Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Engineering 30 (2012) 92 – 99

Procedia Engineering

www.elsevier.com/locate/procedia

International Conference on Communication Technology and System Design 2011

IMPROVED USER NAVIGATION PATTERN PREDICTION TECHNIQUE FROM WEB LOG DATA

V. SUJATHA[a], PUNITHAVALLI[b], a*

[a]*Department of Computer Science,CMS College of Science and Commerce,Coimbatore, India*
[b]*Department Computer Application, SNS Arts and Science College,Coimbatore, India*

Abstract

Web Usage Mining (WUM) is the automatic discovery of user access pattern from web servers. Organizations collect large volumes of data in their daily operations, generated automatically by web servers and collected in server access logs. This paper presents the Prediction of User navigation patterns using Clustering and Classification (PUCC) from web log data. In the first stage PUCC focuses on separating the potential users in web log data, and in the second t 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 for user navigation pattern in web usage mining systems. These results can be use for predicting user’s next request in the huge web sites

Keywords: Web usage mining; Navigation pattern; classification; weblog; clustering; Graph partitioning.

1. INTRODUCTION

Web Mining can be broadly divided into three distinct categories, according to the kinds of data to be mined. They are web content mining, web usage mining and web structure mining. Many web analysis tools exist but they are limited and the efficiency of these tools is still to reach the state of perfection. Clustering and classification are two active areas of machine learning research that is proving to be promising to help with this problem. 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 paper, a solution to

* V.Sujatha. Tel.: +91-9843416962.
E-mail address: sujatha.padmakumar@rediffmail.com.

doi:10.1016/j.proeng.2012.01.838

predict user request from navigation pattern is proposed. A general process of web mining consists of four different stages (Figure 1).

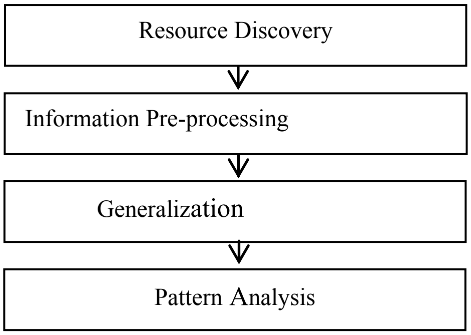


Fig 1: Process of Web Mining

They are (i) resource discovery, (ii) information pre-processing, (iii) generalization, and (iv) pattern analysis. Resource Discovery is task of retrieving information from web resources and documents. Information pre-processing is the transform process of the result of resource discovery. Generalization is used to uncover general patterns from individual and across multiple sites. In this step, machine learning and traditional data mining techniques are typically used. Pattern Analysis is the validation of the mined patterns. Analyzing user navigation pattern is a problem domain under web usage mining, where the general access pattern is tracked from web logs to understand access patterns and usage trends. These analyses can shed light on better structure and grouping of resource providers

The main objective of the proposed system 'Predicting User navigation patterns using Clustering and Classification from web log data (PUCC) is to predict user navigation patterns using knowledge from (i) a Classification process that identifies potential users from web log data and (ii) a clustering process that groups potential users with similar interest and (ii) Using the results of classification and clustering, predict future user requests. For this purpose, amalgamations of various techniques like, pre-processing, classification, clustering, weight-based prediction model are used and are explained in the next section. The rest of the paper is organized as below. Section 3 explains the general framework of the proposed system. Section 4 presents the results obtained while Section 5 concludes the work with future research directions.

2. RELATED WORK

Identifying Web browsing strategies is a crucial step in Website design and evaluation, and requires approaches that provide information on both the extent of any particular type of user behavior and the motivations for such behavior [9].Pattern discovery from web data is the key component of web mining and it converge algorithms and techniques from several research areas. Baraglia and Palmerini (2002) proposed a WUM system called SUGGEST that provide useful information to make easier the web user navigation and to optimize the web server performance. Liu and Keselj (2007) proposed the automatic classification of web user navigation patterns and proposed a novel approach to classifying user navigation patterns and predicting users' future requests and Mobasher (2003) presents a Web Personalizer system which provides dynamic recommendations, as a list of hypertext links, to users. Jespersen et al. (2002) [10] proposed a hybrid approach for analyzing the visitor click sequences. Jalali et al. (2008a [7] and 2008b [8]) proposed a system for discovering user navigation patterns using a graph partitioning model. An undirected graph based on connectivity between each pair of Web pages was considered and weights were assigning to edges of the graph. Dixit and Gadge (2010) [5] presented another user navigation pattern mining system based on the graph partitioning. An undirected graph based on connectivity between Referrer and URI pages was presented along with a pre-processing method to process unprocessed web log file and a formula for assigning weights to edges of the undirected graph.

3. PUCC SYSTEM

The general architecture of the proposed PUCC system is given in Figure 2. The heart of the PUCC 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). Each entry in the web log data include details like the IP address of the computer making the request, user ID, date and time of the request, a status field indicating if the request was successful, size of the file transferred, referring URL (URL of the page which contains the link that generated the request), name and version of the browser being used.

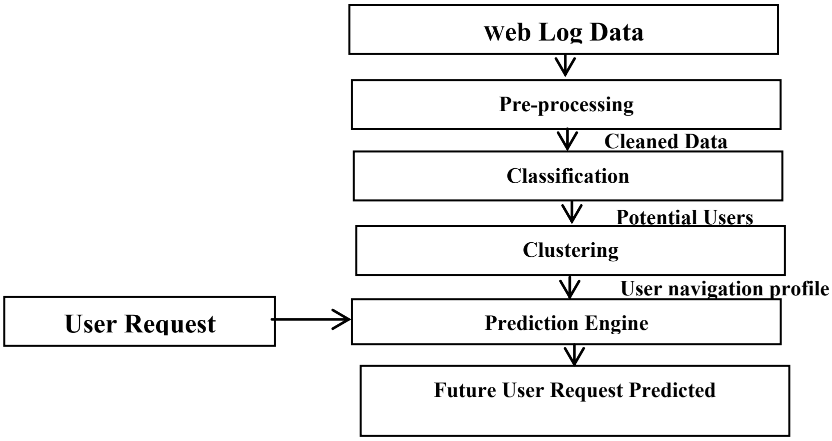


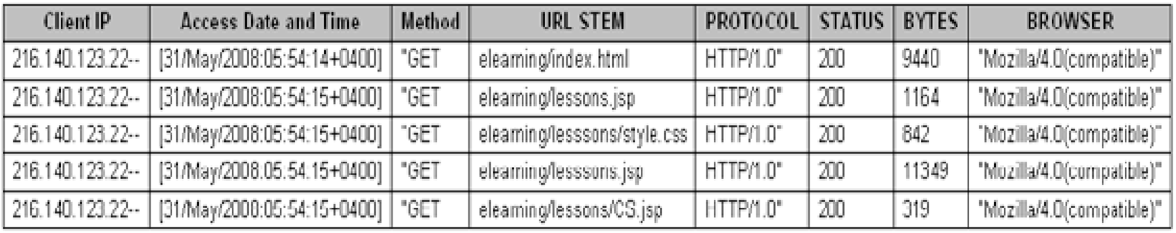
Fig 2: PUCC Model

3.1 Weblog files

Web log file is log file automatically created and maintained by a web server. Every "hit" to the Web site, including each view of a HTML document, image or other object, is logged. The raw web log file format is essentially one line of text for each hit to the web site. At least two log file formats exists: Common Log File format (CLF) and Extended Log File format ([16] for more details), robots do not follow the proposed standard. Thus to delete robot entries the following procedure is used. A sample web log file is shown in Figure 3. The information in web log file represent the navigation patterns of different segments of the overall web traffic, ranging from single-user, single-site browsing behavior to multi-user, multi-site access patterns. Irrespective of the source of collection, the web log file has the following general characteristics.

- The log file is text file. Its records are identical in format.
- Each record in the log file represents a single HTTP request.
- A log file record contains important information about a request: the client side host name or IP address, the date and time of the request, the requested file name, the HTTP response status and size, the referring URL, and the browser information.

A browser may fire multiple HTTP requests to Web server to display a single Web page. This is because a Web page not only needs the main HTML document; it may also need additional files, like images and JavaScript files. The main HTML document and additional files all require HTTP requests.



Client IP	Access Date and Time	Method	URL STEM	PROTOCOL	STATUS	BYTES	BROWSER
216.140.123.22--	[31/May/2008:05:54:14+0400]	"GET	elearning/index.html	HTTP/1.0"	200	9440	"Mozilla/4.0(compatible)"
216.140.123.22--	[31/May/2008:05:54:15+0400]	"GET	elearning/lessons.jsp	HTTP/1.0"	200	1164	"Mozilla/4.0(compatible)"
216.140.123.22--	[31/May/2008:05:54:15+0400]	"GET	elearning/lesssons/style.css	HTTP/1.0"	200	842	"Mozilla/4.0(compatible)"
216.140.123.22--	[31/May/2008.05.54.15+0400]	"GET	elearning/lesssons.jsp	HTTP/1.0"	200	11349	"Mozilla/4.0(compatible)"
216.140.123.22--	[31/May/2008:05:54:15+0400]	"GET	elearning/lessons/CS.jsp	HTTP/1.0"	200	319	"Mozilla/4.0(compatible)"

Fig 3: Sample Web Log File

3.2 Pre-processing

The first step of PUCC is the pre-processing of web log data, where the unformatted log data is converted into a form that can be directly applied to mining process. 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.

3.3 Identification of Potential Users

This step of PUCC focuses on separating the potential users from others. Suneetha and Krishnamoorthy (2010) used decision tree classification using C4.5 algorithm to identify interested users. They use a set of decision rules for this purpose. The algorithm worked efficiently in identifying potential users, but had the drawback that it completely ignored the entries made by network robots. Search engines normally use network robots to crawl through the web pages to collect information. The number of records created by these robots in a log file is extremely high and has a negative impact while discovering navigation pattern. This problem is solved in this paper by identifying the robot entries first before segmenting the user groups into potential and not-potential users.

According to Yu *et al*. (2005), entries in web log made by network robots can be identified by their IP address and agents. But this might require knowledge on all type of agents and search engines, which is difficult to obtain. An alternative way is to study the robots.txt file (located at the website's root directory), as a network robot must always read this file before accessing the website. This is because the robots.txt has the access details of the website and each robot is request to know its access right before scrawling. But this cannot be always relied on since compliance to robot exclusion standard is voluntary and most of the

- Detect and remove all entries which has accessed robots.txt file
- Detect and remove all entries with visiting time of access as midnight (commonly used as the network activity at that time is light)
- Remove entry when access mode is HEAD instead of GET or POST
- Compute browsing speed and remove all entries whose speed exceeds a threshold T1 and number of visited pages exceeds a threshold T2.

The browsing speed is calculated as the number of viewed pages / session time. After handling the network robot entries, a series of decision rules are applied to group the users as potential and not-potential users. Given a set of training data containing valid log attributes, C4.5 classification algorithm is used to classify the users. The attributes selected are time (>30 seconds), number of pages referred in a session (Session time=30 minutes) and the access method used. The decision rule for identifying potential users is "If Session Time > 30 minutes 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.

3.4 Clustering Process

This paper uses a graph partitioned clustering algorithm to group users with similar navigation pattern (Jalali *et al.*, 2005). An undirected graph based on the connectivity between each pair of web pages is used. Each edge in the graph is assigned a weight, which is based on the connectivity time and frequency. Connectivity Time measures the degree of visit ordering for each two pages in a session.

(Equation 1).
$$TC_{a,b} = \frac{\sum_{i=1}^{N} \frac{T_i}{T_{ab}} \text{ x } \frac{f_a(k)}{f_b(k)}}{\sum_{i=1}^{N} \frac{T_i}{T_{ab}}} \qquad (1)$$

T_i is the time duration of i^th^ session that contain both a and b pages, T_{ab} is the difference between requested time of page a and page b in the session, f(k)=k if web page appears in position k. Frequency measures the occurrence of two pages in each sessions (Equation 2).

$$FC_{a,,b} = \frac{N_{ab}}{\mathrm{Max}\{N_a, N_b\}} \qquad (2)$$

Where N_{ab} is the number of sessions containing both page a and page b. N_a and N_b are the number of session containing only page a and page b. Both the formulas normalize all values for time and frequency are between 0 and 1. Both these are considered as two indicators of the degree of connectivity for each pair of web pages and is calculated using Equation (3).

$$W_{a,b} = \frac{2 \text{ x } TC_{ab} \text{ x } FC_{ab}}{TC_{ab} + FC_{ab}} \qquad (3)$$

The data structure can be used to store the weights is an adjacency matrix M where each entry M_{ab} contains the value W_{ab} computed according to (3) .To limit the number of edge in such graph ,element of M_{ab} whose value is less than a threshold are too little correlated and thus discarded. This threshold is named as MinFreq in this contribution.

3.5. Prediction engine

The main objective of prediction engine in this part of architecture is to classify user navigation patterns and predicts users' future requests. The below table shows the effect of pre-processing. This paper uses the Longest Common Subsequence algorithm during prediction. The main aim of LCS is to find the longest subsequence common to all sequences in a set of sequences. This method is discussed in this section. The algorithm works with two features.

- The first property states that if two sequences X and Y both end with the same element, then their LCS will be found by removing the last element and then finding LCS of the shortened sequence.
- The second property is used when the two sequences X and Y does not end with the same symbol. Then, the LCS of X and Y is the longest sequence of LCS (X_n, Ym_{-1}) and LCS (X_{n-1}, Ym).

Thus, the LCS can be formulated using Equation (4).

$$LCS(X_i, Y_i) = \begin{cases} 0 & \textit{if } \mathrm{i=0 \text{ or } j=0} \\ (LCS(X_{i-1}, Y_{j-1}), x_i) & \text{if } \mathrm{x}_i = y_i \\ longest(LCS(X_i, Y_{j-1}), LCS(X_{i-1}, Y_j)) & \text{if } \mathrm{x}_i \neq y_i \end{cases} \qquad (4)$$

```
L[p] = P; // Assign all URLs to a list of web pages.
For each (P_i, P_j) ∈ L[p] do   //for all pair of web pages
 M(i, j) = Weight Formula(P_i, P_j);   //computing the weight based on Equation (3)
 Edge (i, j_) = M (i, j); End for For all Edge (u, v) ∈ Graph (E, V) do
 //removing all edges that its weight is below than MinFreq
   If Edge (u, v) < MinFreq then     Remove (Edge (u, v));
 End if    End for for all vertices (u) do Cluster[i] = DFS (u);
// perform DFS If cluster[i] < MinClusterSize
//remove cluster whose length is below MinClusterSize
Remove (Cluster[i]); End if i = i + 1   end for return (Cluster)
```

Fig 4: Graph Based Clustering Algorithm

To find the longest subsequences common to X_i and Y_j, the elements x_i and y_i are compared. If equal, then the sequence LCS (X_{i-1}, Y_{j-1}) is extended by that element, x_i. If they are not equal, then the longer of the two sequences, LCS (X_i, Y_{j-1}), and LCS (X_{i-1}, Y_j), is retained. If they are both of the same length, but are not identical, then both are retained. A worked out example of LCS is given in http://en.wikipedia.org/wiki/Longest_common_subsequence_problem.

4. EXPERIMENTAL RESULTS

4.1. Pre-processing results

Table I shows the effect of cleaning log data in terms of number of transactions and amount of memory required to store it.

Table: 1 Effect of Pre-processing

No. of Transactions	7000	1420	11987	4320	34000	11381
Memory Used (MB)	1.76	0.48	1.95	0.61	2.47	1.12

4.2. Classification

Figure 5 shows the results after identifying potential users. The log data obtained after cleaning was used for classification.

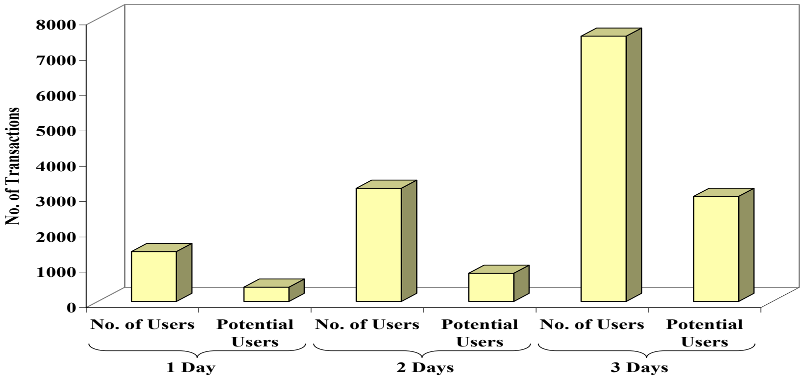


Fig 5: Effect of Pre-processing

4.3. 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 behavior 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 Figure 6.

number of clusters found is another parameter that was used to analyze the performance of the clustering algorithm.

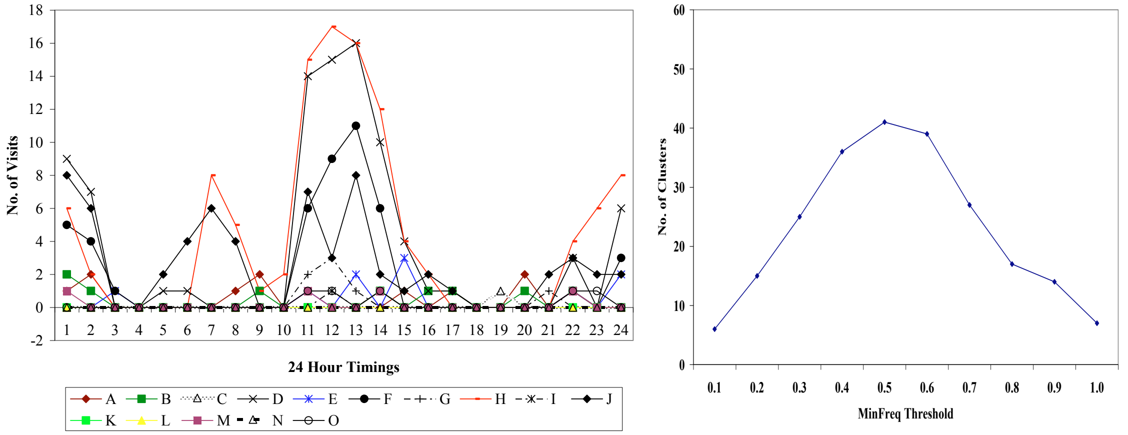


Fig 6: (a) 24 Hours Page Visit Details (b) Effect of Thresholding on number of clusters

The MinFreq is the threshold 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 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.

4.4. 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 Equations (7), (8) and (9) and the results are projected in Figures 9, 10 and 11 respectively.

$$\text{Accuracy} = \frac{|P(as_{np}, T) \cap eval_{np}|}{|P(as_{np}, T)|} \quad (7) \qquad \text{Coverage}(P(as_{np}, T)) = \frac{|P(as_{np}, T) \cap eval_{np}|}{|eval_{np}|} \quad (8)$$

$$F1(P(as_{np}, T)) = \frac{2x\,Accuracy(P(asnp,T))\ \ x\ \ Coverage(P(asnp,T))}{Accuracy(P(asnp,T)) + Coverage(P(asnp,T))} \quad (9)$$

From the figures, 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

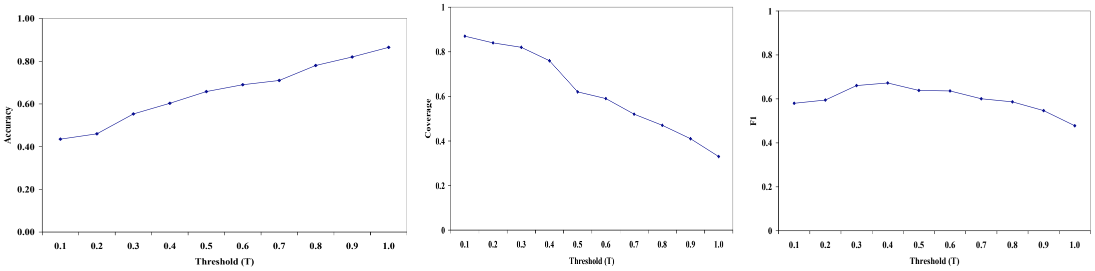


Fig 9: (a) Prediction Accuracy (b) Prediction Coverage (c) Prediction Measure using F1 Value

5. CONCLUSION

In this paper, 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, a graph partitioned clustering algorithm was used to discover the navigation pattern. An LCS classification algorithm 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.

6. REFERENCES

[1] Cat ledge, L. and Pitkow, J., "Characterizing browsing behaviours on the World Wide Web", *Computer Networks and ISDN Systems*,1995, Vol. 27, No. 6, Pp. 1065-1073.

[2] Cooley, R., Srivastava, J. and Mobasher, B. , "Web mining: Information and pattern discovery on the world wide web, Tools with Artificial Intelligence", *Ninth IEEE International Conference on In Tools with Artificial Intelligence*, 1997. Proceedings., Vol. 10, pp. 0558-567.

[3] Eirinaki, M. and Vazirgiannis, M. , "Web mining for web personalization"**,** *ACM Transactions on Internet Technology* (TOIT), 2003, Vol. 3, Issue 1, Pp. 1-27.

[4] http://en.wikipedia.org/wiki/Longest_ common_subsequence_problem, Last Accessed on 27-02-2011.

[5] Huysmans, J., Baesens, B. and Vanthienen, J.), "Web Usage Mining: A Practical Study", *Katholieke Universities Leuven*, Dept. of Applied Economic Sciences (2003).

[6] Inktomi, A, " Web surpasses one billion documents", www.inktomi.com/new/press/billion.html. 2000.

[7] Jalali, M., Mustapha, M., Mamat, A. and Sulaiman, M.N.B. , "A new clustering approach based on graph partitioning for navigation patterns mining", 9th *International Conference on Pattern Recognition*, Pp. 1-4.

[8] Kosala, R. and Blockeel, H. , "Web mining research: A survey. SIGKDD Esplorations", 2000, Vol. 2, No. 1, Pp.1-15.

[9] Kumar, P.R. and Singh, A.K. , "Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval", American Journal of Applied Sciences, 2010, Vol. 7, No.6, Pp. 840-845.

[10] RFC 1413 (2010) Identification Protocol, http://www.rfceditor.org/rfc/ rfc1413.txt.

[11] Schenker, A., Kandel, A., Bunke, H., and Last, M., "Graph-theoretic techniques for web content mining", *Series in Machine Perception and Artificial Intelligence*, 2005, Vol. 62, P. 1.

[12] Srivastava, J., Cooley, R., Deshpande, M. and Tan, P.N. , "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", *SIGKDD Explorations, ACM SIGKDD*, 2000, Vol. 1, Issue 2, Pp. 12-23.

[13] Suneetha, K.R. and Krishnamoorthi, R., "Classification of web log data to identify interested users using decision trees", *International Conference on Computing, Communications and Information Technology Applications*, (CCITA 2010), Coimbatore, India.

[14] WCA (2010) Web characterization terminology and definitions, http://www.w3.org/ 1999/05/WCA-terms.

[15] Yu, J., Ou, Y., Zhang, C. and Zhang, S. (2005) Identifying Interesting visitors through web log classification, IEEE Computer Society, Pp. 55-59.