Clustering of Web Usage Data using Hybrid K-means and PACT Algorithms

T. Vijaya Kumar[1] and **H. S. Guruprasad**[2]

Submitted in May 2014; Accepted in March, 2015

Abstract-Clustering is an exploratory technique that structures the data items into groups based on their similarity or relativeness. Clustering is used in Web usage scenario to form clusters of users showing same behaviour and clusters of pages with similar or related information. The Clustering of users results in the establishment of groups of users with related browsing patterns. The most popular technique to find the clusters is the K-means clustering algorithm. This paper presents a technique to improve the Web session's cluster quality using Subtractive clustering algorithm. In this paper, the Web Session clusters are obtained by using K-means algorithm initialized by subtractive clustering algorithm. The clusters formed are analysed using Profile Aggregation Based on Clustering of Transactions [PACT] algorithm. Web navigational data of the users accessing theWebsitehttp://www.enggresources.com in combined log format taken for a time window of 25 days is used as the raw data for the overall study and analysis.

Index Terms - K-means, Vector matrix, Subtractive clustering, and PACT algorithm.

1.0 INTRODUCTION

World Wide Web has grown into a very powerful and interactive media for the communication of information. Different users geographically located at different places need to access the dissimilar data types efficiently. The navigations of users with web sites generate a huge repository called Web access log file which can be analysed to discover the navigational patterns of the users. The analysis of Web access log file is termed as Web Usage Data mining. Similar to every data mining technique, Web Usage Data mining comprises of three main tasks such as web access logpre-processing, discovery of clusters followed by analysis of discovered clusters. In the pre-processing phase, the unwanted data that are not required for the next phase are removed followed by separation of users and sessions. Cluster discovery phase finds groups of similar pages or users with the similar behavioural patterns. The disseminated information on Web, results in a huge number of links for a search query. There is a need to properly organize these search results. Some search engines cluster these results and present them in an improved manner to the user. The uninterested patterns are filtered out from the user clusters and page clusters in the Cluster analysis phase. The main goal of the Clustering technique is to group together set of items which are similar to each other into the same cluster and dissimilar objects into different clusters. K-means algorithm is used by many researchers to form clusters. K-means is an algorithm with no predefined cluster centroids and non– deterministic in nature. Web access log data mining is a process of drawing out valuable information from Web access log file. The main objective of Web usage data mining is to collect the data, prototype the data as a model to represent the data, analyse the formulated model and visualize the navigational patterns of users. In this paper we have suggested an approach to obtain and analyse the clusters using hybrid K-means and PACT algorithms. First in the pre-processing phase, Server log file is given as the input and the sessions are constructed using conceptual dependency between pages and Web site structure link information which is considered as Web site graph. These identified sessions are represented using an intermediate representation called page-view matrix. Then in the Cluster discovery phase the session clusters are obtained from page-view matrix by using K-means algorithm initialized by subtractive clustering algorithm. Then these found clusters are analysed using PACT algorithm. A brief description about the review of literature is presented in Section 2. The overall architecture of K-means algorithm initialized by subtractive clustering algorithm and the details of PACT algorithm is given in Section 3. The clusters are formed as results and analysed in section 4. Conclusion for our work is briefed in section 5.

[1]*Research Scholar, Department of IS&E, BMSCE, Bangalore, Karnataka, India.*
Corresponding author.
[2]*Professor and Head, Department of CS&E, BMSCE, Bangalore, Karnataka, India.*
E-mail:[1]*vijaykrte@gmail.com and* [2]*hs_gurup@yahoo.com*

2. 0 LITERATURE REVIEW

Recently the application of data mining and artificial neural network techniques to Web log data has fascinated many researchers and they have contributed numerous procedures, tools for Web Usage mining to analyse Web navigation data. Numeral methods have been used to create models of Web navigational data using data mining and artificial neural network approaches. Various Models have been designed based on clustering algorithms, classification techniques, sequential analysis, and Markov models for discovering the knowledge from Web access log data. Web usage data clustering is the process of grouping Web users or Web sessions into clusters so

that users exhibiting the similar navigation behaviour in the same group and dissimilar navigational behaviour in different groups. K-means is deliberated as one of the main algorithms extensively used in clustering. The main advantage of the K-means algorithm is its speed and efficiency compared to other clustering algorithms. Some major drawbacks of K-means algorithm are the number of required cluster centroids must be defined before applying clustering and the random choice of initial cluster centroids. The output of the Clustering technique depends on the random choice of original cluster centroids and different runs may produce different results. In [1] the drawbacks of the standard K-means algorithm, such as the need to compute the distance from each data items to all cluster centroids, is eliminated by introducing two simple data structures. One data structure is used to hold the label of cluster and the other data structure is used to store the distance from every data item to the nearest cluster obtained in each step that can be used to find the distance in the next step. The main downside of K-means is to decide the number of clusters and initializing the centroids for the first iteration. Bashar Al-Shboul et al. have proposed an algorithm which uses genetic algorithms to initialize K-means algorithm [2]. Adaptive Resonance Theory 2 (ART2) neural network is combined with genetic K-means algorithm (GKA) to design a procedure which finds the solution for e- commerce navigational paths [3]. This technique is compared with ART2 followed by K-means and found to be better. The details of clustering algorithms and useful research directions in clustering such as semi-supervised clustering, simultaneous feature selection during data clustering, etc. are provided in [4].Fuzzy clustering [5], also known as soft clustering groups data elements that can be in more than one cluster, and a membership value is associated with each element which will result in the formation of overlapping clusters. The Fuzzy c-means algorithm is initialized by using Subtractive clustering method and experimental results showed that the modified algorithm can decrease the time complexity by reducing the number of iterations, and results in more stable and higher precision classification [6].In [7], an extension for the subtractive clustering algorithm is presented by computing the data point mountain vale. Rather than using conventional method, a kernel – induced distance measure is used in the approach. A model for cluster similar sessions by grouping the similarity matrix using Agglomerative Clustering method is presented in [8]. The session similarity is found by aligning sequences using dynamic programming. A technique for mining Web usage profiles based on subtractive clustering that scales to huge datasets is proposed in [9]. Unlike the clustering based on user description of any input parameter, they have searched in the cluster space for the finest clustering of the given Web access data. Experimental results show that the approach mines the anticipated user profiles much faster than present techniques. A new framework has been suggested in [10] using genetic algorithm and K-means clustering algorithm to improve the cluster quality of Web sessions. Costantinos Etanalyse [11] have built a model for predicting Web page by considering Web access data and Web content with weighted suffix trees. A similarity matrix is considered in the pre-processing procedure by considering the local and global sequence alignment. They have utilized the page content to enhance the proposed scheme. Fuzzy ART neural network is used to enhance the performance of the K-means in [12]. Fuzzy ART neural network technique is used to generate an initial seed value and K-means is applied as the finishing clustering algorithm. Dempster-Shafer's theory which uses evidence or beliefs from dissimilar sources is used to group users into different clusters and generate common user summaries [13]. In [14], Esin Saka et al. have presented a scheme by combining Spherical K-means algorithm and flock of agent based FClust algorithm. Spherical K-means algorithm is mainly used for clustering sparse and high dimensional data. FClust is mainly applicable for representing high dimensional data in a visualization plane. In [15],BamshadMobahser et al. have obtained aggregate usage profiles from the discovered pattern to provide effective recommendation systems for real time Web personalization systems. In [16], Parul Gupta et al. have presented a clustering technique which forms clusters from the set of documents. Every document is assigned with an identifier, so that closer document identifiers are assigned to similar documents. They have proposed an improvement for this clustering algorithm to form super clusters from mega clusters which are formed using similar clusters in a hierarchical clustering process. Their work describes the search process optimization. In [17], Naveen Aggarwal et al. have discussed on the problem of bridging the "semantic gap" between a user's need for meaningful retrieval and the current technology for computational analysis and description of the media content for Integrated Multimedia Repositories. A conceptual framework for agent-based Service Oriented Architecture (SOA) is proposed in [18], which is designed to integrate Service Oriented Architecture with the agent technology & other tactical technologies. In [19], Anil Kumar Pandey et al. have presented mutually exclusive Maximal Frequent Item set discovery based K-Means approach for finding expertise in chosen area of research. Kate A Smith [20], developed LOGSOM to represent Web pages as a two dimensional map using well known Kohonen's Self Organizing Map (SOM). The Web pages are grouped based on the interest of the Web users rather than the content of the Web page. They have considered a transaction group consisting of 235 URLs and treated them as a 235-dimensional vector as input and clustered into K = 9 clusters using K-means algorithm. For SOM output they have considered a 16 X 16 map of 256 nodes.

3.0 SYSTEM DESIGN

The main objective of the proposed system is to cluster the Web usage data using hybrid clustering and analyse the clusters using PACT algorithm. Fig. 1 depicts the overall architecture of the proposed model. We have considered the Webusage navigational data taken for a time window of 25 days of the Website ***http://www.enggresources.com*** for experimental study and discussion. In the Pre-processing Phase Data cleaning, Users Identification and Session Identification are considered

to obtain distinct users and sessions. In Data cleaning the raw server log file is cleaned and only relevant data is taken for further cluster discovery and analysis. Combinations of IP address & user agent are used to identify distinct users. In the next phase Session construction is done based on the time heuristic and navigation approach along with concept hierarchy. Session identification considers all pages accessed by single user and splits all pages into sessions. A sequence of requests made by a single user with a unique IP address on a particular Web domain for a pre-defined period of time is considered as a session. There are several approaches to construct sessions. In time heuristic, if the time spent on a page exceeds a certain threshold, or if the time between two page requests go beyond a threshold time limit then it is assumed

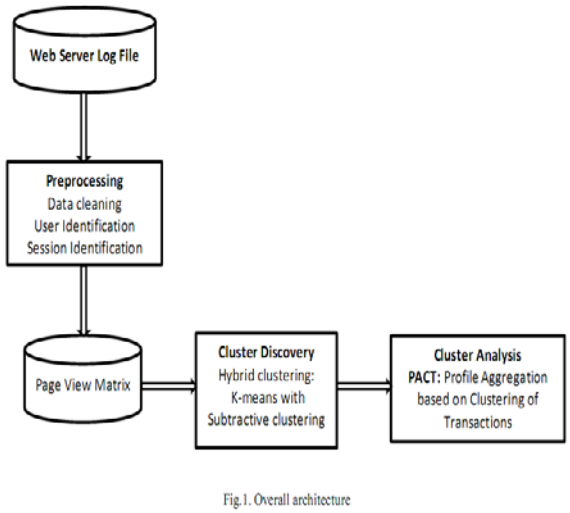

Fig.1. Overall architecture

Figure 1: Overall architecture

that a new session has been created. We have used concept switching and navigation approach with timeout as a criterion for creating user sessions [21]. Then these sessions are represented as click stream matrix. Click stream matrix can be formed by placing the session v/s sequence of pages visited in the respective session. Click stream matrix describes relationship between Web pages and sessions. To form this matrix, first we need to index each unique entry in log file and then form the matrix by placing the sequence of page visited against each session. Click stream matrix is then converted into numerical format called page view matrix. Page view matrix is constructed by building a page set of size n as pages$\{p_1, p_2, p_3, \dots\dots\dots, p_n\}$and user session set of size m as $\{ s_1, s_2, s_3, \dots\dots\dots, s_m\}$ and corresponding entry in the matrix is considered as weight for the page, it can be calculated by number of hits to the page multiplied by page hit weight which consider has 0.01.

Weight of the page $P_{i,j}$ = {frequency of access to page p_j in sessionss_i} * Page hit weight

In the Cluster discovery phase clusters are obtained by K-means clustering algorithm initialized by subtractive clustering algorithm, which takes page view matrix as input and produces the optimal number of clusters as output. K-means algorithm takes page-view matrix and number of clusters as input and it mark each session with cluster it belongs to. The modified K-means algorithm for the Web usage domain can be summarized as follows.

Step1. Consider the data set in whish sessions are represented as page-view matrix and select K points which characterize initial group centroids.

Step2. Calculate the distance between each session and every centroid and assign the session to the centroid cluster with the minimum distance.

Step3. When all sessions have been assigned to clusters, the centroid positions are calculated again by considering the cluster data points.

Step4. Repeat Steps 2 and 3 until there is no change in the centroid positions.

One of the requirements of K-means algorithm is to specify the number of centroids K before the algorithm is applied. It is difficult to guess the number of centroids for a given data set. We have used Subtractive clustering for approximating the number of centroids and the cluster centres in a dataset. The subtractive clustering technique assumes that each data point can be a promising cluster centre. Any data item which has more data items in its vicinity will have more chance of becoming a cluster centroid than data items which have less data items in its neighbourhood. Based on this principle, the potential value for each data item is computed by the following formula:

$$P_i = \sum_{j=1}^{n} e^{-4\|x_i - x_j\|^2/R_a^2}$$

Where x_i, x_j are data items and R_a is a constant value defining the range of the vicinity. The potential of the remaining data itemsx_i, is then revised by

$$P_i \Rightarrow P_i - P_k^* e^{-4\|x_i - x_k\|^2/R_b^2}$$

whereR_b is a positive constant (R_b>R_a). Thus, the data items near the first cluster centre will have greatly condensed potential value, and therefore will have very less chance of to be getting selected as the next cluster centroid. The constant R_b is the radius defining the vicinity that has a lesser potential value thanR_a. The value of R_b is set to be greater thanR_a to avoid getting nearby cluster centres. This process continues until no new cluster centroid is found. The number of clusters and the cluster centroids along with the page-view matrix is given as the input to K-means algorithm to obtain clusters. Then the PACT algorithm is used to analyse these obtained clusters to produce the aggregate profile for each Web transaction cluster. For each cluster we compute the mean vector. The measurement value for each page-view in the mean vector is the ratio between total page-view weights of all transactions and the total transactions in the cluster. The importance of any page p in a cluster is provided by its mean

vector measurement value. Page-views in the mean vector can be sorted according to these measurement values and lower measurement value page-views can be filtered out to obtain set of page view-value pairs that can be used to characterize the group of users showing similar navigational behaviour as aggregate usage summaries. These summaries can be used by the recommendation engines to provide the recommendation.

We can build the aggregate usage summary pr_{cl}, for any cluster *cl*, as a page-view weight pair by computing the mean vector of *cl* using the following formula.

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

Where,

weight (p, pr_{cl}), of the page p within the aggregate usage summary$pr_{cl}$is given by

$$weight(p, pr_{cl}) = \frac{1}{|cl|} \sum_{s \epsilon cl} w(p, s);$$

| *cl* | is the number of sessions in cluster *cl*

$w(p, s)$is the weight of page p in session vector s

An outline of the Hybrid clustering and PACT algorithms for our system is summarized below.

Input: Sessions constructed from the pre-processing phase. Each requested URL is assigned with a unique number.

Output: Web user clusters and Recommendations

Step1. A set of m sessions are constructed from user transactions consisting of subset of n Web pages $\{p_1, p_2, p_3, \ldots\ldots\ldots, p_n\}$. These sessions are converted into page-view matrix.

Step2. Use subtractive clustering algorithm to approximate the optimal numbers of clusters and cluster centroids.

Step3. Cluster the data items usingK-means algorithm.

Step4. Obtain the recommendations using PACT algorithm.

4. Experimental Design and Results

For experimental study and analysis, we have considered the Web usage navigation data from access log files of the Web site ***http://www.enggresources.com*** collected for a time window of 25 days. Concept based Website graph is constructed as an additional input using concept hierarchy and Web site link information. We have used a tool called Web log Filter to remove fields which are not required for further analysis from access log files such as error records, requests for images and multimedia data. For further processing the important fields like IP address of the user who accessed the Web site, timestamp which represents the data and time of access, user agent details like browser information, request page and the page from the request is made called referrer are retained. User separation is considered as the next step in the pre-processing phase. In user separation, IP address and user agent are used to determine the users. In session construction, we have combined two trivial approaches, such as Time based approach and Navigation based approach along with concept name match approach for identifying user sessions. Then these sessions are represented as click stream matrix and later converted this into page view matrix. In Cluster Discovery phase we have obtained clusters by using K-means algorithm initialized by subtractive clustering algorithm. One of the major problems with K-means is to determine the optimal number of clusters and its centres which is done randomly. Subtractive clustering is used to approximately finding the number of clusters and the cluster centroids in a set of data. The subtractive clustering method assumes that each data item can be considered as cluster centre. A data item with more data items in the vicinity will have a higher chance to become a cluster centroid than data points with fewer data items in the vicinity of the data point. PACT algorithm is used to analyse these obtained clusters. The mean vector for each cluster is computed and the measurement value for each page view in the mean vector is computed by taking the ratio between the total page view weights of all transactions and the total transactions in the cluster. Clusters obtained by using K-means algorithm is plotted in Fig 2.The major drawback with K-means algorithm is determining the optimum number of clusters and its centres to initialize the K-means. The hybrid K-means algorithm uses the subtractive clustering algorithm to initialize the K-Means algorithm by providing the optimal number of clusters and its centres. The potential value for each data item is calculated based on the density of nearby data points. K-means initialized by using subtractive clustering algorithm is termed as hybrid K –means clustering and clusters obtained by using hybrid K-means is plotted in Fig 3.

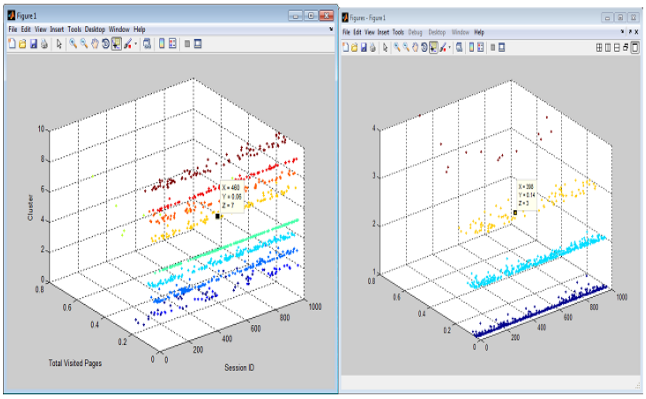

Fig. 2. Clusters obtained using K-means

Fig. 3. Clusters obtained using hybrid K-means

Figure 2: Clusters obtained using K-means

Figure 3: Clusters obtained using hybrid K-means

Fig4 depicts the Cluster-1 user segment interest. The total number of session in the cluster1 is "2286" and threshold considered is "page weight = 0.001". From Fig 4 we can observe that given a new user who shows interest in "Page32", "Page33" and "Page34", this pattern may be used to conclude that the pages "Page28", "Page3" and "Page31"may be recommended to this user. Fig 5 depicts the Cluster-2 user segment interest. The total number of session in the cluster2 is "1812" and threshold considered is "page weight = 0.0002". From Fig 5 we can observe that given a new user who shows interest in "Page28", "Page34" and "Page33", recommendation engine can consider this pattern to conclude that the recommendation engine might recommend any one of the other

pages in the above list to the user based on the order of their weight. Similar results are depicted for Cluster-3 and Cluster-4 user segments with "page weight = 0.0005" in Fig 6 and Fig 7 respectively. From Fig 6 we can observe that given a new user who shows interest in "Page31", "Page32" and "Page33", recommendation engine can consider this pattern to conclude that the recommendation engine might recommend any one of the other pages in the above list to that user based on the order of their weight. From Fig 7 we can observe that given a new user who shows interest in "Page22", "Page23" and"Page27",recommendation engine can consider this pattern

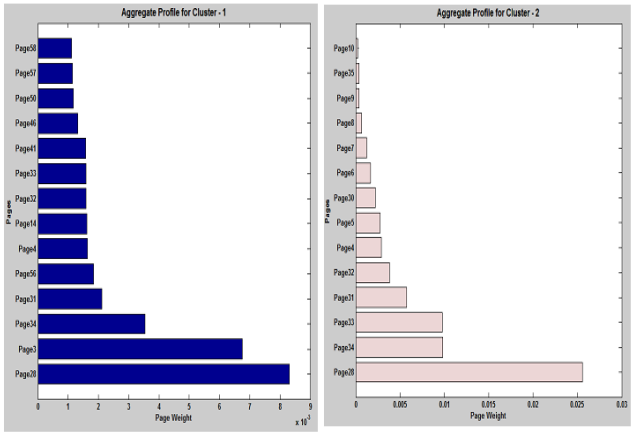

Fig. 4. Cluster-1 user segment interest

Fig. 5. Cluster-2 user segment interest

Figure 4: Cluster-1 user segment interest
Figure 5: Cluster-2 user segment interest

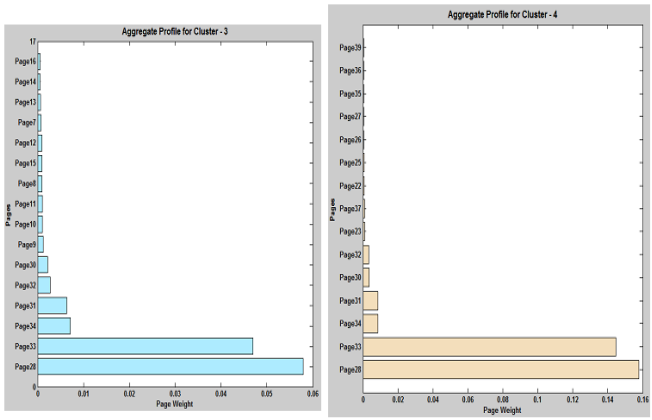

Fig. 6. Cluster-3 user segment interest

Fig. 7. Cluster-4 user segment interest

Figure 6: Cluster-3 user segment interest
Figure 7:Cluster-4 user segment interest

to conclude that the user might belong to this segment and recommendation engine might recommend any one of the other pages in the above list to that user based on the order of their weight.

5.0 CONCLUSIONS

Clustering and Analysis approach for Web usage data using hybrid K-means clustering algorithm and PACT algorithm is presented in our proposed scheme. The sessions for clustering phase are obtained by using conceptual dependency between pages and Website structure link information which is considered as Web site graph. Then these sessions are represented as click stream matrix and later converted this into page view matrix. Then clusters are formed by using K-means clustering algorithm initialized by subtractive clustering algorithm. Then clusters are analysed by using PACT algorithm to give the recommendations. As a future work, we can improve this as a recommendation engine to compare current request with navigation pattern in each cluster and come up with the recommendations.

REFERENCES

[1]. Shi Na, Guan yong, and Liu Xumin, "Research on K-means clustering algorithm" Third International Symposium on Intelligent Information Technology and Security Informatics, 2010 IEEE.

[2]. Bashar Al-Shboul and Sung-HyonMyaeng, "Initializing K-means using Genetic Algorithms" World Academy of Science, Engineering and Technology, 54 2009.

[3]. R. J. Kuo, J. L. Liao and C. Tu, "Integration of ART2 neural network and genetic K-means algorithm for analysing Web browsing paths in Electronic commerce", Decision Support Systems 40 (2005) 355-374, www.sciencedirect.com.

[4]. Anil K. Jain, "Data clustering: 50 years beyondK-means", Pattern Recognition Letters 31 (2010) 651-666.Journalhomepage www.elsevier.com/locate/patrec.

[5]. Zahid Ansari, A Vinay Babu,WaseemAhamed and Mohammed FazleAzeem, "A Fuzzy set theoretic approach to discover user sessions from Web navigational data", Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE.

[6]. Qing yang, Dongxu Zhang, and Feng Tian, "Aninitialization method for Fuzzy c –means algorithm usingsubtractive clustering" 2010 Third international conference on Intelligent Networks and Intelligent Systems.

[7]. Dae-Won Kim, KiYoung Lee, Doheon Lee and Kwang H. Lee, "A Kernel based subtracting clustering algorithm",Pattern Recognition Letters, 26 (2005) 879-891, www.elseveir.com/locate/patree.

[8]. Dr. K. Duraiswamy, and V. ValliMayil, "Similarity Matrix Based Session Clustering by Sequence Alignment Using Dynamic Programming",Computer and Information Science Vol. 1, N0. 3, August 2008, www.ccsenet.org/journal.html

[9]. Bhushan Shankar Suryavanshi, NematollaahShiri, and Sudhir P. Mudur, "An Efficient Technique for Mining Usage profiles Using Relational Fuzzy Subtractive Clustering", Proceedings of 2005 International

Workshop on Challenges in Web Information Retrieval and Integration (WIRI'05)

[10]. N. Sujatha, and K. Iyakutty, "Refinement of Web usage data clustering form K-means with genetic algorithm", European Journal of Scientific Research ISSN 1450-216X Volume 42 Number 3 (2010), Pages.478-490.

[11]. Costantinos Dimopoulos, Christos Makris, YannisPanagis, EvangelosTheodoridis and Athanasios Tsakalidis, "A Web page usage prediction scheme using sequence indexing and Clustering techniques", Data and Knowledge Engineering 69 (2010) 371-382, www.elsevier.com/locate/datak

[12]. Sungjune Park, Nallan C. Suresh, and Bong KeunJeong, "Sequence based clustering for Web usage mining: A new experimental framework and ANN- enhanced K-means algorithm", Elsevier Data and Knowledge Engineering 65 (2008) 512 – 543.

[13]. YunjuanXie and Vir V. Phoha, "Web user clustering from Access log using Belief Function", K-CAP'01, October 22-23, 2001, Victoria, British Columbia, Canada.

[14]. Esin Saka, and OlfaNasraoui, "Simultaneous Clustering and Visualization of Web Usage Data using Swarm-based Intelligence", 20th IEEE International Conference on Tools with Artificial Intelligence.

[15]. BamshadMobasher, Honghua Dai, Tao Luo, and Miki Nakaguva, "Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization", Data mining and Knowledge Discovery 6, 61-82, 2002.

[16]. Parul Gupta and A.K. Sharma, "A Framework for Hierarchical Clustering Based Indexing in Search Engines", BIJIT - BVICAM's International Journal of Information Technology, July – December, 2011; Vol. 3 No. 2; ISSN 0973 – 5658.

[17]. Naveen Aggarwal, Dr.Nupur Prakash and Dr. Sanjeev Sofat, "Mining Techniques for Integrated Multimedia Repositories: A Review", BIJIT - BVICAM's International Journal of Information Technology, January – June, 2009; Vol. 1 No. 1; ISSN 0973 – 5658.

[18]. O. P. Rishi, "Service Oriented Architecture for Business Dynamics: An Agent Based Business Modelling Approach", BIJIT - BVICAM's International Journal of Information Technology, July – December, 2009; Vol. 1 No. 2; ISSN 0973 – 5658.

[19]. Anil Kumar Pandey and T. Jaya Lakshmi, "Web Document Clustering for Finding Expertise in Research Area", BIJIT - BVICAM's International Journal of Information Technology, July – December, 2009; Vol. 1 No. 2; ISSN 0973 – 5658.

[20]. Kate A Smith and Alan Ng, "Web page clustering using a Self-Organizing Map of user navigation patterns", Decision Support Systems 35(2003) 245-256, www.elsevier.com/locate/dsw.

[21]. T. Vijaya Kumar, Dr. H.S. Guruprasad, Bharath Kumar K.M, Irfan Baig and KiranBabu S, "A New Web Usage Mining approach for Website recommendations using Concept hierarchy and Website Graph", IJCEE, ISSN: 1793-8198 (Online Version);1793-8163(print version).