International Journal of Research and Reviews in Wireless Communications (IJRRWC)
Vol. 2, No. 2, June 2012, ISSN: 2046-6447

www.sciacademypublisher.com/journals/index.php/IJRRWC

Data Preprocessing on Web Server Log Files for Mining Users Access Patterns

Thanakorn Pamutha[1], Siriporn Chimphlee[2], Chom Kimpan[1], and Parinya Sanguansat[3]

[1]Faculty of Information Technology, Rangsit University, Muang-Ake, Phathumthani, Thailand
[2]Faculty of Science and Technology, Suan Dusit Rajabhat University, Bangkok, Thailand
[3]Faculty of Engineering and Technology, Panyapiwat Institute of Management, Nonthaburi, Thailand

Email: thanakorn.p@yru.ac.th, siriporn_chi@dusit.ac.th, chom@rsu.ac.th, sanguansat@yahoo.com

Abstract – Web Usage Mining (WUM) is the application of data mining techniques to discover the knowledge hidden in the web log file, such as user access patterns from web data and for analyzing users' behavioral patterns. Thewebsite may likewise be accessed for various website design tasks. Nonetheless, the data stored in the web log file has a large amount of erroneous, misleading, and incomplete information. Preprocessing which is one of the important phases in WUM is needed to transform a log into a set of web user sessions that are suitable for analyses. A sample web log file was collected from the web server at NASA Kennedy Space Center. This study focuses on the preprocessing of the web log file methods that can be used for the task of session identification from web log file. The work in this study also produces statistical information of user session, such as: (1) total unique IPs; (2) total unique pages; (3) total sessions; (4) Session length and (5) the frequency visited pages. After preprocessing completed, the result will be used for mining user access patterns.

Keywords – Usage Mining, Pattern Analysis, Web Log file, Preprocessing, Session Identification

1. Introduction

Web servers log files that can collect activity information when a user accesses to a Web Server. In recent years, [1] the massive influx of information onto World Wide Web has facilitated users not only in retrieving information but likewise in discovering knowledge. However, web users usually suffer from the information overload considering the significant increase and the rapidly expanding growth in the amount of information available on the web. This increase in the size of data available through the web has made it necessary to find intelligent ways to retrieve the data needed and the user's behavioral pattern in collecting the said data. Web data mining [2] is the application of data mining techniques on web data. Web mining is divided into three types. They are 1) Web content mining that deals with the useful discoveries of web contents and services, 2) Web structure mining aims to understand the structure of hyperlinks within the web itself, and 3) web usage mining mineslogsthe data stored in the web server.

Web usage mining [3] is application of data processing techniques that discover usage patterns of users from the available web data. This is to ensure an improved service of web-based applications. The user access log files present very significant information about a web server. It is applied to fix several world problems by discoveringthe interesting user navigational patterns. This eventually leads to applied improvementson website designs using the shortest possible time. Moreover, recommendations on pertinent web content improvements can readily be made by studying the user's web access patterns. WUM has several applications [3, 4] that are utilized in the following areas: 1) it offers users the ability to analyze massive volume of click stream or click flow data, integrate the data seamlessly with translation and demographic data from offline sources. 2) Personalization of data information based on previously accessed pages. These pages can immediately identify the typical browsing behavior of webusersthat could subsequently predictand intuit their desired sites. 3) By determining the access behavioral patterns of web users, the preferred linksmay be identified to improve and auto-intuit (predict) future access. 4) Web usage patterns are used to gather business intelligence to improve customer attraction, customer retention, sales, marketing, and advertisements cross sales. 5) Web usage mining is used in e-Learning, e-Business, e-Commerce, e-Newspapers, e-Government and Digital Libraries. The information gathered through web mining is evaluated by using traditional data mining parameters such as clustering and classification, association, and examination of sequential patterns. Web usage mining comprises three phases namely preprocessing, pattern discovery and pattern analysis. Nevertheless, [5] the data stored in the log files do not present an accurate picture of the users' accesses to the Web Server. Hence, preprocessing of the Web log data is an essential and pre-requisite phase before it can be used for the pattern discovery task.

This study concentrates mostly on the preprocessing stage of web usage mining in determining web access patterns. All the web log files were culled from the NASA web server that could automatically extractparticular sessions of the user's web behavior that includes but is not necessarily limited to removal of images, sharing of multimedia files, style pagefiles, java script files and web robot requests. The contents of this paper is ordered as follows:(2) briefs related works, (3) sources of web logs, web structures, status codes

of Hyper Text Transfer Protocol and various step of Web Usage,(4) Experimental results as shown, (5) conclusions and future works.

2. Related Work

Several researchers were involved in the presentation of the Datapreprocessing phase that immediately involves analyses of web access patterns. This information has significantly improved the web service of various servers. Steps in data preprocessing have been discussed in detail in [1][2-5]. K, Draiswamy and V valli Mayil have analyzed [6] the session reconstructions in Web usage analysis. It follows reactive strategiesof users and provides session identification that utilizes the ExLF of HTTP server. The paper focuses on the methods of obtaining session details such as: (1) Session duration; (2) Page viewing time on each session; (3) Dweb time-staying time of a page; (4) Amount of data carried over the session; (5) Status of the page –success or failure of the retrieving page; and (6) Number of web pages that are accessed in each session. The session details can be used in the pattern analysis phase.

The author [7]identifies user behavior by analyzing the Web server access log file. This paper is concerned with the in-depth analyses of the web log data as culled from NASA website. It stores all the information about a web site, its top errors; it could likewise predict potential visitors to the site. This aids the systems administrator and web designersin predicting the future of the World Wide Web and incorporating these web intuitive methods in fixing systems errors and restoring the corrupted and broken links. The results obtainedfrom this study shall be used for further development and evolvement of websitesthereby maximizing the instant predictability of web trends and designs.C.P.Sumathi,R.Padmaja Valli and T.Santhanam[1] have presented an overview of the different steps involved in the preprocessing state.V.Chitraa and Antony Selvdoss Damamani [8] have served on preprocessing methods for Web usage data. This paper likewise reviews existing works as accomplished in the preprocessing stage. A brief overview of various data mining techniques for discovering patterns, and pattern analysis are discussed. Arshi Shamsi,Rahul Nayak,Pankj Pratap Singh and Mahesh Kumar Tiwar [9] have presented how the logged data in servers is preprocessed. This includes data cleaning, user identification, Sessionization and path completion. Once all the data is stored in, the preprocessed items are thereafter used as web references for future web patterns. Cooley et al.[10] presented the methods for user identification, session web focuses, page view identifications, path completion and episode identifications.

Aye et al [11]proposed a data preprocessing technique to prune noisy and irrelevant data, and to reduce data volume for the pattern discovery phase. This paper mainly focuses on data preprocessing with activities like field data extraction and data cleaning algorithms. Field extraction algorithm performs the process of separating fields from a single line of the log file. Data cleaning algorithm eliminates inconsistent or unnecessary items in the analyzed data. Sanjay Bapu et al [12]introduced an effective and complete preprocessing for web usage mining. This paper describes the effective and complete preprocessing of access stream before actual mining process can be performed. The Log file collected from different sources undergoes different preprocessing phases to make actionable data source. It is used for the automatic discovery of meaningful pattern and relationships from access stream users.Sathiyamoorthy and Murali Bhaskaran[13] introduced a various preprocessing techniques that are carried out at proxy server access Log which generate web access pattern and can also be used for further application.

Ramya and Kavitha [14]proposed a complete preprocessing methodology for discovering patterns in usage mining toimprove the quality of data by reducing the quantity of data.Maheswara Rao and Valli Kumari[15]introduced an extensive research frame work capable of preprocessing web log data completely and efficiently. The learning algorithm of proposed research frame work can separates human user and search engine access intelligently with less time. And also Data Cleaning, User Identification, Sessionization and Path Completion are designed collectively. The framework reduces the error rate and improves significant learning performance of the algorithm.This framework helps to investigate the web user usage behavior efficiently.

3. Web Logs and Web Usage Mining

Web log preprocessing aims to reformat the original web logs to identify all web access sessions.

3.1. Web Logs

A Web logs files[3]are computer files contains requests addressed to web servers.These are recorded in a chronological order. The most popular log file reformats the original file logged through CERN and the NCSA usingthe Common Log Format (CLF). The following is an example form of this web specimen using the server log file: (NASA_access_log_jul95).

```
199.72.81.55 - - [01/Jul/1995:00:00:01 -0400]
"GET /history/apollo/ HTTP/1.0" 200 6245
```

Figure 1. A fragment from the serve logs file (NASA_access_log_jul95).

This reflects the information as follows [3, 7, 16]:

- The Remote IP address: This identifies who had visited the web site;
- User Authentication: Username and password if the server requires user authentication (generally empty and represented by a "-");
- The date and time of the request: This attribute is used to determine how long a visitor has spent on a given page;
- The operation type (GET, POST, HEAD, etc.): The method used for information transfer is noted;
- The requested resource name: The resource accessed by the user. It may be an HTML page, a CGI program, or a script;
- The protocol Version: HTTP protocol being used.
- The request status: The HTTP status code returned to the client, e.g., 200 is "ok" and 404 are "not found";
- The requested page size: The content-length of the

document transferred.

The web access log line from Figure 1 shows that a user from the IP address 199.72.81.55 successfully requested the page “Apollo” on July 01, 1995at 00:00:01 a.m. The HTTP method the remote user used “GET”. The number of bytes returned to the user is 6245.

The status code of hypertext transfer protocol is shown in table 1

Table 1. Status codes of Hypertext Transfer Protocol[7, 17]

Code	Description
101	Switching Protocols
200	OK
201	Created
202	Accepted
203	Non-Authoritative Information
204	No Content
205	Reset Content
206	Partial Content
300	Multiple Choices
302	Found
303	See Other
304	Not Modified
305	Use Proxy
306	(Unused)
307	Temporary Redirect
400	Bad Request
401	Unauthorized
402	Payment Required
403	Forbidden
404	Not Found
405	Method Not Allowed
406	Not Acceptable
407	Proxy Authentication Required
408	Request Timeout
409	Conflict
410	Gone
411	Length Required
412	Precondition Failed
413	Request Entity Too Large
414	Request-URI Too Long
415	Unsupported Media Type
416	Requested Range Not Satisfiable
417	Expectation Failed
500	Internal Server Error
501	Not Implemented
502	Bad Gateway
503	Service Unavailable
504	Gateway Timeout
505	HTTP Version Not Supported
101	Switching Protocols

3.2. Web Usage Mining

Web usage mining[8]is also known as web log mining. It is the application of data mining techniques on websites for an easier facility website design tasks. The main sources of data for web usage mining consist of textual logs that had been gathered bythe web servers all over the world. There are four phasesinvolved in web usage mining: (1) Data Collection: users log data is collected from various sources like server side, client side, proxy servers and so on, (2) Preprocessing : Performs a series of processing of web log file covering data cleaning, user identification, session identification, path completion and transaction identification, (3) Pattern discovery : Application of various data mining techniques to processed data like statistical analysis, association, clustering, pattern matching and so on ,(4) Pattern analysis : once patterns were discovered from web logs, uninteresting rules are filtered out. Analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations.

The fundamental task of web log mining is data preprocessing. It comprises six difference tasks[6] : (1) Data collection, (2) Data cleaning, (3) User identification, (4) Sessions identification, (5) Session reconstruction, or path completion, and (6) Data formatting

3.2.1. Data collection

WUM applications are based on data collections from three main sources[7]: (1) Web server side, (2) The proxy side : Many Internet Service Provider (ISPs) give their customers proxy server services to improve navigation speed through caching. A proxy server collects data of groups of users accessing huge groups of web servers, and (3) the client side: Usage data may also be traced on the client side by using JavaScript, Java applets or even modified browsers.

3.2.2. Data Pre-Processing

Web log preprocessing aims to reformat the original web logs to identify all web access sessions. The web server usually registers all the users’ access activities through theweb server logs. Due to the different server setting parameters, they produce many types of web logs. Nonetheless, the log files contain similar basic information such as client IP address, request time, requested URL, HTTP status code, referrer and so on and so forth.

Generally, several preprocessing tasks are required to be performed before using web mining algorithms on the web server logs. For our work, these include data cleaning, session identification and data formatting. The original server logs are cleaned, formatted, and then grouped into meaningful sessions before being utilized by WUM.

a) *Data Cleaning* The data cleaning process removes the data tracked in web logs that are useless or irrelevant for mining purposes. The request processed by auto search engines, such as Crawler, Spider, and Robot, and requests for graphical page content (e.g., jpg and gif images) are deleted because these image files are auto-downloaded with the requested pages.
b) *User identification* The user identification process analyzes the log file and clusters the users so that every user in the same group has the same access characteristics.
c) *Sessions identification* Once the log files have been cleaned, the next step in the data preprocessing is the identification of the session. Session identification is the process of segmenting the user activity log of each user into groups of page references during one logical period called session. V.Chitraa and Antony Selvdoss Davamani[8] have surveyed three different methods of session identification ,reconstruction and navigation using the web topology.

3.2.3. Time Oriented Heuristics

The simplest methods is time oriented that is derived by accumulating all the logged-in time. The other method is based on a single page stay time. The setof pages visited by a specific user at a specific user at a specific time is called page viewing time. It varies from 25.5 minutes to 24 hours while 30 minutes is the default timeout. The second method

depends on page stay time which is calculated with the difference between two timestamps. If it exceeds 10 minutes the second entry is assumed as a new session. Time based methods are not reliable because users may involve in some other activities after opening the web page and factors such as busy communication line, loading time of components in web page. Content size of web pages is not considered.

3.2.4. Navigation-Oriented Heuristics

Uses web topology in graph formatting. It provides webpage connectivity but it does not offer a hyperlink capability between two consecutive page requests. If a web page is not connected with previously visited page in a session, it is considered as a different session. Cooley proposed a referrer based heuristics on the basis of navigation in which referrer URL of a page should exist in the same session. If no referrer is found, the first page is considered a new session.

Marathe Dagadu Mitharam[18] has described Session captures in two ways.Time Oriented is depends on the Tie stamps or date and time of request in the server log file. The two time oriented session has two types (1) The difference between First request and last request is <= 30 minutes, (2) The difference between First request and ext., request is <= 10.

V. Sathiyamoorthi and V.Murali Bhaskaran [2] have presented data preparation techniques for Web Usage Mining. Many commercial products use 30 minutes as a default timeout but had established a timeout of 25.5 minutes based on the data provided.

C Gomathi M Morthi and K duraiswamy[3] have created session Identifications. The session identification process begins with the IP address of a computer which is a unique identity. The session provides a log-in process for time usage, the number of website visits using the URL address and the number of sessions that the user had logged in a particular website.

C.P.Sumathi, R.Padmaja Valli and T.Santhanam[1] have presented the session identification process of all the pages accessed by a user using various sessions. It is assumed that the user has started a new session if the time between two pages requests exceeds the given time limit. In general, a 30 minutes default timeout is considered also Shaimaa Ezzat Salama[19] has used a 30 minutes to identified session for Web instruction detection.

L.K. Joshila Grace,V.Maheswari and Dhinaharan Nagamalai[5] have described, the session done by using the time stamp details of the web pages. The total time used by each user to view each web page is likewise noted. This can also be done by noting down the user id those who have visited the web page and had traversed through the links of the web page. Session is the time duration spent in the web page.

3.2.5. Path completion

The path completion process identifies unique user sessions by adding important accesses that are not recorded in the access log. Finally, in the data preprocessing task, the sequence of identified pattern may be stored in the relevant data structures.

3.2.6. Data formatting

The data formatting module is the final module of preprocessing. The data should be appropriately formatted according to the type of mining tasks undertaken[1]. Information which is viewed irrelevant or unnecessary for the analysis may not be included in the resultant session file.

4. Working Scheme

In the working scheme of this paper two main modules are involved: data cleaning and session identification.

4.1. Data Cleaning

Web Log files may contain a number of records corresponding to automatic requests originated by web robots, a large amount of erroneous, misleading, and incomplete information. The web log files [20] involved filtering out requests that were requested by robot or spider or crawlers. These[6] are programs that automatically download complete websites by following every hyperlink on every page within the site in order to update the index of the search engine. Request created by web robots are notconsidered as used data and consequently, need to be removed.

In this step, The entries [7] that have status of “error” or “failure” should be removed, then some access records generated by automatic search engine agent should be identified and removed from the access log and also this process removes requests concerning non-analysed resources such as images, multimedia files, and page style files. For example, requests for graphical page content(*.jpg & *.gif image) and requests for any other file which might be included into a web page or evennavigation sessions performed by robots and web spiders.

The following is the algorithm[21] which can be used to data cleaning:

Algorithm DataCleaning (LogFile: Web log file; LogFile: Web log file)

```
Begin
 While not eof (LogFile) Do
   LogRecord = Read (LogFile)
   If ((LogRecord.Cs-url-stem <> gif,jpeg,jpgcss,js))
     AND (LogRecord.Cs-method= ‘GET’) AND
     (LogRecord.Sc-status = (200) AND
     (LogRecord.User-agent <> Crawler, Spider,
     Robot))
Then Write (LogFile, LogRecord)
    End If
  End While
End
```

4.2. Session Identification

The session identification splits all the pages accessed by the IP address which is a unique identity and a timeout; whereby the time between page requests exceeds a certain time limit.It is assumed that the user has started a new session. 30 minutes default timeout is considered. Figure 4 shows a fragment from session identification result. The following is the algorithm [21]:

Algorithm Session den (LogFile: Web log file; SessionFile: Session files)

```
Begin
```

```
SessionSet={}
UserSet = {}
k=0
While not eof (LogFile) Do
  LogRecord=Read (LogFile)
If (LogRecord.time-taken>30min OR
      LogRecord.UserID not in UserSet)
  Then
     k = k+1
     Sk = LogRecord.Url
     SessionSet = SessionSet U {Sk}
Write (SessionFile, SessionSet)
   End If
  End While
End
```

5. Experimental Setup and Result

We run our experiments on a system with a 2.20 GHz Intel(R) Core(TM) i-7 processor and 8 GB DDR RAM running Windows 7 Home Premium. All the preprocessing was done using MATLAB. The web access log was collected from the web server at NASA Kennedy space Centre [22] form 00:00:00 Aug 1, 1995 through 23:59:59 Aug 31, 1995, a total of 31 days. In this period there are totally 1,569,898 requests recorded by the log file. Our implementation is developed using MATLAB.

5.1. Data Cleaning

At the end of the data cleaning step, The result files contained robot 28,180 records, Multimedia objects 997,568 records, corrupt requests 8 records, failed requests 173,417 records and the cleaned log file contains 415,017 records instead of the initial 1,569,898 records, which means that the size of the file was reduced to 26.44% of its initial size. Table2 shows the data cleaning screen after the first step of preproessing.

Table 2. Results of Preprocessed Data Cleaning

Statistics	Number of record
Original Size	1,569,898
Satisfied Requests	1,396,473
Corrupt Requests	8
Failed Requests	173,417
Multimedia Objects	997,568
Other requests	102,608
Robot Requests	28,180
Web pages	469,714
Cleaned web pages	415,017
Reduced Size	415,017
Percentage in reduction	26.44

Table 3. Information extracted in the data cleaning module

Request category	Number of record
Multimedia Objects	997,568
.jpg	29,963
.ico	0
.jpeg	4,151
.gif	945,826
.png	0
.bmp	33
.mp3	0
.wav	1,635
.avi	1
.mpeg	0
.wmv	0
.mid	0
.tif	0
.swf	0
.ram	0
.rm	0
.mpg	15,771
.map	123
.zip	27
.perl	38
Others	
.PDF	70
.TXT	21,267
.DOC	6
.PL	33,396
.XBM	46,354
.COM	70
Otherwise	1,445
Methods	
-GET	156,5812
-POST	111
-HEAD	3,967

Figure 3 shows the cleaned data after the first step of preprocessing.

IP Address	Time Stamp	The requested name
'slppp6.intermind.net'	'01/Aug/1995:00:00:10'	'/history/skylab/skylab.html'
'133.43.96.45'	'01/Aug/1995:00:00:16'	'/shuttle/missions/sts-69/mission-sts-69.html'
'kgtyk4.kj.yamagata-u.ac.jp'	'01/Aug/1995:00:00:17'	'/'
'd0ucr6.fnal.gov'	'01/Aug/1995:00:00:19'	'/history/apollo/apollo-16/apollo-16.html'
'ix-esc-ca2-07.ix.netcom.com'	'01/Aug/1995:00:00:19'	'/shuttle/resources/orbiters/discovery.html'
'www-c8.proxy.aol.com'	'01/Aug/1995:00:00:24'	'/shuttle/countdown/'
'slppp6.intermind.net'	'01/Aug/1995:00:00:32'	'/history/skylab/skylab-1.html'
'uplherc.upl.com'	'01/Aug/1995:00:00:43'	'/shuttle/missions/sts-71/mission-sts-71.html'
'133.43.96.45'	'01/Aug/1995:00:00:46'	'/shuttle/resources/orbiters/endeavour.html'
'uplherc.upl.com'	'01/Aug/1995:00:00:55'	'/shuttle/resources/orbiters/atlantis.html'
'uplherc.upl.com'	'01/Aug/1995:00:01:13'	'/shuttle/resources/orbiters/challenger.html'
'uplherc.upl.com'	'01/Aug/1995:00:01:17'	'/history/apollo/apollo-17/apollo-17.html'

Figure 3. A fragment from of the cleaned data.

5.2. Session Identification

The session identification splits all the pages accessed by the IP address which is a unique identity and a timeout, It is the time between page requests exceeds a certain limit, it is assumed that the user has started a new session. 30 minutes default timeout is considered. Figure 4 shows a fragment from session identification result. A fraction of log files with user sessions shown in figure 4. There were a total of 71,242 unique visiting IP addresses, 935 unique pages and 130,976 sessions.

Figure 5: shows the session length distribution for the NASA dataset after the session is identified. The vertical axis stands for the percentage of occurrence of the number of session length. The horizontal axis is marked with the length of session. The figure shows, for each session length, the statistical distribution of the session having that many consecutive requests on the server.

ID	IP Address	Time Stamp	The requested resource name
1	'***.novo.dk'	'09/Aug/1995 :03:02:48'	'/shuttle/missions/sts-69/mission-sts-69.html'
1	'***.novo.dk'	'09/Aug/1995 :03:03:52'	'/shuttle/countdown/'
1	'***.novo.dk'	'09/Aug/1995 :03:05:38'	'/shuttle/countdown/liftoff.html'
1	'***.novo.dk'	'09/Aug/1995 :03:07:40'	'/shuttle/countdown/lps/fr.html'
2	'@.scimaging.com'	'31/Aug/1995 :01:57:29'	'/software/winvn/winvn.html'
3	'001.msy4.communique.net'	'30/Aug/1995 :02:55:47'	'/software/winvn/winvn.html'
4	'007.thegap.com'	'09/Aug/1995 :15:36:28'	'/shuttle/countdown/'
5	'01.ts01.zircon.net.au'	'21/Aug/1995 :04:46:06'	'/facts/faq04.html'
6	'01-dynamic-c.wokingham.luna.net'	'27/Aug/1995 :20:08:50'	'/shuttle/countdown/countdown.html'
6	'01-dynamic-c.wokingham.luna.net'	'27/Aug/1995 :20:12:59'	'/shuttle/countdown/count.html'
6	'01-dynamic-c.wokingham.luna.net'	'27/Aug/1995 :20:15:45'	'/shuttle/countdown/lps/bkup-intg/bkup-intg.html'
6	'01-dynamic-c.wokingham.luna.net'	'27/Aug/1995 :20:16:15'	'/shuttle/countdown/lps/ab/ab.html'
6	'01-dynamic-c.wokingham.luna.net'	'27/Aug/1995 :20:18:34'	'/shuttle/countdown/liftoff.html'
6	'01-dynamic-c.wokingham.luna.net'	'27/Aug/1995 :20:36:38'	'/shuttle/countdown/lps/fr.html'
6	'01-dynamic-c.wokingham.luna.net'	'27/Aug/1995 :20:37:22'	'/facilities/lcc.html'
6	'01-dynamic-c.wokingham.luna.net'	'27/Aug/1995 :20:37:41'	'/facilities/tour.html'

Figure 4. A fragment from session identification result.

As a session becomes longer, the percentage of web sessions decrease dramatically. The short sessions take a large proportion of the whole data set. This chart shows that a significant number of sessions only consist of one or two request and there are not too many web user sessions which extend over 10 visits thus the average session length is rather low. This is because most of web users are casual users who access only a few pages then leave. However, there is still a sizable of requests for sessions with lengths greater than three.

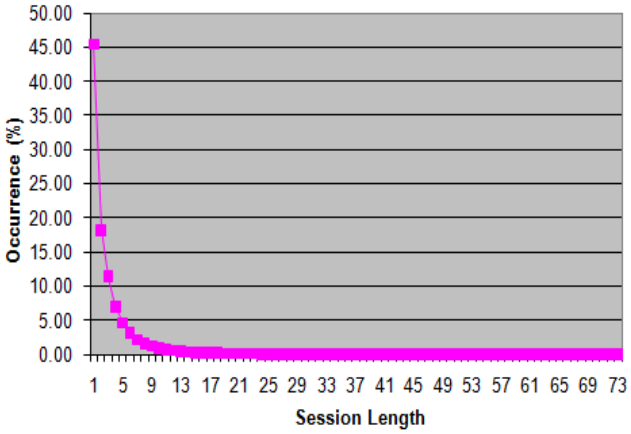

Figure 5. The session length distribution of data set

Figure 6 shows the relationship between the frequency of each page visited by the user and the page ID, Each web page has different number of times accessed by the user. Since each web page is not of the same interest and users always make the similar transaction, Thus, The user's access history can be used to use for mining user access patterns.

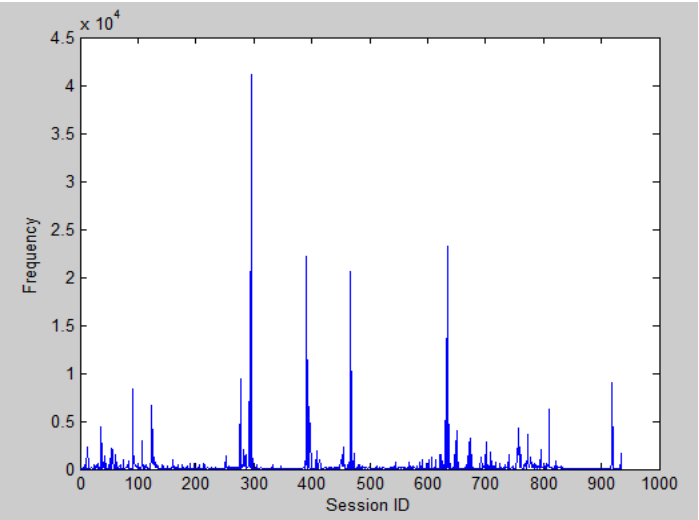

Figure 6. A Frequency chart for the frequency visited sessions.

The top 10 of the most frequency visited pages showed in table 5.

Table 5. The most frequency Top 10 sessions visited

Session ID	Session	number
295	'/ksc.html'	41109
634	'/shuttle/missions/sts-69/mission-sts-69.html'	23203
391	'/shuttle/countdown/'	22128
467	'/shuttle/missions/missions.html'	20613
277	'/history/history.html'	9412
919	'/software/winvn/winvn.html'	8973
91	'/history/apollo/apollo.html'	8297
292	'/images/'	7039
124	'/history/apollo/apollo-13/apollo-13.html'	6661
396	'/shuttle/countdown/liftoff.html'	6571

In the following experiments will be used for mining user access pattern in a future.

6. Conclusion

Data preprocessing is one of the important and prerequisite phases in WUM. This paper presents a brief introduction to WUM, apart from the data mining technologies and also the implementation of the preprocessing of web log files in NASA's web server. This study focuses on methods that can be used for the task of session identification from web log files. The work in this study also produces statistical information of user session.

After preprocessing is completed, the result will be used for mining user access pattern, the future work involves various data transformation tasks that are likely to influence the quality of the discovered patterns resulting from the mining techniques like Association, Clustering, and classificationsthat may be applied only on to a group of sessions according to assumptions of users' intentions.

Acknowledgment

The authors gratefully thank the anonymous referees and collaborators for their substantive suggestions. We also acknowledge research support from Rangsit University.

References

[1] C.P. SUmATHI, R.P.V.a.T.S., "AN OVERVIEW OF PREPROCESSING OF WEB LOG FILES FOR WEB USAGE MINING,"Journal of Theoretical and Applied Information Technology, Vol. 34, No. 1, 2011.

[2] Bhaskaran, V.S.a.V.M., "Data Preparation Techniques for Web Usage Mining in World Wide Web-An Approach,"International Journal of Recent Trends in Engineering, Vol 2, No.4, 2009.

[3] C, G., M. M, and D. K, "Preprocessing of Web Log Files in Web Usage Mining,"The Icfai Journal of Information Technology, 35J-2008-03-06-01(35J-2008-03-06-01): pp. 55-66, 2008.

[4] Mohd Helmy Abd Wahab, M.N.H.M., Hafizul Fahri Hanafi,Mohamad Farhan and Mohamad Mohsin."Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm,"World Academy of Science, Engineering and Technology 48, 2008.

[5] L.K. Joshila Grace, V.M.a.D.N., "ANALYSIS OF WEBLOGS AND WEB USER IN WEB MINING," International Journal of Network Security & Its Applications(IJNSA), Vol.3, No.1, 2011.

[6] K, D.a.V.v.M., "Session Reconstruction in Web Usage Analysis," The Icfai Journal of Information Technology, 2008.

[7] R.Krishnamoorthi, K.R.S.a.,"Identifying User Behavior by Analyzing Web Server Access Log File," IJCSNS International Journal of Computer Science and Network Security, Vol.9, No. 4, 2009.

[8] V.Chitraa, A.S.D., "A Survey on Preprocessing Methods for Web Usage Data," (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010.

[9] Arshi Shamsi, R.N., Pankj Pratap Singh and Mahesh Kumar Tiwar,"Web Usage Mining by Data Preprocessing," International Journal of Computer Science And Technology, Vol.3, No. 1, 2012.

[10] R, C.R.S.J.a.D., "Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data," Ph.D Thesis, University of Minnesota, 2000.

[11] Aye.T.T, "Web Log Cleaning for Mining of Web Usage Patterns," International Conference on computer Research and Development, IEEE, DOI:, 2011. 10.1109/ICCRD.2011.5764181.

[12] Sanjay Bapu Thakare et al., "An Effective and Complete Preprocessing for Web Usage Mining," International Journal on Computer Science and Engineering, Vol. 2, No. 3, p. 848 - 851, 2010.

[13] Bhaskaran, S.V.a.M.,"Data Preprocessing Techniques for Pre-Fetching and Caching of Web Data trhough Poxy Server," International Journal of Computer Science and Network Security, Vol. 11, No.11, pp.92-98, 2011.

[14] G, R.C.a.K., "An Efficient Preprocessing Methodoloy for Discovering Patterns and Clustering of Web Users using a Dynamic ART1 Neural Network," Fifth International Conference on Information Processing, 2011. Springer-Verlag.

[15] Maheswara Rao.V.V.R and Valli Kumari.V, "An Enhanced Pre-Processing Research Framework for Web Log Data Using a Learning Algorithm," Computer Science and Information Technology, DOI:, pp. 1-15, 2011. 10.5121/csit.2011.1101.

[16] format, W.W.W.c.t.C.L.F., http://www.w3.org/Daemon/User/Config/.Logging.html#common-logfile-format,1995, 1995.

[17] Overview, I.H.T.P., http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html.

[18] Mitharam, M.D., "Preprocessing in Web Usage mining," International Journal of Scientific & Engineering research, Vol.3, No.2, 2012.

[19] Salamn, S.E., "Web Server Logs Preprocessing for Web Intrution Detection," Computer and Information Science, Vol.4,No.4, 2011.

[20] CHIMPHLEE, S., "Predicting Next Page Access by Markov Models,"THESIS Ph.d., pp. 41-54, 2006.

[21] F.Yuan, L.W.a.G.Y., "Study on Data Preprocessing Algorithm in Web Log Mining," Proceeding of the Second International Conference on Machine Learning and Cybenetics, 2003.

[22] http://ita.ee.lbl.gov/html/contrib/NASA-HTTP.html, *NASA Logs Files.*