WEB SERVER LOGS TO ANALYZING USER BEHAVIOR USING LOG ANALYZER TOOL

S.Padmaja [1], Dr.Ananthi Sheshasaayee [2]

[1] *Research Scholar, Research and Development Centre, Bharathiar University, Coimbatore, (India).*

[2] *Associate Professor and Head, Department of Computer Science,*

Quaid-e-Millath Government College for women, Chennai , (India).

ABSTRACT

Web usage mining performs mining on Web usage data, or Web logs. A Web log is a listing of page reference data. Sometimes it is referred to as clickstream data because each entry corresponds to a mouse click. These logs can be examined from either a client perspective or a server perspective.So in order to provide better service along with enhancing the quality of websites, it has become very important for the website owner to better understand their customers. This is done by mining web access log files to extract interesting patterns.Web Usage mining deals with understanding the user behavior. The user behavior can be analyzed with the help of Web Access Logs that are generated on the server while the user is accessing the website. A Web access log contains the various entries like the name of the user, his IP address, number of bytes transferred timestamp, URL etc. A different types of Log Analyzer tools exist which help in analyzing various things like users navigational pattern, the part of the website the users are mostly interested in etc. The present paper analyses the use of such log analyzer tool called Web Log Expert for ascertaining the behavior of users with the help of sample data.

Keywords: Log Files, User Behavior, Pattern Recognition, Log Analyzer Tools,User Behavior

I INTRODUCTION

Prediction is the data mining technique that is used to predict missing or unavailable data. In a way, classification that is used to predict class labels can be treated as prediction when numerical data are predicted. Prediction differs from classification by the fact that is used only for numerical data prediction as opposed to classification that predicts class labels. The goal of data mining is to produce new knowledge that the user can act upon. It does this by building a model of the real world based on data collected from a variety of sources which may include corporate transactions, customer histories and demographic information, process control data and relevant external databases such as credit bureau information or weather data. The results of the model building are a description of patterns and relationships in the data that can be confidently used for prediction.

II PATTERN MATCHING

Pattern matching or pattern recognition finds occurrences of a predefined pattern in data. Pattern matching is used in many diverse applications. A text editor uses pattern matching to find occurrences of a string in the text being edited. Information retrieval and Web search engines may use pattern matching to find documents containing a predefined pattern (perhaps a keyword). Time series analysis examines the patterns of behaviour in data obtained from two different time series to determine similarity. Pattern matching can be viewed as a type of classification where the predefined patterns are the classes under consideration. The data are then placed in the correct class based on a similarity between the data and the classes.

III CONTENTS OF A WEB LOG FILE

A file produced by a Web server to record activities on the Web server. It usually has the following features:

- The log file is text file. Its records are identical in format.
- Each record in the log file represents a single HTTP request.
- A log file record contains important information about a request: the client side host name or IP address, the date and time of the request, the requested file name, the HTTP response status and size, the referring URL, and the browser information.
- A browser may fire multiple HTTP requests to Web server to display a single Web page. This is because a Web page not only needs the main HTML document; it may also need additional files, like images and JavaScript files. The main HTML document and additional files all require HTTP requests.
- Each Web server has its own log file format.
- If your Web site is hosted by an ISP (Internet Service Provider), they may not keep the log files for you, because log files can be very huge if the site is very busy. Instead, they only give you statistics reports generated from the logs files.

Nearly all of the major Web servers use a common format for their log files. These log files contain information such as the IP address of the remote host, the document that was requested, and a timestamp. The syntax for each line of a log file is:

```
site logName fullName [date:time GMToffset] "req file proto" status length
```

Because that line of syntax is relatively meaningless, here is a line from a real log file:

```
204.31.113.138 - - [03/Jul/1996:06:56:12 -0800]  "GET /PowerBuilder/Compny3.htm HTTP/1.0" 200 5593
```

Even though the line is split into two, here, you need to remember that inside the log file it really is only one line.

Each of the eleven items listed in the above syntax and example are described in the following list.

- **Site**-either an IP address or the symbolic name of the site making the HTTP request. In the example line the remotehost is 204.31.113.138.
- **LogName**-login name of the user who owns the account that is making the HTTP request. Most remote sites don't give out this information for security reasons. If this field is disabled by the host, you see a dash (-) instead of the login name.
- **Full Name**-full name of the user who owns the account that is making the HTTP request. Most remote sites don't give out this information for security reasons. If this field is disabled by the host, you see a dash (-) instead of the full name. If your server requires a user id in order to fulfil an HTTP request, the user id will be placed in this field.
- **Date-date** of the HTTP request. In the example line the date is 03/Jul/1996.

- **Time-time** of the HTTP request. The time will be presented in 24-hour format. In the example line the time is 06:56:12.
- **GMToffset**-signed offset from Greenwich Mean Time. GMT is the international time reference. In the example line the offset is -0800, eight hours earlier than GMT.
- **Req-HTTP** command. For WWW page requests, this field will always start with the GET command. In the example line the request is GET.
- **File-path** and filename of the requested file. In the example line the file is /PowerBuilder/Compny3.htm. There are three types of path/filename combinations.

IV TYPES OF LOG ANALYZER TOOLS

There are a lot of Web log analysis tools out there, and many are free. This is a list of some of the best free log analysis and Web analytics tools available.

4.1Web Log Expert

WebLog Expert is a fast and powerful access log analyzer. It will give you information about your site's visitors: activity statistics, accessed files, paths through the site, information about referring pages, search engines, browsers, operating systems, and more. The program produces easy-to-read reports that include both text information (tables) and charts. View the WebLog Expert sample report to get the general idea of the variety of information about your site's usage it can provide.The log analyzer can create report in HTML, PDF and CSV formats. It also includes a web server that supportsdynamic HTML reports.

WebLog Expert can analyze logs of Apache and IIS web servers. It can even read GZ and ZIP compressed log files so you won't need to unpack them manually.The program features intuitive interface. Built-in wizards will help you quickly and easily create a profile for your site and analyze it.

4.2 Deep Log Analyzer

Deep Log Analyzer is the best free Web analytics software I've found. It is a local log analysis tool that works on your site logs without requiring any codes or bugs on your site. It's not as fancy as Google Analytics, but it does offer a few extra features. Plus, if you need more features, there is a paid version you can upgrade to. Advanced and affordable web analytics solution for small and medium size websites. We can analyze web site visitors' behavior and get complete website usage statistics in several easy steps. With our website statistics and web analytics software you will know exactly where your customers came from, how they moved through your site and where they left it. This comprehensiveknowledge will help you to attract more visitors to your site and convert them to satisfied customers.

4.3Google Analytics

Google Analytics is one of the best free Web log analysis tools available. There are a few reports that are not included, but the graphs and well-defined reports make it very nice. Some people don't like giving a large corporation like Google such direct access to their site metrics. And other people don't like needing a bug placed on the Web pages in order to track them.

4.4AWStats

AWStats is a featureful tool that generates advanced web, streaming, ftp or mail server statistics, graphically. This log analyzer works as a CGI or from command line and shows you all possible information your log contains, in few graphical web pages. It uses a partial information file to be able to process large log files, often and quickly. It can analyze log files from all major server tools like Apache log files (NCSA combined/XLF/ELF log format or common/CLF log format), WebStar, IIS (W3C log format) and a lot of other web, proxy, wap, streaming servers,mail servers and some ftp servers.AWStats is a free software distributed under theGNU General Public License. You can have a look at this license chart to know what you can/can't do.As AWStats works from the command line but also as a CGI, it can work with all web hosting providers which allow Perl, CGI and log access.

4.5 W3Perl

W3Perl is a CGI based free Web analytics tool. It offers the ability to use a page bug to track page data without looking at log files or the ability to read the log files and report across them.

4.6 Power Phlogger

Power Phlogger is a free Web analytics tool that you can offer to other users on your site. This tool uses PHP to track information. But it can be slow.Powerphlogger is a complete counter hosting tool. It lets you offers counter service to others from your site. It’s built on PHP and requires a MySQL server. Your members don’t need any PHP support on their webserver. They just pass the required data through JavaScript to PPhlogger that is hosted on the server.

4.7 BBClone

BBClone is a PHP based Web analytics tool or Web counter for your Web page. It provides information about the last visitors to your site tracking things like: IP address, OS, browser, referring URL and more.

4.8 Visitors

Visitors is a command line free log analysis tool. It can generate both HTML and text reports by simply running the tool over your log file. One interesting feature is the real time streaming data you can set up.Visitors is a very fast web log analyzer for Linux, **Windows**, and other Unix-like operating systems. It takes as input a web server log file, and outputs statistics in form of different reports. The design principles are very different compared to other **software** of the same type.

4.9 Webalizer

Webalizer is a nice little free Web log analysis tool that is easily ported to many different systems. It comes with several different languages for reports and a bunch of stats to report on. ***The Webalizer*** is a fast, free **web** server log file analysis program. It produces highly detailed, easily configurable usage reports in HTML format, for viewing with a standard web browser.

4.10 Analog

Analog is a widely used free Web log analysis tool. It works on any Web server and is fairly easy to install and run if you understand how your server is administered. It has a lot of good reports and with another cool can be made even prettier.

4.11RealTrackerPersonal

RealTracker uses a code that is placed on your Web pages to track your pages, similar to Google Analytics. It offers a bunch of different reports but the real benefit to this tool is that it's easy to add toyour pages andeasy to read the results. And if you need more features, you can upgrade to the professional or enterprise versions.

4.12 Webtrax

Webtrax is a free Web analytics tool that is very customizable, but not as well programmed as it could be. The author admits that there are some issues, and it doesn't appear to be under active support at this time. But it does support a number of reports and proides good information from your log files.Webtrax is a log file analysis program for NCSA web server logs. It works best on logs that include the "referrer"and "browser"info, such as the "NCSA Combined Format." Webtrax reads a web server's log file and produces up to twenty different graphical and tabular reports of hits, and the activities of individual site visitors, including what pages they viewed and for how long.Webtrax's output is extremely customizable.

4.13 Dailystats

Dailystats is a free Web analysis program that is not intended to be your complete analytics package. Instead, Dailystats wants to give you a small sub-set of statistics that are useful for reviewing on a regular basis - such as daily. It provides information on entry pages, pageviews of each page, and referrer log analysis.

4.14 Relax

Relax is a free Web analytics tool that tells you just who is referring people to your site. It looks at search engines and search key words as well as specific referral URLs to give you precise information on who is sending customers to your site. It's not a complete analytics package, but it works well for referral information.

4.15 Piwik

Piwik is an open source alternative to Google Analytics. It is very flashy with an Ajax or Web 2.0 feel to it. One of the nicest features is that you can build your own widgets to track whatever data you want to track. It runs on your PHP Web server, and requires that you have PHP PDO installed already. But if you have that it's fairly easy to install and get up and running.

4.16 StatCounter

StatCounter is a Web analytics tool that uses a small script that you place on each page. It can also work as a counter and display the count right on your page. The free version only counts the last 100 visitors, then it resets and starts the count over again. But within that limitation, it provides a lot of stats and reports.

4.17 SiteMeter

The free version of SiteMeter offers a lot of good stats and reports for your site. It only provides information on the first 100 visitors, and then after that it resets and starts over. But if you need more information than that, you can upgrade to the paid version of SiteMeter. Like other non-hosted analytics tools, SiteMeter works by inserting a script on every page of your site. This gives you real-time traffic but some people are concerned about privacy implications.

4.18 MyBlogLog

MyBlogLog is a tool with many different features. The analytics are not very robust, but they are not intended to be. In fact, the goal of the MyBlogLog analytics is to provide you with information about where your visitors are going when they leave your site. This can help you to improve your site so they don't leave as quickly. I wouldn't recommend MyBlogLog as your only analytics tool, but it does a good job on the stats it provides.

4.19 WebLog Expert Lite

WebLog Expert Lite is a free Apache and IIS log analyzer, light-weight version of WebLog Expert. It allows you to quickly and easily analyze your log files and get information about your site's visitors: activity statistics, what files visitors accessed information about referring pages, search engines, browsers, operating systems, and more.

V TYPES OF LOG FILES

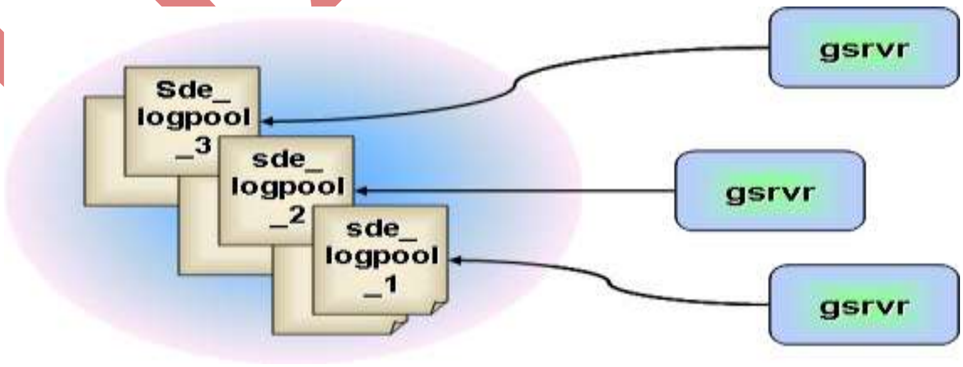
Traditionally there are four types of server logs: Transfer Log, Agent Log, Error Log and Referrer Log [6]. The Transfer and the Agent Log are said to be standard whereas the error and referrer log are considered optional as they may not be turned on. Every log entry records the traversal from one page to another, storing user IP number and all the related information. There are three types of log files:

5.1 Shared Log Files

For SQL Serverdatabases, the defaults are session log files created in tempdb. Each user owns two tables: SDE_logfiles and SDE_logfile_data. Shared log files are created the first time a user's selection exceeds the required threshold (100 features in ArcGIS). If you use shared log files, remember:

- Users require CREATE TABLE permission.
- If permissions are insufficient, selections cannot be made.
- Log files are not checked on connection.
- 5.2. Session log files

Session log files are dedicated to a single connection, not a database user. You can arrange for a pool of session log files to be shared among users, and gsrvrs can also create their own session log files on the fly. As mentioned in the previous section, session log files are the default for SQL Server databases.



Using session log files dramatically reduces contention for log files, since a single connection is using the log file. Similarly, since only one connection is using the log file, the table will not grow as large as it would when dozens or hundreds of connections are using the same log file.

And finally, some delete optimizations can be made in special circumstances. If only one log file is open, the log file table can be truncated rather than deleted. Truncating a table is orders of magnitude faster than deleting.

5.2 Stand-alone log files

Stand-alone log files are useful in several situations. The advantage of this log file configuration is that it can always be truncated and will never grow beyond the size of a single selection set. Of course, there has to be a user configurable limit on how many of these stand-alone log files can be created by a connection. For example, if 100 users are creating selection sets on 50-layer maps, an unrestricted growth of stand-alone log files would result in 5000 log file tables being created. To avoid this, set the MAXSTANDALONELOGFILE parameter appropriately.

Log Types based on network traffic

Log Type	Description
Traffic	The traffic logs records all traffic to and through the FortiGate interface. Different categories monitor different kinds of traffic, whether it be external, internal, or multicast.
Event	The event logs record management and activity events within the device in particular areas: System, Router, VPN, User, WAN, and WiFi. For example, when an administrator logs in or logs out of the web-based manager, it is logged both in System and in User events.
Antivirus	The antivirus log records virus incidents in Web, FTP, and email traffic.
Web Filter	The web filter log records HTTP Fort iGATE log rating errors including web content blocking actions that the Fort iGATE unit performs.
Intrusion	The intrusion log records attacks that are detected and prevented by the FortiGate unit.
Email Filter	The email filter log records blocking of email address patterns and content in SMTP, IMAP, and POP3 traffic.
Vulnerability Scan	The Vulnerability Scan (Netscan) log records vulnerabilities found during the scanning of the network.
Data Leak Prevention	The Data Leak Prevention log records log data that is considered sensitive and that should not be made public. This log also records data that a company does not want entering their network.
VoIP	The VoIP log records VoIP traffic and messages. It only appears if VoIP is enabled on the Administrator Settings page.

VI RESULTS AND INTERPRETATIONS

In this study, we have analyzed thesample log files of Web server of with the help of weblog Expert analyzer tool. The sample log files consists the data fora month. In thisduration log files have stored 200MB data and we have got 26.4 MB data after preprocessing. We have determined different types of errors that occurred in web surfing. Statistics about hits, page views, visitors and bandwidth are shown below.Log analyzer tools are required as they help extensively in analyzing the information about visitors, top errors which can be utilized by system administrator and web designer to increase the effectiveness of the web site.

General Statistics: In this section we get general information pertaining to the website like how many times the website was hit, an average of hits in a day, page views, bandwidth etc. It enlists all the general information which one should know related to a website.

Summary of Result

Hits

Total Hits	1,891,700
Visitor Hits	1,891,700
Spider Hits	0
Average Hits per Day	67,560
Average Hits per Visitor	11.62

Requests

Cached Requests	132,627
Failed Requests	10,966

Views

Total Page Views	613,115
Average Page Views per Day	21,896
Average Page Views per Visitor	3.77

Visitors

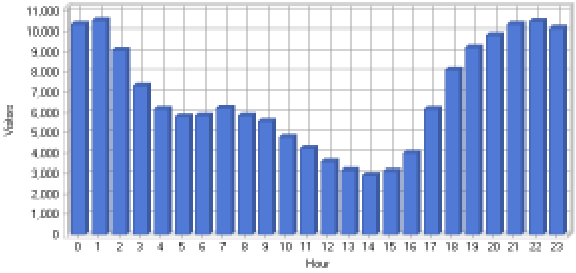
Total Visitors	162,754
Average Visitors per Day	5,812

Bandwidth

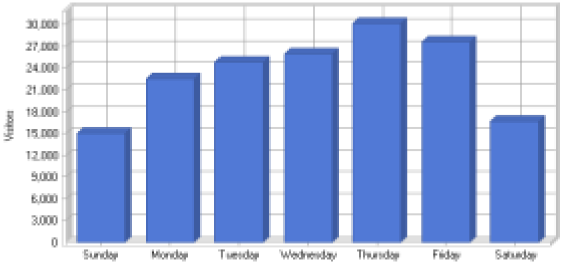
Total Bandwidth	36.04GB
Visitor Bandwidth	36.04GB
Spider Bandwidth	0GB
Average Bandwidth per Day	1.29GB
Average Bandwidth per Hit	19.98GB
Average Bandwidth per Visitor	232.18GB

Activity Statistics

Activity by Hour of Day

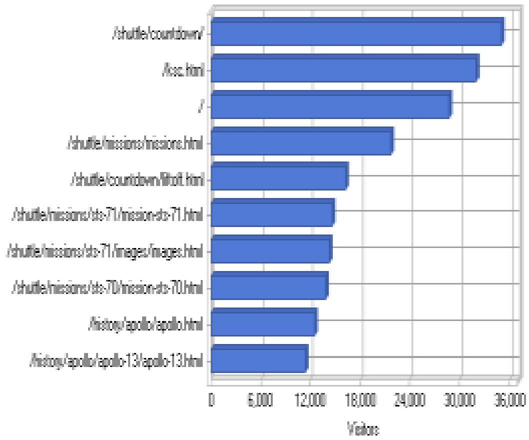


Activity by Day of Week

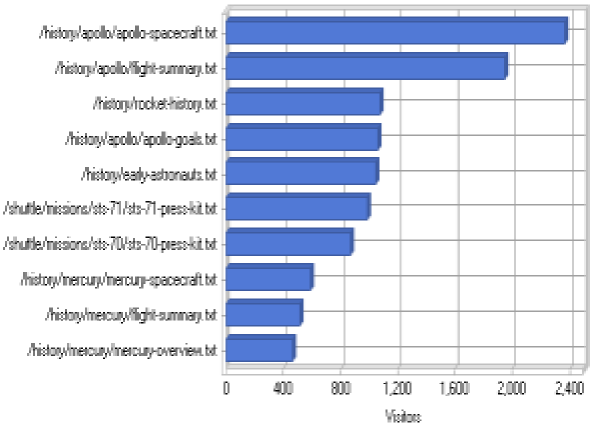


Access statistics

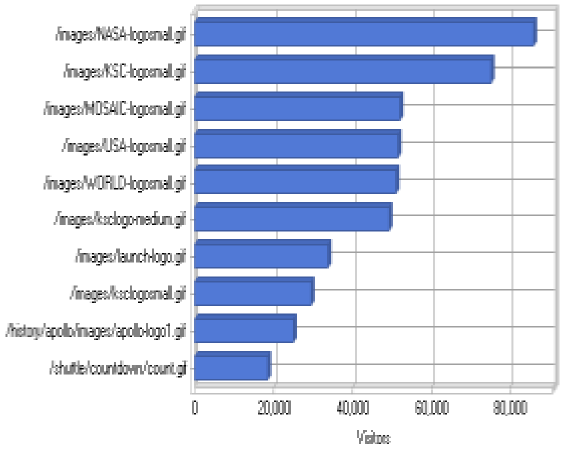
Most Popular Pages



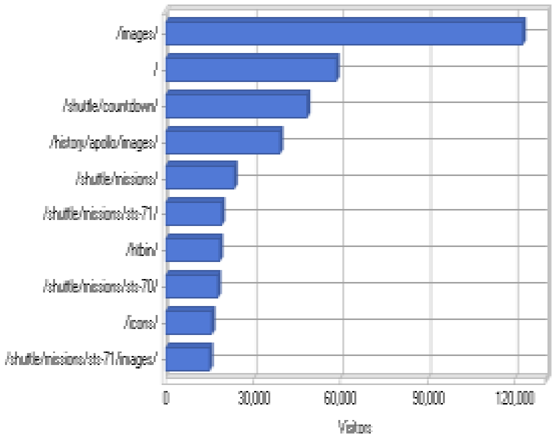
Most Downloaded Files



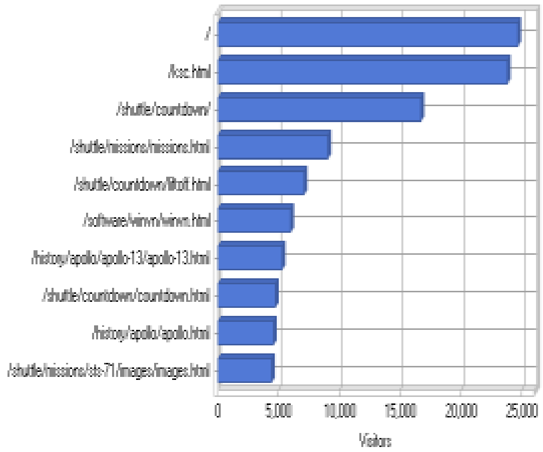
Most Requested Images



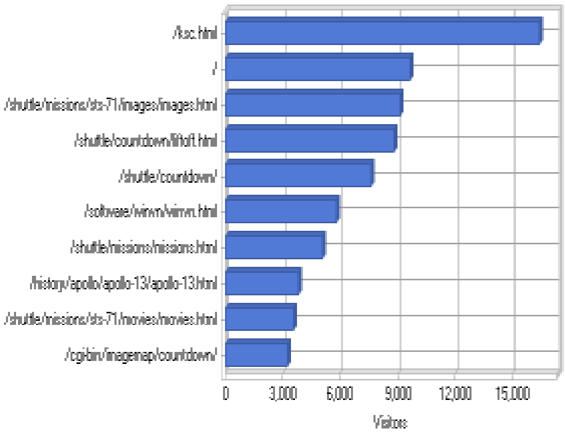
Most Requested Directories



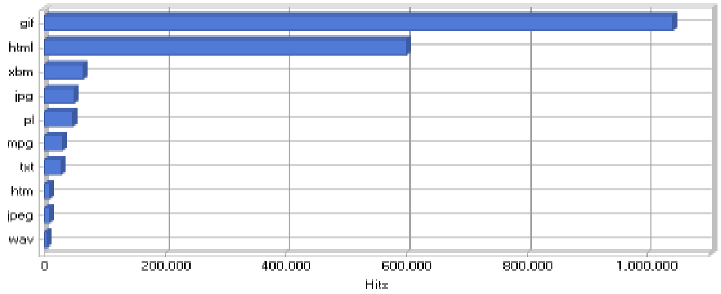
Top Entry Pages



Top Exit Pages

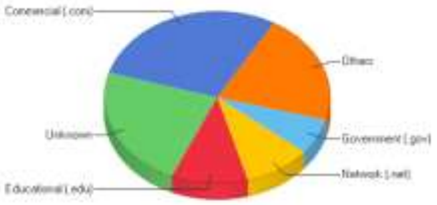


Most Requested File Types

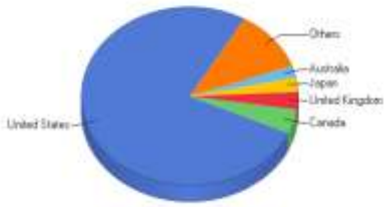


Visitors

Top-Level Domains



Most Active Countries



Referrers

Top Referring SitesSites

No Referrer	162,754
Total	162,754

Referring URLs

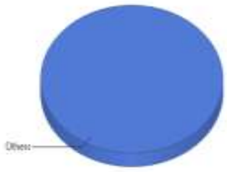
Top Referring URLs

URLVisitors

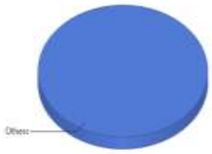
No Referrer	162,754
Total	162,754

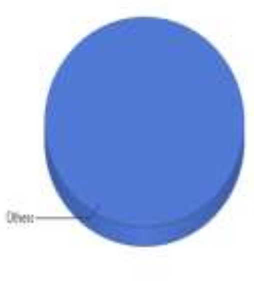
Browsers

Most Used Browsers

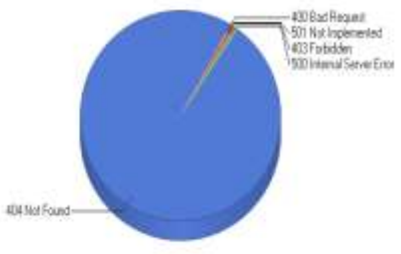


Most Used Operating Systems





Device Types



Error Types

VII CONCLUSION

An important research area in Web mining is Web usage mining which focuses on the discovery of interesting patterns in the browsing and navigation data of Web users. In order to make a website popular among its visitors, System administrator and web designer should try to increase its effectiveness because web pages are one of the most important advertisement tools in international market for business. The obtained results of the study can be used by system administrator or web designer and can arrange their system by determining occurred system errors, corrupted and broken links. In this study, analysis of web server log files of Web Log Expert Log Analyzer tool.One important use of patterns is to summarize data, since the pattern collections together with the quality values of the patterns can be considered as summaries of the data. This paper presents an overview of Log files, Content of a Log files and variety of log files etc. Web Log Analyzer tools are a part of Web Analytics Software.They take a log file as an input, analyze it and generate results. Web Log Expert was taken to analyze the web logs of the website as it provided extensive features that too in the free edition. The results were examined and are being tried to incorporate in the website of the user. Such log analyzer tools should be widely used as they help a lot in understanding the customer behavior to the analysts.

REFERENCES

1] “Mining the Web- Discovering knowledge from Hypertext data” by SoumenChakrabarti, Morgan Kaufmann Publishers.

2] http://www.herongyang.com/Windows/Web-Log-File-IIS-Apache-Sample.html.

3]“Identifying User Behavior by Analyzing Web Server Access Log File” byK. R. Suneetha, Dr. R. Krishnamoorthi, International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009pp 327-333.

4] “An Overview of Preprocessing on Web Log Data for Web Usage Analysis”by Naga Lakshmi, Raja SekharaRao , Sai Satyanarayana ReddyInternational Journal of Innovative Technology and Exploring Engineering ISSN: 2278-3075, Volume-2, Issue-4, March 2013. Pp 274-279

5]“Web Usage Mining: A Survey on Pattern Extraction from Web Logs” Ankita KusmakarSadhna Mishra,International Journal of Advanced Research in Computer Science and Software Engineering Volume 3, Issue 9, September 2013 ISSN: 2277 128X pp834-838.

6] “Introduction to Data Mining with Case Studies: Web Data Mining” by G.K. Gupta, PHI Learning Private Limited, pp. 231-233, 2011.

7] “Extraction of Frequent Patterns from Web Logs using Web Log Mining Techniques” byKanwal Garg, PhD. Rakesh Kumar, and Vinod Kumar,International Journal of Computer Applications (0975 – 8887) Volume 59– No.10, December 2012pp 19-26.

8] http://www.google.com/analytics/.

9]“Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool” by NehaGoel and C.K. Jha, PhD, *International Journal of Computer Applications (0975 – 8887) Volume 62– No.2, January 2013*Pp29-34.

10] “Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data” By Bing Liu, Springer publications.

11] “Identification of Human Behavior using Analysis of Web log Data Mining”, by Dr. PunitGoyalIPASJ International Journal of Information Technology, Volume 1, Issue 1, June 2013 ISSN 2321-5976pp 1-7.

12] http://www.weblogexpert.com/.

13] https://www.google.co.in/#q=free+web+log+analyzer.