Clustering Of Web Usage Data Using Chameleon Algorithm

T.Vijaya Kumar[1], Dr. H.S.Guruprasad[2]

[1]Research Scholar, Department of IS & E, BMSCE, Bangalore, Karnataka, India.

[2]Professor and Head, Department of CS & E, BMSCE, Bangalore, Karnataka, India.

ABSTRACT: Clustering is a discovery process in data mining which groups set of data items, in such a way that maximizes the similarity within clusters and minimizes the similarity between two different clusters. These discovered clusters depict the characteristics of the underlying data distribution. Clustering is useful in characterizing customer groups based on purchasing patterns, categorizing web documents that have similar functionality. In this work, graph-based clustering is proposed to form clusters based on web usage patterns. First sessions are constructed using time oriented approach. Based on the constructed sessions and page requests, adjacency matrix is created. Then data points are generated using adjacency matrix as input. Chameleon clustering algorithm takes data points as input and forms clusters. Chameleon uses a two phase approach to find the clusters. In the first phase, it uses a graph partitioning algorithm to cluster the data items into several relatively small sub-clusters. In the second phase, it uses an algorithm to form genuine clusters by repeatedly combining these sub clusters. Then these clusters are plotted on a plane using MATLAB where different clusters are distinguished by distinct colours and distinct symbols. In this paper, the server log files of the Website www.enggresources.com is considered for overall study and analysis.

KEYWORDS: Web usage mining, K Nearest Neighbor graph, Chameleon algorithm.

I. INTRODUCTION

Clustering is a process of grouping a given set of unlabelled patterns into a number of clusters such that similar patterns are assigned to one cluster. Each pattern can be represented by a vector having many parameters. Most of traditional clustering algorithms use either interconnectivity or closeness to merge two clusters. The key feature of Chameleon algorithm is it considers both interconnectivity and closeness in identifying the most similar pair of clusters. Chameleon also uses a novel approach to model the degree of interconnectivity and closeness between each pair of clusters. The complete detail of Chameleon hierarchical clustering algorithm is given in [1]. Two clusters are merged only if the inter-connectivity and closeness between two clusters are high relative to the internal inter-connectivity of the clusters and closeness of items within the clusters. Chameleon is a clustering algorithm based on K Nearest Neighbour (KNN) graph. In KNN graph, two data points far away from each other may reside in two disconnected sub-graphs and even if two points are in the same sub-graph, they may be separated in the following partition sets. Chameleon first constructs a KNN graph of the data set, and applies hMETIS to partition the graph in to clusters. It then defines the relative interconnectivity between two clusters and relative closeness within a cluster, and repeatedly merges clusters until the specified number of clusters is achieved. Chameleon uses a two phase approach to find the clusters. In the first phase, it uses a graph partitioning algorithm to cluster the data items into several relatively small sub-clusters. In the second phase, it uses an algorithm to form genuine clusters by repeatedly combining these sub clusters. The rest of the paper is organized as follows. Section 2 gives a brief description about the related work. Section 3 depicts the proposed model. The experimental results of the proposed approach are given in section 4 and section 5 concludes the discussion.

II. RELATED WORK

Web usage mining helps in modelling access behaviour of web users by extracting useful information from web server log files. Web usage mining refers to the automatic discovery and analysis of patterns in clickstream and associated data collected or generated as a result of user interaction with web resources on one or more web sites. Several researchers are working on Web Usage mining and have contributed various methodologies, tools for Web Usage mining. Models based on association rules, clustering algorithms, sequential analysis and Markov models have

been used for discovering the knowledge from Web usage data. All these models are predominantly based on usage information from Web usage data alone. Significant improvement can be achieved by making use of domain knowledge, which is usually available from domain experts, content providers, and Web designers. Cooley et al. in [2, 3], covered Web usage mining process & various steps involved in it. It serves as the primary thesis to understand fundamentals of Web usage mining. Norwati Mustapha et al. [4], have proposed a model for mining user's navigation pattern based on Expectation modelling algorithm and used it for finding maximum likelihood estimates of parameters in probabilistic models. In [5], Rajhans Mishra et al. have adopted the similarity upper approximation based clustering of web logs using various similarity metrics. In [6], K.Santhisree et al. have presented a technique to cluster web transactions based on the set similarity measures from web log data which identifies the behaviour of the user's page visits and order of occurrence of visits. They have formed the Web data Clusters using the Similarity Upper Approximations. In [7], 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 in the original feature domain and the flock-based algorithm which iteratively adjusts the position and speed of dynamic flock's agents on a visualization plane. The hybrid algorithm decreases the complexity of FClust from quadratic to linear with further improvements in the cluster quality. In [8], S Park et al. have investigated the use of fuzzy ART neural network, to enhance the performance of the K-Means algorithm. A major limitation with fuzzy ART is the unrestricted growth of clusters. Fuzzy ART networks sometimes produce numerous clusters each with only a small number of members. The authors have used fuzzy ART as an initial seed generator and K-Means as the final clustering algorithm. In [9], Santosh K et al. have developed a clustering algorithm that groups users according to their Web access patterns. The algorithm is based on the ART1 version of adaptive resonance theory. ART1 offers an unsupervised clustering approach that adapts to the changes in user's access patterns over time without losing earlier information. It applies specifically to binary vectors. They have compared their algorithm's performance with the traditional K-Means clustering algorithm and showed that the ART1-based technique performed better in terms of intra cluster distances. In [10], Antonia S et al. have developed a variant of the classical SOM called Growing Hierarchical SOM (GHSOM). They have suggested a new visualization technique for the patterns in the hierarchical structure. In [11], Kate A. Smith et al. have developed LOGSOM, a system that utilizes Kohonen's Self Organizing Map to organize Web page in to two dimensional maps. The organization of the Web pages is based on the user's navigation behaviour rather than the content of the Web pages. In [12], T.Vijaya Kumar et al. have proposed a framework for finding useful information from Web Usage Data that uses Self Organizing Maps (SOM). Sessions are constructed using the concept hierarchy and the link information. Then they have used SOM to form the cluster. In [13], K.Santhisree et al. have presented a technique to cluster web transactions based using rough set DBSCAN clustering algorithm. They have considered datasets of msnbc.com and found user access patterns, the order of visits of the hyperlinks of each user and inter cluster similarity among the clusters. In [14], Pradeep Kumar et al. have presented a new indiscernibility-based rough agglomerative hierarchical clustering algorithm for sequential data. In their approach, similarity upper approximation is used to form initial clusters and the concept of constrained similarity approximation is used to form subsequent clusters. They have compared their approach with the traditional hierarchical clustering algorithm on msnbc.com web navigation data set. In [15], Mehrbad Jalali et al. have proposed a novel based approach on Longest Common Subsequence algorithm for classifying user navigation patterns for predicting user's future requests. In [16], J. Vellingiri et al. have proposed a novel clustering algorithm called Fuzzy Possibilistic C-Means algorithm, which integrates extended partition entropy and inter class resemblance to predict the user behaviour. In [17], K. Suresh et al. have used improved Fuzzy C-Means (FCM) algorithm to cluster the Web Usage data from msnbc.com. In [18], Subhash K. Shinde et al. have developed architecture for online recommendation in Web Usage Mining System. They have proposed an architecture of Online Recommendation in Web Usage Mining (OLRWMS) for enhancing accuracy of classification by interaction between classifications, evaluation, current user activities and user profiles. In [19], Supriya Kumar De et al. have used soft computing technique such as rough approximation based clustering to cluster web transactions In [20], Costantinos Dimopoulos et al. have considered the problem of web page usage prediction in a web site by modelling user's navigation history and web page content with weighted suffix trees. Each navigation session is described by a weighted suffix tree. Then the similarity matrix is constructed which represents the similarity of each pair (i, j) of sequences. Then the clusters are formed by utilizing the content of the web pages in order to enrich the quality of the clustering. The weighted suffix tree is used to capture the most important navigational experience which works like a prediction model without explicitly storing strings and probabilities of occurrences. They have performed evaluation of the proposed scheme with experiments on various web site log files and web pages. In [21], Chen Wu et al. have presented

Semantic Web Log Model (SWLM) to improve the efficiency and accuracy of web log mining. In [22], H. Hannan Inbarani et al. have proposed an approach for feature selection based on rough set theory for Web Usage mining. In [23], Sebastian A Rios et al. have used concept based approach to add semantics in to the mining process. They have applied the proposed solution to a real web site to produce offline enhancements of contents and structure and compared their approach with four different Web Usage Mining methods. In [24], Yunjuan Xie et al. have presented a novel approach to clustering Web site users into different groups and generating common user profiles using the concept of mass distribution in Dempster-Shafer's theory. In [25], Jinfeng Li et al. have proposed a two stage clustering algorithm, named Chameleon Based on Clustering Feature Tree (CBCFT), which hybridizes the Clustering tree algorithm BIRCH with Chameleon algorithm for customer segmentation.

III. PROPOSED ALGORITHM

The main goal of the proposed system is to find web user clusters from web server log files. The proposed Web Usage Mining System is shown in Fig 1. The WUM system in the proposed approach is partitioned into two modules. In the first module User identification and Session constructions are considered for Data preprocessing phase. In the Second phase Chameleon clustering algorithm is used to form the clusters.
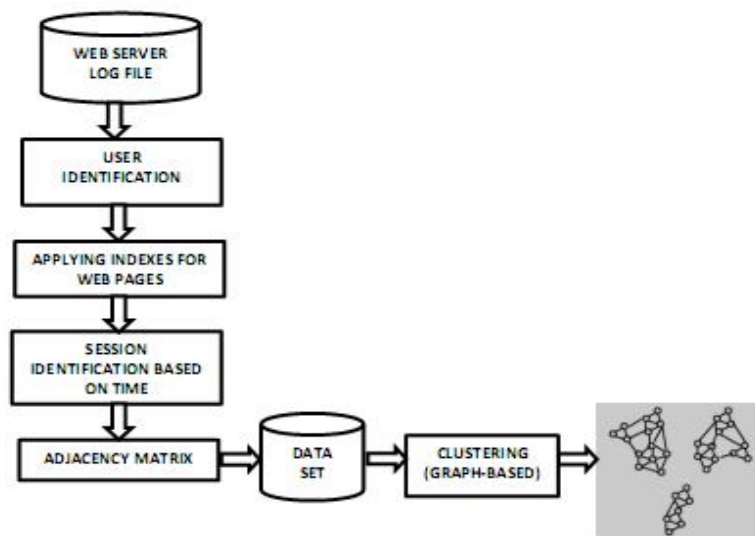

Fig.1. Architectural overview of the system

Web log data is usually diverse and voluminous. This raw log data must be assembled into a consistent, integrated and comprehensive form, in order to be used for pattern discovery. As in most data mining applications, data pre-processing involves removing and filtering redundant and irrelevant data, removing noise, transforming and resolving any inconsistencies, identifying unique users and finally constructing user sessions. There are several ways to identify individual users. The most obvious solution is to assume that each combination of IP address and browser identifies a single user. Then sessions are constructed from the individual user's data. The session is termed as "user activity over a span of time on web site". Here sequence of web page requests made by single users on particular web domain for a specific time period is identified. This is termed as session construction. Pattern discovery is the next stage of web usage mining. In this phase, frequent access patterns are determined from reconstructed sessions. In our work, this phase is divided into 2 phases namely constructing adjacency matrix and identifying clusters. The adjacency matrix is constructed from the sessions considering sessions as rows and pages as columns. Using the adjacency matrix the data points are generated which is considered as input for the clustering phase. There are several types of clustering and

graph-based clustering is considered since it supports all types of data and forms the clusters of different size, shape and different density. Under graph-based clustering Chameleon algorithm is used to merge the clusters. CHAMELEON is an agglomerative hierarchical clustering where the objects are considered as nodes and proximity between nodes are considered as weights. The m by m proximity matrix for m data points can be represented as a dense graph in which each node is connected to all others and the weight of the edge between any pair of nodes reflects their pairwise proximity. For most data sets, objects are highly similar to a small number of objects and weakly similar to most other objects. This property is used to sparsify the proximity matrix by setting many of the low similarity values to 0 before starting the actual clustering process. This sparcification can be performed by removing all links that have a similarity below a specified threshold and by keeping only links to the k nearest neighbours of the point. This approach creates a K Nearest Neighbour graph. Most of the Agglomerative clustering techniques operate by merging the two most similar clusters based on the cluster similarity. The cluster similarity is computed either by using relative closeness or by relative interconnectivity. Chameleon operates by merging the clusters that best preserve the cluster self-similarity with respect to relative interconnectivity and relative closeness [26]. Relative interconnectivity is the absolute interconnectivity of two clusters normalized by the internal connectivity of the clusters. Two clusters are combined if the points in the resulting cluster are almost as strongly connected as points in each of the original clusters. Relative closeness is the absolute closeness of two clusters normalized by the internal closeness of the clusters. Two clusters are combined only if the points in the resulting cluster are almost as close as in each of the original clusters. Chameleon is an agglomerative clustering algorithm which partitions the data using efficient graph partitioning algorithm and combines it with a novel hierarchical clustering algorithm that uses the notion of closeness and interconnectivity together with the local modelling of clusters to form the final clusters. CHAMELEON algorithm consists of two phases, namely Partitioning and Merging. In partition phase, it uses multilevel partition algorithm hMETIS which is built in multilevel partitioning algorithm for hyper graph partition. hMETIS enhances the speed and efficiency of algorithm. In the Merging phase, the sub clusters are merged based on the similarity value which depends on relative interconnectivity and relative closeness. Fig 2 provides an overview of the overall approach used by Chameleon algorithm to obtain the clusters.
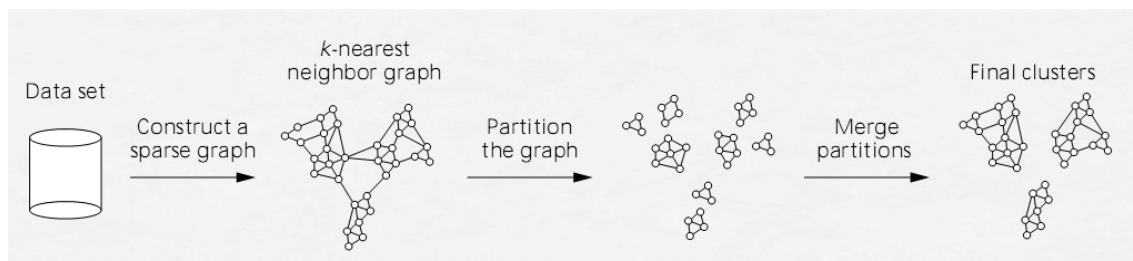

Fig.2. Two – phase approach used in Chameleon algorithm.

IV.   **PSEUDO CODE**

An outline of the proposed system is summarized below.

Step1. Indexing the web page
    Input: User identified log file
    Output: indexed file
    For Each User U
        Add request page to array A
    End for
    For each request in array A
        Remove redundancy in array
    End for

For each request in array A

    Add request page and index value to Hash table

    Write request page and index value to IPindex file

End for

Step2. Session construction

    Input: User identified log file & indexed file

    Output: Session file

    For each User U based on Distinct (ip+browser)

    For each request r of User U

        If (time_of_current_req – time_of_first_req < 30 mins)

          Add this request to current session

        Else

          Create new session & Write session details (ip, time, req) to session file

          curr_req_time=curr_req_time - 30

   End For

   End For

Step3. Generate adjacency matrix and data points.

Step4. Obtaining clusters using Chameleon algorithm

    Input: data points obtained by adjacency matrix

    Output: cluster file

   1: Build a k-nearest neighbour graph of the data points

   2: Partition the graph using multi-level graph partitioning algorithm hMETIS

   3: repeat

   4: Merge the clusters that best preserve the cluster self-similarity with respect to Relative interconnectivity and

    Relative closeness

   5: until No more clusters can be merged

V. SIMULATION RESULTS

For experimentation, data from Server log files of the Web site www.enggresources.com collected from 19-07-2009 to 11-08-2009 is considered for the overall study and analysis. Error records, requests for images and multimedia files are removed from Server log file by using a tool called Web log filter. IP address, timestamp, user agent, request and referrer are retained for further processing. User Identification is considered as the next step. In user identification, IP address and user agent are used. That is, a combination of IP address and user agent is used to identify a unique user. In session construction, Time oriented approach is used for identifying user sessions. Session timeout threshold is set as 30 minutes. Each Web page is assigned with unique index. Then sessions are identified based on time out and it is written on to a separate file. Every unique session is also given unique index. Now the file contains session number and pages referred in that particular session. The adjacency matrix is constructed from the sessions considering sessions as rows and pages as columns. Using the adjacency matrix the data points are generated which is considered as input for the clustering phase. Then Chameleon algorithm is used to obtain the clusters. A KNN graph of the data set is constructed and hMETIS is applied to partition the graph in to clusters. Then the relative interconnectivity between two clusters and relative closeness within a cluster, are used to repeatedly merge the clusters until the specified number of clusters is achieved. Mat lab is used to plot the clustering data points on a plane where the different clusters are distinguished by distinct colours and distinct symbols. The output for the data set with ten clusters is shown in Fig 3. Fig 4 shows output for the same data set with six clusters. Apart from Time oriented and Navigation oriented, sessions can also be also obtained by using concept matching approach [27]. Fig 5 shows the output for the clusters obtained from sessions which are constructed by using concept hierarchy and web site graph with ten cluster numbers.
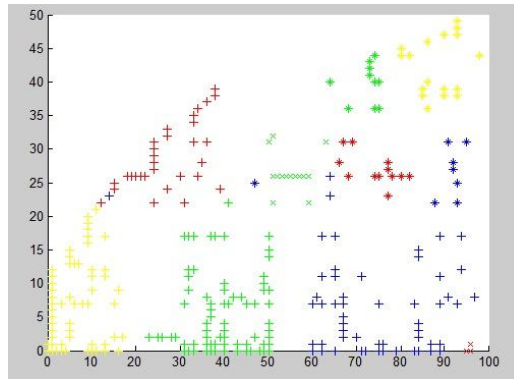
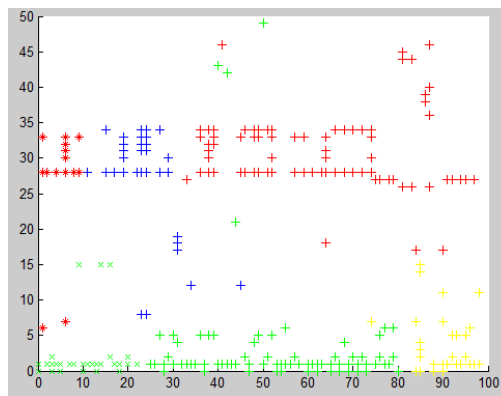Fig.3. Clusters plotted on plane using mat lab(stop cluster number=10)

Fig.4. Clusters plotted on plane using mat lab (stop cluster number=6)
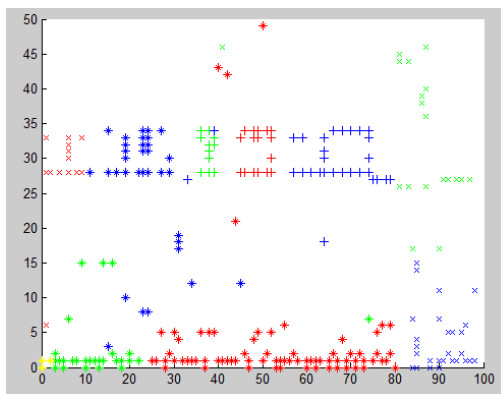

Fig.5. Clusters plotted on plane using mat lab (stop cluster number=10)

International Journal of Innovative Research in Computer and Communication Engineering

VI. CONCLUSION AND FUTURE WORK

Clustering process groups the related data together to form clusters. Chameleon is agglomerative hierarchical clustering algorithm that overcomes the limitations of existing clustering algorithms. The main feature of Chameleon algorithm is that it considers both interconnectivity and closeness in identifying the most similar pair of clusters. In this paper, Chameleon algorithm is used to form the clusters from the data points. First the sessions are constructed using time–oriented approach. Sessions are transformed into adjacency matrix and data points. Then Chameleon algorithm is used to find clusters which are plotted on plane using MATLAB. As a future work rough set theory and semantic information can be used to find relative interconnectivity and relative closeness in order to obtain the most similar pair of clusters.

REFERENCES

1. George Karypis, Eui-Hong (Sam) Han and Vipin Kumar "Chameleon: Hierarchical clustering using dynamic modelling", Data Mining 0018-9162/99©1999 IEEE.
2. R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: information and pattern discovery on the World Wide Web", Ninth IEEE International Conference on Tools with Artificial Intelligence, Newport Beach, CA, USA, 1997, Pages 558-567.
3. J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan, "Web usage mining: discovery and applications of usage patterns from Web data", ACM SIGKDD Explorations Newsletter, Volume 1, Pages 12-23, 2000.
4. Norwati Mustapha, ManijehJalali, and Mehrdad Jalali, "Expectation Maximization Clustering Algorithm for User Modelling in Web Usage Mining Systems", European Journal of Scientific Research ISSN 1450-216X Volume 32 Number.4 (2009), Pages.467-476.
5. Rajhans Mishra and Pradeep Kumar, "Clustering Web Logs Using Similarity Upper Approximation with Different Similarity Measures",International Journal of Machine Learning and Computing, Vol. 2, No. 3, June 2012.
6. K.Santhisree, and Dr.A.Damodaram, "Clustering on Web usage data using Approximations and Set Similarities" 2010 International Journal of Computer Applications (0975 – 8887) Volume 1 – No. 4.
7. Esin Saka, and Olfa Nasraoui, "Simultaneous Clustering and Visualization of Web Usage Data using Swarm-based Intelligence", 20th IEEE International Conference on Tools with Artificial Intelligence.
8. Sungjune Park, Nallan C. Suresh, and Bong Keun Jeong, "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.
9. Santosh K.Rangarajan, Vir V.Phoha, Kiran S.Balagani, Rastko R.Selmic and S.S. Iyengar, "Adaptive Neural Network Clustering of Web Users", IEEE 2004 0018-9162/04.
10. Antonio S, Jose D. Martin, Emilio S, Alberto P, Rafael M and Antonio, "Web mining based on growing hierarchical Self Organizing Maps: Analysis of a real citizen Web portal", Expert Systems with applications 34(2008)2998-2994 www.elsevier.com.
11. Kate A. Smith, Alan Ng, "Web page clustering using a self-organizing map of user navigation patterns", Elsevier Decision Support Systems 35 (2003) 245– 256.
12. T. Vijaya Kumar, Dr. H. S. Guruprasad, "Clustering Web Usage Data using Concept hierarchy and Self Organizing Maps", International Journal of Computer Applications (0975 – 8887) Volume56No.18,October 2012 www.ijcaonline.org.
13. K. Santhisree, A. Damodaram and S. V. Appaji, "An Enhanced Dbscan algorithm to Cluster Web Usage Data using Rough sets and Upper Approximations", International Journal of Computer Science and Communication, Vol. 1, No. 1, January-June2010, pp263-265.
14. Pradeep Kumar, P. Radha Krishna, Raju S. Bapi, and Supriya Kumar De, "Rough clustering of sequential data", Data and Knowledge Engineering 63 (2007) 183-199, www.elsevier.com/locate/datak.
15. Mehrdad Jalai, Norwati Mustapha, Md. Nasir B Sulaiman and Ali Mamat, " A web Usage Mining approach Based on LCS Algorithm in Online Predicting Recommendation Systems", 12th International Conference Information Visualization, 1550-6037/08, DOI 10.1109/iV.2008.40.
16. J. Vellingiri and S. Chenthur Pandian, "Fuzzy Possibilistic C-Means Algorithm for Clustering on Web Usage Mining to Predict the User Behavior", European Journal of Scientific Research ISSN 1450-216X Vol.58 No.2 (2011), pp. 222-230.
17. K. Suresh, R. Madana Mohana and A Rama Mohan Reddy, "Improved FCM algorithm for Clustering on Web Usage Mining", IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 1, January 2011, ISSN (online) 1684-8014, www.IJCSI.org.
18. Subhash K. Shinde and Dr. U.V.Kulkarni, "A new Approach for On Line Recommender System in Web Usage Mining", 2008 International Conference on Advanced Computer Theory and Engineering", 978-0-7695-3489-3/08, DOI 10.1109/ICACTE.2008.72.
19. Supriya Kumar De and P. Radha Krishna, "Clustering web transactions using rough approximation", Fuzzy sets and Systems 148(2004) 131-138, www.elsevier.com/locate/fss.
20. Costantinos Dimopoulo, Christos Makris, Yannis Panagis and Evangelos Theodoridis, " 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.
21. Chen Wu, Jun Dai, Qi Feng Li and Jun Wu Zhu, " A Semantic Clustering Algorithm oriented to Web Log", Proceedings of the Fifth International Conference on Machine Learning and Cybernetics, Dalian, 13-16 August 2006.
22. H. Hannah Inbarani, K. Thangavel and A. Pethalakshmi, " Rough set based Feature Selection for Web Usage Mining", International Conference on Computational Intelligence and Multimedia Applications 2007.
23. Sebastian A. Rios and Juan D. Velasquez, "Semantic web usage mining by a Concept based approach for Off-line Web Site Enhancements", 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.
24. Yunjuan Xie and Vir V. Phoha, "Web user clustering from Access log using Belief Function", K-CAP'01, October 22-23, 2001, Victoria, British Columbia, Canada.
25. Jinfeng Li, Kanliang Wang and Lida Xu, "Chameleon based on Clustering Feature Tree and its application in Customer Segmentation", Ann Oper Res (2009) 168: 225-245, DOI 10.1007/s10479-008-0368-4.
26. Pang-Ning Tan, Vipin Kumar and Michael Steinbach, "Introduction to Data Mining", Dorling Kindersley (India) Pvt. Ltd, Pearson Education, ISBN 978-317-6463-3.
27. T. Vijaya Kumar, Dr. H.S. Guruprasad, Bharath Kumar K.M, Irfan Baig and Kiran Babu S, "A New Web Usage Mining approach for Website recommendations using Concept hierarchy and Website Graph", International Journal of Computer and Electrical Engineering (IJCEE, ISSN: 1793-8198 (Online Version);1793-8163(print version).

BIOGRAPHY

T.Vijaya Kumar is a research scholar at Information Science and Engineering department, BMS College of Engineering, VTU, Belgaum, Karnataka, India. His research interest includes Data Mining, Knowledge Discovery from Web usage data, Theoretical Foundation of Computer Science, Artificial Neural Network and Cloud Computing.

Dr.H.S.Guruprasad is currently working as Professor and Head of Computer Science and Engineering department, BMS College of Engineering, Bangalore, India. He has done his PhD in the area of Network Communication. He has more than twenty two years of teaching experience. He has been awarded with Rashtriya Gaurav award and Best Citizen of India award. His research interest includes Multimedia Communications, Sensor Networks, and Cloud Computing, Algorithms, Android Operating System, Data Mining and Knowledge discovery from Web usage data.