A Relational Based Fuzzy Clustering to Mine user Profiles for Web Directory Personalization

R. Lokesh Kumar, T. Gopalakrishnan, P. Sengottuvelan

Abstract - The log data at a search engine can be used to analyze users' search behavior and to develop search technologies to improve users' search experiences. Web usage mining performs mining on web usage data or web logs. A web log is a listing of page reference data/clickstream data. The behavior of the web page readers is imprint in the web server log files. By using the sequence of pages a user accesses, a user profile could be developed thus used in personalization. With personalization, web access or the contents of web page are modified to better fit the desires of the user and also to identify the browsing behavior of the user can improve system performance, enhance the quality and delivery of Internet Information services to the end user, and identify the population of potential customers. For this purpose a new clustering based approach is used, The proposed algorithm finds the meaningful behavior patterns extracted by applying efficient clustering algorithm, to log data. It is proved that performance of the proposed system is better than that of the existing algorithm. The proposed algorithm can provide popular information from web page visitors for web personalization.

Keywords - User profiles, web log data, clustering, Web Personalization.

I. INTRODUCTION

The ease and speed with which user transactions can be carried out over the Web have been a key driving force in the rapid growth of e-commerce. The ability to track user browsing behavior down to individual mouse clicks has brought the vendor and end customer personalization is applicable to any Web browsing activity, not just e-commerce. Web personalization can be defined as any action that tailors the Web experience to a particular user, or set of users. The experience can be something as casual as browsing a Web site or as (economically) significant as trading stocks or purchasing a car. The actions can range from simply making the presentation more pleasing to anticipating the needs of a user and providing customized information. To date, most personalization systems for the Web have fallen into three major categories. Manual decision rule systems, collaborative filtering systems, and content-based filtering agents. Manual decision rule systems, such as Broadvision allow Web site administrators to specify rules based on user demographics or static profiles (collected through a registration process), or session history. The rules are used to affect the content served to a particular user. Collaborative filtering systems, such as Firefly [11], and Net Perceptions typically take explicit information in the form of user ratings or preferences, and, through a correlation engine, return information that is predicted to closely match the users' preferences. Content-based filtering approaches such as those used by WebWatcher [5][6] rely on content similarity of Web documents to personal profiles obtained explicitly or implicitly from users.

The new generation of Web personalization tools is attempting to incorporate techniques for pattern discovery from Web usage data. For example, some collaborative filtering systems such as Net Perceptions are experimenting with obtaining implicit user ratings from usage data. Web usage mining systems run any number of data mining algorithms on usage or clickstream closer than ever before. It is now possible for vendors to personalize their product messages for individual customers on a massive scale with the rapid progress of World Wide Web (WWW) technology a huge number of web log access log records are being collected. It is not easy to perform systematic analysis on such huge amount of data, however many people realized the potential usage of data to make effective use of web access history for server performance, system design improvement, or customer targeting in electronic commerce [4] [2]. With site mining, the overall quality and effectiveness of the pages at the site can be evaluated. The different modes of usage called user profiles can be discovered using a clustering that extract access patterns from the clickstreams stored in web log files.

II. RELATED WORK

Web usage mining has been used extensively for Web personalization. A number of personalized services employ which could complicate the construction of rich hierarchical models. This method makes the assumption that usage data are collected from the navigation of users within the Web directory [23]. Thus, its applicability to independent services such as a Web portal is questionable.

A knowledge discovery framework for building Web directories according to the preferences of user communities are build. Community Web directories are more appropriate than personal user models for personalization across Web sites, since they aggregate statistics for many users under a predefined thematic taxonomy, thus making it possible to handle a large amount of data, residing in a sparse dimensional space. This is the first attempt to construct aggregate user models, i.e., communities, using navigational data from the whole Web [1]. This address only the problem of "local overload." by combining thematic with usage information to model the user communities and leads to our approaches a new method that combines crisp clustering with probabilistic models.

Manuscript Received June 12, 2012.

R. Lokesh Kumar, Department of Information Technology, Bannari Amman Institute of Technology, Sathyamagalam, Tamil Nadu ,India

T.Gopalakrishnan, Dept. of Information Technology, Bannari Amman Institute of Technology, Sathyamagalam, Tamil Nadu ,India

P. Sengottuvelan ,Dept. of Information Technology, Bannari Amman Institute of Technology, Sathyamagalam, Tamil Nadu ,India

The other method presented a framework for mining, tracking, and validating evolving multifaceted user profiles on Web sites that have all the challenging aspects of real-life Web usage mining, including evolving user profiles and access patterns, dynamic Web pages, and external data describing ontology of the Web content [22].A multifaceted user profile summarizes a group of users with similar access activities and consists of their viewed pages, search engine queries, and inquiring and inquired companies.

The choice of the period length for analysis depends on the application or can be set, depending on the cross-period validation results[7]. Even though it focus on scalability, the latter can be addressed by following an approach similar to where Web clickstreams are considered as an evolving data stream, or by mapping some new sessions to persistent profiles and updating these profiles, hence eliminating most sessions from further analysis and focusing the mining on truly new sessions.

III. CLUSTER ENSEMBLE METHODOLOGY

Let X = {$x_1 \ldots ; x_n$} be a set of N data points and $\prod$ = { $\prod$1, $\prod$ 2,…. $\prod$ n) be a cluster ensemble with M base clusterings, each of which is referred to as an ensemble member. Each base clustering returns a set of clusters such ci ={c_1,c_2…c_n} is the number of clusters in the ith clustering[16]. For each x in $\prod$ denotes the cluster label to which the data point x belongs. In the ith clustering, Ci (or "Cij") if x E Cij. The problem is to find a new partition of a data set X that summarizes the information from the cluster ensemble

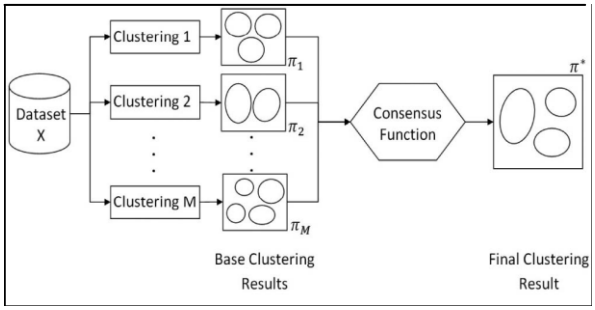

Fig. 1 shows the general framework of cluster ensembles.

Essentially, solutions achieved from different base clustering are aggregated to form a final partition. This metalevel methodology involves two major tasks of: 1) generating a cluster ensemble, and 2) producing the final partition, normally referred to as a consensus function.

Ensemble Generation Methods

It has been shown that ensembles are most effective when constructed from a set of predictors whose errors are dissimilar [17]. To a great extent, diversity among ensemble members is introduced to enhance the result of an ensemble [14] [16]. Particularly for data clustering, the results obtained with any single algorithm over much iteration are usually very similar. In such a circumstance where all ensemble members agree on how a data set should be partitioned, aggregating the base clustering results will show no improvement over any of the constituent members. As a result, several heuristics have been proposed to introduce artificial instabilities in clustering algorithms, giving diversity within a cluster ensemble. The following ensemble generation methods yield different clusterings of the same data, by exploiting different cluster models and different data partitions.

Homogeneous ensembles

Base clustering are created using repeated runs of a single clustering algorithm, with several sets of parameter initializations, such as cluster centers of the k-means clustering technique [21].

Random-k

One of the most successful techniques is randomly selecting the number of clusters (k) for each ensemble member.

Data subspace/sampling

A cluster ensemble can also be achieved by generating base clusterings from different subsets of initial data. It is intuitively assumed that each clustering algorithm will provide different levels of performance for different partitions of a data set [19]. Practically speaking, data partitions are obtained by projecting data onto different subspaces, choosing different subsets of features or data sampling.

Heterogeneous Ensembles

A number of different clustering algorithms are used together to generate base clustering.

Mixed heuristics

In addition to using one of there for mentioned methods, any combination of them can be applied as well [20].

A. A Novel Link-Based Approach

Existing cluster ensemble methods to categorical data analysis rely on the typical pairwise - similarity and binary cluster-association matrices, which summarize the underlying ensemble information at a rather coarse level.

Many matrix entries are left "unknown" and simply recorded as "0." Regardless of a consensus function, the quality of the final clustering result may be degraded. As a result, a linkbased method has been established with the ability to discover unknown values and, hence, improve the accuracy of the ultimate data partition [9]. In spite of promising findings, this initial framework is based on the data point data point pairwise-similarity matrix, which is highly expensive to obtain. The link-based similarity technique, SimRank [10], that is employed to estimate the similarity among data points is inapplicable to a large data set.

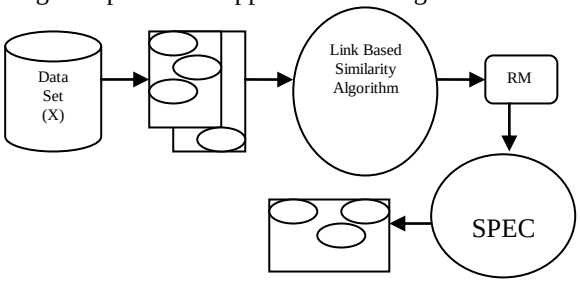

Fig.2 The link-based cluster ensemble framework: 1)

a cluster ensemble $\prod = \{\prod_1, \prod \ldots \prod_M\}$ is created from M base clustering, 2) a refined cluster-association matrix is then generated from the ensemble using a link-based similarity algorithm, and 3) a final clustering result ($\prod_*$) is produced by a consensus function of the spectral graph partitioning.

To overcome these problems, a new link-based cluster ensemble (LCE) approach is introduced herein. It is more efficient than the former model, where a BM-like matrix is used to represent the ensemble information. The focus has shifted from revealing the similarity among data points to estimating those between clusters. A new link-based algorithm has been specifically proposed to generate such measures in an accurate, inexpensive manner. The LCE methodology is illustrated in Fig. 3. It includes three major steps of: 1) creating base clusterings to form a cluster ensemble generating a refined cluster-association matrix (RM) using a link-based similarity algorithm, and producing the final data partition by exploiting the spectral graph partitioning technique as a consensus function.

B. Creating A Cluster Ensemble

Following the study in [8], the first type of cluster ensemble transforms the problem of categorical data clustering to cluster ensembles by considering each categorical attribute value (or label) as a cluster in an ensemble. Let $x=\{x_1,x_2 \ldots x_n\}$ be a set of N data points, $A=\{a_1,a_2..a_M\}$ be a set of categorical attributes, and $\pi = \{\pi_1,\pi_2 \ldots \pi_n\}$ be a set of M partitions. Each partition is generated for a specific categorical attribute ai ε A. With this formalism, categorical data X can be directly transformed to a cluster ensemble without actually implementing any base clustering. While single-attribute data partitions may not be as accurate as those obtained from the clustering of all data attributes, they can bring about great diversity within an ensemble. Besides its efficiency, this ensemble generation method has the potential to lead to a high-quality clustering result.

Unlike the previous case, the following two ensemble types are created from base clustering results, each of which is obtained by applying a clustering algorithm to the categorical data set. For this study, the k-modes technique [12] is used to generate base clusterings, each with a random initialization of cluster centers. In particular to a full-space ensemble, base clusterings are created from the original data, i.e., with all data attributes. To introduce an artificial instability to k-modes, the following two schemes are employed to select the number of clusters in each base clusterings: 1) Fixed-k, k =[√N] (where N is the number of data points), and 2) Random-k, k ε 2 [√N].

C. Generating A Refined Matrix

Several cluster ensemble methods, both for numerical [12] [15] [18] and categorical data are based on the binary cluster-association matrix. Each entry in this matrix BM (xi,cl) ε (0,1) f0; 1g represents a crisp association degree between data point xi 2 X and cluster C_1 . According to Fig. 2 that shows an example of cluster ensemble and the corresponding BM, a large number of entries in the BM are unknown, each presented with "0." Such condition occurs when relations between different clusters of a base clustering are originally assumed to be nil. In fact, each data point can possibly associate to a certain degree within several clusters of any particular clustering. These hidden or unknown associations can be estimated from the similarity among clusters, discovered from a network of clusters.

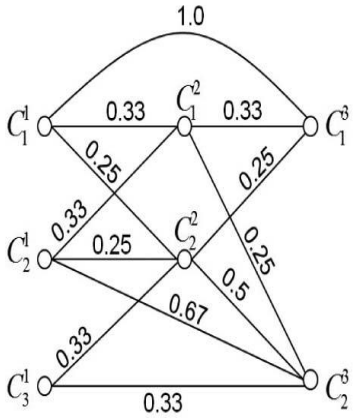

Fig 3. An example of a cluster network, where each edge is marked with its weight.

IV. A NEW LINK-BASED SIMILARITY ALGORITHM

Given a cluster ensemble of a set of data points X, a weighted graph G = (V,W) can be constructed, where V is the set of vertices each representing a cluster and W is a setof weighted edges between clusters. Formally, the weight assigned to the edge w_{xy} ε W, that connects clusters Cx; Cy ε V , is estimated by the proportion of their overlapping members.

$$Wxy = \frac{Lx \cap Ly}{LxULy}$$

Where Lz ε X denotes the set of data points belonging to cluster Cz ε V. Fig. 4 shows the network of clusters that is generated from the example given in Fig. 2. Note that circle nodes represent clusters and edges exist only when the corresponding weights are nonzero[13][15]. For WTQ, can be modified to discriminate the quality of shared triples between a pair of clusters in question. The quality of each cluster is determined by the rarity of links connecting to other clusters in a network. With a weighted graph G presented in Fig. 4, the WTQ measure of clusters Cx; Cy 2 V with respect to each triple Ck 2 V is estimated by

$$WTQ^k_{XY} = \frac{1}{W_k}$$

Here, W_k is defined as $W_k = \sum_{\forall t \varepsilon Nk} w_{tk}$, where Nkε V denotes the set of clusters that is directly linked to the cluster C_k, such that ∀ Ct $\sum \varepsilon N_{k,} W_{tk} \varepsilon W$. The accumulative WTQ score from all triples (1 . . . q) between clusters C_x and C_y can be found as follows:

The WTQ algorithm is summarized below:

ALGORITHM: WTQ (G,Cx, Cy)

G = (V ,W), a weighted graph, where Cx, $C_y \varepsilon V$.
Nk εV , a set of adjacent neighbors of $C_k \varepsilon$ V
$W_k = \sum \varepsilon N_{k,} W_{tk} \varepsilon W$
WTQ_{xy}, the WTQ measure of Cx {and} Cy;
(1) $WTQ_{xy} \leftarrow 0$
(2) For each C ε Nx
(3) If C ε Ny

(4) $WTQ_{xy} \leftarrow WTQ_{xy} + 1 / We$

(5) Return WTQ_{xy}

Following that, the similarity between clusters Cx and Cy can be estimated by

$$Sim(C_x,C_y) = \frac{WTQ_{xy}}{WTQ_{MX}} \text{ X DC}$$

V. RESULTS AND DISCUSSION

A. Datasets

The experiments are conducted on two anonymous web log datasets obtained from UCI machine learning repository [1].

B. The Microsoft anonymous web data.

This dataset has preprocessed web logs of the site www.microsoft.com [24]. It records 37,711 randomly selected anonymous users of the site of which 32,711 are given as training set and 5000 as test set. For each user, the data lists all the areas of the web site that user visited in a one-week timeframe. There are 294 attributes, which are the areas of the site. Here short transactions of size less than 4 are removed for our purpose, as the average transaction length is about 4. Accordingly for both the datasets 5000 user transactions are used for training purpose and 500 transactions are used for testing.

C. Evaluation of clustering techniques based on profile effectiveness

The WAVP of the profiles generated by each of the clustering algorithms is calculated for both the data sets. Here most profiles generated using subtractive clustering algorithm has higher WAVP. Even though FCM, K-means and subtractive clustering each has 2 profiles above 0.6%, most profiles of subtractive clustering are in between 0.5 and 0.6. Next to subtractive clustering, FCM has performed well with most of its profiles placed between 0.4 and 0.5. Kmodes also has got most of its profiles (up to 60 %) in this range i.e., 0.4 and 0.5 but it has no profile having WAVP>0.6%

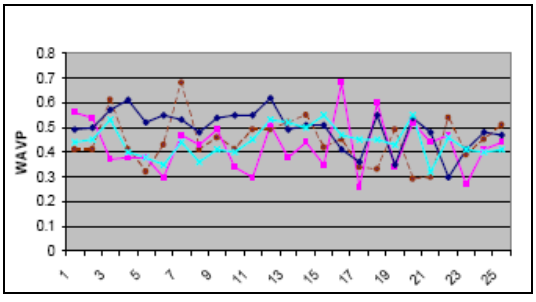

Fig 4. For *msweb* data, WAVP of almost all profiles obtained using Similarity based clustering are between 0.3 and 0.6. 5 out of 7 profiles obtained are between 0.4 and 0.5%. 0.5. But there is no profile with WAVP lying between 0.5 and 0.6.

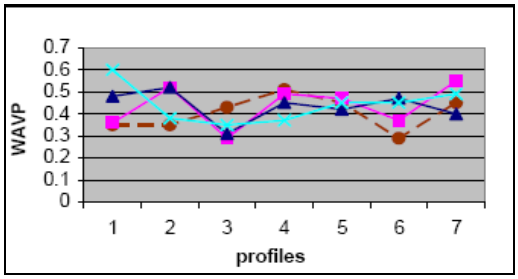

Fig 5 For *msweb* data, WAVP of almost all profiles obtained using Similarity based clustering are between 0.6 and 0.7. 5 out of 7 profiles obtained are between 0.4 and 0.5%. 0.5. But there is no profile with WAVP lying between 0.5 and 0.7.

Table 1.Shows the Results of Clustering on the msweb data after the new link based method.

Threshold	Coverage	Precision	F_1	Avg No of Records
0.2	0.56	0.22	0.28	10.32
0.3	0.5	0.23	0.28	7.47
0.4	0.37	0.28	0.28	4.44
0.5	0.27	0.31	0.25	2.68
0.6	0.14	0.24	0.16	0.41
0.7	0.05	0.11	0.06	0.24

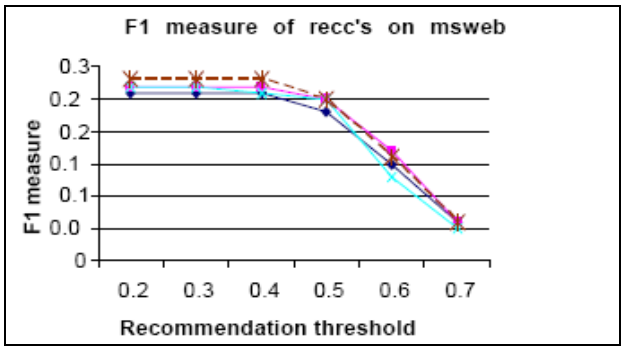

Fig 5 measures the F1 records on msweb

Results on Microsoft data

The coverage values are decreasing as the recommendation threshold increases and Precision values are increasing.

VI. CONCLUSION

This paper presents a novel highly effective Relational based approach to categorical data clustering. It transforms the original categorical data matrix to an information-preserving numerical variation (RM), to which an effective graph partitioning technique can be directly applied. The problem of constructing the RM is efficiently resolved by the similarity among categorical labels (or clusters), using the Weighted Triple-Quality similarity algorithm. The empirical study, with different ensemble types, validity measures, and data sets,

suggests that the proposed link-based method usually achieves superior clustering results compared to those of the traditional categorical data algorithms and benchmark cluster ensemble techniques. The prominent future work includes an extensive study regarding the behavior of other link-based similarity measures within this problem context. Also, the new method will be applied to specific domains, including medical data, tourism sets.

REFERENCES

1. B. Mobasher, R. Cooley, and J. Srivastava, "Automatic Personalization Based on Web Usage Mining," Comm. ACM, vol. 43, no. 8, pp. 142-151, 2000.
2. J. Srivastava, R. Cooley, M. Deshpande, and P.T. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data," SIGKDD Explorations, vol. 1, no. 2, pp. 12-23, 2000.
3. D. Pierrakos, G. Paliouras, C. Papatheodorou, and C.D. Spyropoulos, "Web Usage Mining as a Tool for Personalization: A Survey," User Modeling and User-Adapted Interaction, vol. 13, no. 4, pp. 311-372, 2003.
4. G. Paliouras, C. Papatheodorou, V. Karkaletsis, and C.D. Spyropoulos, "Discovering User Communities on the Internet Using Unsupervised Machine Learning Techniques," Interacting with Computers J., vol. 14, no. 6, pp. 761-791, 2002.
5. G. Xu, Y. Zhang, and Y. Xun, "Modeling User Behaviour for Web Recommendation Using lda Model," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence and Intelligent Agent Technology, pp. 529-532, 2008.
6. W. Chu and S.-T.P. Park, "Personalized Recommendation on Dynamic Content Using Predictive Bilinear Models," Proc. 18th Int'l Conf. World Wide Web (WWW), pp. 691-700, 2009.
7. A.L.N. Fred and A.K. Jain, "Combining Multiple Clusterings Using Evidence Accumulation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 835-850, June 2005.
8. S. Monti, P. Tamayo, J.P. Mesirov, and T.R. Golub, "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data," Machine Learning, vol. 52, nos. 1/2, pp. 91-118, 2003.
9. N. Iam-On, T. Boongoen, and S. Garrett, "Refining Pairwise Similarity Matrix for Cluster Ensemble Problem with Cluster Relations," Proc. Int'l Conf. Discovery Science, pp. 222-233, 2008.
10. T. Boongoen, Q. Shen, and C. Price, "Disclosing False Identity through Hybrid Link Analysis," Artificial Intelligence and Law, vol. 18, no. 1, pp. 77-102, 2010.
11. N. Nguyen and R. Caruana, "Consensus Clusterings," Proc. IEEE Int'l Conf. Data Mining (ICDM), pp. 607-612, 2007.
12. A.P. Topchy, A.K. Jain, and W.F. Punch, "Clustering Ensembles: Models of Consensus and Weak Partitions," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 12, pp. 1866-1881, Dec. 2005.
13. C. Boulis and M. Ostendorf, "Combining Multiple Clustering Systems," Proc. European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 63-74, 2004.
14. B. Fischer and J.M. Buhmann, "Bagging for Path-Based Clustering," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 25, no. 11, pp. 1411-1415, Nov. 2003.
15. Domeniconi and M. Al-Razgan, "Weighted Cluster Ensembles: Methods and Analysis," ACM Trans. Knowledge Discovery from Data, vol. 2, no. 4, pp. 1-40, 2009.
16. Z. Fern and C.E. Brodley, "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning," Proc. Int'l Conf. Machine Learning (ICML), pp. 36-43, 2004.
17. Strehl and J. Ghosh, "Cluster Ensembles: A Knowledge Reuse Framework for Combining Multiple Partitions," J. Machine Learning Research, vol. 3, pp. 583-617, 2002.
18. Ayad and M. Kamel, "Finding Natural Clusters Using Multiclusterer Combiner Based on Shared Nearest Neighbors," Proc. Int'l Workshop Multiple Classifier Systems, pp. 166-175, 2003.
19. N. Fred and A.K. Jain, "Combining Multiple Clustering Using Evidence Accumulation," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 27, no. 6, pp. 835-850, June 2005.
20. S. Monti, P. Tamayo, J.P. Mesirov, and T.R. Golub, "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualization of Gene Expression Microarray Data," Machine Learning, vol. 52, nos. 1/2, pp. 91-118, 2003.
21. N. Iam-On, T. Boongoen, and S. Garrett, "Refining Pairwise Similarity Matrix for Cluster Ensemble Problem with Cluster Relations," Proc. Int'l Conf. Discovery Science, pp. 222-233, 2008.
22. Olfa Nasraoui, Maha Soliman, Esin Saka "A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites" IEEE transactions on knowledge and data engineering, vol. 20, no. 2, February 2008.
23. Dimitrios Pierrakos, Georgios Paliouras, Personalizing Web Directories with the Aid of Web Usage Data" IEEE transactions on knowledge and data engineering, vol. 22, no. 9, September 2010.
24. A UCI machine Learning Repository
25. http://archive.ics.uci.edu/ml/datasets.html

AUTHORS PROFILE

Mr.R.Lokeshkumar received B.E Degree in Computer Science & Engineering Anna University , Chennai in 2006 and M.Tech Degree in Information Technology from Faculty of Engineering and Technology, Anna University, Coimbatore in 2009. Currently he is working as Assistant Professor in the Department of IT, Bannari Amman Institute of Technology, Sathyamangalam. He is doing part time research in Data Mining at Anna University, Coimbatore. His current research focuses on Data Mining, Data Base Systems,Web Mining . He is a member of ISTE.

T.Gopalakrishnan received B.Tech Degree in Information Technology from Anna University, Chennai in 2005 and M.E. Degree in Computer and Communication from Anna University, Chennai in 2008. Currently he is working as Assistant Professor in the Department of IT, Bannari Amman Institute of Technology, and Sathyamangalam. He is doing part time research in Data Mining at Anna University, Coimbatore. His current research focuses on Software Mining, Business Intelligence Data Mining , and Grid Computing.

Dr.P.Sengottuvelan received M.Sc., Degree in Computer Technology from Periyar University, Salem in 2001 and Master of Philosophy in Computer Science from Bharathiar University, Coimbatore in 2003 and M.E. degree in Computer Science & Engineering from Anna University, Chennai in 2004. He also received his Ph.D in degree in Computer Science & Engineering Vinayaka Missions University, Salem in 2010. Since 2004, he is working as Associate Professor in the Department of IT, BIT, Sathyamanagalam. His current research focuses on Concurrent Engineering, Multi Agent System networks, Constraint Management Agents He is member of IACSIT, ACEEE, IAENG and Life Member of FUWA and ISTE.