Web usage mining using Soft Computing Technique

Shilpa Shrivastava
M.E (CS)
(goodshilpa8@gmail.com)

Shweta Modi
Asst .Prof (CS)
(modi_shweta84@yahoo.com)

Shri Ram Institute of Technology, Jabalpur

Abstract: Web usage mining attempts to discover useful knowledge from the secondary data obtained from the interactions of the users with the Web. Web usage mining has become very critical for effective Web site management, creating adaptive Web sites, business and support services, personalization, and network traffic flow analysis and so on. The study of ant colonies behavior and their self-organizing capabilities is of interest to knowledge retrieval/ management and decision support systems sciences, because it provides models of distributed adaptive organization, which are useful to solve difficult optimization, classification, and distributed control problems, among others. Previous study on Web usage mining using a concurrent Clustering, Neural based approach approach has shown that the usage trend analysis very much depends on the performance of the clustering of the number of requests. In this paper, a novel approach Neural Gas is introduced kind of neural network, in the process of Web Usage Mining to detect user's patterns. The process details the transformations necessaries to modify the data storage in the Web Servers Log files to an input of NG**.**

Key words: Web Usage Mining, Clustering, NG, Web Server Log File

I. 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.

II. 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) preprocessing,
(ii) pattern discovery
(iii) Pattern analysis [15]

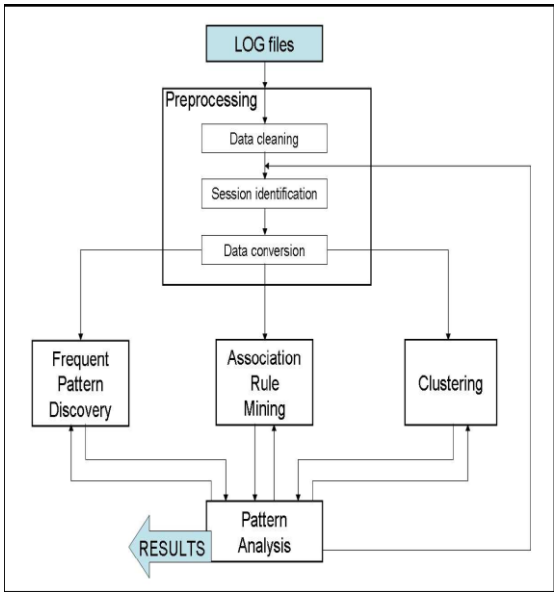

Figure 1 shows the block diagram

In the preprocessing 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 preprocessed 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 behavior. 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 behavior 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 [Keim, 2002; Ankerst, 2001]

The artificial neural networks (ANN), try to simulate the action doing by the human brain; RNA has the possibility of get abstraction of data and work with incomplete data or with errors, RNA has knowledge and can adapt it; and operate in real time [Grosser, 2004; Daza, 2003]. RNA is built by a common part called neurons. These units of processing are interconnected; each neuron has it this activation threshold. The learning in RNA is built by the adjustment of activation threshold in each neuron [Roy, 2000; Abidi, 1996; García Martínez *et al*., 2003]

III. RELATED WORK ON WEB USAGE MINING

The analysis of web user behaviors is known as web Usage Mining (WUM) that is to say, the application of data mining techniques to the problem of learning web usage patterns. The WUM is a relatively new research field that mainly focus on the study of the stream of users requests (or sessions) but that deals more generally with any user interactions with web sites (inserting or editing text in a web page (Kay, 2006), printing document).Many works have been conducted in the last ten years to extract, analyze, model or predict the web users information needs on a web site. Cooley et al. (Cooley 1997) proposed some heuristics to prepare the web log file, to filter and to reconstruct the web sessions. Then, numerous approaches have been proposed to help understanding the web users behaviors and inferring their motivations from their requests on web servers. Cooley et al. (Cooley 1999) introduced the web Miner system that allows to filter web log files, to reconstruct web sessions as transaction vectors, to compute association rules from the sessions and to request the set of rules with an SQL-like language so that an expert of the web site can extract easily meaningful information. Similarly, Spiliopoulou et al. (Spiliopoulou 1998) described the web WUM system, which also introduces an SQL-like language to request new behavioral rules or mine
from an aggregated tree representation of the web sessions. This work has been extended in (Spiliopoulou 1999) where the authors distinguish profiles from episodic or first time visitors, usual visitors and clients to apply modifications on the web site to increase the number of visitors that are clients. Masseglia et al. (Masseglia 1999a, 1999b) proposed webTool, an expert system that relies on sequential patterns extraction and uses the incremental method ISEWUM. The system aims at reorganizing web sites or at predicting web pages like Davison (Davison 2002, 1999). Perkowitz and Etzioni (Perkowitz 1999) reorganize web sites via the generation of a thematic index page using a conceptual clustering algorithm named PageGather. Some other works (Labroche 2003, Baraglia 2002, Heer 2001, Fu 1999, Yan 1996) apply clustering algorithms to discover homogeneous groups of (web) sessions. The underlying idea of these methods is that the algorithm should group in the same cluster the sessions that correspond to the users that navigate similarly, that is to say, that have the same motivation and interests. The objective can be either to discover users that accessed the same resources on the web site or to find clusters of web pages that co-occur frequently in the

same sessions. This kind of approach allows defining a profile of typical web site access from each discovered cluster and can highlight information that is not directly accessible in the data. The extracted profiles can be used for web pages recommendation purposes or dynamic web site organization. Yan et al. (Yan 1996) use the First Leader clustering algorithm to create groups of sessions. The sessions are described as hits vectors in which each component corresponds to a web page and indicates the number of times it has been accessed during the session. The weakness of the method stands in the use of the First Leader algorithm. Although very fast, it is dependant on the order in which web sessions are processed and may need to parameter its maximal number of clusters. Estivill-Castro et al. (Estivill-Castro 2001) uses a k-Means-like algorithm that relies on a median rather than a mean to estimate the cluster centers. In (Nasraoui 2002), the authors use a fuzzy C Medoids approach, derived from the fuzzy C Medoids algorithm FCMdd described in (Krishnapuram 2001), to deal with the uncertainty and inaccuracy of the web sessions. The new algorithm is a linear version of FCMdd algorithm that takes into account only the p objects that most belong to a cluster to compute its new medoid. Nevertheless, a problem still remains: the analyst has to specify the number of expected clusters (even if it is possible to overestimate this number and only keep the most relevant clusters at the end). To automatically evaluate this number of expected clusters Heer and Chi (Heer 2002) propose a simple heuristic that computes the stability of the partitions for different number of clusters. The method works well but is extremely time consuming and thus may not be applicable in a real study context. In (Nasraoui 1999), the authors propose the CARD relational clustering algorithm (Competive Agglomeration for Relational Data) that is able to determine the appropriate number of clusters starting from a large number of small clusters. In (Suryavanshi 2005), the authors propose an incremental variant of a subtractive algorithm that allows updating partitions from previous analysis. This method determines automatically the number of clusters according to an estimation of the potential of each object to be a cluster center based on the density of the other objects in its neighborhood. Labroche et al. (Labroche 2003) describe a relational clustering algorithm inspired by the chemical recognition system of ants named AntClust. In this model, each artificial ant possesses an odor representative of its nest membership called “label” and a genome, which is associated with a unique object of the data set. The algorithm simulates meetings between artificial ants according to behavioral rules to allow each ant to find the label (or nest) that best fits its genome. AntClust does not need to be initialized with the expected number of clusters and runs in linear time with the number of objects. AntClust has also been successfully applied to the web sessions clustering problem with various representations of sessions. More recently, Labroche (Labroche 2006) introduces the Leader Ant algorithm that relies on the same biological model and that is between 5 to 8 times faster than AntClust with similar clustering error values on benchmark data sets. Recently, Mobasher (Mobasher 2006) proposed an overview of the main data mining approaches that have been used to mine user profiles in a web personalization perspective (association rules, sequential patterns and clustering methods).

El-Shishiny, H.Sobhy Deraz,S.Badreddin (2008) use the artificial Neural Network to predict software aging phenomenon and analyze resource data collected on a typical running software system using Multilyaer Perceptron algoirhtm.

Santhi, S.Shrivasan.P (2009) use Back progation algorithm (BPA) has been applied to learn the navigated web pages by different users at different session. the performance of the BPA in prediting the next possible web pages is about 90%

IV. INTRODUCTION TO NEURAL NETWORK

An Artificial Neural Network (ANN) is an information-processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the

information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems. An ANN is configured for a specific application, such as pattern recognition or data classification, through a learning process

Neural networks, with their remarkable ability to derive meaning from complicated or imprecise data, can be used to extract patterns and detect trends that are too complex to be noticed by either humans or other computer techniques. Neural Network has other advantage

1. Adaptive learning: An ability to learn how to do tasks based on the data given for training or initial experience.
2. Self-Organization: An ANN can create its own organization or representation of the information it receives during learning time.
3. Real Time Operation: ANN computations may be carried out in parallel, and special hardware devices are being designed and
4. Manufactured which take advantage of this capability.
5. Fault Tolerance via Redundant Information
   Coding: Partial destruction of a network leads to the corresponding degradation of

V. PROBLEM DESCRIPTION IN WEB USAGE MINING

A. Logs processing

A critical step in the identification of user's habit in web sites is the cleaning and transformation of Web server Log files; and user's session's identification

B. Log files cleaning

Cleaning Web server Log files has a lot of steps when one user request a page, this request is added to the Log File, but, if this page has images, in they will be added in the Log file. This is the same for any resource in the page, for example JavaScript's, flash animations, videos, etc. In most of the cases these resources aren't necessary for the detection of user's habits; for this reason is good cat this records from the log file; to do this task we only need to search records by file extension. To give a little list we can consider cut extensions with jpg, jgeg, gif, js, css, swf, avi, mov, etc.

In some case is proper to filter page inserted in others with frames; in other way is common to generate pages dynamically. Errors code in HTTP is used too filter records in the Logs files, the most common errors in HTTP are: error code 200, 4003 (recourse not found), 403 (access denage), and 500 (internal server error).

C. Users identifications

After the log files cleaning, we need to identify user's sessions. We have some methods to detect sessions each one pros and cons. One method is detecting the use of cookies [Eirinaki & Vazirgiannis, 2003; Huysmans et al., 2003; Kerkhofs, 2001]. W3C [WCA] define cookies as "data sent by the server to the client, data locally storage in cookies and is send to the server with each request". In other words the cookies are HTTP headers in string format. Cookies are used to identify users behind server's access, and what resources the user accesses. One problem with this method is; the users can lock the use of cookies, and the server after that can't storage information locally in the user machine; other problem is; the user can delete the cookies. Another method to identify users is using Identd [Eirinaki & Vazirgiannis, 2003]. Identd is a protocol defined in RFC 1413 [RFC 1413], this protocol permits detect to a user connected by the unique TCP connection.

The problem with Identd is the terminal user needs to configure with the Identd support. Other method is detecting the users in log files by the IP direction registered in each record. Another method is the explicit user's registration each the user time accesses to the site. At last we can detect users with the users name added in the log file in filed name.

D. User session's identification

After identify the users, we need to identify the sessions. To do this we can divide the access of the same users in sessions. It's difficult to detect when one session is finish and start another. To detect sessions is common use of time between requests; if two

requests are called in of time frame, we can suppose that these requests are in the same session; in other way below of time frame we can consider two different sessions. A good time frame is 25.5 minutes.

F. User's habit identification

After that all log processing, we can start to detect the user's habit.

VI. PROPOSED APPROACH:-

In the present work, this paper proposes the use of Neural Gas Algorithm to identify the Web Log file. This kind of artificial neural network will be tried to gather the users by patterns of pages accesses. To obtain this result we need to process the Web Log files to identify users and session of users; after that with this session's user, we'll train the ANN. The selection of Neural Gas algorithm is so because it isn't necessary to supervise to the training.

A. Neural Gas Algorithm

Neural Gas - a biologically inspired adaptive algorithm, coined by Martinetz and Schulten, 1991. It sorts for the input signal according to how far away they are. A certain number of them are selected by distance in order, and then the number of adaptation units and strength are decreased according to a fixed schedule.

Algorithm

The rough steps of the Neural Gas algorithm can be specified as

Assuming that we have a distribution p(ζ)for which a Neural Gas model has to be created. The following parameters are needed for the Algorithm initialization.

λ_i,λ_f and E_i,E_f, and t_{max}

λ_i,λ_f are used to set the rate at which learning rate E converges E_i,E_f are the initial and final learning rate E respectively. t_{max} is the time till which the process continues.

Step 1. Create a Set A to contain N units each with a vector reference from p(ζ). Also initialize the time parameter to 0.

$A=\{C_1,C_2,...C_N\}$, $t=0$

Step 2. Get a random value from the distribution p(ζ) and call it X.

Step 3. Line up all the elements from A in relation with their nearness to X, with the nearest coming first and the farthest the last.

Thus line up A's vectors such that for Cp,Cm,Co... the corresponding vectors $W_p,W_m,W_o,...$

$||W_p-X|| <=||W_m-X||<=||W_o-X||$ holds true

The norm || || usually taken is the square norm.

Step 4. Change the vectors for Cp,Cm,Co...

$\Delta Wi = E(t)*h_\lambda(k_i(X,A))*(X-Wi)$

Where,

$\lambda(t) = \lambda i(\lambda_f/\lambda_i)^{(t/tmax)}$

$E(t) = Ei(E_f/E_i)^{t/tmax}$

$h_\lambda = e^{(-k/\lambda(t))}$

Step 5. Increment t

$t=t+1$

Step 6. $t<t_{max}$ return to step 2

Features of Neural Gas

1. The neural gas model does not delete a node and also does not create new nodes.
2. The neural gas model will require fine tuning of the λ parameters especially to achieve a good convergence rate and stable model.
3. It is topology representing neural network, that is after reaching convergence ($>t_{max}$), the network node's vector would be representing the distribution being modeled
4. Time efficiency is better in Neural Gas
5. By adapting not only the closest feature vector but all of them with a step size decreasing with increasing distance order, compared to k-means clustering a much more robust convergence of the algorithm can be achieved.

VII. CONCLUSION

In this work, we study the possible use of the neural networks learning capabilities to

classify the web traffic data mining set. the discovery of useful knowledge, user information and server access patterns allows Web based organizations to mining user access patterns and helps in future developments, maintenance planning and also to target more rigorous advertising campaigns aimed at groups of users. We can conclude that; to identify common patterns in Web, the Neural Gas Algorithm is better than KMeans. NG has a better group of users. With K-Means, NG we get a few information about user's habits. In other way NG build some gathering with a great quantity of user's sessions for the same users.

REFRENCES: -

[1] Abraham, A., Ramos, V. (2003). Web Usage Mining Using Artificial Ant Colony Clustering and Linear Genetic Programming.

[2] R. Kosala, H. Blockeel, Web Mining Research: A Survey, SIGKKD Explorations, vol. 2(1), July 2000.

[3] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, SIGKKD Explorations, vol.1, Jan 2000.

[4] Cernuzzi, L., Molas, M.L. (2004). Integrando diferentes técnicas de Data Mining en procesos de Web Usage Mining. Universidad Católica "Nuestra Señora de la Asunción". Asunción. Paraguay.

[5] Chau, M.; Chen, H., "Incorporating Web Analysis Into Neural Networks: An Example in Hopfield Net Searching", IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews, Volume 37, Issue 3, May 2007 Page(s):352 – 358

[6] Raju, G.T.; Satyanarayana, P. S. "Knowledge Discovery from Web Usage Data: Extraction of Sequential Patterns through ART1 Neural Network Based Clustering Algorithm", International Conference on Computational Intelligence and Multimedia Applications, 2007, Volume 2, Issue , 13-15 Dec. 2007 Pages :88 -92

[7] Jalali, Mehrdad Mustapha, Norwati Mamat, Ali Sulaiman, Md. Nasir B. , " A new classification model for online predicting users' future movements", in International Symposium on Information Technology, 2008. ITSim 2008 26-28 Aug. 2008, Volume: 4, On page(s): 1-7, Kuala Lumpur, Malaysia

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