International Journal of Engineering Science Invention
ISSN (Online): 2319 – 6734, ISSN (Print): 2319 – 6726
www.ijesi.org || PP—12-16

A New Approach For Web usage Mining using Artificial Neural network

Smaranika Mohapatra[1], Kusumlata Jain[2], Neelamani Samal[3]
[1](*Department of Computer Science & Engineering, MAIET, Mansarovar, Jaipur, India)*
[2]*(Department of Computer Science & Engineering, MAIET, Mansarovar, Jaipur, India)*
[3]*(Department of Computer Science & Engineering, GIET, Bhubaneswar, Odisha ,India)*

Abstract: *Web mining includes usage of data mining methods to discover patterns of use from the web data. Studies on Web usage mining using a Concurrent Clustering has shown that the usage trend analysis very much depends on the performance of the clustering of the number of requests. Despite the simplicity and fast convergence, k-means has some remarkable drawbacks. K-means does not only concern with this but with all other algorithms, using clusters with centers. K-means define the center of cluster formed by the algorithms to make clusters. In k-means, the free parameter is k and the results depend on the value of k. there is no general theoretical solution for finding an optimal value of k for any given data set. In this paper, Multilayer Perceptron Algorithm is introduced; it uses some extension 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 Multilayer Perceptron*

Keywords: *Web Usage Mining, Clustering, Web Server Log File. Neural networks, Perceptron, MLP*

I. Introduction

Web mining has been advanced into an autonomous research field. Web mining includes a wide range of applications that points at discovering and extracting hidden information in data stored on the Web. Web Mining can be distributed into three different categories depending on which part of the Web is to be mined. These three categories are Web content mining, Web structure mining and Web usage mining [4]. Web content mining aims to extract/mine useful information or knowledge from web page contents. It is the job of originating useful information available on-line. There are various types of Web content which can give 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 goal of Web content mining is to provide an effective measure to help the users to find the information they inquire. Web content mining constitutes the task of standardizing and grouping the documents and providing the search engines for collecting the various documents by keywords, categories, contents. Web structure mining is the way 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 [7].

1.1. Web Usage Mining

The objective of Web usage mining is to find new patterns in the activities of users. That improves the services needs of the users by some techniques like dynamic link handling, by page recommendation. The aim of a portal is to supply useful information to the user .Every user means money as per business prospects for commercial and business portals. Thus the goal of each owner of a website is to make his site more attractive for the user. In result the response time of each single site have to be kept below 2s. Moreover, some extras have to be contributed such as supplying dynamic content or links or recommended pages for the users that are possible of interest for a given user. Grouping/Clustering of the user activities collected in different types of log files is an important issue in the Web community [4]. There are three types of log files that can be helpful for Web usage mining. The log files are collected on the server side, client side and proxy servers. The way of storing these log files at different places is difficult for Web mining. But it’s also true , that really reliable results will be available only if one has data from all three types of log files [7].Choosing the three types of log files also has a reason as the server side does not contain the records/logs of those Web page accesses which are cashed on the proxy or client servers

Major algorithms and methodologies are based work for server side. Web usage mining consists of three steps [2]:

1. **Data Collection and Pre-Processing** -
Data is collected from server side and client side. These basically include name and IP of the host, date and time of query request. Using JavaScripts or Applets the data can be tracked.
Data Pre-processing is a method of transforming or converting the raw data into some certain format that can be easily be processed for the purpose. The importance of this stage includes user identification and session identification.

2. Pattern discovery
In this technique the IP addresses that are identified are converted to domain names, the URLs are converted to page titles. This step searches for patterns in the data using techniques such as association rule.

3. Pattern analysis
This step is the last step of Web usage Mining. The aim of this process is to remove the patterns and to extract the patterns from the output of pattern discovery process. It uses the tools like SQl, OLAP, etc

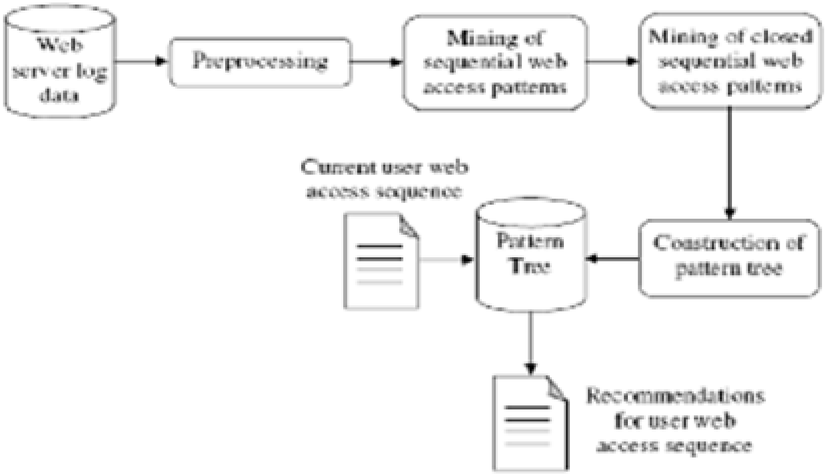
Web server log data
Preprocessing
Mining of sequential web access patterns
Mining of closed sequential web access patterns
Current user web access sequence
Pattern Tree
Construction of pattern tree
Recommendations for user web access sequence

Fig 1 Block Diagram of Pattern analysis

II. Related Work

Many data mining techniques are available:

In [1], authors have proposed ,"Alternative Approach to Tree-Structured Web Log Representation and Mining", It first changes the tree-structured data into a flat representation that conserves the structured and attribute-value information. A "Dynamic Modeling by Usage Data for Personalization Systems "[2],is proposed ,in this approach it is generating an offline model using usage data as an information source will dramatically improve the recommendation performance and the online response efficiency in personalization system.

"A Survey of Soft Computing in Intelligent Information Retrieval Systems"[3]. The author provides an in-depth survey of challenges in the design of intelligent information retrieval systems, pointing out some similarities and differences in the core data mining and web based search operations

In [4], the authors have proposed "Web Mining: Information and Pattern Discovery on the World Wide Web". In this work, author defines what is Web mining and presents an overview of the various research issues, techniques, and development efforts.

"An Applicable Method for Evaluation of Web-learning Platforms Content Quality and Usage"[5] as proposed states that the quantity and quality of e-Learning educational content is shown by its usage. In this, old and new performance metrics and a simple performance prediction algorithm, for the measurement of both educational content as exposed by the educators and the usage of such educational material by the learners, are presented.

In [6], authors have proposed that "Web Usage Mining Statistical Classification and Fuzzy Artificial Neural Networks". Ways to web usage based classification were surveyed, and the artificial neural networks, in particular, Multilayered Perceptrons , were found to be effective and relevant in the problem of classification.

In [7], authors have presented "Web Usage Mining: A Survey on Pattern Extraction from Web Logs". As the web increases along with number of users, it is required for the website owners to better understand their customers so that they can provide much better service, and also enhance the quality of the website.

III. Neural Networks

An Artificial Neural Network (ANN) is an idea based on the belief that is inspired by the biological neurons and the nervous system of a human. It's an information-processing methodology. ANNs are made up of multiple nodes which imitate biological neurons of a human brain It comprises of a large number of interconnected processing elements known as neurons together which can solve many problems in the computing world. ANNs are made up f multiple nodes which imitate biological neurons of a human brain. The various applications of ANNs are pattern recognition or data classification. The advantages of a Neural Network are:

1. Relatively easy to use.
2. Adaptive learning: An ability to learn how the jobs are to be done based on the given data for the training.
3. Self-Organization: An ANN can make its own organization or illustration of the information which is received during learning
4. Real Time Operation: ANN computation can be concurrent and can be run parallel to the other methodologies used basing upon the learning
5. Fault Tolerance through Redundant Information

IV. Problem Descriptions In Web Usage Mining:

In this section, it's been described about the problems related with web usage mining.

Web usage Mining is the application of techniques used in data mining to find out the patterns from Web data to understand and serve the needs of web-based applications.

4.1 Logs Processing:

An important step in the identification and finding of user's usage in web sites and habits is the cleaning and mutation of Web server Log files; and the user's session's identification.

4.2 Log Files Cleaning

Cleaning the Web server Log files has several of ways and steps when one user requests a page, then this request is added to the log file, but, if the page has images, they will be added to the Log file. In some of the cases, it is valid to filter a page inserted in others with frames; as it's common to generate the pages dynamically. In HTTP, the error codes are used to filter the records in the Log files, the most frequent errors in HTTP are: error code 200,403(access denied/forbidden), 400(Bad Request), 502(Bad Gateway).

4.3 User's identifications

After the Log files are cleaned, we are required to describe the user's sessions. Some techniques are available to detect the merits and demerits of each session. One of the techniques is analyzing and detecting the use of cookies. The cookies are HTTP headers in a string format. These are useful to identify the users who access the server and what resources the user is accessing. But still there exist a drawback in this methodology which is; the users can lock the use of the cookies, and the server cannot store the information locally in the user side machine; another demerit of this above method is; the user can easily delete the cookies. There is another way to identify the user by using Identd. Identd is a protocol in RFC 1413[RFC 1413], this rule allows detecting a user connected by the unique TCP connectivity [10].

Second method is by detecting the users in log files by the IPs registered in each record. We can also detect users with the users name if included in the log file.

4.4 User Session's identification

After the users have been identified, we are required to find the sessions. For this we can segregate the access of same users in sessions. It's hard to detect when one session ends and another starts. To detect the sessions in common we are required to check the time between two requests; if the two requests in of time frame, we can assume that these requests are in same session.

V. Proposed Work:

In the present work, a method is proposed the using multi-layer perceptron which is helpful in identifying the user's habits. This kind of neural network can be used to find the users patterns of different page access. To find this result required to use the Web Log files to identify the users and session of user's; and this session of users will be helpful to train this Neural Network.

5.1 Multilayer Perceptron and its Algorithm

This section discusses the architecture of a feed-forward network that can is used generally to accelerate the research works. A multilayer feed forward neural network is an interconnection in which data can flow in a single direction, from the input data to the outputs. The number of layers in a neural network is the number of layers of perceptrons. The Feed-Forward Neural Network architecture has the ability to approximate most problems with high accuracy. Error calculation method is used to train neural network. Researchers have investigated many error calculations in an effort to find a calculation with a short training time appropriate for the network's application. MLP is an approach based on error-correction learning rule. Error consists of two pass through the various layers of the network, a forward pass and a backward pass. In forward pass the input is applied to the nodes and the effect travels from layer to layer. Then a set of outputs is generated as actual response of the network. During this pass the weights of the network are fixed. During back pass the weights are all adjusted accordingly with the error-correction algorithm. Then the actual response is found out from the desired result. This error is then passed backward through the network again reverse to the direction of weights. Again the weights are adjusted to make the actual output more near to the desired output.

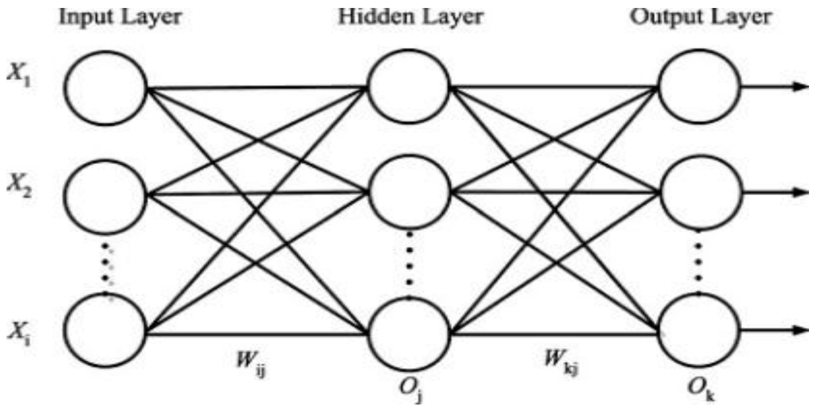

Fig 2: Architecture of MLP

5.2 Algorithm

The algorithm used for the learning is based on back-propagation as discussed above. Algorithm can be written and coded in any programming language. We are assuming the use of sigmoid function f (net).

Algorithm:

(1) Initialize threshold values and synaptic weights to small random values.
(2) The input is Ip = i0,i1,i2,i3,…..,in-1 and the desired output is Dp = d0,d1,d2,….,dm-1 where n is the number of input nodes and m is the number of output nodes. w0 , is the bias , and a0 is 1.For the pattern, Ip and Dp represents the associated patterns. For pattern classification, Dp is assumed to be zero except for one element to 1 that represents the class that Ip is in
(3) Calculating the Actual output
opj = f[w0 x 0 + w1 x 1 + w2 x 2 + …. + wnin]
This response is passed to next layer as input. The final output is apj.
(4) Now we have reached the output layer through forward pass, now again we will be using the output and will move backwards.
wij(t+1)= wij(t)+ αepjapj, where α is the gain and epj is error term for pattern e on node j
For output :
epj=kapj(1-apj)(t-apj)
For hidden layers :
opj= kapj(1-apj)[(ep0wj0 + ep1wj1 + ep2wj2 +….+ epkwjk]
where the sum is over the k nodes in layer after j
(5) Using majority of the nearest neighbors as the prediction value of the new value [6].

VI. Conclusion

In this work, we have seen the usage of neural networks and its learning abilities to recognize and classify the web data mining. The use of useful knowledge and information, user information's and the various server access allows the Web based companies to mine the user access patterns and supports in future maintenance and developments and to aim more advertising campaigns aimed at a group of users. To do this we were required to identify the common patterns in Web, where MLP serves a better solution. In comparison to K-means and SOM, MLP is a better option for classification. In future, this algorithm can be useful to be implemented for efficient results.

References

[1] Fedja Hadzic, Michael Hecker, "Alternative Approach to Tree-Structured Web Log Representation and Mining", *IEEE International Conferences on Web Intelligence and Intelligent Agent Technology*, pp.235-242, 2011.

[2] Saeed R. Aghabozorgi, and Teh Ying Wah, "Dynamic Modeling by Usage Data for Personalization Systems",*IEEE 13th International Conference on Information Visualization*, pp.450-455, 2009.

[3] Mohd Wazih Ahmad, "A Survey: Soft Computing in Intelligent Information Retrieval Systems", *IEEE 12th International Conference on Computational Science and Its Applications*, pp. 26-34, 2012.

[4] R. Cooley, B. Mobasher, and J. Srivastava, "Web Mining: Information and Pattern Discovery on the World Wide Web", *In Proceedings of the 9th IEEE International Conference on Tool with Artificial Intelligence*, pp.558-567, 1997.

[5] Ioannis Kazanidis, Stavros Valsamidis, "An Applicable Method for Evaluation of Web-Learning Platforms Content Quality and Usage", *Proceedings of the IEEE 16th Panhellenic Conference on Informatics*, pp. 186-191, 2012.

[6] Prakash S Raghvendra, Shreya Roy Chowdhury, "Web Usage Mining Statistical Classification and Fuzzy Artificial Neural Networks", *International Journal Multimedia and Image Processing*, **1,** pp. 9-16, March 2011.

[7] S. K. Pani, L. Panigrahy, Web Usage Mining: "A Survey on Pattern Extraction from Web Logs", *International Journal of Instrumentation, Control & Automation* (*IJICA*),**1,** pp. 15-23, 2011.

[8] Li Chaofeng, Research on Web Session Clustering, JOURNAL OF SOFTWARE, Vol.4, pp.460-468, July2009.

[9] C.P. Sumathi, R. Padmaja Valli, "Automatic Recommendation of Web Pages in Web Usage Mining", *IJCSE*, **2,** pp. 3046-3052, 2010. https://tools.ietf.org/html/rfc1413.http://lion.disi.unitn.it/reactive-search/thebook/ node58.html