ISSN No. 0976-5697

Volume 8, No. 5, May – June 2017

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Dynamic Recommendation System Using Enhanced K-means Clustering Algorithm for E-commerce

Ankush Saklecha
Institute of Engineering & Technology
Devi Ahilya University, Indore, India

Jagdish Raikwal
Institute of Engineering & Technology
Devi Ahilya University, Indore, India

Abstract: E-commerce organizations are growing day by day over time in terms of both business and data. Maximum organizations rely on these e-commerce websites to attract new customers and maintain existing ones. Dynamic recommendation system can be used to achieve this goal. It works towards improving the result factors of product priority displayed over the users search records. . This paper focuses on providing real-time dynamic recommendations to all registered users of the website. Here the dynamic recommendation technology is proposed, which uses enhanced K-Means Algorithm to generate item recommendation. By compiling the real time e-commerce data and comparing the system with existing K-means algorithm, the effectiveness of the proposed system is evaluated. The results prove that the proposed system provides good quality, accuracy and reduces the limitations of the conventional recommendation system. The experimental evaluation is measured on precision, recall and accuracy for proving the robustness of the system.

Keywords: Data Mining; Clustering; K-means Clustering; Recommendation; E-commerce; knowledge;

I. INTRODUCTION

With the increase in the use of e-commerce sites, maximum companies are selling their product via e-commerce sites and customers purchase via e-commerce sites increase day by day. Customers become overloaded with multiple choices. It has become very challenging for the users to find the items of their interest without wasting a lot of time. When a user tries to find an item using search engines, for example, Google Search engine and Yahoo Search engine, the user needs to type the exact name of the item. The data in the internet is huge which makes it very difficult for the user to find the items of his interest. Hence, there is a need for a system which learns the likes and dislikes of the user and generates recommendations based on his interest[10]. There are so Many algorithms used when we are designing a recommender system. Recommender systems employ Information Filtering technique that focuses on providing the recommendations of the items to the users that are likely to be of the user's interest. We can define recommender system : "if I is the set of all possible items that can be recommended and U is the set of users, then there exists a function F : (U*I)→R where R is a totally ordered set of positive integers or real numbers within a definite range" The goal of this paper is to study recommendation engines and recognize the problems of traditional recommendation engines and algorithms used in developing a recommendation engine and to develop a web based recommendation engine by making use of "***Improved K-Means Algorithm based on Collaborative filtering (CF) approach***"[2]. The system works on the Improved K-Means Clustering approach of identifying the users belonging to different age groups and classifying them into the appropriate cluster based on K means algorithm, to predict recommendations of unseen items to the users belonging to similar cluster. The system makes use of cluster of ages (of users) to evaluate the similarity between users. Two users are said to be similar if they fall in the same cluster. The results show that the system successfully shows the recommendation to the users of similar clusters[4].

II. LITRETURE REVIEW

Wang Shunye et al[3]Motivated by the problem of random selection of initial centric and similarity measures ,the researcher presented a new K-means clustering algorithm which is based on dissimilarity. This improved k-means clustering algorithm basically consists of 3 steps. The first step discussed is the construction of the dissimilarity matrix i.e dm.Secondly, Huffman tree based on the Huffman algorithm is created according to dissimilarity matrix. The Huffman tree gives the initial centroid as a output. Lastly the k-means algorithm is applies to initial centroids to get k cluster as output.Iris,Wine and Balance Scale datasets are selected from UIC machine learning repository to test the proposed algorithm. As compared to usual k-means the proposed algorithm gives better accuracy and results.

Pallavi Purohit and Ritesh Joshi et al[7] proposed an better approach for unique K-means clustering algorithm due to its certain limitations. The K-means algorithm gives poor performance because of the choice of initial centroids randomly. This algorithm improves the performance and cluster quality of original k-means algorithm.

Juntao Wang et al [9] presented a better k-means clustering algorithm to solve the problem of outlier detection of present k-means algorithm. To deal with thus problem this algorithm uses noise data filter[9]. Density based outlier detection method is applied on the data to be clustered so as to remove the outliers. The motive of this method is that the outliers may not be engaged in computation of initial cluster centers. In the next step Aristidis Likas proposed a fast global k-means algorithm which is applied to the output generated previously. Using Iris, Wine, and Abolone datasets the results between k-means anad improved k means are compared. The Factors clustering accuracy and clustering time are used to test. The disadvantage of the improved k -means is that while dealing

with large data sets, it will cost more time Md.Nasim Akhtar, Md. Sohrab Mahmud and Md. Mostafizer Rahman, [6] gave an algorithm to compute improved initial centroids which is based on heuristic method. The newly offered algorithm results in highly accurate clusters with decrease in computational time. In this algorithm author firstly compute the average score of each data points that consists of multiple attributes and weight factor. Merge sort is used to sort the output that was previously generated and then data points are separated into k number of clusters. Finally the closed possible data point of the mean is taken as initial centroid. With the help of experiment performed, it shown that the algorithm reduces the number of iterations to allocate data into a cluster. But the algorithm still deals with the problem of assigning number of desired cluster as input.

Shuhua Ren and Alin Fan [10] detailed k- means clustering algorithm on the basis of coefficient of variation which is defined as ratio of standard deviation to the mean value. Existing k -means algorithm uses Euclidean distance as the similarity metric which gives inaccurate results due to the effect of useless data. To overcome with this problem, proposed algorithm uses coefficient of weight factor to elicit the effect of ouliers. Weight values are assigned to all the features in clustering to remove irrelevant, noisy data so as to raise cluster quality. The results are evaluated using popular data sets i.e. Iris,Wine and Balance scale. The results prove that the modified algorithm presents more clustering accuracy and the number of iterations required for clustering is less than original k-means. The problem faced by proposed algorithm is that the number of clusters required as output is needed to be initially defined.

SongJie Gong and Zhejiang[11] proposes that 'personalized recommendation systems' are generally utilized in e-commerce websites to gives recommendations to its users. According to that letter, the recommendation system uses collaborative filtering techniques which have been successful in providing recommendations. Techniques to solve common problems in recommended systems, i.e. excess and scalability are suggested in this paper.

Robert M Bell and Yehuda Koren[13] state that recommender systems provide recommendations to the users based on past user-item relationship. Neighbors are calculated based on the previous user-item relationship, which makes prediction easier. The weight of all the neighbors is calculated differently and for many interactions to provide a solution to the problem, they are interconnected between each other simultaneously. The proposed method has been asked to provide recommendation in 0.2 milliseconds. Training takes a lot of time in large-scale applications. The proposed method was tested on Netflix data, which contained 2.8 million queries that were processed in 10 minutes.

Micheal Pazzani[14] discusses about recommending data sources for news articles or web sites after learning the taste of the user by learning his profile. Various types of information have been mentioned in this paper which can be considered to learn the user's profile. Depending on the ratings given by a user for different sites, the ratings that other users have given to those sites and suggest demographic information about users. This paper explains how the above information can be added to provide recommendations for users.

III. PROBLEM STATMENT

The K-Means cluster analysis algorithm that can be implemented in any field. In addition, it can provide descriptive information about the sub-groups of the population who share the same pattern of feedback. However, K- Means has some disadvantages in general[7].

First of all, we need to specify the number of clusters but we do not know the true number of clusters and to find the right number of clusters which represent the true number of population clusters, it is quite subjective. On top of that, your results can change based on the location of observation, which are randomly chosen as primary centroids. K-means cluster analysis is not recommended if you have too many explicit variables. If you have a lot of clear variables, then you have to use a different clustering algorithm which can handle them better. K-means clustering that it believes that the underlying clusters in the population are spherical, different, and are of approximately equal size. Consequently, there is a tendency to identify groups with these characteristics, it will not work even if the long numbers in size are the same or not.

A. Limitations:

Given at an integer K, K-means division the data set into K non-overlapping clusters. It replaces K "centroids" or "prototype" in densely populated areas of data space. Every observation is assigned to the nearest centroid ("Minimal distance rule"). All comments in a cluster are all that are closer to any other Centroids (eg lower picture) compared to a given centroid.

Limitation 1: Handling Empty Clusters: One of the problems with the previously-given basic K-means algorithm is that empty clusters can be obtained if the cluster is not allocated any points during the assignment phase[3]. If this happens, then selecting a replacement centroid requires a strategy, otherwise, the square error will be larger than necessary.

Limitation2: Outliers: When outliers are present, the resulting cluster centroids (prototypes) may not be representative of them as well as the SSE will be high as well.

Limitation 3: Difficult to measure the no of clusters: The user has to select the value of K, the number of clusters. Although for 2D data this choice is easily visual inspection by it, it is not so high dimension data for so much, and the number of clusters May be appropriate.

In Conclusion**:** [1] Difficult to predict K-Value. **[**2]With global cluster, it didn't work well. **[**3]Different initial partitions can result in different final clusters. [4] It does not work well with different sizes and clusters of different density (in basic data).

IV. PROPOSED WORK

The architecture proposed describes user interacting with website and recommendations provided to the user while browsing the pages through the website.

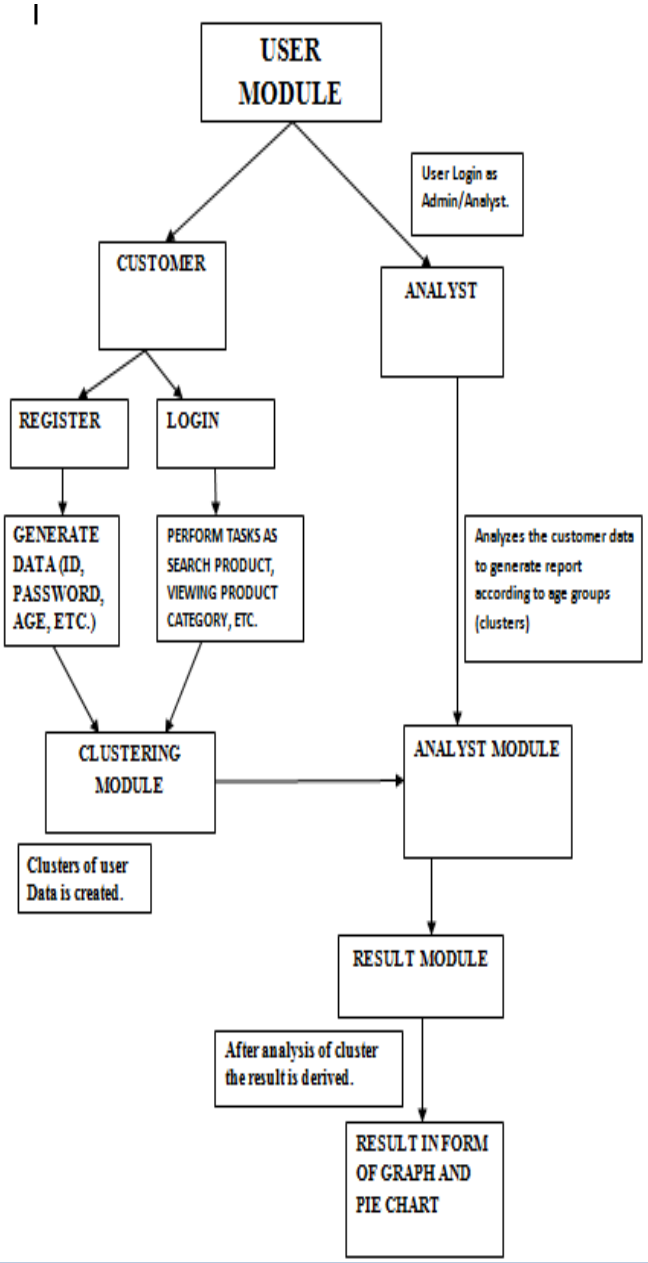

Figure 1: Proposed Recommendation System

The following Proposed Improved K – means algorithm is used your Cluster creation in proposed recommendation system.

A. Flow Chart for Enhanced K means Algorithm

The following flow chart represents execution steps of enhanced k means clustering algorithm. This paper shows that using this enhanced algorithm the accuracy of system will improve.

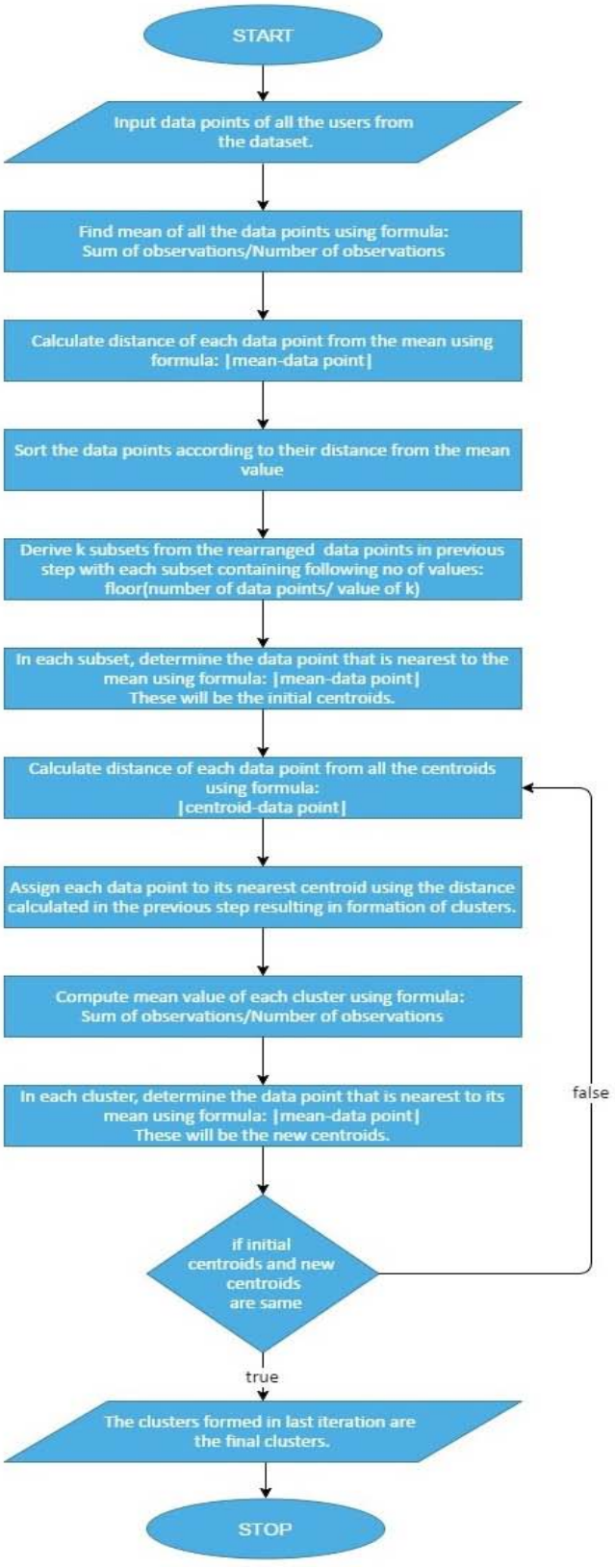

Figure 2: Enhanced K means Algorithm Flow Chart

V. RESULT ANALISIS

The Proposed System is Implemented Using XAMPP server, myPhpAdmin and Eclipse platform. Improved K-means Clustering algorithm is the backbone behind the Proposed

System. System tries to group similar objects in one cluster and the dissimilar objects far from each other. For validating our Proposed Algorithm web portal was developed that offer the most popular books for the registered customer. The database used contains real time data given by the customer at the time of registration. It contains record of more than 2000 books. Figure3. Illustrate the GUI of BestSeller portal offering books of various categories. Figure4. Illustrate the Tabular form of Cluster created by the Proposed algorithm and shows recommend books according to that.

Figure 3. GUI of Proposed System

Category	Cluster1: (3 to 9 years)	Cluster2: (10 to 15 years)	Cluster3: (16 to 20 years)	Cluster4: (21 to 25 years)	Cluster5: (26 to 30 years)	Cluster6: (31 to 35 years)	Cluster7: (36 to 40 years)	Cluster8: (41 to 45 years)	Cluster9: (46 to 49 years)	Cluster10: (50 to 53 years)
Books for Children	49.5 %	30.8 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Literature and Fiction	0.0 %	0.0 %	4.8 %	25.0 %	13.3 %	9.1 %	0.0 %	0.0 %	0.0 %	18.5 %
Biography	3.0 %	0.0 %	0.0 %	12.5 %	6.7 %	0.0 %	0.0 %	0.0 %	0.0 %	22.2 %
Crime and Thriller	6.1 %	0.0 %	0.0 %	0.0 %	6.7 %	9.1 %	0.0 %	0.0 %	0.0 %	0.0 %
Spiritual	10.1 %	0.0 %	0.0 %	0.0 %	6.7 %	0.0 %	0.0 %	0.0 %	0.0 %	3.7 %
Education	2.0 %	15.4 %	4.8 %	25.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Travel	7.1 %	0.0 %	0.0 %	12.5 %	0.0 %	9.1 %	0.0 %	0.0 %	0.0 %	7.4 %
Cookbooks	0.0 %	0.0 %	0.0 %	0.0 %	6.7 %	9.1 %	6.2 %	26.3 %	0.0 %	0.0 %
Teens and Young Adults	7.1 %	46.2 %	76.2 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Health and Fitness	0.0 %	0.0 %	4.8 %	0.0 %	6.7 %	0.0 %	0.0 %	36.8 %	15.0 %	22.2 %
Romance	13.1 %	0.0 %	0.0 %	12.5 %	6.7 %	0.0 %	43.8 %	36.8 %	20.0 %	0.0 %
Inspirational	1.0 %	0.0 %	0.0 %	0.0 %	13.3 %	9.1 %	0.0 %	0.0 %	15.0 %	3.7 %
Computer	0.0 %	7.7 %	4.8 %	12.5 %	6.7 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
History	0.0 %	0.0 %	0.0 %	0.0 %	6.7 %	9.1 %	0.0 %	0.0 %	0.0 %	22.2 %
Sci-fi and Fantasy	0.0 %	0.0 %	0.0 %	0.0 %	6.7 %	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %
Business	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	18.2 %	50.0 %	0.0 %	15.0 %	0.0 %
Psychology	0.0 %	0.0 %	4.8 %	0.0 %	0.0 %	9.1 %	0.0 %	0.0 %	0.0 %	0.0 %
Language and Culture	0.0 %	0.0 %	0.0 %	0.0 %	0.0 %	9.1 %	0.0 %	0.0 %	20.0 %	0.0 %
Classics	1.0 %	0.0 %	0.0 %	0.0 %	6.7 %	9.1 %	0.0 %	0.0 %	0.0 %	0.0 %

Figure 4: Clusters Creation of Enhanced K means Algorithm for BestSeller Portal

A. Quality Evaluation Parameter :

Accuracy is the parameter used to evaluate the effectiveness of proposed system with respect to all three techniques. A matrix is constructed to show accuracy:

Table 1. Evaluation Parameter

Items	***Recommended item***		
	Item Recommended by the system	***Item not Recommended by the system***	
Expected item	True Positive	False Negative	
Not an expected item	False Positive	True Negative	

Based on the recommendation matrix we calculate recall, precision and accuracy as shown below:

- Recall can be defined as a fraction of all relevant items that are recommended by the system.

Recall =True Positive (TP)/(False Negative (FN)+True Positive (TP))

- Precision is a factor of all the recommended products that are relevant.

Precision =True Positive (TP)/(False Positive (FP)+True Positive(TP))

- The accuracy is the ratio of true positives to the sum all recommended products.

Accuracy = (TP+ TN)/(FP+ FN+TP+ TN)

According to the expected outcomes the value of recall , precision and accuracy should be better than the old k means algorithm. Also the number of iterations should be specifically less than the old k means algorithm which would make the new k means more efficient in terms of space and time complexity.

Accuracy of the System was measured by collecting feedback from the customer based on evaluation metrics. Figure5. Illustrate the values of True positive(TP),True Negative(TN), False Positive(FP), False Negative(FN) obtained after the

	user_name	TP	TN	FP	FN
☐	john	1	0	0	0
☐	mary	1	0	0	0
☐	william	0	1	0	0
☐	helen	1	0	0	0
☐	george	1	0	0	0
☐	Anna	0	1	0	0
☐	george	0	0	1	0
☐	anna	1	0	0	0
☐	ruth	1	0	0	0
☐	Charles	1	0	0	0
☐	Robert	0	0	0	1
☐	Elizabeth	1	0	0	0
☐	Joseph	1	0	0	0
☐	Dorothy	1	0	0	0
☐	Marie	0	1	0	0
☐	frank	1	0	0	0
☐	Thomas	1	0	0	0
☐	Henry	1	0	0	0
☐	Alice	1	0	1	0
☐	Willie	1	0	0	0
☐	Walter	0	1	0	0
☐	Harry	1	0	0	0
☐	Fred	0	1	0	0
☐	Annie	0	1	0	0
☐	Rose	1	0	0	0
☐	Albert	1	0	0	0
☐	Carl	0	1	0	0

customer feedback. Using this Values the value of Average Recall, Average Precision and Average Accuracy is calculated.

	user_name	TP	TN	FP	FN
☐	Edna	0	1	0	0
☐	Albert	0	0	1	0
☐	Mildred	1	0	0	0
☐	Frances	1	1	0	0
☐	Clarence	0	1	0	0
☐	Rose	0	1	0	0
☐	Fred	1	1	0	0
☐	Annie	1	0	1	0
☐	Harold	0	1	0	0
☐	Grace	1	1	0	0
☐	Paul	0	0	1	0

Figure 5: TP,TN,FP,FN based on customer feedback

So Total value of True Positive(TP) = 133 , Total value of True Negative(TN) = 81 , Total value of False Positive(FP) = 34 , Total value of False Negative(FN) = 10

- ***Recall*** = 133/(10+133) = 0.93
- ***Precision*** =133/(34+133) = 0.79
- ***Accuracy*** = (133+81)/(133+81+34+10) = 0.82

Result shows that improved K means algorithm provide accuracy approximate 82 percent.

VI. CONCLUSION

In this Research work the main focus on providing good quality recommendation system to the registered users. The important thing about the project is it provide dynamic recommendation to the registered user i.e. clusters vary time to time depend upon the number of users registered and the according to their age. The proposed recommendation system minimizes the false positive error that occurs frequently in traditional system. Results prove that accuracy achieved using improved k-means that is 82 to 85 percent is better that the old k-means algorithm. The recommendation system has the potential to attract the new customer and maintain the existing one. The proposed work represents age based clustering method that improved K-means clustering algorithm performance and accuracy in the area of recommending products such as books to users. This paper concludes that increasing efficiency of K-mean algorithm and Users find better results corresponding to their views and purchases of books. The proposed system is also used for other recommendation like movie, music, electronic items etc.

VII. REFERENCES

[1] Nikita Jain, Vishal Srivastava "DATA MINING TECHNIQUES: SURVEY PAPER"IJRET eISSN: 2319-1163 | pISSN: 2321-7308.

[2] Fahim A.M., Salem A.M., "Efficient enhanced k-means clustering algorithm", Journal of Zhejiang University Science, 1626 – 1633, 2006.

[3] Wang Shunye "An Improved K-means Clustering Algorithm Based on Dissimilarity" 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)Dec 20-22,2013, Shenyang, China IEEE.

[4] Sanjay garg, Ramesh Chandra Jain, "Variation of k-mean Algorithm: A study for High Dimensional Large data sets", Information Technology Journal5 (6), 1132 – 1135, 2006.

[5] Navjot Kaur, Jaspreet Kaur Sahiwal, Navneet Kaur "EFFICIENT K-MEANSCLUSTERING ALGORITHM USING RANKING METHOD IN DATA MINING" ISSN: 2278–1323 International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012.

[6] Md. Sohrab Mahmud, Md. Mostafizer Rahman, and Md.Nasim Akhtar "Improvement of K-means Clustering algorithm with better initial centroids based on weighted average" 2012 7th International Conference on Electrical and Computer Engineering 20-22 December, 2012, Dhaka, Bangladesh, 2012 IEEE.M. Young, The Technical Writer's Handbook. Mill Valley, CA: University Science, 1989.

[7] Pallavi Purohit "A new Efficient Approach towards k-means Clustering Algorithm", International journal of Computer Applications,Vol 65-no 11,march 2013.

[8] Friedrich Leisch1 and Bettina Gr un2, "Extending Standard Cluster Algorithms to Allow for Group Constraints", Compstat 2006, Proceeding in Computational Statistics, Physica verlag, Heidelberg, Germany,2006.

[9] Juntao Wang & Xiaolong Su "An improved K -Means clustering algorithm" 2011 IEEE.

[10] Shuhua Ren &Alin Fan "K-means Clustering Algorithm Based on Coefficient of Variation" 2011 4th International Congress on Image and Signal Processing2011 IEEE.

[11] SongJie Gong and Zhejiang "Joining User Clustering and Item Based Collaborative Filtering in Personalized RecommendationService",2009, 10746963,DOI: 10.110 9/IIS.2009.70, IEEE, Haikou, China

[12] K. A. Abdul Nazeer, M. P. Sebastian "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm"2009, Proceedings of the World Congress on Engineering 2009 Vol I WCE 2009, July 1 - 3, 2009, London, U.K."

[13] Robert M Bell and Yehuda Koren "Scalable Collaborative Filtering with Jointly Derived Neighborhood Interpolation Weights", Seventh IEEE International Conference on Data Mining.

[14] Micheal Pazzani, "A Framework for Collaborative, Content-Based and Demographic Filtering". University of California, 2004

.