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. 5, Special Issue 4, February 2018

Enhanced Self Organizing Map Algorithm for Web Usage mining through Neural Network

Prof.A.S.Sridhar
H.O.D.
Department of Computer Science and Applications
Jaya-Arakkonam Arts and Science College
Arakkonam, Tamilnadu, India
assridhar200704@gmail.com

***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.**

I. Introduction

The proposal shares dissimilar goals with many those agents, our approach is automatic that it does not require users explicit input. Moreover, it takes a systematic approach to collect and comprehend user activities. It provides 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. It proposes 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 . 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.

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.

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 predictive Web caching. Thus, Web mining can be categorized into three different classes based on which part of the Web is to be mined and these three categories are Web content mining, Web structure mining and Web usage mining. Web content mining 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 or to improve the Web structure and Web server performance For this reason a model of the users (User Model - UM) have to be built based on the information gained from the log data.

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. 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 have to be split into sessions. In general the timeout for determining a session is set to 30 minute. 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 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 tries to understand the patterns detected in before step. The most common techniques is data visualization applying filters, zooms, etc

2 Literature Review

Liao Hui et al (2015) have discussed above the clustering algorithm analysis of web users with Dissimilarity and SOM neural Network. International journal of advanced research in computer science and software engineering, web usage mining using self organized Map.

Jin and Zhou (2015) have discussed above the categorizing the web mining based on different classes, these three categories are web content mining, web structure mining and web usage mining. 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 or to improve the Web structure and Web server performance. For this reason a model of the users (User Model - UM) have to be built based on the information gained from the log data.

Kohone (2014) have discussed above the 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.

Mac-Queen algorithm represents each of *k* clusters *Cj* by the mean (weighted average) *cj* of its point (introduced K-mean alogrithm this clustering 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.

Teuvo Kohonen(2010) 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 dimensions. The SOM network integrates dimensions reducing and clustering in one network.

In the survey, different SOM algorithms are used for mining the Web data .the clustering the different log files are explained by different author's .the literature survey gives the following outcomes.

- 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 dimensions.

- The algorithm analysis of web users with Dissimilarity and SOM neural Network. clustering

- The categorizing the web mining based on different classes by three categories are web content mining, web structure and web usage.

- The 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.

- Mac-Queen algorithm represents each of *k* clusters *Cj* by the mean (weighted average) *cj* of its point (introduced K-mean alogrithm This clustering 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.

3 Objectives of Research

The objective of the research work is ESOM (Enhance self organizing map) method uses the web data by dissimilarity clustering which optimize SOM neural network, learning and improve clustering effect. The object of the proposed research work is:

- This proposed work provides the output nodes affecting the clustering effect about data set.
- Proposed work link the output node weight, select the initialization values which reduces the difficulties in clustering.
- Proposed work is combination of not only on log files also user registration information such as age, gender, income, region etc.,
- This proposed work ESOM can estimate the centre and the number of clustering data set by"dissimilarity computing".
- Proposed work optimizes SOM neural network learning and improves clustering.

4 Description of Research Work

4.1 *Enhance Self Organized Map*

The Self-Organizing Map (SOM) 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 A 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 high dimensional data.

Figure 1 shows the mapping from a one-dimensional input to a two-dimensional array.

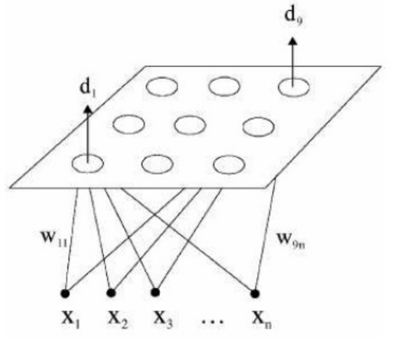


Figure 1: The Mapping from a one-dimensional input to a two-dimensional array .The SOM network organizes itself by competing representation of the samples. Neurons are also allowed to change themselves in hoping to win the next competition. This selection and learning process makes the weights to organize themselves into a map representing dissimilarities. 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 ESOM 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 it 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.

4.2 *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. [7] proposed a system which is an innovative congestion control algorithm named FAQ-MAST TCP (Fast Active Queue Management Stability Transmission Control Protocol) is aimed for high-speed long-latency networks. Four major difficulties in FAQ-MAST TCP are highlighted at both packet and flow levels. The architecture and characterization of equilibrium and stability properties of FAQ-MAST TCP are discussed. Experimental results are presented comparing the first Linux prototype with TCP Reno, HSTCP, and STCP in terms of throughput, fairness, stability, and responsiveness. FAQ-MAST TCP aims to rapidly stabilize high-speed long-latency networks into steady, efficient and fair operating points, in dynamic sharing environments, and the preliminary results are produced as output of our project. The Proposed architecture is explained with the help of an existing real-time example as to explain why FAQ-MAST TCP download is chosen rather than FTP download.

4.3 *ESOM-based Web page clustering*

This approach can be divided into three steps: data reprocessing, Web page mapping, and clustering analysis. In the data pre-processing 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 pre-processed into data abstractions for further processing. After the data pre-processing step, ESOM 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, approach uses the user's current navigation pattern. Moreover, ESOM 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 it identified in the pre-processing 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.

4.4 *Functioning of ESOM in 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. It describes 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, it only wants 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 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 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

4.5 *Design of 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.

5 Organization of the Thesis

The thesis contains six chapters and organized as follows

Chapter 1 discuss the introduction about the Self organizing mapping algorithm, Web mining.

Chapter 2 explains the work done by numerous researchers about different methods of clustering and log file mining.

Chapter 3 presents Enhance self organizing mapping method architecture.

Chapter 4 explains the implementation of ESOM in web mining.

Chapter 5 analyses the experimental results.

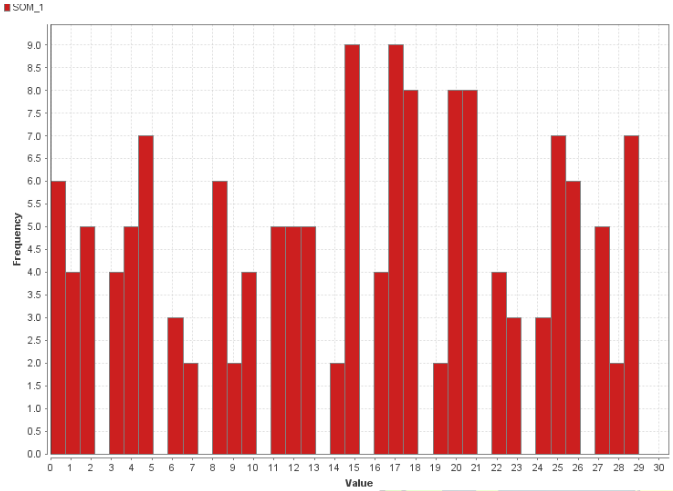
Chapter 6 deals the conclusion and scope for future work.

6 Results and discussion

The following are the finding with respect to the algorithm implemented. By using the R tool simulated the data find that the approach indeed results a very meaningful ESOM network in the sense that the Web pages are organized into clusters based on the similarity of their usage. Within a cluster, it can see that users are indeed likely to navigate Web pages within the same node, even though the ESOM was given no information about the directory structure of the server and the contents of the Web pages. The ESOM 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.

In ESOP approach dimensions reduced by reducing the no of attributes

Comparison between SOP and ESOP



6 Conclusion

An Enhanced Self-Organizing Map (ESOM) used for mining Web log data. Starting from the raw Web log data that is available in any Web server, we pre-processed 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 irs data demonstrate that approach is very useful in a specified domain.

The results of the clusters generated from the ESOM network shows that approach can effectively discover usage patterns. Results can also be used to predict the user's browsing. It can be effectively reduces the dimension of data.

References

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

[2] Huang. X, F. Peng, A. An, and 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] 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.

[4] Jin. X, Zhou, and 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.

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

[6] Kohonen. T, 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.

[7] Christo Ananth, S.Esakki Rajavel, I.AnnaDurai, A.Mydeen@SyedAli, C.Sudalai@UtchiMahali, M.Ruban Kingston, "FAQ-MAST TCP for Secure Download", International Journal of Communication and Computer Technologies (IJCCTS), Volume 02 – No.13 Issue: 01 , Mar 2014, pp 78-85.

[8] Liu. J, 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.

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

[10] Nakagawa. M and 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.

[11] Nasraoui. O and 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.

[12] 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.

[13] Ypma. A and 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.

[14] Zhang Y.,X. Yu, and J. Hou, "Web communities: Analysis and construction,"

Berlin Heidelberg, 2006.