PREPROCESSING OF WEB USAGE DATA FOR LOG ANALYSIS

Bhawesh Kumar Thakur, Syed Qmar Abbas, Mohd Rizwan Beg, Sheenu Rizvi

Abstract— Frequent interaction of users with the World Wide Web to access wide ranges of information and available services, it become necessary to handle web usage data in an efficient manner.
Web usage mining incorporate data generated at both client side as well as server side. But generally web log preprocessing are meant for the data available at server that is actually generated by the clients when they made any request to the server. The knowledge retrieved from preprocessing can be used to make a comfortable environment to the frequently visiting customer.
Preprocessing of web log data is one of the important activities that are required to make web log data in proper format and then can be used for further knowledge discovery process
This paper comprehensively addresses the activities involved to preprocess the web log data stored at server side. Preprocessed data are then useful to generate intelligent knowledge to personalize the web user's environment.

Index Terms— web mining; web usage mining; log data; data preprocessing; personalization; data abstractions; log analysis.

1. INTRODUCTION

The main aim of Web usage data processing is to extract the knowledge kept in the web log files of a Web server. By using statistical and data mining approach to the Web log files, useful guide about the users' navigational behavior can be recognized, such as user and transaction clusters, as well as possible correlations between Web pages and user clusters [3]. Web usage mining is used to improve the design of web site, personalization of contents, etc. User's activities can also be stored into a log file. There are several types of log: Server log, Proxy server log, Client/Browser log. These log files are used by web usage mining to analyze and discover useful patterns. The process of web usage mining involves three interdependent steps:

1. Data preprocessing,
2. Pattern discovery and
3. Pattern analysis.

Among these steps, Data preprocessing plays an important role because of nature of log data is unstructured, redundant and noisy. To improve later phases of web usage mining like Pattern discovery and Pattern analysis several data preprocessing techniques such as Data Cleaning, User Identification, Session Identification, Path Completion etc. have been used [4].

- *Bhawesh Kuamr Thakur is currently working toward the PhD degree in computer science at the Department Of Computer Science & Engineering, Integral University, Lucknow, India, bhawesh_k_thakur@yahoo.co.in*
- *Syed Qamar Abbas completed his Master of Science (MS) from BITS Pilani. His PhD was on computer-oriented study on Queueing models, qrat_abbas@yahoo.com*
- *Prof. Dr. M. Rizwan Beg is M.Tech & Ph.D in Computer Sc. & Engg., rizwanbeg@gmail.com*
- *Sheenu Rizvi is currently working toward the PhD degree in computer science at the Department Of Computer Science & Engineering, Integral University, Lucknow, India, sheenu_r@yahoo.com*

2. WEB MINING

Two different approaches were taken in initially defining web mining:

1. A "process-centric view," which define web mining as a sequence of tasks.
2. A "data-centric view," which define web mining in terms of the types of web data that was being used in the mining process.

The second approach has become more usable.
Web Mining is one of the data mining approaches that collect useful information automatically from the web documents and web services..

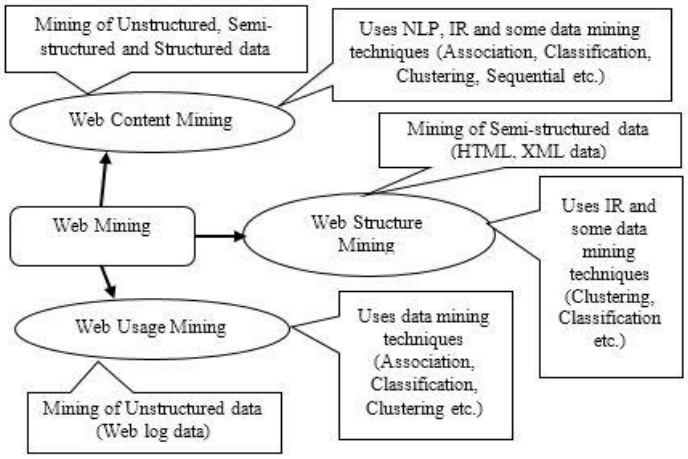


Figure 1: Classification of web mining

2.1 WEB USAGE DATA

With Web Usage Mining, data can be collected in server logs, browser logs, proxy logs, or obtained from an organization's database. These data collections differ in terms of the location of the data source, the kinds of data available, the segment of population from which the data was collected, and methods of implementation.

There are many kinds of data that can be used in Web Mining [1, 2, 10].

1. Content: The visible data in the Web pages or the information which was meant to be imparted to the users. A major part of it includes text and graphics (images).

2. Structure: Data which describes the organization of the website. It is divided into two types. Intra-page structure information includes the arrangement of various HTML or XML tags within a given page. The principal kinds of inter-page structure information are the hyper-links used for site navigation.

3. Usage: Data that describes the usage patterns of Web pages, such as IP addresses, page references, and the date and time of accesses and other information depending on the log format.

The results extracted using web usage mining can also be used in different purposes which are given as follows [6]:

a) Recommendations: It is an approach to evaluate user's past activities and current request to recommend user for selecting items or viewing some pages. It is heavily used for e-commerce web-sites to recommend some items and services to users.

b) Pre-fetching and Caching: The result of web usage mining can be used for improvement of web applications and web server performance. It means pre-fetching and caching of pages helps to improve server's response time.

c) Web-site design improvement: Easy accesses of web pages are important issues in designing of web-sites. Web usage mining provides user's behavioral feedback to improve design of web application.

d) Business intelligence: Collecting business intelligence from web usage data is vital for online E-commerce web-sites. Main issues with this are customer retention, cross sales, customer attraction and customer departure. tables and figures will be processed as images. You need to embed the images in the paper itself. Please don't send the images as separate files.

2.2 DATA SOURCES

The data sources used in Web Usage Mining may include web data repositories like:

1. Web Server Logs – These are logs which maintain information related to page requested by the user. The W3C uses a standard structure for web log files, but other formats also exist. More current entries are generally added to the end of the file. Data about the user request, including client IP address, request date/time, page requested, HTTP code, bytes served, user agent, and referrer are added. These data can be collected into a single file, or maintained into different logs, like an access log, error log, or referrer log. Web server log file do not store user specific information. So, these files are not accessible to general users. These are only for the purpose of the admin users. A statistical analysis of the server log may be performed to monitor traffic patterns by time of day, day of week, referrer, or user agent. Efficient web site administration, proper resources hosting and the refining the efforts can be improved by the web server log analysis.

2. Proxy Server Logs - A Web proxy is a caching mechanism which lies between client browsers and Web servers. It helps to reduce the load time of Web pages as well as the network traffic load at the server and client side. Proxy server logs contain the HTTP requests from multiple clients to multiple Web servers. This may serve as a data source to discover the usage pattern of a group of anonymous users, sharing a common proxy server.

3. Browser Logs – Various browsers like Mozilla, Internet Explorer etc. can be modified or various JavaScript and Java applets can be used to collect client side data. This implementation of client-side data collection requires user cooperation, either in enabling the functionality of the JavaScript and Java applets, or to voluntarily use the modified browser. Client-side collection scores over server-side collection because it reduces both the robot and session identification problems.

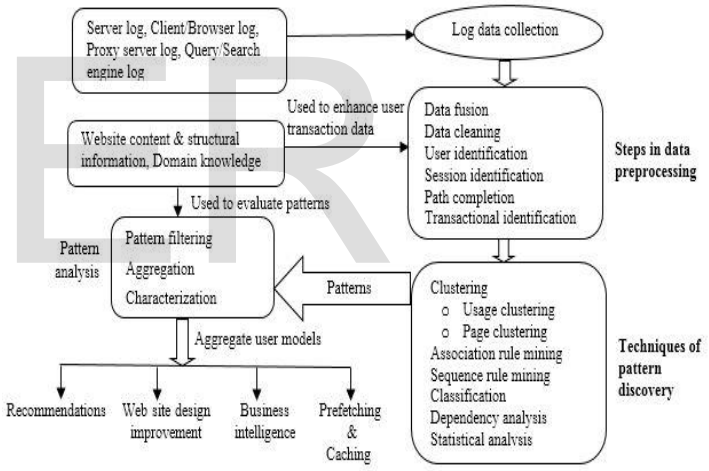


Figure 2: Web usage mining process

2.3 WEB SERVER LOG

All information related to web pages accesses are kept in the access log of the host Web server. The records of a web log file consist of several fields that follow a proper structure. The fields of the common log format are as follows:

Remote comp rfc123 authuser date "request" status bytes

Where, "remotecomp" is the remote computer name or IP address. "rfc123" is the remote log name. "authuser" is the username with which the user can be authenticated, it is available when using password-protected pages are there. "date" is the date and time of the web page request. "request" is the request came from the client i.e. the file, the name, and the method used to retrieve it. "status" is the HTTP status code returned to the client, indicating whether the file was successfully retrieved and if not, what error message was returned. "bytes" is the content-length of the documents transferred. If,

any of the fields cannot be identified a minus sign (-) is marked.

An enhanced format for Web server log files, called the "extended" log file format are also used, which is influenced by the requirement to support the collection of data for demographic analysis and for log summaries. This format allows customized log files to be stored in a readable format. The major addition to the common log format is that a number of fields are added to it. The most important are: referrer, it is the URL the client was visiting earlier requesting that URL, user agent, which is the software tool the client wants to be using, and cookie, in case of the site visited uses cookies.

```
...
looney.cs.umn.edu han - [09/Aug/1996:09:53:52 -0500] "GET /~mobasher/courses/cs5106/cs510611.html HTTP/1.0" 200 9370
mega.cs.umn.edu njain - [09/Aug/1996:09:53:52 -0500] "GET / HTTP/1.0" 200 3291
mega.cs.umn.edu njain - [09/Aug/1996:09:53:53 -0500] "GET /images/backgnds/paper.gif HTTP/1.0" 200 3014
mega.cs.umn.edu njain - [09/Aug/1996:09:53:53 -0500] "GET /images/misc/footer.jpg HTTP/1.0" 200 13355
mega.cs.umn.edu njain - [09/Aug/1996:09:54:12 -0500] "GET /cgi-bin/Count.cgi?df=CS-home.dat&dd=C&ft=1 HTTP/1.0" 200 646
mega.cs.umn.edu njain - [09/Aug/1996:09:54:18 -0500] "GET /~advisor HTTP/1.0" 302
mega.cs.umn.edu njain - [09/Aug/1996:09:54:19 -0500] "GET /~advisor/ HTTP/1.0" 200 487
looney.cs.umn.edu han - [09/Aug/1996:09:54:28 -0500] "GET /~mobasher/courses/cs5106/cs510612.html HTTP/1.0" 200 14072
mega.cs.umn.edu njain - [09/Aug/1996:09:54:31 -0500] "GET /~advisor/csci-faq.html HTTP/1.0" 200 13786
looney.cs.umn.edu han - [09/Aug/1996:09:54:47 -0500] "GET /~mobasher/courses/cs5106/princip.html HTTP/1.0" 200 6965
moose.cs.umn.edu mobasher - [09/Aug/1996:09:55:50 -0500] "GET /~suharyon/lisa.html HTTP/1.0" 200 654
moose.cs.umn.edu mobasher - [09/Aug/1996:09:55:53 -0500] "GET /~suharyon/line/line16.gif HTTP/1.0" 200 1423
moose.cs.umn.edu mobasher - [09/Aug/1996:09:55:57 -0500] "GET /~suharyon/joko1.jpg HTTP/1.0" 200 30890
...
```

Fig 3 Sample Information from access logs

2.4 WEB DATA ABSTRACTIONS

A Web site is considered as a collection of interconnected web pages which include a host page kept at the same network place. A user uses client program that interactively access and deliver resources. In the Web environment, a user is one who accessing files from a Web server, using a browser. A user session is defined as a enclosed set of user clicks across one or more Web servers. A server session is defined as a collection of user request to a single Web server during a user session. It is also called a visit. A pageview is defined as the visual interpretation of a Web page in a specific environment at a specific point in time i.e. a pageview consists of several items, such as frames, text, graphics, and scripts that construct a Web page. A clickstream is a sequential series of pageview requests, made from a single user.

Preprocessing of Web Usage Data

There are several key technical issues that must be considered during this phase in the context of the Web personalization process. It is required for Web log data to be cleaned and preprocessed in order to use them in the later phases of the process. The first task in the preprocessing phase is data preparation. Based on the application, Web log data may require to be cleaned from entries involving pages that returned an error or graphics file accesses. Except some cases such information, should be eliminated from a log file. Furthermore, crawler activity can be filtered out, because such entries do not give useful information. Another problem is related with caching. Accesses to cached pages are not kept in the Web log, so, such information is missed. Caching is mostly dependent on the client-side technologies used. In such cases, cached pages can usually be inferred using the referring information from the logs. Further, a useful phase is to perform pageview identification, i.e. determining which page file accesses contribute to a single pageview. Such a decision is application-oriented.

Most important of them is the user identification. There are several approaches to identify individual visitors. The most apparent solution is to assume that each IP address identifies a single visitor. This is not very proper because, for example, a visitor may access the Web from different computers, or if a proxy is used then many users may use the same IP address. An additional assumption can then be made, that consecutive page accesses from the same host during a certain time interval come from the same user. More accurate approaches for a user identification of unique visitors are the use of cookies or similar mechanisms or the requirement for user registration. But, a latent problem in using such methods might be the users are not interested to share personal information. After a user is identified, the next step is to perform session identification, by dividing the clickstream of each user into sessions. The typical solution is to set a minimum timeout and assume that consecutive requests within it belong to the same session, or set a maximum timeout, where two consecutive requests that go above it belong to different sessions.

In the WWW (World Wide Web) environment, a large amount of traffic can be generated between clients and web server. The traffic from a client to a server is a URL. The traffic from server to client is a named HTML file that will be interpreted and displayed on the client screen.

The web usage log probably is not in a format that is usable by mining applications. As with the data may need to be reformatted and cleaned. Steps that are part of the preprocessing phase include cleansing, user identification, session identification, path completion, and formatting.

- Let P be a set of literals, called pages or clicks, and U be a set of users. A Log is a set of triplets

{(u1,p1,t1)…(un,pn,tn)} where ui □U, pi □P, and ti is a time stamp.

Standard log data consist of the following:-Source site, destination site, and time stamp. A common technique for a server site to divide the log records into sessions.

A session is a set of page references from one source site during one logical period. Historically, a user logging into a computer, performing work, and then logging off the login would identify a session and log off represent the logical start and end of the session.

- **Data Cleaning**

To clean a server log to remove irrelevant items is very important for Web log analysis. The discovered information are only useful if the data represented in the server log gives a correct scenario of the user accesses to the Web site. The HTTP protocol maintains an individual connection for each file requested from the Web server.

- **User Identification**

Next, unique users must be identified. This is a very complex task because of the existence of caches, firewalls, and proxy servers. The Web Usage Mining methods that rely on user cooperation or by the automatic identification of web user are the ways to deal with this problem.

- **Session Identification**

For logs that cover long duration of time, it is most probable that users will visit the Web site multiple times. The objectives of session identification are to break the web page visits of each user into individual sessions. The simplest method for this is through a timeout, where if the time between page requests crossed a certain threshold, it is supposed that the user is starting a new session. Many commercial products use 15-30 minutes as a default timeout.

- **Path Completion**

Another problem in reliably identifying unique user sessions is determining if there are vital accesses that are not recorded in the web access log. This problem is referred to as path completion. Methods similar to those used for user identification can be used for path completion.

- **Formatting**

Once the appropriate pre-processing steps have been applied to the server log, a final preparation module can be used to properly format the sessions or transactions for the type of data mining to be accomplished.

Let P = (p1,p2…..pu) is the sequence of Web pages accessed from a certain IP between t1 and tu. Then , a user session v(t1,tf),tf tu, is defined as: v(t1,tf)=(p1,p2,….pf) : (Δt=tj-ti-1 δ, 1<j f) ∧ (f = u∨tf+1-tf>δ), where δ is a predefined time threshold.

There are many problems associated with the pre-processing activities, and most of these problems centre around the correct identification of the actual user. User identification is complicated by the use of proxy servers, client side caching, and corporate firewalls. Cookies can be used to assist in identifying a single user regardless of machine used to access the web.

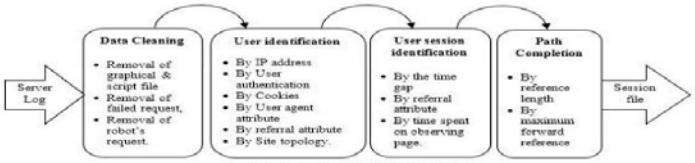


Figure 4: Web usage data preprocessing

Algorithm-1 for preprocessing of web log data

Step-1 (Definition-1)

A Log is a set of triplets {(u1,p1,t1)…(un,pn,tn)} where ui□ U (set of users), pi□P (set of pages/ clicks), and ti is a time stamp.

Step-2 Data cleaning

Elimination of the items (URLs) deemed irrelevant in data set. (By checking the suffix of the URL name)

Step-3 User Identification

Here, I assume user cooperation is necessary to identify unique user (unique IP address). System requires user registration for assigning unique-ID for each user

Step-4 Divide the web log data records into sessions (logical period, here 15 minutes) for generation of concept hierarchy. (Definition-2)

Let L be a Log. A session S is an ordered list of pages accessed by a user i.e. S={(p1,t1),(p2,t2),….(pn,tn)}, where there is a user ui □ U such that {(ui,p1,t1), (ui,p2,t2),⋯.(ui,,pn,tn)} □L.Here ti □ tjiff i<j.

Step-5 Formatting

Apply final preparation module to format properly the session file.

Step-6 OUT PUT (after applying above algorithm manually on test data set)

We get □ = {T1, T2, T3⋯Tm} // Set of transactions for unique user And U = {URL1, URL2, URL3…URLn} // set of unique URLs appearing in the preprocessed log.

2.5 LOG ANALYSIS

Log analysis take as input raw Web log data and process them to extract statistical information. Such information includes various statistics for the site activity such as total number of visits, average number of hits, successful/failed/redirected/cached hits, average view time, and average length of a path through a site, analytical statistics such as server errors, and page not found errors, server statistics which includes top pages visited, entry/exit pages, and single access pages, referrers statistics such as top referring sites, search engines, and keyword etc., user demographics such as top geographical location, and most active countries/cities/organizations, client statistics such as visitor's Web browser, operating system, and cookies, and so on.

This information is used by web site manager for improving the system performance, facilitating the site modification task, and providing support for marketing decisions. However, most advanced Web mining systems process this information to extract more complex observations that deliver knowledge, utilizing data mining techniques such as association rules, clustering, and classification etc..

1. Number of Hits: This number usually signifies the number of times any resource is accessed in a Website. A hit is a

request to a web server for a file like web page, image, Cascading Style Sheet, etc. At the point when a web page is transferred from a server then the number of "hits" is same as the number of files requested. Therefore, one page load may be greater than one hit because frequently pages are made up of other images and other files which heap up the number of hits counted.

2. Number of Visitors: A "visitor" means who navigates to our website and browses one or more pages on our site.

3. Visitor Referring Website: The referring website gives the information or URL of the website which referred the particular website in consideration.

4. Visitor Referral Website: The referral website gives the information or URL of the website which is being referred to by the particular website in consideration.

5. Time and Duration: This information in the server logs give the time and duration for how long the Website was accessed by a particular user.

6. Path Analysis: Path analysis gives the analysis of the path a particular user has followed in accessing contents of a Website.

7. Visitor IP address: This information gives the Internet Protocol (I.P.) address of the visitors who visited the Website.

8. Browser Type: This information gives the information of the type of browser that was used for accessing the Website.

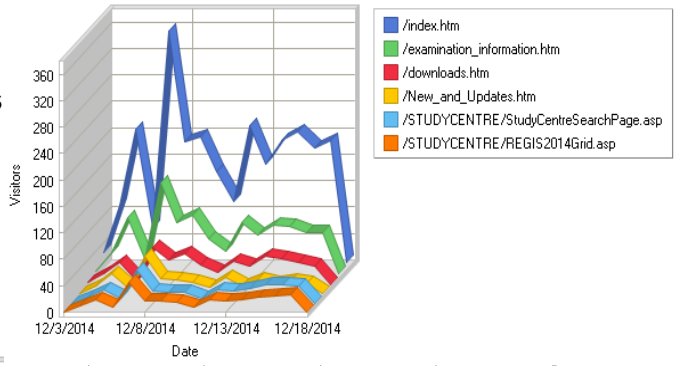
3.4.1 Statistical Analysis

For statistical analysis web log data is collected from the web server of MCU website for time period 03/12/2014 to 17/12/2014 into 15 files of size 18.7 MB. After that Web Log Expert Lite tool [5] is used to analyze the log file and corresponding results are shown below in figures and

Total Hits	83,575
Visitor Hits	81,355
Spider Hits	2,220
Failed Requests	15,311
Cached Requests	13,128
Total Page Views	11,785
Total Unique IP	2,872
Most popular page after home page	/examination_information.htm/
Top Search Engine	Google
Top Search Phrase	Makhanlal
Most used browser	Internet Explorer
Most used operating system	Windows XP
Most occurred error type	404: file not found

Table 1: General Statistics from the web log expert lite tool

Figure 5 shows the number of daily visitors who accessed website during the day, as it is clear from the figure the average visitors per day around 233 but last day no of visitors is less. Figure 6 shows that most of the visitors visited between 10:00 a:m and 7:00 p:m



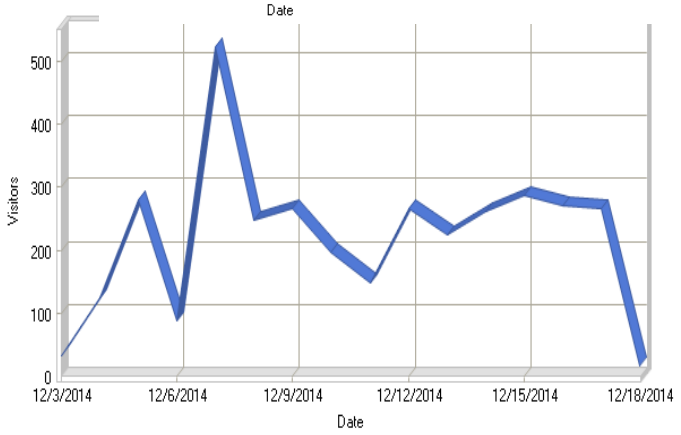


Figure 5: Daily Visitors

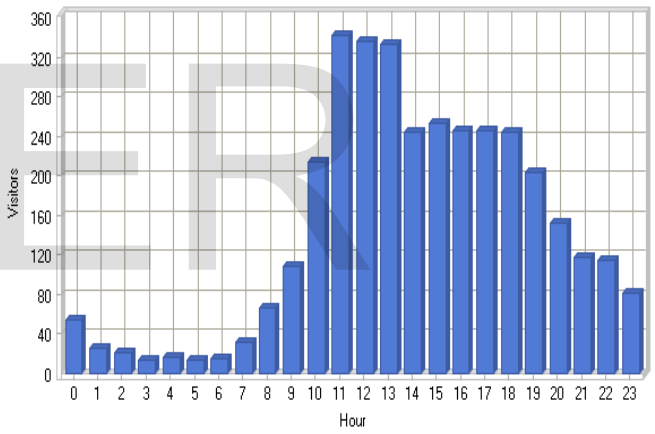
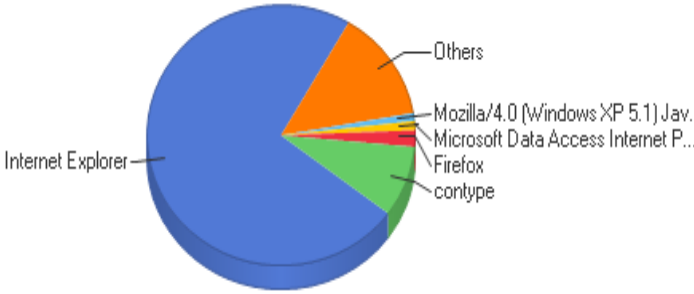


Figure 6: Activity by Hour of Day

Figure 7 shows accessed pages of website by visitors. Among them most popular page is /examination_information.htm page after the home page due to collection of log from the month of December which is generally treated as month for exam. This analysis is useful to arrange the pages of web site to facilitate fast accessing for visitors.

Figure 7: Daily Page Access

Figure 8 shows the top referring sites from where requests of pages of the websites have been done. Top most site is lbef.com after the home page.



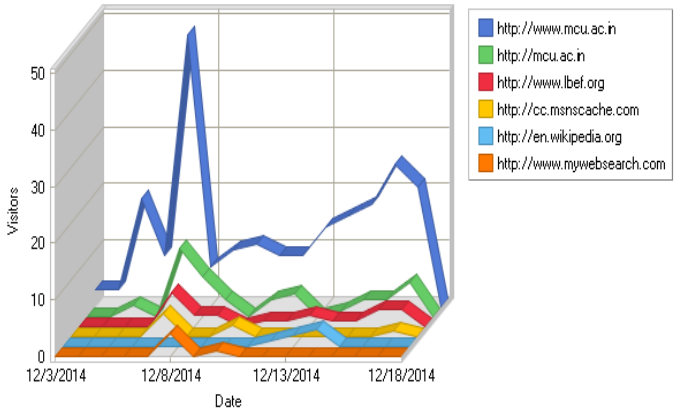


Figure 8: Daily Referring Sites

Figure 9 shows that the top most search engine used by visitors. Mostly used search engine is Google. Total 609 visitors have used search engines to access the web site. Among them 530 visitors have preferred Google. Others are MSN, Yahoo etc.

Figure 10 shows that daily used browser by visitors. Approx.73% of visitors have used Internet Explorer and rest of them used Firefox, Mozilla etc

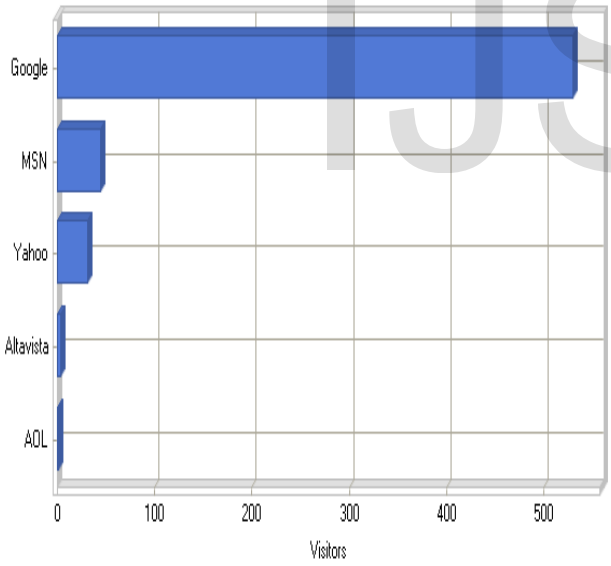


Figure 9: Top Search Engines

Figure 10 shows that used operating systems by visitors. Mostly visitors around 40 % have used Windows XP operating system and 20% have preferred Windows 98. Others are Windows 2000, Windows ME, Windows server2003, Linux etc

Figure 10: Most Used Browsers

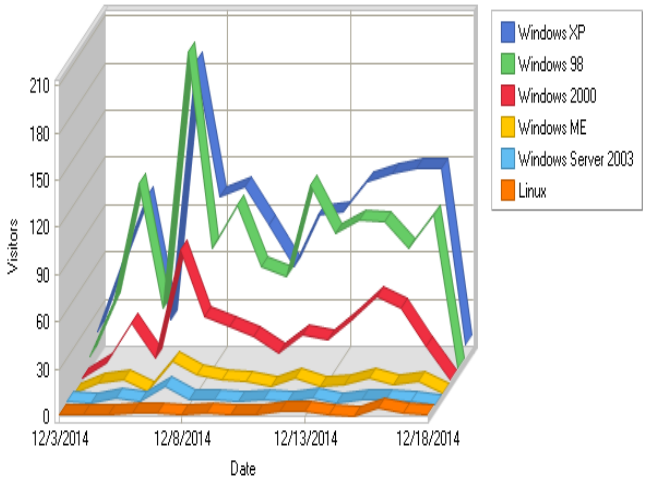


Figure 11: Daily Used Operating Systems

Figure 12 includes http request errors. Among 83,575 requests 15,311 are failed request which is approximately 18% of total requests. Most popular error is 404: Page not found.

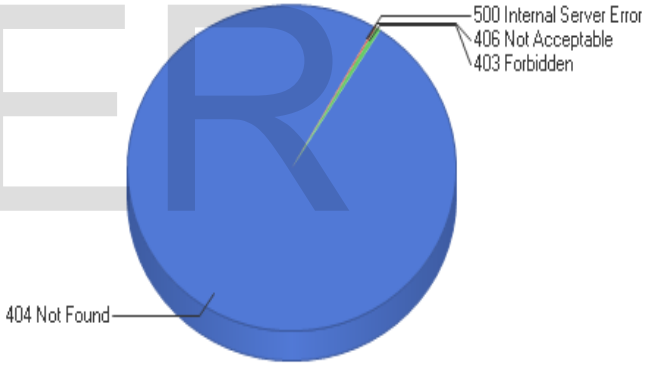


Figure 12: Error Types

3. CONCLUSION

Web log data is a collection of huge information. Many interesting patterns are available in the web log data. But it is very complicated to extract the interesting patterns without preprocessing phase. Preprocessing phase helps to clean the records and discover the interesting user patterns and session construction. Data preprocessing is an important task of Web log mining application. Therefore, data must be processed before applying data mining techniques to discover user access patterns from web log. The data preparation process is often the most time consuming as it includes different phases as data cleaning, user identification, session identification, and path

completion. The preprocessed data is then available for further pattern discovery and pattern analysis.

REFERENCES

[1]. Han, Jiawei and Kamber, Micheline, (2001), Data Mining Concepts and Techniques, ELSEVIER Morgan Kaufmann, USA.
[2]. Dunham Margaret H., (2003), Data Mining: Introductory and Advanced Topics, Pearson Education, India.
[3]. Eirinaki, Magdalini And Vazirgiannis, Michalis, (2003) 'Web Mining for Web Personalization', ACM Transactions on Internet Technology, Vol. 3, No. 1, pp 1–27.
[4]. Srivastava, Mitali and Garg, Rakhi and Mishra, P. K., (2014), 'Preprocessing Techniques in Web Usage Mining: A Survey' International Journal of Computer Applications, Volume 97– No.18.
[5]. http://www.weblogexpert.com, WebLog Expert Lite Tool, version 9.0,(2002-2015).
[6]. Srivastava, Jaideep and Cooley, Robert and Deshpande, Mukund and Tan, Pang-Ning, (2000), 'Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data', ACM SIGKDD, Vol.1(2), pp 12-23.
[7]. Mobasher, Bomshad, (2004), 'Web Usage MiningPersonalization',Practical Handbook of Internet Computing, Chapman and Hall/CRC, pp 1-31.
[8]. Pabarskaite, Zidrina & Raudys, Aistis, (2007), 'A process of knowledge discovery from web log data: Systematization and critical review', Journal of Intelligent Information Systems,Vol 28(1), pp 79-104.
[9]. Joshi, Anupam and Krishnapuram, Raghu (2000), 'On Mining Web Access Logs', ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, pp 1-11.
[10]. Chitraa, V. and Davamani, Dr. Antony Selvdoss (2010), 'A Survey on Preprocessing Methods for Web Usage Data', International Journal of Computer Science and Information Security,Vol. 7, No. 3, pp 78-83.