Capturing Web Log and Performing Preprocessing of the User's Accessing Distance Education System

Mr. Shivkumar Khosla[1], Mrs. Varunakshi Bhojane[2]
[1, 2] (Department of Computer Engineering, Mumbai University, India)

ABSTRACT: *In this paper, we have introduced a concept of capturing different web log file, while the user is accessing the Distance Education System website. Web log file can be further used in pattern discovery and pattern analysis process. Web log file is saved in text (.txt) format with "comma" separated attributes. Log files can't be directly used for pattern discovery process because it consists of irrelevant and inconsistent access information. Therefore there is need of Web log preprocessing which includes different techniques such as field extraction, data cleaning, data filtering, and data summarization. We have discussed different types of web log files and preprocessing techniques.*

Keywords: *Data Preprocessing, Web Log Mining, Web Log File, Web Personalization, and Web Usage Mining.*

I. Introduction

With the rapid development of World Wide Web (WWW), Web application is increasing at enormous speed and its users are increasing at exponential speed. Distance Education System has been widely used in the field of education. Students try to learn from web so, there is a need to provide students with the personalize teaching in Distance Education System which help them to understand the concept in much more better way than the traditional distance education system.

Web Log mining provides a log or access information of the users accessing the Distance Education System. Web Log Mining is an effective and mostly used technique of Web Usage Mining. Most Application performs Web Log Mining to improve web personalization, future prediction and also increasing performance in terms of response from the Web server.

1. Web Log Content

There are three main sources to get the web log file [Fig 1.] such as [2]

1. Client Log File
2. Proxy Log File
3. Server Log File

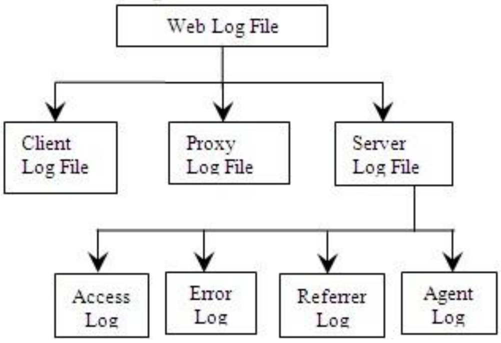

Fig.1 Types of Web log File

Client Log Files mostly consist of authentic i.e. username and password but it does not consist of any of the browser information.

Proxy Log File used to capture the user access data i.e. it capture the pages that are being accessed by the user. Proxy server is in many-manycardinality since there are many users accessing many pages.

Server Log Files are in relationship of many to one since there is only one web server response to many users. Server log file do not record the cashed pages requested.
Different types of Server Log File include:

a. Referrer Log
b. Error Log
c. Agent Log
d. Access Log

Referrer Log file contains information about the pages that is being referrer. **Error LogFile** records the errors of web site especially page not found error (404 File not found).

Agent Log File records the information about the website user's browser, browser version & Operating System.

Access Log file records all the click, hits and accesses made by the user to the website.

Web log File describes about different types of Log that can be captured. The Log captured should be stored in a specific format separated by comma, so that it can further use for processing. The Format which is used in the paper is Common Web log format [Table1] in a customized manner [2].

Table1 Attributes of log file and description

Attributes	**Description**
ClientIP	Client Machine IP address
Time	Time of Transaction
Date	Date When user made access
Server Site name	Internet Service name
Server Computer name	Server Name
Server Port	Server port configured for data transmission
Server Client Status	Status code returned by server 200, 404
User Agent	Browser Types that client used
Referrer	Link from where Client jump to site
Client Server URI Stem	Targeted Default Web page
URI Query	Client Query Which start "?"
Client Server Host	Host Header Name

$ClientIP ,$time, $scriptfilename, $serverport, $servername, $referrer, $statuscode, $uri, $useragent, $querystring, $host, $documentroot

Table2 Example

127.0.0.1, 2012-08-8 14:18:11, des/index.php, 80, local host,http://localhost/des/index.php?subj=1, 200, Mozilla/5.0 (X11; Linux x86_64) Chrome,subj=1, localhost, /var/www.

While we are capturing the log each attributes is separated by comma which make it easier retrieve of the attributes during preprocessing and the complete attributes are stored in the text file (.txt) which also allow for fastest retrieve of information from the server.

II. Data Preprocessing

Preprocessing converts the captured log data in to the valuable information which can be given for further pattern discovery. In this phase main steps includes are [1]:

1. Extraction the attributes from the web log which is located in web server
2. Cleaning the web logs and removing the redundant and irrelevant information.
3. Manage the data and put it in relational database or data warehouse.

2.1 Field Extraction

In Field Extraction we extract the attributes from the log file which is separated by character ‘,’. The attributes that are extracted is stored in the log table which is relational database. Process flow is shown in Fig.2.

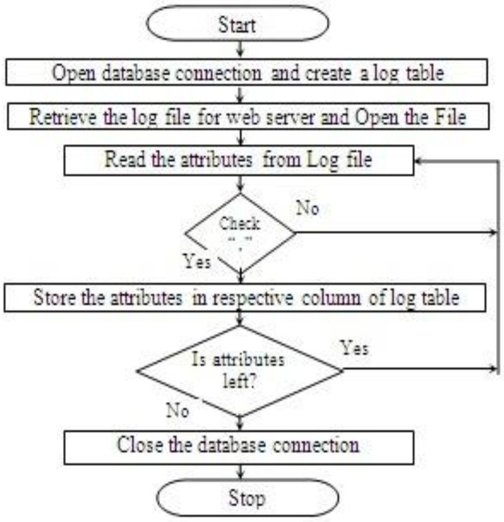
Start
Open database connection and create a log table
Retrieve the log file for web server and Open the File
Read the attributes from Log file
Check ","
No
Yes
Store the attributes in respective column of log table
Is attributes left?
Yes
No
Close the database connection
Stop

Fig.2 Process Flow for Field Extraction

2.2 Data Cleaning and Data Filtering

Data cleaning eliminates irrelevant or unnecessary records stored in log table. Since website will be accessed by millions of users. Data Cleaning is used to remove the records while analyzing the data. It removes all the jpg, gif and css files which the user has accessed & also the failed HTTP status code (404 Page Not Found). Data Filtering provides with the more better cleaning by removing the repeated pages that the user have accessed in the referrer string and also filtering help to reduce the path’s accesses of the page by splitting the Referrer String and getting valuable information.

By performing Data Cleaning and Filtering errors files, inconsistent data, missing data, repeatable data will be detected and removed to improve the quality of data. Process flow is shown in Fig.3.

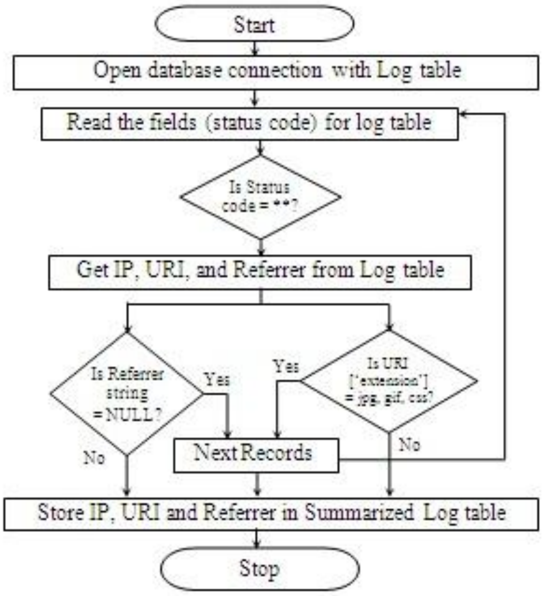
Start
Open database connection with Log table
Read the fields (status code) for log table
Is Status code = **?
Get IP, URI, and Referrer from Log table
Is Referrer string =NULL?
Yes
Yes
Is URI ['extension'] = jpg, gif, css?
No
Next Records
No
Store IP, URI and Referrer in Summarized Log table
Stop

Fig.3 Process Flow of Data Cleaning and Filtering

2.3 User Identification

User Identification is based identify the user with the User Id which help to predict the user behavior uniquely i.e. pages the user have accesses along with the IP address of the user through which user was accessing the sites.

User Id of the user also helps in Session Identification i.e. if the session is set then the user is registered user and can start recording the pages the user accessing and also provide the user with dynamism of the pages but if session is not set the user is allow to access the pages but do not record any of the log information of the user which in turn reduces the log information of the non-registered user.

2.4 Data Summarization

In this paper data summarization is describe by the graphical representation of the log that are being captured. Since, useful and consistent records remain in the Summarized log table which helps in further pattern discovery process.

Summarized log table when compared with the log table provides information such as Total no. of records, Total no. of registered User, Total no. of URL, Visited count of the registered user.

III. Results

3.1 Comparison

We have captured a log for a month ’19-07-2012’ to ’22-08-2012’ of distance education system and depending upon the log captured we had following analysis. [Table3] shows the comparison between log table and summarized log table. [Table4] shows the Number of user that have uniquely access the distance education system. [Table5] shows summarized statistical report. The Following Figures are showing total no. of records in Log table and summarized log table [Fig. 4], Number of access of the unique users [Fig. 5], File Size in (KB) [Fig. 6].

Table3 Comparison between Log Table (LT) and Summarized Log Table (SLT)

Attributes	**LT**	**SLT**
File Size (KB)	2154	519
Number of Records	343254	12519

Table4 Number of Access of the Unique Users

Client IP address	**No. of Access**
172.16.11.101	34
172.16.11.102	40
172.16.11.103	19
172.16.11.104	26
172.16.11.105	53
172.16.11.106	13
…	…
…	…

Table5 Summary Statistical Report

Status	**Records**
Errors	1289
Css	868
Jpeg	2548
Hits	1524785

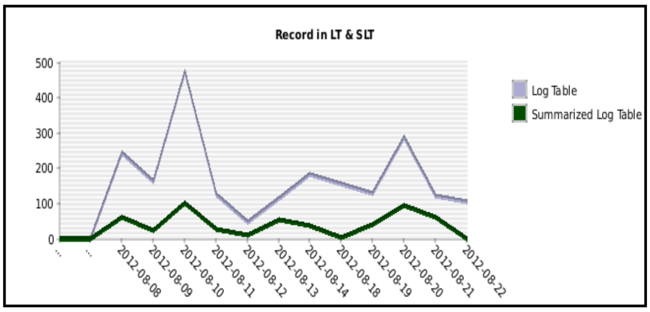

Fig.4 Total no. of records in Log table and Summarized log table (Date-wise)

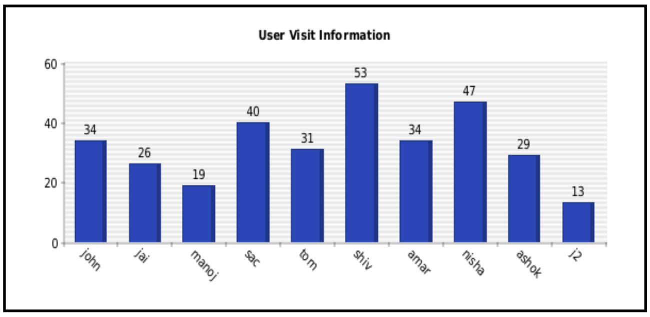

Fig.5 No. of accesses of unique users

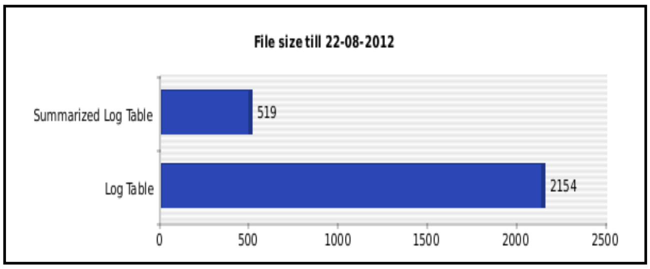

Fig.6 File size in (KB)

3.2 Analysis

From the Results after capturing the web log and representing it in graphical manner, we can get the conclusion as log file size is reduced to ¼ of the actual log captured and record in the file are also correspondingly reduced.

IV. Conclusion

The information after Data Preprocessing can be given to pattern discovery process which includes different data mining techniques such as clustering, classification and association rules and since irreverent information is removed therefore it speeds up the execution time and provides with valuable information to the users.

It also allows the user to have different access pattern due to which better accessibility is achieved. We can also record the information in the cookies and when the user ends its session we can then transfer the complete information from the cookies to the database.

REFERENCES

[1] TheintT heintAye,Web log Cleaning for Mining of Web Usage Patterns, *IEEE 978-1- 61284-840-2/11*

[2] Tasawar Hussian, Dr. Sohail Asghar, Dr, Hayyer Masood, Web Usage Mining: A Survey on Preprocessing of Web log File, *International Conference 978-1-4244080306/10*

[3] Xinjin Li, Sujing Zhang, Application of Web usage mining in e-learning platform, *2010 International Conference on E-Business and E-Government*

[4] R. Cooley, B. Mobasher, and J. Srivastava, Data preparation for mining World Wide Web browsing patterns, *Knowledge and Information Systems, Vol. 1, No. 1*, 1999, pp. 5-32.

Proceedings Papers:

[5] Yuan, F., L.-J. Wang, et al. (2003). Study on Data Preprocessing Algorithm in Web Log Mining. Proceedings of the Second International Conference on Machine Learning and Cybernetics, Wan, 2-5 November 2003