is used for factorizing the data sets to find optimal solution. In the sparse user matrix, the user rating has been predicted for the user u for item i is computed as given in the Equation 1.

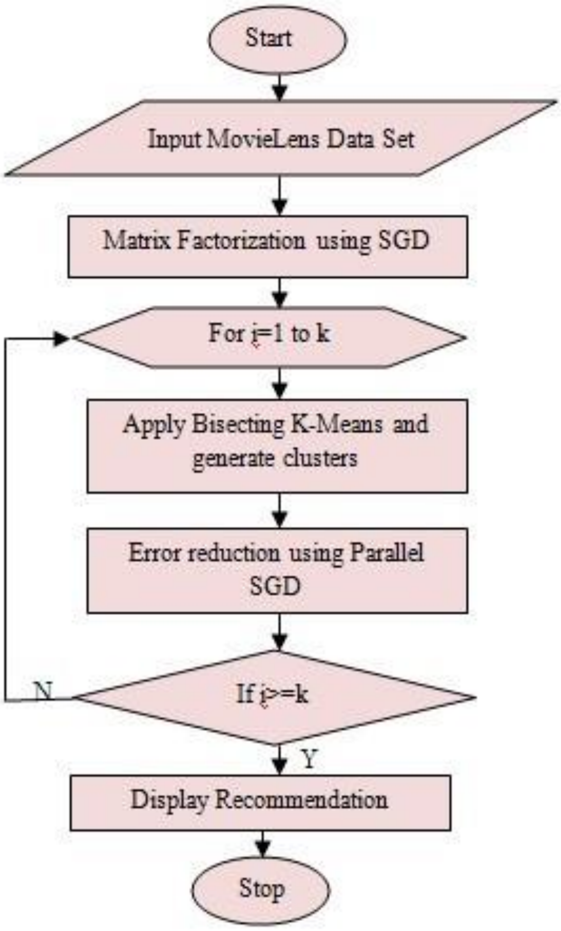

Start
Input MovieLens Data Set
Matrix Factorization using SGD
For i=1 to k
Apply Bisecting K-Means and generate clusters
Error reduction using Parallel SGD
If i>=k
N
Y
Display Recommendation
Stop

Fig. 1. Flow Chart for Movie Recommendation System

The rating by the user u for the item i in the sparse matrix R is computed as given in the Equation 1.

$$R'_{u,i} = \sum_{k=1}^{k} P_{u,k} * Q_{k,i} \qquad (1)$$

Where $P_{u,k}$denotes the latent vector for user and $Q_{k,i}$ denotes the latent vector for items. Movie rating values for the item i, mentioned by the user u, has been expressed as a cross product of the latent vectors of users and latent vectors of item. Features projected from the base matrix for expressing the cross product are considered as latent factors. Increase in latent factors improves the personalization in recommendation.

But algorithm may suffer with over fitting problem as the number of latent factors got increased for consideration. This can be resolved by adding regularization parameter. The difference in the user item rating in the base matrix and in the product matrix is considered as $E_{u,i}$. This difference error is

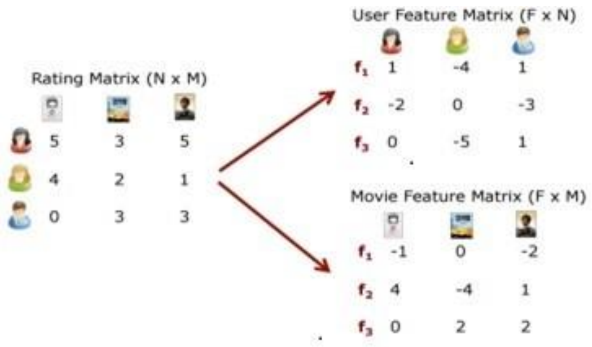


Fig. 2. Matrix factorization process (Source: RecSys by Juan Diego DiazWordPress.com)

computed as squared difference between the actual rating and the estimated rating, as estimated rating may be greater than the actual rating. The overall error in rating is computed as per following equation 2.

$$E_{u,i}^2 = \left(R_{u,i} - R'_{u,i}\right)^2 \qquad (2)$$

By adding the regularization factors to the error estimation equation, the update in the equation has been shown in the following equation 3

$$E_{u,i}^2 = \left(R_{u,i} - R'_{u,i}\right)^2 + \frac{\beta}{2}\sum_{k=1}^{k}\|P\|^2 + \|Q\|^2 \qquad (3)$$

Where, factored matrices P and Q give better approximation of R. Accuracy in the movie recommendation can be increased by adding the natural biasing factors with respect to users and items. The error estimation adding the user and the item related bias factors is computed as in the equation 4.

$$E_{u,i}^2 = \left(R_{u,i} - R'_{u,i}\right)^2 + \left(b + b_u + b_i + \frac{\beta}{2}\sum_{k=1}^{k}\|P\|^2 + \|Q\|^2\right) \quad (4)$$

Where, b_u is user bias and b_i shows the bias factor by item, b is the global bias.

B. Bisecting K-Means

Bisecting K-means is a combination of K-Means and hierarchical clustering approaches. Algorithm takes all the training data as a single cluster for processing. At the first step, it bisects the set applies K-means clustering algorithm to divide the training data set into two clusters by limiting the cluster numbers to two. In the subsequent steps, each sub cluster is going to apply the recursive steps to generate the clusters by bisecting the parent cluster into two partitions. This continues till K number of partitions has been created. Bisecting steps introduce parallelism to apply computation. Each cluster would be then processed by applying K-means to derive the best possible solution. When the trained model is used for prediction, algorithm starts computing points with child cluster nodes and best points are used for further iteration. Prediction process continues following selected clusters till leaf nodes.. The fig.3 shows the flow of the prediction process across the iterations

The centroid of each cluster is updated in every iteration based on Euclidean distance. For the lager data sets, the overall process of distance estimation with minimal error has been computed using Stochastic Gradient Descent approach. The algorithm illustrates the working of the combined approach of Bisect K-means and SGD algorithms.

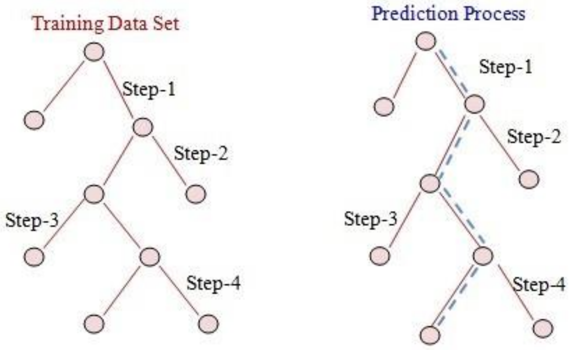


Fig. 3. Train and Prediction Process

Algorithm. 1 Parallel SGD with Bisect K-Means

```
Combined approach of Bisecting K-Means and SGD
Initialization:
Make entire data set as a single cluster
Initialize Total clusters: $K$
Initialize Clusters: $C = 0$
($C = K$) Perform bisection on the training data set
$for$ $i \leftarrow 1$ to $t$    do Bisect selected cluster using
 K-Means algorithm (2-means);
Select the cluster having lowest intra cluster distance;
Add this cluster to cluster group;
C= C + 1;
$for$  $i \in [1,2,3,\ldots,K]$  in parallel Do
    $for$ $i \leftarrow 1$ to $t$ do Update the weight function;
   Select  j $\in [1,2,3,\ldots,m]$ uniformly at random;
$w_t = w_{t-1}\lambda\Delta w C^j(w_{t-1})$
 Compute $W = \frac{1}{k}\sum_{k=1}^{k} w_t$
Return $W$
```

III. IMPLEMENTATION

Implementation using Pyspark with GPU and RDD-based API used from Pyspark.mllib. We have considered 3 attributes user-id, movie-id and m-rating for generating recommendations. The data-set has been processed in two stages. In the first stage, the raw data-set has been supplied for Matrix factorization. Stochastic Gradient Descent (SGD) method reduces the factorization error. Base matrix is factorized into two lower dimension rectangular matrices. The second stage uses Bisect K-Means and SGD for reducing the prediction error followed with validation by generating the recommendation list.The perceptions have been taken by working framework on Ubuntu platform, Apache Hadoop 2.3.4, python 3.6, Apache Spark 1.6.0 and JDK 1.8 has been used for distributed and parallel computations.

IV. RESULT AND DISCUSSION

The Fig. 4 and Fig. 5 shows the observed results when SGD has been applied to reduce the possible RMSE value. The error rate during matrix factorization has been recorded by changing the latent factors from 6 to 20 and keeping the regularization factors in the range [0.001, 0.2]. The training data set size considered to observe the results is 1MB and 256MB.

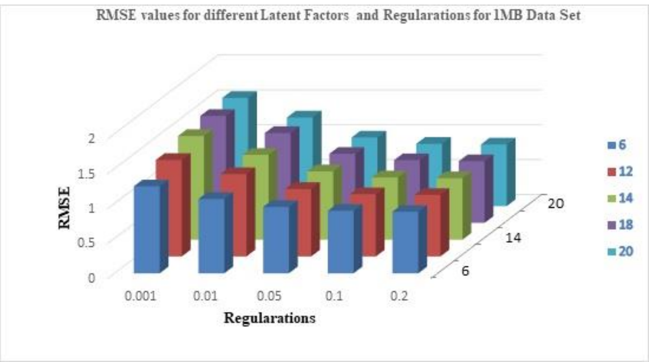


Fig. 4. Matrix factorization using SGD for 1M data set

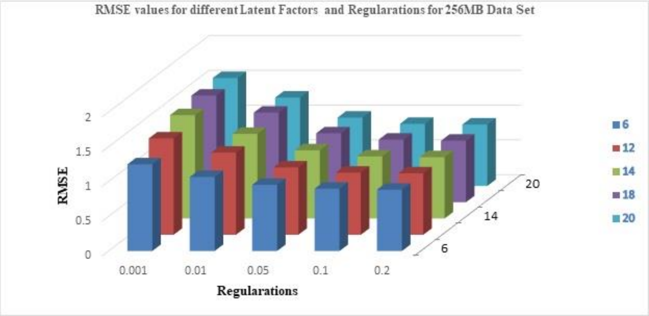


Fig. 5. Matrix factorization using SGD for 256M data set

Movie recommendation has been generated by using the combined approach of Bisect K-Means and SGD algorithms. Bisecting K-means bisects the training data set in every iteration and the prediction process proceeds with the partition having smaller value for the error. Minimization of the prediction error has been carried out using SGD algorithm. The derived results have been compared with existing approaches uses K-means and Bisect K- means method. The comparison and analysis of computational cost are shown in Fig.6.

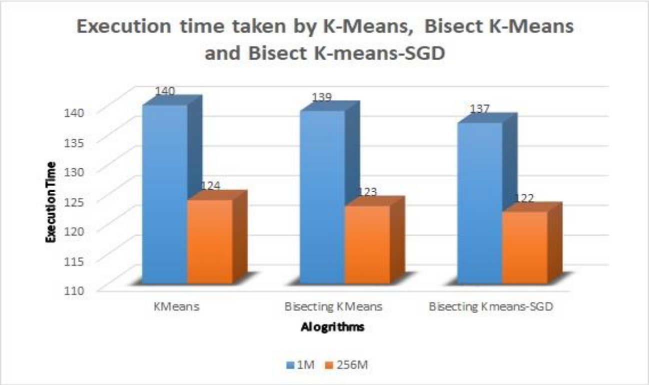
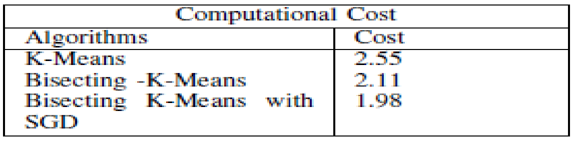


Fig.6. The Execution time by different Recommendation approaches

Table. I
Time Taken to Generate Recommendation by Different



Computational Cost	
Algorithms	Cost
K-Means	2.55
Bisecting -K-Means	2.11
Bisecting K-Means with SGD	1.98

Approaches

ISSN: 2005-4297 IJCA

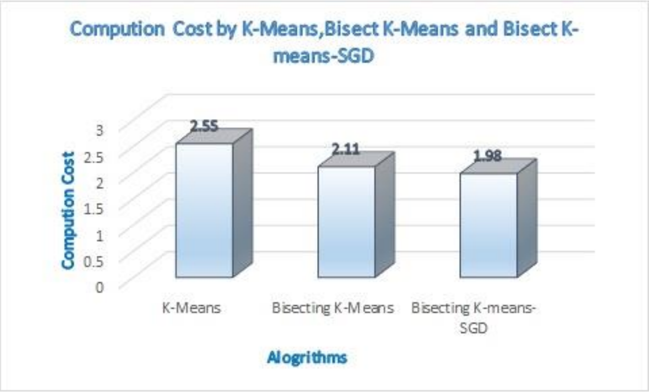
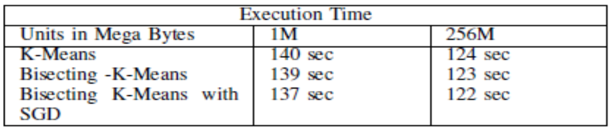


Fig. 6. Comparison and Analysis of Computational Cost

Cost function values for the different approaches have been recorded by generating recommendation on 1MB data set is shown in the Table I. The observed results have shown the following values. The minimum value for the cost function shown by K-Means algorithm is 2.55. Bisect K-Means algorithm reduces the cost value to 2.11 and the combined approach of Bisect K-Means with parallel SGD, cost value gets reduced to 1.98. Thus, the combined approach of Bisect K-Means and Parallel SGD has shown a better accuracy than the other two currently used approaches. The execution time taken by different approaches is shown in the Table II. Execution time for K-Means, Bisect K-Means and for the combined approach of Bisect K-Means and SGD has been recorded for generating recommendation for 1MB data-set and 256MB data-sets.

From the Fig.7 it is very clear that the time taken by the combined approach of Bisect K-Means and SGD is comparatively less than other approaches. When only K-means is applied for generating the recommendation, all the iterations have to go through all the generated clusters for the detection of best predicted value. Hence consume more time than other approaches. Bisect K-means applies clustered approach by limiting the processing of clusters to 2^l, where l specifies the number of levels to continue with iterations. The number of evaluations per iteration is half of the computations in K-Means because of bisection applied at each level. When Bisect K-Means is combined with SGD, the SGD works in parallel on each bisected cluster and speeds up the computation and reduces the overall prediction error. Thus consumes less computational time and the increased accuracy in generating movie recommendation.

TABLE II
Time Taken To Generate Recommendation By Different



Execution Time		
Units in Mega Bytes	1M	256M
K-Means	140 sec	124 sec
Bisecting -K-Means	139 sec	123 sec
Bisecting K-Means with SGD	137 sec	122 sec

Approaches

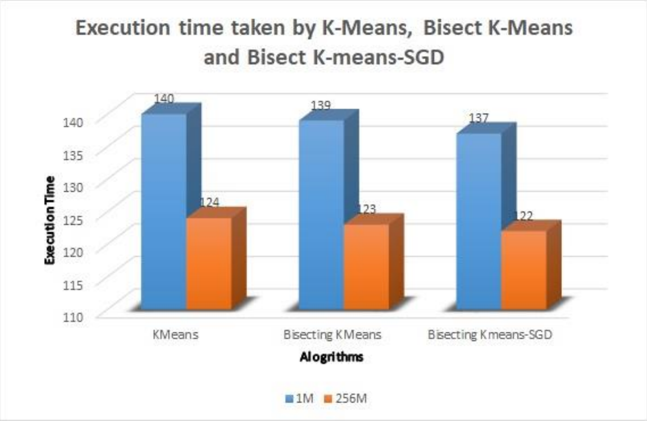


Fig. 7. The Execution time by different Recommendation approaches

V. CONCLUSION

The paper has a main focus to propose a Movie Recommendation System with increased accuracy and lesser execution time than the approaches used currently for the same purpose. Training data-set is generated by using matrix factorization using Stochastic Gradient Descent (SGD) approach. SGD reduces the factorization error to the minimal value. Generated training set data is supplied to generate Movie Recommendation. We have used a combined approach of Bisect K-Means and Parallel SGD to reduce the overall execution time and to increase the accuracy of the recommended value by reducing the value of the cost function. The experimental results have shown that the current approach has a better accuracy and lesser time consumption than the current approaches to generate Movie Recommendation.

REFERENCES

1. GoksuTyysuzoglu and ZerrinIsik. A Hybrid Movie Recommendation Computing Academic Research (IJCAR)) Volume 7, Number 2 , pages 29-37.ISSN 2305-9184 , April 2018.
2. Jeffrey Lund and Yiu-Kai Ng. ' Movie Recommendations Using the Deep Learning Approach, In IEEE International Conference on Information Reuse and Integration(IRI),July-2018, DOI: 10.1109/IRI.2018.00015
3. Raymond Li,Samira Ebrahimi Kahouet.al 'Towards Deep Conversational Recommendations', In 32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montral, Canada.
4. Nagamanjula R, A.Pethalakshmi, A Novel Scheme for Movie Recommendation System using User Similarity and Opinion Mining, In International Journal of Innovative Technology and Exploring Engineering (IJITEE) ,ISSN: 2278-3075, Volume-8 Issue-4S2 March, 2019
5. Bei-Bei CUI, Design and Implementation of Movie Recommendation System Based on Knn Collaborative Filtering Algorithm, In ITM Web of Conferences,12, 04008 (2017),DOI: 10.1051/ 7120 ITA 2017 ITM Web of Conferences itmconf/201 4008,ITA 2017
6. Lili Zhao, Zhongqi Lu et.al, Matrix Factorization for Movie Recommendation, In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16)
7. Zehra C ATALTEPE, Mahiye ULUYAGMUR et. al, Feature selection for movie recommendation, In Turk J ElecEng Comp Sci (2016) 24: 833 848, TUB ITAK doi:10.3906/elk-1303-189
8. O-Joun Lee, Jason J. Jung, Mahiye ULUYAGMUR et. al, Explainable Movie Recommendation Systems by using Story-based Similarity, In ExSS 18, March 11, Tokyo, Japan. CEUR-WS.org/vol 20168/exss5.pdf
9. SajalHalder, A.M. Jehad Sarkar et.al , Movie Recommendation System Based on Movie Swarm, In CGC2012:804-809.
10. Wei Yang et al , User's Interests-Based Movie Recommendation in Heterogeneous Network In International Conference on Identification, Information and Knowledge in Internet of Things (IIKI)-2015. Beijing, China DOI: 10.1109/IIKI.2015.23

[11] Will Serrano, Intelligent Recommender System for Big Data Applications Based on the Random Neural Network, Big Data Cogn. Computing- 2019, 3, 15; DOI:10.3390/bdcc3010015

[12] Rahul Katarya , Om Prakash Verma , An effective collaborative movie recommender system with cuckoo Search, Egyptian Informatics Journal, DOI:10.1016/j.eij.2016.10.002

[13] HumaSamin ,TayyabaAzim Knowledge Based Recommender System for Academia Using Machine Learning', A Case Study on Higher Education Landscape of Pakistan, IEEE. Translations and content mining DOI:10.1109/ACCESS.2019.2912012

[14] Debani Prasad Mishra, Subhodeep Mukherjee, SubhenduMahapatra, Antara Mehta ,Analysis of Movie Recommender System using Collaborative Filtering, International Journal of research trend in engineering and research, DOI: 10.23883/IJRTER.2017.3232.ZPBXJ

[15] Donghui Wang, Yanchun Liang Dong Xu, Xiaoyue Feng, RenchuGuan ,A content-based recommender system for computer science publications,Elsevier-knowledge based system,DOI: 10.1016/j.knosys.2018.05.001

[16] Ching-Seh, Mike-Wu, Deepti Garg, Unnathi Bandary,'Movie Recommendation System Using Collaborative Filtering', Conference: 2018 IEEE 9th International Conference on Software Engineering and Service Science (ICSESS), November 2018,DOI: 10.1109/ICSESS.2018.8663822

[17] Hung-Wei Chen,Yi-Leh Wu, Maw-Kae Hor, Cheng-Yuan Tang,'Fully content-based movie recommender system with feature extraction using neural network', Conference: 2017 International Conference on Machine
Learning and Cybernetics (ICMLC),DOI: 10.1109/ICMLC.2017.8108968

[18] Arpita Jain ,Santosh K,'Collaborative Filtering for Movie Recommendation using RapidMiner', International Journal of Computer Applications 169(6):29-33 July 2017, DOI: 10.5120/ijca2017914771

[19] Manoj Kumar, D,K Yadav, Ankur Singh, Vijay K R,'A Movie Recommender System: MOVREC', International Journal of Computer Applications 124(3):7-11 August 2015, DOI: 10.5120/ijca2015904111

[20] Badrul Sarwar,George Karypis, Joseph Constan, John Riedl, ,' Itembased Collaborative Filtering Recommendation Algorithms', August 2001,Proceedings of ACM World Wide Web Conference, DOI - 10.1145/371920.372071

[21] Anshu Sang, Santhosh K, ' Item-based Collaborative Filtering Recommendation Algorithms,', August 2001,Proceedings of ACM World Wide Web Conference, DOI - 10.1145/371920.372071

[22] Noratiqah Mohd ariff, Mohd Aftar Abu Bakar, Nurul Farhanah Rahim, ' Comparison between content-based and collaborative filtering recommendation system for movie suggestions', AIP Conference Proceedings 2013(1):020057 October 2018

[23] G. Krishna Kishore, D. Suresh Babu, 'Recommender System based on Customer Behaviour for Retail Stores', IOSR Journal of Computer Engineering, May 2017

[24] Rabi Narayan Behera, Anindita Chakraborty,Progga Laboni Saha, Sujata Dash, 'Hybrid Movie Recommendation System based on PSO based Clustering Hybrid Movie Recommendation System based on PSO based Clustering', International Journal of Control Theory and Applications 10(18) January 2017

[25] Abhishek Mahata, Nandini Saini, Sneha Saharawat, Ritu Tiwari, 'Intelligent Movie Recommender System Using Machine Learning', Intelligent Human Computer Interaction: 8th International Conference, ICI 2016, Pilani, India, December 12-13, 2016, Proceedings (pp.94-110), January
2017