Comparison of UWAD Tool with Other Tools Used for Preprocessing

Nirali Honest
Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Charotar University of Science and Technology (CHARUSAT), Changa, India.

Dr. Bankim Patel
Shrimad Rajchandra Institute of Management and Computer Application Uka Tarsadia University, Bardoli, India.

Dr. Atul Patel
Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, Charotar University of Science and Technology (CHARUSAT), Changa, India.

Abstract

The purpose of Preprocessing phase is to produce accurate data with speed and accuracy. Interested patterns can be discovered based on the quality of data generated after the preprocessing phase. Designing this phase taken maximum effort in the entire process of Web Usage Mining (WUM), as it focuses on reducing the quantity of data but not compromising with the quality of data. The paper focuses the implementation concerns of preprocessing phase using tool Log parser lizard and focuses on the problems faced in using the tool. It also shows the usefulness of University Website Access Domain (UWAD) tool.

1. Introduction

Due to huge information and lack of unified structure information retrieval is difficult. Most users may not have good knowledge of the structure of the information network, and may easily get fed up by taking many access hops and losing impatience when waiting for the information [1]. Web is the single largest data source in the world, due to heterogeneity and lack of structure of web data, mining is a challenging task [2]. Preprocessing of log fie is complex and laborious job and it takes 80% of the total time of web usage mining process as whole [3]. We cannot negate the importance of preprocessing step in web usage mining. Paying due attention to preprocessing step, improves the quality of data [4], furthermore, preprocessing improves the efficiency and effectiveness of other two steps of WUM such as pattern discovery and pattern analysis.Information about internet user is stored in different raw log files. Doru Tanasa, et al. [5] focus on web server logs from several web sites, generally belonging to the same organization. An important organization might have several web servers for its web sites. Fang Yuan, et al. [6] mainly focus on analyzing visiting information from logged data in order to extract usage pattern, which can be classified on to three categories: similar user group, relevant page group and frequency accessing paths.

Web Server logs are plain text (ASCII) files, that is independent from the server platform. Mohd Helmy Abd Wahab, et al. [7] discusses on the types of logs, but traditionally there are four types of server logs:

- Transfer Log
- Error Log
- Agent Log
- Referrer Log

Each HTTP protocol transaction, weather completed or not, is recorded in the logs and some transactions are recorded in more than one log. Transfer log and error log are standard. The referrer and agent logs may or may not be "turned on" at the server or may be added to the transfer log file to create an "extended" log file format. Currently, there are three formats available to record log files:-

- W3C Extended Log files Format
- Microsoft IIS Log File
- NCSA Common Log files Format

The W3C Extended log file format, Microsoft IIS log file format, and NCSA log file format are all ASCII text formats. This proposed research assumes that server uses W3C Extended Log File Format to record log files

This proposed research will be experimented based on following parameters.

- Preprocessing Accuracy
- Hit Ratio
- Bandwidth usage
- Access pattern for particular events
- Ease of Use

2. Comparison

The initial testing of the below algorithms were carried out using Log Parser Lizard tool, which supports cleaning files, exporting files, writing queries and generating reports. But the drawback

with the tool is that you need to write queries for all the operations and generate the intermediate results and also need to use MS Excel to perform other utility tasks, which is very tedious and prone to error. Customization of reports based on the molded mining process is not available. Another objective is to generate per page frequency but the pages designed with CMS are generated with page ID and not page Name, so biding of ID and Name is not done by Log Parser Lizard, certain tools like state counter is a tool which support the generation of per page frequency but it generates report with page ID and not with name so it may not be informative. So taking these points into consideration we have prepared our own tool for our mining process namely UWAD (University Website Access Domain), and now the above algorithm steps are implemented using this tool. Below listing shows the algorithm used and the queries written for achieving certain tasks. The next section shows the results generated using this UWAD tool. Figure 1 shows the use of tool for selecting the file to pre-process and after that the summary of details of the cleaning process. Table3 shows the summary of the Data Cleaning process. Figure 2 reflects the cleaned log data.

2.1 Data Cleaning

a) Download the W3C Extended Log file from internet.
b) Parse the raw log file according to delimiter (space) and convert it to appropriate fields of W3C Extended log file format.
c) Remove all other entries which have other then .html, .asp,.aspx,.php extensions. These also includes log entries which do not have any URL in the URL entry.
d) Remove log entries having code other then 200 and 304 from the file.
e) Remove entries with request methods except GET and POST.
f) Remove web crawlers, robots, Spiders.

For testing this steps b to d Log Parser Lizard tool is used by giving the appropriate queries.

a) Download the W3C Extended Log file from internet. The sample file is a log of 28th November,2012 of charusat website.

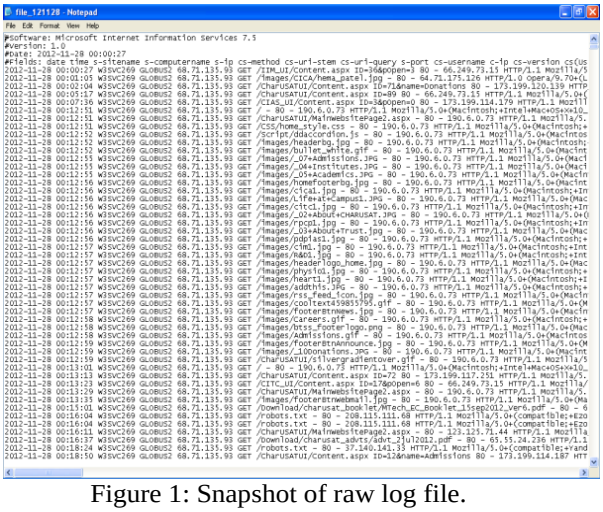

Figure 1: Snapshot of raw log file.

b) Parse the raw log file according to delimiter (space) and convert it to appropriate fields of W3C Extended log file format. For testing this step Log Parser Lizard is used and below query is executed.

Select * from C:\logs\file_121128.log;

LogFilename	RowNumber	date	time	s-sitename	s-computername	s-ip	cs-method	cs-uri-stem
C:\logs\file_121128.log	5	11/28/2012	0:00:27	W3SVC269	GLOBUS2	68.71.135.93	GET	/IIM_UI/Content.aspx
C:\logs\file_121128.log	6	11/28/2012	0:01:05	W3SVC269	GLOBUS2	68.71.135.93	GET	/images/CICA/hema_patel.jpg
C:\logs\file_121128.log	7	11/28/2012	0:02:04	W3SVC269	GLOBUS2	68.71.135.93	GET	/CharUSATUI/Content.aspx
C:\logs\file_121128.log	8	11/28/2012	0:05:17	W3SVC269	GLOBUS2	68.71.135.93	GET	/CharUSATUI/Content.aspx
C:\logs\file_121128.log	9	11/28/2012	0:07:36	W3SVC269	GLOBUS2	68.71.135.93	GET	/CIAS_UI/Content.aspx
C:\logs\file_121128.log	10	11/28/2012	0:12:51	W3SVC269	GLOBUS2	68.71.135.93	GET	/
C:\logs\file_121128.log	11	11/28/2012	0:12:51	W3SVC269	GLOBUS2	68.71.135.93	GET	/CharUSATUI/MainWebsitePage2.a
C:\logs\file_121128.log	12	11/28/2012	0:12:51	W3SVC269	GLOBUS2	68.71.135.93	GET	/CSS/home_style.css
C:\logs\file_121128.log	13	11/28/2012	0:12:52	W3SVC269	GLOBUS2	68.71.135.93	GET	/Script/ddaccordion.js
C:\logs\file_121128.log	14	11/28/2012	0:12:52	W3SVC269	GLOBUS2	68.71.135.93	GET	/images/headerbg.jpg
C:\logs\file_121128.log	15	11/28/2012	0:12:52	W3SVC269	GLOBUS2	68.71.135.93	GET	/images/bullet_white.gif

(Rows:26380 Time taken: 00:02:29)

Figure 2: Snapshot of log file separated in different fields.

Steps c and d are performed together. Remove all other entries which have other then .html, .asp,.aspx,.php extensions and also Remove log entries having code other then 200 and 304 from the file.

select LogFilename,date,time,cs-method,cs-uri-tem,cs-uri-query,c-ip,s-ip,sc-status,time-taken,s-port,cs-version, **cs**(User-Agent),sc-status,sc-bytes,cs-bytes, time-taken **FROM** C:\logs\file1.csv**where** cs-uri-stem like '%.htm' and (sc-status=200 or sc-status=304) or (cs-uri-stem like '%.asp' and (sc-status=200 or sc-status=304)) or(cs-uri-stem like '%.php' and (sc-status=200 or sc-status=304))or (cs-uri-stem like '%.aspx' and (sc-status=200 or sc-status=304))

LogFilename	RowNumber	date	time	s-sitename	s-computername	s-ip	cs-method	cs-uri-stem
C:\logs\file_121128.log	5	11/28/2012	0:00:27	W3SVC269	GLOBUS2	68.71.135.93	GET	/IIM_UI/Content.aspx
C:\logs\file_121128.log	6	11/28/2012	0:01:05	W3SVC269	GLOBUS2	68.71.135.93	GET	/images/CICA/hema_patel.jpg
C:\logs\file_121128.log	7	11/28/2012	0:02:04	W3SVC269	GLOBUS2	68.71.135.93	GET	/CharUSATUI/Content.aspx
C:\logs\file_121128.log	8	11/28/2012	0:05:17	W3SVC269	GLOBUS2	68.71.135.93	GET	/CharUSATUI/Content.aspx
C:\logs\file_121128.log	9	11/28/2012	0:07:36	W3SVC269	GLOBUS2	68.71.135.93	GET	/CIAS_UI/Content.aspx
C:\logs\file_121128.log	10	11/28/2012	0:12:51	W3SVC269	GLOBUS2	68.71.135.93	GET	/
C:\logs\file_121128.log	11	11/28/2012	0:12:51	W3SVC269	GLOBUS2	68.71.135.93	GET	/CharUSATUI/MainWebsitePage2.:
C:\logs\file_121128.log	12	11/28/2012	0:12:51	W3SVC269	GLOBUS2	68.71.135.93	GET	/CSS/home_style.css
C:\logs\file_121128.log	13	11/28/2012	0:12:52	W3SVC269	GLOBUS2	68.71.135.93	GET	/Script/ddaccordion.js
C:\logs\file_121128.log	14	11/28/2012	0:12:52	W3SVC269	GLOBUS2	68.71.135.93	GET	/images/headerbg.jpg
C:\logs\file_121128.log	15	11/28/2012	0:12:52	W3SVC269	GLOBUS2	68.71.135.93	GET	/images/bullet_white.gif

(Rows:2498 Time taken: 00:00:08)

Figure 3: Snapshot of cleaning log file with particular file extension and status code.

e) Remove entries with request methods except GET and POST.

select LogFilename,date,time,cs-method,cs-uri-stem,cs-uri-query,c-ip,s-ip,sc-status,time-taken,s-port,cs-version,**cs**(User-Agent),sc-status,sc-bytes,cs-bytes,time-taken **FROM** C:\logs\file2.csv **where** cs-method like 'GET' or cs-method like 'POST'

LogFilename	RowNumber	date	time	s-sitename	s-computername	s-ip	cs-method	cs-uri-stem
C:\logs\file_121128.log	5	11/28/2012	0:00:27	W3SVC269	GLOBUS2	68.71.135.93	GET	/IIM_UI/Content.aspx
C:\logs\file_121128.log	6	11/28/2012	0:01:05	W3SVC269	GLOBUS2	68.71.135.93	GET	/images/CICA/hema_patel.jpg
C:\logs\file_121128.log	7	11/28/2012	0:02:04	W3SVC269	GLOBUS2	68.71.135.93	GET	/CharUSATUI/Content.aspx
C:\logs\file_121128.log	8	11/28/2012	0:05:17	W3SVC269	GLOBUS2	68.71.135.93	GET	/CharUSATUI/Content.aspx
C:\logs\file_121128.log	9	11/28/2012	0:07:36	W3SVC269	GLOBUS2	68.71.135.93	GET	/CIAS_UI/Content.aspx
C:\logs\file_121128.log	10	11/28/2012	0:12:51	W3SVC269	GLOBUS2	68.71.135.93	GET	/
C:\logs\file_121128.log	11	11/28/2012	0:12:51	W3SVC269	GLOBUS2	68.71.135.93	GET	/CharUSATUI/MainWebsitePage2.:
C:\logs\file_121128.log	12	11/28/2012	0:12:51	W3SVC269	GLOBUS2	68.71.135.93	GET	/CSS/home_style.css
C:\logs\file_121128.log	13	11/28/2012	0:12:52	W3SVC269	GLOBUS2	68.71.135.93	GET	/Script/ddaccordion.js
C:\logs\file_121128.log	14	11/28/2012	0:12:52	W3SVC269	GLOBUS2	68.71.135.93	GET	/images/headerbg.jpg
C:\logs\file_121128.log	15	11/28/2012	0:12:52	W3SVC269	GLOBUS2	68.71.135.93	GET	/images/bullet_white.gif

(Rows:2497 Time taken: 00:00:07)

Figure 4: Snapshot of cleaning log file for particular method Get and Post.

f) Remove web crawlers, robots, Spiders.

select LogFilename,date,time,cs-method,cs-uri-stem,cs-uri-query,c-ip,s-ip,sc-status,time-taken,s-port,cs-version,**cs**(User-Agent),sc-status,sc-bytes,cs-bytes,time-taken **FROM** C:\logs\file3.csv **where** not **cs**(User-Agent) like '%spider%'

select LogFilename,date,time,cs-method,cs-uri-stem,cs-uri-query,c-ip,s-ip,sc-status,time-taken,s-port,cs-version,**cs**(User-Agent),sc-status,sc-bytes,cs-bytes,time-taken **FROM** C:\logs\file4.csv **where** not **cs**(User-Agent) like '%crawler%'

select LogFilename,date,time,cs-method,cs-uri-stem, cs-uri-query,c-ip,s-ip,sc-status,time-taken,s-port,cs-version, **cs**(User-Agent),sc-status,sc-bytes,cs-bytes, time-taken **FROM** C:\logs\file5.csv **where** not **cs**(User-Agent) like '%robot%'

LogFilename	date	time	cs-method	cs-uri-stem	cs-uri-query
C:\logs\file_121128.log	11/28/2012	0:00:27	GET	/IIM_UI/Content.aspx	ID=36&pOpen=3
C:\logs\file_121128.log	11/28/2012	0:02:04	GET	/CharUSATUI/Content.aspx	ID=71&name=Donations
C:\logs\file_121128.log	11/28/2012	0:05:17	GET	/CharUSATUI/Content.aspx	ID=89
C:\logs\file_121128.log	11/28/2012	0:07:36	GET	/CIAS_UI/Content.aspx	ID=3&pOpen=0
C:\logs\file_121128.log	11/28/2012	0:12:51	GET	/CharUSATUI/MainWebsitePage2.aspx	
C:\logs\file_121128.log	11/28/2012	0:13:13	GET	/CharUSATUI/Content.aspx	ID=72
C:\logs\file_121128.log	11/28/2012	0:13:23	GET	/CITC_UI/Content.aspx	ID=17&pOpen=6
C:\logs\file_121128.log	11/28/2012	0:13:29	GET	/CharUSATUI/MainWebsitePage2.aspx	
C:\logs\file_121128.log	11/28/2012	0:18:50	GET	/CharUSATUI/Content.aspx	ID=12&name=Admissions
C:\logs\file_121128.log	11/28/2012	0:20:12	GET	/CharUSATUI/Content.aspx	ID=42&name=About_Trust
C:\logs\file_121128.log	11/28/2012	0:21:31	GET	/RND_UI/Content.aspx	ID=4
C:\logs\file_121128.log	11/28/2012	0:21:36	GET	/CharUSATUI/Content.aspx	ID=5&name=About_University
C:\logs\file_121128.log	11/28/2012	0:21:36	GET	/CharUSATUI/Content.aspx	ID=52

(Rows:2197 Time taken: 00:00:21)

Figure 5: Snapshot of cleaning log file with particular file extension and status code.

2.2 User Identification

a) Read record from the cleaned log file.
b) If new IP address then add new record the IP address, browser and OS details and increment the count of number of users.
c) If IP address is already present then compare the browser and OS details if not same then increment the count of number of users.

a) Read record from the cleaned log file.

Select c-ip,cs-version,**cs**(User-Agent),time **from** C:\ulogs\file6.csv

c-ip	cs-version	cs(User-Agent)
66.249.73.	HTTP/1.1	Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html)
66.249.73.	HTTP/1.1	Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html)
190.6.0.73	HTTP/1.1	Mozilla/5.0+(Macintosh;+Intel+Mac+OS+X+10_7_4)+AppleWebKit/537.11+(KHTML
66.249.73.	HTTP/1.1	Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html)
190.6.0.73	HTTP/1.1	Mozilla/5.0+(Macintosh;+Intel+Mac+OS+X+10_7_4)+AppleWebKit/537.11+(KHTML
157.55.35.	HTTP/1.1	Mozilla/5.0+(compatible;+bingbot/2.0;++http://www.bing.com/bingbot.htm)
66.249.73.	HTTP/1.1	Mozilla/5.0+(compatible;+Googlebot/2.1;++http://www.google.com/bot.html)
157.55.35.	HTTP/1.1	Mozilla/5.0+(compatible;+bingbot/2.0;++http://www.bing.com/bingbot.htm)
157.55.35.	HTTP/1.1	Mozilla/5.0+(compatible;+bingbot/2.0;++http://www.bing.com/bingbot.htm)
157.55.35.	HTTP/1.1	Mozilla/5.0+(compatible;+bingbot/2.0;++http://www.bing.com/bingbot.htm)
157.55.35.	HTTP/1.1	Mozilla/5.0+(compatible;+bingbot/2.0;++http://www.bing.com/bingbot.htm)
157.55.32.	HTTP/1.1	Mozilla/5.0+(compatible;+bingbot/2.0;++http://www.bing.com/bingbot.htm)
65.55.24.2	HTTP/1.1	Mozilla/5.0+(compatible;+bingbot/2.0;++http://www.bing.com/bingbot.htm)

(Rows:2196 Time taken: 00:00:01)

Figure 6: Snapshot of reading records from stored intermediate file.

b) If new IP address then add new record the IP address, browser and OS details and increment the count of number of users and If IP address is already present then compare the browser and OS details if not same then increment the count of number of users.

Select c-ip,cs-version,**cs**(User-Agent) **from** C:\ulogs\ulist1.csv order by c-ip

c-ip	cs-version	cs(User-Agent)	time
1.187.23.195	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1)+App	4:59:58
1.187.23.195	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1)+App	5:01:02
1.187.23.195	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1)+App	5:01:24
1.187.23.195	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1)+App	5:02:16
1.187.23.195	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1)+App	5:08:32
1.187.23.195	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1)+App	4:59:37
1.187.23.195	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1)+App	5:08:53
1.187.23.195	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1)+App	5:09:32
1.187.23.195	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1)+App	4:58:44
1.187.23.195	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1)+App	4:57:52
1.187.23.195	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1)+App	4:57:13
1.187.4.41	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1;+WO	16:09:02
1.187.4.41	HTTP/1.1	Mozilla/5.0+(Windows+NT+6.1;+WO	16:10:52

(Rows:2197 Time taken: 00:00:01)

Figure 7: Snapshot of reading records based on combination of IP, version and useragent.

Select distinct c-ip,cs-version,**cs**(User-Agent) **from** C:\ulogs\ulist1.csv

c-ip	cs(User-Agent)	cs-version
1.187.23.195	Mozilla/5.0+(Windows+NT+6.1)+App	HTTP/1.1
1.187.4.41	Mozilla/5.0+(Windows+NT+6.1;+WO	HTTP/1.1
1.23.135.238	Mozilla/5.0+(Windows+NT+6.1;+WO	HTTP/1.1
1.38.24.180	Mozilla/5.0+(SymbianOS/9.4;+Series	HTTP/1.1
1.38.24.54	Mozilla/5.0+(Linux;+U;+Android+2.3.	HTTP/1.1
1.38.24.62	Mozilla/4.0+(compatible;+MSIE+8.0;	HTTP/1.1
1.38.24.71	Mozilla/5.0+(Linux;+U;+Android+4.0.	HTTP/1.1
1.38.24.79	Mozilla/5.0+(Linux;+U;+Android+2.3.	HTTP/1.1
1.38.24.95	Mozilla/4.0+(compatible;+MSIE+8.0;	HTTP/1.1
1.38.25.103	Mozilla/5.0+(iPhone;+CPU+iPhone+	HTTP/1.1
1.38.26.126	Mozilla/5.0+(Linux;+Android+4.1.1;+	HTTP/1.1
1.38.26.5	OneBrowser/3.5/Mozilla/5.0+(Linux;+	HTTP/1.1
1.38.27.111	Mozilla/5.0+(Linux;+U;+Android+2.3.	HTTP/1.1

(Rows:352 Time taken: 00:00:00)

Figure 8: Snapshot of reading records with unique IP.

2.3 Session Identification

a) Read record from the log file.
b) If there is a new user, then there is a new session.
c) In one user session, if the refer page is null we can draw a conclusion that there is a new session.
d) If the time between the page requests exceeds a certain limits (30 Minutes), it is assumed that the user is starting a new session.

.

c-ip	cs(User-Agent)	time-taken	Total time-taken	Minutes	Minutes_MOD_30	Session_Count
1.187.23.195	Mozilla/5.0+(Windows+NT+6.1)+App	35564	130783	2	2	2
1.187.4.41	Mozilla/5.0+(Windows+NT+6.1;+WO	1120	2350	0	0	1
1.23.135.238	Mozilla/5.0+(Windows+NT+6.1;+WO	870	870	0	0	1
1.38.24.180	Mozilla/5.0+(SymbianOS/9.4;+Series	9077	15237	0	0	1
1.38.24.54	Mozilla/5.0+(Linux;+U;+Android+2.3.	13042	19181	0	0	1
1.38.24.62	Mozilla/4.0+(compatible;+MSIE+8.0;	2197	2197	0	0	1
1.38.24.71	Mozilla/5.0+(Linux;+U;+Android+4.0.	8383	8383	0	0	1
1.38.24.79	Mozilla/5.0+(Linux;+U;+Android+2.3.	5020	103474	2	2	2
1.38.24.95	Mozilla/4.0+(compatible;+MSIE+8.0;	12243	39976	1	1	1
1.38.25.103	Mozilla/5.0+(iPhone;+CPU+iPhone+	1878	15192	0	0	1
1.38.26.126	Mozilla/5.0+(Linux;+Android+4.1.1;+	1226	1976	0	0	1
1.38.26.5	OneBrowser/3.5/Mozilla/5.0+(Linux;+	3560	3560	0	0	1
1.38.27.111	Mozilla/5.0+(Linux;+U;+Android+2.3.	4478	4478	0	0	1
1.38.28.224	Mozilla/5.0+(Windows+NT+6.1;+WO	47928	47928	1	1	1
100.2.222.165	Mozilla/5.0+(Macintosh;+Intel+Mac+	284	1269	0	0	1
101.2.41.142	Mozilla/5.0+(Linux;+Android+4.0.4;+	2979	2979	0	0	1
101.63.112.164	Mozilla/5.0+(Windows+NT+6.1;+WO	1067	1067	0	0	1

Figure 9 : Snapshot of Calculating Session count

3. Experimental Results

Log parser lizard tool is very useful tool but, for performing steps with the tool we need to perform lots of intermediate tasks like, managing to write queries, storing intermediate results and performing other utility tasks using MS Excel, for cleaning a single file. All these activities include lots of extra overhead so it becomes a error prone and time consuming process. Tool. Customization based on requirement of molding process in not available and pages designed with Master Page [8][9]concept are not supported by most of the tools[10].

After using both the tools the results vary in the processing time and the elimination of redundantly writing the same queries over different files, which aids ease and speed in generating results. Table 1 show the result derived after applying the algorithm using Log parser Lizard tool.

Stage of Preprocessing	**No. of Web Objects Retrieved**
Initial size of file	12886 KB (12.5 MB)
Size of file after cleaning	608 KB
Before Cleaning	26380
After Cleaning	2226
Total time taken for data cleaning	2.41 minutes + extra overhead (writing queries, generating results, saving intermediate files, applying other utilities) (approximately 5 minutes)

(Table-1 Result Analysis of Proposed Cleaning Process)

Table 2 shows the result after applying the algorithm in the tool UWAD. It reduces time in processing the records and it saves the data in the database for further use.

Stage of Preprocessing	No. of Web Objects Retrieved
Total records in the file	26380
Total records after cleaning	2226
Total errors found in the file	3884
Total reduction in file size	91.56%
Total time taken for data cleaning	146.441 Seconds (2.44 minutes)

(Table-2 Result Analysis of Proposed Cleaning Process using UWAD Tool)

		Time Taken to Clean File	
File Name	No. of Records	Log Parsar Lizard	UWAD
A	21852	00:02:01 + overhead	2.24
B	30589	00:02:15 + overhead	2.39

*overhead : writing queries, storing intermediate data, performing other utility tasks

(Table-3 Comparison of Tools Used)

Standard as well as Visual reports can be generated using UWAD tool with customized requirements for the mining process as shown in Figures 10 and 11.

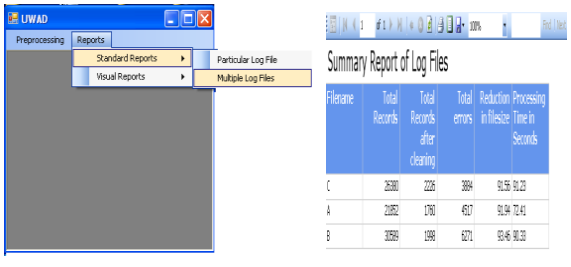

Figure 10 (a): Snapshot of selecting report.
(b): Snapshot of summary report.

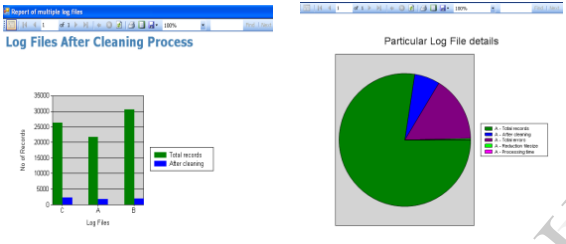

Figure 11 (a): Snapshot Cleaned log files.
(b): Snapshot of Particular file details.

Conclusion

The purpose of research focuses on the key areas like, Reduction of efforts by the automation of a number of tasks, Easier association of mining goals with the results that can be obtained and Possibility of involvement of people having less technical skills w.r.t mining. With the use of existing tools this areas are not satisfied optimally so UWAD provides the necessary foundation upon which further extended pattern based on the molded mining process can be generated. Next pattern will focus on per page frequency of pages designed using CMS and master page concept.

Acknowledgment

The authors thank Charotar University of Science and Technology (CHARUSAT) for providing necessary resources to accomplish this study.

References

[1] HAN Jia-Wei, MENG Xiao-Feng, WANG Jing etc. Research on Web Mining. Journal of Computer ResearchLkDevelopment, 2001,38(4): 405-414.
[2] Bing Liu , Web Content Mining ,The 14th International World Wide Web Conference (WWW-2005),May 10-14, 2005, Chiba, Japan.,
[3] Pabarskaite, Z. (2002). Implementing Advanced Cleaning and End-User Interpretability Technologies in Web Log Mining. 24th Int. Conf. information Technology Interfaces /TI 2002, June 24-27, 2002, Cavtat, Croatia.
[4] Han, J. and M. Kamber (2006). Data Mining: Concepts and Techniques. A. Stephan. San Francisco,, Morgan Kaufmann Publishers is an imprint of Elsevier.
[5] Doru Tanasa and Brigitte Trousse, "Advanced Data Preprocessing for Intersites Web Usage Mining ", IEEE Computer Society, March/April, 2004.
[6] Fang Yuan,Li-Juan Wang,Ge Yu,"Study on Data Preprocessing Algorithm in Web Log Mining ", Proceedings of the Second International Conferences on Machine Learning and Cybernetics, Xi'an, 2-5 November 2003.
[7] Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, et. Al, "Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm" , World Academy of Science, Engineering and Technology, 2008.
[8] Master Page Architecture and working, found at,http://msdn.microsoft.com/enus/library/wtxbf3hh .aspx
[9] Master Page Information, found at , http://www.w3schools.com/aspnet/aspnet_masterpages.asp
[10]Nirali Honest, Dr. Bankim Patel, Dr. Atul Patel, Article: Sessionization Process for the Pages Designed with the Concept of CMS in the International Journal of Advanced Research in Computer Science and Software Engineering , (IJARCSSE) – ISSN: 2277 128X, Volume 3, Issue. 6, September 2013.For papers published in translation journals, please give the English citation first, followed by the original foreign-language citation [6].