A SURVEY ON WEB USAGE MINING WITH NEURAL NETWORK AND PROPOSED SOLUTIONS ON SEVERAL ISSUES

[1] MR. JAYKUMAR M. JAGANI, [2] PROF. KAMLESH PATEL

[1]M.Tech. [Computer Engineering] Student, School Of Engineering R.K. University, Katurbadham, Tramba, Rajkot, Gujarat
[2]M.E. [Computer Engineering], School Of Engineering R.K. University, Katurbadham, Tramba, Rajkot, Gujarat
Jay.jagani009@gmail.com, kamlesh.patel@rku.ac.in

ABSTRACT**:** *Web data is expanding day by day; find a proper way & to manage it becomes mandatory. Extract useful knowledge from WWW data is considered as web mining. It is mainly concern with 3 types, what the content is(content mining),how should be the structure(structure mining)and how and where and how much usage of web data (usage mining).Web usage mining has very emerging implications as network traffic control and flow analysis, adaptive website management, personalization, creation of adaptive websites etc. Neural network is matched with ant colony behavior and have capability of self organization and adaptive learning. Such concept is used for information retrieval and output and decision support system. It is also used for complex classification, optimization and distributed control problems [1]. Such methods are useful in analysis of usage pattern depends on throughput of clustering of all requests [1]. In this paper, we have introduced solutions for self-organizing and growing network that helps in information retrieval from huge web data and also discussed various neural network algorithms i.e. GNG, ART model etc. Web log data files are input for neural algorithms and expected outcome would be optimal representation of network that further used for easy knowledge extraction in web usage mining.*

Keywords—Web usage mining, GNG, Web log data, classification, clustering, artificial neural networks.

1: INTRODUCTION

Nowadays the usage of web resources and data are growing exponentially as numbers of users are increasing every day. Growing data storage and usage make the network (WWW) complex and may unhandle in future. So, few efficient techniques of web usage mining are useful for retrieving knowledge from huge data available on the web. Web mining basically classifies in 3 major categories [2], content mining, structure mining and usage mining. These techniques are used depending upon what to mine from the web.

Now, focus on web usage mining that helps to deal with certain web scaling problems such as user trend analysis of surfing, traffic flow analysis, distributed control handling, web traffic management and many more[1]. Session tracking and website reorganization, distributed traffic sharing on distributed servers can be identified [2] and analysis based on web data can be possible using concepts of neural network. So, applying data mining techniques on web logs, server log files resulted in useful usage path extraction, session tracking, session duration, number of session creations, adaptive web sites and website reorganization[1],[2], [3],[5].

As the Web data increasing over the time, to find the useful pattern or knowledge efficiently, Ant colony behavior, self organizing [2] concept should be used for the network. Such neural network concept is very useful for adapting manageable usage mining from Web. Neural network is far different from static networks in which each node is self-intelligent, hence the network becomes intelligent. So, web users can use this network more and more. This concept is widely useful for web usage mining for extracting information for web traffic analysis on live servers and frequent usage path analysis and many more concepts. In this paper, the discussion is based on mining process and parameters and use of neural network and many such algorithms to overcome from various issues and have a great mining result.

2: PARAMETRES OF WEB USAGE MINING

Web usage mining is based on certain parameters such as, ipaddress, size of data, time stamps (creation / release), pages, method [5], distance, authenticated users, Location of web servers & clients, etc. According to such parameters, the process of path extraction, session tracking, and usage duration and frequency usage of various services has been carried out. By using this various parameters we can extract useful knowledge such as traffic analysis, congestion assumptions when traffic is more in particular region which causes low speed reply or session time out scenarios. So, according to the analysis based on above parameters we can extract hidden and useful information about web usage and it would be very useful for business analysis, traffic control and so on.

2.1: PROCESS OF WEB USAGE MINING

The process of web usage mining is shown as per figure 1,

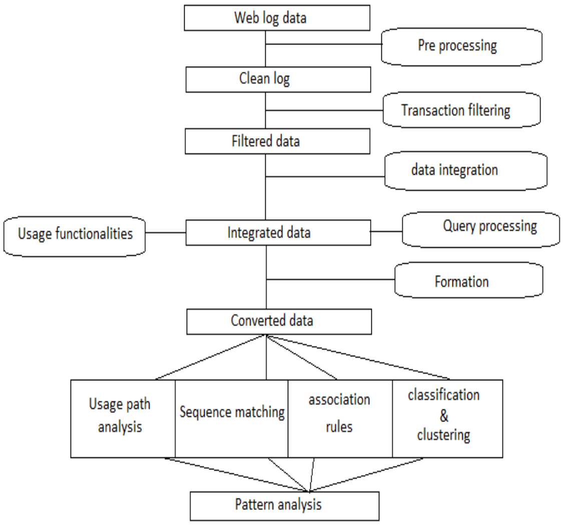


Figure-1 General architecture of web usage mining [5]

According to the figure 1 the processing steps and various techniques are similar to the data mining process. Just difference in data mining and web mining is that at the initial level in data will come from various data bases and warehouses and in web mining data will come from server log files.

2.2: COLLECTION OF INFORMATION

Web usage mining applications can gather data mainly from 3 sources [3]. (1)Web servers (2) Proxy servers (3) Web clients [3]. The largest source of web data is web servers, the huge mass of data can be available there. In web servers, data is generally presented in slandered common log format, extended log format and LogML [3],[6]. For example **ECLF** (Extended common log format) is generally used in web servers.

ECLF format

Ip add.	rfc 931	Auth user	Req. time stamp	Req. format	Status
Bytes	Referrer	User agent			

Table 1 (various columns of format)

Important terms

Ipaddress- network address of user machine
Rfc 931- remote login name of user
Status- as success / page not found like errors
User agent- software or browser (web client)
Authuser – original user name
Bytes – size of transferred information

2.3 VARIOUS ISSUES IN DATA USAGE

Caching: The data is stored on the server in cache hierarchy. It is possible to mismatch in the local cache data access patterns and web server log records [5]. e.g. user has visited page hierarchy as page 1, page2, page1, page3 but due to data in caching server has recorded log as page1, page2, page3 as second time access of page 1 would directly been from cache. So the second entry of page 1 is missed from log [5]. So, we cannot say that log that every time 100% correct data. Thus, caching is a very big issue for accessing web data.
CGI data: CGI is referred as Common Gateway Interface and it is used to pass variables and user entered data to respective server. CGI has a functionality to hide the username and value pairs from URI. So, the data is accessed by whom cannot be tracked by usage mining methods.
Session identification: Tracking and finding the session creation and usage duration and especially when parallel login with same account through different machines makes identification complex.
Dynamicity of Pages: Dynamic pages may change their content according to user request or fixed time interval. So, even minor change in content makes the log data huge as result.
Transaction uniqueness: Issues in identifying unique users and their unique transactions as same account multiplicity is available.

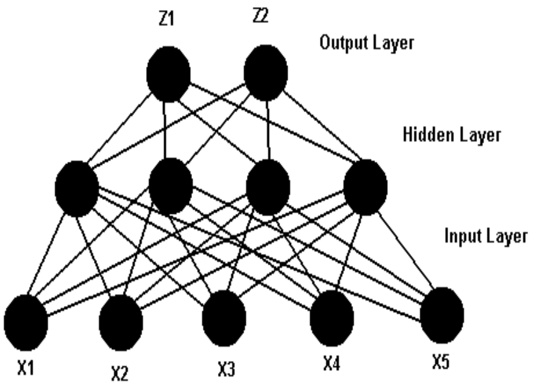
3. NEURAL NETWORK COMES IN ACTION

The use of neural network is an efficient technique for web usage mining to extract hidden knowledge from web-data in easy and efficient manner.

3.1 INTRODUCTION TO ARTIFICIAL NEURAL NETWORK

Artificial neural network is a knowledge processing paradigm that inspired from biological nervous system [1]. The main feature of this technology is its unique structure of system. ANN is the combination of large number of highly connected processing elements, called neurons in medical science (brain), working together to solve a particular problem.
ANN is configured to solve a particular problem such as, pattern reorganization or classification [1]. Neural networks are highly capable to do the things such as meaning derivation from complex data. It is mostly used in finding usage trends that are sometimes very common but complex and even not carried out by machines.
Neural network has many advantages: **Self adaptive learning-**How to deal with problems based on initial training or experience from the network. **Self organization-**Artificial neural network creates its own organization and behavior and representation of knowledge it accepts while

learning. **Real time operations-**Neural network computation may carry out parallel but for that special hardware are required to design. **Fault tolerance via multiple information copies -**partial destroys or failure of network cannot affect the performance of the network.



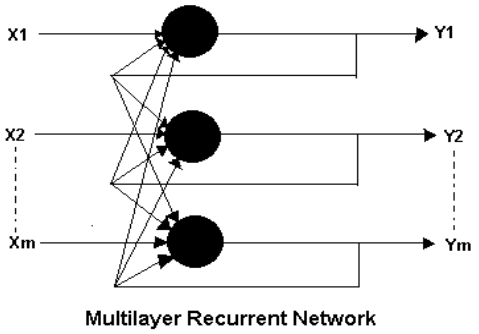


Figure -2 multi layer recurrent networks

As shown in the figure 2, neural networks are basically works in three layers as input, hidden and output layer. And multi layer recurrent network is there; it is very useful for web usage mining especially for traffic analysis and distributed control

4. ISSUES IN VARIOUS STAGES OF WEB USAGE MINING

From figure 1 we can see many stages of the process of web usage mining.

4.1 LOG PROCESSING

Logs are present on the web in ECLF format (Table 1). So, to extract information from huge logs for various purposes where each purpose having parameters is difficult.

4.2 LOG CLEANING

Server data may not always been clean, there may be many redundant entries in one server from the same client. e.g. A client may use same server with diff browser at the same time that may create duplication in the log data. So, managing such situations is necessary and server logs must be cleared before further processing.

4.3 RECOGNIZATION OF USER

Once log is cleaned, move forward to identify the user. User can be identified through its session. There are many methods such as using cookie [7] or using Identd [protocol in RFC 1413] [7].

4.4 SESSION TRACKING

When one user is using multiple sessions then it is somewhat difficult to detect because user is in which session when one session is ending and other is going to start [2]. We can identify using time stamp value. But processing from a large log data and many sessions, it is very complex and difficult.

4.5 USAGE TREND ANALYSIS

After processing everything there is a need to analyze what is the trend of user, or particular application.

5. IDEAS TO RESOLVE ISSUES IN USAGE MINING

In general, web usage mining dealing with unsupervised data where there are few issues as mentioned in section-4, to overcome these issues some of the proposed work is as below:

5.1GROWING NATURAL GAS (GNG)

It is one type of artificial intelligent network, inspired by self organizing map and introduced in 1991 by Thomas maninetz and Klaus schulten. The natural gas is one algorithm for finding optimal representation using feature vectors [2]. The algorithm named as “natural gas” because during adaptation process, they distributes themselves like a gas in the whole network data space. The parameters of GNG are constant in time and as it is incremental, so no need to determine number of nodes a priori [1].

Growing natural gas approach is very much useful in analysis like **user trend analysis.** GNG is one type of neural network which acquires the users by identifying the pattern of page accesses by users. To get the outcomes we need to do some process on web log files to identify users and session of users. According to the session’s user, we have to train ANN [1]. Now select self organization method [1] because using this method doesn’t require supervising the training. The method of self organizing the multilayered recurrent neural network is designed and used to train sample logical neural networks for web usage mining [1]. For heuristic self organization methods, concept of ANN is evolutionary. However, the complexity of the synthesized neural network can’t be optimal as the results of the heuristic self-organization methods depend upon the defined configuration of the Selection-criterion and the freedom of the candidate-structures selection [1].

5.2 ART MODEL MECHANISMS

ART approach does web log analysis via introducing ART structure for huge, widely distributed, highly heterogeneous, semi structured, interconnected and evolving hyper text information repository of WWW [3]. ART system has two sub systems, (1) attention subsystem (2) orienting subsystem. Stabilization of learning and activation occurs in attention sub system. In this method, bottom-up input activation and top down expectation [3]. Orienting subsystem used to handle the mismatch occurring in attention sub system.

Properties of ART

ART model has very basic four properties. (1) Self scaling computation units (2) self adjusting memory search (3) Previous patterns directly access their respective category (4) The system behaves as a teacher and according to environment, it changes its vigilance.

Use of ART in Web Usage Mining

The ART model was anticipated for unsupervised clustering of binary data [3]. It has one layer neural network in its attention subsystem. Moving towards the process of ART, it has fixed no of input neurons to understand that no of dimensions and no of output neurons to map with same amount of maximum clusters. Initially output neurons are not assigned. Once output neuron trained from a pattern, it will become assigned. The activation function is calculated at all assigned output neurons. The input and output is connected by both top-down, and bottom-up weights [3]. The main steps of this approach can be as follows: (1) Web log data collection (2) data pre-processing (3) clustering unsupervised data (4) Web usage mining after above steps [3].

5.3 EXPECTED OUTCOMES

Using GNG, the possible result makes the network intelligent and stable for other users. Network is able to handle large set of web data and optimal representation of the network is possible which is very useful for user trend analysis and usage path frequency analysis and popularity of web clients.

Using ART, the best results can be achieved, specially its feature of dynamic vigilance parameter and bottom up input action and top down expectation approach [8]. Huge amount of web logs can easily been classified using ART model. ART model can classify and cluster any type of complex log data on the basis of specific analysis such pattern identification and session tracking.

6. CONCLUSION

After the survey on web usage mining, for all the mentioned issues and parameters, it is confirmed that the approach of neural network can be very useful because of its adaptive learning nature. The proposed solutions in the form of GNG and ART model are very useful for handling large web log data and to perform usage trend analysis, path analysis and classification based on data respectively. Using neural network concept session creation, session duration, user trend analysis and identification of patterns, etc. are possible with efficiency. So, as growing the web data, neural network approach will play an important role in today and upcoming days in knowledge extraction from web-logs.

7. REFRENCES

[1] Sonali muddalwar, Shashank Kawan, “Applying artificial neural networks in web usage mining”, international journal of computer science and management research, vol 1 issue 4 [Nov-12]

[2] Anshuman Sharma, “Web usage mining Neural network”, international journal of reviews in computing, vol 9, [10th april, 2012].

[3] Valishali A. Zilpe, Dr. Mohammad Atique, “Neural network apprach for web usage mining”, ETCSIT, published in IJCA [2011]

[4] Farhad F. Yusifov, “web traffic mining using neural networks”, world acedamy of science, Engineering and technology, 21, [2008]

[5] Jaydeep Srivastava, “Web Mining: Accomplishments and future directions”, http://www.cs.unm.edu/faculty/srivastava.html

[6] John R. Punin, Mukkai S. Krishnamoorthy, Mohammed J. Zaki, “Web usage mining- language and algorithms”, rensseluer polytechnic institute, troy NY 12180.

[7] Eirinaki and vazirgiannis, 2003, Huysmans et al., 2003.

[8] S.Sherma, M. Varshney, “An efficient approach for web log using ART”, ICEMT [2010].