Volume 1, No. 1, August 2010

Journal of Global Research in Computer Science

RESEARCH PAPER

Available Online at www.jgrcs.info

Users' Navigation Pattern Discovery using Ant Based Clustering and LCS Classification

Mrs. K. Devipriyaa*[1] and Dr. (Mrs.) B.Kalpana[2]
*[1]Avinashilingam Deemed University for Women, Coimbatore, India.
Email: krdevipriyaa@gmail.com
[2]Avinashilingam Deemed University for Women, Coimbatore, India.
Email: kalpanabsekar@gmail.com

***Abstract*:** Web Usage Mining (WUM) is the process of extracting knowledge from Web user’s access data by exploiting Data Mining techniques. It mines the secondary data (web logs) derived from the users' interaction with the web pages during certain period of Web sessions. In the present research work, a hybrid method is proposed, which uses the ant-based clustering and LCS classification methods to find and predict user’s navigation behaviour. The proposed system works in two phases, (i) the offline phase and (ii) the online phase. The offline phase takes care of preprocessing and clustering, while the classification and prediction is performed during the online phase. Ant-based clustering method used to discover or extract user’s navigational pattern from web log files. The LCS classification algorithm uses the knowledge from offline stage and predicts the users’ next request.

***Keywords*:** web usage mining; pre-processing; pattern discovery; pattern analysis; ant-based clustering; LCS classification

INTRODUCTION

Web mining is the application of data mining techniques to web-based data for the purpose of learning or extracting knowledge. Based on several researches done in the area of web mining, we can broadly classify it into three domains: web content mining, web structure mining, and web usage mining. Web usage information mining could help to engage new customers, maintain current customers, track customers who are leaving web site, and so on [1]. A general web usage mining system (Figure 1) consists of three steps, namely, preprocessing, pattern discovery and pattern analysis.

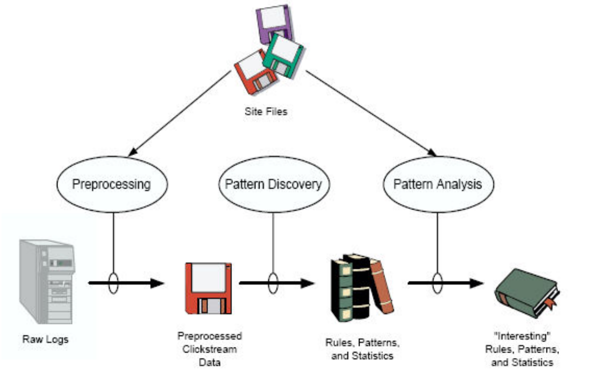

Site Files
Preprocessing
Pattern Discovery
Pattern Analysis
Raw Logs
Preprocessed Clickstream Data
Rules, Patterns, and Statistics
"Interesting" Rules, Patterns, and Statistics

Figure 1: Knowledge Discovery

Data preprocessing is responsible for converting the usage, content, and structure information contained in the web log file into a format that is suitable for pattern discovery. Pattern discovery draws upon methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. Pattern analysis is the last step in the overall Web Usage mining process. The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase.

A clustering algorithm discovers groups in the set of documents such as documents within a group or more similar than document across groups. The clustering method used in an ant-based clustering method[12], where artificial ants act as agents, which communicate and influence themselves through the configuration of objects on the floor. Thus, the agents construct groups of similar objects or construct clusters. The classification algorithm based on Longest Common Sequence algorithm is used [9]. The main aim of this algorithm is to use the knowledge from clustering stage and predict the users’ next request. It uses a weight matrix to calculate the LCS and the path with least weight is chooses to predict the next request.

In the following section we give an overview over the related work. Section 3 explains the methodology in more detail. Section 4 goes into detail how we implement the proposed method i.e. the experimental procedure of the proposed method and results are shown. We concluded our work in Section 5.

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) [12] 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)[13] proposed the automatic classification of web user navigation patterns and proposed a novel approach to classifying user navigation patterns and predicting users’ future requests. Mobasher *et al.* (2000) [15] and Nakagawa and Mobasher (2003) [16] presents a WebPersonalizer 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 preprocessing method to process unprocessed web log file and a formula for assigning weights to edges of the undirected graph. Ant-based clustering due to its flexibility and self-organization has been applied in a variety of areas from problems arising in e-commerce to circuit design, and text-mining to web-mining, etc (Jianbin *et al*., 2007) [11]. The various works proposed in this area with particular emphasize on web usage mining, clustering and classification was provided in this section. In this present work, research work is one another attempt made to propose a hybrid system that uses clustering and classification methods to discover the user's navigation pattern and analyze them from the server's web log file.

METHODOLOGY

The navigation pattern of an online user often reflects user's mental model and for this reason, websites owners and developers pay more attention to the navigation patterns. In order to study these patterns effectively, browsing patterns gathered from a site is very important. Several solutions have been proposed and the usage of clustering and classification is more frequently used on such solutions. This research work is one another attempt made to propose a hybrid system that uses clustering and classification methods to discover the user's navigation pattern and analyze them from the server's web log file.

Preprocessing

Preprocessing of a web log file simply reformats the entries of a log file into a form that can be used directly by the subsequent steps of the log analyzer. The preprocessing is performed in four steps as given below. (i) Cleaning (ii)User identification (iii) Session identification and (iv) Formatting.

In the **Cleaning step**, (i.e.,) cleaning of data, unwanted data will be deleted. Examples of unwanted data include requests for images, Javascripts, flash animations, video, etc. These data are not required for user navigation and hence are deleted form the log file. The identification of users through the IP address is the most frequently used method as it is simple, easy to capture and is never empty. Session Identification is performed using session timeout value. The original log file is thus pruned from unwanted data and is formatted which is done in the **last step**. This formatted data is taken as input in the next process of web log analysis.

Pattern Discovery

Several methods and techniques have already been developed for this step. Some of the frequently used solutions are statistical analysis, clustering, classification, association rules, sequent patterns and dependency modeling. Etminani *et al.* (2009) [12] used an Ant-based clustering approach to discover navigation patterns for discovering navigation pattern. This method has some issues when combined with log file knowledge discovery. The issues are

(i) One of the critical *problems* of *ant-based clustering* is the high-execution times.

(ii) The second issue is the inability of ant-based algorithm to detect the completion of clustering process.

(iii) Another important issue is that one ant can produce the same results as many ants in the models function like stochastic sampling algorithms and the result is repeated and are considered during clustering.

All the above mentioned problems motivated the present research work to use a classification algorithm, after the ant-based clustering. The classification algorithm used is the Longest Common Subsequence (LCS), proposed by Jalali *et al.[9]*. The study divides the pattern discovery and analysis phase into two phases, the online and offline phase. The offline phase consists of analyzing the log file and producing the clusters. The online phase, on receiving a new user request, the URL and session to which the user belongs are identified and are classified to the correct cluster. Both the ant-based clustering algorithm and the LCS algorithm are explained in the next section.

Pattern Analysis

The final stage is the analysis of the patterns discovered in the previous step. This is done in two stages:

(i) Validation : To identify relevant rules or patterns from which interesting patterns can be discovered

(ii) Interpretation : Mathematical interpretations which can be used by scientists to discover knowledge

Ant-based clustering (offline Phase)

The ant algorithm is mainly based on the version described in Handl and Meyer (2002)[6]. Deneubourg et al. in [4] proposed ant-based clustering and sorting. In the case of ant-based clustering and sorting, two related types of natural ant behaviors are modeled. When clustering, ants gather items to form heaps. And when sorting, ants discriminate between different kinds of items and spatially arrange them according to their properties. Lumer and Faieta in [14] proposed ant-based data clustering algorithm (shown in Figure 2), which resembles the ant behavior described in [4]. As shown in Figure 2, the agents (ants) and data are randomly initialized on a toroidal grid. By moving agents,Data is sorted according to its neighbors. The picking and dropping probabilities, given a grid position and a particular data item i, are computed using the density functions:

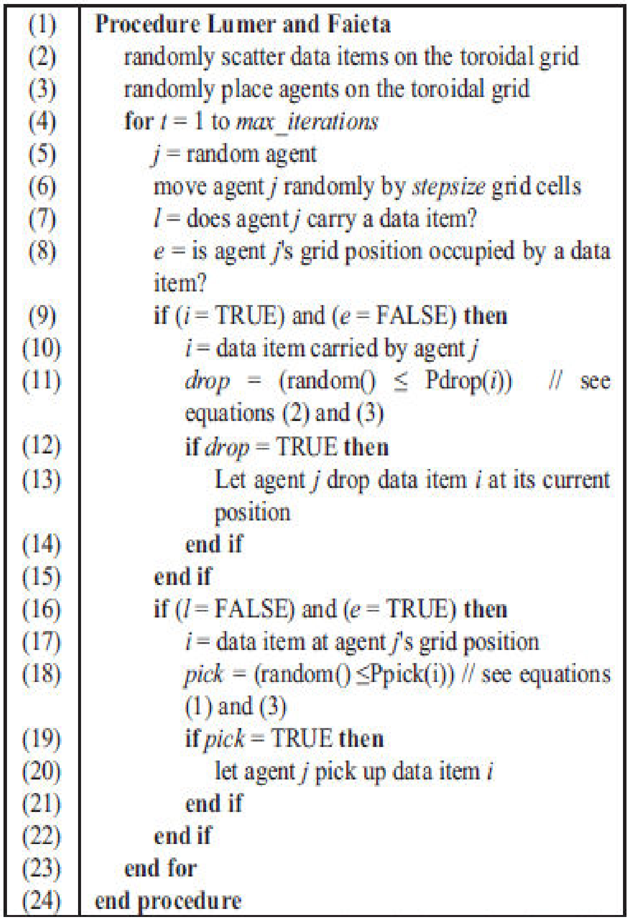
$$p_{\text{pick}}(i) = \left(\frac{k^*}{K^* + f(i)} \right)^2 \qquad (1)$$

$$p_{\text{drop}}(i) = \begin{cases} 2f(i) & \text{if } f(i) < k^- \\ 1 & \text{otherwise} \end{cases} \qquad (2)$$

where k+ and k_ are constants, and f(i) is a neighborhood function:

$$f(i) = \max\left(0, \frac{1}{\sigma^2} \sum_{j \in L} \left(1 - \frac{d(i,j)}{\alpha} \right) \right)$$
(3)

where, d(i, j) _ [0, 1] is a measure of the dissimilarity between data points i and j , __[0, 1] is a data-dependent scaling parameter, and 2 is the size of the local neighborhood L.



```
Procedure Lumer and Faieta
  randomly scatter data items on the toroidal grid
  randomly place agents on the toroidal grid
  for $t$ = 1 to $max\_iterations$
    $j$ = random agent
    move agent $j$ randomly by $stepsize$ grid cells
    $l$ = does agent $j$ carry a data item?
    $e$ = is agent $j$'s grid position occupied by a data item?
    if ($i$ = TRUE) and ($e$ = FALSE) then
      $i$ = data item carried by agent $j$
      $drop$ = (random() ≤ Pdrop($i$))  // see equations (2) and (3)
      if $drop$ = TRUE then
        Let agent $j$ drop data item $i$ at its current position
      end if
    end if
    if ($l$ = FALSE) and ($e$ = TRUE) then
      $i$ = data item at agent $j$'s grid position
      $pick$ = (random() ≤Ppick(i)) // see equations (1) and (3)
      if $pick$ = TRUE then
        let agent $j$ pick up data item $i$
      end if
    end if
  end for
end procedure
```

Figure 2: Ant based algorithm

Online Phase – LCS Approach

During the online phase, when a new request arrives at the server, the URL requested and the session to which the user belongs are identified, the underlying knowledge base is updated, and a list of suggestion is appended to the requested page. The online phase is shown in Figure 3.

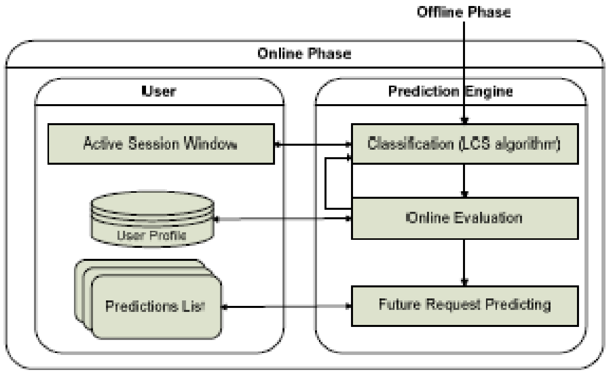


Figure 3: Online Phase

From the clustering results, we have a set of clusters np' = <np_1, np_2, …np_n> where np_i = <P_1, P_2, …, P_k> where k is the set of web pages identified as user navigation patterns and 1 ≤ i ≤ n. Sequence W' = <P_1, P_2, …P_m> is a current active session and m is size of active session window.

Before classifying an active session to construct the prediction list, the pages in active session windows is sorted based on values stored in the co-occurrence matrix M. After this step, for building the prediction list, the system must find the cluster based on LCS algorithm. After applying this algorithm, the system finds a cluster with highest degree of LCS in respect to sequence W'.

When the prediction engine finds more than one cluster based on LCS algorithm, then the prediction engine selects a cluster in such a way that, if the difference between positions of last elements of longest common subsequence founded in the cluster and the position of first element of this sequence is minimized, the system chooses this cluster. In this module, if the first page in the next user activity is different with prediction list, it needs again to classify with new user activities.

RESULTS AND DISCUSSION

In order to test the effectiveness of the proposed system, server web log data file was obtained. The system was tested with several data collected from 90 days for easy discussion, experiments projected here are from one day, that is, data collected on 29-12-2009.

As mentioned in section 3, the preprocessing is conducted in four steps, namely (i) Cleaning (ii) User Identification (iii) Session Identification and (iv) formatting.

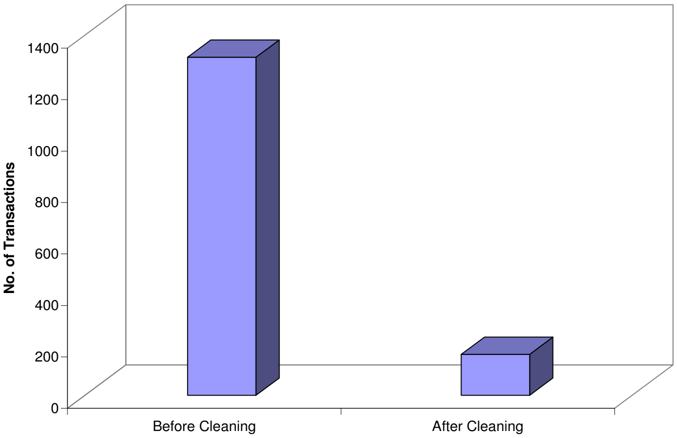
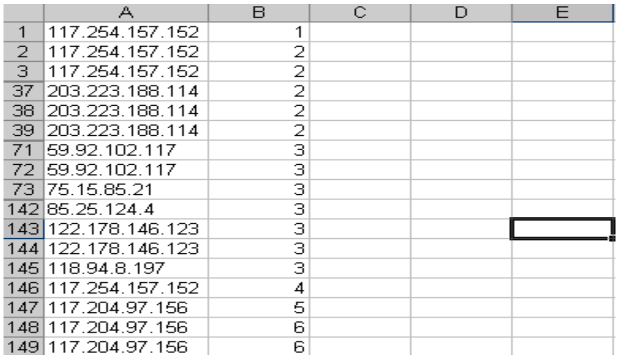


Figure 4: Effect of cleaning step on raw web log file

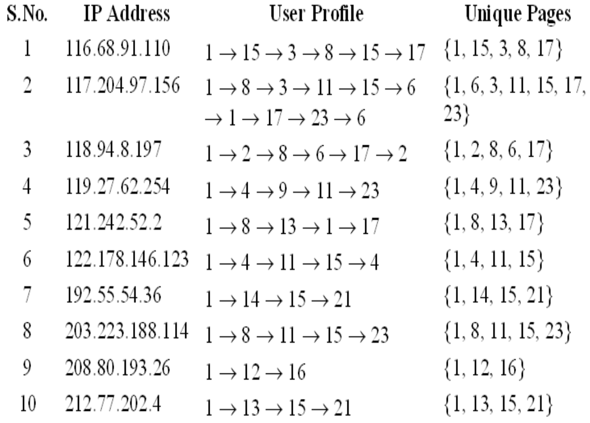
Ant Based clustering is an application for clustering similar interested users into a single class. The result after such grouping is shown in Figure 5.



	A	B	C	D	E
1	117.254.157.152	1			
2	117.254.157.152	2			
3	117.254.157.152	2			
37	203.223.188.114	2			
38	203.223.188.114	2			
39	203.223.188.114	2			
71	59.92.102.117	3			
72	59.92.102.117	3			
73	75.15.85.21	3			
142	85.25.124.4	3			
143	122.178.146.123	3			
144	122.178.146.123	3			
145	118.94.8.197	3			
146	117.254.157.152	4			
147	117.204.97.156	5			
148	117.204.97.156	6			
149	117.204.97.156	6			

Figure 5: Cluster Groups

Knowledge gained from the clustering results include the number of visits made to a single webpage, website traffic, , most frequently viewed pages cluster information, the users navigation behaviour, etc. The user profile and path accessed by the users are shown in figure 6.



S.No.	IP Address	User Profile	Unique Pages
1	116.68.91.110	1 → 15 → 3 → 8 → 15 → 17	{1, 15, 3, 8, 17}
2	117.204.97.156	1 → 8 → 3 → 11 → 15 → 6 → 1 → 17 → 23 → 6	{1, 6, 3, 11, 15, 17, 23}
3	118.94.8.197	1 → 2 → 8 → 6 → 17 → 2	{1, 2, 8, 6, 17}
4	119.27.62.254	1 → 4 → 9 → 11 → 23	{1, 4, 9, 11, 23}
5	121.242.52.2	1 → 8 → 13 → 1 → 17	{1, 8, 13, 17}
6	122.178.146.123	1 → 4 → 11 → 15 → 4	{1, 4, 11, 15}
7	192.55.54.36	1 → 14 → 15 → 21	{1, 14, 15, 21}
8	203.223.188.114	1 → 8 → 11 → 15 → 23	{1, 8, 11, 15, 23}
9	208.80.193.26	1 → 12 → 16	{1, 12, 16}
10	212.77.202.4	1 → 13 → 15 → 21	{1, 13, 15, 21}

Figure 6 Extracted navigation patterns

To predict the user's next request, LCS classification algorithm was used. The LCS finds the longest navigation sequence cluster that matches with the user's referral URL

For example, Table I shows the navigation pattern of four users belonging to the same cluster, constructed over 3 sessions.

Table I: Sample pattern

IP Address	URL Navigation Pattern
117.204.97.156	3 5 6
192.55.54.36	3 6 7 8 9 10 11
59.92.110.200	2 6 12 13 14 15 16 17 18 19
121.242.52.2	3 6 9 8
122.162.209.77	3 6
116.68.91.110	2 3 6 8 11 20 24
122.160.76.157	6 8 28
118.94.8.197	3 6 8
88.80.205.215	3 6 8
122.178.146.123	3 6 8 31 32

In order to make pattern analysis and prediction, the LCS algorithm calculates a weight matrix with each pattern discovered. The weight W for edge E can be computed as:

$$W_{ij} = \frac{N_{ij}}{\max \{N_i, Nj\}} \qquad (1)$$

According to the LCS rules, the weight with the lowest value predicts the next request more accurately. The system was tested in similar fashion for different threshold values and the accuracy was calculated according to Equation (4.2).

$$\text{Accuracy} = \frac{\text{Total number of correct predictions}}{\text{Total Number of records}} \text{x100} \qquad (2)$$

The accuracy of the system while testing with threshold values ranging from 0.1 to 1.0. From the results, it is apparent that the proposed ant-based clustering when combined with LCS produces better result when compared to the Jalali's system [9], which uses graph partitioning clustering algorithm and LCS algorithm. The result shows a trend that when the threshold value increases, the accuracy also increases and the maximum accuracy achieved by the proposed system is 74%.

CONCLUSION

The results thus prove that the application of clustering and classification have positive impact during user navigation pattern discovery process and gives its best result in both finding the navigation pattern and predicting future request. Future research can also try to combine clustering and association rules to discover more knowledge from the clustered data. Different clustering algorithms can also be investigated to improve the trend analysis and knowledge discovery.

REFERENCES

[1] Abraham. Natural Computation for Business Intelligence from Web Usage Mining, Proceeding of Seventh International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNAC2005), pp. 3-11, 2005.

[2] Baraglia, R. and Palmerini, P. (2002) SUGGEST: A web usage mining system, Proc. of IEEE Int'l Conf. on Information Technology: Coding and Computing, P.282.

[3] Clark, L., Ting, I.H., Kimble, C., Wright, P. and Kudenko, D. (2006) Combining ethnographic and clickstream data to identify user Web browsing strategies, Information Research, Vol. 11, No. 2, Pp. 1-9.

[4] Deneubourg, J.L., Goss, S., Franks, N., Sendova–Franks, A., Detrain, C. and Chretien, L. (1990) The Dynamics of Collective Sorting Robot–Like Ants and Ant–Like Robots. From Animals to Animals, Proc. of the 1st Int. Conf. on simulation of Adaptive Behaviour, Pp. 356–363.

[5] Dixit, D. and Gadge, J. (2010) A New Approach for Clustering of Navigation Patterns of Online Users, International Journal of Engineering Science and Technology, Vol. 2, No.6, Pp. 1670-1676.

[6] Handl, J. and Meyer, B. (2002) Improved ant-based clustering and sorting in a document retrieval interface, Proceedings of the Seventh International Conference on Parallel Problem Solving from Nature, Vol. 2439 of LNCS, Springer-Verlag, Berlin, Germany, Pp. 913–923.

[7] Jalali, M., Mustapha, M., Mamat, A. and Sulaiman, M.N.B. (2008a) A new clustering approach based on graph partitioning for navigation patterns mining, 9th International Conference on Pattern Recognition, Pp. 1-4.

[8] Jalali, M., Mustapha, N., Mamat, A., Sulaiman, N.B. (2008b) Web user navigation pattern mining approach based on graph partitioning algorithm, Journal of Theoretical and Applied Information Technology, Pp. 1125-1131

[9] Jalali, M., Mustapha, N., Sulaiman, N.B. and Mamat, A. (2008c) A web usage mining approach based on LCS algorithm in online predicting recommendation systems, 12th International Conference Information Visualization, IEEE Computer Society, Pp. 302-307.

[10] Jespersen S.E., Thorhauge J., and Bach T. (2002), A Hybrid Approach to Web Usage Mining, Data Warehousing and Knowledge Discovery, LNCS 2454, Y. Kambayashi, W. Winiwarter, M. Arikawa (Eds.), Pp. 73-82.

[11] ianbin, C., Jie, S. and Yunfei, C. (2007) A New Ant-based Clustering Algorithm on High Dimensional Data Space, Part 12, Complex Systems Concurrent Engineering, Springer London, Pp. 605-611.

[12] Kobra Etminani. Web Usage Mining: users' navigational patterns extraction from web logs using Ant-based Clustering Method, Proceedings of IFSA-EUSFLAT pp.396-401, 2009

[13] Liu, H. and Keselj, V. (2007) Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests, Data and Knowledge Engineering, Vol. 61, Issue 2, Pp. 304-330.

[14] E. Lumer, B. Faieta, Diversity and adaptation in populations of clustering ants. Proceeding of the third international conference on simulation of adaptive behaviour, pp. 501–508, MIT Press, 1994

[15] Mobasher, B., Cooley, R. and Srivastava, J. (2000) Automatic Personalization Based on Web Usage Mining, Communications of the ACM, Vol. 43 , Issue 8, Pp: 142 - 151

[16] Nakagawa, M. and Mobasher, B. (2003) A hybrid web personalization model based on site connectivity, Proc. of WebKDD, Pp.59-70.