Neural Network Approach for Web Usage Mining

Vaishali A.Zilpe[#1], Dr. Mohammad Atique[#2]

[#1]Government College of Engineering, Amravati,
Maharashtra, India.
[#2]Associate Professor, SGBAU, Amravati,
Maharashtra, India

ABSTRACT
From the last few decades, we have witnessed an explosive growth in the information available on the World Wide Web (WWW). Today, web browsers provide easy access to myriad sources of text and multimedia data. More than millions of pages are indexed by search engines, and finding the desired information is not an easy task. The users want to have the effective search tools to find relevant information easily and precisely. The Web service providers want to find the way to predict the users' behaviors and personalize information to reduce the traffic load and design the Web-site suited for the different group of users. This profusion of resources has prompted the need for developing automatic mining techniques on the WWW, thereby giving rise to the term "web mining." The analysis of Web log may offer advices about a better way to improve the offer, information about problems occurred to the users, and even about problems for the security of the site. The key component of this paper is a Web-mining approach for Web-log analysis via introducing ART structure [1] for huge, widely distributed, highly heterogeneous, semi-structured, interconnected, evolving, hypertext information repository of World Wide Web. So, the Web sites automatically improve their organization and presentation by self-learning.

Keywords
ART (Adaptive Resonance Theory), attention-subsystem, orienting-subsystem, Web log, Web mining, Web usage (web-log) mining.

1. INTRODUCTION
The enormous content of information on the World Wide Web makes it obvious candidate for data mining research. The majority of non-expert users find it difficult to keep up with the rapid development of computer technologies, while at the same time they recognize that the Web is an invaluable source of information for their everyday life. As more data are becoming available, there is much need to study web-user behaviors to better serve the users and increase the value of enterprises Web data usually exhibits the following characteristics [2]: the data on the Web is huge in amount, distributed, heterogeneous, unstructured, and dynamic. As more data are becoming available, there is much need to study web-user behaviors to better serve the users and increase the value of enterprises. One important data source for this study is the web-log data. The aim of this study is to extract rule set and constructs prediction models that predict the user's next requests as well as when the requests are likely to happen, based on the web-log data. Web usage mining can be used to support dynamic structural changes of a Web site in order to suit the active user, and to make recommendations to the active user that help him/her in further navigation through the site he/she is currently visiting Furthermore, with the wide application of Internet and E-commerce, web has been turned into an important approach for information acquiring. There is pressing demands on the recommendation systems which could actively provide users with

The paper is organized as follows: in Section 2, we discuss Web-log mining process in detail; in section 3, we describe structure of web log files; in section 4, we discuss about ART1 in detail with its algorithm flowchart, and proposed results; and finally in section 5, conclusion and future-work.

2. OVERVIEW OF WEB-USAGE MINING
Web usage mining (WUM) belongs discovers the mode of user visiting Web page by mining Web log, recognizes user's faith degree, taste, satisfaction degree by analyzing the rules in the log, discovers potential users, enhances service compete capability of the website. Web usage mining mines secondary data generated by the users' interaction with the web. WUM works on user profiles, user access patterns, and mining navigation paths which are being heavily used by e-commerce companies for tracking customer behavior on their sites.

Web usage mining, also known as Web-log mining is the automatic discovery of user access patterns from Web servers. Organizations collect large volumes of data in their daily operations, generated automatically by Web servers and collected in server access logs. Other sources of user information include referrer logs which contain information about the referring pages for each page reference, and user registration or survey data gathered via CGI scripts.

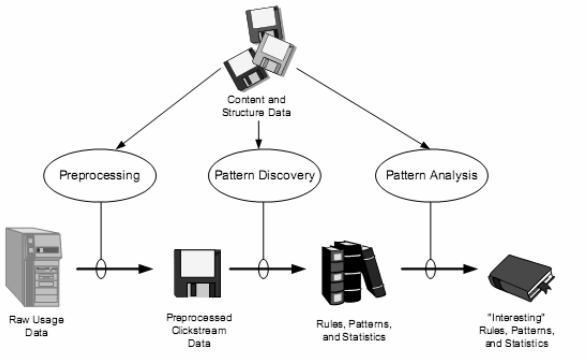

Figure 1. Web usage mining process

2.1 DATA COLLECTION
Web Usage Mining applications are based on data collected from three main sources [3]: (i) Web servers, (ii) Proxy servers, and (iii) Web clients.

Web servers are surely the richest and the most common source of data. They can collect large amounts of information in their log files and in the log files of the databases they use. These logs usually contain basic information e.g. name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc. This information is usually represented

in standard format e.g.: Common Log Format, Extended Log Format, and LogML.

The log files are text files that can range in size from 1KB to 100MB, depending on the traffic at a given a website. In this study, raw log file are collected from Government college of engineering, Amravati website known as www.gcoea.ac.in.

2.2 Data preprocessing

In Data preprocessing [4] phase the web log data must be cleaned, filtered, integrated and transformed in such a way that the irrelevant and redundant data can be removed, user session and transaction can identified. Web log data is usually diverse and voluminous. This data must be assembled into a consistent, integrated and comprehensive view, in order to be used for pattern discovery.

2.3 Pattern discovery

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.

2.4 Pattern analysis

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 [5, 6].

3. STRUCTURE OF WEB-LOG DATA

A log from a Web server (Web log) contains records of users' browsing activities, and is a potentially large source of data on customer preferences [7].Web log data line has format like this:

64.111.11.11 - - [31/Oct/2004:21:45:03 -0800] "GET /cgi-bin/log/source/vs/vs_main.cgi HTTP/1.1" 200 1540096 "http://www.sitename.com/cgi-bin/ai/osp.cgi" "Mozilla/4.7 [en]C-SYMPA (Win95; U)".

Different servers have different log formats. Nevertheless the data in this log fragment is pretty typical of the information available. Let's look at one line from the above fragment (split for easier viewing).

1. IP address: *"64.111.11.11"*

This is the IP address of the machine that contacted our site.

2. Username etc: *"- -"*

Only relevant when accessing password-protected content.

3. Timestamp: "*[31/Oct/2004:21:45:03 -0800]"*

Time stamp of the visit as seen by the web server.

4. Access Request : "GET */cgi-bin/log/source/vs/vs_main.cgi* HTTP/1.1"

The request made. In this case it was a "GET" request (i.e. "show me the page") for the file "/cgi-bin/log/source/vs/vs_main.cgi" using the "HTTP/1.1" protocol. A "HEAD" request fetches only the document header, and is the web equivalent of a "ping" to check your page is still there and hasn't changed.

***5.* Result Status Code:** *"200"*

The resulting status code. "200" is success. If the requested URL didn't exist, this is where the dreaded "404" would have shown up in the log.

***6.* Bytes Transferred:** *"1540096"*

The number of bytes transferred. If this matches the size of the file requested, so this is a successful download. If the number is less, then that would indicate a failed or partial download. Some user agents can fetch files a bit at a time. Each bit will show up as a separate line in the log file, so a series of "hits" whose total adds up to, or exceeds, the file size could indicate a successful download. On the other hand it could indicate someone having trouble connecting to site who has to keep reconnecting.

7. Referrer URL: *"http://www.sitename.com/cgi-bin/ai/osp.cgi"*

The referring page. Not all user agents supply this information. This is the page the visitor is on when they clicked to come to this page. Sometimes this is simply the page the user was looking at when they typed in address into their browser, or clicked on the address in some other software such as a newsreader or an email client.

This information is very useful to webmasters, as it allows them to measure which sites are driving traffic to their site. It also represents a small loss of privacy, as it lets us see where visitors are coming from.

***8.* User Agent:** *"Mozilla/4.7 [en]C-SYMPA (Win95; U)"*

The "User Agent" identifier. The User Agent is whatever software the visitor used to access this site. It's usually a browser, but it could equally be a web robot, a link checker, an FTP client or an offline browser. The "user agent" string is set by the software manufacturer, and can be anything they choose to be. In this case "Mozilla/4.7" probably means Netscape 4.7, "[en]" probably implies it's an English version, "Win 95" indicates Windows 95 etc, etc. Well-behaved web bots and spiders will usually use this string to identify themselves, their web site and an email address.

4. PROPOSED SCHEME

The objective is to provide an acceptable solution at low cost by seeking for an approximate solution to problems. Soft computing methodologies (involving fuzzy sets, neural networks, genetic algorithms and rough sets) hold promise in Web mining.

The proposed approach includes Web-log analysis via introducing ART structure for huge, widely distributed, highly heterogeneous, semi-structured, interconnected, evolving, hypertext information repository of World Wide Web. ART architecture models can self-organize in real time producing stable recognition while getting input patterns beyond those originally stored. ART is a family of different neural architectures where, the most basic architecture is ART1 (Carpenter, and Grossberg, 1987). ART1 can learn, and recognize binary patterns. ART2 (Carpenter, and Grossberg, 1987) is a class of architectures categorizing arbitrary sequences of analog input patterns. ART is used in modeling such as invariant visual pattern recognition where biological equivalence is discussed in 1990.

An ART system consists of two subsystems, an attention-subsystem, and an orienting subsystem (Figure 2). The stabilization of learning and activation occurs in the attention-subsystem by matching bottom-up input activation, and top-down expectation. The orienting subsystem works like a novelty detector. It controls the attention-subsystem when a mismatch occurs in the attention-subsystem.

4.1 Properties of ART

An ART system has four basic properties.

1. Self-scaling computational units. The attention subsystem is based on competitive learning enhancing pattern features but suppressing noise.
2. Self-adjusting memory search. The system can search memory in parallel, and adaptively change its search order.

3. Already learned patterns directly access their corresponding category.
4. The system can adaptively ovulate attentional vigilance using the environment as a teacher. If the environment disapproves the current recognition of the system, it changes this parameter to be more vigilant.

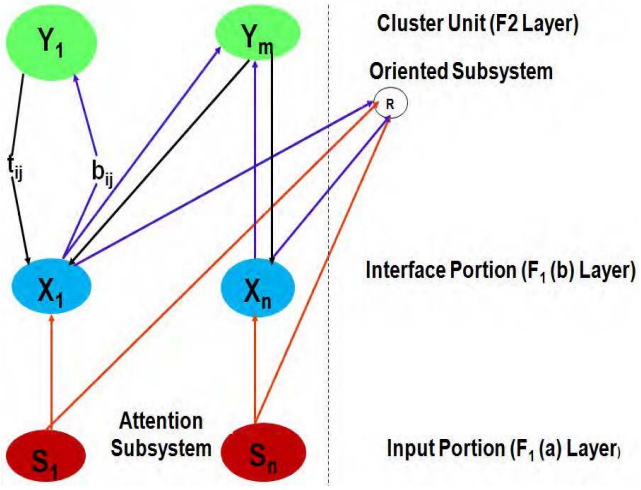

Figure 2. Basic Structure of ART

5. ART-1

There are two models of ART-1, a slow-learning, and a fast-learning one. The slow learning model is described by in terms of differential equations while the fast learning model uses the results of convergence in the slow learning model.

ART-1 is the first version of ART-based networks proposed by Carpenter, and Grossberg. The network was intended for unsupervised clustering of binary data. It has two major subsystems: attention-subsystem, and orienting subsystem. The attention-subsystem is a one layer neural network. It has D input neurons to learn D-dimensional data, and C output neurons to map C maximum clusters. Initially all output neurons are uncommitted. Once an output neuron learned from a pattern, it becomes committed. The activation function is computed at all committed output neurons. The input and output is connected by both top-down, and bottom-up weights.

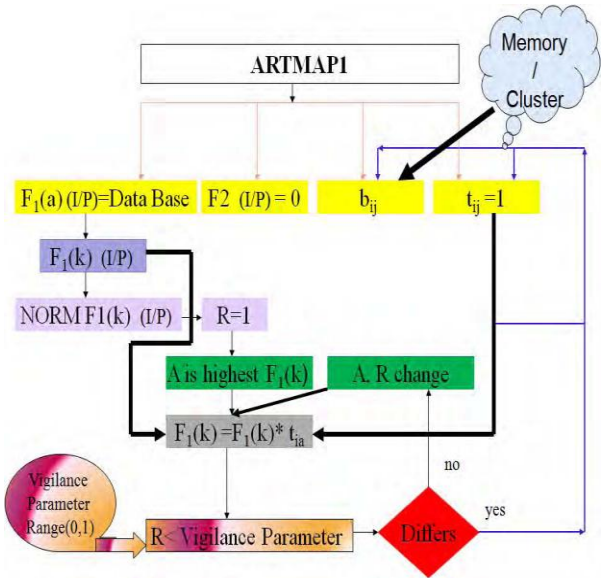

Figure 3. ART1 Algorithm.

Three major steps of this approach can be integrated as follows:-

a. **Web-log data collection**: The logs we research are of W3C Extended Log File Format under IIS5.0 environment. Web log data is collected from the server of website for the period of one month for experimental purpose

b. **Data pre-processing**: We can use database software Access and Java programming language to implement the preprocessing work. Also web-log file preprocessing tools such as WEBMINER, AWStat can be used for data cleaning, user identification and path completion.

c. **Web–usage mining from web-log files**: The final step of web-usage mining can be implemented using neural network approach via. Adaptive resonance network algorithm (Figure 3).

6. PROPOSED RESULTS

If any Web-mining researches apply this ART1, then can easily obtained best result than any implemented Web mining techniques because of vigilance parameter, top-down and bottom-up weights as per study of S. Sharma, M. Varshney [1] .

Because of huge amount of Web-log, it is infeasible to classify them by hand, so by using ART model, we can analyze them in supervised learning. Using this concept, adaptive analysis of web-log data can be done using ART model.

7. CONCLUSION AND FUTURE WORK

Accurate Web usage information could help to attract new customers, retain current customers, improve cross marketing/sales, effectiveness of promotional campaigns, tracking leaving customers and find the most effective logical structure for their Web space. User profiles could be built by combining users' navigation paths with other data features, such as page viewing time, hyperlink structure, and page content [8].

Another variant of ART can produce better result than this one. So, any web-mining researcher can implement such algorithms to obtain more beneficial outputs. In future, ART can also be implemented with all previous techniques like semantic Web log, hybrid information filtering, fuzzy immunity clonal selection neural network, and fuzzy multi-set to build Multi-pass ART, and provide more efficient result.

8. REFERENCES

[1] S. Sharma, M Varshney, "An Efficient approach for web log mining using ART", *International Conference on Education and Management Technology*, 2010 (ICEMT 2010).

[2] Zhang Y.,X. Yu, and J. Hou, "Web communities: Analysis and construction," *Berlin Heidelberg*, 2006.

[3] S. Chakrabarti, M. van den Berg, and B. Dom, "Focused crawling: A new approach to topic-specific web resource discovery," presented at the *8th World Wide Web Conf.*, Toronto, ON, Canada, May 1999.

[4] N Tyagi,A. Solanki and S. Tyagi, "An algorithmic approach to data preprocessing in web usage mining", *Int. journal ofInformation technology and knowledge management*, July-December 2010, Volume 2, No. 2, pp. 279-283.

[5] 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.

[6] 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.**

[7] J. D. Vel'asquez and V. Palade, "Adaptive Websites: A Knowledge Extraction From Web Data Approach," *IOS Press,* Amsterdam, NL, 2008.

[8] Heer, J. and Chi E.H., Identification of Web User Traffic Composition using Multi- Modal Clustering and Information Scent, In *Proc. of the Workshop on Web Mining,* SIAM Conference on Data Mining, pp.51-58, 2001.