Missouri University of Science and Technology

Scholars' Mine

Engineering Management and Systems Engineering Faculty Research & Creative Works

Engineering Management and Systems Engineering

01 Jan 2003

# Web Personalization using Neuro-Fuzzy Clustering Algorithms

Kartik Menon

Cihan H. Dagli
*Missouri University of Science and Technology*, dagli@mst.edu

## Recommended Citation

K. Menon and C. H. Dagli, "Web Personalization using Neuro-Fuzzy Clustering Algorithms," *Proceedings of the 22nd International Conference of the North American Fuzzy Information Processing Society, 2003*, Institute of Electrical and Electronics Engineers (IEEE), Jan 2003.
The definitive version is available at https://doi.org/10.1109/NAFIPS.2003.1226840

# Web Personalization using Neuro-Fuzzy Clustering Algorithms

Kartik Menon and Cihan H. Dagli
Smart Engineering Systems Laboratory
University of Missouri – Rolla
kartik@umr.edu, dagli@umr.edu

## Abstract

*Different users have different needs from the same web page and hence i is necessary to develop a system which understands the needs and demands of the users. Web server logs have abundant information about the nature of users accessing it. In this paper we discussed how to mine these web server logs for a given period of time using unsupervised and competitive learning algorithm like Kohonen's self organizing maps (SOM) and interpreting those results using Unified distance Matrix (U-matrix).These algorithms help us in efficiently clustering users based on similar web access patterns and each cluster having users with similar browsing patterns. These clusters are useful in web personalization so that it communicates better with its users and also in web traffic analysis for predicting web traffic at a given period of time.*

## 1. Introduction

Designing a web site is a complex problem. Different people have different needs and requirements from the web site. Even a single user can have different goals from the same web site. A website may have been designed for one primary purpose but maybe used in unexpected ways. Consider for example the website www.yahoo.com, it was started primarily as an email server but over the span of time is used by different people for different reasons other than just emails. Hence, an adaptive (self evolving) web site [11][12] are site that improve themselves by learning from user access patterns would be quite helpful in this corporate world of ease and comfort to make web pages more accessible, highlight interesting links, connect related pages, and cluster similar documents together. This process of creating an adaptive website is called web personalization.

For web personalization there is a constant need to understand the users and cluster users having similar web traversal patterns. Fuzzy clustering algorithms like fuzzy c-means [1][2][3][5] and Kohonen's Self Organizing Maps (SOM) [6][7][8] help us in clustering the users into different categories. This paper discusses how Kohonen's SOM could be used in our web mining architecture for generating clusters. The learning process being competitive and unsupervised, help us in categorizing the users without any prior knowledge about the user. The paper discusses a graphic visualization tool called Unified Distance Matrix (U-Matrix) [13] for interpreting results generated by Kohonen's SOM. This tool is helpful in identifying clusters which otherwise would have been difficult to identify using a crude SOM on any data using human eye. This cluster information generated by SOM and U-Matrix can be used for web personalization and in predicting web traffic as well as user behavior in a particular day [14][15]. These clusters also help in predicting the nature of the users in a given period of time depending on his previous web access patterns.

## 2. Web Personalization Architecture

The web personalization architecture is clearly explained in Figure 1. Web users interact directly with the server maintained by the owner of the web site.

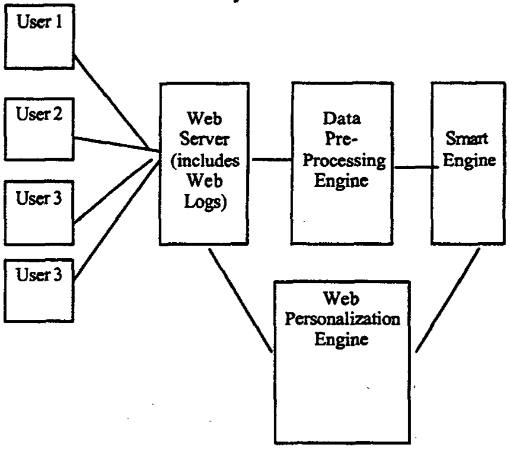
User 1
User2
User 3
User3
Web Server (includes Web Logs)
Data Pre-Processing Engine
Smart Engine
Web Personalization Engine

Figure 1: Web Personalization Architecture

0-7803-7918-7/03/$17.00 © 2003 IEEE

As a result the behavior of the user is recorded in the web server logs. These web logs provide an abundant opportunity to observe users interacting with that site. Web log data is crude and hence needs to be pre-processed. A typical Apache Web log record looks like:- 131.151.83.11 [26/Apr/2003:17:36:34 -0500] "GET /web_mining.html HTTP/1.1" 200 1982 "-" "Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; .NET CLR 1.0.3705)" where131.151.83.11 is the machine address of the user accessing the webpage, 26/Apr/2003:17:36:34 is the date and time the user accesses the webpage, GET is the get or post methods, web_mining.html is the webpage accessed. This log record can be used for automated analysis to create adaptive websites. This data is pre-processed in the Data Preprocessing engine for filtering , numerical modeling, smoothing and data cleaning.

This pre-processed data is fed into the Smart Engine where various clustering algorithms like Fuzzy c means and Kohonen's SOM could be used to identify natural groupings of data. After 'learning' about the user access pattern using Smart tools the data is fed into the Web Personalization Engine for creating adaptive web sites in other words rearrange the web site to improve the structure and presentation. Web Personalization Engine also will predict the web traffic as well as the nature of users accessing the web site during given period of time.

In this paper we will be concentrating primarily the on the Smart Engine and how Kohonen's SOM can be used on the preprocessed data to identify clusters with the given set of users in a given period of time.

## 3. Unified Distance Matrix (U-Matrix)

U-matrix representation of the Self-Organizing Map visualizes the distances between the neurons. The distance between the adjacent neurons is calculated and presented with different colorings between the adjacent nodes. A dark coloring between the neurons corresponds to a large distance and thus a gap between the codebook values in the input space. A light coloring between the neurons signifies that the codebook vectors are close to each other in the input space. Light areas can be thought as clusters and dark areas as cluster separators. This can be a helpful presentation when one tries to find clusters in the input data without having any a priori information about the clusters.

*Algorithm*

```
Step 1: Take the n*n*m weight matrix where n is the size of the topological map and m is number of attributes
Step 2:  For all neurons (i=1 ..n, j=1..n) do
Step3:  Calculate vector difference to all eight neighboring neurons on the map to
        [(i-1, j-1), (i-1, j), (i+1, j), (i, j-1), (i, j+1), (i+1, j-1), (i+1, j), (i+1, j+1)]
Step 4: sum it up
Step 5: divide through eight
Step 6: store the result
Step 7: get n*n*1 matrix
```

The results of U-Matrix can be used to easily interpret results of SOM which otherwise is difficult to interpret. The shade differences indicate cluster difference. These visualizations help in cluster identification.

## 4 Data Preprocessing Engine

In this Engine each web document is parsed for web links to other documents, which in turn were parsed sequentially in order to gather the links which can be traversed from that particular document. Each web document is assigned a weight scalar value according to its parent document. The root webpage has a weight of 1 and all the webpage it refers to directly have a weight having value having starting with 1 and decimal in accordance to its order of placement in the root page . This is carried on for the entire list of web documents in the web server. In this way there are no two web pages with same weights. Figure 3 illustrates how weights are allocated to each web document.

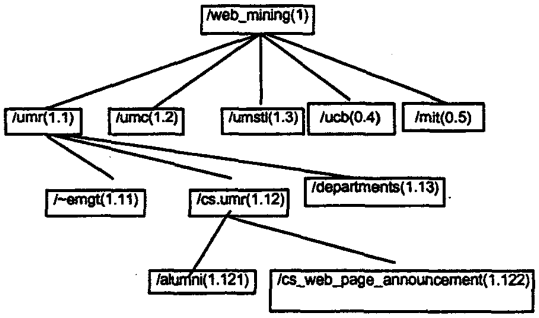

Figure 2: Web Document Weight Hierarchy

After assigning each web page a weight, the web logs are parsed to capture information about the user. Apache web server was used to provide the web services and each web document was stored in the web server so that if the user goes to a web document is can be traced by mining the web logs. Each record is scanned to parsed to get following parameters which is served as a feature vector to the SOM algorithm: the total weight which is the sum of weights of all the web page the user traverses i.e. $\sum weight$ , number of web pages the user traverses, time being the total time the user takes on a particular web page and its children i.e. $\sum time$ and weight-time is weighted time i.e. $\sum weight \times time$ .

The web log file consisted of 8 records having different web traversal patterns for each of these 8 users. The log file was parsed to gather information about the 8 users. The results of the pre-processing engine generated the feature vector which is shown as Table 1.

Table 1. Feature Vector for SOM

| User | Total Weight | Number of URLs | Time | Weight-Time |
|---|---|---|---|---|
| 1 | 5.32 | 5 | 130 | 133.32 |
| 2 | 3.8 | 3 | 15 | 18 |
| 3 | 7.92 | 6 | 31 | 41.14 |
| 4 | 6.69 | 6 | 19 | 22.18 |
| 5 | 6.1 | 6 | 15 | 19.44 |
| 6 | 4.92 | 4 | 13 | 13.82 |
| 7 | 7.57 | 7 | 16 | 20.61 |
| 8 | 9.6 | 7 | 36 | 49.4 |

## 5. Results and Analysis of Smart Engine

In the Smart Engine the preprocessed data is analyzed using several smart and intelligent tools like Fuzzy Logic and Artificial Neural Networks. We used the Kohonen's SOM in this engine to cluster the data. The algorithm takes the feature vector of Table 1 as its input. Repeated testing were conducted for same sets of data with different by changing the input parameters like the neuron grid dimensions, learning parameters, topology and neighborhood distance. The algorithm provided best results for "grid top" topology, learning parameter of 0.9, number of epochs is 1000 and neighborhood distance of 1 which is calculated as the Euclidean distance. Figure 3 shows that the feature vector mapping to the 6 x 6 neuron grid. The testing data used for identification of clusters is same as the training data which is also the original data set.

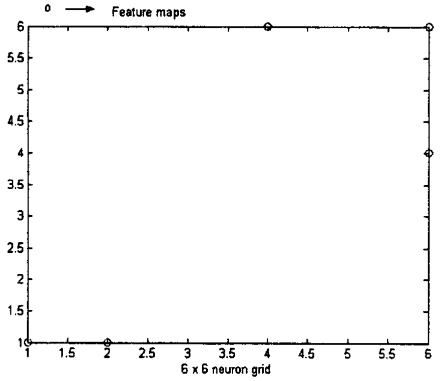

Figure 3: Feature VectorNeuron Map for 6 x 6 neuron grid

The weights generated by SOM were interpreted using U-Matrix as explained in section 3 to give results as shown in Figure 4 for same 6 x 6 neuron grid. For simplicity purposes the users were categorized into three clusters. Each cluster constituting of users having a similar web access pattern. The cluster scale was decided on the basis of the weights provided by the SOM and U-Matrix. We see that from the figure that for 6 x 6 neuron grid the total weight value for all the neurons ranges from 0-4500, we can divide the U-Matrix results in 3 clusters 0-1500 being cluster1, 1501-3000 being cluster2 and 3001-4500 being cluster3.

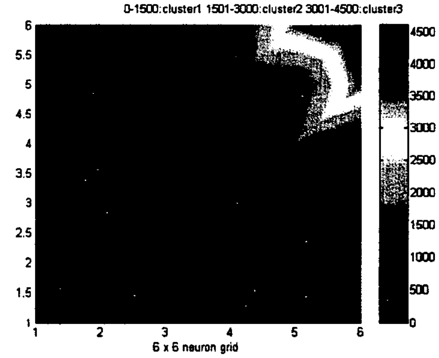

Figure 4: SOM results for 6 x 6 neuron grid for 6 x 6 neuron grid

These clusters can also be visually separated as wherever there exists a change in the shading there is a cluster difference in that region. For 5 x 6 neuron grid the feature vector to neuron mapping is shown in Figure 5 and the clusters are shown in Figure 6.

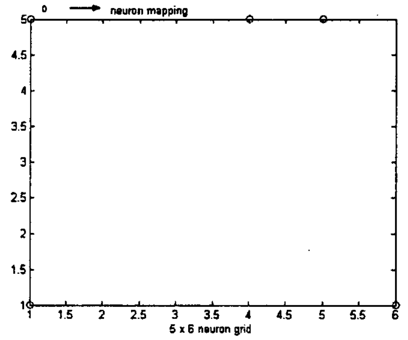

Figure 5: Feature Vector Neuron Map for 5 x 6 neuron grid

By juxtaposing the neurons mappings to feature vector as well as the cluster differences as shown by the U-Matrix results it becomes easy to identify which cluster each of the eight users belongs to. The cluster identification of the users as generated by the 6 x 6 neuron grid is shown in Table 2 and for 5 x 6 neuron grid is shown in Table 3.

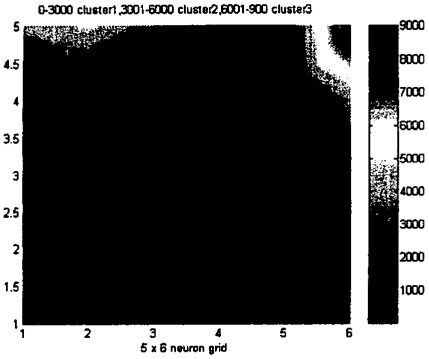

Figure 4: SOM results for 6 x 6 neuron grid for 6 x 6 neuron grid

Table 2. User clusters for 6 x 6 neuron grid

| User | Cluster |
|---|---|
| 1 | 3 |
| 2 | 3 |
| 3 | 3 |
| 4 | 1 |
| 5 | 3 |
| 6 | 1 |
| 7 | 3 |
| 8 | 2 |

Table 3. User cluster for 5 x 6 neuron grid

| User | Cluster |
|---|---|
| 1 | 3 |
| 2 | 3 |
| 3 | 3 |
| 4 | 1 |
| 5 | 3 |
| 6 | 1 |
| 7 | 3 |
| 8 | 1 |

As shown in the above tables provided that the remaining parameters being same the two sets of results have approximately 88% stability in their cluster identification. For the overall testing on the test data using different parameters provided approximately 50% stability in the different set of results.

## 5. Web Personalization Engine (WPE)

WPE takes the results and analysis provided by the Smart Engine to provide changes or modifications to the web page for the users to increase the popularity of the web page by tailoring it more to the needs of the users accessing it. Hence these methods will help the web master to develop adaptive web pages which can be self evolving based on the previous web traversal of the user. These cluster information can also help the WPE in predicting web traffic as well as user behavior in a particular day. For example based on this clustered information it is possible to find how many people will be accessing a particular web page a given day or given period of time depending on the past queries that each user has made on that page during that given period of time. The clusters can also help in finding the nature of the users in a given period of time depending on his previous web access patterns but that remains as future work. This is useful because certain group of people may use the web page more than others and hence needed to be allocated more bandwidth to avoid congestion. Most of the work in WPE remains as future work.

## 6. Conclusion and Future Work

In this paper we talked about how Kohonen's SOM and U-Matrix can be used in our architecture to cluster users based on their web access traversal patterns. SOM is an easy way of clustering similar web access patterns for different user sessions. The use of Euclidean distance was very helpful to learn more about these web access patterns. U-Matrix provided easy results and plots which was highly interpretable. We observe that that SOM and U-Matrix provided almost stable results which is 88% for the two case of 5 x 6 and 6x 6 neuron grid provided all other parameters same and 50% for overall test results. The results provided by Kohonen's SOM can be used in our WPE for web personalization to create adaptive web sites and in decisions for allocation of resources for users belonging to different clusters. These results also help in understanding how different groups of users interact with the particular website.

## 7. References

[1] Bezdek James C., "*Pattern Recognition with Fuzzy Objective Function Algorithms*". Plenum, New York, 1981.

[2] Bezdek James C. A Convergence "*Theorem for the Fuzzy ISODATA Clustering Algorithms*". TPAMI, 2(1):1--8, 1980.

[3] Bezdek James C., Hathaway Richard J., Sabin Michael J., and Tucker William T. "*Convergence Theory for Fuzzy c-Means: Counterexamples and Repairs*". SMC, 17(5):873--877, 1987.

[4]. Ham Fredric M., Kostanic Ivica, "Principles of neuro computing for science and engineering", McGraw Hill Publications, 2001

[5] Joshi A. and Krishnapuram R., *"On Mining Web Access Logs".* In Proceedings of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery 2000, pp. 63--69, 2000.

[6] Kohonen T.: "*Automatic Pattern Recognition*",A Challenge to Computer Technology. IFIP Congress 1980: 641-649

[7].Kohonen T.: The self-organizing map. Neurocomputing 21(1-3): 1-6 (1998)

[8] Lagus K, Honkela T, Kaski S, Kohonen T, "*Self-Organizing Maps of Document Collections: A New Approach to Interactive Exploration.*", KDD 1996: 238-243

[9] Mobasher B., Cooley R. and Srivastava J., "*Creating Adaptive Web Sites Through Usage-Based Clustering of URLs*", Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99), November 1999

[10] Paliouras G, Papatheodorou C., Karkaletsis V. and Spyropoulos C.D. , "*Clustering the Users of Large Web Sites into Communities*," Proceedings Intern. Conf. on Machine Learning (ICML), pp. 719-726, Stanford, California, 2000.

[11].Perkowitz M. and Etzioni O., "*Adaptive Web Sites: an AI Challenge*" , Proceedings of IJCAI,1997

[12] Perkowitz M. and Etzioni O. ,"*Adaptive Web Sites : Conceptual Cluster Mining*" , Proceedings of IJCAI,1999

[13] Vesanto J, Himberg J, Siponen M, Simula O." *Enhancing SOM based data visualization*", Proceedings of the 5th International Conference on Soft Computing and Information/Intelligent Systems. Methodologies for the Conception, Design and Application of Soft Computing

[14]Wang X., Abraham A. and Smith K.A., "*Web Traffic Mining Using a Concurrent Neuro-Fuzzy Approach*", 2nd International Conference on Hybrid Intelligent Systems, Soft Computing Systems: Design, Management and Applications, IOS Press Netherlands, pp. 853-862, 2002.

[15] Wang X., Abraham A. and Smith K.A, "*Soft Computing Paradigms for Web Access Pattern Analysis* ", Proceedings of the 1st International Conference on Fuzzy Systems and Knowledge Discovery, pp. 631-635, 2002.