# Preprocessing Techniques in Web Usage Mining: A Survey

Mitali Srivastava
Department of Computer Science, Banaras Hindu University, Varanasi

Rakhi Garg
Computer Science Section, MMV, Banaras Hindu University, Varanasi

P. K. Mishra
Department of Computer Science, Banaras Hindu University, Varanasi

## ABSTRACT
Due to huge, unstructured and scattered amount of data available on web, it is very tough for users to get relevant information in less time. To achieve this, improvement in design of web site, personalization of contents, prefetching and caching activities are done according to user's behavior analysis. User's activities can be captured into a special file called log file. There are various 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: Data preprocessing, Pattern discovery and Pattern analysis. Among these steps, Data preprocessing plays a vital role because of unstructured, redundant and noisy nature of log data. 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. In this paper all these techniques are discussed in detail. Moreover these techniques are also categorized and incorporated with their advantage and disadvantage that will help scientist, researchers and academicians working in this direction.

## Keywords
Data mining, Web mining, Web usage mining, Data preprocessing.

## 1. INTRODUCTION
With the enormous growth of web there is a huge volume of structured, unstructured, semi- structured, heterogeneous, dynamic, distributed and high dimensional data available on web pages. So accessing relevant information with speed is a challenging task today. Several issues like multimedia data , scalability and temporal arises due to dynamic and diverse nature of data .While interaction with web various problems like finding useful information, personalization of information, to learn about consumers or individual users, creating new knowledge from the information available on web arises [1,2]. To solve these problems many techniques from Information retrieval (IR), Database, Natural Language Processing (NLP), Web mining are used directly or indirectly [4, 5]. Among them web mining has emerged as most popular and effective technique to overcome above problems in last few decades. Web mining is an application of data mining to extract uncover, relevant, hidden information on web. Web mining can be categorized into three classes based on content, structure and usage of web pages which is shown in Figure 1 [1, 27].

Apart from structural information and content information of web site, server logs are also considered as valuable source of information. Every time when a server of a website receives a request from web user, an entry is recorded in log file which is automatically stored and maintained by web server. Web usage mining is a field of study where these log files are analyzed and mined to generate useful patterns.

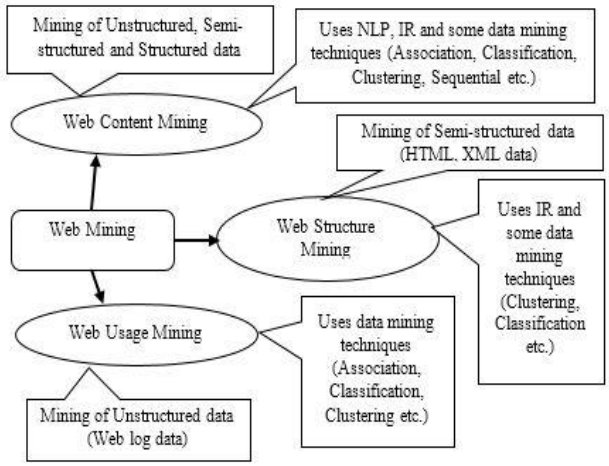



**Figure 1: Classification of web mining with influenced discipline [1, 27]**

Basically these kind of patterns are useful to characterize web users (user's navigational behavior) but the results obtained from web usage mining can also be exploited in various applications which are described below [6, 9]:

a) **Recommendations:** It is a process to analyze user's past behavior and current behavior to recommend new user for purchasing products or viewing certain pages. It is extensively used by commercial web-sites to recommend some products and services to users.
b) **Prefetching and Caching:** Web usage mining can be used to improve performance of web applications and web servers i.e. prefetching and caching of pages helps to improve response time of server.
c) **Web-site design improvement:** Ease of use is one of the important issues in designing of web-sites. Web usage mining gives user's behavior feedback to improve design of web application. Adaptive web sites is one of emerging application of this type.
d) **Business intelligence:** Extracting business intelligence from web usage data is important for online commercial web-sites. Main issues with this are customer retention, cross sales, customer attraction and customer departure.

Generally web usage mining processes includes three main steps Data preprocessing, Pattern discovery and Pattern analysis. Among them preprocessing has been considered as one of the essential step in web usage mining.

In this paper various techniques applied in preprocessing step of web usage mining are reviewed with their advantage and

disadvantage. Section II describes web usage mining and various data sources with their standard formats. After that section III discussed about different preprocessing techniques with their sub steps: data cleaning, user identification, session identification and path completion. Further section IV have included literature review of preprocessing techniques described in section III. It also includes statistical analysis of log data. At last section V concludes the paper.

# 2. WEB USAGE MINING

Web usage mining is one of the application of data mining which is used to mine of log files to discover useful patterns which can be further exploited in better personalization, improving navigations, recommendations, and recognition of web sites and attracting more advertisements etc. The web usage mining process is elaborated in Figure 2 [6, 11].Web usage mining generally uses basic data mining algorithms such as Association rule mining, Sequential rule mining, Clustering, Classification etc. for pattern discovery phase.

Due to high raise in number of transactions, *Association rule mining* is the most basic data mining technique to be used in web usage mining to find association between web pages. It refers to the set of pages that are accessed together in a single server session. This information can be useful to restructure the web site. *Clustering* is the most suitable techniques to analyze huge data sets by making clusters of those data. It groups the items based on some similar properties shared by them. In web usage mining clustering can be used in two ways i.e. usage cluster and page cluster. Further this type of analysis can be used as a base for recommendation systems and web personalization. *Classification* is used in web usage mining to group users according to predefined class labels based on their browsing history. It can be used in efficient personalization, profile building. *Sequential rule mining* is useful to predict user's navigation behavior which is further used in prefetching and catching to improve server's response time [8]. These data mining techniques can't be directly applied to log files due to unstructured, redundant, noisy nature of log data. So preprocessing of log files is an important step of web usage mining and it takes almost 80% time of whole web usage mining process [15].

## 2.1 Data Sources

Typically three main data sources are used to collect log data for web usage mining. Those are Server log, Proxy server log, Client/ Browser log as shown in Figure 3:

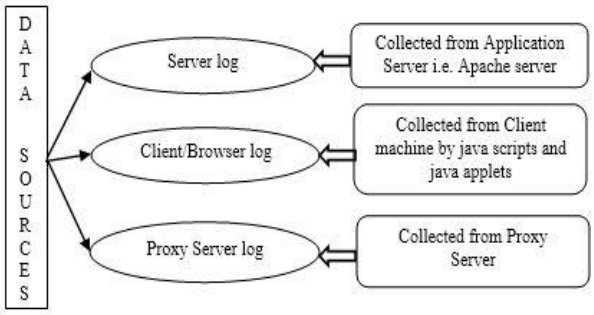



**Figure 3: Types of log data and its sources**

### *2.1.1 Server Log*

When an internet user request a particular page on web, an entry is logged into a special file called server log file. This file is not accessible by general internet user, only administrative person or server owners can access these files [8]. Server logs are considered as a richest and reliable source of information to predict user's behavior but it lacks with many quality factors such as completeness and privacy issues. Generally web server cached pages to provide fast accessing of pages in order to increase its response time. If a page is available in cache then no entry is logged in server log file for that particular page when it is requested by the web user. Path completion techniques have been used to resolve this issue [10]. Different web server provide various format of log files such as Common log format, IIS standard/extended log format, Combined/Extended common log format, Log markup language (LogML), because of different setting parameters[8,28].

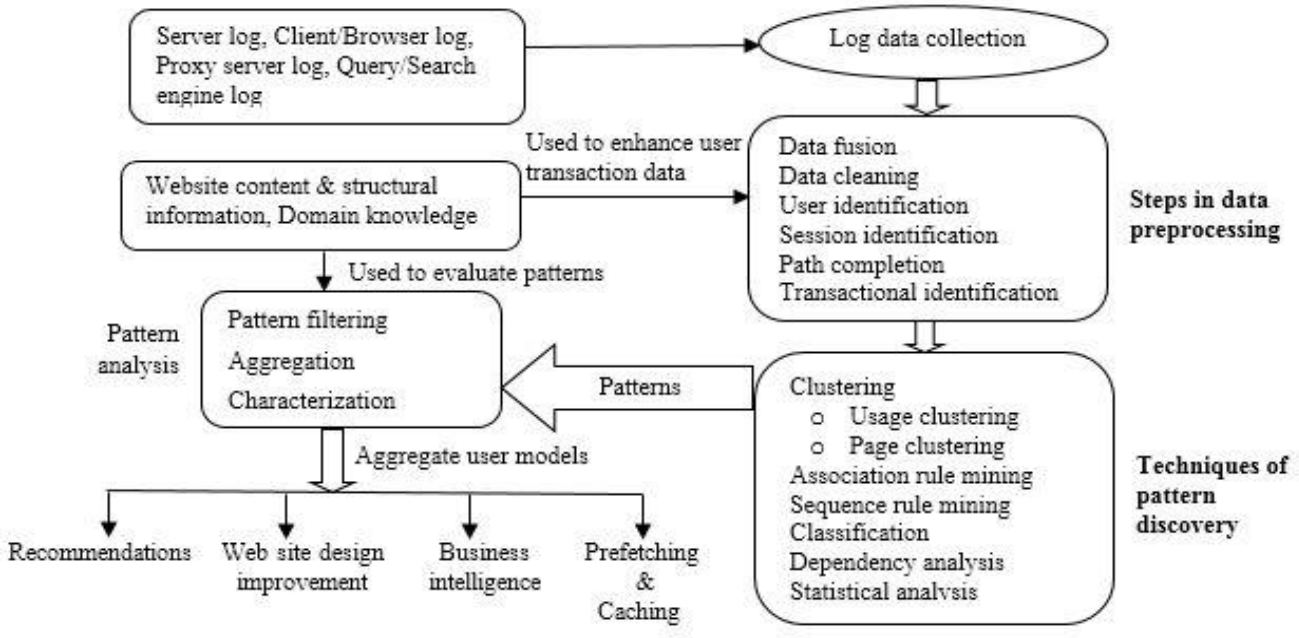



**Figure 2**: **Web usage mining process and its applications [6, 11]**

Among them common Log format are commonly used. Common log format is a standard non-customized format (fixed no of attributes) suitable for http web sites. This type of log includes user's *IP address/hostname, rfcname, log name, date with time zone, page access method, PATH, http version, server response code and byte received* [8, 11]. On the other hand extended log format is a customizable log file format which can add some additional attributes like *referrer_url, http_user_agent* and *cookies* [3, 8]. Table-1 provides the description of attributes of extended common log format and Figure 4 shows the snapshot of extended common log format [12, 28].this snapshot shows the single entry of the log file.

It is taken from Banaras Hindu University web server (Apache/2.2.3 (Red Hat)). In that entry cookie information is not present because the cookies are not enabled by apache server.

**Table 1: Attributes of Extended Common Log Format**

| | Attribute | Description |
|---|---|---|
| Attributes of *common log format* | IP Address/Host address | It identifies the visitor's machine address. |
| | rfcname | Identifies user's authentication. a "-" character shows that this field is empty. |
| | logname | Provides the login name of user. a "-" character shows that this field is empty. |
| | date with time zone | Returns the date and time of user's request. |
| | page access method | Describes the mode of request. It can be (GET, PUT, HEAD or POST) |
| | PATH | Give the path of requested page on server |
| | http version | Returns the version of http protocol. |
| | server response code | Provides the status of response given by server i.e. response code 404 represents file not found |
| | byte received | Denotes the size of file sent from server |
| | referrer_url | Provide the URL from where requested page is coming. When this field value is not present it is shown by "-". |
| | user_agent | Identifies user's operating system and browser's version |
| | cookies | It is a piece of information sent to visitors by web server to identify the details of a particular user. |

### *2.1.2 Client/Browser Log*

Web log data can also be collected from client machine by integrating java applets to the website, writing java scripts or even modified browsers. Client side logs are useful to tackle problems related with server logs like web page caching, session reconstruction [13, 14].

### *2.1.3 Proxy Server Log*

A Proxy server is a server which act as an intermediary between user's requests to other web servers. They are generally used for caching services to improve navigation speed, administrative control and security. Collecting proxy level usage data is similar as collecting server level data. However it also suffers from problems like caching and user identification. It is considered as most complex log source to predict a particular user access [14].

## 3. DATA PREPROCESSING

The main steps of web usage mining process are Data preprocessing, Pattern discovery and Pattern analysis [11]. Among them preprocessing is considered as a more complex and time consuming process due to diverse nature of log data. It has been observed that preprocessing of log file takes more time than other phases of web usage mining process [15]. It is necessary to perform preprocessing of log file to improve efficiency and scalability of basic data mining techniques applied on log data. Web log data preprocessing can be done in several steps: Data Fusion, Data cleaning, Page view identification, User identification, Session Identification, Path completion, Transaction identification and Formatting [7, 11]. Some techniques like Data cleaning, User identification, Session identification and Path completion have been discussed in detail one by one in following sub-sections.

## 3.1 Data cleaning

Web log file contains plenty of information into which some information are not relevant for web usage mining purpose so removal of these records is an essential step [12]. *First step* in cleaning is to remove unsuccessful http request that is recognized by status code field in log entries. The status code below 200 and above 299 having entries should be removed [28]. *The second step* is the removal of graphical contents (audio, video, images) as they are downloaded with the requested page even if they are not explicitly requested by the users [3]. Graphical files are easily identified due to its file extension (jpg, gif etc.). It is important to remove these kind of requests because they are just increasing the size of log file and nothing to do with analysis of user's navigational behavior. *The third step* is to remove log entries created by web robots (sometimes referred as web spiders or web crawlers). Robots are a special type of software which is used by various search engines to update its indexed pages by accessing pages of a particular website in a periodic time interval [11]. Robot's detection and removal is not easy as graphical contents removal. Some techniques for robot detection are [16]:

i) Checking user agent field where most robots declares themselves.
ii) Checking remote host name.
iii) Checking request of robot .txt file.

Apart from above techniques some heuristic techniques are also applied based on non-human behavior characteristics [17]. These heuristic based techniques are listed below:

i) Request for same URL is repeated by same host.
ii) Time interval between requests is too short.
iii) All request from single host whose referrer URLs are empty.

The simplest way to recognize robot request is to monitor navigation pattern of user. If a particular user accesses all pages of the web site it may be robot's requests [29].

## 3.2 User Identification

User identification is one of the complicated task due to existence of local/external proxy servers, cache systems, cooperate firewalls and shared internet [7, 8]. There are several methods to identify unique user is discussed below:

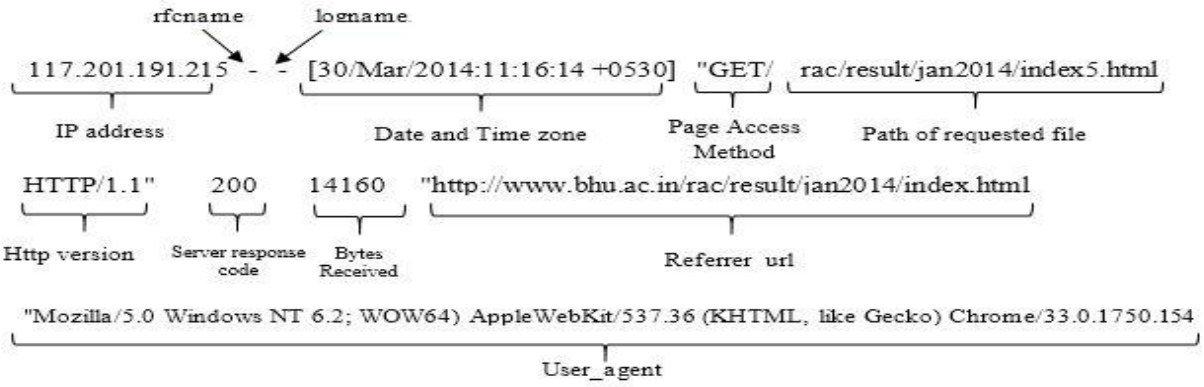



**Figure 4: Sample of Extended common log format (Collected from web server of BHU website)**

### *3.2.1 User identification by IP address*

IP address is used to assign a unique address to devices (computer, printers etc.) participating in network. IP address is logged into log file when a user hits a page. This address can be used to distinguish different users. But in case of proxy server when many users request a particular page then web site server logged same IP address (Proxy server IP) into the log file. Practically different users are accessing that page. Caching also creates problem to identify unique user. Whenever a user try to access previously accessed page, browser display pages from local cache and no entry is logged into the log file.

Several methods have been proposed to identify correct user and to address above issues. Prabarskaite [18] solve this issue by rejecting entries into log file if they are from proxy servers. According to him requests coming from proxy servers can be identified by domain name .i.e. if request is through proxy server the domain definition contains "proxy or cache" word. The problem with this approach is that it miss some important patterns by proxy users by rejecting their entries in log file.

### *3.2.2 User Identification by authentication data*

Registration data can also be used to identify unique user .This is best way to identify unique user by using logname and rfname attributes of log file (Table 1) if authentication data (username and password) is asked while requesting a page. This method is not so much popular because user always try to avoid such type of web sites [19].

### *3.2.3 User Identification by cookies*

Cookie is a piece of data sent by a web server to client machine when user request a web page for the first time. This information stored in a text file on client machine with the browser. Cookies can contain useful information regarding user so it is possible to correctly identify unique user by using it[9].There are some scenario where cookies doesn't work i.e. some browsers does not support cookies, some browsers disables cookie, cookies are deleted by the users and cookies are not logged by the web servers or deleted by the servers.

### *3.2.4 User Identification by client information*

Mostly researchers use some heuristic techniques to identify unique user. One of them is to look agent field of Log file which contains operating system name and browser's name with version. If two requests with same IP/host addresses have different browser's name or operating system then there is possibility that theses request from two different users [7]. Although this method is not reliable and results in confusion i.e. if a user is visiting two pages of web site by using different browsers simultaneously on single machine then this method will consider that two request by different user even they are from single user.

### *3.2.5 User Identification by site topology*

This method uses structural site topology of web site to identify unique user. Cooley et al. [7] has assumed that if a user request a page that is not accessible through its previously requested pages is considered as a new user. This can be done by using referrer attribute of extended log format and link information from site topology. Some situation where this approach results in confusion i.e. if user make a request by using bookmarked pages which are not connected via links.

## 3.3 Session Identification

Once user is identified there is need to identify sessions. Session is set of requests done by single user for defined duration to a particular web site. Basically there are two ways to find sessions regarding particular user [11]:

i) By using authentication information from users such as cookies mechanism or embedded session id.
ii) By applying some heuristic techniques.

These two methods are also called as "proactive" and "reactive" methods [20]. Proactive strategies creates session based on session_id collected from cookies. It creates some cookies related issues which is listed in user identification methods. In Reactive strategy sessions are constructed from web log information [20]. Most of the researchers have worked on the reactive methods because proactive strategy relies on user's cooperation. Some reactive methods are discussed below in detail.

### *3.3.1 By the Time gap*

When the time gap between two consecutive requests by the same user is greater than certain threshold then a new session is created. Creation of new session can be represented mathematically by given equation [11].

$$\boldsymbol{s.t_{n+1} - s.t_n \geq time_{threshold}, \quad then\ new\ session\ created} \qquad (i)$$

Where $s.t_{n+1}$ and $s.t_n$ are time stamp of two consecutive request. The most popular threshold value used by many researchers is 25.5minute. However this can vary from 10 minutes to 2 hour [3]. This value can be determined by several parameters like site topology, application type etc. Most of the commercial websites and open source tools takes 30 minutes

threshold value. *Adaptive or dynamic threshold* can also be used to improve efficiency of session construction [21, 22].

### *3.3.2 By the Referrer Attribute*

Session can be identified by referrer attribute in extended log format (Table 1). Suppose x and y are two requests for consecutive pages by same user and $x \in S$ (a session), if referrer of y was invoked previously in that session S then y is added in session S otherwise a new session is created with y as a first requested page [14].

### *3.3.3 By the Time spent on observing page*

Cooley [7] categorizes pages into two groups: *Information pages* and *Navigational pages* based on time spent on these pages. *Information pages* are those pages in which users are interested and *Navigational pages* are those pages which helps (for navigational purpose only) to user's to reach at information pages. Users spend more time on informational pages than navigational pages. The duration of time spend on navigational pages are smaller. If percentage of navigation page is assumed in log file, then maximum length of navigation page is given by the formula:

$$q = -\frac{\ln(1-\gamma)}{\mu} \qquad (ii)$$

Where q denotes threshold value of navigational pages, γ represents the percentage of navigational pages and μ denotes the mean value of observed duration time for all pages in log file [7].

## 3.4 Path Completion

After identify unique user session there is need to determine important page accesses that are not logged into the log file due to presence of client side or proxy side caching. If a user accesses a page by using back button in browser then it return copy of that page which is stored in cache. This kind of accessing does not record any entry in log file that causes problem of missing references hence path completion techniques are required to fill these entries in log file [23]. To find missing references there is need of referrer attribute of log file and site topology of that web site. If the URL of referrer attribute is not same as the previous requested page then that path is incomplete. This shows that user have used back button to visit that page. The summarization of all data preprocessing techniques are given in Figure 5:

# 4. LITERATURE REVIEW & CRITICAL ANALYSIS

In above sections various techniques used in preprocessing of log files have been discussed. Commonly used sub-steps of preprocessing are Data cleaning, User identification, Session identification and Path completion. Different researchers have introduced various preprocessing techniques to improve efficiency and scalability of pattern discovery techniques. Some of them are discussed below:

Cooley et al. [7] have proposed methods for data cleaning, user identification, session identification and transaction identification. Although their methods are good enough but some heuristics are not appropriate for complex web sites.

Prabarskaite [15] proposed a better cleaning methodology. According to him standard cleaning methodology is not appropriate for frame pages containing websites. He applied two approaches: advanced cleaning to improve web log mining and filtering to remove irrelevant links. In this preprocessing process author did not perform any other steps of preprocessing like user identification session identification etc.

Tanasa et al. [24] divides preprocessing process in four steps: Data fusion, Data cleaning, Data structuration and Data summarization. In Data fusion author joined multiple log files from different web servers and also from site maps into a single log files. After that they anonymized log file by encrypting host name. Further Data cleaning is performed by removing requests for non-analyzed resource such as multimedia files (images, audio, video etc.) and robot's generated requests In Data structuration part author have completed user identification by authentication data or IP address, Session identification by host and agent, Page view identification by site map etc. At last Data summarization step includes pattern analysis part by using data generalization and aggregation. They did not considered unsuccessful request in data cleaning phase which is also required to remove to get rid of unnecessary calculations in later phases of web log mining processes.

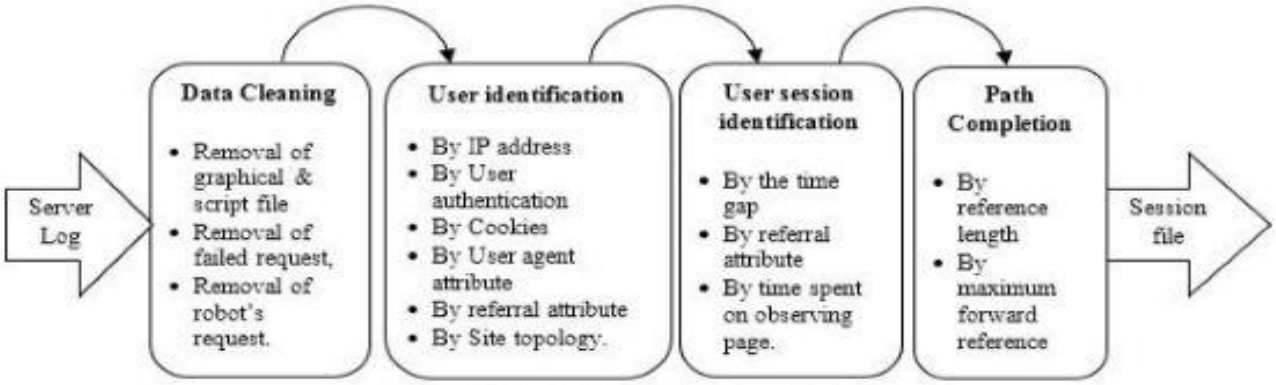



**Figure5: Web usage data preprocessing techniques**

Castellano et al. [25] developed a tool LODAP (Log Data Preprocessor) which takes log file as input and gives statistical analysis and user sessions as output. This tool is divided into three modules: Data cleaning module, Data structuration module and Data filtering module. Data cleaning module removes multimedia files, status code, and robot's request from log files. In Data structuration module users are identified by authentication data/IP address and sessions are identified by time based heuristics. The maximum elapsed value for session identification has set to 30 minutes and

minimum to 2 seconds between two consecutive requests. Further in Data filtering most requested pages are retained and least requested pages are dropped out based on threshold value. Authors have completed almost all the steps of data preprocessing except path completion which is also an important step in case of cache/proxy server. They should also include some more attributes of log file with IP address for user identification effectively.

Robert et al. [26] introduced a new concept called integer programming for better session identification .This method generates session simultaneously and produced session better match to an empirical distribution.

Yen li et al. [23] proposed an approach for path completion by combining Maximal forward reference length and Reference length algorithm. First Maximal forward reference is used to find the sequence of page in user access path and it is also used to identify the page, and finally Reference length algorithm is used to find whether the page is informative page or auxiliary page. Lastly by using referrer field complete path has been built.

Xiang–ying li [30] has proposed an algorithm named CSIA (Client and Session Identification algorithm) for identification of user and sessions. This algorithm includes comprehensive approach by combining IP address, topology, browser version and referrer page to identify unique user with better accuracy and efficiency. He proposed his algorithm in JAVA language framework as it is good for space utilization. However this algorithm is suffering with decrease in operating rate due to consideration of many factors for identifying user. The summarization of literature review is given in Table 2.

**Table 2: Summary of Literature Review of Preprocessing Techniques**

| Author | Preprocessing techniques | Focused on | Remarks |
|---|---|---|---|
| Cooley et al. [7] | Data Cleaning, User Identification, Session Identification, transaction identification | Transaction Identification | Proposed heuristics are not suitable for complex web sites |
| Prabarskaite [15] | Advance data cleaning, Filtering and data visualization | Data cleaning | Did not perform any other preprocessing technique like user identification and session identification etc. |
| Tanasa et al. [24] | Data fusion, Data cleaning, Data structuration and Data summarization | All except path completion | Ignored the removal of wrong http request status code |
| Castellano et al. [25] | Data cleaning module, Data structuration module and Data filtering module | All except path completion | Included almost all steps of data preprocessing. |
| Robert et al. [26] | Data cleaning and filtering, User identification, Session Identification | Session Identification | Better session creation simultaneously by using integer programing |
| Yen li et al. [23] | Data cleaning, User identification, Session Identification and path completion | Path Completion | Combined two approaches Maximal forward reference length and Reference length to find out completed path |
| Xiang-ying li[30] | Data cleaning, Client Identification, Session Identification and Path Completion | Client Identification and Session Identification | High accuracy and high efficiency but poor operating rate. |

## 4.1 Statistical Analysis

For statistical analysis log data is collected from the web server of Banaras Hindu University website for time period 23/03/2014 06:44:04 to 30/03/2014 11:16:06 into a file of size 1.5 GB. After that Web Log Expert Lite tool [31] is used to analyze the log file and corresponding results are shown below in figures and Table 3:

**Table 3: General Statistics from the web log expert lite tool**

| | |
|---|---|
| Total Hits | 6,660,811 |
| Visitor Hits | 6,604,320 |
| Spider Hits | 56,491 |
| Failed Requests | 738,089 |
| Cached Requests | 554,513 |
| Total Page Views | 855,437 |
| Total Unique IP | 162,022 |
| Most popular page after home page | /admission/ |
| Top Search Engine | Google |
| Top Search Phrase | BHU |
| Most used browser | Google chrome |
| Most used operating system | Windows 7 |
| Most occurred error type | 404: file not found(735,556) |

Table 3 includes general statistics like total hits, visitor hits, spider hits, failed requests, cached requests, total page views, total unique IP, most popular page after home page, most used search engine, most used operating system and most occurred error type.

Figure 6 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 30,000 but last day no of visitors is less. This is due time of the log as it is taken till 11 am for last day. .

**Figure 6: Daily visitors**

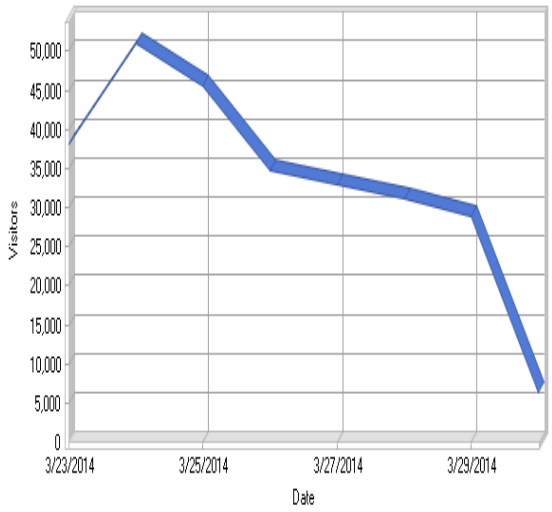



Figure 7 shows accessed pages of website by visitors. Among them most popular page is /admission page after the home page due to collection of log from the month of March. This analysis is useful to arrange the pages of web site to facilitate fast accessing for visitors.

**Figure 7: Daily page access by visitors**

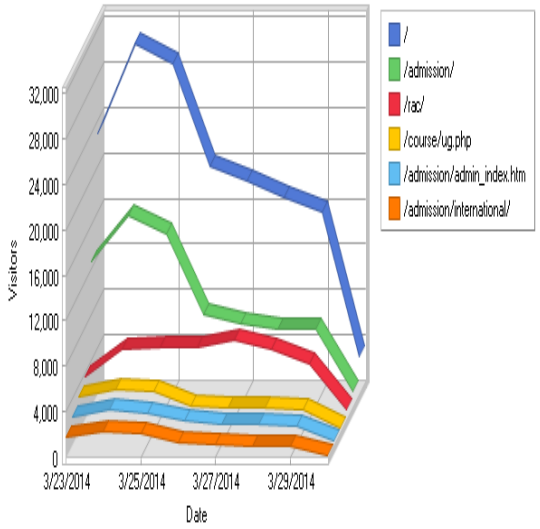



Figure 8 shows the top referring sites from where requests of pages of the websites have been done. The top most site is free jobalert.com after the home page. This is due to the time duration of log collection when university was doing recruitments.

**Figure 8: Daily used referring sites**

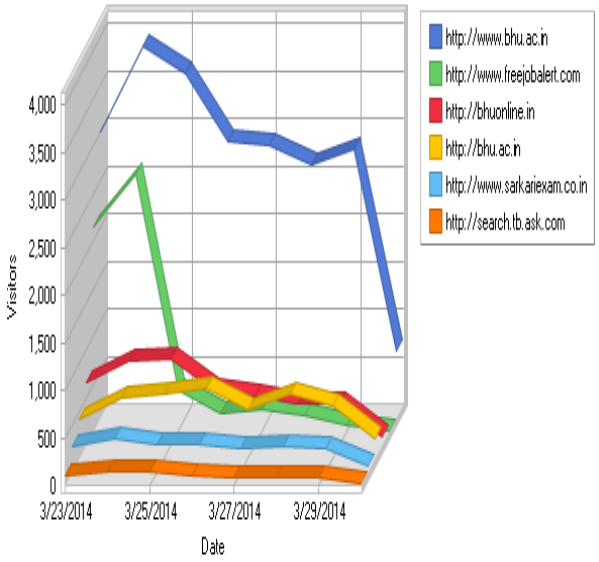



Figure 9 shows that the top most search engine used by visitors. Mostly used search engine is Google. Total 131391 visitors have used search engines to access the web site. Among them 126,737 visitors have preferred Google. Others are Bing, Yahoo etc.

**Figure 9: Daily used search engines**

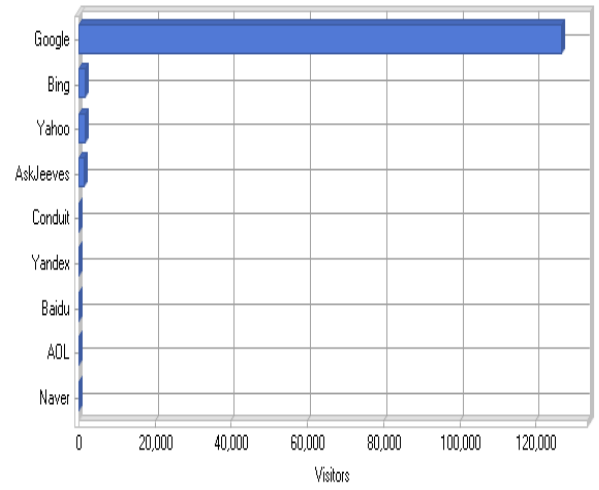
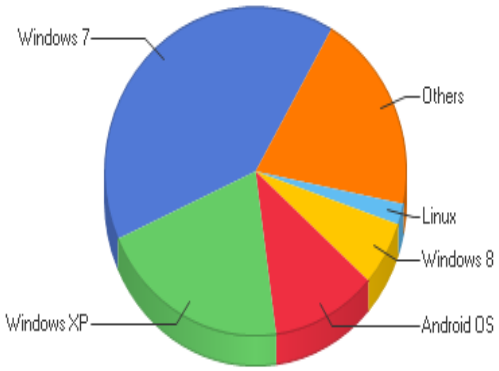


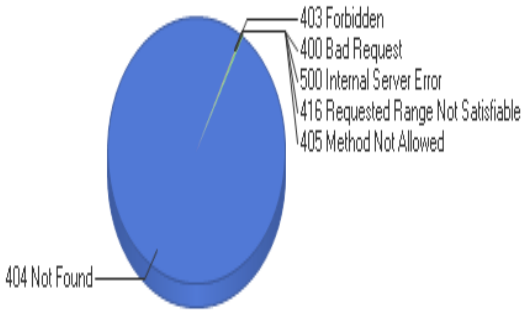

Figure 10 shows that used operating systems by visitors. Mostly visitors around 40 % have used Windows 7 operating system and 20% have preferred windows XP. Others are Linux, Android etc.

**Figure 10: Used operating systems**




At last Figure 11 includes http request errors. Among 6,660,811 requests 738,089 are failed request which is approximately 11% of total requests. Most popular error is 404: Page not found.

**Figure 11: Error types**




From the above analysis it is observed that one day total hits is around 60 million which is a huge number of hits. But in these request around 11% of total hits is failed requests and around 1% is spider hits. So there is need to remove these irrelevant data to get rid of unnecessary calculation for further phases of mining.

## 5. CONCLUSION

In last few decades web has become an informational hub for users. Thus analysis of user's behavior is becoming more and more important for e-commerce companies to provide better services to customers and visitors. Web usage mining is a field of study where user's activity is analyzed and processed to generate useful patterns. Due to irrelevant data in log file, preprocessing is considered as an essential step in web usage mining. In this paper different steps of preprocessing: Data cleaning, User identification, Session identification, and Path completion have been discussed. Web usage mining depicts various challenging problems for preprocessing of log data. High dimensionality and large volume of data results in high computational complexity of mining process. So there is need to compress data without losing essential information regarding user's behavior. Apart from that, preprocessing techniques and proposed heuristics are also facing relevancy issues and no robust techniques are present to solve them. For example, sometimes due to privacy concern, cookies and user authentication data is not available in log file to correctly identify user. To correctly identify unique user some heuristics techniques are proposed but they are suffering with exceptions. Further in future, combination of two or more user identification techniques can be used to make better user identification. This paper concludes that various applied data preprocessing techniques with their advantages and disadvantages and draws conclusion and research directions in future.

## 6. REFERENCES

[1] R. Kosala, H Blockeel (2000), Web Mining Research: A Survey in ACM SIGKDD Explorations, Vol.2 Issues 1, Page(s):1-15.

[2] B. Singh, H. K. Singh (2010), Web Data Mining Research: A Survey in Computational Intelligence and Computing Research (ICCIC), IEEE International Conference, Page(s): 1-10.

[3] Chintan R. Varnagar, Nirali N. Madhak, Trupti M. Kodinariya (2013). Web Usage Mining: A Review on Process Methods and Techniques in Information Communication and Embedded Systems (ICICES), IEEE International Conference, Page(s): 40 – 46.

[4] Qingyu Zhang, Richard Segall (2008), Web mining: a survey of current research, techniques and software in International Journal of Information Technology & Decision Making Vol. 7, No. 4 Page(s) 961-965.

[5] R. Cooley, B. Mobasher, J. Srivastava (1997), Web mining: information and pattern discovery on World Wide web in Tools with artificial intelligence Ninth IEEE International Conference, Page(s): 558-567.

[6] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-Ning Tan (2000), Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data, Vol.1 Page(s): 12-23.

[7] R. Cooley, B. Mobasher, J. Srivastav (1999), Data preparation for mining world wide web browsing pattern in Journal of Knowledge and Data Engineering Workshop, IEEE, Vol.1 Page(s): 5-32.

[8] Zidrina Pabarskaite, Aistis Raudys (2007), A process of knowledge discovery from web usage data: Systemization and critical review in Journal of Intelligent Information System, Springer Vol.28 Issue.1 Page(s): 79-104.

[9] F. Facca, P. Lanzi (2005), Mining interesting knowledge from weblogs: a survey in Data and Knowledge Engineering, Vol. 53 Issue 3, Page(s): 225–241.

[10] Yuan, F., L.-J. Wang, et al. (2003), Study on Data Preprocessing Algorithm in Web Log Mining in Proceedings of the Second International Conference on Machine Learning and Cybernetics, Vol. 1 Page(s): 28-32.

[11] Bing Liu,Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data, 2nd ed. 2011

[12] Tasawar Hussain (2007), Hierarchical sessionization at preprocessing level of WUM based on swarm intelligence in 6th International Conference on Emerging Technologies, IEEE Page(s): 21-26.

[13] C. Shahabi, F. Banaei-Kashani (2002), A Framework for Efficient and Anonymous Web Usage Mining Based on Client-Side Tracking in WEBKDD Third International Workshop on Mining Web Log Data, Page(s): 113-144.

[14] D. Pierrakos, G. Paliouras, C. Papatheodorou, and C. D. Spyropoulos (2003), Web usage mining as a tool for personalization: A survey in User Modeling and User Adapted Interaction journal, Vol. 13 Issues. 4 Page(s): 311-372.

[15] Pabarskaite Z (2002), Implementing advanced cleaning and end-user interpretability technologies in web log mining in 24th International Conference on Information Technology Interfaces (ITI), Vol. 1 Page(s): 109-113.

[16] P.-N. Tan, V. Kumar (2000) Modeling of web robot navigational patterns, in: WEBKDD Web Mining for Ecommerce Challenges and Opportunities, Second International Workshop.

[17] Berendt, B. spiliopoulou M (2000), Analyzing navigation behavior in Web sites integrating multiple information systems in VLDB Journal, Special Issue on Databases and the Web, Vol. 9 Page(s): 56-75.

[18] Pabarskaite Z (2003), Decision trees for web log mining in Intelligent Data Analysis Journal, Vol. 7 Issue. 2 Page(s): 141–155.

[19] Renata Ivancsy, and Sandor Juhasz (2007), Analysis of Web User Identification Methods in World Academy of Science Engineering and Technology, Vol. 34, 2007.

[20] Berendt, B., Mobasher, B., Nakagawa, M., & Spiliopoulou, M. (2002). The impact of site structure and user environment on session reconstruction in web usage analysis in 4th WebKDD Workshop on Knowledge Discovery in Databases Edmonton.

[21] M. Chen, A.S. LaPaugh, J.P. Singh (2002), Predicting category accesses for a user in a structured information space in Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, Page(s): 65–72.

[22] J. Zhang, Ali A. Ghorbani (2004), The Reconstruction of user session from a server log using improved time oriented heuristic in 2nd Annual Conference on Communication Networks and Service Research IEEE, Page(s): 315-322.

[23] Yan LI (2008), Research on path completion technique in web usage mining in International Symposium on Computer Science and Computational Technology, IEEE, Vol. 1 Page(s): 554-559.

[24] D. Tanasa, B. Trousse (2004), Advanced Data Preprocessing for Intersites Web Usage Mining in IEEE Intelligent Systems, Vol. 19 Issues. 2 Page(s): 59-65.

[25] G. Castellano, A. Fanelli, M. Torsello, LODAP: A Log Data Preprocessor for Mining Web Browsing Patterns in Proceedings of the 6th Conference on Artificial Intelligence, Knowledge Engineering and Data Bases, Page(s): 12–17.

[26] R. F. Dell (2008),Web user session reconstruction using integer programming in International Conference on Web Intelligence and Intelligent Agent Technology, IEEE/ACM/WIC, Vol. 1 Page(s): 385-388.

[27] Atul Kumar Srivastava, Mitali Srivastava, Rakhi Garg, P. K. Mishra (2014), Comparative Study of Web Page Ranking Algorithms in IJETCAS, ISSN (Print): 2279-0047, ISSN (On-line): 2279-0055, Issue 7, Vol. 1 Page(s): 26-32.

[28] Wahab, M. H. A., M. N. H. Mohd, et al. (2008), Data Preprocessing on Web Server Logs for Generalized Association Rules Mining Algorithm. World Academy of Science, Engineering and Technology.

[29] P.E. Román, G. L'Huillier, J.D. Velásquez (2010), Web usage mining advanced Techniques in Web Intelligence, Springer (2010), Page(s): 143–165.

[30] Xiang-ying Li (2013), Data Preprocessing in Web Usage Mining in 19th International Conference on Industrial Engineering and Engineering Management Page(s): 257-266.

[31] Web Expert Lite Tool version 8.4, www.weblogexpert.com.