International Journal of Pure and Applied Mathematics
Volume 117 No. 15 2017, 1207-1213
ISSN: 1311-8080 (printed version); ISSN: 1314-3395 (on-line version)
url: http://www.ijpam.eu
Special Issue

ijpam.eu

Self Organizing For Dimensional Reduction in Web Mining

1.C.SADHANA(SP14CAD502)
Research Scholar
Computer Science and Application
St peters University
2.Dr.L.Mary Immaculate Sheela
Professor,
Department of Information Technology,
Pentecost University College,
Ghana.

Abstract

Data mining is a form of extracting data available in the internet. Web mining is a part of data mining. Web mining adopts much of the data mining techniques to discover potentially useful information. Web mining analysis relies on three general sets of information pervious usage patterns, degree of shared content and inter memory association link structure corresponding to three subset in web mining namely Web usage mining ,Web content mining, Web structure mining respectively. proposal shares dissimilar goals with many those agents, our approach is automatic that it does not require users explicit input. Moreover, we take a systematic approach to collect and comprehend user activities. We provide a general framework for collecting, mining, and search/query personal usage data, which may be employed by various agents. Web usage mining is used to analyze the behavior of websites users. It involves automatic discovery of user access patterns from one or more web servers. It contains four processing stages including data collection, preprocessing, pattern discovery and analysis. The web content mining refers to the discovery of useful information from web contents which include text, image, audio, video etc. The mining of link structure aims at developing techniques to take advantage of the collective .It includes extraction of structure data from web pages, identification, similarity and integration of data with similar meaning. There are two common tasks involved in web mining they are Clustering and Classification.Neural based approach is used to analysis the performance of the clustering of the number of request. We propose an approach "ENHANCED SELF ORGANIZTION MAP" which is data visualization technique; it reduces the dimensions of data through the use of neural network. In previous study on SOM plot the similarities of data by grouping similar data items together, so they reduces dimension and display similarities SOM organize sample data, which are usually surrounded by similar samples ,similar samples are not always near each other .In ESOM we use users Clustering mining algorithm. ESOM can estimate the center and the number of clustering data set by" dissimilarity computing", it optimizes SOM neural network learning and improve clustering effect.

Introduction

The expansion of the World Wide Web (Web forshort) has resulted in a large amount of data that isnow in general freely available for user access.The different types of data have to be managedand organized that they can be accessedefficiently. Therefore the application of datamining techniques on the Web is now the focus ofan increasing number of researchers. Several data Mining methods are used to discover the hiddeninformation in the Web. However, Web mining does not only mean applying data miningtechniques to the data stored in the Web. Thealgorithms have to be modified to better suit thedemands of the Web. New approaches should beused better fitting to the properties of Web data.Furthermore, not only data mining algorithms, butalso artificial intelligence, information retrievaland natural language processing techniques can beused efficiently. Thus, Web mining has beendeveloped into an autonomous research area. Webmining involves a wide range of applications thataim at discovering and extracting hiddeninformation in data stored on the Web. Another Importantpurpose of Web mining is to provide a mechanismtomake the data access more efficiently andadequately. The third interesting approach is todiscover the information which can be derivedfrom theactivities of users, which are stored in log files for example for predictive Web caching [1]. Thus, Web mining can be categorized into threedifferent classes based on which part of the Webis to be mined [2], [3], and [4]. These threecategories are Web content mining, Web structuremining and Web usage mining. Web contentmining [7], [6] is the task of discovering usefulinformation available on-line. There are differentkinds of Web content which can provide usefulinformation to users, for example multimediadata, structured (i.e. XML documents), semi structured (i.e. HTML documents) andunstructured data (i.e. plain text). The aim of Webcontent mining is to provide an efficient mechanism to help the users to find theinformation they seek. Web content miningincludes the task of organizing and clustering the documents and providing search engines foraccessing the different documents by keywords,categories, contents. Web structure mining is theprocess of discovering the structure of hyperlinks within the Web. Practically, while Web content mining focuses on the inner-documentinformation, Web structure mining discovers thelink structures at the inter-document level. Theaim is to identify the authoritative and the hubpages for a given subject. Web usage mining isthe task of discovering the activities of the userswhile they are browsing and navigating throughthe Web. The aim of understanding the navigationpreferences of the visitors is to enhance thequality of electronic commerce services(ecommerce), to personalize the Web portals [7]or to improve the Web structure and Web serverperformance [3]. For this reason a model of theusers (User Model - UM) have to be built based on the information gained from the log data.

Web Usage Mining

The aim of Web usage mining is to discoverpatterns of user activities in order to better servethe needs of the users for example by dynamiclink handling, by page recommendation etc. The aim of a Web site or Web portal is to supply theuser the information which is useful for him.There is a great competition between the differentcommercial portals and Web sites because everyuser means eventually money (throughadvertisements, etc.). Thus the goal of each ownerof a portal is to make his site more attractive forthe user. For this reason the response time of each single site have to be kept below 2s. Moreoversome extras have to be provided such assupplying dynamic content or links orrecommending pages for the user that are possibleof interest of the given user. Clustering of the useractivities stored in different types of log files is akey issue in the Web community. There are three types of log files that can be used for Web usagemining [4]. Log files are stored onThe server side, on the client side and on theproxy servers. By having more than one place forstoring the information of navigation patterns ofthe users makes the mining process more difficult.Really reliableresults could be obtained only if one has datafrom all three types of log file. The reason for thisis that the server side does not contain records ofthose Web page accesses that are cached on theproxy servers or on the client side. Besides the logfile on the server, that on the proxy server provides additionalInformation. However, the page requests stored inthe client side are missing. Yet, it is problematicto collect all the information from the client side.Thus, most of the algorithms work based only theServer Side data. Web usage mining consists of threemain steps:
(i) Pre-processing(ii) Pattern discovery (iii) Pattern analysis

In the pre-processing phase the data have to beCollected from the different places it is stored(client side, server side, proxy servers). Afteridentifying the users, the click-streams of eachuser has to be split into sessions. In general thetimeout for determining a session is set to 30minute [5]. The pattern discovery phase meansapplying data mining techniques on thepre-processed log data. It can be frequent pattern mining, association rulemining or clustering. In this paper we are dealingonly with the task of clustering web usage log. Inweb usage mining there are two types of clustersto be discovered: usage clusters and page clusters.The aim of clustering users is to establish groupsof users having similar browsing behaviour. Theusers can be clustered based on several Information. In the one hand, the user can berequested filling out a form regarding theirinterests, for example when registration on the web portal. The clustering of the users can beaccomplished based on the forms. On the otherhand, the clustering can be made based on theinformation gained from the log data collectedduring the user was navigating through the portal. Different types of user data can be collected usingthese methods, for example (I) characteristics ofthe user (age, gender, etc.), (ii) preferences and interests of the user, (iii) user's behaviourpattern. The aim of clustering web pages is to have groupsof pages that have similar content. Thisinformation can be useful for search engines orfor applications that create dynamic index pages. The last step of the whole web usage miningprocess is to analyze the patterns found during thepattern discovery step. Web Usage Mining try tounderstand the patterns detected in before step. The most common techniques is data visualizationapplying filters, zooms, etc.
Web Usage Mining(WUM) consists in the analysis of the way a web site is browsed by its users so as to improve it(in a very broad sense).The practical goals include to a few; improving the performances of web servers with intelligent caching and proxy that anticipate user requests; improving the structure of a site based on user typical navigation, for instance by creating automatically site map and bypassing links; introducing recommendations, such as the proposed by on-line bookstores, based on recognition of typical browsing sessions or of buying patterns.

Self Organized Map

The Self-Organizing Map (SOM) [17] was developed by professor Kohonen. It is one of the most popular neural network models. It belongs to the category of competitive learning networks Based on unsupervised learning, which means that no human intervention is needed during the learning and that little need to be known about the characteristics of the input data. SOM is used for clustering the data without knowing the class. The SOM can be used to detect features inherent to the problem and thus has also been called SOFM, the Self-Organizing Feature Map, Provides a topology preserving mapping from the high dimensional space to map units. Map units, or neurons, usually form a two-dimensional lattice and thus the mapping is a mapping from high dimensional space onto a plane. The property of topology preserving means that the mapping preserves the relative distance between the points. Points that are near each other in the input space are mapped to nearby map units in the SOM. The SOM can thus serve as a cluster analyzing tool of highdimensional data.

Self Organising Algorithm

1. Select output layer network topology.Initialize current neighbourhood distance, D (0), to a positive value.
2. Initialize weights from inputs to outputs to small random values.
3. Let t=0
4. While computational bounds are not exceeded do (t<=1).
i) Select an input sample ti, k.
ii) Compute the square of the Euclidean distance of ti, k. Fromweight vectors (wj) associated with each output node.
ti, k - wj, k (t))2
iii) Select output node *j** that has weight vector with minimumvalue from step 2.
iv) Update weights to all nodes within a topological distancegiven by D(t) from *j**, using the weight update rule: wj (t+1) = wj (t) + n(t)(tl- wj (t))
v) Increment t.
5. End while.

Dissimilarities of WUM

A difficulty included by the proposed representation is that N(the number of sessions)is in general quite high(more than 16000 in the proposed application) .

2 Related Algorithms

2.1 SOM

TeuvoKohonen [4] introduced the SOM network thatreduced the dimensions of data through the use of selforganizingneural networks. The SOM network producesa map of usually one or two dimensions which plot thesimilarities of the data by grouping similar data items together.This mapping process reduces the problem dimensions.The SOM network integrates dimensions reducingand clustering in one network. Figure 1 shows the mappingfrom a one-dimensional input to a two-dimensional array.

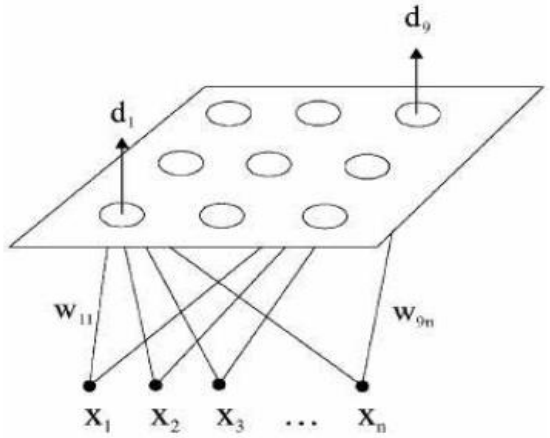


Figure 1: The Mapping from a one-dimensional input to atwo-dimensional array [11].The SOM network organizes itself by competing representationof the samples. Neurons are also allowed tochange themselves in hoping to win the next competition.This selection and learning process makes the weights toorganize themselves into a map representing similarities.The algorithm of the SOM network is shown as follows:

1. Initialize Map
2. Set $t = 0$ and repeat the following steps until $t > 1$
Randomly select a sample
Get best matching unit
Scale neighbors
Increase t by a small amount
3. End for

The first step in constructing a SOM is to initialize theweight vectors. From there the algorithms select a sample vector randomly and search the map of weight vectors tofind the weight that can represent the sample best. Sinceeach weight vector has a location, it also has neighbouringweights that are close to it. The chosen weight is rewardedto perform better than a randomly selected sample vector.In addition to this reward, the neighbours of the weight arealso rewarded. From this step we increase t some small amount because the number of neighbours and how mucheach weight can learn decreases over the time. This wholeprocess is then repeated a large number of times, usually atleast 1000 times.
The main advantage of using the SOM network is thatSOM automatically (self-organizing) clusters documents. The SOM network also can be applied to a large scale ofdata.

2.2 K-Means

The *k*-means algorithm was introduced by J. Mac-Queen, and it had been one of the most popular clustering Algorithms. This clustering algorithm represents each of*k* clusters *Cj*by the mean (weighted average) *cj*of its point (called centroid). It initially selects clusters such thatpoints are mutually farthest apart. Next, it examines eachpoint and assigns it to one of the clusters depending on theminimum distance. The centroid's position is recalculatedevery time a point is added to the cluster. This process willbe repeated until all the points are grouped into the *k* clusters.However, this algorithm does not work well if thereare large differences in the data set. The equation for kmeansalgorithm is in Equation 1 and 2.

$$u_{ij} \in U_{c\times n}; Ci = \frac{1}{n_i}\sum_{j=1}^{n} X_j \qquad (1)$$

$$\min J = \sum_{i=1}^{c}\sum_{j=1}^{n} u_{ij}\,\|X_j - C_i\|^2 \qquad (2)$$

In equations (1) and (2), *Xj*represents each point *j*'sco-ordinatesand *uij*represents the hypothetical belongingof point *j* into cluster i (i.e., *uij*= 1 if *j* belongs to cluster *i*;*uij*= 0 if *j* belongs to any other cluster different from *i*)

3 SOM-based Web page clustering

Overview of the SOM Algorithm
We have a spatially *continuous input space*, in which our input vectors live. The aim is
to map from this to a low dimensional spatially *discrete output space*, the topology of which is formed by arranging a set of neurons in a grid. Our SOM provides such a nonlinear transformation called a *feature map*.
The stages of the SOM algorithm can be summarised as follows:
1. *Initialization* – Choose random values for the initial weight vectors w*j*.
2. *Sampling* – Draw a sample training input vector x from the input space.
3. *Matching* – Find the winning neuron *I*(x) with weight vector closest to input vector.
4. *Updating* – Apply the weight update equation D*wji t Tj I t xi wji*=h( ) , (x)( ) (-).
5. *Continuation* – keep returning to step 2 until the feature map stops changing.
Dissimilarity/Similarity metric: Similarity is expressed in terms of a distance function, which is typically metric:*d*(*i, j*)
There is a separate "quality" function that measures the "goodness" of a cluster.
The definitions of distance functions are usually very different for interval-scaled, boolean, categorical, ordinal and ratio variables.
It is hard to define "similar enough" or "good enough"
the answer is typically highly subjective.
Similarity and Dissimilarity Between Objects
Distances are normally used to measure the similarity or dissimilarity between two data objects
Some popular ones include: *Minkowski distance*:
where $i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ and $j = (x_{j1}, x_{j2}, \ldots, x_{jp})$ are two *p*-dimensional data objects, and *q* is a positive integer
If *q = 1*, *d* is Manhattan distance

$$d(i,j) = |x_{i1} - x_{j1}| + |x_{i2} - x_{j2}| + \ldots + |x_{ip} - x_{jp}|$$

If q = 2, *d* is Euclidean distance:
Properties
$d(i,j) \geq 0$
$d(i,i) = 0$
$d(i,j) = d(j,i)$
$d(i,j) \leq d(i,k) + d(k,j)$
Also one can use weighted distance, parametric Pearson product moment correlation, or other disimilarity measures.
Dissimilarity between Binary Variables

gender is a symmetric attribute
the remaining attributes are asymmetric binary
let the values Y and P be set to 1, and the value N be set to 0

3.2 Data Pre-processing

There are several pre-processing tasks to be donebefore executing the data mining algorithms on the Webserver logs. These processes include data formatting, useridentification, session identification, and transaction identification.The original server logs are formatted and groupedinto meaningful transactions before being processed by themining system. We describe each of these processes in thefollowing paragraphs.Data formatting The access log is saved to keep arecord of every request made by the users. Since our mainpurpose is to facilitate more effective and efficient navigation,we only want to keep the log entries with information

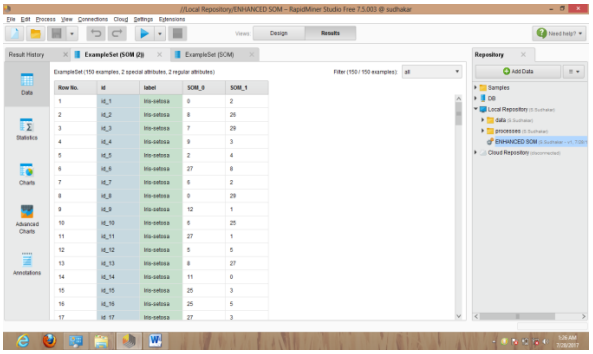
relevant to our purpose of organizing theWeb pages. Someirrelevant log entries are deleted from the log file.Sometimes a user requests a page that does not exist.This will create an error entry in the log. Since we are organizingthe existing Web URLs, we are not interested inthis kind of error entries, and hence these error entries shallbe deleted. A users request to view a particular page oftenresults in several log entries because the page consists of several materials such as graphics or small applets. However,we are only interested in, and hence only keep, whatthe user explicitly requests because we intend to design asystem that is user-oriented.
User identification The task of identifying uniqueusers is greatly complicated by the existence of localcaches, corporate firewalls, and proxy servers. Therefore,some heuristics are commonly used to help identify uniqueusers. We use the machines IP addresses to identify uniqueusers.
User-session identification For logs that span a longperiod of time, it is very likely that different users may usethe same machine to access the server Web sites. Therefore,we differentiate the entries into different user-sessionsthrough a session timeout. That is, if two time stamps betweenpage requests exceeds a certain limit, we assume thepages are requested by two different user-sessions, eventhough the IP address is the same.Transaction identification The transactions are identifiedusing maximal forward references. Each time a backwardreference is made, a transaction is identified. A newforward reference indicates the next transaction for that session

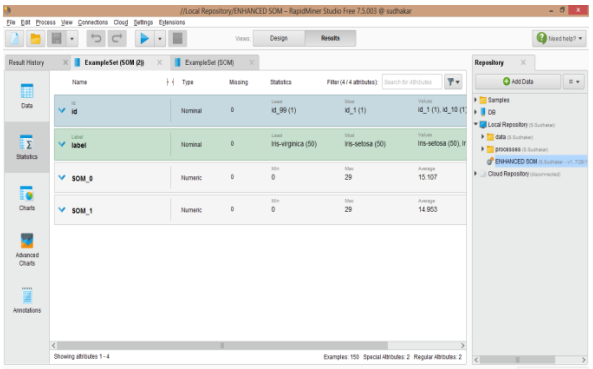
3.3 Web Page Mapping

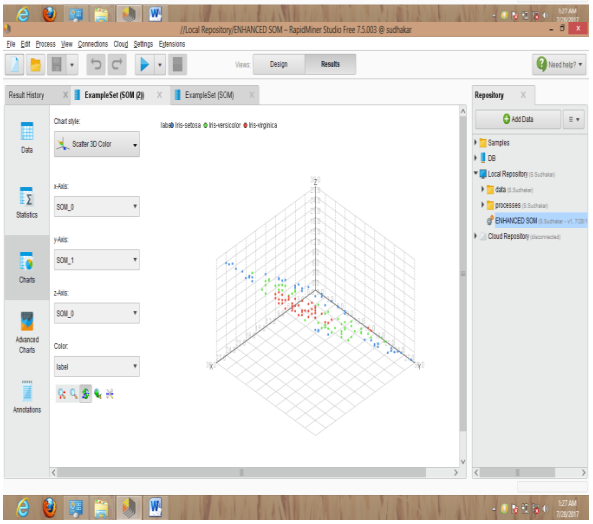
K-Means Clustering After the user sessions andtransactions are identified, we make a two-dimensional arrayin which each row is arranged for a transaction and eachcolumn is for a URL. Initially, the URLs that appear in atransaction are set to one in the corresponding row, and restvalues are set to zero.Initially, k transactions are selected at random for the$k$ clusters. Then the means of the k clusters will be calculated.Afterwards,the distance between every transactionand the k clusters is calculated using the means of the $k$clusters. A transaction will be grouped into the cluster towhich the distance is the shortest.For each of these k clusters, we sum up the values ofeach column and calculate its new mean. The mean valuesare used as the weights for the groups, which are used toindicate the similarity between groups. The algorithm willbe repeated until the weights become stable.SOMThe$k$ groups of transactions and the set ofunique URLs are the input to the SOM network. The inputis represented by a two-dimensional m by k matrix, where$m$ is the number of unique URLs and k is the number oftransaction groups.

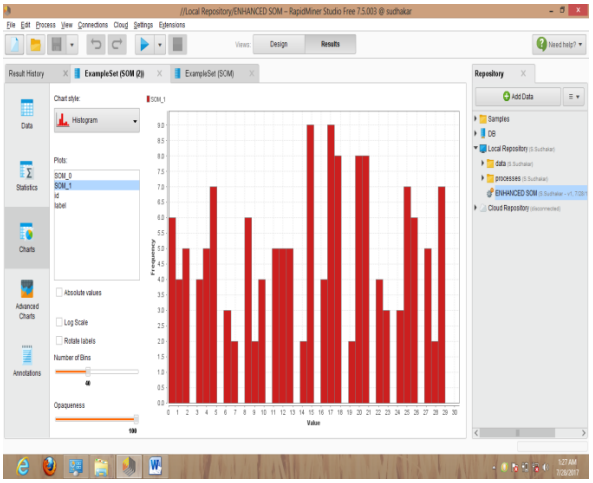
4 Experimental Results

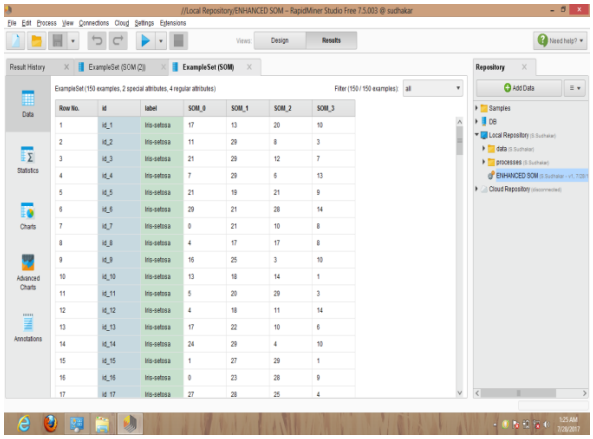
We used Web log file for October, 2006 from theas our test data. The data size is about30MB with about 300,000 entries. Table 1, 2 and 3 showsthe example of user identifications, session identifications,and transaction identifications.The number of unique URLs generated by preprocessingis 188. We used a fixed value of 20 as the numberof clusters, so the input to the SOM network is a 188by 20 array. We have tested different parameters for theSOM network as follows: ® varies from 0.2 to 0.9 and*!*











varies from 1 to 40 where ® represents the learning rate and*!* determines the number of times a URL being presentedwithin one learning cycle before the neighborhood size isdecreased. In our algorithm, there are 18 learning cyclesfor organizing the Web pages. In particular, we decreasedthe neighborhood size from its initial value of 17 to 0. Table4 and 5 shows the SOM map with (® = 0.1, *!*= 40) and(® = 0.5, *!*= 40), respectively.From our experimental results, we find that, with*!*= 40, the two-dimensional array maps display clearestcontesting. Table 6 shows part of the clusters with ® = 0.1and *!*= 40. The SOM mapping self cluster the web pagewithout prior knowledge.To assess the effectiveness of our approach, we inspectedthe SOM map. We find that the approach indeed resultsa very meaningful SOM network in the sense that theWeb pages are organized into clusters based on the similarityof their usage. Within a cluster, we can see that users are indeed likely to navigate Web pages within the same node,even though the SOM was given no information about thedirectory structure of the server and the contents of theWebpages. The SOM network has placed Web pages togetherwhen they are commonly accessed by the users in the sametransactions.Although it has been proven that clustering Webpages based on their contents is very effective and useful,it may be more advantageous to organize the Web pagesin a user-pattern-based clustering. In such a way, the Webpages are organized for humans to search in a more effectiveand efficient manner due to its simplicity. Analysis the usage patterns of Web users can play an important role in assisting other users.

5 Conclusions

We introduced a Self-Organizing Map (SOM) approachto the study of mining Web log data. Starting fromthe raw Web log data that is available in any Web server,we preprocessed it into distinct user transactions. We usedthe classical *k*-means algorithm to classify the URLs intoclusters based on users' browsing history. The experimentalresults based on the data from the Web log of the serverof our CS department demonstrate that our approach is very useful is a specified domain. The results of the clusters generated form the SOM network shows that our approach caneffectively discover usage patterns. Our results can also beused to predict the user's browsing behavior based on thepast experience.

0	2	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0
0	3	2	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	2	0	0	0	0	2	0	0
0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
2	0	0	0	2	0	5	2	0	0	0	0	1	0	3	0	0	0
0	8	0	1	0	2	0	0	0	33	0	0	0	2	4	1	0	0
0	2	0	0	1	3	0	0	22	0	0	0	0	0	0	0	0	0
0	0	0	0	2	0	6	0	0	0	0	0	0	2	0	0	0	0
0	1	2	1	1	1	3	0	0	1	0	0	0	3	0	0	0	0
0	0	0	0	1	4	6	0	12	1	1	119	0	0	0	0	1	0
2	0	0	0	0	0	0	2	0	1	0	1	0	0	0	1	0	1
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 4: SOM map with α=0.1 and ω=40

0	2	2	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
0	0	1	0	0	0	0	0	0	0	0	0	0	0	2	0	0	0
0	0	1	0	0	0	0	0	0	3	0	0	0	0	0	0	0	0
0	0	116	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0
0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
0	6	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
0	35	3	2	0	1	3	0	0	1	0	0	0	0	1	3	0	0
0	2	0	0	0	1	0	2	0	0	0	0	0	0	3	1	1	0
24	12	0	0	0	0	0	0	0	0	0	0	0	0	4	0	1	0
1	4	6	0	16	3	0	0	0	0	0	0	0	0	0	0	2	0
0	2	6	0	0	5	2	1	0	0	0	0	0	0	0	0	0	0
2	1	2	2	4	0	0	0	0	0	0	0	0	0	0	0	0	0
0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 5: SOM map with α=0.5 and ω=40

References

[1] M. Baglioni, U. Ferrara, Romei A., S. Ruggieri, andF. Turini.Preprocessing and mining web log data forweb personalization.In *the 8th Natational Conferenceof the Italian Association for Artificial Intelligence*,2003.

[2] X. Huang, F. Peng, A.An, and D. Schuurmans.Dynamicweb log session identification with statisticallanguage models.*Journal of the American Society forInformation Science and Technology*, 55(14):1290–1303, 2004.

[3] X. Jin, Y. Zhou, and B. Mobasher. Web usage miningbased on probabilistic latent semantic analysis.In *Proceedings of the ACM SIGKDD Conference onKnowledge Discovery and Data Mining*, Seattle,WA,2004.

[4] T. Kohonen. *Self-Organization and Associative Memory*.Springer-Verlag, New York, 1988.

[5] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi,J. Honkela, V. Paatero, and A. Saarela. Self organizationof a massive document collection.*IEEE Transactionson Neural Networks*, 11(3):574–585, 2000.

[6] J. Liu, S. Zhang, and J. Yang.Characterizing webusage regularities with information foraging agents.*IEEE Transactions on Knowledge and Data Engineering*,2004(16):566–584, 2004.

[7] Rosa Meo, Pier Luca Lanzi, Maristella Matera, andRoberto O Esposito. Integrating web conceptual modelling and web usage mining. In *Proceedings of thesixth WEBKDD workshop: Webmining and Web UsageAnalysis*, pages 105–115, Seattle, WA, 2004.

[8] B. Mobasher, H. Dai, and M. Tao. Discovery andevaluation of aggregate usage profiles for web personalization.*Data Mining and Knowledge Discovery*,6:61–82, 2002.

[9] M. Nakagawa and B. Mobasher. A hybrid web personalization model based on site connectivity. In*Proceedings of the International Workshop on WebKnowledge Discovery and Data Mining*, pages 59–70, 2003.

[10] O. Nasraoui and C. Petenes.Combining web usagemining and fuzzy inference for website personalization.In *Proceedings of the InternationalWorkshop onWeb Knowledge Discovery and Data Mining*, pages37–46, 2003.

[11] Kate A. Smith and Alan Ng. Web page clustering usinga self-organizing map of user navigation patterns.*Decision Support Systems*, 35(2):245–256, 2003.

[12] JaideepSrivastava, Robert Cooley, MukundDeshpande, and Pang-Ning Tan. Web usage mining: Discoveryand applications of usage patterns from webdata.*SIGKDD Explorations*, 1(2):12–23, 2000.

[13] Zhong Su, Qiang Yang, Hong-Jiang Zhang, XiaoweiXu, and Yu-Hen Hu. Correlation-based documentclustering using web logs. In *34th Hawaii InternationalConference On System Sciences*, pages 5022–5027, Hawaii, 2001. IEEE Computer Society.

[14] A. Ypma and T. Heskes. Categorization of web pagesand user clustering with mixtures of hidden markovmodels. In *Proceedings of the International Workshopon Web Knowledge Discovery and Data Mining*,Edmonton, Canada, 2002.

[15] S. Sharma, M Varshney, "An Efficient approach for web log mining using ART", *International Conference on Education and Management Technology*, 2010 (ICEMT 2010).

[16] Zhang Y.,X. Yu, and J. Hou, "Web communities: Analysis and construction," *Berlin Heidelberg*, 2006.

[17] S. Chakrabarti, M. van den Berg, and B. Dom, "Focused crawling: A new approach to topic-specific web resource discovery," presented at the *8th World Wide Web Conf*., Toronto, ON, Canada, May 1999.

[18] N Tyagi,A. Solanki and S. Tyagi, "An algorithmic approach to data preprocessing in web usage mining", *Int. journal ofInformation technology and knowledge management*, July-December 2010, Volume 2, No. 2, pp. 279-283.

[19] R. Cooley, B. Mobasher, and J. Srivastava. "Web mining: Information and pattern discovery on the World Wide Web", *Technical Report TR 97-027*, University of Minnesota, Dept. of Computer Science, Minneapolis, 1997.

20. Aguilar.J.S, Ruiz.R, Riquelme J.C and Gir´aldez.R, (2001) SNN:A Supervised Clustering Algorithm, in: 14th *Int. Conf. on Industrial & Engineering Applications of Artificial Intelligence & Expert Systems (IEA/AIE 2001): Lecture Notes in Artificial Intelligence, Springer-Verlag*, Budapest, Hungary, June 4–7,2001, pp. 207–216.

21. Cooley.R. (2000) Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. PhD thesis, University of Minnesota, May 2000.

22. Graham .J and Starzyk J, (2008) A Hybrid self Organizing Neural Gas Network , *IEEE World Conference on Computational Intelligence (WCCI'08) June 1-6*

23. Jirayusakul .A and Auwatanamongkol .S (2000) A supervised growing neural gas algorithm for cluster analysis.

24. Kohonen.T,(1995). Self-Organizing Maps, Berlin, Germany: *Springer*,

25. Martinet.Mz, Berkovich.S and Schulten.K, (1993). Neural-gas network for vector quantization and its application to time series prediction, *IEEE Trans. Neural Networks* 4(1993)558-569.

26. Pedrycz.W and Vukovich.G, (2004) Fuzzy clustering with supervision, Pattern Recognition **37**(7), 1339–1349.

27. Qu.Y and Xu.S, (2004) Supervised cluster analysis for microarray data based on multivariate Gaussian mixture, Bioinformatics **20**(12), 1905–1913.

28. Slonim.N and Tishby.N, (1999) Agglomerative information bottleneck, in: *Proceedings of the 13th Neural Information Processing Systems*, (NIPS).

29. SonaliMuddalwar and ShashankKawar, (2012) Applying Artificial neural network in Web Usage Mining, International Journal of Computer Science and Management Research Vol 1 Issue 4 November 2012.

30. Yu.F, Sandhu..K, and Shih .M. (2000) A generalization-based approach to clustering of web usage sessions. In Proc. of the 1999 KDD