Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 1, January 2015, pg.78 – 84

RESEARCH ARTICLE

WEB PAGE CLUSTERING USING SELF-ORGANIZING MAP

A.M.Sote[1], S.R.Pande[2]

[1]Department of Electronics and Computer Science RTM Nagpur University Nagpur, India

[2] Department of Computer Science, SSESA's Science College, Nagpur, India

[1] amar_sote@rediffmail.com; [2] srpande65@rediffmail.com

Abstract - The continuous growth in the size and use of the Internet is creating difficulties in the search for information. A sophisticated method to organize the layout of the information and assist user navigation is therefore particularly important. In this paper, we evaluate the feasibility of using a self-organizing map (SOM) to mine web log data and provide a visual tool to assist user navigation. The organization of the web pages is based solely on the user's navigation behaviour, rather than the content of the web pages. The resulting map not only provides a meaningful navigation tool (for web users) that is easily incorporated with web browsers, but also serves as a visual analysis tool for webmasters to better understand the characteristics and navigation behaviour of web users visiting their pages.

Keywords - Clustering, Self-Organising Map, Web Log File, Web Usage Mining

I. INTRODUCTION

The size and use of the internet thus creates difficulties in the search for information. as result, when a user enters a keyword in a search engine, the returned result is often a large list of web pages, many of which are irrelevant pages, moved pages, abandoned pages, etc. therefore, a sophisticated method to organize the layout of the information is important, particularly as the internet grows in size. The purpose of this study is to assist information retrieval on the internet by applying data mining techniques. Particularly, we focus on web usage mining, applying data mining techniques to web server logs. The use of data mining in this domain can be seen as the application of a new technology to an acknowledged problem.

Navigational aids were devised not long after the Internet became popular, including technologies such as Supercard that inspected the hypertext of where the user had been, and helped the user navigate through the hypertext based on this prior activity [1]. More recently though, researchers have been using the techniques of data mining to assist information retrieval on the Internet.

One of the most well-known data mining approaches is WEBSOM [2,3,4], which is a system using Kohonen's Self-Organizing Map [5,6] to organise web documents into a two-dimensional map, according to their document content. Documents which are similar in content are located in similar regions 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. For example, a group of web pages about data mining may appear in one cluster but it is unlikely that there are any pages on industrial mining (for example, oil or coal) in the cluster, because the self-organizing map (SOM) clusters by content rather than keyword. In addition, the location of the representing node indicates the closeness (similarity) of the documents represented.

There are several advantages to using the SOM to cluster documents, rather than people due to the objectivity of the process. In addition, the process is automatic (hence the name ''self-organizing''). It can thus be done on a large scale and therefore saves labour costs. It also facilitates search by concept instead of search by keyword. However, for

the system to accurately reflect the needs of users, the organization of the web documents should also take into account the feedback from users. While it is useful to have a system to organize the web pages in a content-driven manner, it may be more advantageous to organize the web pages in a web-user-oriented manner. After all, the web documents are organized so that humans can search in a more effective and efficient manner. The system should realise that users who visit data mining web pages may also like to visit web pages about self-organization, for instance. The usage patterns of web users can therefore play a role in assisting other users.

Smith and Ng [7] used a LOGSOM system to cluster Web pages using Web logs. The clustering is based on user's browsing history instead of contents of Web pages. The system provided a visual tool to demonstrate the relationship between webpages.

In this paper we focus on a self-Organizing Map (SOM) approach for web usage mining which is one of the types of web mining techniques. Different from other content-based web page clustering approaches [4, 8, 9], the SOM-based approach clustering for web pages based on the user's browsing patterns. The goal of this article is to study the feasibility of the SOM-based approach on web page clustering. The paper is organised as follows: in section 2 we discusses on web mining using Self-Organising Map, section 3 discusses SOM algorithms, section 4 discusses SOM based web page clustering in section 5 we discusses on experimentation and finally in section 6 we conclude the papers.

II. WEB MINING USING SELF- ORGANIZING MAP

Applying SOM on natural language data means doing data mining on text data, for instance Web documents [10]. The role of SOM is to cluster numerical vectors given at input and to produce a topologically ordered result. The main problem of SOM as applied to natural language is the need to handle essentially symbolic input such as words. If we want SOM to have words as input then SOM will arrange the words into word categories. But what about the input (training) vector associated to each input word? What should be the vector components, i.e. the attributes of a word? Similarity in word appearance is not related to the word meaning, e.g. "window", "glass", widow". We have chosen to classify words by SOM, Creating thus word category maps. The attributes of the words in our experiments were the count of the word occurrences in each document in a collection of documents.

Consequently, we have chosen to represent the meaning of each word as related to the meanings of text passages (documents) containing the word and, symmetrically, the semantic content of a document as a bag-of-words style function of the meanings of the words in the document. The lexical-semantic explanation of this contextual usage meaning of words is that the set of all the word contexts in which a given word does and does not occur provides a set of mutual constraints that captures the similarity of meaning of words and passages (i.e. documents, contexts) to each other. The measures of word-word, word-passage and passage-passage relations are well correlated with several cognitive phenomena involving semantic similarity and association [11].

The meaning of semantically similar words is expressed by similar vectors. After training a SOM on all the words in a collection of documents, where the vectorial coding of words represents the contextual usage , the result self-organizing map groups the words in semantic categories. There are also other possibilities to code words, which lead to grammatical or semantic word categories [12,13,4].

III. SELF-ORGANIZING MAP

Teuvo Kohonen [14] introduced the SOM network that reduced the dimensions of data through the use of self-organizing 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 dimension The SOM network integrates dimensions reducing and clustering in one network. Figure 1 shows the map-ping from a one-dimensional input to a two-dimensional array.

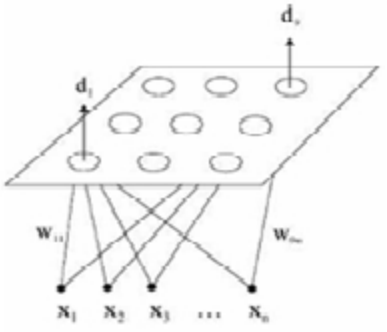

Fig.1: The Mapping from a one-dimensional input to a two-dimensional array [7].

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 neighbours
   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 algorithm selects a sample vector randomly and searches 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.

IV. SOM-BASED WEB PAGE CLUSTERING

A. Overall Architecture

Generally, our approach can be divided into three steps: data preprocessing, Web page mapping, and clustering analysis. Figure 2 shows these three steps[15].

In the data preprocessing step, a couple of methods are used to identify users, sessions, and transactions. The Web site topology is also identified in this step. In general, in this step, the raw Web data should be preprocessed into data abstractions for further processing.

After the data preprocessing step, SOM is used to cluster pages from similar navigating patterns. Unlike other Web personalization systems that usually find pages belonging to the same cluster based on the contents of the pages, our approach uses the user's current navigation pattern. Moreover, our SOM network uses the k-means clustering algorithm where more than one cluster will be considered at the same time for further analysis.

In the clustering analysis step, results from the Web page mapping step are stored in two-dimensional arrays. The Web site topology we identified in the preprocessing step will be used to filter patterns containing pages of a certain usage type. Clustering analysis can help the developer to get user's Web browsing patterns and predict the users move when they brows some particular sites.

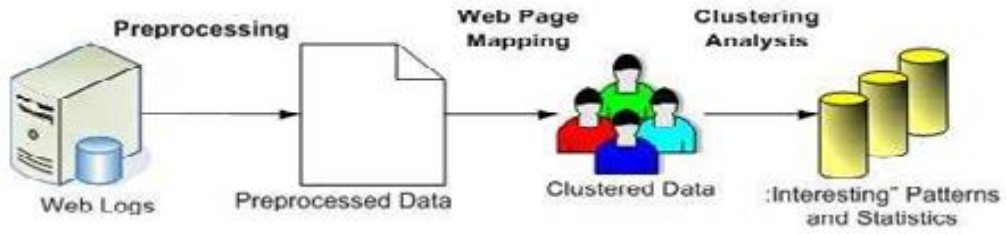

Fig.2 The Mapping from a one dimension input to the two-dimensional array [7]

B. Data Preprocessing

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 the Web 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 along 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 be-tween page requests exceed 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.

C. 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.

V. EXPERIMENTATION

We used Web log file from the http://nielsen-netratings.com as our test data. The data size is about 50 MB 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 pre-processing is 190. We used a fixed value of 20 as the number of clusters, so the input to the SOM network is a 190 by 20 array. We have tested different parameters for the SOM network as follows: α varies from 0.2 to 0.9 and ω varies from 1 to 40 where α represents the learning rate and ω determines the number of times a URL being presented within one learning cycle before the neighbourhood size is decreased. In our algorithm, there are 18 learning cycles for organizing the Web pages. In particular, we decreased the neighbourhood size from its initial value of 17 to 0. Fig.2 and Fig.3 shows the SOM map with ($\alpha = 0.1$, $\omega = 40$) and ($\alpha = 0.5$, $\omega = 40$), respectively.

From our experimental results, we find that, with $\omega = 40$, the two-dimensional array maps display clearest contesting. Table 4 shows part of the clusters with $\alpha = 0.5$ and $\omega = 40$. 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 the Web 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.

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	127,151,155,156,231,232,279,280,281 291,307,308,323,348,360,381,382,383,389, 390,441,814,1273,1274,1275,1276,1277, 1278,1286,1287,1289

Table 4: part of clusters with α=0.5 and ω=40

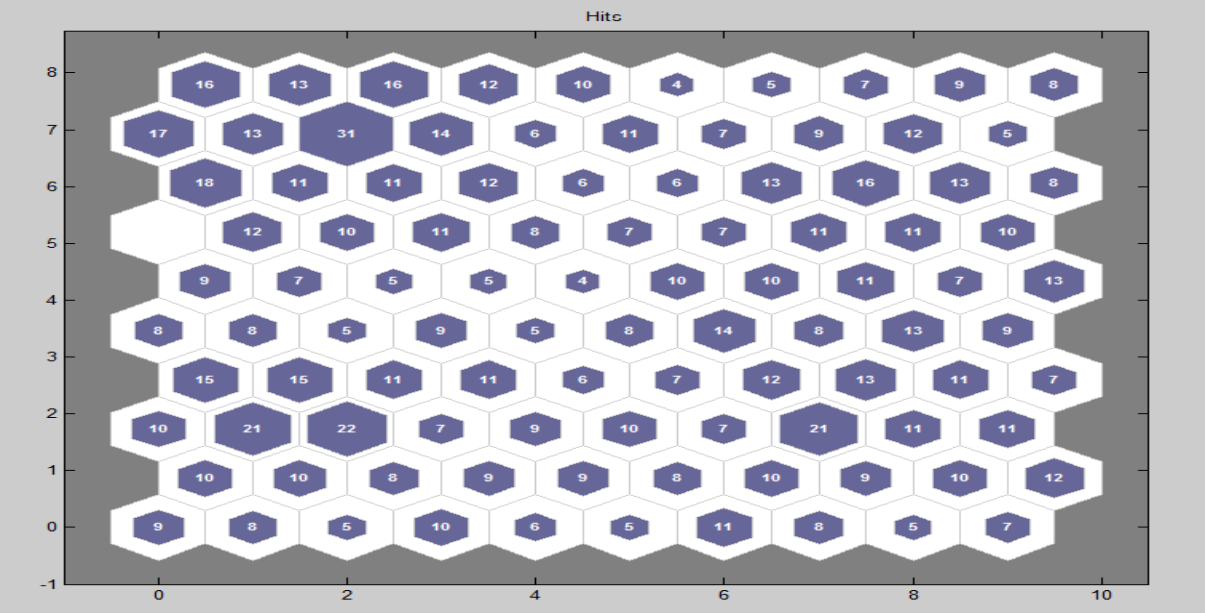

Fig. 2 SOM Map with α=0.5 and ω=40

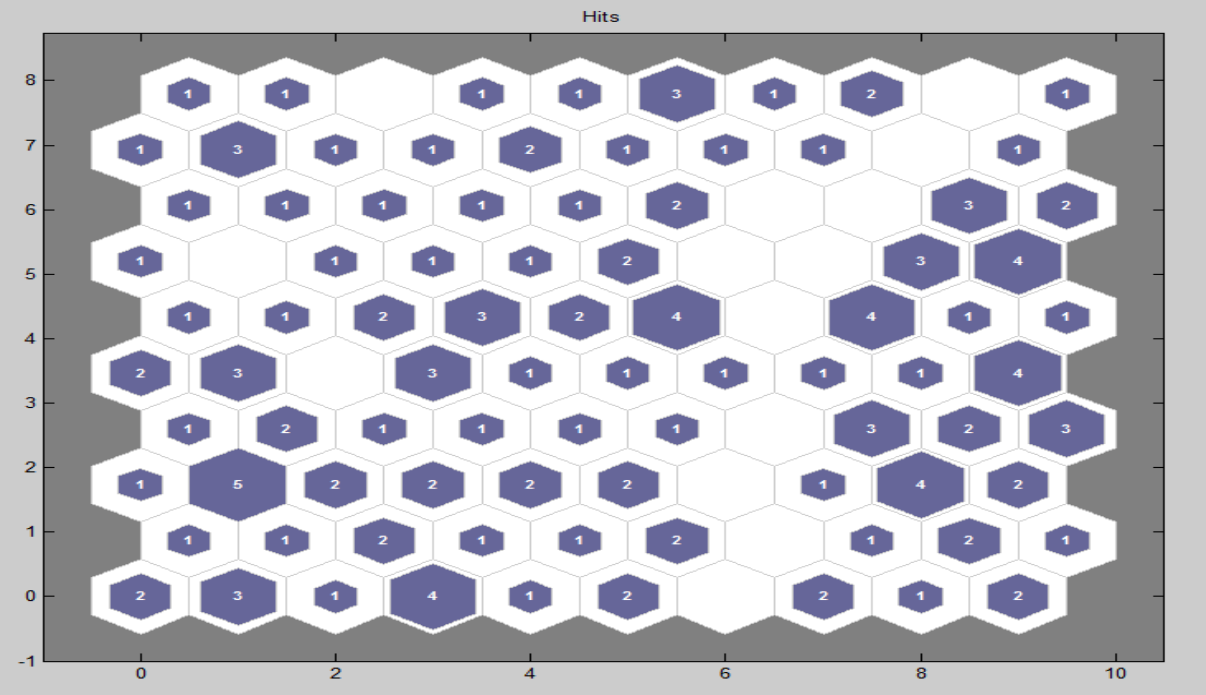

Fig. 3 SOM Map with α=0.1 and ω=40

VI. CONCLUSION

We introduced a Self-Organizing Map (SOM) approach to the study of web mining implementing on Web log data. Starting from the raw Web log data that is available in any Webserver, 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 that data 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. The aim of this study has been to demonstrate the feasibility of the approach within a controlled domain, and lay the foundations for future research in this area. We will also investigate the benefits of combining content information with usage patterns and consider the use of temporal information to enable the SOM to adapt over time to changing user navigation patterns.

REFERENCES

[1] J. Nielsen, *The art of Navigating in Hypertext*, Communications of the ACM 33 (3) (1990) 296 - 310.
[2] K. Lagus, T. Honkela, S. Kaski, T. Kohonen, *Self-organizing maps of document collections: a new approach to interactive exploration*, Second International Conference on Knowledge Discovery and Data Mining, AAAI Press, Menlo Park, CA, 1996, pp. 238 - 243.
[3] S. Kaski, T. Honkela, K. Lagus, T. Kohonen, *WEBSOM — self-organizing maps of document collections*, Neuro Computing 21 (1 - 3) (Oct. 1998) 101 - 117.
[4] T. Kohonen, S. Kaski, K. Lagus, J. Salojarvi, J. Honkela, V. Paatero, A. Saarela, *Self organization of a massive document collection*, IEEE Transactions on Neural Networks 11 (3) (May 2000) 574 - 585.
[5] T. Kohonen, *Construction of similarity diagrams for phonemes by a self-organizing algorithm*, Technical Report TKK-F-A463, Helsinki University of Technology, Espoo, Finland (1981).
[6] T. Kohonen, *Self-organized formation of topologically correct feature maps, Biological Cybernetics* 43 (1982) 59 - 69.
[7] 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.
[8] 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.
[9] 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.
[10] K. Lagus, ,*Text retrieval using self-organized document maps*, Technical Report A61, Helsinki university of Technology, Laboratory of Computer and Information Science (2000).
[11] T.K. Landauer, P.W. Foltz,. And D. Laham,. *Introduction to Latent Semantic Analysis*, Discourse Processes, vol. 25, 1998, pp. 259-284.
[12] T. Honkela. *Self-organizing maps in natural language processing*, Ph.D. thesis, Neural Networks Research Center Helsinki University of Technology, Finland (1997).
[13] T.Kohonen, ,J. Hynninen,, J. Kangas, and J. Laaksonen, *SOM_PAK: The self-organizing map program package*, Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, 1996.
[14] T. Kohonen. *Self-Organization and Associative Memory.* Springer-Verlag, New York, 1988.
[15] Dehu Qi, Chung-Chih Li *Self-Organizing Map based Web Pages Clustering using Web Logs*, School of Information technology, Illinois State University, USA 2005.

