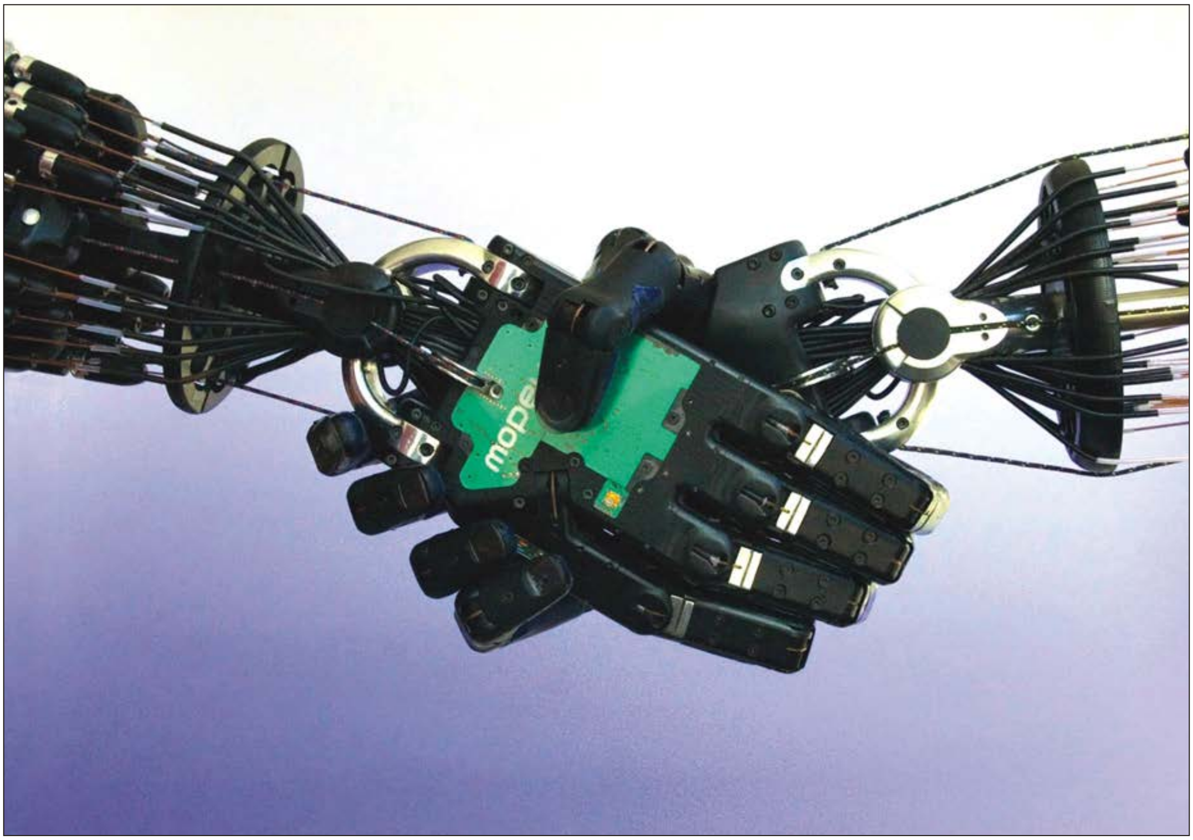


Journal of Artificial Intelligence

ISSN 1994-5450

Journal of Artificial Intelligence 6 (3): 210-219, 2013
ISSN 1994-5450 / DOI: 10.3923/jai.2013.210.219

Applicability of Ensemble Clustering and Ensemble Classification Algorithm for User Navigation Pattern Prediction

[1]V. Sujatha, [2]M. Punithavalli and [2]V. Thavavel

[1]Department of Computer Science, [2]Department of Computer Application, CMS College of Science and Commerce Ramakrishna Engineering College Karunya University, Coimbatore, India

Corresponding Author: V. Sujatha, Department of Computer Science, CMS College of Science and Commerce Ramakrishna Engineering College Karunya University, Coimbatore, India

ABSTRACT

Web Usage Mining (WUM) is used to discover user navigation pattern from Web log data. This study presents the Prediction of User navigation patterns using Clustering and Classification from web log data. In the first stage Predicting user navigation pattern using Clustering and Classification (PUCC) focuses on separating the potential users in web log data and in the second stage clustering process is used to group the potential users with similar interest and in the third stage the results of classification and clustering is used to predict the user future requests. The experimental results represent that the approach can improve the quality of clustering and classification by applying the ensemble model to group the clustering and classification algorithm for user navigation pattern in web usage mining systems.

Key words: Web usage mining, user session, weblog data, clustering, classification

INTRODUCTION

WWW refers to open development phase of the internet which provides a way to gather information from web server. Every hit to a web page is stored in the web log files. Web usage mining is the application of data mining techniques to discover usage pattern from Web log data, in order to understand and better serve the needs of Web-based applications. In this study Clustering separates a web log data into groups with similar features. Classification, on the other hand, a data is assigned to a predefined labelled category, if it has more features similar to that group. Both areas are used for knowledge discovery.

In this study, a solution to predict user request from navigation pattern is proposed. The main objective of the proposed system 'Predicting User navigation patterns using three Clustering and three Classification from web log data, Predicting User navigation pattern using Clustering and Classification (PUCC) (Sujatha and Punithavalli, 2012) is to predict user navigation patterns using knowledge from (1) A classification process that identifies potential users from web log data and (2) A clustering process that groups potential users with similar interest and (2) Using the results of classification and clustering, predict future user requests. For this purpose, three clustering algorithms like Ant based clustering, Graph partitioning and Pairwise Nearest Neighbour algorithm are ensemble and three classification algorithm like Maximum Likelihood algorithm, Longest Common Subsequence algorithm and Markov Chain Model are ensemble and are used and explain in the next section.

This study presents a user navigation pattern algorithm that enhances the work of Jalali *et al.* (2008). Longest common subsequence algorithm for classifying user navigation patterns for predicting users' future requests.

RELATED WORK

A model of user browsing pattern that separates web page references into those made for navigation purposes and those for information content purposes (Cooley *et al.*, 1997). Finding the effective way to infer demographic information about people who use the internet (Masand and Spiliopoulou, 2000). Web query log analysis may be significantly shifted depending on the fraction of agents (Buzikashvili, 2007). Web traversal pattern mining involves discovering users' access patterns from web server access logs. Lee and Yen (2008) put forth a web usage mining technique based on LCS algorithm for online predicting recommendation systems (Jalali *et al.*, 2008) it provides a whole framework and findings in mining Web usage navigation from Web log files of a genuine Web site which has every challenging characteristics of real-life (Nasraoui *et al.*, 2008). Kosala and Blockeel (2000) proposed a system for discovering user navigation patterns using ensemble model for clustering.

METHODOLOGY

The general architecture of the proposed system is given in Fig. 1. The heart of the system is the web log data, which stores all the successful hit made in the Internet. A hit is defined as a request to view a HTML document or image or any other document. The web log data are automatically created and can be obtained from either client side server or proxy server or from an organization database (Srivastava *et al.*, 2000).

Weblog files: Web log file (Fig. 2) is log file automatically created and maintained by a web server.

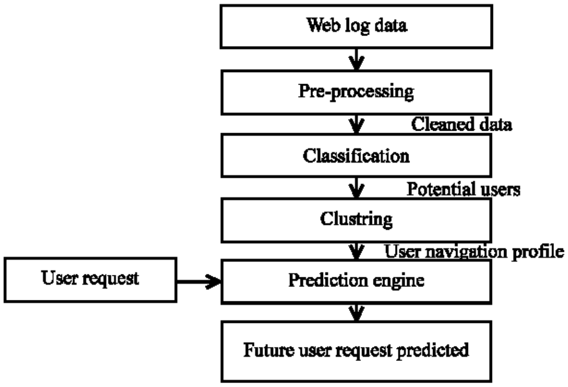
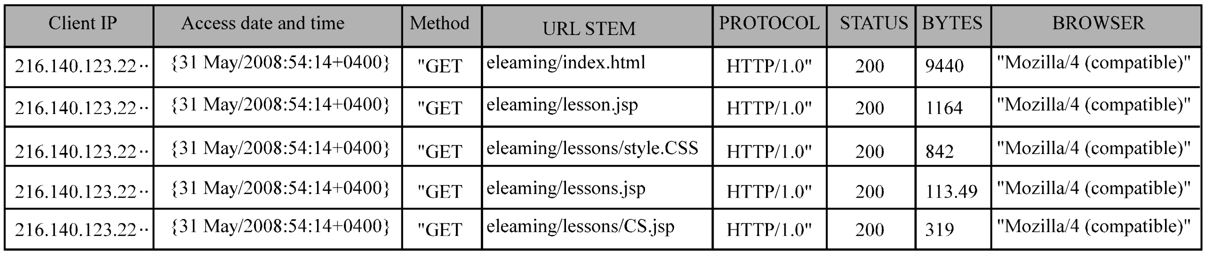


Fig. 1: Predicting user navigation using clustering and classification model



Client IP	Access date and time	Method	URL STEM	PROTOCOL	STATUS	BYTES	BROWSER
216.140.123.22··	{31 May/2008:54:14+0400}	"GET	eleaming/index.html	HTTP/1.0"	200	9440	"Mozilla/4 (compatible)"
216.140.123.22··	{31 May/2008:54:14+0400}	"GET	eleaming/lesson.jsp	HTTP/1.0"	200	1164	"Mozilla/4 (compatible)"
216.140.123.22··	{31 May/2008:54:14+0400}	"GET	eleaming/lessons/style.CSS	HTTP/1.0"	200	842	"Mozilla/4 (compatible)"
216.140.123.22··	{31 May/2008:54:14+0400}	"GET	eleaming/lessons.jsp	HTTP/1.0"	200	113.49	"Mozilla/4 (compatible)"
216.140.123.22··	{31 May/2008:54:14+0400}	"GET	eleaming/lessons/CS.jsp	HTTP/1.0"	200	319	"Mozilla/4 (compatible)"

Fig. 2: Sample web log file

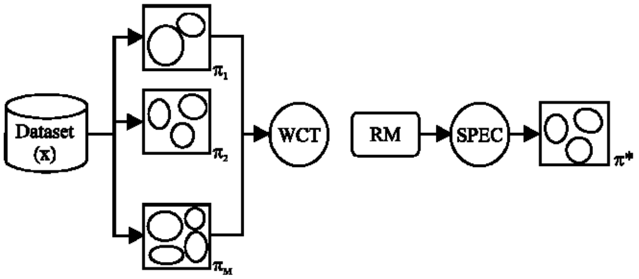


Fig. 3: Cluster ensembles model

Pre-processing: The pre-processing steps include cleaning, user identification and session identification. Cleaning is the process which removes all entries which will have no use during analysis or mining.

Identification of potential users: Sujatha and Punithavalli (2012) focuses on separating the potential users from others. Suneetha and Krishnamoorthi (2010) used decision tree classification using C4.5 algorithm to identify interested users. The attributes selected are time (>30 seconds), number of pages referred in a session (Session time = 30 min) and the access method used. The decision rule for identifying potential users is "If Session Time >30 min and Number of pages accessed >5 and Method used is POST then the classify user as "Potential" else classify as "Not-Potential". The purpose of introducing classification is to reduce the size of the log file. This reduction in size will help for efficient clustering and prediction.

Cluster ensemble methodology: Let X = $\{x_1..., x_n\}$ 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 (Eirinaki and Vazirgiannis, 2003). For each x in $\prod$ denotes the cluster label to which the data point×belongs. In the ith clustering, Ci (or "C_{ij}") if x E C_{ij}. The problem is to find a new partition of a data set X that summarizes the information from the cluster ensemble.

Figure 3 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. To a great extent, diversity among ensemble members is introduced to enhance the result of an ensemble (Fischer and Buhmann, 2003; Fern and Brodley, 2004). 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 clustering's of the same data, by exploiting different cluster models like ant based clustering and graph partition with different partitions (Fig. 4).

Weighted connected-triple (WCT) algorithm: Given a cluster ensemble $\prod$ 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 in $\prod$ and W is a set of weighted edges between clusters. Formally, the weight assigned to the edge w_{ij} connecting clusters C_i and C_j V is estimated in accordance with the proportion of their overlapping members:

$$w_{ij} = \frac{|XC_i \cap XC_j|}{|XC_i \cup XC_j|} \tag{1}$$

where, $XCi \subset X$ denotes the set of data points belonging to cluster Ci. Instead of counting the number of triples as a whole number, the weighted connected-triple method regards each triple as the minimum weight of the two involving edges:

$$WCT^{k}{}_{ij=} \min (w_{ik}, w_{jk}) \tag{2}$$

where, $WCT^{k}{}_{ij}$ is the count of the connected-triple between clusters $C_i;C_j \in V$ whose common neighbor is cluster $Ck \in V$. The count of all q(1 = q<8) triples between cluster C_i and cluster C_j can be calculated as follows:

$$q = \frac{WCT_{ij} = \Sigma WCT^{k_{ij}}}{K = 1} \tag{3}$$

Following that, the similarity SimWCT (i, j) between clusters C_i and C_j can be estimated as follows, where WCTmax is the maximum WCTxy value of any two clusters within the cluster ensemble$\prod$

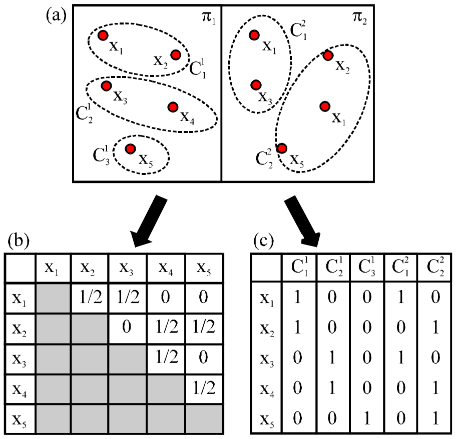
$$Sim = \frac{WCT\ (i, j) = WCTij}{WCTmax} \tag{4}$$

Algorithm: WTQ (G,Cx, Cy)

```
G = (V, W), a weighted graph, where Cx, Cy$\varepsilon$V
Nk $\varepsilon$V, a set of adjacent neighbors of Ck$\varepsilon$ V
Wk = $\Sigma\varepsilon$Nk,Wtk$\varepsilon$W
WTQxy, the WTQ measure of Cx {and} Cy;
 (1) WTQxy$\leftarrow$0
 (2) For each C$\varepsilon$Nx
 (3) If C $\varepsilon$ Ny
 (4) WTQxy$\leftarrow$WTQxy+1/We
 (5) Return WTQxy
```

Following that, the similarity between clusters Cx and Cy can be estimated by a relational based fuzzy clustering:

$$Sim\ (Cx,Cy) = \frac{WTQxy \times DC}{WTQMX}$$



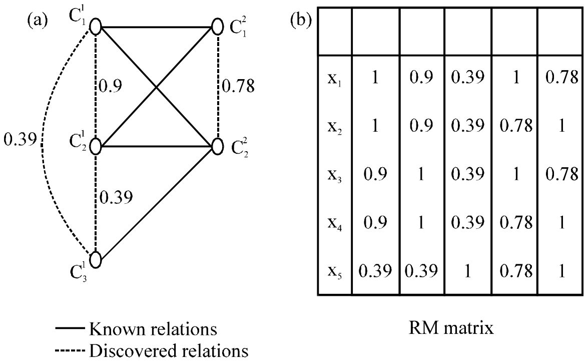
(b)

	x_1	x_2	x_3	x_4	x_5
x_1		1/2	1/2	0	0
x_2			0	1/2	1/2
x_3				1/2	0
x_4					1/2
x_5					

(c)

	C^1_1	C^1_2	C^1_3	C^2_1	C^2_2
x_1	1	0	0	1	0
x_2	1	0	0	0	1
x_3	0	1	0	1	0
x_4	0	1	0	0	1
x_5	0	0	1	0	1

Fig. 4(a-b): An example of (a) Cluster ensemble of samples {x1... x5} that consists of two base clustering (π1 = {$C^1{}_1$, $C^1{}_2$, $C^1{}_3$} and π2 = {$C^2{}_1$, $C^2{}_2$}), (b) Corresponding pairwise similarity matrix and (c) binary cluster association matrix (BM), respectively



x_1	1	0.9	0.39	1	0.78
x_2	1	0.9	0.39	0.78	1
x_3	0.9	1	0.39	1	0.78
x_4	0.9	1	0.39	0.78	1
x_5	0.39	0.39	1	0.78	1

RM matrix

Fig. 5(a-b): (a) Details of disclosed WCT similarities/relations with DC being 0.9 and the and (b) Resulting Refined cluster association matrix (RM)

Applying a consensus function to RM: Having obtained a refined cluster-association matrix (RM) with the aforementioned link-based similarity algorithm, a graph-based partitioning method is exploited to obtain the final clustering. This consensus function requires the underlying matrix to be initially transformed into a weighted bipartite graph. Formally, given an refined cluster association matrix (RM). Figure 5 representing associations between N samples and P clusters in an ensemble ∏, a weighted bipartite graph G = (V, W) can be constructed, where V = $V^X \cup V^C$ is a set of vertices representing both samples V,X and clusters V,C and W denotes a set of weighted edges that can be defined as follows:

- wij: 0 when vertices vi, vj V X, i.e., correspond to samples
- wij: 0 when vertices vi, vj∈V C, i.e., correspond to clusters
- wij: RM (vi, vj) when vertices vi∈V X and vj∈V C. The bipartite graph G is bi-directional such that wij is equivalent to wji. Given such graph, the Spectral Graph Partitioning (SPEC) method is applied to generate a final data partition. This is a powerful method for decomposing an

undirected graph, with good performance being exhibited in many application areas, including protein modeling, information retrieval and identification of densely connected online hyper textual regions. Principally, given a graph G = (V, W), SPEC firstly finds the K largest eigenvectors u1,...,uK of W, which are used to formed another matrix U (i.e., U = (u1,..., uK), whose rows are then normalised to have unit length. By considering the row of U as K-dimensional embedding of the graph vertices, SPEC applies k-means to these embedded points in order to acquire the final clustering result. The ensemble will group the user navigation pattern in the offline phase where the Longest Common Subsequence algorithm is used to predict the user future request in the online phase.

Prediction engine: The main objective of prediction engine in this part of architecture is to classify user navigation patterns and predicts users' future requests. From the ensemble clustering results, a set of clusters C is obtained. Initially, the pages in active session window are sorted based on values stored in the co-occurrence matrix M. During prediction, the system attempts to find the cluster with highest degree of ensemble classifier in respect to sequence. When the prediction engine finds more than one cluster based on ensemble classifier algorithm, then the prediction engine selects a cluster in such a way that, if the difference between positions of last elements of longest path founded in the cluster and the position of first element of this sequence is minimized, the system chooses this cluster. This study uses the Ensemble classifier algorithm during prediction. The main aim of ensemble classifier is to find the majority of voting from all sequences in a set of sequences. For this purpose the ensemble clustering and classification models is combined as given in Fig. 6.

The above Fig. 6 shows the Prediction of User navigation patterns using Ensemble model of Clustering and Classification from web log data. Pre-processed Weblog data are given as an input to the a clustering process that groups potential users with similar interest and by using the results of ensemble model of Clustering and classification the result can predict future user requests. Clustering and classification is used to first group the same characteristics navigation pattern for accessing web page and secondly classifier is used to predict the future request from the cluster group. The ensemble model has been used to lead an effective improvement of the results of clustering and classification algorithm.

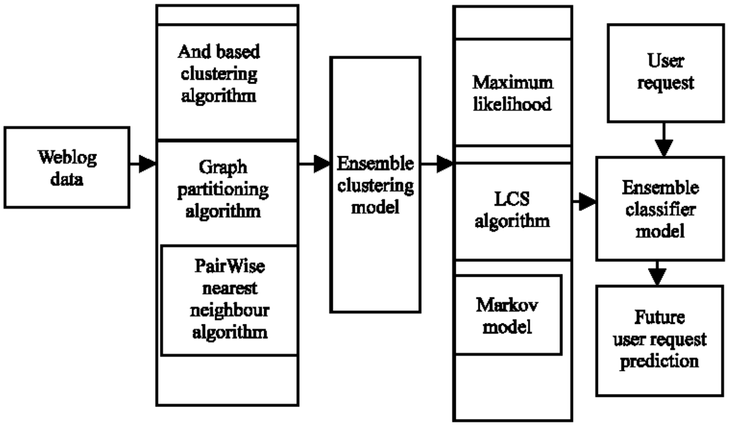

And based clustering algorithm
Weblog data
Graph partitioning algorithm
PairWise nearest neighbour algorithm
Ensemble clustering model
Maximum likelihood
LCS algorithm
Markov model
User request
Ensemble classifier model
Future user request prediction

Fig. 6: Ensemble model for user navigation pattern

EXPERIMENTAL RESULTS

This dataset has preprocessed web logs of the site www.microsoft.com. It records 34,000 randomly selected anonymous users of the site of which 32,711 are used for preprocessing out of which 11,381 weblog data are preprocessed and are used to find the potential user with 5000 are used for testing the ensemble model for clustering and classification. The Graph partition clustering algorithm as offline phase and LCS as online phase has obtained 73% to predict the next user request. In this experimental result ensemble model of clustering and classification is used to obtain an increase of 0.87% of accuracy in the prediction of user next request when the threshold value was 0.9.

Clustering results: Clustering results provide with various forms of knowledge extracted from the log data. These include number of visits made to a single webpage, webpage traffic, most frequently viewed page and navigation pattern of the users. The web log data contained 15 unique web pages which are assigned codes for clarity. The number of visits made by the browsers in 24 hours to these 15 pages is presented in Fig. 7 number of clusters found is another parameter that was used to analyze the performance of the clustering algorithm.

The MinFreq is the threshold in Fig. 8 used to remove low correlated edges and MinClustersize is assumed to be 1. From the graph, it can be understood that the threshold value 0.5 seems optimal

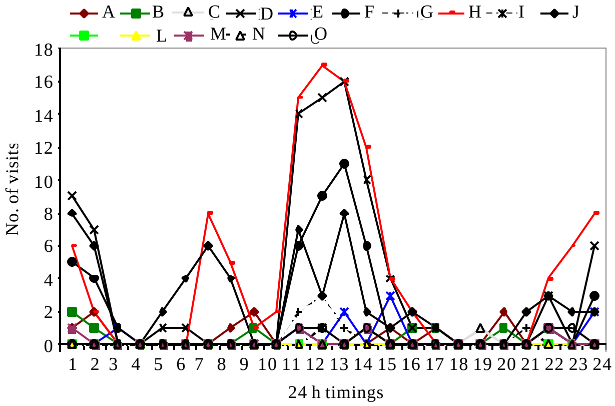


Fig. 7: Twenty four hours page visit details

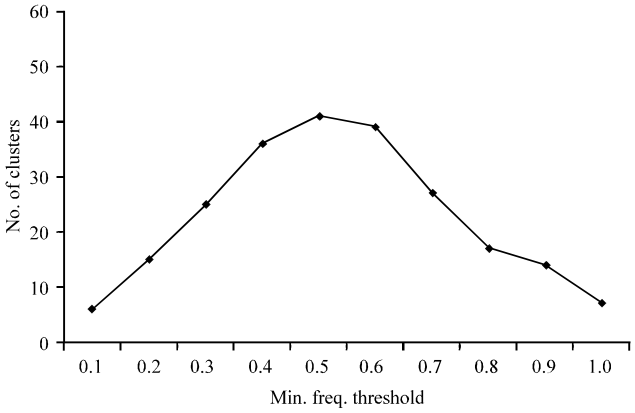


Fig. 8: Effect of Thresholding on number of clusters

for the dataset tested. The test was repeated with varying size of web log data and similar results were found. The clusters resulting when threshold value was set to 0.5 was used during the prediction step.

Prediction results: The performance of the prediction engine was evaluated using three performance parameters, namely, accuracy, coverage and F1 Measure. The navigation patterns are identified from the clusters generated from the previous step and each pattern is divided into two sets. The first set is used for generating prediction and the second set is used to evaluate the predictions. Let as$_{np}$ denote the navigation pattern obtained for the active session's' and let T be a threshold value. The prediction set is denoted as P (as$_{np}$, T) and the evaluation set is denoted as eval$_{np}$. The three parameters can then be calculated using Eq. 5, 6 and 7 and the results are projected in Fig. 10(a-c), respectively:

$$\text{Accuracy} = \frac{|\,P(as_{np},T)\cap eval_{np}\,|}{|\,P(as_{np},T)\,|} \tag{5}$$

$$\text{Coverage }(P\ (as_{np}\ T) = \frac{|\,P(as_{np},T)\cap eval_{np}\,|}{|\,eval_{np}\,|} \tag{6}$$

$$\text{F1 }(P\ (as_{np}, T) = \frac{2\times \text{Accuracy}(P(asnp,\ T))\times \text{Coverage}(P(asnp,\ T))}{\text{Accuracy}(P(asnp,\ T)) + \text{Coverage}(P(asnp,\ T))} \tag{7}$$

From the below Fig. 9, 10 and 11 it could be understood the accuracy of prediction increases with increase in threshold. In the present study, the best accuracy obtained is 0.87% when the threshold value was 0.9.

Figure 9 proves that increase in the size of Live Session window (LSW) increases the accuracy and results shown in Fig. 10 proves that increase in the size of LSW coverage decreases and Fig. 11 shows increases more than minimum Threshold i.e., equal to 3 that means website requires change in sequence of their arrangement of web pages it means site requires to reshuffle the arrangement of web pages.

The prediction model (Fig. 12) that combines hybrid clustering and hybrid classification provides the maximum accuracy when compared with Ant based clustering, Graph partitioning

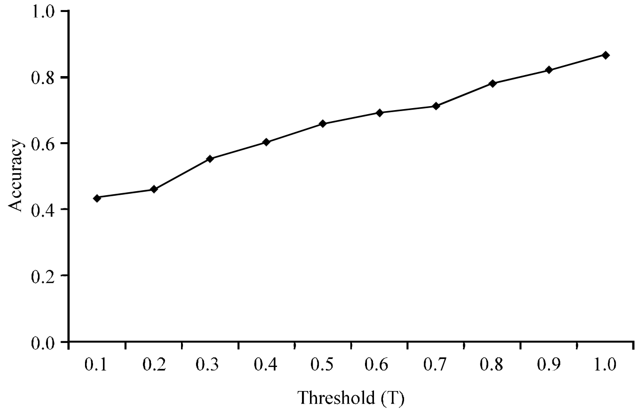


Fig. 9: Prediction accuracy with threshold value from 0.1-1.0

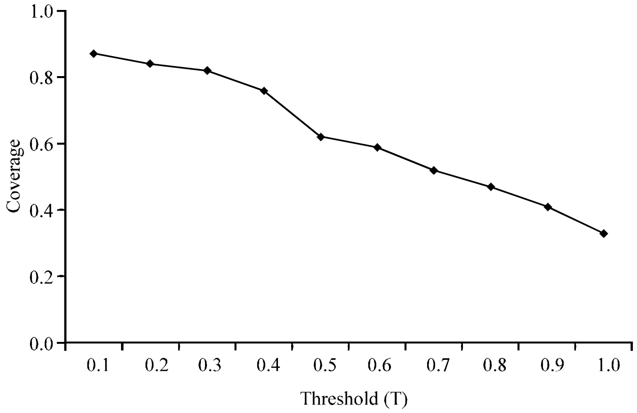


Fig. 10: Prediction coverage with threshold value from 0.1-1.0

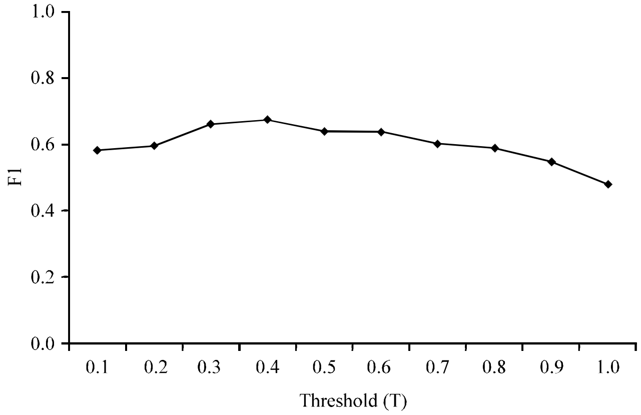


Fig. 11: Prediction measure using F1 value with threshold value from 0.1-1.0

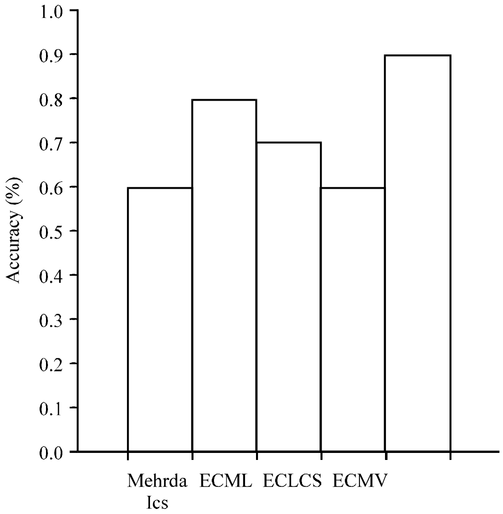


Fig. 12: Prediction accuracy for ensemble model for clustering and classification

and Pairwise Nearest Neighbour algorithm and three classification algorithm like Maximum Likelihood algorithm, Longest Common Subsequence algorithm and Markov Chain Model.

CONCLUSION

In this study, a usage navigation pattern prediction system was presented. The system consists of four stages. The first stage is the cleaning stage, where unwanted log entries were removed. In the second stage, cookies were identified and removed. The result was then segmented to identify potential users. From the potential user, an ensemble clustering model was used to discover the navigation pattern. An ensemble classifier was then used to predict future requests. The experimental results prove that the proposed amalgamation of techniques is efficient both in terms of clustering and classification. In future, the proposed work will be compared with existing systems to analyze its performance efficient. Plans in the direction of using association rules for prediction engine are also under consideration.

REFERENCES

Buzikashvili, N., 2007. Sliding window technique for the web log analysis. Proceedings of the 16th International Conference on World Wide Web, May 8-12, 2007, New York, USA., pp: 1213-1214.

Cooley, R., B. Mobasher and J. Srivastava, 1997. Grouping web page references into transactions for mining world wide web browsing patterns. Proceedings of the IEEE Knowledge and Data Engineering Exchange Workshop, November 4, 1997, Newport Beach, CA., USA., pp: 2-9.

Eirinaki, M. and M. Vazirgiannis, 2003. Web mining for web personalization. ACM Trans. Internet Technol., 3: 1-27.

Fern, X.Z. and C.E. Brodley, 2004. Solving cluster ensemble problems by b ipartite graph partitioning. Proceedings of the 21st International Conference Machine Learning, July 2004, Banff, Canada, pp: 36-43.

Fischer, B. and J.M. Buhmann, 2003. Bagging for path-based clustering. IEEE Trans. Pattern Anal. Mach. Intell., 25: 1411-1415.

Jalali, M., M. Mustapha, A. Mamat and M.N.B. Sulaiman, 2008. A new clustering approach based on graph partitioning for navigation patterns mining. Proceedings of the 19th International Conference on Pattern Recognition, December 8-11, 2008, Tampa, FL., pp: 1-4.

Kosala, R. and H. Blockeel, 2000. Web mining research: A survey. ACM SIGKDD Explorat. Newslett., 2: 1-15.

Lee, Y.S. and S.J. Yen, 2008. Incremental and interactive mining of web traversal patterns. Inform. Sci., 178: 287-306.

Masand, B. and M. Spiliopoulou, 2000. Web Usage Analysis and User Profiling. Springer, New York, USA., ISBN-13: 9783540678182, pp: 1-6.

Nasraoui, O., M. Soliman, E. Saka, A. Badia and R. Germain, 2008. A web usage mining framework for mining evolving user profile in dynamic web sites. IEEE Trans. Knowl. Data Eng., 20: 202-215.

Srivastava, J., R. Cooley, M. Deshpande and P.N. Tan, 2000. Web usage mining: Discovery and applications of usage patterns from web data. ACM SIGKDD Explorat., 1: 12-23.

Sujatha, V. and Punithavalli, 2012. Improved user navigation pattern prediction technique from web log data. Procedia Eng., 30: 92-99.

Suneetha, K.R. and R. Krishnamoorthi, 2010. Classification of web log data to identify interested users using decision trees. Proceedings of the International Conference on Computing Communications and Information Technology Applications, January 21-23, 2010, Coimbatore, India.