



Survey on Preprocessing in Web Server Log Files

Mrs. Ujawala M. Patil

Associate Professor, Department Of Computer Engineering
R.C.Patel Institute of Technology
Shirpur, India
patil_ujawala2003@rediffmail.com

Priyanka V. Patil*

ME Scholar, Department of Computer Engineering
R.C.Patel Institute of Technology
Shirpur, India
ce.priyankapatil@gmail.com

Abstract: The World Wide Web is a system of hypertext documents accessed via the Internet. In that web pages may contain text, images, videos, and other multimedia and navigate between them via hyperlinks. World Wide Web gives large information to internet user. World Wide Web is a huge repository of web pages and links. When user accesses websites are recorded in web logs file. Web server log file is a simple plain text file. Display of log file data in different format like W3C Extended log file format, NCSA common log file format, IIS log file format. To improve quality of data, log file should be preprocessed. Log files usually contain noisy and unnecessary data. Preprocessing reduce log file size also increase quality of available data. log file is input for mining algorithm. It gives detailed discussion about web log file, web log file format. In this paper we survey about data preprocessing of web log file.

Keywords: preprocessing; web log file; web log file format.

I. INTRODUCTION

The Internet carries a vast range of information resources and services. Web site is a collection number of web pages grouped under the same domain name. When user accesses website, log file are created. Log file recorded information about each user. Tremendous uses of web, web log files are growing at faster rate. Web server maintain web log file. Standard web server like Apache and IIS.

Log files are located in different location like web server, web proxy server, and Client browser. Log file record information in three different formats like W3C, NCSA, and IIS format. Preprocessing log file have different steps data cleaning, user identification, session identification, path completion [1]. This paper is organized in section II explain overview of web log file, in section III location of web log file, section IV we present some log file formats, section V preprocessing web log file, in section VI we discuss some related work, section VII finally, we conclude.

II. WEB LOG FILE

Web server log file is a simple plain text file which record information about each user. Log file contain information about user name, IP address, date, time, bytes transferred, access request. A web log file provides clear idea about user. Analyzing log file are used to detecting attacks on web. When user accesses website, text file created automatically. Log file gives significant information to web server.

Log file gives information about:

- Which pages were requested in website?
- How many bytes sent to user from server?
- What type of error occurs?

When user submit request to a web server that activity are recorded in web log file. Log file used for debugging purpose. Web log file reside in web server.

A. Type of web Server Logs:

- Access log file:** Data of all incoming request and information about client of server. Access log records all requests that are processed by server.
- Error log file:** list of internal error. Whenever an error is occurred, the page is being requested by client to web server the entry is made in error log [2]. Access and error logs are mostly used, but agent and referrer log may or may not enable at server.
- Agent log file:** Information about user's browser, browser version.
- Referrer log file:** This file provides information about link and redirects visitor to site.

III. LOCATION OF LOG FILE

Web log file can be located in 3 different places:

A. Web Server:

Web log gives most accurate and complete usage of data. Data of web log files are sensitive, personal information so server keeps them closed.

B. Web Proxy Server:

This is intermediate server between client and web server. Client send request to web server via proxy server. web proxy server takes HTTP request from user, gives them to web server, then result passed to web server and return to user.

C. Client Browser

Log file can reside in client's browser window itself. HTTP cookies used for client browser [3]. These HTTP cookies are pieces of information generated by a web server and stored in user's computer, ready for future access.

IV. LOG FILE FORMAT

The web log files record information in three different formats:

- W3C Extended log file format
- NCSA common log file format
- IIS log file format

NCSA and IIS log file format the data logged for each request is fixed. W3C format allows user to choose properties, user want to log for each request.

A. W3C extended Log file Format:

This format is customizable. ASCII text format with variety of different field, it is default log file format on IIS server. Field are separated by space, time is recorded as GMT (Greenwich Mean Time).

In W3C format of year is YYYY-MM-DD. W3C log format can be customized that is administrators can add or remove fields depending on what information want to record.

```
#Software: Microsoft Internet Information Services 7.5
#Version: 1.0
#Date: 2012-01-09 03:56:27
#Fields: date time cs-method cs-uri-stem c-ip cs-version sc-status
2012-01-09 3:56:27 GET /WebSite/ ::1 HTTP/1.1 200
```

Figure1. Example of W3C log file format

In Fig. 1 shows that

#software –version of IIS that is running

#version-the log file format

#Date-recording date and time of first log entry.

#fields: this is not standard format because field can be customized. in this format fields : date , time, client IP Address, Method , URI stem, Protocol status, protocol version

The entry designates on 9 Jan, 2012 at 3:56 am, a user with HTTP version 1.1 and GET is client server method. 200 is HTTP status code.

Field that are selected but there no information then ‘-’ is placed.

B. NCSA Common Log File Format:

National Center For Supercomputing Application format. NCSA is fixed format, it cannot customized. It is available for website but not for FTP site. Format of year is DD/MM/YYYY. Fields are separated by space, time is local time.

```
::1 - - [09/Jan/2012:10:00:30 +0530] "GET /Website/
HTTP/1.1" 200 1107
```

Figure 2. Example of NCSA log file format

Fig. 2 means that “Get” method the page called “website” at the date and time previously specified. The request was successfully executed and this is clear from the “200” status code. The object returned to the user was in size of 1107 bytes.

C. IIS Log File Format:

IIS format is not customized, it is fixed ASCII format. Its records more information than NCSA format. Fields are

separated by comma, easy to read. Time recorded in local time.

Fields in IIS are Client IP address, user name, date, time, service and instance, server name, server IP address, time taken, client bytes sent, serve bytes sent, service status code, windows status code, request type, target of operation, parameters.

```
::1, -, 1/9/2012, 9:57:42, W3SVC1, JAY-PC, ::1, 5, 904, 1153,
200, 0, POST, /WebSite/default.aspx, -,
```

Figure 3. Example of IIS log file format

Fig. 3 shows date , time , W3SVC1 is service and instance , JAY-PC is server name, service status code is 200, 0 is windows status code, request type POST.

V. PREPROCESSING WEB LOG FILE

Why preprocessing is necessary, because Log file contain noisy & ambiguous data which may affect result of mining process. Some of web log file data are unnecessary for analysis process and could affect detection of web attack. Preprocessing reduce log file size also increase quality of available data. So preprocessing steps come before applying mining algorithm. There are 3 steps in preprocessing

A. Data Cleansing:

In this step remove noisy and unnecessary data. Remove log entry nodes contain extension like jpg, gif means remove request such as multimedia files, image, page style file. Remove successful status code 200.

B. User Identification:

This step identify individual user by using their IP address. If new IP address, there is new user. If IP address is same but browser version or operating system is different then it represents different user.

C. Session Identification:

Each user spends total time in each web page. Session means time duration spent in web pages.

D. Path Completion:

Path Completion should be used acquiring the complete user access path.

VI. RELATED WORK

L.K. Joshila Grace et al.[2] etc. gives detailed discussion about web log file, their creation, access procedures, various algorithm used and addition parameter used in log file. This paper gives a detailed description of how log file is being processed in web usage mining process. Web usage mining consist of three main steps preprocessing, pattern discovery, pattern analysis.

K.R.Suneeta et. al. [3] studied the user access log files of NASA web server were analyzed to system administrator and web designer to arrange their system by determining occurred system errors, corrupted, broken links. Author describes application of web usage mining. Web usage mining is application of data mining techniques to discover usage patterns from web data.

The three main stages of web usage mining are data preprocessing, pattern discovery, and pattern analysis.

Data preprocessing involves removal of unnecessary data. In pattern discovery data mining techniques used in order to extract pattern of usage from web data. Pattern analysis is to extract interesting pattern from output of pattern discovery by removing irrelative pattern. Pattern analysis is final step of web usage mining.

G. Castellano *et al.* [4] present LODAP tool (log data pre-processor) which design and implemented in order to perform preprocessing of log file. LODAP takes input log file related to website and output a database containing pages visited by user and identified user sessions. LODAP tool reducing size of web log file and grouping web request into a number of user sessions.

Li. Chaofeng [5] presents several data preparation techniques to improve performance of data preprocessing. Also this paper has presented data preprocessing task that are necessary for performing web usage mining, application of data mining and knowledge discovery techniques to WWW server access log .Web usage mining is to discover usage patterns from web data, also to usage logs of large data repositories. Result of web usage mining can be used in system improvement, site modification.

Shaimaa Ezzat Salama *et al.* [6] Focused on many works have been devoted to preprocess data in log file for web usage mining. But in these they discuss on preprocessing. Web log file for intrusion detection.

It involves integrating data from multiple log files into one single file. They were combining them in XML file. After that preprocessing steps are applied.

XML files are less complicated; require less storage space than relational database. It is more readable, structured than text format. When two different log files converted into one unified XML file, after that preprocessing steps come.

S.F.Yusufovna [7] focuses on several data mining techniques that can aid in the process of intrusion detection. In this paper provide information about different data mining techniques are feature selection, machine learning, classification techniques, clustering techniques, statical techniques. This data mining techniques have been proposed enhancement of intrusion detection.

VII. CONCLUSION

This paper gives detailed look about web log file, its format, and its location. These web log files records information of each user request and analyze web server log files to detect web attacks. Advantages of log files that data is easily available to be analyzed. Log files usually contain noisy and ambiguous data and preprocessing involves removal of unnecessary data. Preprocessing web log file is used in data mining techniques, also used in intrusion detection system as input to detect intrusion.

VIII. REFERENCES

- [1] V. Chitraa, Dr. A.S. Davamani. "A Survey on Preprocessing Methods for Web Usage Data", IJCSIS, 2010
- [2] L.K. Joshila Grace, V.Maheswari, Dhinaharan Nagamalai. "Analysis of web logs and web user in web Mining", IJNSA, 2011
- [3] K.R. Suneetha, Dr. R. Krihnamoorthi. "Identifying User Behavior by Analyzing Web Server Access Log File", IJCSNS, 2009
- [4] G. Castellano, A. M. Fanelli, M. A. Torsello. "Log Data Preparation for Mining Web Usage Patterns", IADIS International Conference Applied Computing, 2007
- [5] Li. Chaofeng. "Research and Development of Data Preprocessing in Web Usage Mining", International Conference on Management Science and Engineering, 2006
- [6] Shaimaa Ezzat Salama, Mohamed I. Marie , Laila M. El-Fangary & Yehia K. Helmy "web server logs for preprocessing for web intrusion detection", Published by Canadian Center of Science and Education, 2011.
- [7] S. F. Yusufovna "Integrating Intrusion Detection System and Data Mining", International Symposium on Ubiquitous Multimedia Computing, 2008

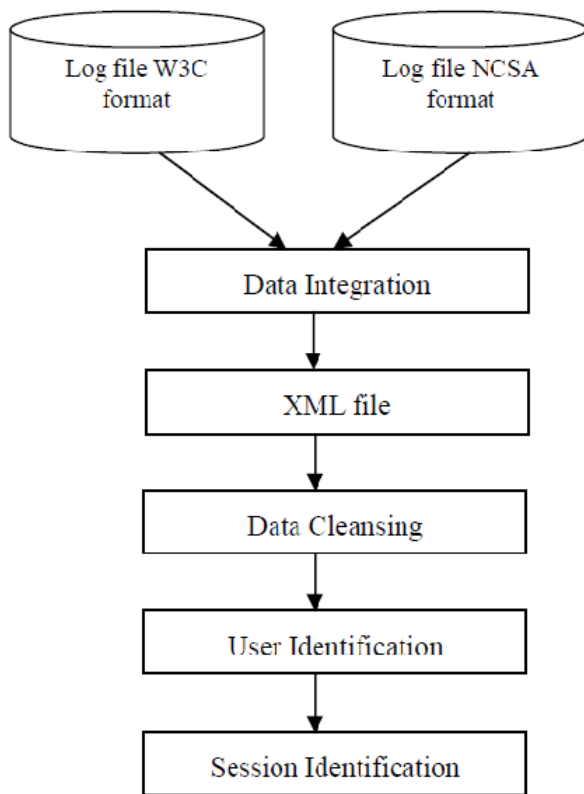


Figure 4. Log files preprocessing [6]