Aggregate Profiling for Recommendation of Web Pages using SOM and K-means Clustering Techniques

Shveta Kundra Bhatia
Research Scholar
Dept. of Computer Science
University Of Delhi, India

Harita Mehta
Research Scholar
Dept. of Computer Science
University Of Delhi, India

Veer Sain Dixit
Dept. of Computer Science
ARSD College
University Of Delhi, India

ABSTRACT
Since, number of users are increasing exponentially so proper analysis of such data by devising efficient algorithms is essential which ultimately helps in determining the life time value of customers and judging the effectiveness of promotional campaigns as well. Better services and quality can be provided by mining the web access log files. In this paper,we have shown that with the help of clustering techniques, Self Organized Feature Maps and K-Means useful knowledge is extracted. We have also proposed to derive the interest and behavior of a significant group of users by applying the concept of "Aggregate Usage Profile". Further, this technique has been used for looking frequently accessed pages for recommendations.

Keywords
Web Usage Mining, K-Means, Self-Organizing Feature Maps and Aggregate Usage Profile

1. INTRODUCTION
Web Usage Mining [7, 8, 13, 15] discovers meaningful patterns from data generated by Client-Server transactions. Web Usage Mining research mainly focuses on the data from the Web server side. The logs are pre-processed to group requests from the same user into sessions. A session contains the requests from a single visit of a user to the Web site. During the pre-processing, irrelevant information for Web Usage Mining such as background images and unsuccessful requests are ignored. The users are identified by the IP addresses in the log and all requests from the same IP address within a certain time-window are put into a session. Clustering of similar sessions help in doing study of the users having similar interests by using which groups of web pages can be recommended to users of a similar category.

We have organized the research work into seven sections.Section 2 deals with literature review of research work done related to K-means and SOM clustering techniques and step wise procedure of Web Usage Mining. Section 3, discusses about SOM and K-Means clustering techniques used for Web Usage Mining. In Section 4, Aggregate Profiling with pageview-weight pairs is addressed. In section 5, there is a discussion on implementation part of the said problem. Section 6consists of analysis of results. Finally, Section 7 concludes thepaper along with future research prospects.

2. LITERATURE REVIEW
Jianhan [22] in the year 2002 used the Citation Cluster Algorithm that constructs a conceptual hierarchy of the web site. Borges and M-Levene [23] suggested a dynamic clustering based method in the year 2004 that represented a collection of user web navigation sessions. Self-Organizing Feature Maps were used by Paola Britos, Damian Mastinelli, Herman Merlino, Raman Gracia, and Martinez [3] in the year 2007 to compare results on two different web sites and also detailed the transformations necessary to modify the data storage in the web server log files. Mehrdad Jalali, Norwati Mustspha, Ali Mamat, Md. Nasir B. Suleiman [24] in 2008 applied graph partitioning to establish an undirected graph based on connectivity between each pair of web pages and also proposed formulas for assigning weights to edges of the graph. A web based recommender system was developed by Mehdi, Hosseini [25] in the year 2008 to predict user's intention and their navigation behaviors. Chu-Hui Lee, Yu-Hsiang Fu [26] predicted users browsing behavior and improved the two level prediction models to achieve higher hit ratio by using Hierarchical Agglomerative Clustering in 2008. Kobra Etminani Mohammad-R, Akbarzadeh-T, Noorali Raeeji Yanehsari [27] extracted frequent patterns for pattern discovery and displayed the results in an interpretable format by applying ant based clustering algorithm. Illustration of how to use DTS, T-SQL and other tools under SQL server 2000 to realize data transfer, cleaning, user recognition, session recognition and usage of OLAP as well as DM to realize mode discovery and mode analysis was studied and applied by De Min Dong[28] in 2009. In 2010 N. Sujatha, K. Iyakutty [29] modified the iterative K-Means algorithm converging it to a better local minimum byproposing a GA based refinement algorithm to improve cluster quality.

2.1 Web Usage Mining

Web usage mining [7, 8, 13, 15, 5, 18] includes the following five major steps:

1. Data Collection: By collecting data from web server logs, proxy server logs, cookies and meta data integrity and authenticity of data is maintained. In this paper the data used is web server log data which has 5999 entries and can be found at www.vtsns.edu.rs.

2. Data Pre-processing: This step includes transformation of web log data by cleaning and structuring data to prepare for pattern extraction. Pre-processing in our work has been done by the sawmill tool.

3. Pattern Discovery: Patterns of interest could be extracted using techniques such as statistical analysis, association rules, clustering, classification, sequential pattern detection and dependency modeling. Our focus in this paper is on clustering [10] techniques SOM and K-Means.

4. Pattern Analysis [4]: Elimination of rules that are irrelevant and analysis of extracted patterns using techniques such as visualization, OLAP, data and knowledge querying or usability analysis. In our paper, pattern analysis is done using Aggregate Usage Profile.

5. Applications: After generating [17] analyzed patterns by using some efficient method leads to providing recommendations for the browsing process, personalization of web sites and improvement of web site designs.

3. INTRODUCTION TO SELF ORGANIZING FEATURE MAPS AND K-MEANS ALGORITHM

SOM [1, 2, 3] is a kind of unsupervised learning technique of Neural Networks [1, 2, 3, 14] which helps in reducing the high dimensional data into low dimensional data and visualizes that. Based on competitive learning principles, SOM helps in clustering data together for analysis and in clustering similar sessions together. By analyzing these clusters we can find frequently accessed pages by a set of similar users.

3.1 SOM Algorithm

1. Assign random values to weight vectors of a neuron.
2. Provide an input vector to the network.
3. Traverse each node in the network
   a) Find similarity between the input vector and the network's node's weight vector using Euclidean Distance.
   b) Find the node that produces the smallest distance which is the Best Matching Unit (BMU)
4. Update the nodes in the neighborhood of BMU by changing the weights using the following equation:

$$Wv(t+1) = Wv(t) + \Theta(t)\alpha(t)(D(t) - Wv(t))$$

Where,

- t denotes current iteration
- λ is the limit on time iteration
- **Wv** is the current weight vector
- *D* is the target input
- $\Theta(t)$ is the neighborhood function
  In this algorithm neighborhood function has been derived using Gaussian function.
- $\alpha(t)$ is learning rate due to time

5. Increment t and repeat from step2 while $t < \lambda$.

The k sessions and the set of m unique URLs are the input to the SOM network. The input is represented by a two dimensional matrix of order m x k.

3.2 K-means Algorithm

K-means [3, 29] is also considered to be one of the important tools for clustering problems.

K-means works using the following steps:

1. Place K objects points into the space that are to be clustered. object points always represent initial group centroids.
2. Assign each object point to the group that has *the closest* centroid.
3. When all object points have been assigned, re-calculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the object points into groups from which the metric to be minimized can be calculated.

The algorithm aims to minimize an objective function:

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

Where $\|x_i^{(j)} - c_j\|^2$ is a chosen distance measure between a data point and the cluster centre. It is an indicator of the distance of the n data points from their respective cluster centers.

4. AGGREGATE USAGE PROFILE

The useful information gained from the web log data by applying SOM and K-means represent user segments based on various attributes. The transaction clusters obtained are not an effective means of capturing aggregated view of common user patterns. Our approach creates an aggregate view of each cluster to compute the centroid of URLs in each cluster. The dimension value for each pageview is computed by finding the ratio of the sum of the pageview weights across sessions to the total number of sessions in the cluster. Given a transaction cluster cl, Aggregate Profile as a set of pageview-weight pairs is computed as follows:

$$pr_{cl} = \{(p, weight(p, pr_{cl}))l\ weight(p, pr_{cl}) \geq \mu\}$$

Where

$$weight(p, pr_{cl}) = \left(\frac{1}{|cl|}\right)\sum w(p, s); s\ \varepsilon\ cl$$

l*cl*l is the number of transactions in cluster cl; w(p, s) is the weight of page p in session vectors s of cluster cl and μ is used to focus only on those pages in the cluster that appear in sufficient number of vectors in the cluster. These aggregate representations can be used for predictive modeling and in applications such as recommender systems [11].

5. IMPLEMENTATION

5.1 Data Set

Data used in the experiment is taken from a log file containing information about all web requests to an institution's official website on November 16, 2009. Each line in the web usage log file contains the following information like date/time of request, hit type, page, hostname, referrer, server domain, authenticated user, server response, page views, size etc. The raw web log file, we used for the experiment contained 5999 web requests. This file can be found at http://www.vtsns.edu.rs/maja/vtsnsNov16.

5.2 Data Preparation and Session Identification

To prepare the web log data [19] for the mining process, it needs to be cleared of irrelevant requests; and transformed to a format that can be fed into the clustering algorithm. For pre-processing and creation of sessions we used a tool called Sawmill. Sawmill is a software package for pre-processing, session creation, statistical analysis and reporting of log files. It has been developed by Flowerfire Inc. in 1998 in the C language and has the capability to work on various different platforms. Sawmill [21] computes session information by tracking the page, date/time, and visitor id for each page view. When a session view is requested, it processes all of these page views at the time of the request. Sawmill groups the hits into initial sessions based on the visitor id by assuming that each visitor contributes to a session. In the next step sorting by time is performed for a click-by-click record of each visitor. A session timeout interval of 30 minutes is considered for generating final sessions and sessions longer than 2 hours are eliminated.

5.3 Transforming the file format

The sessions that were generated using sawmill were integrated together into a matrix where the rows define the sessions and the columns the URLs. The matrix was encoded using scripts written in the tcl language in the following format:

URLs → **Session ID ↓**	**X1**	**X2**	**...**	**X43**	**Label**
1	1	0			Session1
2					Session2
N					Session n

The above matrix has entries 0 or 1 depending on the following conditions:

i) 1 if the page Xi has been accessed in the session id j.
ii) 0 if the page Xi has not been accessed in the session id j.

6. EXPERIMENTAL RESULTS

The raw web log file we used for the experiment contained 5999 web requests. Using the Sawmill tool on our web log data led to the creation of sessions and 110 unique URLs. 72.9% of the total sessions were exported into a .csv format with the help of scripts in tcl language as rest of the sessions had only either one or two page views. Further we optimized our matrix and 59.1% of the sessions and 43 unique URLs were used for experimentation. The optimization was performed on the basis of sessions having less than 3 pageviews and pages that were viewed 5 or less than 5 times have been removed. The optimized matrix was used for clustering using the Self-Organizing Feature Maps and K-Means algorithms. We used the Spice-SOM [30] tool and SPSS software for implementation of the respective algorithms. Applying the two algorithms we can see that clusters with similarity among sessions have been obtained and can be used for prediction of pages to a user of similar interests. Clusters of sizes 10, 15 and 20 were generated using both the techniques to study the variation in percentage sessions lying in each cluster. Further, we applied aggregate usage profile to calculate the weights of pages lying in a single cluster. The threshold value μ was taken as 0.3 to from groups of URLs that can be recommended to a set of similar users.

6.1 Results

We apply the two techniques of SOM and K-Means and study the distribution of sessions by varying the number of clusters as 10, 15 and 20.

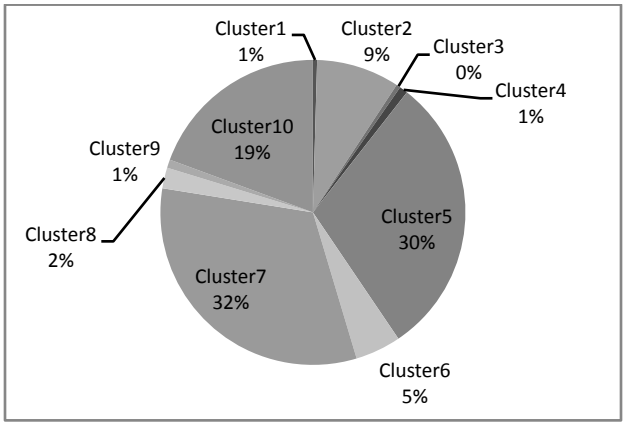

Fig 1: Percentage sessions using K-Means with K=10

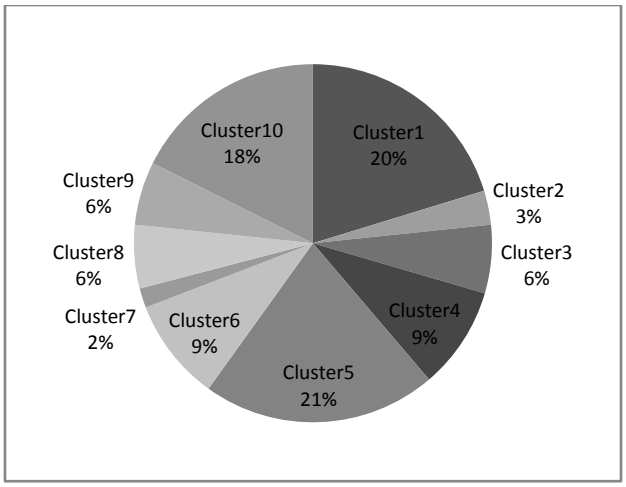

Fig 2: Percentage sessions using SOM with output matrix 5 x 2

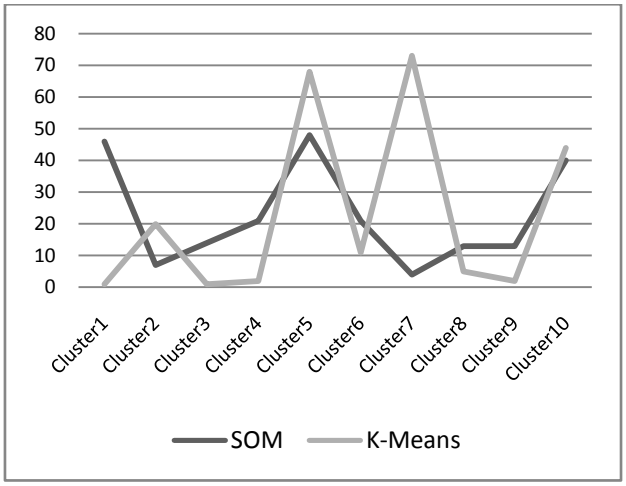

Fig 3: Comparison of number of session's for 10 clusters using SOM and K-Means

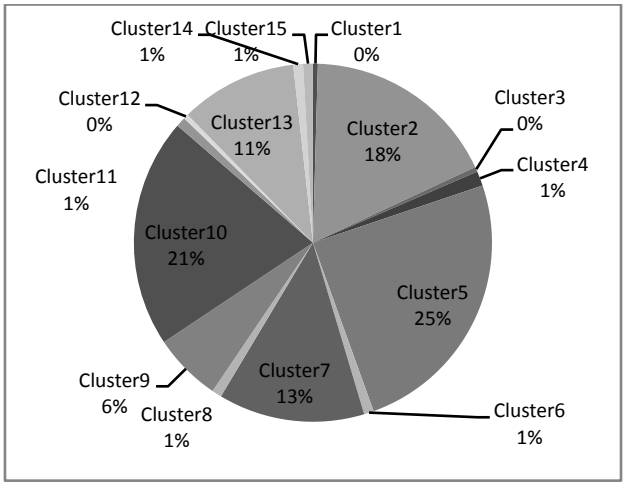

Fig 4: Percentage sessions using K-Means with K=15

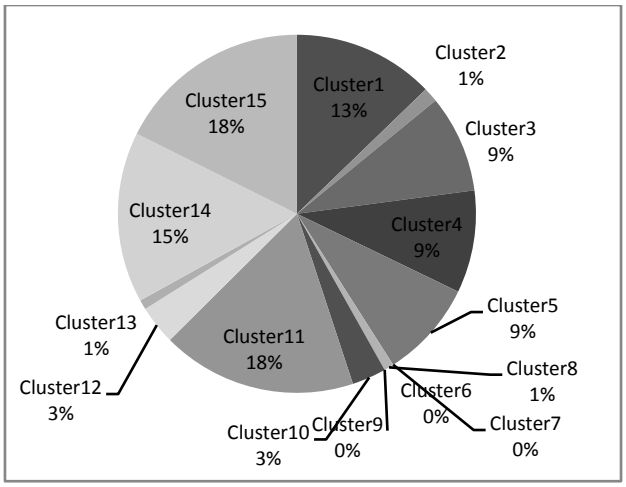

Fig 5: Percentage sessions using SOM with outputmatrix 5 x 3

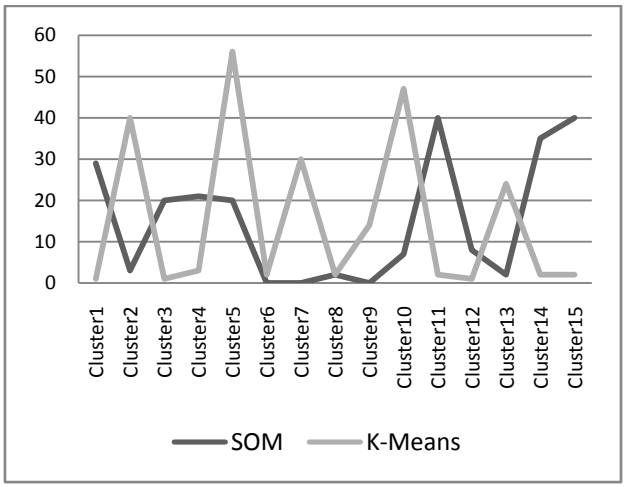

Fig 6: Comparison of number of session's for 15 clusters using SOM and K-Means

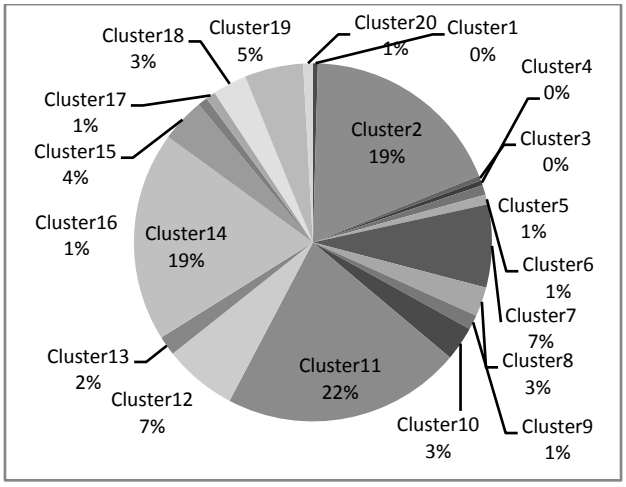

Fig 7: Percentage sessions using K-Means with K=20

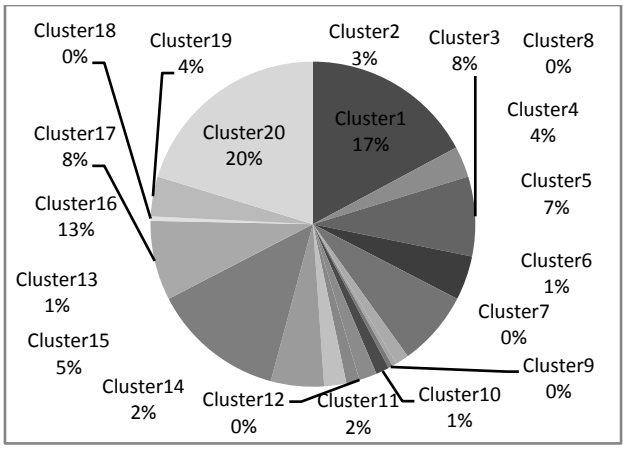

Fig 8: Percentage sessions using SOM with output matrix 5 x 4

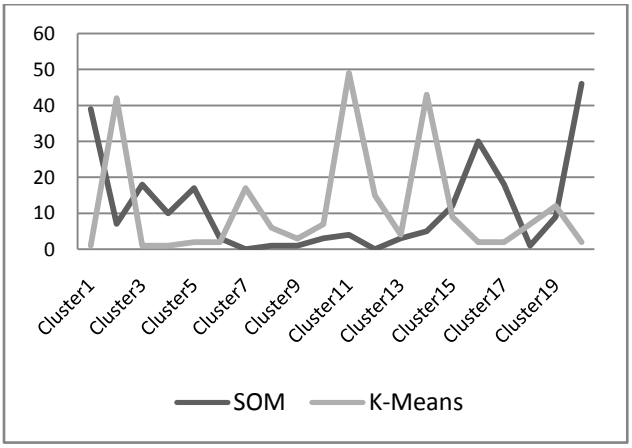

Fig 9: Comparison of number of session's for 20 clusters using SOM and K-Means

In the above graphs figure 1 and 2 depict the percentage sessions in the clusters using SOM and K-Means for number of clusters being 10. Figure 4 and 5 depict the percentage sessions in the clusters using SOM and K-Means for number of clusters being 15. Figure 7 and 8 depict the percentage sessions in the clusters using SOM and K-Means for number of clusters being 20. Figure 3, 6 and 9 represent the comparison of number of session's for 10, 15 and 20 clusters respectively using SOM and K-means. By observation as the value of K increases, the distribution of sessions almost remains similar in K-Means where a few clusters have very large number of sessions and the rest of the clusters have very few sessions in them. The SOM algorithm has some clusters with zero sessions in it as the value of K increases. Now by applying Aggregate Usage Profile on the clusters obtained by SOM and K-means algorithms we get the following results.

Table 1: Aggregate Usage Profile for cluster 13 with pageview and weights using SOM

Pageview	Weight
X1	1.00
X2	1.00
X3	1.00
X7	0.66
X8	0.66
X9	0.33
X10	0.33
X13	0.33
X14	0.33
X21	0.33
X23	0.33
X27	0.33

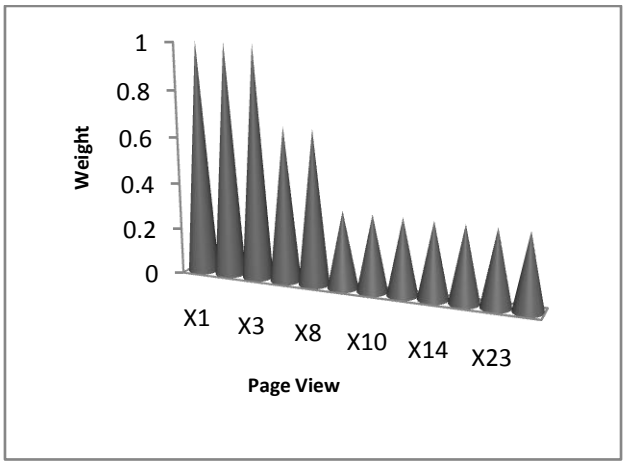

Fig 10: Recommendations using aggregate usage profile from cluster 13 of 5x4 output distribution using SOM

Table 2: Aggregate Usage Profile for cluster 15 with pageviewand weights using K-Means.

Pageview	Weight
X1	1.00
X19	0.67
X21	0.33
X24	0.56
X37	0.44
X40	0.44

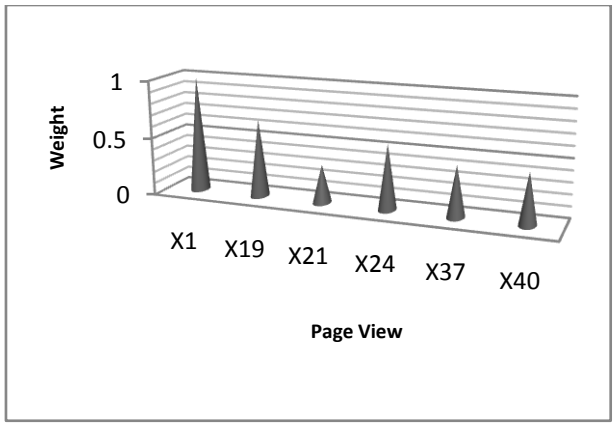

Fig 11: Recommendations using aggregate usage profile from cluster 15 out of 20 clusters using K-Means

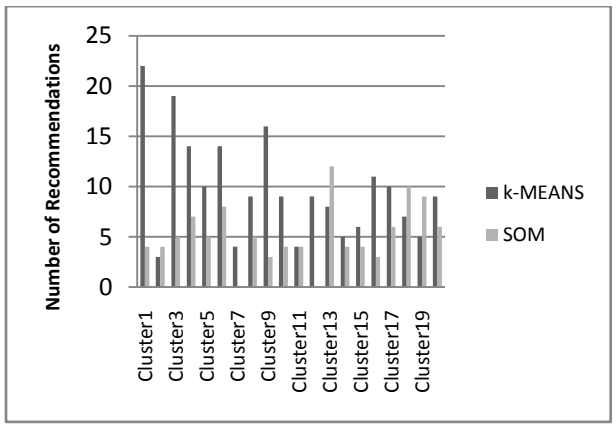

Fig 12: Comparison of number of recommendations using SOM and K-Means for 20 clusters

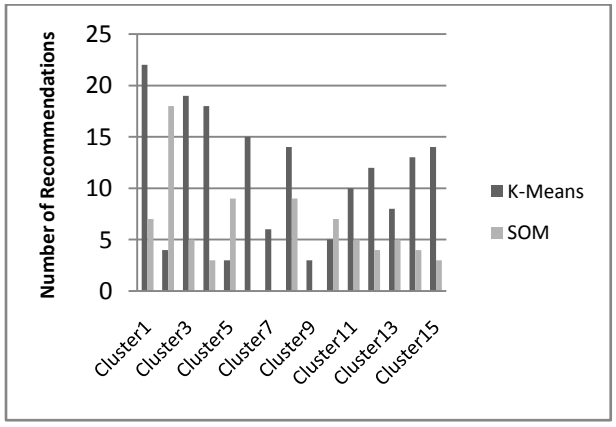

Fig 13: Comparison of number of recommendations using SOM and K-Means for 15 clusters

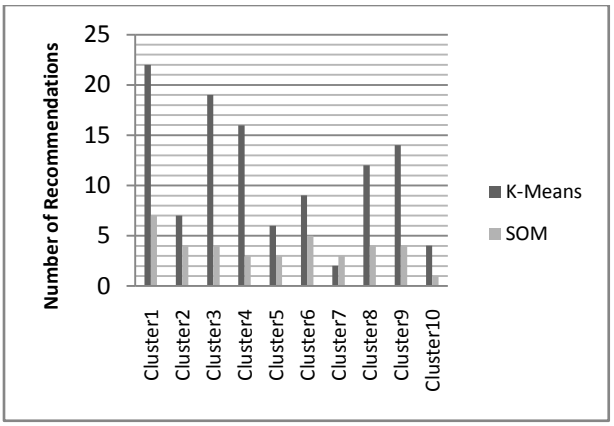

Fig 14: Comparison of number of recommendations using SOM and K-Means for 10 clusters

Figure 10 and 11 above represent the recommendations using Aggregate Usage Profile for cluster 13 and 15 along with their weights and URL's names. Figure 12, 13 and 14 display the comparison of the number of recommendations using SOM and K-means for 20, 15 and 10 clusters respectively. In summary, the graphs above show that Aggregate Usage Profile that was applied on clusters obtained from SOM and K-Means techniques can be an effective way for personalization on the basis of pages viewed by the users for recommendation. We can see that Aggregate Usage Profile captures relevant patterns that can be used for usage based recommendations.

7. CONCLUSION AND FUTURE WORK

The number of recommendations obtained using K-Means is very large with an anomaly that all the recommendations are part of a single session as the Aggregate Usage Profile values are 1 for almost all recommended URLs. Whereas Aggregate Usage Profiles for URLs obtained using SOM are better in terms of recommendation because theirAggregate Profile values lay between 0.33 and 0.89; indicating that more number of sessions have the same URLs (users with similar interests). By observation it is found that the recommendation process is better using SOM as compared to K-Means in terms of Aggregate Usage Profiling.

In future, we will work on the complexity and performance of recommendations generated along with the usage of other Web Usage Mining clustering algorithms to generate better Aggregate Usage Profiles with voluminous data sets; and develop algorithms to enhance and compare the performance of clustering processes on the basis of various index values by removing unwanted clusters.

8. REFERENCES

[1] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero and A. Saarela. Self-organization of a massive document collection. IEEE Transactions on Neural Networks, 11(3):574–585, 2000.

[2] Kate A. Smith and Alan Ng. Web page clustering using a Self-organizing map of user navigation patterns. Decision Support Systems, 35(2):245–256, 2003.

[3] Web Usage Mining Using Self Organized Maps, Paola Britos, Damián Martinelli, Hernan Merlino, Ramón García-Martínez IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.6, June 2007.

[4] X. Wanga, A. Abraham, K. A. Smitha. Intelligent web traffic mining and analysis. Journal of Network and Computer Applications, vol. 28, 2004, pp. 147–165.

[5] R. Iváncsy, I. Vajk, Different Aspects of Web Log Mining. 6th International Symposium of Hungarian Researchers on Computational Intelligence. Budapest, Nov., 2005.

[6] R. Kosala, H. Blockeel, Web Mining Research: A Survey, ACM SIGKKD Explorations, vol. 2(1), July 2000.

[7] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKKD Explorations, vol.1, Jan 2000.

[8] B.Moshaber, R. Cooley, J. Srivastava, Automatic Personalization Based on Web Usage Mining, Communications of the ACM, vol.43 (8), 2000.

[9] S. K. Pal, V Talwar, P Mitra. Web Mining in Soft Computing Framework: Relevance. State of the Art and Future Directions. IEEE Trans. on Neural Networks, vol.13 (5), 2002, pp. 1163–77.

[10] Dehu Qi, Chung-Chih Li, Self-Organizing Map based Web Pages Clustering using Web Logs. Conference of Software Engineering and Data Engineering 2007, 265-270.

[11] Ranieri Baraglia and Fabrizio Silvestri, "An Online Recommender System for Large Web Sites", Web Intelligence, 2004 Proceedings. IEEE/WIC/ACM International Conference on 20-24 Sept. 2004

[12] Web Data Mining Research: A Survey, Brijendra Singh, Hemant Kumar Singh.IEEE 2010 Conference.

[13] S. Santhi, Dr. Purushothaman Srinivasan, "An Improved Usage Mining using Back Propagation Algorithm with Functional Update", 2009 IEEE International Conference Advance Computing Conference (IACC 2009), 978-1-4244- 2928-8/09.

[14] Hanan Ettaher Dagez &Mhd Sapiyan Baba, "Applying Neural Network Technology in Qualitative Research for Extracting Learning Style to Improve E-Learning Environment, The IEEE International Conference, 978-1-4244-2328-6/08 2008.

[15] Hafidh Ba-Omar, Ilias Petrounias and Fahad Anwar, "A Framework ofWeb Usage Mining to Personalize E-Learning", Seventh IEEE International Conference on Advanced Learning Technologies (ICALT 2007) 0-7695-2916-X/07 007 IEEE.

[16] Zurina Muda and Ros Emiliana Kartina Mohamed "Adaptive User Interface Design In Multimedia Courseware" IEEE 0-7803-9521-2/06 2006.

[17] Ekaterina Vasilyeva, Mykola Pechenizkiy, Seppo Puurone "Towards the Framework of Adaptive User Interfaces for eHealth", Proceedings of the 18th IEEE Symposium on Computer-Based Medical Systems (CBMS'05) 1063-7125/05 2005.

[18] Pattern Discovery of Web Usage Mining. Nina, S.P. ; Rahman, M. ; Bhuiyan, K.I. ; Ahmed,K.; Comput. Sci. & Eng., Shahjalal Univ. of Sci. & Technol., Sylhet, Bangladesh. In: Computer Technology and Development, 2009. ICCTD '09. International Conference on 13-15 Nov. 2009 Volume: 1, 499

[19] Tsuyoshi Murata and Kota Saito "Extracting Users Interests from Web Log Data", Proceedings of the 2006 IEEE/WIC/ACM International Conference of Web Intelligence (WI 2006 Main Conference Proceedings) (WI'06) 2006 IEEE.

[20] Web usage mining: Discovery of the users' navigational patterns using SOM.Etminani, K. ; Delui, A.R. ; Yanehsari, N.R. ; Rouhani, M. ; Dept. of Comp. Eng., Ferdowsi Univ. of Mashhad, Mashhad, Iran. In:Networked Digital Technologies, 2009. NDT '09. First International Conference on 28-31 July 2009. Page 224

[21] SAWMILL: http://www.sawmill.net

[22] Page Cluster: Mining conceptual link hierarchies from Web log files for adaptive Web site navigation: Jianhan Zhu, Jun Hong, John G. Hughes; published in Journal ACM Transactions on Internet Technology.Volume 4 Issue 2, May 2004

[23] A Dynamic Clustering-Based Markov Model for Web Usage Mining; Jos_e Borges, Mark Levene Birkbeck, May 26, 2004

[24] Jalali, M., Mustapha, N., Mamat, A., and Sulaiman, M.N. A new clustering approach based on graph partitioning for navigation patterns mining. In Proceedings of ICPR. 2008, 1-4.

[25] Mehdi Hosseini, Hassan Abol Hassani, "Mining Search Engine Query Log for evaluating Content and Structure of a Web Site" in Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence.

[26] Chu-Hui Lee, Yu-Hsiang Fu, "Web Usage Mining Based on Clustering of Browsing Features," isda, vol. 1, pp.281-2862008 Eighth International Conference on Intelligent Systems Design and Applications, 2008

[27] KobraEtminani,Mohammad-R. Akbarzadeh-T., Noorali Raeeji Yanehsari, "Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method", IFSA-EUSFLAT 2009

[28] DeMin Dong, "Exploration on Web Usage Mining and its Application", International Workshop on Intelligent Systems and Applications, Pp. 1-4, 2009

[29] N. Sujatha, K. Iyakutty, "Refinement of Web usage Data Clustering from K-means with Genetic Algorithm", European Journal of Scientific Research ISSN 1450-216X Vol.42 No.3 (2010), pp.464-476

[30] Written by Cao Thang in Soft Intelligent Laboratory, Ritsumeikan University, 2003-200

[31] HaritaMehta,Shveta Kundra Bhatia, Punam Bedi,V.S.Dixit, "Collaborative Personalized WebRecommender Systemusing Entropy based Similarity Measure", International Journal of Computer Science Issues, Vol. 8, Issue 6, No 3, November 2011.

9. AUTHORS PROFILE

Shveta Kundra Bhatia is a Research Scholar and working as an Assistant Professor in the Department Of Computer Science, Swami Sharaddhanand College, University of Delhi. Her research area is Web Usage Mining and is currently pursuing PhD under Dr. V.S. Dixit from Department of Computer Science, University of Delhi.

Harita Mehta is a Research Scholar and working as an Assistant Professor in the Department of Computer Science, Acharya Narender Dev College, University of Delhi. Her research area is Web Recommender Systems and is currently pursuing PhD under Dr. V.S. Dixit from Department of Computer Science, University of Delhi.

Dr. V. S. Dixit is working as senior Assistant Professor in the Department Of Computer Science, AtmaRam Sanatam Dharam College, University of Delhi. His research area is Queuing theory, Peer to Peer systems, Web Usage Mining and Web Recommender systems. He is currently engaged in supervising the research scholars (Ms. Harita Mehta and Ms. Shveta Kundra Bhatia) for PhD. He is Life member of IETE.