Clustering Web Usage Data using Concept Hierarchy and Self Organizing Map

T.Vijaya Kumar
Research Scholar, Department of IS & E
BMS College of Engineering, Bull Temple Road
Bangalore - 560019, Karnataka, India

H.S.Guruprasad, PhD.
Professor and Head, Department of IS & E
BMS College of Engineering, Bull Temple Road
Bangalore - 560019, Karnataka, India

ABSTRACT
Clustering Web Usage data is one of the important tasks of Web Usage Mining, which helps to find Web user clusters and Web page clusters. Web user clusters establish groups of users exhibiting similar browsing patterns and Web page clusters provide useful knowledge to personalized Web services. Different types of clustering algorithms such as partition based, distance based, density based, grid based, hierarchical and fuzzy clustering algorithms are used to find clusters from Web usage data.These clustering algorithms require more space and time forlargerWeb server log files. K-Means algorithm has been one of the most widely used algorithms for clustering Web usage data due to its computational performance. Although K-Means algorithm is relatively fast and efficient compared to other clustering algorithms, it has some major drawbacks. The number of clusters must be specified in advance. The initial cluster centroids are selected randomly. Clustering result depends on the selection of randomly selected initial cluster centroids and different runs on the same input data might produce different results. K-Means algorithm is sensitive to noisy data and outliers. Recent studies have supported the use of neural networks such as Adaptive Resonance Theory (ART) and Self Organizing Maps (SOM) for real world data mining problems. Among the architectures and algorithms suggested for neural networks, the SOM has special property of effectively creating spatially organized internal representations of various features of input data and their abstractions. In this paper we propose a framework for finding useful information from Web Usage Data that uses SOM. First we have constructed the sessions using concept hierarchy and link information. Then SOM is used to cluster the sessions. We provide experimental results to show the benefits of using concept hierarchy for synaptic weights and clustering Web usage data using SOM. In this paper, we have considered the server log files of the Website www.enggresources.com for overall study and analysis.

General Terms
Web mining, Neural network.

Keywords
Concept hierarchy, Web usagemining, Concept based Website graph, Self-Organizing Maps, Synaptic weight vector.

1. INTRODUCTION
The growth of World Wide Web over the last two decades has resulted in a large amount of data that is available for user access. These different types of data have to be managed and organized in such a way that they can be accessed by users efficiently and effectively. Meanwhile, the substantial increase in the number of Websites presents a challenging task for Website administrators to organize the contents to serve the needs of users. Website administrators may want to know how they can attract visitors, which pages are being accessed most or least frequently, which part of Website is most or least popular and need enhancement, etc. Web mining can be used to discover and extract useful information from the World Wide Web documents and services in order to better understand and serve the needs of Web-based applications. Web mining research can be classified into three categories, namely Web Content Mining (WCM), Web Structure Mining (WSM) and Web Usage Mining (WUM). WCM deals with the extraction of knowledge from Web contents like text, image, and audio or video data. WSM is the process of using graph theory to analyze the node and connection structure of a Web site. WUM deals with the automatic discovery of user access patterns from one or more Web servers. Web Usage mining contains three main tasks namely Data preprocessing, Cluster discovery and Cluster analysis. Data preprocessing consists of data cleaning,data transformation, and data reduction. Data cleaning routines work to clean the data by filling in missing values, smoothing noisy data, and resolving inconsistencies in the data. In data transformation, the data are transformed or consolidated into forms appropriate for mining. Data reduction techniques can be applied to obtain a reduced representation of the data that is much smaller in volume, yet closely maintains the integrity of the original data. Cluster discovery deals with formation of groups of users exhibiting similar browsing patterns and obtaining groups of pages that are accessed together. Cluster analysis filters out uninteresting patterns from the user clusters and page clusters found in the Cluster discovery phase. Cluster analysis is the final stage of Web usage mining, which is a descriptive method used to analyze the Web usage data and customer behaviors. The general summary of the overall user behavior can be obtained from this phase resulting in aggregate user models that can be used as input to applications such as recommendation engines, visualization tools, and Web analyzers and report generation tools. Clustering is a data mining technique that groups together a set of items having similar characteristics. In the Web usage domain, two kinds of interesting clusters such as user clusters and page clusters can be discovered. Clustering algorithms require more space and computation for large data set. The finer details of data may be lost due to representation of data into fewer groups. Recent studies have supported the use of neural networks such as Adaptive Resonance Theory (ART) and Self Organizing Maps (SOM) for real world data mining problems. A Self Organizing Map (SOM) is an unsupervised neural network model that can be used for applications of data clustering and visualization. SOM is considered as one of the most common neural network methods for cluster analysis. In SOM, data in the input data space will be projected onto prototype vectors on the grid so that two vectors which are projected onto prototypes are more likely to belong to the same cluster. SOM has prominent visualization properties.In this paper we propose a framework for finding useful information from Web Usage Data that uses SOM. First in the

preprocessing phase the sessions are constructed using concept hierarchy and link information. Then in the Cluster discovery phase the user clusters and page clusters are obtained from sessions by using SOM. We provide experimental results to show the benefits of using concept hierarchy and link information for session construction and clustering Web usage data using SOM. The rest of the paper is organized as follows. Section 2 gives a brief description about the related work. Section 3 covers details of Data Preprocessing using concept hierarchy and Web site topology. The details of Self Organizing Maps are discussed in Section 4. The experimental design and results are discussed in section 5. Finally we give our conclusion in section 6.

2. LITERATURE SURVEY

Various data mining methods have been have been used to generate models of usage patterns. Models based on association rules, clustering algorithms, sequential analysis and Markov models have been used for discovering the knowledge from Web usage data. All these models are predominantly based on usage information from Web usage data alone. Significant improvement can be achieved by making use of domain knowledge, which is usually available from domain experts, content providers, Web designers and the Web page itself [1]. In [2, 3], Cooley et al. covered Web usage mining process & various steps involved in it. It serves as the primary thesis to understand fundamentals of Web usage mining. Along with the server log file other sources of knowledge such as site content or structure and semantic domain knowledge can be used in Web usage mining [4]. In [5], Natheer Khasawneh et al. have presented new techniques for preprocessing Web log data including identifying unique users and sessions by making use of Website ontology. In [6], Kobra etminani et al. proposed an idea to obtain details of how Website's link structure can be used for tracking Web user's activities. In [7], Sebastian A. Rios et al. have shown the use of concept-based approach using semantics in Web usage mining. In [8], Murat Ali Bayir et al. have proposed a novel framework, called Smart-Miner for Web usage mining problem which uses link information for producing accurate user sessions and frequent navigation patterns. Norwati Mustapha et al. [9], have proposed a model for mining user's navigation pattern based on Expectation modeling algorithm and used it for finding maximum likelihood estimates of parameters in probabilistic models. In [10], Saeed R. Aghabozorgi et al. have proposed an off line dynamic model using clustering algorithm and periodic user transaction for mining user behavior prediction for Web personalization system. A complete framework for mining evolving user profiles in dynamic Websites is proposed in [11]. They also described how to enrich the discovered user profiles with explicit information need that is inferred from search queries extracted from Web log data. In [12], Jiyang Chen et al. have proposed a visualization tool to visualize Web graphs, representations of Web structure overlaid with information and pattern. They also proposed Web graph algebra to manipulate and combine Web graphs and their layers in order to discover new patterns in an ad hoc manner. In [13], Esin Saka et al. have proposed a hybrid approach which combines the strengths of Spherical K-Means algorithm for fast clustering of high dimensional datasets in the original feature domain and the flock-based algorithm which iteratively adjusts the position and speed of dynamic flock's agents on a visualization plane. The hybrid algorithm decreases the complexity of FClust from quadratic to linear with further improvements in the cluster quality. In [14], S Park et al. have investigated the use of fuzzy ART neural network, to enhance the performance of the K-Means algorithm. A major limitation with fuzzy ART is the unrestricted growth of clusters. Fuzzy ART networks sometimes produce numerous clusters each with only a small number of members. The authors have used fuzzy ART as an initial seed generator and K-Means as the final clustering algorithm. In [15], Santosh K et al. have developed a clustering algorithm that groups users according to their Web access patterns. The algorithm is based on the ART1 version of adaptive resonance theory. ART1 offers an unsupervised clustering approach that adapts to the changes in user's access patterns over time without losing earlier information. It applies specifically to binary vectors. They have compared their algorithm's performance with the traditional K-Means clustering algorithm and showed that the ART1-based technique performed better in terms of intra cluster distances. In [16], Antonia S et al. have developed a variant of the classical SOM called Growing Hierarchical SOM (GHSOM). They have suggested a new visualization technique for the patterns in the hierarchical structure. In [17, 18], T. Kohonen et al. have used Kohenen's Self Organizing Map to organize Web documents into a two dimensional map according their document content. Documents which are similar in content are located in similar regions on the map. In [19], Kate A. Smith et al. have developed LOGSOM, a system that utilizes Kohonen's Self Organizing Map to organize Web page in to two dimensional maps. The organization of the Web pages is based on the user's navigation behavior rather than the content of the Web pages.

3. DATA PREPROCESSING

Data preprocessing [20] comprises of, merging of log files from different Web servers, Data cleaning, Identification of users, sessions, and visits, Data formatting and Summarization. Data cleaning consists of removing superfluous data from log file. We have considered the server log files of the Website www.enggresources.com . The server log file is in Extended Combined Log Format which is an extension of Common Log Format with two extra fields, referrer & user agent. User identification deals with identifying unique clients to Web server. A combination of IP address & user agent is used to identify users uniquely. User identification can also be done using client side cookies. But, due to privacy reasons, cookies can be disabled by users, and not every Website employ cookies. Session identification is considered as the next step. A session is a sequence of requests made by a single user with a unique IP address on a particular Web domain during a specified period of time.

Time oriented approach: The most basic session definition comes with Time Oriented Heuristics which are based on time limitations on total session time or page-stay time. They are divided into two categories with respect to the thresholds they use:

- In the first one, the duration of a session is limited with a predefined upper bound, which is usually accepted as 30 minutes. In this type, a requested page can be appended to the current session if the time difference between requested page and the first page in the current session doesn't exceed total session duration time. Otherwise, a new session is assumed to start with the requested page.
- In the second time-oriented heuristic, the time spent on any page is limited with a threshold. This threshold value is accepted as 10 minutes. If the timestamps of two consecutively accessed pages is greater than the threshold, the current session is terminated after the former page and a new session starts with the latter page.

Navigation oriented approach: Navigation-Oriented approach [21, 22] uses link information of Website graph

which is present in concept based Website graph constructed by using Website knowledge. In this approach, it is necessary to have a hyperlink between every two consecutive Web page requests.

Let $p = [\, p_1, p_2, p_3, \ldots\ldots p_k, p_{k+1}, \ldots\ldots, p_n]$be a session containing Web pages with respect to their timestamp orders. In this session, for every pagep_k, except the initial page p_1, there must be at least one page p_j in the session which is referring to $p_k$and has a smaller timestamp than p_k. Topology constraint forces to consider user navigation according to some path in Website graph.

Concept-matching approach: This approach considers concepts of Web pages from concept based Website graph. Adding a page p_{n+1} to a session $[p_1, p_2, \ldots\ldots\ldots, p_n]$is performed as follows: If the concept names of pages p_n&p_{n+1} are same,then addp_{n+1} to the current session else create a new session and add $p_{n+1}$to new session. That is concept switching is taken as one more criteria for breaking session [23]. A brief outline of the session construction algorithm is explained below.

Session construction algorithm
Input: Cleaned log file and Concept based Website graph
Output: Session file
foreach user based on distinct ip & user_agent
for each request of current user
if(time diff between cur & prev request < pagestay time threshold**&**time diff between cur & first request of session< session time out **&**Link_constraint(prev-request,current-request)**&**Concept_Match(prev-request,current-request))
Add this request to Current Session.
else
Write previous session to session file,
Add this request to New Session.
endif
end for
Write All Session indexes of Current user to session navigation file.
end for
end of algorithm.

4. SELF-ORGANIZING MAPS

The Self Organizing Map (SOM) is an excellent tool for experimental data mining. It projects input vectors on prototypes of a low dimensional regular grid that can be effectively utilized to explore properties of the data. A brief outline of the SOM concept is explained below.

Let m denote the dimension of the input space.

Let an input vector selected at random from the input space be denoted by $x = [x_1, x_2, x_3, \ldots\ldots\ldots, x_m]^T$.

The synaptic weight vector of each neuron in the network has the same dimension as the input space. Let the synaptic weight vector of neuron$j$ be denoted by $w = [\, w_{j1}, w_{j2}, w_{j3}, \ldots\ldots\ldots, w_{jm}]^T j = 1,2,3, \ldots\ldots\ldots, l.$

where l is the total number of neurons in the network.

To find the best match of the input vector x with the synaptic weight vectorw_j, compute the inner products ${w_j}^T x$ for $j = 1,2,3, \ldots\ldots\ldots l$ and select the largest. This assumes that the same threshold is applied to all the neurons. Thus by selecting the neuron with the largest inner product ${w_j}^T x$ the location where the topological neighbourhood of excited neurons is to be centered can be determined. This is mathematically equivalent to minimizing the Euclidean distance between the vectors $x$and w_j.

Let $i(x)$ denotes the neuron that best matches the input neuron x.

Then $i(x) = \arg min_j \; | x - w_j |$where $j = 1, 2, \ldots\ldots l$

This sums up the essence of competition process among the neurons. The particular neuron i that satisfies this condition is called the best matching or winning neuron for the input vector x. A continuous input space of activation patterns is mapped on to a discrete output space of neurons by a process of competition among the neurons in the network. Depending on the application of interest, the response of the network could be either the index of the winning neuron or the synaptic weight vector that is closest to the input vector in a Euclidean sense. The winning neuron locates the center of a topological neighborhood of cooperating neurons. A neuron that is firing tends to excite the neurons in its immediate neighborhood more than those farther away from it. The topological neighborhood around the winning neuron i decays smoothly with lateral distance. A typical choice of the topological neighborhood centered on winning neuron and encompassing a set of cooperating neuron is the Gaussian function. For the network to be Self-Organizing, the synaptic weight vector w_j of neuron j in the network is required to change in relation to the input vector x. A modified Hebbian Hypothesis can be used to change the synaptic weight vector w_j of the neuron j [24].

An outline of the revised SOM algorithm for our system is summarized below.

Input: Sessions constructed from the preprocessing phase. Each requested URL is assigned with a unique number.

Output: Web user clusters and Web page clusters.

Step1. **Initialization:** Choose random values for the weight vectors w_j. Here we have selected the synaptic weight vectors from the available set of session vectors based on the concept hierarchy and Web site topology.

Step2. Sampling: Draw an input vector sample x with a certain probability. The dimension of vector x is equal to m. We assume that there is a set of$n$Web pages $\{\, p_1, p_2, p_3, \ldots\ldots\ldots, p_n\}$ and a set of m user transactions.

Step3. Similarity Matching: Find the best matching winning neuron $i(x)$ at time step n by using minimum distance Euclidean Criterion.

$i(x) = \arg min_j \; | x - w_j |$where$j = 1, 2, \ldots\ldots l$

Step4. Updating: Adjust the synaptic weight vectors of all neurons by using the update formula

$w_j(n+1) = w_j(n) + \eta(n) h_{j,i(x)}(n)\,(x(n) - w_j(n))$

Where $\eta(n)$ is the learning rate parameter, and $h_{j,i(x)}(n)$ is the neighborhood function centered around the winning neuron $i(x)$. Both $\eta(n)$and $h_{j,i(x)}(n)$ are varied dynamically during learning for best results.

Step5. Continuation: Continue with step 2 until no noticeable changes in the feature map are observed.

5. EXPERIMENTAL DESIGN

Server log file from the Web site www.enggresources.com is taken for our experimental study and concept based Website graph is constructed as an additional input. Error records, requests for images and multimedia files are removed from Server log file by using a tool called Web log filter. IP address, timestamp, user agent, request and referrer are retained for further processing. User Identification is considered as the next step. A combination of IP address and user agent is used to identify the uses uniquely. In session construction, we have combined two trivial approaches, Time oriented approach and Navigation oriented approach along with concept name match approach for identifying user sessions. Page stay time threshold and session timeout threshold are set as 10 and 30 minutes respectively. Each Web page is assigned with unique index. And, every unique session

is also given unique index.Domain knowledge can exist in various forms such as concept hierarchy, Website topology and semantic classification. This knowledge in conjunction with Web usage data can be used to improve the knowledge discovery process. We have considered Web site specific factors and Web Usage Mining specific factors for our study and analysis. Web site specific factors such as concept hierarchy and Website graph, Web site specific factors are the ones over which the Web site administrator has little control. Hence the objective is to compare and evaluate the clustering performance using different representation schemes. Web Usage Mining specific factors are the number of user sessions, the minimum session length and the number of resulting clusters. Web Usage Mining factors can be decided by the analyst conducting the mining process. Perhaps the research most relevant to our study is Self-Organization of massive document collection called WEBSOM and the Web page clustering using SOM called LOGSOM.WEBSOM, uses Kohenen's Self-organizing maps to organize Web documents into a two dimensional map according to their document content. Documents which are similar in content are located in similar region on the map. This method is very effective because the system is able to automatically organize the documents into meaningful clusters according to their content. LOGSOM utilizesSOM to mine Web log data and provides a visual tool to assist user navigation. The system uses Kohenen's Self-organizing maps to organize Web pages into a two dimensional map. The organization of the Web pages is based solely on the user navigation behavior, rather than the content of the Web pages. The resulting map not only provides a meaningful navigational tool that is easily incorporated with Web browsers, but also serves as a visual analysis tool for Web masters to better understand the characteristics of navigation behaviors of Web users visiting the pages. The approach used in LOGSOM is different from WEBSOM, since it clusters the Web pages according to the user's navigation behavior rather than the Web content. Our approach uses concept hierarchy and Website topology for synaptic weight vectors and SOM for clustering the preprocessed data.

5.1 Results and Discussion

We have considered Web Server log file from the Web site www.enggresources.com for our experimental study and concept based Website graph is constructed as additional input. Error records, requests for images and multimedia files are removed from Server log file by using a tool called Web log filter. Usually this process removes requests concerning non-analyzed resources such as images, multimedia files, and page style files (*.CSS) etc. IP address, timestamp, user agent, request and referrer are retained for further processing. In user identification, IP address and user agent are used. That is, a combination of IP address and user agent is used to identify a unique user. In session construction, we have combined two trivial approaches, Time oriented approach and Navigation oriented approach along with concept name match approach for identifying user sessions. Page stay time threshold and session timeout threshold are set as 10 and 30 minutes respectively. Each Web page is assigned with unique index. And, every unique session is also given unique index. 10217 users and 25814 sessions were discovered from pre-processing. Then based on the number of times a page is repeated in a session and the link information each page is assigned with a unique number between 0 and 1.The synaptic weight vector matrix w_j is initialized using concept hierarchy and Website link information. Then sessions constructed from the previous phase as taken as the input variable x and is compared with each synaptic weight vector w_j at each successive iteration. Then the best matching weight vector is updated to match even more closely to the current input x. The different weight vectors tend to become spatially tuned to different domains of the input variable x. The computer simulation presented here is intended to illustrate the effect that the weight vector matrix tends to approximate to the input vector. The Web site administrator can initialize the weight matrix based on the concept hierarchy and Website topology. It is important to note that both the directory structure and content of the Web pages are used as the inputs along with the server log file to the processing part of our system. The Web site administrator can analyze the adaptive changes in the weight vector matrix for different input data set. The choice of weight matrix may favor one particular orientation in the output map, since the output mainly depending on the initial value of weight matrix. This can be resolved by considering different set of Website specific factors and Web usage mining factors for the computer simulation. Fig. 1(a) and Fig. 1(b) depict the Initial synaptic weight vector matrix and Synaptic weight vector matrix after 100 epochs respectively. The weight vectors appear as points in the same coordinate system as that in which the input variable vector x are represented, in order to indicate to which unit each w_j value belongs. The points corresponding to the $w_j$vector values have been connected by a lattice of lines conforming to the topology of the array. A line connecting two weight vectors used to indicate that the corresponding vectors are adjacent [25]. The Website considered for our study contains 95 URLs. Fig. 2(a) and 2(b) shows a typical SOM produced by our system. The blank node contains no URLs, while those that are numbered indicate the number of URLs contained within each node. For example, in 2(a), the node at the end of the first row is numbered 27 because it contains 27 URLs. The two important parameters of the Self Organizing Maps are the learning rate parameter $\eta(n)$ and the number of repetitions of the process of learning within one cycle, which is the number of times each URL is presented within one cycle before the neighborhood size is decreased. We have considered the learning rate parameter to begin with 0.1, and decrease gradually but remain above 0.01. We found that while there are some differences in the quality of maps for different values of learning rate $\eta(n)$,the quality becomes similar when the number of iterations is more. Although the groups of Web pages may be located at different corners of the maps, the similar pages are still grouped together even though they may be located at different corners of the maps. Fig 3(a) and 3(b) show session clusters and weight planes respectively for 6000 sessions with session length 16. The number in each node indicates the number of sessions contained within each node. A Sample of SOM experiments conducted with varying Web usage mining specific factors such as session length and number of user sessions, are shown in Fig 4(a), 4(b), 5(a), 5(b), 6(a) and 6(b).

6. CONCLUSION

We have developed a system that utilizes Kohonen's SOM to cluster Web usage data.Formation of clusters is based on the user's navigation behavior and concept based Website graph.Web user clusters and Web page clusters formed by SOM is more refined than K-Means algorithm as SOM clusters the data in a two dimensional plane. Although SOM is also capable of presenting the data in three dimensional spaces, two dimensional planesare most frequently used due to the ease of visualization.One major disadvantage with SOM is it only reveals the number of data points in each neuron but it cannotdiscriminate the exact data points. Our future

research will address this issue by enhancing the visualization technique by providing the data distribution in each neuron.In our system, sessions are constructed using the concept hierarchy and Website graph. The sessions can be enhanced by incorporating Web page content information along with the concept hierarchy. Our future research also covers the benefits of combining Web page content information along with concept based Website graph.

0.6311	0.8221	0.6296	0.0134	0.0547	0.3073	0.0354	0.9195	0.9180
0.4282	0.3604	0.5642	0.4078	0.6639	0.8829	0.1951	0.8474	0.1377
0.3206	0.4077	0.9830	0.0809	0.0590	0.0889	0.6698	0.9733	0.6935
0.0850	0.7197	0.9288	0.4940	0.1281	0.3987	0.7480	0.2090	0.7147
0.4267	0.2460	0.2384	0.5240	0.3040	0.7221	0.4753	0.0511	0.7165
0.0450	0.8315	0.8198	0.8728	0.9481	0.5278	0.6211	0.7537	0.2973
0.1738	0.7346	0.0776	0.2363	0.6011	0.5216	0.5125	0.5206	0.6766
0.7595	0.6937	0.2941	0.5843	0.4038	0.0444	0.1857	0.5404	0.0002
0.5015	0.6201	0.5084	0.0754	0.1920	0.3682	0.4909	0.1679	0.8864
0.7289	0.5302	0.5542	0.5416	0.6696	0.5885	0.5324	0.0027	0.1280
0.8299	0.8381	0.8812	0.8354	0.4316	0.9550	0.1526	0.6377	0.8016
0.0629	0.3397	0.2012	0.1249	0.2426	0.3640	0.7349	0.0723	0.0361
0.6758	0.2093	0.6097	0.2618	0.2197	0.6193	0.1493	0.6861	0.9192
0.3362	0.5259	0.8254	0.1994	0.2351	0.0842	0.1680	0.4317	0.7738
0.1299	0.2063	0.9881	0.4022	0.5529	0.2031	0.9065	0.1398	0.9984
0.6940	0.6193	0.3163	0.6481	0.6956	0.2377	0.6538	0.5073	0.1263
0.2383	0.8444	0.1548	0.9756	0.8585	0.0149	0.5132	0.6924	0.7471

Fig. 1(a): Initial weight vector matrix

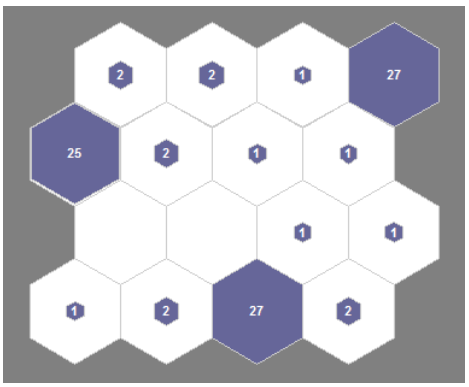

Fig. 2(a): SOM user clusters

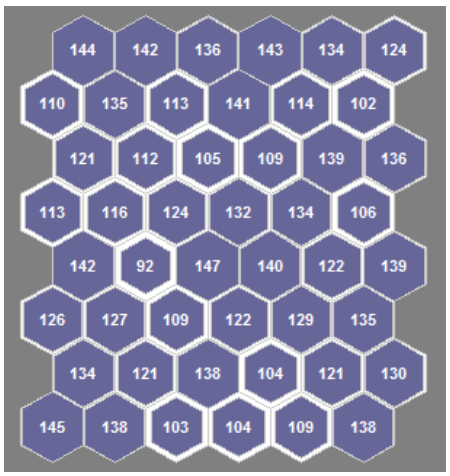

Fig. 3(a): SOM Session clusters

0.6311	0.4963	0.6296	0.0134	0.4573	0.4566	0.0958	0.9195	0.9180
0.4282	0.4125	0.5642	0.4078	0.6105	0.6620	0.3860	0.8474	0.1377
0.3206	0.6245	0.9830	0.0809	0.4884	0.6007	0.8042	0.9733	0.6935
0.0850	0.6141	0.9288	0.4940	0.3951	0.4321	0.6498	0.2090	0.7147
0.4267	0.3324	0.2384	0.5240	0.6141	0.9615	0.8421	0.0511	0.7165
0.0450	0.6268	0.8198	0.8728	0.5672	0.3546	0.6618	0.7537	0.2973
0.1738	0.4583	0.0776	0.2363	0.4663	0.7212	0.2979	0.5206	0.6766
0.7595	0.3719	0.2941	0.5843	0.7962	0.4272	0.3843	0.5404	0.0002
0.5015	0.6042	0.5084	0.0754	0.4482	0.1912	0.5091	0.1679	0.8864
0.7289	0.3107	0.5542	0.5416	0.5539	0.5376	0.5898	0.0027	0.1280
0.8299	0.5493	0.8812	0.8354	0.4633	0.2280	0.2702	0.6377	0.8016
0.0629	0.7568	0.2012	0.1249	0.5671	0.2504	0.2930	0.0723	0.0361
0.6758	0.4603	0.6097	0.2618	0.4515	0.4927	0.1148	0.6861	0.9192
0.3362	0.7090	0.8254	0.1994	0.3093	0.8503	0.2357	0.4317	0.7738
0.1299	0.4439	0.9881	0.4022	0.4079	0.2743	0.4809	0.1398	0.9984
0.6940	0.2349	0.3163	0.6481	0.4593	0.6554	0.2047	0.5073	0.1263
0.2383	0.7146	0.1548	0.9756	0.5057	0.2425	0.1258	0.6924	0.7471

Fig. 1(b): Weight vector after 100 epochs

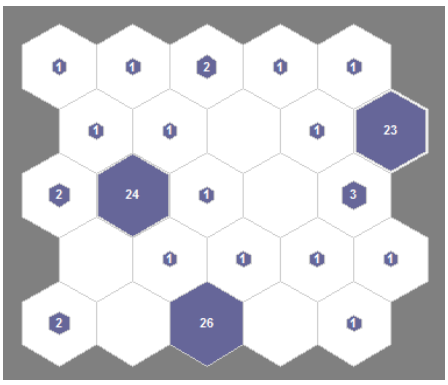

Fig. 2(b): SOM user clusters

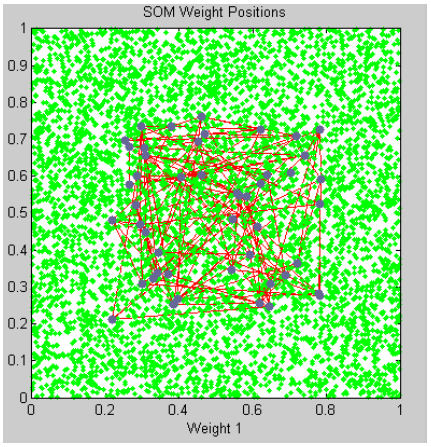

Fig. 3(b): SOM weight plane

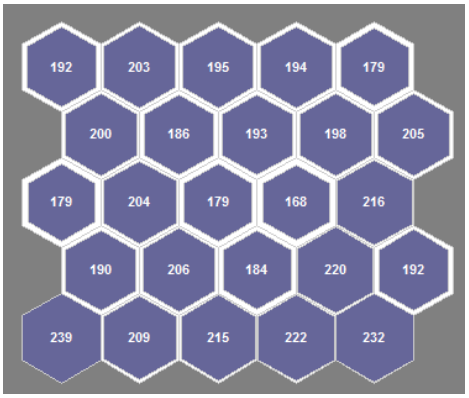

Fig. 4(a): SOM Session clusters

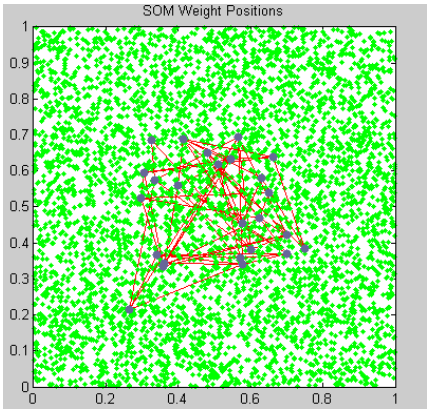

Fig. 4(b): SOM weight plane

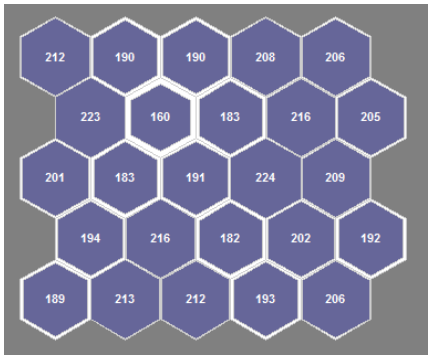

Fig. 5(a): SOM Session clusters

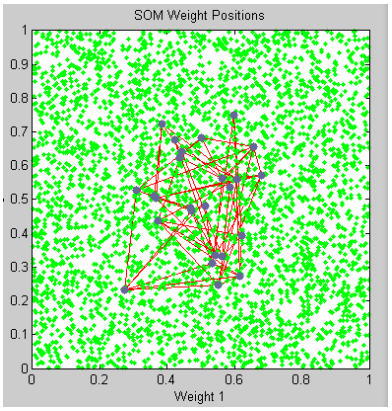

Fig. 5(b): SOM weight plane

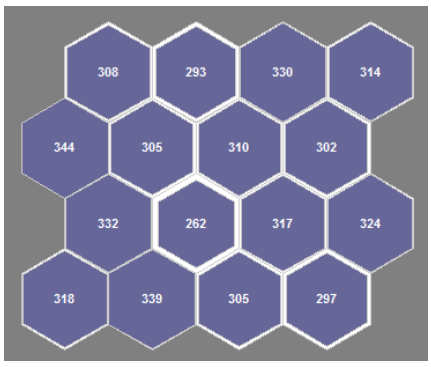

Fig. 6(a): SOM Session clusters

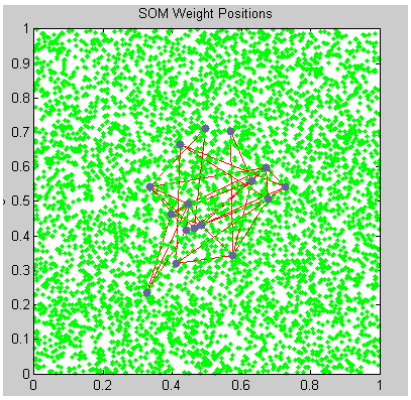

Fig. 6(b): SOM weight plane

7. REFERENCES

[1] Kalyan Beemanapalli, Jaideep Srivastava, and Sigal Sahar, "Incorporating Concept Hierarchies into Usage Mining Based Recommendations", WEBKDD'06, August 20, 2006, Philadelphia, USA, ACM.

[2] R. Cooley, B. Mobasher, and J. Srivastava, "Web mining: information and pattern discovery on the World Wide Web", Ninth IEEE International Conference on Tools with Artificial Intelligence, Newport Beach, CA, USA, 1997, Pages 558-567.

[3] J. Srivastava, R. Cooley, M. Deshpande, and P. N. Tan, "Web usage mining: discovery and applications of usage patterns from Web data", ACM SIGKDD Explorations Newsletter, Volume 1, Pages 12-23, 2000.

[4] Bamshad Mobasher, Chapter: 12, "Web Usage Mining in Data Collection and Pre-Processing", ACM SIGKKD 2007 Pages 450-483.

[5] Natheer Khasawneh and Hien-Chung Chan, "Active User-Based and Ontology-Based Weblog data preprocessing forWeb Usage Mining",IEEE/ WIC/ACM International Conference 2006.

[6] Kobra etminani, Amin, and Noorali Rouhani, "Web usage Mining:Discovery of the user's navigational patterns using SOM", IEEE 2009.

[7] Sebastian A. Rios, and Juan D.Velasquez, "Semantic Web Usage Mining by a Concept-based approach for Off-line Web Site Enhancements", IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology 2008.

[8] Murat Ali Bayir, Ismail Hakki Toroslu, Guven Fidan, and Ahmet Cosar, "Smart Miner: A New Framework for Mining Large Scale Web Usage Data", ACM 2009.

[9] Norwati Mustapha, Manijeh Jalali , and Mehrdad Jalali, "Expectation Maximization Clustering Algorithm for User Modeling in Web Usage Mining Systems", European Journal of Scientific Research ISSN 1450-216X Volume 32 Number.4 (2009), Pages.467-476.

[10] Saeed R. Aghabozorgi, and Teh Ying Wah, "Dynamic Modeling by Usage Data for Personalization Systems", 13th International Conference on Information Visualization IEEE 2009.

[11] Olfa Nasraoui, Maha Soliman,Esin Saka,Antonio Badia, and Richard Germain, "Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites", IEEE transactions on knowledge and data engineering, Volume. 20, Number. 2, February 2008.

[12] Jiyang Chen, Lisheng Sun, Osmar R.Zaiane, and Ranidy Goeble, "Visualizing and Discovering Web Navigational Patterns", Seventh International Workshop on the Web and Databases (Web DB 2004),June17-18, 2004, Paris, France.

[13] Esin Saka, and Olfa Nasraoui, "Simultaneous Clustering and Visualization of Web Usage Data using Swarm-based Intelligence", 20th IEEE International Conference on Tools with Artificial Intelligence.

[14] Sungjune Park, Nallan C. Suresh, and Bong Keun Jeong, "Sequence based clustering for Web usage mining: A new experimental framework and ANN-enhanced K-Means algorithm", Elsevier Data and Knowledge Engineering 65 (2008) 512 – 543.

[15] Santosh K.Rangarajan, Vir V.Phoha, Kiran S.Balagani, Rastko R.Selmic and S.S. Iyengar, "Adaptive Neural Network Clustering of Web Users", IEEE 2004 0018-9162/04.

[16] Antonio S, Jose D. Martin, Emilio S, Alberto P, Rafael M and Antonio, "Web mining based on growing hierarchical Self Organizing Maps: Analysis of a real citizen Web portal", Expert Systems with applications 34(2008)2998-2994 www.elsevier.com

[17] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela,V. Paatero,A. and Saarela, "Self organization of a massive document collection", IEEE Transactions on Neural Networks 11 (3)(May 2000) 574– 585.

[18] Samuel Kaski, Timo Honkela, Krista Lagus, and Teuvo Kohonen, "WEBSOM-Self organizing maps of document collections", Neurocomputing 21(1998) 101-117 Elsevier.

[19] Kate A. Smith, and Alan Ng, "Web page clustering using a self-organizing map of user navigation patterns", ElsevierDecision Support Systems 35 (2003) 245– 256.

[20] G.T.Raju, and P. S. Satyanarayana "Knowledge Discovery from Web Usage Data: Complete Preprocessing Methodology", IJCSNS International Journal of Computer Science and Network Security, Volume.8 Number.1, January 2008.

[21] C. Shahabi and F. B. Kashani, "Efficient and anonymous Web-usage mining for Web personalization", INFORMS Journal on Computing, 15(2) Pages 123-147, 2003.

[22] M. Spiliopoulou, B. Mobasher, B. Berendt, and M. Nakagawa, "A framework for the evaluation of session reconstruction heuristics in Web usage analysis", INFORMS Journal on Computing, 15(2), Pages 171-190, 2003.

[23] T.Vijaya Kumar, Dr. H.S. Guruprasad, Bharath Kumar K.M, Irfan Baig and Kiran Babu S,"A New Web Usage Mining approach for Website recommendations using Concept hierarchy and Website Graph", International Journal of Computer and Electrical Engineering (IJCEE, ISSN: 1793-8198 (Online Version);1793-8163(print version).

[24] Simon Haykin, "Neural Neworks – A Comprehensive Foundation", Prentice-Hall, Inc-1999.

[25] Teuvo Kohonen, "The Self Organizing-Map", Proceedings ofthe IEEE, VOl.78, No.9,September1990.