Web Log Clustering Approaches – A Survey

Mrs. G. Sudhamathy
Department of Computer Applications
Velammal College of Engineering & Technology
Madurai, Tamil Nadu 625 009, India
sudhamathi10@hotmail.com

Dr. C. Jothi Venkateswaran
Department of Computer Science
Presidency College
Chennai, Tamil Nadu 600 025, India
jothivenkateswaran@yahoo.co.in

***Abstract*— As more organization rely on the Internet and the World Wide Web to conduct business, the proposed strategies and techniques for market analysis need to be revisited in this context. We therefore present a survey of the most recent work in the field of Web usage mining, focusing on three different approaches towards web logs clustering. Clustering analysis is a widely used data mining algorithm which is a process of partitioning a set of data objects into a number of object clusters, where each data object shares the high similarity with the other objects within the same cluster but is quite dissimilar to objects in other clusters. In this work we discuss three different approaches on web logs clustering, analyze their benefits and drawbacks. We finally conclude on the most efficient algorithm based on the results of experiments conducted with various web log files.**

Keywords-Web Mining; Web Usage Mining; Web Logs; Clustering

I. INTRODUCTION

Web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs. Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web access logs of different web sites can help understand the user behavior and the web structure, thereby improving the website design. Log record has lots of useful information such as URL, IP address, time and so on. Analyzing and discovering log could help us to find more potential users of the web site and trace service quality of the site. The large majority of methods that have been used for pattern discovery from Web data are clustering methods. Clustering has been used for grouping users with common browsing behavior, as well as grouping Web pages with similar content.

In this paper we study, discuss three different approaches on web logs clustering and analyze their benefits and drawbacks. The three web logs clustering algorithms are: Web logs clustering based on Fuzzy Logic, Temporal Cluster analysis of web log data and the Particle Swarm Optimization based web logs clustering. Based on the experimental evaluation with several web log files and on different parameters we have arrived at a conclusion that the Fuzzy Logic for web logs clustering is the most efficient from different perspectives.

II. RELATED WORK

The first approach discussed is a Fuzzy Clustering Algorithm that produces the design mentality of the electronic commerce websites. This algorithm is simple, effective and easy to realize, it is suitable to the web usage mining demand of constructing a low cost B2C website. The second approach is on temporal web logs clustering. Temporal web usage mining is the analysis of cluster behavior over time and it can reveal additional information on the web site usage. There can be two different temporal changes in cluster analysis - change in cluster compositions and change in cluster memberships. TCMM – Temporal Cluster Migration Matrices is a framework useful for the analysis of changes in nature of the web site usage and loyalty of web site users. TCMM also serves as a visualization tool for analysis of results of temporal data mining. In the third approach we apply swarm intelligence to the existing web usage clustering techniques and propose the new PSO based Clustering Algorithm for clustering web usage sessions. PSO clustering algorithm is better than the standard K-means clustering algorithm. The efficiency of the data mining algorithms can be enhanced using optimization techniques. One such optimization technique is applying swarm intelligence to data mining techniques. This

swarm intelligence takes its inspiration from the social and cognitive properties of the vertebrates and insects. This is implemented using software components called Multi Agent Systems that are communicating in a highly decentralized environment. Their cooperative behavior ensures that they converge on an optimum solution.

III. APPROACHES TO WEB LOG CLUSTERING

A. Fuzzy Clustering Algorithm

Step1: Collect the web logs from authenticated sources.

Step2: Pre-treat the web logs

– Clean, format, identify users, sessions, split the web log information into Browser Id, Client Ip / User Id, URL and Timestamp.

Step3: Find the count of user's visits to different pages.

Step4: Establish topology of web site by finding:

a. V = {URL_1, URL_2, URL_n} – set of all URLs
b. R = {<URL_1, URL_2>, <URL_2, URL_3>, …}
   - ordered hyperlink set of pages

Step5: Establish the matix of users visiting pages:

$$B_{m\times n}=\begin{pmatrix} URL_1 \\ URL_2 \\ \cdots \\ URL_m \end{pmatrix}\begin{pmatrix} User_1 & User_2 & \cdots & User_n \end{pmatrix}=\begin{pmatrix} A_{11} & A_{12} & \cdots & A_{1n} \\ A_{21} & A_{21} & \cdots & \\ & \vdots & & \\ A_{m1} & A_{m2} & & A_{mn} \end{pmatrix}$$

- Value of A_{ij} = No. of time the URL_i visited by the $User_j$
- Row Vector B[x, j] = situation of all users visiting the URL_x
- Column Vector B[i, y] = situation of all URLs being visited by the $User_y$
- Sum of rows S_i = No. of times all users visiting the URL_i
- Sum of columns S_j = No. of times all URLs being visited by the Userj

Step6: This matrix can be considered as a relational table and we can use SQL to find the useful information as listed below:

a. First N pages that are mostly accessed
   - Calculate the total no. of times all users visits the pages using the formula
   $$S_i = \sum_{j=1}^{n} A_{ij}$$
   - Compose the aggregate S = {S_1, S_2, ... S_m}, where i = 1, ... m
   - Arrange the values S_1, S_2 in S in descending order, so that the values in the front of the list represent the most important pages.

b. First N users that has the most visiting time
   - Calculate the time spent by every user visiting all the pages of the web site using the formula:
   $$S_j = \sum_{i=1}^{m} A_{ij}$$
   - Compose aggregate of S = {S_1, S_2, ...S_n}, where j = 1,...n
   - Arrange the values of S_1, S_2 in S in descending order, so that the values in the front of the list are the users with most visiting time.

c. Pages which the users are mostly interested in.
   - Calculate the time spent by every user visiting all the pages of the web site using the formula:
   $$S_j = \sum_{i=1}^{m} A_{ij}$$
   - Compose aggregate of S = {S_1, S_2, ...S_n}, where j = 1,...n
   - Calculate the rate of a user visiting all the pages of this website as per the formula:
   $$r_{ij} = A_{ij} / \sum_{j=1}^{n} S_j$$
   - Using this compose a matrix R_{mxn} as below:

$$R_{m\times n}=\begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ & \vdots & & \\ r_{m1} & r_{m2} & \cdots & r_{mn} \end{pmatrix}$$

- Calculate the time the user spends in each page of this web site as per the formula - Settling time = End time – Begin time.
- Using this compose a matrix T_{mxn} as below

$$T_{m\times n}=\begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ t_{21} & t_{22} & \cdots & t_{2n} \\ & \vdots & & \\ t_{m1} & t_{m2} & \cdots & t_{mn} \end{pmatrix}$$

- Take the values from both the above matrices, R & T and arrange them in descending order of the values of rate and time.
- From this we can obtain the pages which the users are mostly interested in.
- That is the visiting rate of these user interested pages are high and the settling time is also longer.

d. Visit characteristics of specific users
   - Each users frequent access path
   - Each users frequent access time

e. Correlated web pages
f. Similar user community

B. Temporal Cluster Migration Matrices Algorithm

TCMM = < H, C, n, Ω: H → C_n >, Where, H is the set of users, C is a set of cluster labels and n represents the number of time periods.

Ω: H → C_n is a mapping that describes the sequence of cluster memberships for each user over n time periods.

Web site users are identified by a seven digit number: H = {h | 1000000 <= h <= 9999999}

The users are grouped into five clusters based on their visiting patterns in a month.

These clusters can be named as:

(i) Loyal big visitors (C = 1)
(ii) Loyal moderate visitors (C = 2)
(iii) Semi-loyal big visitors (C = 3)
(iv) Semi loyal moderate visitors (C = 4)
(v) Infrequent visitors (C = 5)

Hence C = {1, 2, 3, 4, 5} – represents the cluster labels;

n = 6; say, the cluster behavior analyzed for six months.

Step1: Collect the web log data of a web site for six months from authenticated sources.

Step2: Pre-treat the web logs

– Clean, format, identify users, sessions, split the web log information into Browser Id, Client Ip / User Id, URL and Timestamp.

Step3: Find the count of pages visited by each user in each month.

Step4: Based on the count of pages visited by each user, the users are categorized into different clusters identified above.

Step5: This clustering is done for each month.

Step6: Using the clustering results the TCMM matrix is constructed as per the format mentioned in Table I.

Step7: Perform analysis on the TCMM table data using SQL and get the below results.

a. Number of loyal big visitors in each month obtained by the below queries.
b. View of overall clustering for a given period.
c. The list of customers who where loyal throughout the study period.

d. Detect the changes in the customer loyalty (Increase in the cluster label during different iterations of clustering)
(i) Loyalty decreases from 5th Month to 6th Month
(ii) List of customers whose loyalty declined from fourth month to fifth month and the decline in loyalty sustained in the sixth month.

Step8: The results of the above queries can be visualized in a better way by presenting those using bar graphs and pie charts.

TABLE I. SAMPLE TCMM MATRIX

ID	M1	M2	M3	M4	M5	M6
'2221725'	1	1	2	1	1	2
'6361765'	3	3	3	3	1	3
'7777777'	2	1	1	1	1	3
'7371767'	3	1	2	3	1	3

For web personalization we have to concentrate on individual visitors. The company may want to encourage loyal visitors by giving special offers. Another important marketing strategy is to detect the changes in customer loyalty. Business is worried about the attrition rate of their best customers. Let us assume that the cluster labels for the users represent their desire to the business. Then customer's attrition rate will be evident from their increasing cluster labels during repetitive application of clustering. Such a customer may be a potential target for promotional material.

In the result say one customer is always in cluster 3, but moves to cluster 1 in M5 and then again moves to cluster 3 in M6. This is not an indication of potential attrition, but an indication that this customer has the potential to move up to cluster 1. Hence such customers need a different type of campaign than the attrition campaign. Customers who are oscillating between the clusters 1, 2 and 3 can be ignored. Serious indication of attrition is the continued decline of loyalty; that is the customer declined from cluster 1 to cluster 2 and then continued to be in cluster 2 for the subsequent months.

To execute any analysis and to find the result for any set of data, it will be efficient to embed the SQLs in a programming language such as Java. This will help the analyzers to execute more generic and complex queries. TCMM can be applied for a retail ecommerce site or marketing databases.

C. PSO Based Clustering Algorithm

PSO stands for Particle Swarm Optimization. Particles represent individual solutions. Swarm is a collection of particles that represents the solution space.

In PSO the swarm is initialized to a uniform solution set. The particles move through this solution space with velocity v. They maintain the best personal position pBest, best position found by a particle and the global best position gBest, best fitness position for all particles. The velocity of the particles is influenced by the social and cognitive components.

- $V_i(t)$ is the velocity of the particle i at time t
- $X_i(t)$ is the position of the particle i at time t
- w is the inertia weight of the particles
- q_1, q_2 are the vectors representing the cognitive and social components
- r_1, r_2 are the random numbers between 0 and 1
- Range of velocities of the particles is $[-V_{max}, V_{max}]$

$$V_i(t+1) = [w * V_i(t)] + [q_1 * r_1 * (pBest - X_i(t))] + [q_2 * r_2 * (gBest - X_i(t))] \quad (1)$$

$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (2)$$

Stopping criteria for PSO is the maximum number of iterations reached or minimum error requirement reached.

Web session clustering exploits the two main dimensions of web usage data, the Time dimension and Browsing sequence dimension. As per the parameters listed in the Table II, the Time and Browsing sequence dimensions can be combined to identify the relative importance of a visit.

The approach discussed in PSO Based Web Session Clustering applies the PSO optimization to the Web Session Clustering. In this approach each Session is considered as a particle.

Step1: Collect the web logs from authenticated sources and pre-treat the web logs – identify sessions.

Step2: Initially a set of particles are created and uniformly mapped to the input sessions.

Each particle consists of the below attributes:

Particle Id - Uniquely identifies a particle.

DistanceFromEachSession - Array that store the distance of a particle to each session at a particular iteration. WonSessionVectors - Array that represents the session vectors won by a particle at a given iteration. SessionAttributeValues - Represents the current values of the attributes of the particle in the form of a data vector. A session vector consists of Session Id, Client Ip address, Count of Pages visited in the session, Session length (End_time – Start-time), Total Bytes downloaded.

PBest - The position of the nearest session to the particle achieved so far.

TABLE II. WEB SESSION CLUSTERING DIMENSIONS – POTENTIAL PARAMETERS

	Web Session Clustering Dimensions	
	Time	***Browsing Sequence***
Potential Parameters	Total Session Time	Page Visit Sequence
	Time on Each Page	Visit Sequence Similarity
	Average Time per Request	Visit Sequence Length
	% of Session Time on each Page	Topical Sequence of the User

Step3: Initialize the particles for the above attributes.

Step4: Now these particles start moving from their initial position, guided by the cognitive and self organizing component. Each position move is considered as one iteration.

Step5: After each iteration the swarm calculates all its parameters and the swarm organizes itself according to the new data vectors won by each particle.

The cognitive component of the algorithm is encoded as:

$$(pBest - X_i(t)) \quad (3)$$

The self organizing component is encoded as:

$$(Y_i(t) - X_i(t)) \quad (4)$$

Where, Yi(t) is the current value of the particle and Xi(t) is the initial value of the particle. The value of the pBest is calculated based on the particles distance to the current centroid session.

After every iteration, the swarm changes its position by winning the nearest particles. This winning of nearest particle by the swarm is achieved using the Euclidian distance measure.

The Euclidian distance measure is calculated using:

$$d(x_n, z_i) = \| x_n - z_i \| \quad (5)$$

Step6: The iterations are repeated until there are no significant changes in the position of the particles or the number of maximum iterations reached or no movement of data vectors from one data cluster to another is observed.

IV. CRITICAL EVALUATION AND ANALYSIS

Of the three approaches we can see that the Temporal Cluster Migration Matrices approach is just to categorize the web users into different clusters and to study their cluster migration behavior over a period of time. Thus it does not deal on grouping the web site pages, rank the most interesting pages, frequently accessed web site path and correlated web pages. Hence this approach is mostly suitable for online retail web pages to study the customer behaviors and change the marketing promotion strategies accordingly.

On the contrary the Fuzzy Clustering approach can be applied to study any aspect of E-commerce web sites starting from ranking the users based on their visit time and visit frequency, ranking the web pages based on the visiting rate and time spent on the pages, analyze the users visit characteristics to identifying the correlated web pages. It is an easy to apply and simple approach that provides any required knowledge. It is flexible to adapt to complex analysis as well.

The third approach being different from the rest two approaches is based on the PSO optimization technique that is applied on the web session clustering concept. In this we consider the time and the browsing sequence dimensions of a session and apply the PSO concept iteratively to obtain a more accurate session clusters with less intra cluster distance than the k-means clustering algorithm. Thus this clustering can provide knowledge on the

user behavior and common access patterns. Even though it provides all the required knowledge on web usage mining the approach is found to be quite complex and less practical to apply.

Hence after critical evaluation and analysis of the three approaches we can see that the fuzzy clustering logic is simple, effective, and practical to apply and provides all the required knowledge about web usage mining. We can check if the outcome of our analysis holds after experimentation of real time web log data.

V. EXPERIMENTAL RESULTS

To assess the performance of the three clustering approaches we tested the algorithms on the eretailstore web log file for a period of six months, 07/2010 to 12/2010. The logs were processed through the preprocessing steps and were analyzed using the three algorithms. The Fig 1 shown below shows the bar graph representing the runtime taken for clustering of web logs by the three algorithms. The Fig 2 shows the optimized efficiency percentage of the three approaches. Fig 3 compares the memory utilization of Fuzzy, Temporal and PSO algorithms.

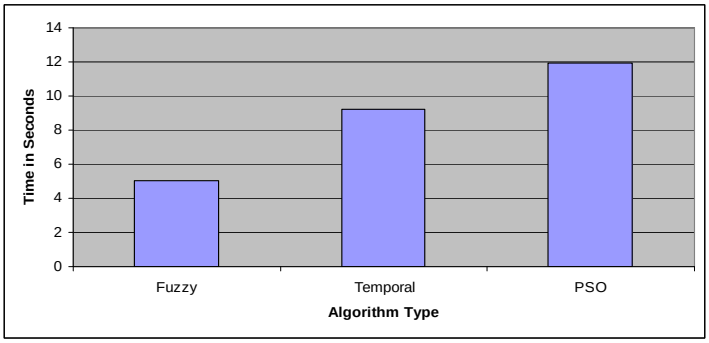

Figure 1. Run Time Comparison of Fuzzy, Temporal and PSO Approaches

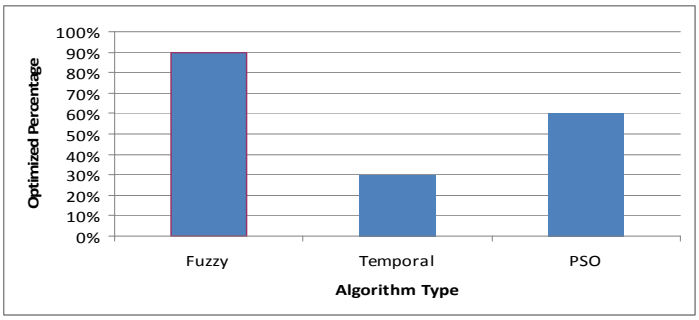

Figure 2. Optimized Percentage Comparison of Fuzzy, Temporal and PSO Approaches

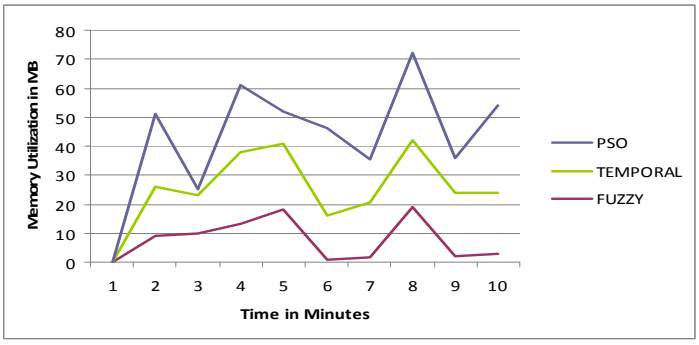

Figure 3. Memory Utilization Comparison of Fuzzy, Temporal and PSO Approaches

Hence after studying the experimental results we can conclude that our critical evaluation holds good and we can conclude based on various aspects that Fuzzy Web Log Clustering is the efficient approach that can be used for web usage mining, especially for E-commerce sites.

VI. CONCLUSION

With the growth of Web based application, specifically electronic commerce, there is significant interest in analyzing Web usage data to better understand Web usage, and apply the knowledge to better serve users. This has lead to a number of open issues in Web Usage Mining area. This article provides a survey of three web logs clustering approaches focusing on its application to Web Personalization. This survey aims to serve as a source of ideas for people working on personalization of information systems. It proposes the easy, simple, best approach, the Fuzzy Clustering to be used for user behavior pattern discovery. This outcome is based on experimental evaluation of several web log files over periods. For future work we should explore the use of Fuzzy logic along with temporal to explore the interesting dimension of the change in web usage behaviors.

REFERENCES

[1] Qingtian Han, Xiaoyan Gao, Wenguo Wu, "Study on Web Mining Algorithm Based on Usage Mining", Computer-Aided Industrial Design and Conceptual Design, 2008. CAID/CD 2008. 9th International Conference on 22-25 Nov. 2008

[2] Lingras, P., Hogo, M. and Snorek, M. "Temporal Cluster Migration Matrices for Web Usage Mining".*In Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence*, 2004.

[3] Shafiq Alam, Gillian Dobbie, Patricia Riddle, "Particle Swarm Optimization Based Clustering of Web Usage Data, " wi-iat, pp.451-454, 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, 2008.

[4] Chen, M. S., Park, J. S., and Yu, P. S., "Efficient Data Mining for Path Traversal Patterns", *IEEE Transactions on Knowledge and Data Engineering*, March/April, 1998, pp. 209-221.

[5] R. Cooley, B. Mobasher and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", *Proceedings of the 9th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'97)*, November.

[6] Hong T, Chiang M, Wang S H, "Mining weighted browsing patterns with linguistic minimum supports", 2002 *IEEE International Conference on Systems, Man and Cybernetics*, 2002,Yasmine Hammamet, Tunisia, pp. 635-639.

[7] O. R. Zaiane, M. Xin and J. Han, "Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs", *Advances in Digital Libraries Conf*, 1998, Santa Barbara, CA, pp.19-29.

[8] Cooley R, Mobasher B, Srivastava J. "Data preparation for mining world wide web browsing patterns", *Knowledge and Information System*, 1999, pp.5-32.

[9] C. Antunes and A Oliveira, "Temporal Data Mining: An Overview", Proceedings of KDD 2001 Workshop on Temporal Data Mining,http://www.acm.org/sigkdd/kdd2001/ Workshops/ ano.pdf, 2001.

[10] T. Joachims, R. Armstrong, D. Freitag, and T. Mitchell, "Webwatcher: A learning apprentice for the world wide web", in *AAAI Spring Symposium on Information Gathering from Heterogeneous, Distributed Environments*, 1995.

[11] I. V. Cadez, D. Heckerman, C. Meek, P. Smyth, and S. White, "Model-based clustering and visualization of navigation patterns on a Web site", *Journal of Data Mining and Knowledge Discovery*, 7(4), 2003.

[12] M. Hogo, M. Snorek, and P. Lingras, "Temporal Web Usage Mining", *Proceedings of 2003 IEEE/WIC InternationalConference on Web Intelligence*, 2003, pp. 450-453.

[13] P. Lingras,R. yan, and M. Hogo, "Rough Set Based Clustering: Evolutionary, Neural, and Statistical Approaches", *Proceedings of First Indian International Conference of Artificial Intelligence*, Hyderabad, India, 2003.

[14] B. Mobasher, R. Cooley, and J. Srivastava, "Automatic Personalization Based on Web Usage Mining", h*ttp://maya.cs.depaul.edu/ ˜ mobasher/ personalization/*, 2000.

[15] Z. Pawlak, "Information Systems: Theoretical Foundations", *Informations Systems*, 6, 1981, pp. 205-218.

[16] M. Perkowitz and O. Etzioni, "Adaptive web sites: Conceptual cluster mining", in Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, 1999.

[17] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan, "Web Usage Mining", *Discovery and Applications of Usage Patterns from Web Data, in SIGKDD Explorations*, 1(2), 2000, pp. 1-12.

[18] X. Yao, "Research Issues in Spatio-temporal Data Mining", *http://www.ucgis.org/Visualization/whitepapers/Yao-KDVIS2003.pdf*, 2003.

[19] Y. Fu, K. Sandhu, and M-Y Shih, "A Generalization-Based Approach to Clustering of Web Usage Sessions", Revised Papers from the International Workshop on Web Usage Analysis and User Profiling, p.21-38, 1999.

[20] C. Shahabi, A. Zarkesh, J. Adibi, and V. Shah. "Knowledge Discovery from Users Web-Page Navigation", In workshop on Research Issues in Data Engineering, England, 1997.

[21] A. Banerjee, and J. Ghosh, "Click stream clustering using Weighted Longest Common Subsequence", Proceedings of the 1st SIAM International Conference on Data Mining: Workshop on Web Mining (2001).

[22] Ajith Abraham, "Natural Computation for Business Intelligence from Web Usage Mining", (SYNASC'05), pp. 3-10, 2005.

[23] S. Schockaert , M. De Cock, C. Cornelis, E. E. Kerre, "Clustering Web Search Results Using Fuzzy Ants", IJIS, Volume 22, Issue 5 , Pages 455 - 474

[24] J. Chen, H. Zhang, "Research on Application of Clustering Algorithm Based on PSO for the Web Usage Pattern", Wireless Communications, Networking and Mobile Computing, 2007.

[25] J. Kennedy, and R. C. Eberhart, "Particle Swarm Optimization", Proc. Of IEEE ICNN, Vol. IV, Perth, Australia (1995) 1942-1948

[26] S.C. M. Cohen, and L. N. de Castro, "Data Clustering with Particle Swarms," IEEE Congress on Evolutionary Computations, Vancouver, BC, Canada, 2006.

AUTHORS PROFILE

G. Sudhamathy has obtained an undergraduate degree B.Sc. (Spl) Mathematics from Madurai Kamaraj University of India in May 1995. She holds a post graduate degree Master of Computer Applications (MCA) at Madurai Kamaraj University of India in April 1998. She is

currently working as Assistant professor in the Department of Computer Applications in Velammal College of Engineering and Technology, Madurai, Tamil Nadu, India. She has a rich industrial experience of about 10.5 years working in various multinational Information Technology companies. She also has around 2 years of academic and research experience. Her research interests are in web usage mining and its application in E-commerce. She has published 1 paper in ACM International journal and have presented a paper in an International conference. She is a member of IEEE Society and ISTE.

Dr. C. Jothi Venkateswaran is the Head of the Department of Computer Applications, Presidency College, Chennai, Tamil Nadu, India. His area of research interests are data mining, image mining, software engineering and database management systems. He has been serving for more than 24 years in teaching and more than 10 years in research field. He has published many articles in the National and International Journals and have presented papers in many conferences. He is a reputed consulting editor in the Institute of Innovative Indian Research. He is a special officer of Directorate of Collegiate Education, Tamil Nadu, India.