ISSN:0975-9646

Multipurpose Web Log Analyzer: A Survey

Lalit Agrawal

Department of Computer Science & Engineering,
Acropolis Institute of Technology and Research
Indore bypass road Mangliya square

***Abstract*— Web is a large and dynamic domain of knowledge and discovery. Domain of web is a dynamic and the working with this domain is quite expensive and time consuming so we select a sub-domain known as web mining. There are lots of tools, techniques and methods are proposed and implemented for web log mining and they have its own importance. In this project we work over the information gathering from log mining. we emphasize on a new approach over the information gathering by the mining of web access logs or web usage data.**

Keywords — Web Mining, Web Log Analyzer

I. INTRODUCTION

The World Wide Web (WWW) is one of the most important medium that provides an interface to store, share and distribute information. At present, the figure for Google is index of 8 billion web pages. This global medium is today used in every part of the world, which has led the web designers to think of the latest web technologies and the techniques to keep their website secure.

In this competitive era of today, it has become necessary to meet the user requirements always as there are multiple websites for a particular domain and users compare different website to select the best one for their use. Web log techniques may be used to cope up with such issues to drive the business for any web based system.

The intense use of the Web has provided an opportunity to study user and system behaviour by exploring Web access logs. Web logs, generally associate to as blogs, are studied as online diaries published and maintained by individual users (bloggers), bloggers daily activities reports.

Web mining is primarily aimed at deriving actionable knowledge from the Web through the application of various data mining techniques. Web Usage Mining is the discovery of user access patterns from Web server access logs. It is a suitable technique to discover and extract interesting knowledge/patterns from Web. Web Mining is the application of data mining techniques refers to the overall process of discovering potentially useful and previously unknown information or knowledge from the Web data. The information gathered can be classified into three broad categories of web mining namely: Web Content Mining, Web Structure Mining and Web Usage Mining.

Web Content Mining: Web content mining focus on the extraction of knowledge from the content of web pages and therefore the target data consist of multivariate type of data contained in a web page as images, text, documents, multimedia etc.

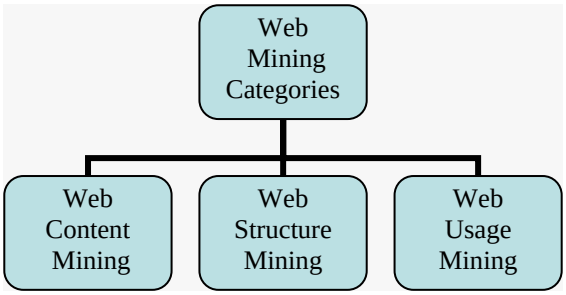

Fig.1 Web Mining Categories

Web Structure Mining: Web structure mining corresponds with the connectivity of websites and the extraction of knowledge from hyperlinks of the web.

Web Usage Mining: Web Usage Mining is the extraction of knowledge from user navigation data when they visit a website. The target data are requests from users recorded in special files stored in the website's servers called log files.

Web Usage Mining (WUM) consists of three main steps: Pre-processing, knowledge discovery and results analysis. The goal of the pre-processing step is to transform the raw web log data into a set of user profile. Every profile takes a sequence or a set of URLs representing a user session. Our approach is to mine the access of the internet in an efficient way. The system mines web log records to discover profile of user and interactions with Web site are its main goal. Uses of Ontology techniques for capturing, modeling and analyzing the behavioral patterns of users in different areas more efficiently. Clustering and Classification algorithm is applied to predict future depending on the current analysis outcomes. -

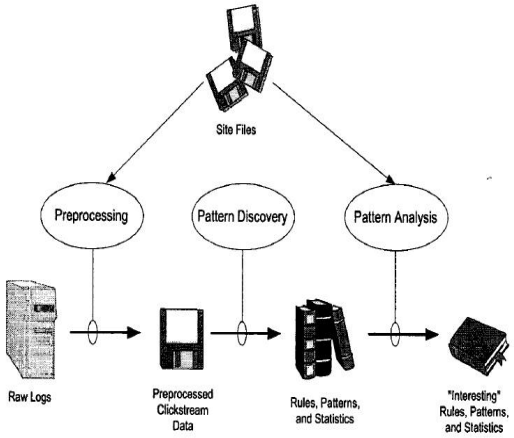

Fig.1 High Level Web Usage Mining Process

Now we will describe the short description of web usage mining process with diagram-
Procedure Data preparation: Web log data are pre-processed in order to identify users, sessions, page views and so on.
Pattern discovery: Statistical methods, as well as data mining methods (such as association rules, sequential pattern discovery, clustering, and classification) are applied in order to detect interesting patterns.
Pattern analysis phases: The patterns are stored and analysis is done in this phase.

A. Goals of Multipurpose Log Analyzer:

- User differentiation from the same IP source.
- Provide the Ontology mapping scheme for fast working with log file.
- Web usage based mining on frequent item sets and our proposed algorithms are the core step of the system.
- Improve the performance of log mining technique.

B. Log Management and intelligence:

Each access to a Web page recorded in the access log of the Web server that hosted it. Fields that follow a predefined format. Example of Web Log Data-

```
161.184.77.234 - - [18/Sep/2001:01:24:44 +0000] "GET
/images/home.gif HTTP/1.1" 304 -
"http://www.123loganalyzer.com/buy.htm" "Mozilla/4.0
(compatible; MSIE 5.5; Windows 98)"
161.184.77.234 - - [18/Sep/2001:01:24:45 +0000] "GET
/images/buy_now-a.gif HTTP/1.1" 200 1388
"http://www.123loganalyzer.com/buy.htm" "Mozilla/4.0
(compatible; MSIE 5.5; Windows 98)"
161.184.77.234 - - [18/Sep/2001:01:25:22 +0000] "GET
/features.htm HTTP/1.1" 304 -
"http://www.123loganalyzer.com/buy.htm" "Mozilla/4.0
(compatible; MSIE 5.5; Windows 98)"
```

Figure 3: Sample Web Server Log

Common Log Format: A typical configuration for the access log might look as follows.

161.184.77.234 - - [18/Sep/2001:01:24:45 +0000] "GET /images/buy_now-a.gif HTTP/1.1" 200 1388 http://www.123 loganalyzer.com /buy.htm "Mozilla/4.0 (compatible; MSIE 5.5; Windows 98)"

The format above is from an Apache log. Depending on the type of server the site is on, the log entries may look different. Thousands (or even hundreds of thousands) of entries such as the one above are placed into a plain text file, called the server log.

The above log entry includes the following information:

- IP address of the requesting computer 161.184.77.234. This is not the user's IP address, but rather the address of the Host machine they've connected to.
- Date and time of the request: [18/Sep/2001:01:24:45].That's September 18, 2001 at 1:24:45 pm and the time zone is 5 hours behind GMT, which is Eastern Standard Time in the USA (this is because the server is in that time zone, not the user.)
- The full HTTP request: " GET /images/buy _nowa.gif HTTP/1.1"
  a. Request method: GET
  b. Requested file: /images/buy_ now -a .gif
  c. HTTP Protocol version: HTTP/1.1
- HTTP Response Code: 200. This particular code Means the request was ok
- Response size: 1388 bytes. This is the size of the file that was returned.
- Referring document: http://www.123loganalyzer.com/buy.htm.
- User-Agent String (Browser & Operating system information):"Mozilla/4.0 (compatible ; MSIE 5.5;Windows 98)

II. RELATED WORK

1. Many different kinds of tools are designed and developed to extract important information from the web log file.

TABLE 1
DIFFERENT WEB LOG ANALYZER TOOLS

Name	Firma	Type	Comments
Web Log Parse	ACME Labs Software	Log files Processing	Extract specific fields from a web log file, support different web log format.
Web Log	Darryl C. Bergdorf	Log files analysis Tools	Keep track of activity on your site by month, week day, page view byte transfer etc.
Analog	University of Cambridge	Log files Analyzer	It tells which page are most popular, which country people visited from, etc

2. The most of tool extract same data from the log. Thus required a new tool by which administrator

can extract more and different information from the analysis of web log file. There are some problem related to existing software are-

- They don't work for user personalization.
- They don't work over the frequent item sets of user navigation with the website.
- They are not able to predict what response will be generated by the server for user request.
- They perform analysis on huge amount of data it takes more time to process the data.

3. Without data mining techniques there is no sense of massive data provided by WWW. Our work is to emphasize on the mining of web access log & web usage data. Our system is intended to provide for Web- Site Maintainers, Web Analysers, Pre-fetched Systems, Web Personalized Systems and Recommender Systems. Our information extraction is based on user's purposes, date and site of web log data.
4. Application of data mining techniques to the World Wide Web, referred to as Web mining, has been the focus of several recent research projects and papers. However, there is no established vocabulary, leading to confusion when comparing research efforts. The term Web mining has been used in two distinct ways. The first, called Web content mining is the process of information discovery from sources across the World Wide Web. The second, called Web mage mining, is the process of mining for user browsing and access patterns.
5. Web usage mining is the area of web mining which deals with the extraction of interesting knowledge from web log information produced by web servers. Web usage mining techniques can be applied for web log analysis. Web access data, traditionally, are stored in the server log files. Several web usage mining approaches have been presented for exposing usage patterns with the most prominent ones being clustering, association rule, and sequential pattern mining.
6. Data mining is the process of exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules. If the conception of computer algorithms being based on the evolutionary of the organism is surprising, the extensiveness with which these methodologies are applied in so many areas is no less than astonishing. At present data mining is a new and important area of research and ANN itself is a very suitable for solving the problems of data mining because its characteristics of good robustness, self-organizing adaptive, parallel processing, distributed storage and high degree of fault tolerance. The commercial, educational and scientific applications are increasingly dependent on these methodologies.
7. Introduced less than twenty years ago, the Web has become the environment where people of all ages, languages and cultures conduct their daily digital lives. Working or entertaining, learning or socializing, home or on the road, individually or as a group, Web users are ubiquitously surrounded by an infrastructure of devices, networks and applications. This infrastructure combined with the perpetually growing amount of every imaginable type of information supports the user's intellectual or physical activity. Whether searching, using or creating and disseminating the information, users leave behind a great deal of data revealing their information needs, attitudes, personal and environmental facts. Web designers collect these artifacts in a variety of Web logs for subsequent analysis.

III. PROPOSED WORK

Problem Formulation:

1. Many different kinds of tools are designed and develop to extract important information from the web log file and most of them extract same data.
2. Thus required a new tool by which administrator can extract more and different information from the analysis of web log files.
3. They are not supported ontology mapping to find information of specific domain.
4. They are not support user profiling which help to identify different user and their request.
5. They are work with specific domain of web usage mining.

Solution Domain**:**

To solve the above described issue we propose the following solutions:

1. Import input data from log files in both IIS and tomcat server log.
2. Extract the formal information related to log file.
3. Apply ontology mapping to get specific information from data.
4. Apply data mining technique to recognize and identify different type of users and their request.
5. Our approach is to construct a hydride web log analyzer, it cover many the application area of web usage mining.
6. Site Modification: In term of both site content and structure.
7. Business Intelligence: Usage Characterization based on content, structure and usage.
8. The contribution of our proposed system real world problems, such as to improve Web sites/search engine usage, to make additional topics or product recommendation to attract more users, study and analysis of users' behavior from multiple points of view.

IV. CONCLUSIONS

Our proposed framework for Web Usage Analyser is intended to apply in many ways such as for Web Site Maintainers, Web Analysers, Pre-fetched Systems, Web Personalized Systems and Recommender Systems because we are mining based on user's purposes, date, and site of web log data then presents the results upon different dimensions. It can also be used in statistical analysis of the information about most often used web sites to inform the particular clients and can be used in Pre-fetched system.

REFERENCES

[1] Framework for Multi-purpose Web Log Access Analyzer 978-1-4244-6349-7/10/$26.00 _c 2010 IEEE V3-289

[2] From Web Mining to Social Multimedia Mining 978-0-7695-4375-8/11 $26.00 © 2011 IEEE DOI 10.1109/ASONAM.2011.32

[3] Data Mining of WHO Data Warehouse with PASW Modeler

[4] Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data

[5] Kosala, R., Blockeel, H., (2000). Web Mining Research: A Survey, *ACM 2*(1):1-15.

[6] Cooley, R., Mobasher, B. Srivastava, J., (1997). Web Mining: Information and Pattern Discoveryon the World Wide Web, 9th International Conference on Tools with Artificial Intelligence(ICTAI'97) New Port Beach, CA, USA, IEEE Computer Society, 558-567.

[7] Non-Redundant Sequential Association Rule Mining and Application in Recommender Systems, 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology

[8] Automatic web personalization system for user profiling.

[9] Ontology-Based Integration of Information —A Survey of Existing Approaches. H.Wache, T. V¨ogele, U. Visser, H. Stuckenschmidt,G. Schuster, H. Neumann and S. H¨ubner

[10] An Implementation of ID3 --- Decision Tree Learning Algorithm Wei Peng, Juhua Chen and Haiping Zhou Project of Comp 9417: Machine Learning University of New South Wales, School of Computer Science & Engineering,Sydney, NSW 2032, Australia