ISSN NO. 2320-5407

Journal homepage: http://www.journalijar.com

INTERNATIONAL JOURNAL OF ADVANCED RESEARCH

RESEARCH ARTICLE

Comparative Analysis of Web Usage Data using SOM and K-means algorithms

***Shilpa M Patil[1], T Vijaya Kumar[2], Dr. H S Guruprasad[3]**

1. PG Scholar, Dept. of CSE, BMSCE, Bangalore, INDIA
2. Associate Professor, Dept. of CSE, BIT, Bangalore, INDIA
3. Professor and Head, Dept. of CSE, BMSCE, Bangalore, INDIA

Manuscript Info

Manuscript History:

Received: 15 August 2015
Final Accepted: 26 September 2015
Published Online: October 2015

Key words:

Breeding season, Buffaloe,
Post- partum anoestrus,

Corresponding Author

Shilpa M Patil

Abstract

Web data is becoming very popular for the transformation and distribution of valuable information which can be freely accessible by users. Hence Web is becoming too large and diverse. Organization of data on the Web for the efficient access has become a big challenge for the Web site administrators. So there is a need to apply data mining and neural network techniques to extract information from the Web for the better organization of Web data. Web usage mining is one of the main research areas which focus on extracting valuable information from the Web by using Web usage data. Web usage mining is a part of data mining that is much needed to find out patterns or clusters with help of user's session and behaviour. Web usage mining process starts with pre-processing followed by clustering of data and finally visualization of clusters effectively. We have considered, Web usage mining to find required information by analysing Web usage data using two Self Organizing Map and K-means. As to have a comparison between both methods in clustering of Web usage data we need to initially prepare Web navigational data available for clustering. So we will start with pre-processing of log file to remove unwanted data followed by removal of redundant data and later separating the users as well as sessions depending on time interval. In pre-processing phase the sessions are formed based on time interval taken by each of the user. Once all the sessions are formed we will make use of Self Organizing Map (SOM) algorithm to segregate them into different clusters using the weight matrix. To compare SOM, with K-means algorithm clusters are formed using K-means algorithms. Clusters formed by both the algorithms are visualized by using JFree tool. Finally, the charts obtained by both the algorithms are compared to analyse the clusters.

INTRODUCTION

Web is a current powerful platform for discovering knowledge and to gain of needful information by analysing the Web data. Web mining, as a part of data mining includes various types and one among them is Web usage mining. Web usage mining is the application of data mining to discover required and interesting patterns from Web data which gives clear understanding of Web based applications. Web usage mining process starts with pre-processing of raw data, followed by clustering of data and finally visualization of the formed clusters effectively. Clustering is a technique which groups the complete data set into clusters such that data objects in each cluster will have some similarity. K-means is one such well known and most used clustering method. Usually K-means is used to cluster very large set of data. In K-Means algorithm, we need to calculate the distance between each of data object and cluster centre in every iteration. As K-means is partitioning and iterative method the clusters formed will be independent and compact. There will be random selection of K centres in the first iteration and obtain clusters by

taking data objects which is nearest to the cluster centres. In the subsequent iterations cluster centres are updated based on the nearest data objects. Generally Euclidean distance is considered to find out distance between each data object and cluster centre. So in this way similar data objects are brought together to form clusters. Similarly SOM is used to cluster the data with an idea of assigning the same number of instances to each session. SOM is one of the neural network algorithms which not only cluster the data but also reduces the dimensionality of data. SOM makes use of weight matrix that is obtained from previous steps. Weight matrix includes weights calculated for each session. So here we illustrate the feasibility of using SOM to mine the Web log data of user navigation. And finally we need to analyze the pattern that is nothing but visualizing the clusters formed by SOM. And this is done by using visual studio and through line chart and pie chart for both algorithms. Later we will have clear comparison between both algorithms. A brief description about the work proposed by other researchers is presented in Section 2. The overall architecture of K-means algorithm and the details of SOM are given in Section 3. The comparative study of clusters formed using SOM and K-means are discussed in section 4. Conclusion and future work are briefed in section 5.

Literature Review

Several authors have contributed by providing various clustering models to use to discover valuable information from Web usage data. In [1], Paweł Weichbroth, et al., have provided a framework for mining the Web navigational data patterns in order to know the management of Web usage data. In [2], Nawal Seal et al., have used data techniques to analyze learner behaviour in Educational systems. They have defined various static variables along with SCORM content tree and multidimensional graphs in their approach. In [3], Laura Hollink, et al., characterized Web sites in terms of meaning of queries that lead to know about large data sets on Web. They demonstrated exploitation of such links for effective mining of data along with how patterns can be used effectively. In [4], R.M. Suresh, et al., have discussed the importance of data pre-processing methods along with the required steps. In [5], the authors have developed a Visualization tool called LOGSOM, using a Self-Organizing Map of user navigation patterns. Cluster analysis is used very frequently in data mining. Most of the researchers have used K-means as the clustering technique. Improvement for K-means algorithm is proposed by many authors. Subtractive clustering algorithm can be used to estimate the number of clusters and the cluster centres in a set of data. The subtractive clustering method assumes that each data point is a potential cluster centre. A data point with more neighboring data will become a cluster centre than points with fewer neighboring data [6]. In [7], Shi Na et al., have proposed an improvement by storing information obtained in each repetition that can be used in the next iteration, and combining K-means with other techniques like Affinity Propagation is proposed in [8]. In [9], Anil K. Jain has provided the details of clustering algorithms and useful research directions in clustering such as semi-supervised clustering, simultaneous feature selection in during data clustering , large scale data clustering etc. The reduction of dimensionality for K-means clustering algorithm using feature selection and feature extraction is proposed in [10]. K-means initially starts with partition of data. As k-means is sensitive to initial condition, Juanying Xie et al., [11] have proposed a new global k-means algorithm that takes less time to run than K-means. They have proposed a function to select a candidate centre for next cluster and effectively reduced the computational time. There are various tools to visualize web access requests. In [12], Jiyang Chen et al., have proposed visual data mining system that allows interactive investigation of web data as well as ad-hoc knowledge data patterns discovery of web navigational behaviour. Web site specific factors such as concept hierarchy and website graphs can be included with navigational data to achieve better results for web page recommendations [13]. In [14], Esin Saka et al. have proposed a hybrid approach which combines the strengths of Spherical K-means algorithm for fast clustering of high dimensional datasets and the flock-based algorithm.

System Design

The block diagram used in our work is depicted in Fig. 1. We have analyzed the data by comparing the clusters formed by Self Organizing Map and k-means. First we identify user and sessions from the Web server log file and then construct a graph and discover homogeneous groups with the help of SOM and K-means clustering techniques. Then the clusters are t visualized for analysing the formed clusters. Web usage mining is a process of extracting the user behaviour on a Website by analysing the server log file. Web usage mining initially starts with pre-processing of log file where the cleaning of data takes place along with removal of unwanted and redundant data. User identification and Session identification is done based on time heuristics and navigation heuristics. Once the sessions

are formed we make use of Self Organizing Map clustering algorithm to form the cluster taking weight matrix as input. Clusters are also formed using K-means algorithm. Hence the pattern discovery after pre-processing of data will be done using SOM algorithm and also using K-means algorithm to form the clusters. Clustering divides the data into groups as clusters where data objects in each cluster will have some similarity. As K-means is partitioning and iterative method the clusters formed will be independent and compact. There will be random selection of K centres and next starts with taking each data object which is nearest to that and Euclidean distance is considered to find out distance between each data object and cluster centre. So in this way similar data objects are brought together to form clusters. SOM is used to cluster the data based on similarity, with an idea of assigning the same number of instances to each session. SOM is one of the neural network algorithms which not only cluster the data but also reduces the dimensionality of data. Here SOM makes use of weight matrix that is obtained from previous steps. Weight matrix includes weights calculated for each session. So here we illustrate the feasibility of using SOM to mine the Web log data of user navigation. Next step is to analyze the pattern that is nothing but visualizing the clusters formed by SOM. And this is done by using visual studio and through line chart and pie chart for both algorithms. Later we will have clear comparison between both algorithms. To visualize the formed of clusters from both algorithms we make use of Jfree charts. Jfree charts provide one of the best ways to represent data by using charts like pie chart and line chart. So that we can also have differences between them and can compare both methods.

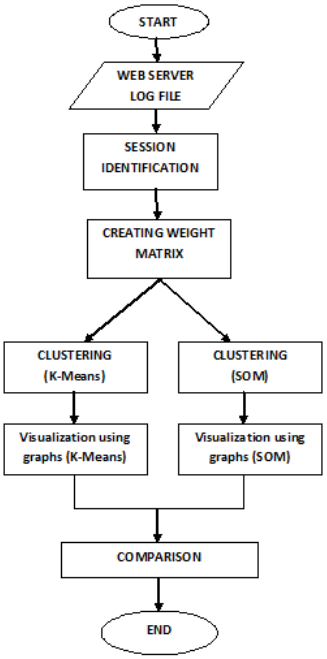

Fig.1: Clustering and Visualization using SOM and K-means

The K- means algorithm used in our approach is summarized as follows:

Input: The session weight matrix which is obtained in, pre-processing stage, Number of clusters to be formed, cluster centre initialization and distance matrix.

Output: various clusters with closest centroid.

The K-means clustering algorithm for Web usage data comprises the following four steps.

Step 1: Choose K initial cluster centres (representing k groups) randomly from the data set which includes around 2398 sessions formed in pre-processing steps.

Step 2: Assign all sessions to their closest cluster (measuring from the cluster centre). This is done by

presenting each data point x and calculate the similarity (distance) d of this input entries y of each cluster centre j. the closest cluster centre to a data point x is the cluster centre with minimum distance to the data point x.

Step 3: Recalculate the centre of each cluster as the centroid of all the sessions in each cluster.

Step 4: If the new centres are different from the previous ones, repeat steps 2, 3 and 4. Otherwise terminate the algorithm.

So K-means starts with initial partition of K clusters and assign patterns to these clusters in order to reduce or decrease the squared error.

The SOM algorithm used in our approach is summarized as follows.

Input: Sessions formed using the concept hierarchy and link in sequence where each URL is assigned with a distinctive or sole index.

Output: Clusters representing the sessions with similar behaviour.

Step 1: Choose random values for the session weight matrix.

Step 2: Select an input vector sample with a definite possibility. And also there is a set of Web pages along with a set of user transactions.

Step 3: For each session, should calculate the Euclidean distance between the input vector and the weight vector. Winning neuron will be the index value of the weight vector having minimum Euclidean distance with the input vector.

Step 4: Adjust the weight vectors of all neurons by using the update formula.

Step5: Continue with step 2 until no obvious changes in the feature chart are observed.

Result and Discussion

Results for SOM algorithm

The result of Self Organizing Map clustering algorithm on Web usage data is shown in Fig.2 and Fig 3. The results show how sessions are separated into various clusters. There are total 2398 sessions after pre-processing of raw log data and these are categorized into various clusters according to algorithm. As per the SOM algorithm some clusters are dense where as others contain very few or even null. This can be clearly understood by the graph. In line chart below numbers of sessions are plotted against each cluster. And in pie chart whole arc length is divided into number of clusters and each cluster shows the percentage of sessions in each cluster.

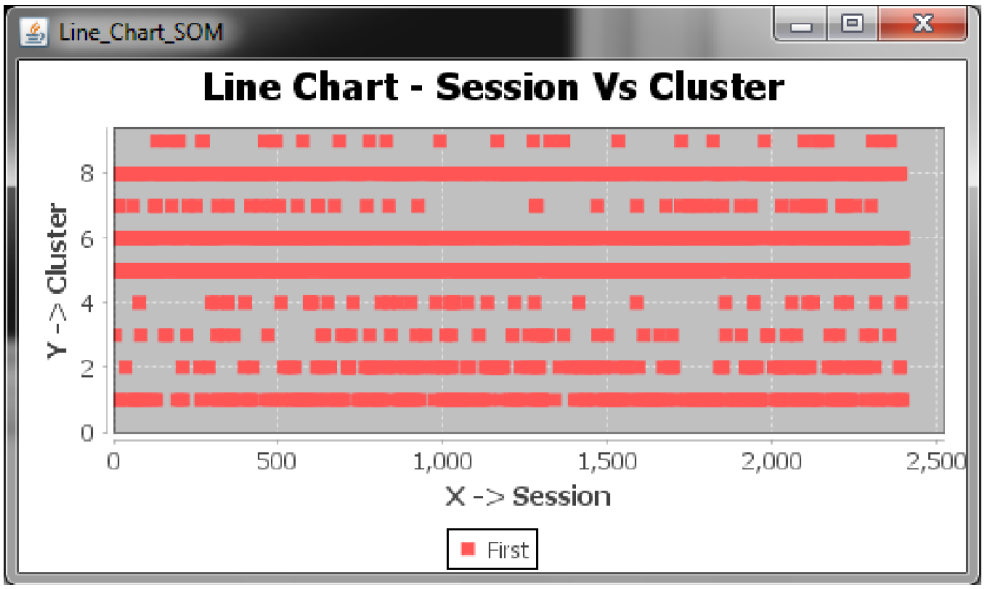

Fig. 2: Line chart for SOM

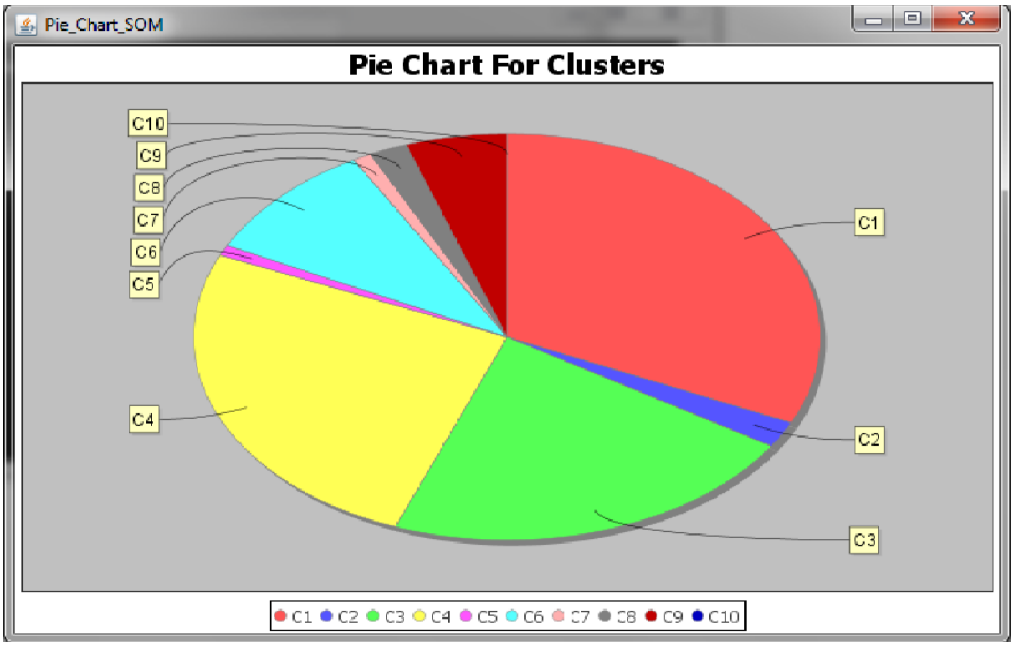

Fig. 3: Pie chart for SOM

Results for K-means algorithm

The results for K-means algorithms are shown in Fig. 4 and Fig 5. According to K-means algorithm the number of clusters is given initially and by taking distance matrix as inputs it groups the data into different clusters. The formed sessions and clusters are represented using line chart and pie chart. The line chart represents sessions in particular cluster. And pie chart shows the percentage sessions in each cluster.

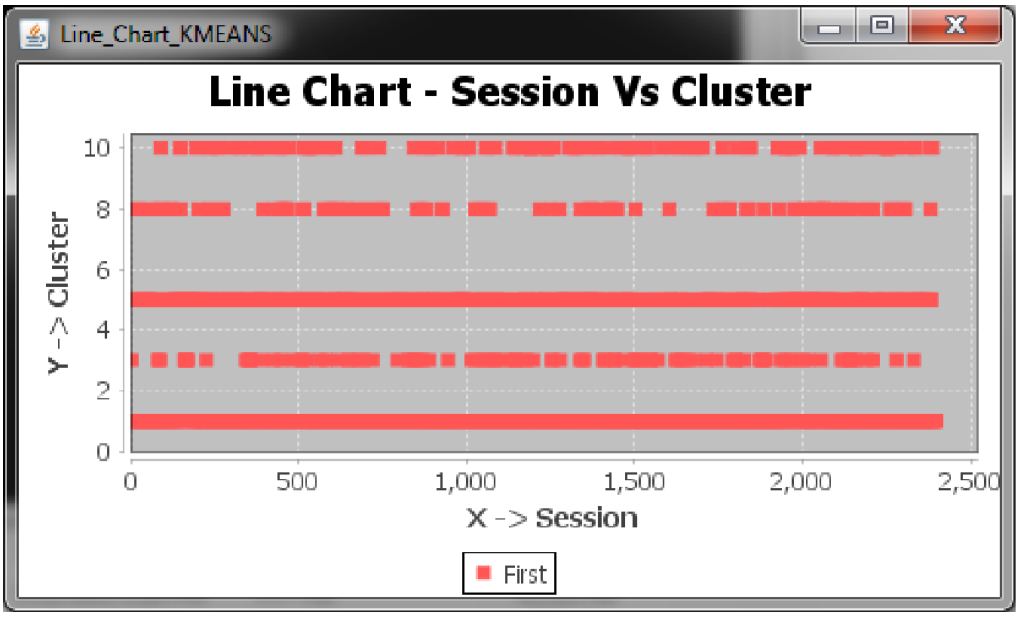

Fig.4: Line chart for K-means

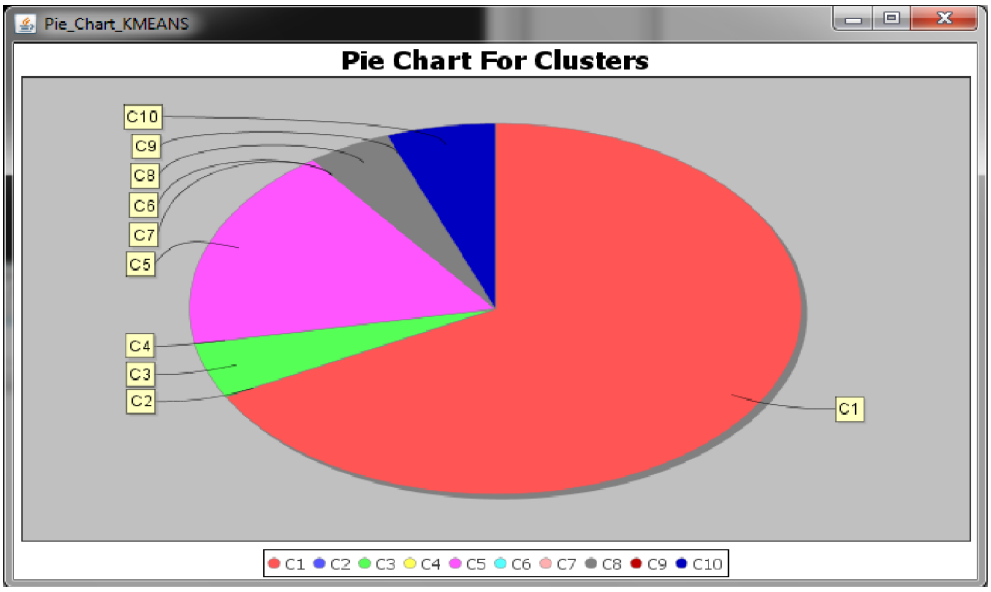

Fig. 5: Pie chart for K-means

Comparison between both algorithms

The graph given in fig. 6 shows the comparison of SOM and K-means clustering algorithms when applied on Web usage navigation data. The red line indicates SOM algorithm and blue line indicates K-means algorithm. Comparison results show that the sessions and the total number of sessions in clusters formed by using SOM algorithm may be different from the sessions and the total number of sessions in clusters formed by K-means algorithm. The clusters contain different number of sessions and even some of the clusters are null. The cluster formation by both algorithms mainly varies in containing set of sessions. Each cluster in both algorithms have corresponding sessions thus varies in number of sessions. According to K-means algorithm cluster one has 1600 sessions out of 2398 total sessions, whereas same cluster has 800 sessions in case of SOM algorithm. Similarly cluster four has zero sessions in K-means and 600 in case of SOM, this is because distance between centers in K-means for cluster 4 is minimum. Hence the cluster formation in SOM is different from K-means algorithm for Web usage data.

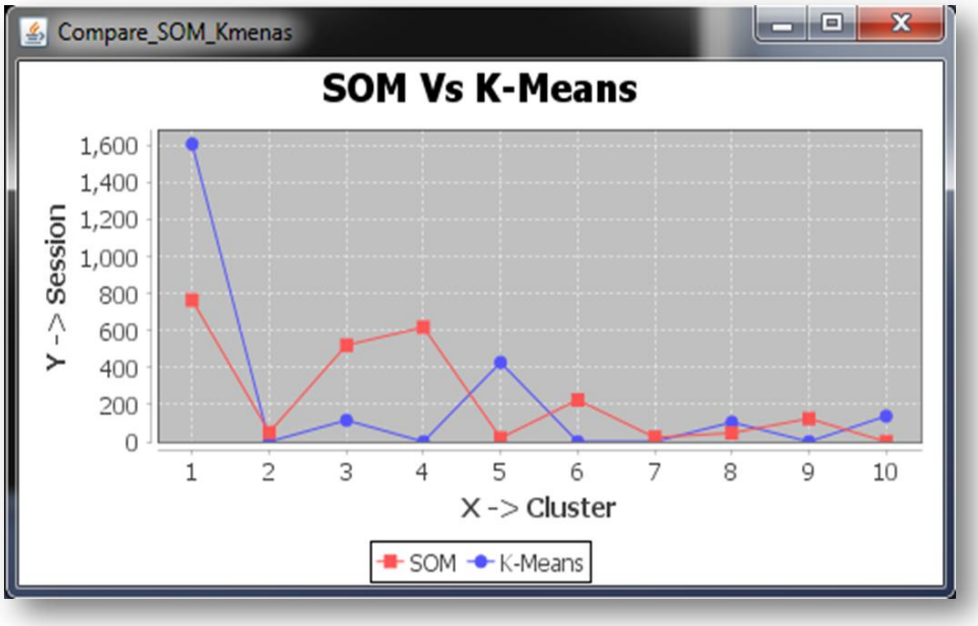

Fig. 6: Comparison of SOM and K-means

Conclusion

In this paper, we have used SOM and K-means algorithms to cluster and visualize Web usage data. Results show that SOM algorithm works better than K-means for Web navigational data. As K-means is very sensitive to initial values, clustering may not be stable. SOM provides superior clustering quality than K-means. In future work, we will combine SOM and K-means algorithm to attain more accurate and appropriate clusters from Web usage data.

ACKNOWLEDGEMENT

The work reported in this paper is supported by the college through the TECHNICAL EDUCATION QUALITY IMPROVEMENT PROGRAMME [TEQIP-II] of the MHRD, Government of India.

References

Paweł Weichbroth, Mieczysław Owoc and Michał Pleszkun (2012). Web User Navigation Patterns Discovery from WWW Server Log Files, Computer Science and Information Systems (FedCSIS), Federated Conference, 9-12 Sept. 2012.

Nawal Seal, Abdelaziz Marzak and Hicham Behja (2013). Web Usage Mining, International Journal of Computer Science issues, Volume 10, Issue 2, No 2, March 2013.

Laura Hollink, Peter Mika and Roi Blanco (2013). Web Usage Mining with Semantic Analysis, International World Wide Web conference, Republic and Canton of Geneva, Switzerland, 13 May 2013.

R.M. Suresh and R. Padmajavalli (2006). An Overview of Data Pre-processing in Data and Web Usage Mining, Digital Information Management, First International Conference, December 2006.

Kate A. Smith and Alan Ng (2003). Web page clustering using a Self-Organizing Map of user navigation patterns, Decision Support Systems, No. 35, pp 245 - 256.

T. Vijaya Kumar and Dr. H. S. Guruprasad (2015). Clustering of Web Usage Data using Hybrid K-means and PACT Algorithms, BIJIT - BVICAM's International Journal of Information Technology BIJIT, July - December 2015, Vol. 7, No. 2; ISSN 0973 - 5658.

Shi Na, Liu Xumin and Guan yong (2010). Research on k-means Clustering Algorithm, Third International Symposium on Intelligent Information Technology and Security Informatics on 978-0-7695-4020-7/10 2010 IEEE.

Yan Zhu, Jian Yu and Caiyan Jia (2009). Initializing K-means Clustering Using Affinity Propagation, Hybrid Intelligent Systems, 2009. HIS '09. Ninth International Conference, Volume 1, 12-14 August 2009.

Anil K. Jain (2010). Data clustering: 50 years beyond K-means, Pattern Recognition Letters 31 (2010) 651-666.

Christos Boutsidis, Anastasios Zouzias, Michael W. Mahoney and Petros Drineas (2015). Randomized Dimensionality Reduction for k-Means Clustering, IEEE Transactions on Information Theory, Vol. 61, No. 2, February 2015.

Juanying Xie and Shuai Jiang (2010). A simple and fast algorithm for global K-means clustering, ETCS 2010, Second International workshop, 6-7 March 2010.

Jiyang Chen, Tong Zheng, William Thorne, Osmar R. Zaiane and Randy Goebel (2007). Visual Data Mining of Web Navigational Data, Information Visualization 2007, International Conference, 4-6 July 2007.

T. Vijaya Kumar and Dr. H. S. Guruprasad (2012). Clustering Web Usage Data using Concept hierarchy and Self Organizing Maps, International Journal of Computer Applications (0975-8887), Volume 56, No.18, October 2012, www.ijcaonline.org.

Esin Saka and Olfa Nasraoui (2008). Simultaneous Clustering and Visualization of Web Usage Data using Swarm-based Intelligence, 20th IEEE International Conference on Tools with Artificial Intelligence, Dayton, OH, 3-5 Dec 2008, pp 539-546, DOI: 10.1109/ICTAI.2008.100.