ISSN 2394-3777 (Print)
ISSN 2394-3785 (Online)
Available online at www.ijartet.com

International Journal of Advanced Research Trends in Engineering and Technology (IJARTET)
Vol. 4, Special Issue 4, March 2017

Enhanced Self Organizing Map Algorithm for Web Usage Mining Through Neural Network

C.Sadhana
Research Scholar
St Peters University,

Dr.L.Mary Immaculate Sheela
Professor
Department of Computer Application
R.M.D Engineering College

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.

Keywords— web mining, SOM, clustering, Classification

INTRODUCTION

The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized that they can be accessed efficiently. Therefore the application of data mining techniques on the Web is now the focus of an increasing number of researchers. Several data

Mining methods are used to discover the hidden information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified to better suit the demands of the Web. New approaches should be used better fitting to the properties of Web data. Furthermore, not only data mining algorithms, but also artificial intelligence, information retrieval and natural language processing techniques can be used efficiently. Thus, Web mining has been developed into an autonomous research area. Web mining involves a wide range of applications that aim at discovering and extracting hidden information in data stored on the Web. Another

Important purpose of Web mining is to provide a mechanism to make the data access more efficiently and adequately. The third interesting approach is to discover the information which can be derived from the activities of users, which are stored in log files for

example for predictive Web caching [1]. Thus, Web mining can be categorized into three different classes based on which part of the Web is to be mined [2], [3], and [4]. These three categories are Web content mining, Web structure mining and Web usage mining. Web content mining [7], [6] is the task of discovering useful information available on-line. There are different kinds of Web content which can provide useful information to users, for example multimedia data, structured (i.e. XML documents), semi

structured (i.e. HTML documents) and unstructured data (i.e. plain text). The aim of Web content mining is to provide an efficient mechanism to help the users to find the information they seek. Web content mining includes the task of organizing and clustering the documents and providing search engines for accessing the different documents by keywords, categories, contents. Web structure mining is the process of discovering the structure of hyperlinks

within the Web. Practically, while Web content mining focuses on the inner-document information, Web structure mining discovers the link structures at the inter-document level. The aim is to identify the authoritative and the hub pages for a given subject. Web usage mining is the task of discovering the activities of the users while they are browsing and navigating through the Web. The aim of understanding the navigation preferences of the visitors is to enhance the quality of electronic commerce services (ecommerce), to personalize the Web portals [7] or to improve the Web structure and Web server performance [3]. For this reason a model of the users (User Model - UM) have to be built based on the information gained from the log data.

A. *Abbreviations and Acronyms*

Self-Organizing Map (SOM)

Web Usage Mining

The aim of Web usage mining is to discover patterns of user activities in order to better serve the needs of the users for example by dynamic link handling, by page recommendation etc. The aim of a Web site or Web portal is to supply the user the information which is useful for him. There is a great competition between the different commercial portals and Web sites because every user means eventually money (through advertisements, etc.).

Thus the goal of each owner of a portal is to make his site more attractive for the user. For this reason the response time of each single site have to be kept below 2s. Moreover some extras have to be provided such as supplying dynamic content or links or recommending pages for the user that are possible of interest of the given user. Clustering of the user activities stored in different types of log files is a key issue in the Web community. There are three types of log files that can be used for Web usage mining [4]. Log files are stored on The server side, on the client side and on the proxy servers. By having more than one place for storing the information of navigation patterns of the users makes the mining process more difficult. Really reliable results could be obtained only if one has data from all three types of log file. The reason for this is that the server side does not contain records of those Web page accesses that are cached on the proxy servers or on the client side. Besides the log file on the server, that on the proxy server provides additional Information. However, the page requests stored in the client side are missing. Yet, it is problematic to collect all the information from the client side. Thus, most of the algorithms work based only the Server Side data. Web usage mining consists of three main steps:

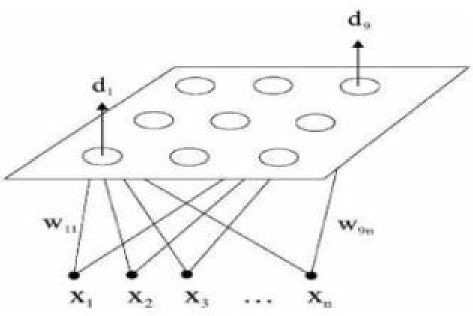
(i) Pre-processing (ii) Pattern discovery (iii) Pattern analysis

In the pre-processing phase the data have to be Collected from the different places it is stored (client side, server side, proxy servers). After identifying the users, the click-streams of each user has to be split into sessions. In general the timeout for determining a session is set to 30 minute [5]. The pattern discovery phase means applying data mining techniques on the pre-processed log data. It can be frequent pattern mining, association rule mining or clustering. In this paper we are dealing only with the task of clustering web usage log. In web usage mining there are two types of clusters to be discovered: usage clusters and page clusters. The aim of clustering users is to establish groups of users having similar browsing behaviour. The users can be clustered based on several Information. In the one hand, the user can be requested filling out a form regarding their interests, for example when registration on the web portal. The clustering of the users can be accomplished based on the forms. On the other hand, the clustering can be made based on the information gained from the log data collected during the user was navigating through the portal. Different types of user data can be collected using these methods, for example (I) characteristics of the user (age, gender, etc.), (ii) preferences and interests of the user, (iii) user's behaviour pattern. The aim of clustering web pages is to have groups of pages that have similar content. This information can be useful for search engines or for applications that create dynamic index pages. The last step of the whole web usage mining process is to analyze the patterns found during the pattern discovery step. Web Usage Mining try to understand the patterns detected in before step. The most common techniques is data visualization applying filters, zooms, etc

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.
        From weight vectors (wj) associated with each output node.
        ti, k - wj, k ( t ) )2
    iii) Select output node j* that has weight vector with
        minimum value from step 2.
    iv) Update weights to all nodes within a topological
        distance given 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.
```

- Related Algorithms
- 2.1 SOM
- Teuvo Kohonen [4] introduced the SOM network that reduced the dimensions of data through the use of selforganizing neural networks. The SOM network produces a map of usually one or two dimensions which plot the similarities of the data by grouping similar data items together.This mapping process reduces the problem dimensions.The SOM network integrates dimensions reducing and clustering in one network. Figure 1 shows the mapping from a one-dimensional input to a two-dimensional array.
- Example of a figure SOM. *(SOM network)*



- Figure 1: The Mapping from a one-dimensional input to a two-dimensional array [11].The SOM network organizes itself by competing representation of the samples. Neurons are also allowed to change themselves in hoping to win the next competition. This selection and learning process makes the weights to organize 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 the weight vectors. From there the algorithms select a sample vector randomly and search the map of weight vectors to find the weight that can represent the sample best. Since each weight vector has a location, it also has neighbouring weights that are close to it. The chosen weight is rewarded to perform better than a randomly selected sample vector. In addition to this reward, the neighbours of the weight are also rewarded. From this step we increase t some small
- amount because the number of neighbours and how much each weight can learn decreases over the time. This whole process is then repeated a large number of times, usually at least 1000 times.
- The main advantage of using the SOM network is that SOM automatically (self-organizing) clusters documents. The SOM network also can be applied to a large scale of data.

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 that points are mutually farthest apart. Next, it examines each point and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated every time a point is added to the cluster. This process will be repeated until all the points are grouped into the *k* clusters. However, this algorithm does not work well if there are large differences in the data set. The equation for kmeans algorithm 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*'s co-ordinates and *uij* represents the hypothetical belonging of 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.

3.2 Data Pre-processing

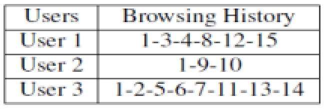
There are several pre-processing tasks to be done before executing the data mining algorithms on the Web server logs. These processes include data formatting, user identification, session identification, and transaction identification.The original server logs are formatted and grouped into meaningful transactions before being processed by the mining system. We describe each of these processes in the following paragraphs. Data formatting The access log is saved to keep a record of every request made by the users. Since our main purpose 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. Some irrelevant 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 organizing the existing Web URLs, we are not interested in this kind of error entries, and hence these error entries shall be deleted. A users request to view a particular page often results 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, what the user explicitly requests because we intend to design a system that is user-oriented.
User identification The task of identifying unique users is greatly complicated by the existence of local caches, corporate firewalls, and proxy servers. Therefore, some heuristics are commonly used to help identify unique users. We use the machines IP addresses to identify unique users.
User-session identification For logs that span a long period of time, it is very likely that different users may use the same machine to access the server Web sites. Therefore, we differentiate the entries into different user-sessions through a session timeout. That is, if two time stamps between page requests exceeds a certain limit, we assume the pages are requested by two different user-sessions, even though the IP address is the same.Transaction identification The transactions are identified using maximal forward references. Each time a backward reference is made, a transaction is identified. A new forward reference indicates the next transaction for that session

3.3 Web Page Mapping

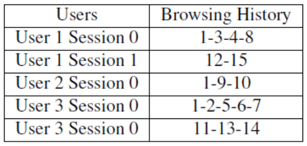
K-Means Clustering After the user sessions and transactions are identified, we make a two-dimensional array in which each row is arranged for a transaction and each column is for a URL. Initially, the URLs that appear in a transaction are set to one in the corresponding row, and rest values are set to 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 transaction and the *k* clusters is calculated using the means of the *k* clusters. A transaction will be grouped into the cluster to which the distance is the shortest.For each of these *k* clusters, we sum up the values of each column and calculate its new mean. The mean values are used as the weights for the groups, which are used to indicate the similarity between groups. The algorithm will be repeated until the weights become stable.SOM The *k* groups of transactions and the set of unique URLs are the input to the SOM network. The input is represented by a two-dimensional *m* by *k* matrix, where *m* is the number of unique URLs and *k* is the number of transaction groups.

4 Experimental Results



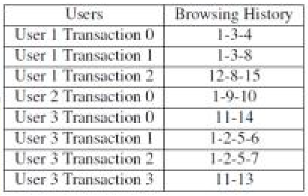
Users	Browsing History
User 1	1-3-4-8-12-15
User 2	1-9-10
User 3	1-2-5-6-7-11-13-14

Table 1: User Identification



Users	Browsing History
User 1 Session 0	1-3-4-8
User 1 Session 1	12-15
User 2 Session 0	1-9-10
User 3 Session 0	1-2-5-6-7
User 3 Session 0	11-13-14

Table 2: Session Identification



Users	Browsing History
User 1 Transaction 0	1-3-4
User 1 Transaction 1	1-3-8
User 1 Transaction 2	12-8-15
User 2 Transaction 0	1-9-10
User 3 Transaction 0	11-14
User 3 Transaction 1	1-2-5-6
User 3 Transaction 2	1-2-5-7
User 3 Transaction 3	11-13

Table 3: Transaction Identification

Cluster Number	Web Pages
8	901 902 903 904 905
10	8 21 23 89 90 133 134 136 168 180 284 285 286 288 289 290 313 319 328 337 338 343 344 351 357 359 374 392 393 394 399 406 410 416 421 434 442 448 454 455 456 466 480 487 498 499 513 583 593 789 1212 1230 1292
11	88 92 130 132 135 141 166 167 177 179 190 191 194 202 211 212 213 303 320 321 325 326 339 342 345 356 368 384 385 386 387 388 391 397 398 403 404 405 409 411 412 413 415 417 418 420 422 427 430 431 432 433 446 447 449 451 452 453 479 490 496 497 503 508 512 515 530 788 846 978 1213 1229 1259 1260 1288 1291 1293 1342
13	127 151 155 156 159 231 232 279 280 281 291 307 308 323 348 360 381 382 383 389 390 441 814 1273 1274 1275 1276 1277 1278 1286 1287 1289

Table 6: Part of clusters with α=0.5 and ω=40

We used Web log file for October, 2006 from the as our test data. The data size is about 30MB with about 300,000 entries. Table 1, 2 and 3 shows the example of user identifications, session identifications, and transaction identifications.The number of unique URLs generated by preprocessing is 188. We used a fixed value of 20 as the number of clusters, so the input to the SOM network is a 188 by 20 array. We have tested different parameters for the SOM 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 presented within one learning cycle before the neighborhood size is decreased. In our algorithm, there are 18 learning cycles for organizing the Web pages. In particular, we decreased the neighborhood size from its initial value of 17 to 0. Table 4 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 clearest contesting. Table 6 shows part of the clusters with ®

= 0.1 and ! = 40. The SOM mapping self cluster the web page without prior knowledge.To assess the effectiveness of our approach, we inspected the SOM map. We find that the approach indeed results a very meaningful SOM network in the sense that the Web pages are organized into clusters based on the similarity of 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 the directory structure of the server and the contents of theWeb pages. The SOM network has placed Web pages together when they are commonly accessed by the users in the same transactions. Although it has been proven that clustering Web pages based on their contents is very effective and useful, it may be more advantageous to organize the Web pages in a user-pattern-based clustering. In such a way, the Web pages are organized for humans to search in a more effective and efficient manner due to its simplicity. Analysis the usage patterns of Web users can play an important role in assisting other users.

Conclusions

We introduced a Self-Organizing Map (SOM) approach to the study of mining Web log data. Starting from the raw Web log data that is available in any Web server, we preprocessed it into distinct user transactions. We used the classical *k*-means algorithm to classify the URLs into clusters based on users' browsing history. The experimental results based on the data from the Web log of the server of 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 can effectively discover usage patterns. Our results can also be used to predict the user's browsing behavior based on the past experience.

References

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

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

[3] X. Jin, Y. Zhou, and B. Mobasher. Web usage mining based on probabilistic latent semantic analysis. In *Proceedings of the ACM SIGKDD Conference on Knowledge 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 organization of a massive document collection. *IEEE Transactions on Neural Networks*, 11(3):574–585, 2000.

[6] J. Liu, S. Zhang, and J. Yang. Characterizing web usage 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, and Roberto O Esposito. Integrating web conceptual modelling and web usage mining. In *Proceedings of the sixth WEBKDD workshop: Webmining and Web Usage Analysis*, pages 105–115, Seattle, WA, 2004.

[8] B. Mobasher, H. Dai, and M. Tao. Discovery and evaluation 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 Web Knowledge Discovery and Data Mining*, pages 59– 70, 2003.

[10] O. Nasraoui and C. Petenes. Combining web usage mining and fuzzy inference for website personalization. In *Proceedings of the InternationalWorkshop on Web Knowledge Discovery and Data Mining*, pages 37–46, 2003.

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

[12] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000.

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

[14] A. Ypma and T. Heskes. Categorization of web pages and user clustering with mixtures of hidden markov models. In *Proceedings of the International Workshop on 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. Sonali Muddalwar and Shashank Kawar, (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