Neural Networks and Web Mining

R. Vasudevan

Dept of ECE, M.A.M Engineering College
Trichy.

ABSTRACT

The World Wide Web (WWW) is huge, universal and it contains unstructured and heterogeneous data. The expectations in growth of World Wide Web get increased in recent years. There are several billions of documents in the form of HTML and XML, pictures, audios, videos and other multimedia files are embedded in the internet and the number is still increasing now-a-days. But one thing is to retrieving the interesting content has become a very difficult task in the web. Hence the web usage mining is used, which one of the techniques of web mining, it is very useful to discover knowledge from secondary data obtained from the interaction from the users with the web. The web usage mining is mining of usage data captured through various logs stored on server, client or proxy. It is very essential for obtaining useful information throughout the web. Artificial Neural Network (ANN) is information processing system made up of large number of processing elements called neurons. In this paper, discuss about basic ideas of web usage mining and artificial neural networks.

1. INTRODUCTION

In general, data mining is the process of analyzing data from different angles and summarizing it into useful information or extracting the useful information from the huge level of data. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases [1]. The other terms for data mining are Knowledge mining from data, Knowledge extraction, Data/Pattern analysis, Data dredging, Data archaeology. Web mining is the application of data mining techniques to discover patterns or trends followed by the user from the web [2].

1.1. REASON FOR WEB MINING

The requirement of web mining is used to store information on World Wide Web (WWW). The information is growing rapidly in the web so this gives what users want is important. It is again needed to reduce the time loss experienced by users while browsing for the required information retrieved.

1.2. AREAS OF WEB MINING

Patterns followed by the users are evaluated from the techniques of web mining and then these patterns are analyzed to get a user desired output. The desired output is fed into the user understandable GUI [3]. In web mining, the areas can be broadly categorized into three ways. It includes:

- Web content mining
- Web structure mining
- Web usage mining

1.2.1. WEB CONTENT MINING

Web Content Mining is the mining, extraction and integration of useful data, information and knowledge from web page content which can be text or multimedia or both [2].

content mining

1.2.2. WEB STRUCTURE MINING

Web Structure Mining is the process of using graph theory to analyze the node and connection structure of a website According to the type of web structural data. Web structure mining can be divided into two kinds: [2].

- Extracting patterns from hyperlinks in the web: A hyperlink is a structural component that connects the web page to a different location.
- Mining the document structure: Analysis of the tree-like structure of page structures to describe HTML of XML tag usage.

structure mining

1.2.3. WEB USAGE MINING

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications [2].

usage mining

1.2.3.1. WEB USAGE MINING PROCESS

Web usage mining is the process of extracting useful information from the server logs e.g. use web usage mining is the process of finding out what users are looking for on the Internet. Some users might be looking at only textual data, whereas some others might be interested in multimedia data. Web usage mining is the application of data mining techniques to discover interesting usage patterns from web data in order to understand and better serve the needs of web based applications. Usage data captures the identity or origin of web users along with their browsing behavior at a website. Web usage mining itself can be classified further depending on the kind of usage data considered: [2]

- Web Server Data: The user logs are collected by the Web Server. Typical data includes IP address, page reference and access time [2].
- Application Server Data: Commercial application servers have significant features to enable e-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs [2].
- Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them thus generating histories of these specially defined events. It must be noted, however, that many end applications require a combination of one or more of the techniques applied in the categories above [2].

1.3. STRUCTURE OF DATA IN WEB LOGS

The log files are text files that can range in size from 1KB to 100 MB, depending on the traffic at a given website. The data will be taken for any particular website at given time. There are various fields in the log data which includes: [4]

- IP address: This is the IP address of the machine that contacted our site
- Username: This is the user that requested that website
- Timestamp: It is the timestamp of the visit
- Access request: It is the request made
- Result status code: This is whether URL was successfully returned or not. A number is saved stating whether request was successfully answered or not
- Bytes transferred: The number of bytes transferred after request was responded to by the server
- Referrer URL: This is the page referred by the user
- User agent: It is the software that the user is using to access the website. It is actually browser used by the user

1.4. PHASES TO PERFORM WEB USAGE MINING

The phases to perform web usage mining are as follows:

- Preprocessing: It is a process of preparing data so that it can be used for Pattern Discovery and Analysis. It includes Cleaning of Server Log files accompanied by identification of user sessions and user habits [5]. It consists of
  - Data field extraction
  - Data cleaning
  - User identification
  - Session identification

- Pattern Discovery: After the data is preprocessed, this data is utilized for discovering homogeneous patterns [3].
- Pattern Analysis: Once the patterns are discovered then these patterns are evaluated and analysis is performed on these patterns and result generated is given to neural network for further processing [5].

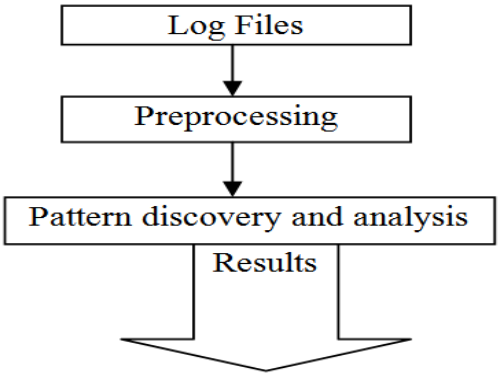

Figure 1: Phases of Web Usage Mining process.

1.5.WEB USAGE MINING ARCHITECTURE

The WEBMINER is a system that implements parts of this general architecture[7,8]. The architecture divides the Web usage mining process into two main parts. The first part includes the domain dependent processes oftransforming the Web data into suitable transaction form. This includes preprocessing, transaction identification, and data integration components. The second part includes the largely domain independent application of generic data mining and pattern matching techniques (such as the discovery of association rule and sequential patterns) as part of the system's data mining engine. The overall architecture for the Web mining process is depicted in Figure 2. Data cleaning is the first step performed in the Web usage mining process. Some low level data integration tasks may also be performed at this stage, such as combining multiple logs, incorporating referrer logs, etc. After the data cleaning, the log entries must be partitioned into logical clusters using one or a series of transaction identification modules. The goal of transaction identification is to create meaningful clusters of references for each user. The task of identifying transactions is one of either dividing a large transaction into multiple smaller ones or merging small transactions into fewer larger ones. The input and output transaction formats match so that any number of modules to be combined in any order, as the data analyst sees fit. Once the domain-dependent data transformation phase is completed, the resulting transaction data must be formatted to conform to the data model of the appropriate data mining task. For instance, the format of the data for the association rule discovery task may be different than the format necessary for mining sequential patterns. Finally, a query mechanism will allow the user (analyst) to provide more control over the discovery process by specifying various constraints. For more details on the WEBMINER system refer to [7, 8]

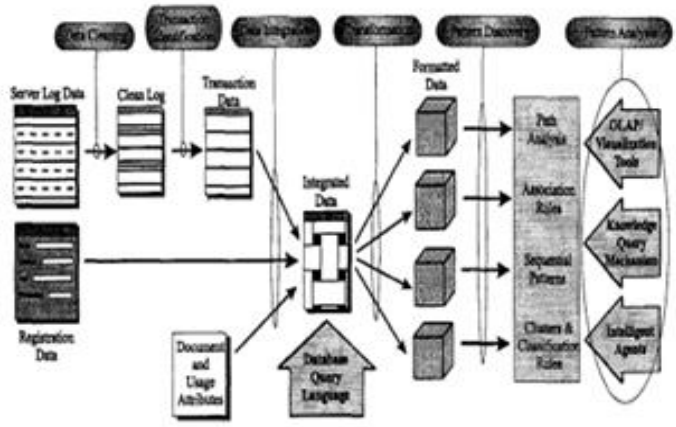

Figure 2: A general architecture for Web Usage Mining

1.6.PROBLEMS FACED WHILE PERFORMING WEB USAGE MINING

The following problems to face while performing web usage mining[6]

- Processing of logs that is cleaning of log files
- Cleaning of log files that is removing data that is not relevant
- Identification of user sessions
- Identification of user habits

2. ARTIFICIAL NEURAL NETWORK:

An artificial neural network is a computational simulation of a biological neural network. These models mimic the real life behavior of neurons and the electrical messages they produce between input (such as from the eyes or nerve endings in the hand), processing by the brain and the final output from the brain (such as reacting to light or from sensing touch or heat). There are other ANNs which are adaptivesystems used to model things such as environments and population.

2.1.FEATURES OF NEURAL NETWORK

The different features of neural network are as follows [9]

1. Nonlinearity: An artificial neuron can be linear or nonlinear. A neural network made up of an interconnection of nonlinear neurons. So neural network itself a nonlinear. Nonlinearity is a special kind of sense that is distributed throughout the network. Nonlinearity is a highly important property, particularly if the underlying physical mechanism responsible for generation of the input signal (e.g. speech signal) is inherently nonlinear.

2. Input-Output mapping:A popular paradigm of learning with a teacher or supervised learning. It involves modification of the synaptic weights of a neural network. It done by applying set of labeled training samples or task examples. Each example consists of a unique input signal and a corresponding desired response. The network is presented with an example picked at random from the set. The synaptic weights of the network are modified to minimize the difference between the desired response and the actual response of the network. The training of network is repeated until it reaches to the steady state it means there is no significant change in the synaptic weights. The previously applied training examples may be reapplied during the training session but in a different order. So network learn from the examples by constructing an input output mapping for the problem.

3. Adaptivity: Neural networks have a built in capability to adapt change in synaptic weights according to the surrounding environment.

4. Evidential response:A neural network can be designed to provide information not only about which particular pattern to select but also about the confidence in the decision made. This information may be used to reject ambiguous patterns and improve the classification performance of the network.

5. Contextual information: It is deal with naturally by a neural network. Knowledge is represented by vary structure and activation state of a neural network. Every neuron in the network is potentially affected by the global activity of all other neurons in the network.

6. VLSI implementability: The massively parallel nature of neural network makes it potentially fast for the computation of certain tasks. So it is well suited for VLSI technology.

7. Uniformity of analysis and design: In neural network all domains of application use the same notations. So it is easy to share theories and learning algorithms between different applications of neural network. It provides seamless integration of modules.

8. Fault tolerance:A neural network implemented in hardware form, has the potential to be inherently fault tolerance. In order to make neural network fault tolerance it is necessary to train network such a way.

9. Neurobiological analogy: The design of a neural network is motivated by analogy with the brain. It gives proof that fault tolerance is fast and powerful. Neurobiologists look to neural networks as a research tool for the interpretation of neurobiological phenomena.

2.2. TYPES OF NEURAL NETWORK ARCHITECTURES

There are fundamentally three different types of architecture of neural network. They are differing byhow neurons of neural network are linked with learning algorithms. Following are different types.

- Single layer feed forward network
- Multilayer feed forward network
- Recurrent network

2.2.1. Single layer feed forward network

The simplest kind of neural network is a single-layer perceptron network, which consists of a single layer of output nodes; the inputs are fed directly to the outputs via a series of weights. In this way it can be considered the simplest kind of feed-forward network. The sum of the products of the weights and the inputs is calculated in each node, and if the value is above some threshold (typically 0) the neuron fires and takes the activated value (typically 1); otherwise it takes the deactivated value (typically -1). Neurons with this kind of activation function are also called artificial neurons or linear threshold units. In the literature the term perceptron often refers to networks consisting of just one of these units [10]

2.2.2. Multilayer feed forward network

This class of networks consists of multiple layers of computational units, usually interconnected in a feed-forward way. Each neuron in one layer has directed connections to the neurons of the subsequent layer. In many applications the units of these networks apply a sigmoid function as an activation function [10]

2.2.3. Recurrent network

A **recurrent neural network** (RNN) is a class of neural network where connections between units form a directed cycle. This creates an internal state of the network which allows it to exhibit dynamic temporal behavior. Unlike feed forward neural networks, RNNs can use their internal memory to process arbitrary sequences of inputs. This makes them applicable to tasks such as un-segmented

connected handwriting recognition, where they have achieved the best known results [11]

2.3. MODEL FOR NEURAL NETWORK APPROACH 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 [6]. Web usage mining consists of four main steps:

A. Data collection
B. Preprocessing,
C. Pattern discovery
D. Pattern analysis

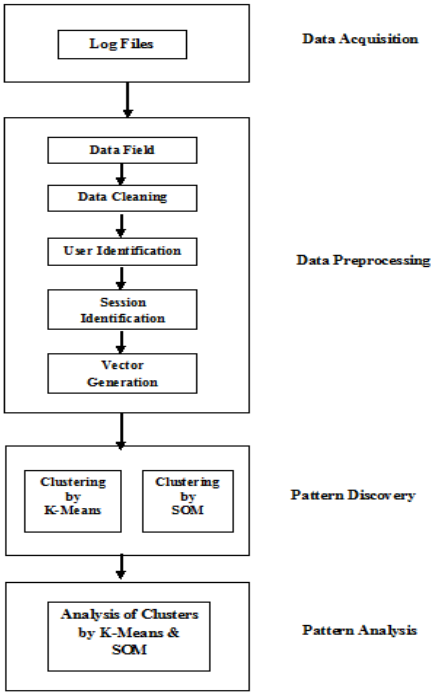

Figure 3: Model for Neural Network Approach for Web Usage Mining.

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. 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 High dimensional data stream contains a tremendous huge amount of data. Such massive amount data contains a large data with high dimensions with data complexity. For example wireless sensor network data, web logs, Google search, etc. Traditional methods are not suitable over high dimensional data as they required very high computation cost for processing data.

2.4. ADVANTAGES OF NEURAL NETWORKS:

1. High Accuracy: Neural networks are able to approximate complex non-linear mappings
2. Noise Tolerance: Neural networks are very flexible with respect to incomplete, missing and noisy data.
3. Independence from prior assumptions: Neural networks do not make a priori assumptions about the distribution of the data, or the form of interactions between factors.
4. Ease of maintenance: Neural networks can be updated with fresh data, making them usefulfor dynamic environments.

5. Neural networks can be implemented in parallel hardware
6. When an element of the neural network fails, it can continue without any problem by their parallel nature

CONCLUSION

This paper has attempted to provide a basic idea about the artificial neural network, web mining. In this paper a general overview of Web usage mining is presented. Web usage mining is used in many areas such as e-Business, e-CRM, e-Services, e-Education, e-Newspapers, e-Government, advertising, Digital Libraries, marketing, bioinformatics and so on. One of the open issues in data mining, in general and Web Mining, in particular, is the creation of intelligent tools that can assist in the interpretation of mined knowledge.

REFERENCES:

[1]http://www.anderson.ucla.edu/faculty/jason.frand/teacher/technologies/palace/datamining.htm
[2] http://en.wikipedia.org/wiki/Web_mining
[3] Anshuman Sharma (2012), "Web usage mining using neural network" International Journal of Reviews in Computing
[4] http://www.web-datamining.net/usage/
[5] SonaliMuddalwarShashankKawar (2012) ,"Applying artificial neural network in web usage mining", Vol 1 Issue 4, International Journal of Computer Science and Management
[6] KetkiMuzumdar, Ravi Mante, PrashantChatur,(2013)" Neural Network Approach for Web Usage Mining" Volume-2, Issue-2, International Journal of Recent Technology and Engineering (IJRTE)
[7] R. Cooley, B. Mobasher, and J. Srivastava. Web mining: Information and pattern discovery on the world wide web. Technical Report TR 97-027, University of Minnesota, Dept. of Computer Science, Minneapolis, 1997.
[8] B. Mobasher, N. Jain, E. Han, and J. Srivastava. Web mining: Pattern discovery from world wide web transactions. Technical Report TR !36-050, University of Minnesota, Dept. of Computer Science, Minneapolis, 1996
[9] Er.Romil.V.Patel,Dheeraj Kumar Singh(2012),Introduction to Integrating Web Mining With Neural Network, IRACST - International Journal of Computer Science and Information Technology & Security (IJCSITS), Vol. 2, No.6, December 2012
[10]http://en.wikipedia.org/wiki/Feedforward_neural_network
[11]http://en.wikipedia.org/wiki/Recurrent_neural_network