ISSN:0975-9646

WEB USAGE MINING USING NEURAL NETWORK APPROACH: A CRITICAL REVIEW

Vaishali A.Zilpe#1, Dr. Mohammad Atique#2
#1Government College of Engineering, Amravati,
Maharashtra, India.
#2Associate Professor, SGBAU, Amravati,
Maharashtra, India

Abstract- **The emergence of WWW has drawn new frontiers for database research. The analysis of Web log files may give information that are useful for improving the services offered by Web portals and information access and retrieval tools, giving information on problems occurred to the users. This study reports on initial findings on a specific aspect that is highly relevant for personalization services: the study of Web user sessions. The key component of this paper is a Web-mining approach for Web-log analysis via introducing ART structure 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.**

Index Terms: **ART (Adaptive Resonance Theory), attention-subsystem, orienting-subsystem, Web log, Web mining, Web usage (web-log) mining.**

I. INTRODUCTION

The enormous content of information on the World Wide Web makes it obvious candidate for data mining research. 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 review is 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 differs from collaborative filtering in the fact that we are not interested in explicitly discovering user profiles but rather usage profiles. 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 individualized information services. Using efficient web-log mining [1] recommendations can be made to the site administrators and designers, regarding structural changes to the site in order to enable more efficient browsing.

The paper is organized as follows: in Section 2, we describe related work in field of Web-log mining; in section 3, we discuss about Web-Usage mining; in section 4, we discuss about web-usage mining architecture; in section 5, we discuss about ART1 in detail with its algorithm flowchart, and proposed results; and finally in section 6, conclusion and future-work.

II.RELATED WORK

Web log mining is one of the important content of web mining, is also a hotspot in data mining domain currently. Web log mining is usually consisted of data preprocessing, pattern discovery and pattern analyzing [3, 4]. During data processing phase, literature [3] presents the method of mining abnormal data, literature [5] presents transaction distinguishes based on most forward reference. During model analyzing phase, Web Watcher [4] tracks users' browse behavior, distinguish the links user may interest in and recommend to them. For every user, Web Watcher firstly describes on interests simply, and then studies this user's interests based on its browser behavior and others' browser behavior, which have similar interests. Literature [5, 6] uses Markov model to create sequence mode to Web fetch and system optimizing. The contribution of understanding Internet user's behavior and identification of relevant pattern is the benefit to predict, and prefetch Web object in Web caching [7].

III.WEB-USAGE MINING

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. Analyzing such data can help organizations determine the life time value of customers, cross marketing strategies across products, and effectiveness of promotional campaigns, among other things. It can also provide information on how to restructure a Web site to create a more effective organizational presence, and shed light on more effective management of workgroup communication and organizational infrastructure. For selling advertisements on the World Wide Web, analyzing user access patterns helps in targeting ads to specific groups of users. Most existing Web analysis tools [8,9] provide mechanisms for reporting user activity in the servers and various forms of data filtering. Using such tools it is possible to determine the number of accesses to the server and to individual files, the times of visits, and the domain names and URLs of users. However, these tools are designed to handle low to moderate traffic servers, and usually provide little or no analysis of data relationships among the accessed files and directories within the Web space. More sophisticated systems and techniques for discovery and analysis of patterns are now emerging. These tools can be placed into two main categories, as discussed below.

A. Pattern Discovery Tools. The emerging tools for user pattern discovery use sophisticated techniques from AI, data mining, psychology, and information theory, to mine for knowledge from collected data. For example, the WEBMINER system [10, 11] introduces a general architecture for Web usage mining. WEBMINER automatically discovers association rules and sequential patterns from server access logs.

B. Pattern Analysis Tools. Once access patterns have been discovered, analysts need the appropriate tools and techniques to understand, visualize, and interpret these patterns, e.g. the WebViz system [12]. Others have proposed using OLAP techniques such as data cubes for the purpose of simplifying the analysis of usage statistics from server access logs . The WEBMINER system [11] proposes an SQL-like query mechanism for querying the discovered knowledge (in the form of association rules and sequential patterns).

IV.WEB USAGE MINING ARCHITECTURE

The general architecture of web usage mining is described here. The WEBMINER is a system that implements parts of this general architecture. The architecture divides the Web usage mining process into two main parts. The first part includes the domain dependent processes of transforming 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 1**.**

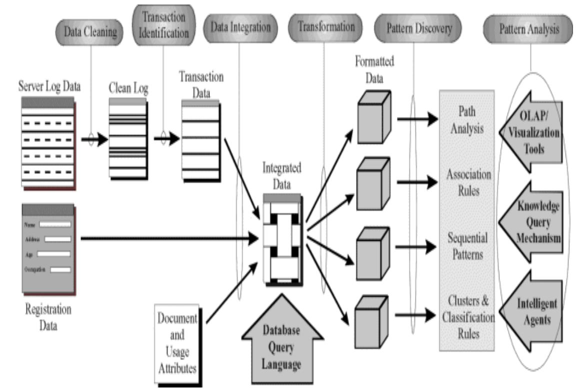

Figure 1**:** General Architecture for Web Usage Mining

1. Data cleaning: 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.

2. Transaction identification: 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.

3. Transformation: 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.

4. 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.

5. 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 [10, 11].

V. FRAGMENT 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 [13].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 (see below) 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 (see below) 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.

VI. 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.

A. *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.

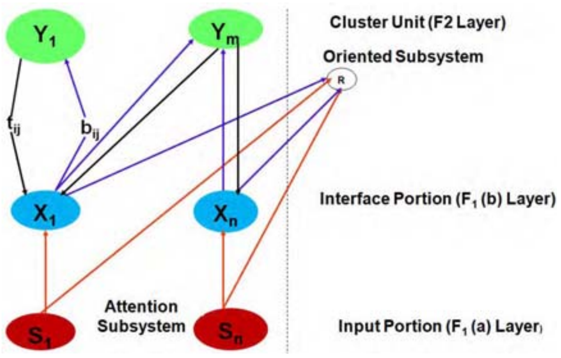

Figure 2: Basic Structure of ART

B. 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.

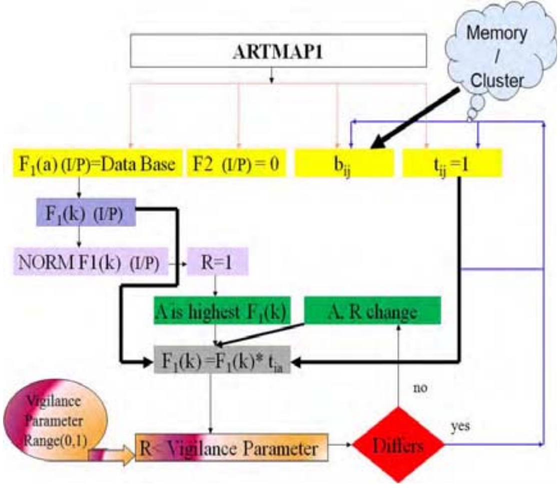

Figure 3: ART1 Algorithm.

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).

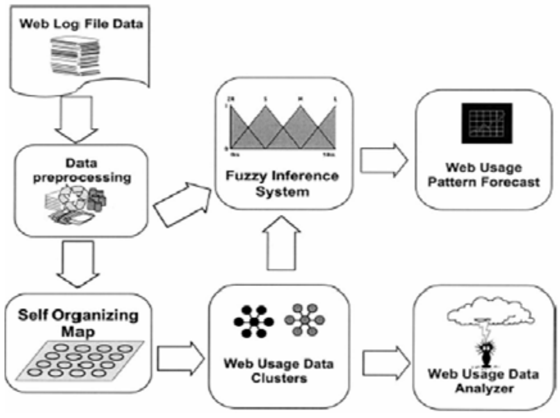

Figure 4: Neuro-Fuzzy Approach for Web Mining

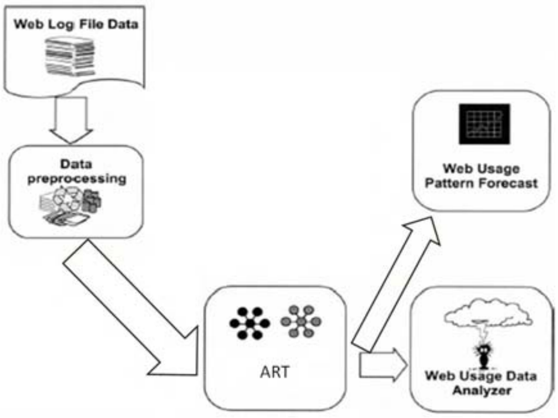

Figure 5: Reduction of Stages on Neuro-Fuzzy after ART implementation.

C. 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.

Also using ART, it is more beneficiary to minimize the number of steps in Web mining as compare to neuro-fuzzy approach. As neuro-fuzzy approach uses five major steps to produce the Web-usage pattern forecast, and Web-usage data analyzer; named Web-log data collection, data preprocessing, self-organizing map, Web-usage data cluster, and fuzzy inference system (Figure. 4). But ART use only three steps as Web-log data collection, data preprocessing, and ART itself (Figure 5).

VII. CONCLUSION AND FUTURE WORK

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.

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.

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] Zaiane O R, Xin M ,Han J. "Discovering Web access patterns and trends by applying OLAP and data mining technology on Weblogs", 1998. **http://citeseer.nj.nec.com/zaiane98discovering.html.**

[4] Cooley, R., Srivastava, J., Mobasher, B.: "Web Mining: Information and Pattern Discovery on the World Wide Web," *Proc. of the 9th IEEE Int. Conf. on Tools with Artificial Intelligence* (1997).

[5] Tim Berners-Lee, James Hendler and Ora Lassila, '"The Semantic Web", *Scientific American*, May 2001.

[6] Joachims T, Freitag D, Mitchell T. Webwatcher: "A tour guide for the world wide web", *In The 15th Intl. Conf. on Artificial Intelligence*, Nagoya, Japan, 1997.

[7] Areerat Sangwattana, "Mining Web logs for prediction in prefetching and caching," *IEEE International Conference on Convergence and Hybrid Information Technology*,2008,pp.1006-1011.

[8] Software Inc. Web trends. **http://www.webtrends.com,** 1995.

[9] Open Market Inc. Open market web reporter. **http://www.openmarket.com,** 1996.

[10] 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.

[11] 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.**

[12] J. Pitkow and Krishna K. Bharat. Webviz: A tool for world-wide web access log analysis. *In First International WWW Conference*, 1994.

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