Results and Discussion

4. RESULTS AND DISCUSSION

Large volumes of data are gathered automatically by the Web servers in order to access log files. The amount of information stored in the log files are growing and using them for knowledge discovery is gaining popularity, as they can provide significant and useful information. Web usage mining is a data mining technique which is used for this purpose. In this research, the problem of Web usage mining is approached through the use of clustering and classification to find the navigation patterns of users while browsing the WWW. Two popular algorithms, namely, Ant-based clustering and LCS classification are used for this purpose. Various experiments were conducted to judge the performance of the proposed system and in this chapter the results obtained are presented.

4.1. DATA ENVIRONMENT

World Wide Web today is the most important media for collecting, sharing and distributing information. In order to test the effectiveness of the proposed system, server web log data file was obtained. The system was tested with several data collected from 90 days for easy discussion, experiments projected here are from one day, that is, data collected on 29-12-2009. Experiments were conducted on a Pentium IV system with 512MB memory, running in Windows environment. The application was developed in MATLAB 7.3.

The results of this study are divided into two sections

1. Preprocessing results – This section discusses the results obtained through the four steps of preprocessing.
2. Pattern Discovery : This section discusses the access pattern and navigation pattern discovery
3. Performance Analysis: This section compares the two proposed algorithms in terms of speed in discovering the navigation pattern.

4.2. PREPROCESSING RESULTS

As mentioned in Chapter 3, the preprocessing is conducted in four steps, namely (i) Cleaning (ii) User Identification (iii) Session Identification and (iv) formatting. The initial log file consisted of 1314 log records and was read into an Excel File. The results of the intermediate steps in preprocessing as well as the clustering and classification results were exported to Excel files. This section shows screen projection of such results. A sample raw log file is shown in Figure 4.1.

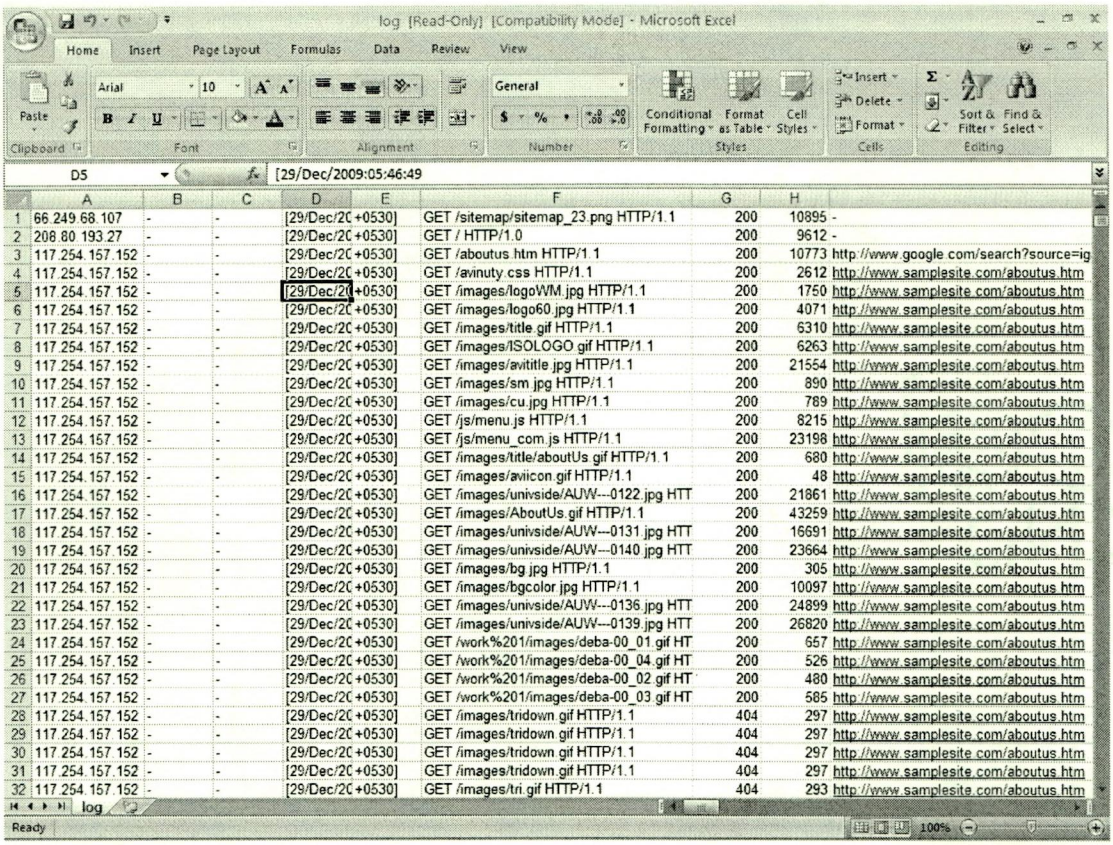
The cleaning step, which is the first step in preprocessing, Is performed to remove unwanted transactions from the log file. In this phase, three types of transactions were removed. They are

1. **Entries with unsuccessful HTTP status codes** – HTTP status codes are used to indicate the success or failure of a requested event, and we only consider successful entries with codes between 200 and 299.

2. **Entries with image and JavaScript transactions** – All web pages request images of the type gif, jpg and png formats which do not contribute anything to user's navigation pattern.

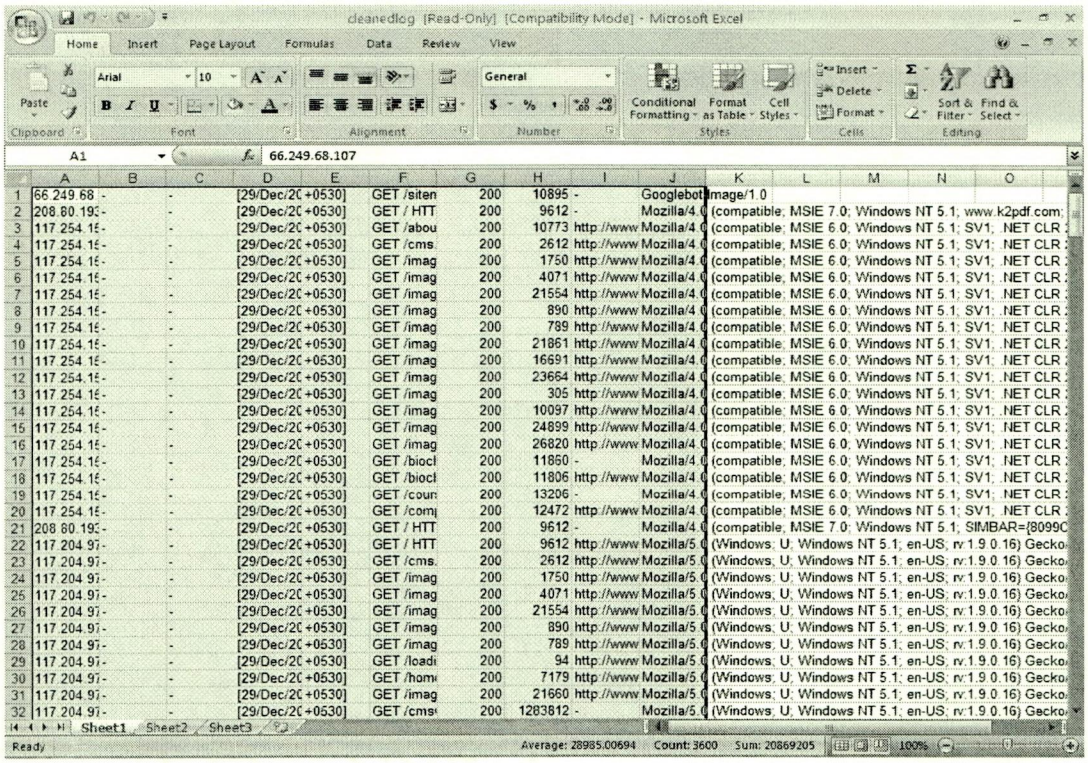
3. **Empty IP Address Field** - Some users, for security reasons, protect system. In such cases, their corresponding IP address is not recorded in the log file. As the user identification is based on IP address, these entries can be removed as they provide no details about navigation.

The resulting web log data after cleaning step is shown in Figure 4.2. It is obvious that the number of records is reduced after cleaning step. The effect of cleaning step of preprocessing on the number of transactions is shown in Figure 4.3.



	A	B	C	D	E	F	G	H
1	66.249.68.107	-	-	[29/Dec/2(	+0530]	GET /sitemap/sitemap_23.png HTTP/1.1	200	10895 -
2	208.80.193.27	-	-	[29/Dec/2(	+0530]	GET / HTTP/1.0	200	9612 -
3	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /aboutus.htm HTTP/1.1	200	10773 http://www.google.com/search?source=ig
4	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /avinuty.css HTTP/1.1	200	2612 http://www.samplesite.com/aboutus.htm
5	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/logoWM.jpg HTTP/1.1	200	1750 http://www.samplesite.com/aboutus.htm
6	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/logo60.jpg HTTP/1.1	200	4071 http://www.samplesite.com/aboutus.htm
7	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/title.gif HTTP/1.1	200	6310 http://www.samplesite.com/aboutus.htm
8	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/ISOLOGO.gif HTTP/1.1	200	6263 http://www.samplesite.com/aboutus.htm
9	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/avititle.jpg HTTP/1.1	200	21554 http://www.samplesite.com/aboutus.htm
10	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/sm.jpg HTTP/1.1	200	890 http://www.samplesite.com/aboutus.htm
11	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/cu.jpg HTTP/1.1	200	789 http://www.samplesite.com/aboutus.htm
12	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /js/menu.js HTTP/1.1	200	8215 http://www.samplesite.com/aboutus.htm
13	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /js/menu_com.js HTTP/1.1	200	23198 http://www.samplesite.com/aboutus.htm
14	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/title/aboutUs.gif HTTP/1.1	200	680 http://www.samplesite.com/aboutus.htm
15	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/avilcon.gif HTTP/1.1	200	48 http://www.samplesite.com/aboutus.htm
16	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/univside/AUW---0122.jpg HTT	200	21861 http://www.samplesite.com/aboutus.htm
17	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/AboutUs.gif HTTP/1.1	200	43259 http://www.samplesite.com/aboutus.htm
18	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/univside/AUW---0131.jpg HTT	200	16691 http://www.samplesite.com/aboutus.htm
19	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/univside/AUW---0140.jpg HTT	200	23664 http://www.samplesite.com/aboutus.htm
20	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/bg.jpg HTTP/1.1	200	305 http://www.samplesite.com/aboutus.htm
21	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/bgcolor.jpg HTTP/1.1	200	10097 http://www.samplesite.com/aboutus.htm
22	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/univside/AUW---0136.jpg HTT	200	24899 http://www.samplesite.com/aboutus.htm
23	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/univside/AUW---0139.jpg HTT	200	26820 http://www.samplesite.com/aboutus.htm
24	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /work%201/images/deba-00_01.gif HT	200	657 http://www.samplesite.com/aboutus.htm
25	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /work%201/images/deba-00_04.gif HT	200	526 http://www.samplesite.com/aboutus.htm
26	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /work%201/images/deba-00_02.gif HT	200	480 http://www.samplesite.com/aboutus.htm
27	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /work%201/images/deba-00_03.gif HT	200	585 http://www.samplesite.com/aboutus.htm
28	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/tridown.gif HTTP/1.1	404	297 http://www.samplesite.com/aboutus.htm
29	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/tridown.gif HTTP/1.1	404	297 http://www.samplesite.com/aboutus.htm
30	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/tridown.gif HTTP/1.1	404	297 http://www.samplesite.com/aboutus.htm
31	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/tridown.gif HTTP/1.1	404	297 http://www.samplesite.com/aboutus.htm
32	117.254.157.152	-	-	[29/Dec/2(	+0530]	GET /images/tri.gif HTTP/1.1	404	293 http://www.samplesite.com/aboutus.htm

Figure 4.1: Raw Web Log File



	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	66.249.68	-	-	[29/Dec/2(	+0530]	GET /siten	200	10895	-	Googlebot	Image/1.0				
2	208.80.193	-	-	[29/Dec/2(	+0530]	GET / HTT	200	9612	-	Mozilla/4.(	(compatible; MSIE 7.0; Windows NT 5.1; www.k2pdf.com;				
3	117.254.15	-	-	[29/Dec/2(	+0530]	GET /abou	200	10773	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
4	117.254.15	-	-	[29/Dec/2(	+0530]	GET /cms.	200	2612	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
5	117.254.15	-	-	[29/Dec/2(	+0530]	GET /imag	200	1750	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
6	117.254.15	-	-	[29/Dec/2(	+0530]	GET /imag	200	4071	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
7	117.254.15	-	-	[29/Dec/2(	+0530]	GET /imag	200	21554	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
8	117.254.15	-	-	[29/Dec/2(	+0530]	GET /imag	200	890	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
9	117.254.15	-	-	[29/Dec/2(	+0530]	GET /imag	200	789	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
10	117.254.15	-	-	[29/Dec/2(	+0530]	GET /imag	200	21861	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
11	117.254.15	-	-	[29/Dec/2(	+0530]	GET /imag	200	16691	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
12	117.254.15	-	-	[29/Dec/2(	+0530]	GET /imag	200	23664	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
13	117.254.15	-	-	[29/Dec/2(	+0530]	GET /imag	200	305	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
14	117.254.15	-	-	[29/Dec/2(	+0530]	GET /imag	200	10097	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
15	117.254.15	-	-	[29/Dec/2(	+0530]	GET /imag	200	24899	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
16	117.254.15	-	-	[29/Dec/2(	+0530]	GET /imag	200	26820	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
17	117.254.15	-	-	[29/Dec/2(	+0530]	GET /bioc	200	11860	-	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
18	117.254.15	-	-	[29/Dec/2(	+0530]	GET /bioc	200	11806	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
19	117.254.15	-	-	[29/Dec/2(	+0530]	GET /cour	200	13206	-	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
20	117.254.15	-	-	[29/Dec/2(	+0530]	GET /com	200	12472	http://www	Mozilla/4.(	(compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR				
21	208.80.193	-	-	[29/Dec/2(	+0530]	GET / HTT	200	9612	-	Mozilla/4.(	(compatible; MSIE 7.0; Windows NT 5.1; SIMBAR={8099C				
22	117.204.97	-	-	[29/Dec/2(	+0530]	GET / HTT	200	9612	http://www	Mozilla/5.(	(Windows; U; Windows NT 5.1; en-US; rv:1.9.0.16) Gecko/				
23	117.204.97	-	-	[29/Dec/2(	+0530]	GET /cms.	200	2612	http://www	Mozilla/5.(	(Windows; U; Windows NT 5.1; en-US; rv:1.9.0.16) Gecko/				
24	117.204.97	-	-	[29/Dec/2(	+0530]	GET /imag	200	1750	http://www	Mozilla/5.(	(Windows; U; Windows NT 5.1; en-US; rv:1.9.0.16) Gecko/				
25	117.204.97	-	-	[29/Dec/2(	+0530]	GET /imag	200	4071	http://www	Mozilla/5.(	(Windows; U; Windows NT 5.1; en-US; rv:1.9.0.16) Gecko/				
26	117.204.97	-	-	[29/Dec/2(	+0530]	GET /imag	200	21554	http://www	Mozilla/5.(	(Windows; U; Windows NT 5.1; en-US; rv:1.9.0.16) Gecko/				
27	117.204.97	-	-	[29/Dec/2(	+0530]	GET /imag	200	890	http://www	Mozilla/5.(	(Windows; U; Windows NT 5.1; en-US; rv:1.9.0.16) Gecko/				
28	117.204.97	-	-	[29/Dec/2(	+0530]	GET /imag	200	789	http://www	Mozilla/5.(	(Windows; U; Windows NT 5.1; en-US; rv:1.9.0.16) Gecko/				
29	117.204.97	-	-	[29/Dec/2(	+0530]	GET /loadi	200	94	http://www	Mozilla/5.(	(Windows; U; Windows NT 5.1; en-US; rv:1.9.0.16) Gecko/				
30	117.204.97	-	-	[29/Dec/2(	+0530]	GET /hom	200	7179	http://www	Mozilla/5.(	(Windows; U; Windows NT 5.1; en-US; rv:1.9.0.16) Gecko/				
31	117.204.97	-	-	[29/Dec/2(	+0530]	GET /imag	200	21660	http://www	Mozilla/5.(	(Windows; U; Windows NT 5.1; en-US; rv:1.9.0.16) Gecko/				
32	117.204.97	-	-	[29/Dec/2(	+0530]	GET /cms	200	1283812	-	Mozilla/5.(	(Windows; U; Windows NT 5.1; en-US; rv:1.9.0.16) Gecko/				

Figure 4.2: After Cleaning Step

The next step comprises the identification of users from their IP addresses (Figure 4.4). During user identification, all the requests entries made through search engines, request from web bots or robots are ignored. As mentioned previously, this step even though identifies users effectively, has the drawback of repeated users IP addresses. The reason behind this is that some users might be browsing over several sessions and therefore, the log file at this stage has an entry for each session, which results as unique user ID in ant-based clustering.

To avoid this, the preprocessing step identifies the user and the user sessions. In this step, all the entries which have been in the same web page for more than 30 minutes were considered as genuine users. The sample file after session identification is shown Figure 4.5 and the number of entries in the log file before and after user and session identification is shown in Figure 4.6.

The last step, (i.e.,), the formatting of the log data resulted from the previous step, consists of assigning numeric codes to the URLs visited. For this initially, unique URLs from the data from Figure 4.5 were found (Figure 4.7). After identifying unique URLs, they were assigned numerical codes. In the web data taken for experimentation, 25 unique URLs were identified which were assigned codes from 1- 25 (Figure 4.8).

The final formatted web log data after preprocessing is shown in Figure 4.9. This data is taken as input by ant-based clustering process. The results obtained during clustering are discussed in the next section.

4.3. CLUSTERING RESULTS

Ant Based clustering is an application for clustering similar interested users into a single class. The result after such grouping is shown in Figure 4.10. Careful analysis of this data reveals two serious drawbacks.

(i) Same user belongs to more than one cluster

(ii) Clusters having the same user details multiple times

Both the problems are solved when ant-based clustering is following by the LCS algorithm.

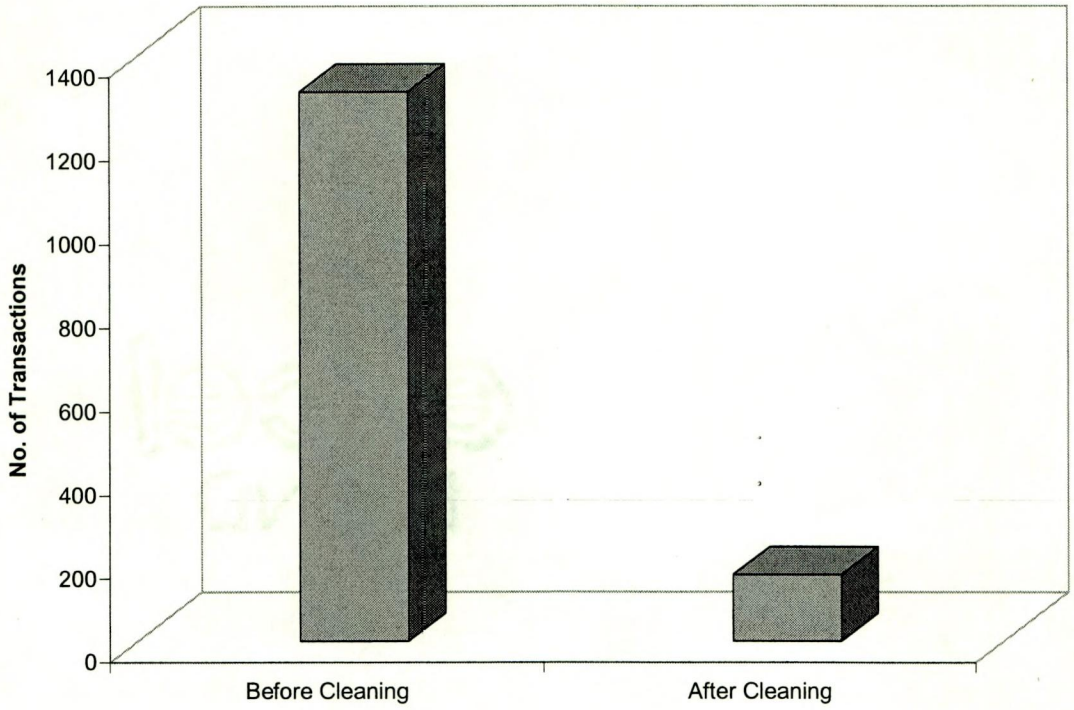


Figure 4.3: Effect of cleaning step on raw web log file

Microsoft Excel - 02uss.xls

File Edit View Insert Format Tools Data Window Help

Type a question for help

Arial 10 B I U 100%

B35 http://www.google.co.in/search?hl=en&source=hp&q=SS+university+coimbatore&meta=&aq=7&oq=AUW+

	A	B	C	D	E
1	66.249.68.107	-	[29/Dec/2009:05:07:13		
2	208.80.193.27	-	[29/Dec/2009:05:15:44		
3	117.254.157.152	http://www.google.com/search?source=ig&hl=en&rlz=1G1GGLQ_ENIN360&q=university&aq=f&oq=&aqi=	[29/Dec/2009:05:46:36		
4	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:46:43		
5	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:46:49		
6	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:46:49		
7	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:46:49		
8	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:46:58		
9	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:46:58		
10	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:47:30		
11	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:47:35		
12	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:47:35		
13	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:47:46		
14	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:47:46		
15	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:47:42		
16	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:47:45		
17	117.254.157.152	-	[29/Dec/2009:05:48:22		
18	117.254.157.152	http://www.samplesite.ac.in/biochemistry.htm	[29/Dec/2009:05:48:30		
19	117.254.157.152	-	[29/Dec/2009:05:50:42		
20	117.254.157.152	http://www.samplesite.ac.in/course.htm	[29/Dec/2009:05:52:03		
21	208.80.193.54	-	[29/Dec/2009:06:13:20		
22	117.204.97.156	http://www.careerforum.in/mba_infoupdate09/mat09/listinst.htm?reloaded=true	[29/Dec/2009:06:31:01		
23	117.204.97.156	http://www.samplesite.ac.in/	[29/Dec/2009:06:31:01		
24	117.204.97.156	http://www.samplesite.ac.in/	[29/Dec/2009:06:31:01		
25	117.204.97.156	http://www.samplesite.ac.in/	[29/Dec/2009:06:31:02		
26	117.204.97.156	http://www.samplesite.ac.in/	[29/Dec/2009:06:31:02		
27	117.204.97.156	http://www.samplesite.ac.in/	[29/Dec/2009:06:31:02		
28	117.204.97.156	http://www.samplesite.ac.in/	[29/Dec/2009:06:31:02		
29	117.204.97.156	http://www.samplesite.ac.in/	[29/Dec/2009:06:31:02		

Sheet1 Sheet2 Sheet3

Draw AutoShapes

Ready

Figure 4.4: User Identification

Microsoft Excel - 03ps.xls

File Edit View Insert Format Tools Data Window Help

B42 http://www.samplesite.ac.in/

	A	B	C	D	E	F	G
1	66.249.68.107	-	[29/Dec/2009:05:07:13				
2	208.80.193.27	-	[29/Dec/2009:05:15:44				
3	117.254.157.152	http://www.google.com/search?source=ig&hl=en&rlz=1G1GGLQ_ENIN360&q=a	[29/Dec/2009:05:46:36				
4	117.254.157.152	http://www.samplesite.ac.in/aboutus.htm	[29/Dec/2009:05:46:43				
5	117.254.157.152	-	[29/Dec/2009:05:48:22				
6	117.254.157.152	http://www.samplesite.ac.in/biochemistry.htm	[29/Dec/2009:05:48:30				
7	117.254.157.152	-	[29/Dec/2009:05:50:42				
8	117.254.157.152	http://www.samplesite.ac.in/course.htm	[29/Dec/2009:05:52:03				
9	208.80.193.54	-	[29/Dec/2009:06:13:20				
10	117.204.97.156	http://www.careerforum.in/mba_infoupdate09/mat09/listinst.htm?reloaded=true	[29/Dec/2009:06:31:01				
11	117.204.97.156	http://www.samplesite.ac.in/	[29/Dec/2009:06:31:01				
12	117.204.97.156	-	[29/Dec/2009:06:31:02				
13	66.249.68.107	-	[29/Dec/2009:07:55:24				
14	8.21.4.254	-	[29/Dec/2009:08:24:27				
15	192.55.54.36	http://www.google.co.in/search?hl=en&source=hp&q=SS+university+coimbatore	[29/Dec/2009:08:24:28				
16	192.55.54.36	http://www.samplesite.ac.in/	[29/Dec/2009:08:24:29				
17	192.55.54.36	http://www.samplesite.ac.in/loading.swf	[29/Dec/2009:08:24:37				
18	192.55.54.36	-	[29/Dec/2009:08:25:25				
19	192.55.54.36	http://www.samplesite.ac.in/courseP.htm	[29/Dec/2009:08:25:26				
20	192.55.54.36	-	[29/Dec/2009:08:27:35				
21	192.55.54.36	http://www.samplesite.ac.in/index.htm	[29/Dec/2009:08:27:53				
22	192.55.54.36	http://www.samplesite.ac.in/admission.htm	[29/Dec/2009:08:27:54				
23	192.55.54.36	-	[29/Dec/2009:08:29:30				
24	67.195.112.180	-	[29/Dec/2009:08:42:24				
25	59.92.110.200	http://www.google.co.in/url?sa=t&source=web&ct=res&cd=2&ved=0CA4QFjAB&	[29/Dec/2009:10:13:45				
26	59.92.110.200	http://www.samplesite.ac.in/	[29/Dec/2009:10:13:48				
27	59.92.110.200	http://www.samplesite.ac.in/contactus.htm	[29/Dec/2009:10:14:06				
28	59.92.110.200	http://www.samplesite.ac.in/contactusheadofthedepartment.htm	[29/Dec/2009:10:14:30				
29	59.92.110.200	http://www.samplesite.ac.in/contactusofficebearers.htm	[29/Dec/2009:10:14:36				

Sheet1 Sheet2 Sheet3

Draw AutoShapes

Ready

Figure 4.5: User Session Identification

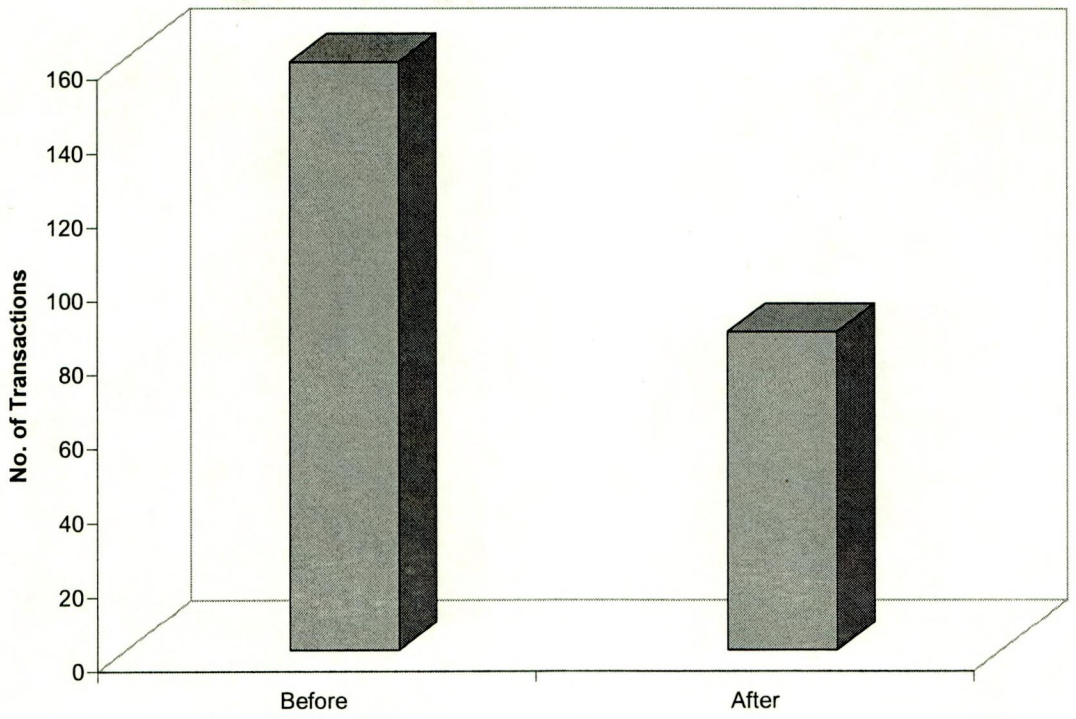


Figure 4.6: Effect of user and session identification on web log data

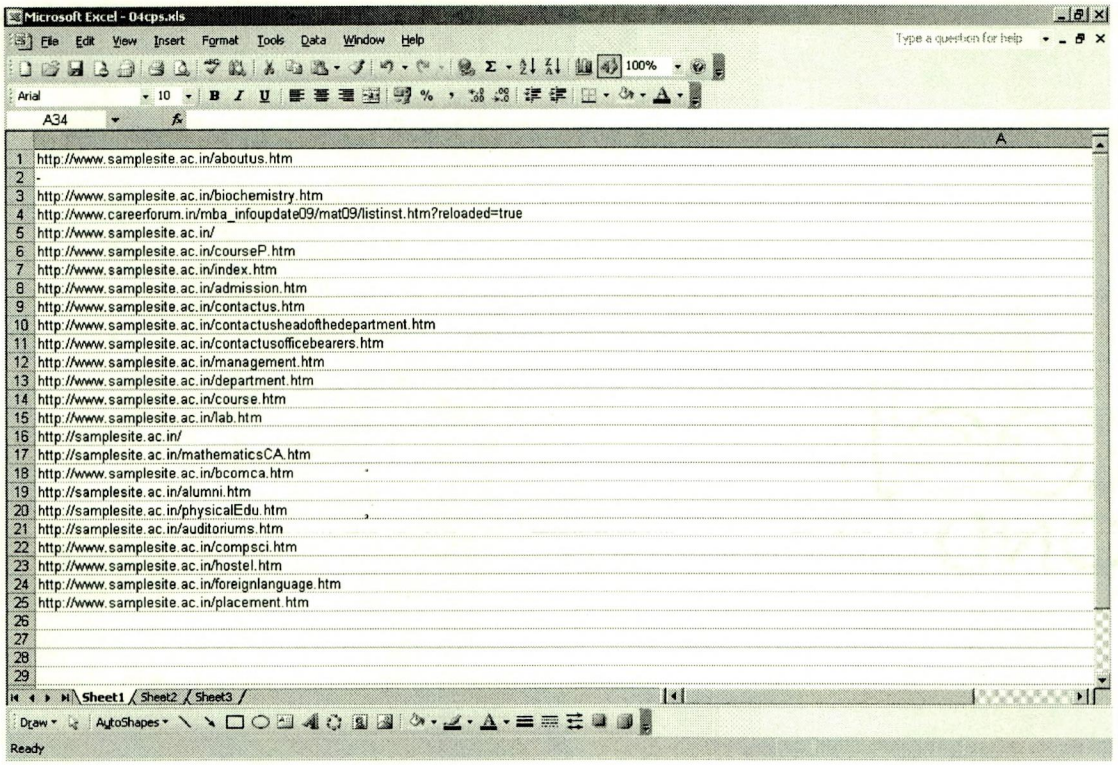


Figure 4.7: Unique URLS identified

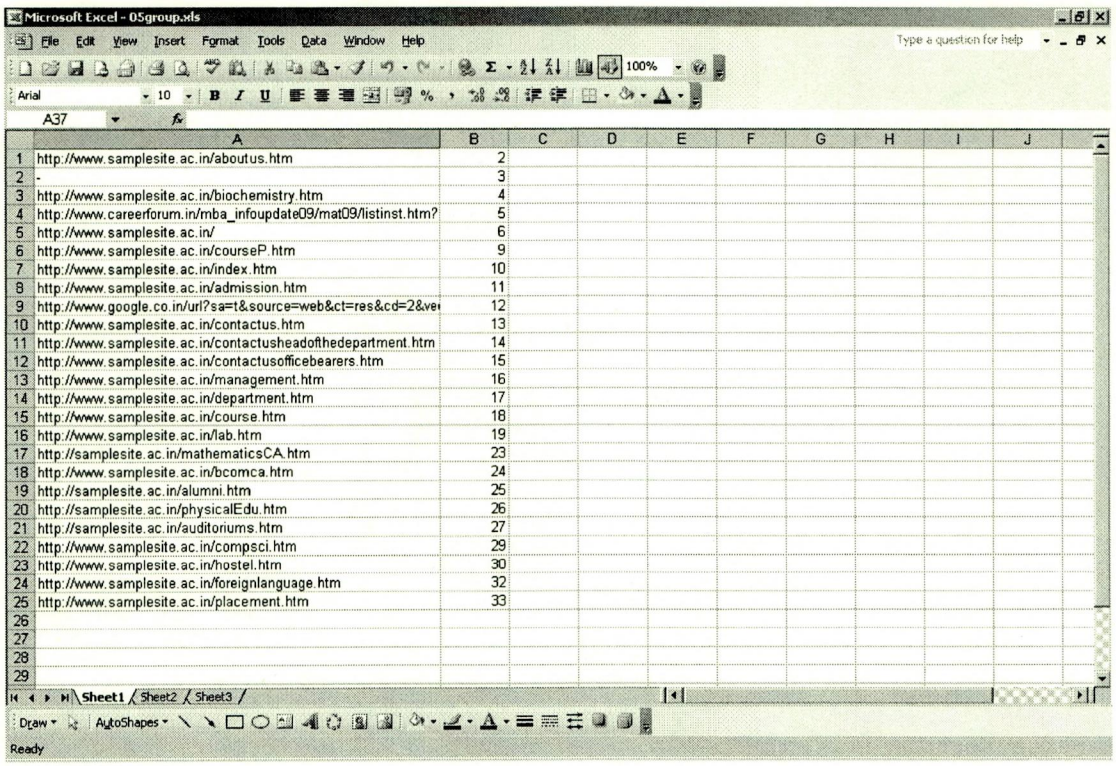
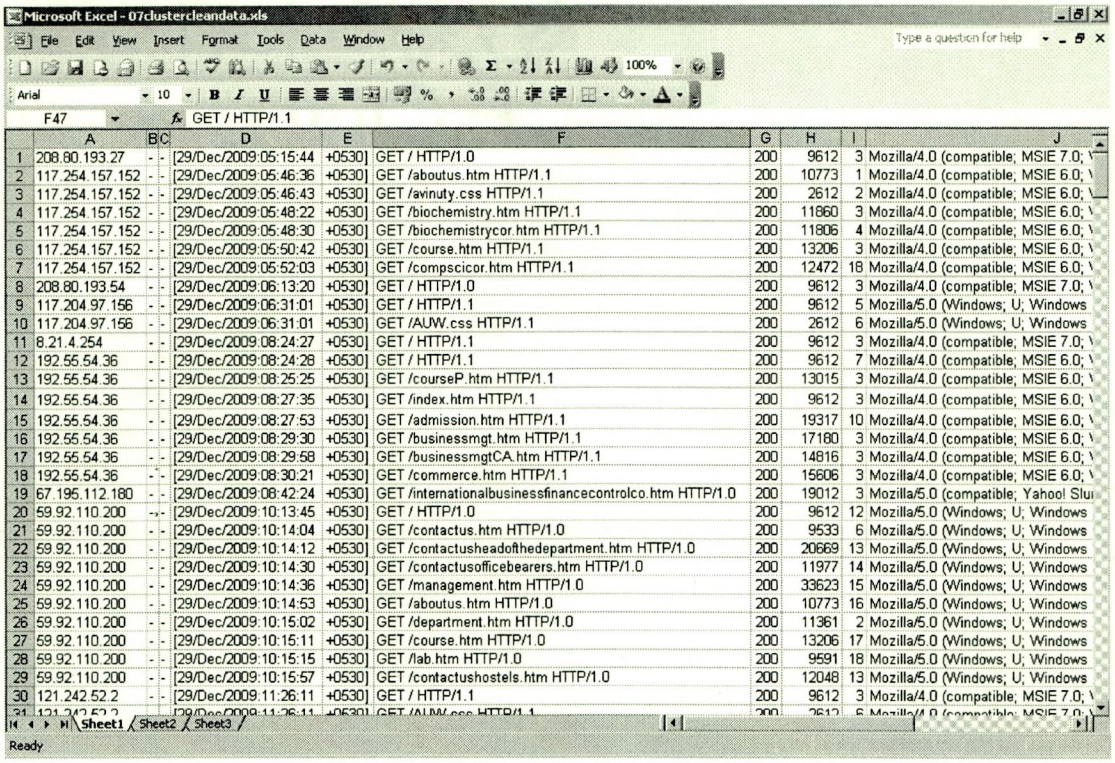


Figure 4.8: Numerical code assigned to each URL



Microsoft Excel - 07clustercleandata.xls

File Edit View Insert Format Tools Data Window Help

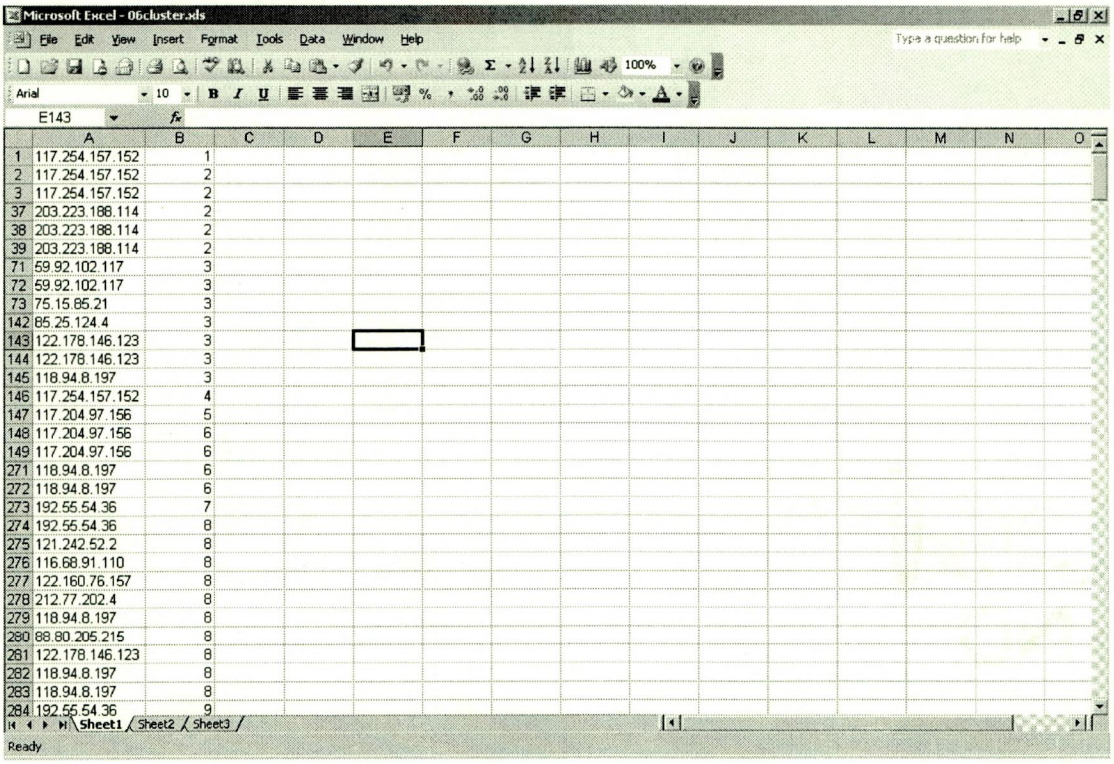
F47 GET / HTTP/1.1

	A	B	C	D	E	F	G	H	I	J
1	208.80.193.27	-	-	[29/Dec/2009:05:15:44	+0530]	GET / HTTP/1.0	200	9612	3	Mozilla/4.0 (compatible; MSIE 7.0; \
2	117.254.157.152	-	-	[29/Dec/2009:05:46:36	+0530]	GET /aboutus.htm HTTP/1.1	200	10773	1	Mozilla/4.0 (compatible; MSIE 6.0; \
3	117.254.157.152	-	-	[29/Dec/2009:05:46:43	+0530]	GET /avinuty.css HTTP/1.1	200	2612	2	Mozilla/4.0 (compatible; MSIE 6.0; \
4	117.254.157.152	-	-	[29/Dec/2009:05:48:22	+0530]	GET /biochemistry.htm HTTP/1.1	200	11860	3	Mozilla/4.0 (compatible; MSIE 6.0; \
5	117.254.157.152	-	-	[29/Dec/2009:05:48:30	+0530]	GET /biochemistrycor.htm HTTP/1.1	200	11806	4	Mozilla/4.0 (compatible; MSIE 6.0; \
6	117.254.157.152	-	-	[29/Dec/2009:05:50:42	+0530]	GET /course.htm HTTP/1.1	200	13206	3	Mozilla/4.0 (compatible; MSIE 6.0; \
7	117.254.157.152	-	-	[29/Dec/2009:05:52:03	+0530]	GET /compscicor.htm HTTP/1.1	200	12472	18	Mozilla/4.0 (compatible; MSIE 6.0; \
8	208.80.193.54	-	-	[29/Dec/2009:06:13:20	+0530]	GET / HTTP/1.0	200	9612	3	Mozilla/4.0 (compatible; MSIE 7.0; \
9	117.204.97.156	-	-	[29/Dec/2009:06:31:01	+0530]	GET / HTTP/1.1	200	9612	5	Mozilla/5.0 (Windows; U; Windows
10	117.204.97.156	-	-	[29/Dec/2009:06:31:01	+0530]	GET /AUW.css HTTP/1.1	200	2612	6	Mozilla/5.0 (Windows; U; Windows
11	8.21.4.254	-	-	[29/Dec/2009:08:24:27	+0530]	GET / HTTP/1.1	200	9612	3	Mozilla/4.0 (compatible; MSIE 7.0; \
12	192.55.54.36	-	-	[29/Dec/2009:08:24:28	+0530]	GET / HTTP/1.1	200	9612	7	Mozilla/4.0 (compatible; MSIE 6.0; \
13	192.55.54.36	-	-	[29/Dec/2009:08:25:25	+0530]	GET /courseP.htm HTTP/1.1	200	13015	3	Mozilla/4.0 (compatible; MSIE 6.0; \
14	192.55.54.36	-	-	[29/Dec/2009:08:27:35	+0530]	GET /index.htm HTTP/1.1	200	9612	3	Mozilla/4.0 (compatible; MSIE 6.0; \
15	192.55.54.36	-	-	[29/Dec/2009:08:27:53	+0530]	GET /admission.htm HTTP/1.1	200	19317	10	Mozilla/4.0 (compatible; MSIE 6.0; \
16	192.55.54.36	-	-	[29/Dec/2009:08:29:30	+0530]	GET /businessmgt.htm HTTP/1.1	200	17180	3	Mozilla/4.0 (compatible; MSIE 6.0; \
17	192.55.54.36	-	-	[29/Dec/2009:08:29:58	+0530]	GET /businessmgtCA.htm HTTP/1.1	200	14816	3	Mozilla/4.0 (compatible; MSIE 6.0; \
18	192.55.54.36	-	-	[29/Dec/2009:08:30:21	+0530]	GET /commerce.htm HTTP/1.1	200	15606	3	Mozilla/4.0 (compatible; MSIE 6.0; \
19	67.195.112.180	-	-	[29/Dec/2009:08:42:24	+0530]	GET /internationalbusinessfinancecontrolco.htm HTTP/1.0	200	19012	3	Mozilla/5.0 (compatible; Yahoo! Slu
20	59.92.110.200	-	-	[29/Dec/2009:10:13:45	+0530]	GET / HTTP/1.0	200	9612	12	Mozilla/5.0 (Windows; U; Windows
21	59.92.110.200	-	-	[29/Dec/2009:10:14:04	+0530]	GET /contactus.htm HTTP/1.0	200	9533	6	Mozilla/5.0 (Windows; U; Windows
22	59.92.110.200	-	-	[29/Dec/2009:10:14:12	+0530]	GET /contactusheadofthedepartment.htm HTTP/1.0	200	20669	13	Mozilla/5.0 (Windows; U; Windows
23	59.92.110.200	-	-	[29/Dec/2009:10:14:30	+0530]	GET /contactusofficebearers.htm HTTP/1.0	200	11977	14	Mozilla/5.0 (Windows; U; Windows
24	59.92.110.200	-	-	[29/Dec/2009:10:14:36	+0530]	GET /management.htm HTTP/1.0	200	33623	15	Mozilla/5.0 (Windows; U; Windows
25	59.92.110.200	-	-	[29/Dec/2009:10:14:53	+0530]	GET /aboutus.htm HTTP/1.0	200	10773	16	Mozilla/5.0 (Windows; U; Windows
26	59.92.110.200	-	-	[29/Dec/2009:10:15:02	+0530]	GET /department.htm HTTP/1.0	200	11361	2	Mozilla/5.0 (Windows; U; Windows
27	59.92.110.200	-	-	[29/Dec/2009:10:15:11	+0530]	GET /course.htm HTTP/1.0	200	13206	17	Mozilla/5.0 (Windows; U; Windows
28	59.92.110.200	-	-	[29/Dec/2009:10:15:15	+0530]	GET /lab.htm HTTP/1.0	200	9591	18	Mozilla/5.0 (Windows; U; Windows
29	59.92.110.200	-	-	[29/Dec/2009:10:15:57	+0530]	GET /contactushostels.htm HTTP/1.0	200	12048	13	Mozilla/5.0 (Windows; U; Windows
30	121.242.52.2	-	-	[29/Dec/2009:11:26:11	+0530]	GET / HTTP/1.1	200	9612	3	Mozilla/4.0 (compatible; MSIE 7.0; \

Sheet1 Sheet2 Sheet3

Ready

Figure 4.9: Formatted web log file after preprocessing



Microsoft Excel - 06cluster.xls

File Edit View Insert Format Tools Data Window Help

E143

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
1	117.254.157.152	1													
2	117.254.157.152	2													
3	117.254.157.152	2													
37	203.223.188.114	2													
38	203.223.188.114	2													
39	203.223.188.114	2													
71	59.92.102.117	3													
72	59.92.102.117	3													
73	75.15.85.21	3													
142	85.25.124.4	3													
143	122.178.146.123	3													
144	122.178.146.123	3													
145	118.94.8.197	3													
146	117.254.157.152	4													
147	117.204.97.156	5													
148	117.204.97.156	6													
149	117.204.97.156	6													
271	118.94.8.197	6													
272	118.94.8.197	6													
273	192.55.54.36	7													
274	192.55.54.36	8													
275	121.242.52.2	8													
276	116.68.91.110	8													
277	122.160.76.157	8													
278	212.77.202.4	8													
279	118.94.8.197	8													
280	88.80.205.215	8													
281	122.178.146.123	8													
282	118.94.8.197	8													
283	118.94.8.197	8													
284	192.55.54.36	9													

Sheet1 Sheet2 Sheet3

Ready

Figure 4.10: Cluster Groups

Knowledge gained from the clustering results include the number of visits made to a single webpage, website traffic, most frequently viewed pages cluster information, and the users navigation behavior. The numerical code and the webpage are given in Table 4.1. The number of visits made by the browsers in 24 hours to these 25 pages is presented in Figure 4.11.

TABLE 4.1

WEBPAGE AND ITS NUMERICAL CODE

WEBPAGE	NUMERICAL CODE
http://www.samplesite.com/index.jsp	1
http://www. samplesite.com /aboutus.htm	2
http://www. samplesite.com /biochemistry.htm	3
http://www. samplesite.com /admission.htm	4
http://www. samplesite.com /contactus.htm	5
http://www. samplesite.com /management.htm	6
http://www. samplesite.com /department.htm	7
http://www. samplesite.com /course.htm	8
http://www. samplesite.com /lab.htm	9
http:// www.samplesite.com /mathematicsCA.htm	10
http:// www.samplesite.com /alumni.htm	11
http:// www.samplesite.com /auditoriums.htm	12
http://www. samplesite.com /compsci.htm	13
http://www. samplesite.com /hostel.htm	14
http://www. samplesite.com /placement.htm	15
http://www. samplesite.com /faculty.htm	16
http://www. samplesite.com /staff.htm	17
http://www. samplesite.com /controller.htm	18
http://www. samplesite.com /syllabus.htm	19
http://www. samplesite.com /library.htm	20
http://www. samplesite.com /homescience.htm	21
http://www. samplesite.com /womenseducation.htm	22
http://www. samplesite.com /mba/management.htm	23
http://www. samplesite.com /compsc/centre.htm	24
http://www. samplesite.com /facilities.htm	25

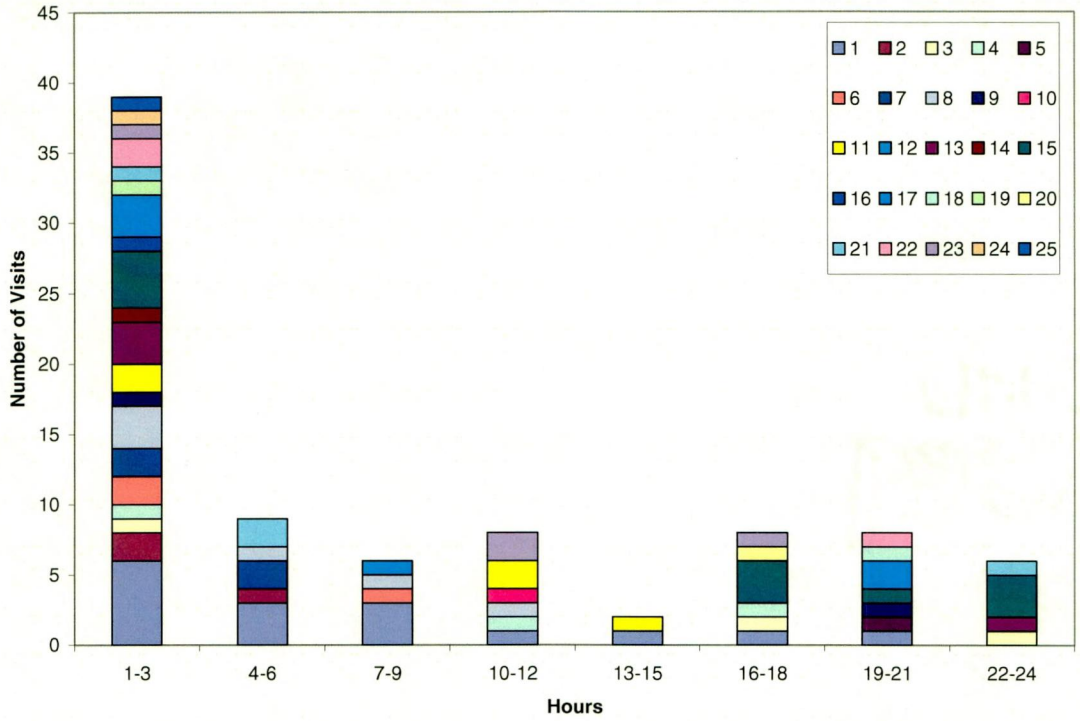


Figure 4.11: Webpage Traffic

From the figure, it can be seen that most of the people view the website during the 1 to 3 hrs, followed by 4 to 6 hrs time slot. One way of interpreting this result is that most of the viewers are non-Indians as the frequently used time slots are mid night time in India. The time slots 10 to 12 hrs, 16 to 18 hrs and 19-21 hrs are also heavily used slots. The 13-15 hrs slots are the minimum used slot.

The user profile and path accessed by the users are shown in Table 4.2.

TABLE 4.2

EXTRACTED NAVIGATION PATTERNS

S.No.	IP Address	User Profile	Unique Pages
1	116.68.91.110	$1 \rightarrow 15 \rightarrow 3 \rightarrow 8 \rightarrow 15 \rightarrow 17$	{1, 15, 3, 8, 17}
2	117.204.97.156	$1 \rightarrow 8 \rightarrow 3 \rightarrow 11 \rightarrow 15 \rightarrow 6 \rightarrow 1 \rightarrow 17 \rightarrow 23 \rightarrow 6$	{1, 6, 3, 11, 15, 17, 23}
3	118.94.8.197	$1 \rightarrow 2 \rightarrow 8 \rightarrow 6 \rightarrow 17 \rightarrow 2$	{1, 2, 8, 6, 17}
4	119.27.62.254	$1 \rightarrow 4 \rightarrow 9 \rightarrow 11 \rightarrow 23$	{1, 4, 9, 11, 23}

5	121.242.52.2	1 → 8 → 13 → 1 → 17	{1, 8, 13, 17}
6	122.178.146.123	1 → 4 → 11 → 15 → 4	{1, 4, 11, 15}
7	192.55.54.36	1 → 14 → 15 → 21	{1, 14, 15, 21}
8	203.223.188.114	1 → 8 → 11 → 15 → 23	{1, 8, 11, 15, 23}
9	208.80.193.26	1 → 12 → 16	{1, 12, 16}
10	212.77.202.4	1 → 13 → 15 → 21	{1, 13, 15, 21}
11	59.92.102.117	2 → 3 → 8 → 11 → 15 → 8 → 17 → 16	{2, 3, 8, 11, 15, 17, 16}
12	59.92.110.200	1 → 8 → 15 → 8	{1, 8, 15}
13	67.195.112.180	1 → 8 → 22 → 13 → 1 → 15	{1, 8, 22, 13, 15}
14	75.15.85.21	10 → 1 → 20	{10, 1, 20}
15	8.21.4.254	1 → 8 → 21 → 24 → 4 → 9 → 24 → 1 → 15	{1, 8, 21, 24, 4, 9, 15}
16	85.25.124.4	13 → 1 → 15 → 19 → 15 → 22	{13, 1, 15, 19, 22}
17	88.80.205.215	1 → 8 → 5 → 25 → 18 → 22	{1, 8, 5, 25,22 }

Further, the data projected also reveals that the most frequently viewed page is the page with code 1, that is, http://www.samplesite.com/index.jsp. This might be because most of the users start their browsing from home page. The next most frequently used pattern is http://www.samplesite.com /placement.htm with numerical code 15, followed by http://www. samplesite.com /course.htm with code 8. This shows most of the users apart from visiting home page are mostly interested about the placement facilities and various courses.

4.4. CLASSIFICATION RESULTS

To predict the user's next request, LCS classification algorithm was used. The LCS finds the longest navigation sequence cluster that matches with the user's referral URL. The user's referral navigation is shown in Figure 4.12.For example, Table 4.3 shows the navigation pattern of four users belonging to the same cluster, constructed over 3 sessions.

TABLE 4.3

SAMPLE PATTERN

IP Address	URL Navigation Pattern
117.204.97.156	3 5 6
192.55.54.36	3 6 7 8 9 10 11
59.92.110.200	2 6 12 13 14 15 16 17 18 19
121.242.52.2	3 6 9 8
122.162.209.77	3 6
116.68.91.110	2 3 6 8 11 20 24
122.160.76.157	6 8 28
118.94.8.197	3 6 8
88.80.205.215	3 6 8
122.178.146.123	3 6 8 31 32

In order to make pattern analysis and prediction, the LCS algorithm calculates a weight matrix with each pattern discovered. Table 4.4 shows the common sequences and their corresponding weights for the incoming referral address 3 → 6 for threshold 0.1.

TABLE 4.4

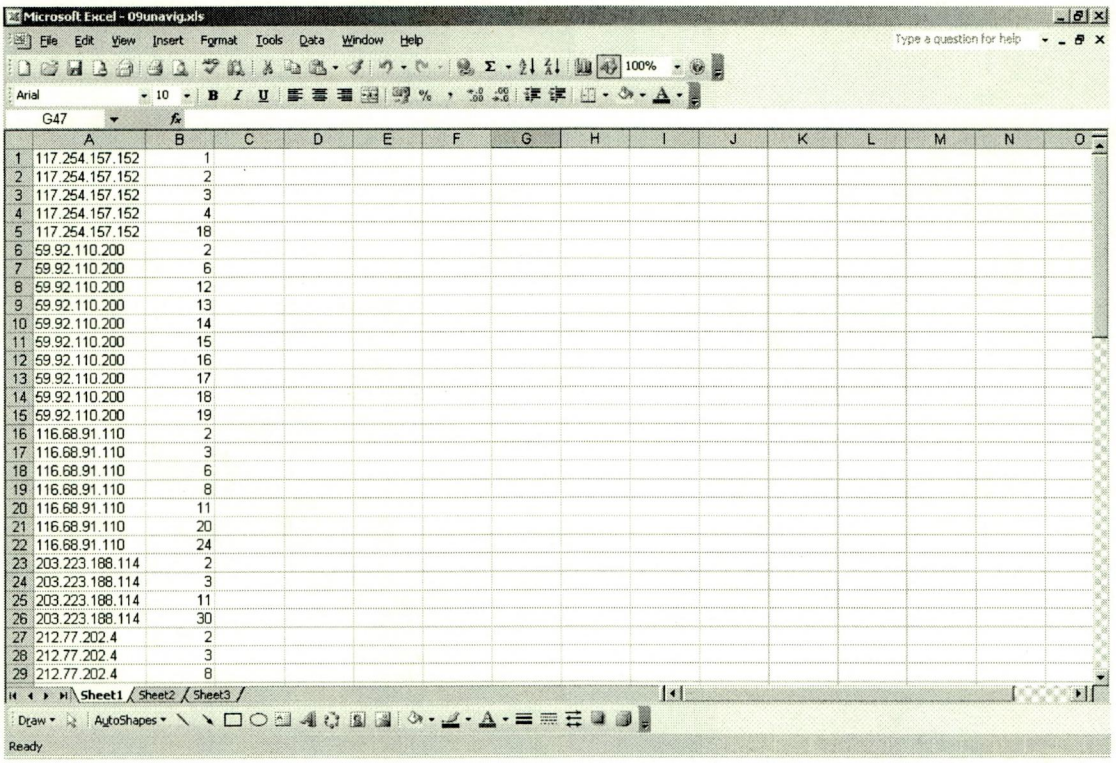
COMMON SEQUENCES AND WEIGHT

IP Address	URL Navigation Pattern	Weight
192.55.54.36	3 6 7 8 9 10 11	0.25
121.242.52.2	3 6 9 8	0.25
122.162.209.77	3 6	0.33
116.68.91.110	2 3 6 8 11 20 24	0.25
118.94.8.197	3 6 8	0.17
88.80.205.215	3 6 8	0.25
122.178.146.123	3 6 8 31 32	0.25

The table shows the IP addresses having the pattern 3 → 6 along with referral URL navigation pattern. According to the LCS rules, the weight with the lowest value predicts the next request more accurately. According to this, the prediction is made to web page with code 8. The system was tested in similar fashion for different threshold values and the accuracy was calculated according to Equation (4.1).

$$\text{Accuracy} = \frac{\text{Total number of correct predictions}}{\text{Total Number of records}} \text{x100} \qquad (4.1)$$

The accuracy of the system while testing with threshold values ranging from 0.1 to 1.0 is presented in Figure 4.13.



	A	B
1	117.254.157.152	1
2	117.254.157.152	2
3	117.254.157.152	3
4	117.254.157.152	4
5	117.254.157.152	18
6	59.92.110.200	2
7	59.92.110.200	6
8	59.92.110.200	12
9	59.92.110.200	13
10	59.92.110.200	14
11	59.92.110.200	15
12	59.92.110.200	16
13	59.92.110.200	17
14	59.92.110.200	18
15	59.92.110.200	19
16	116.68.91.110	2
17	116.68.91.110	3
18	116.68.91.110	6
19	116.68.91.110	8
20	116.68.91.110	11
21	116.68.91.110	20
22	116.68.91.110	24
23	203.223.188.114	2
24	203.223.188.114	3
25	203.223.188.114	11
26	203.223.188.114	30
27	212.77.202.4	2
28	212.77.202.4	3
29	212.77.202.4	8

Figure 4.12: Referral Pattern

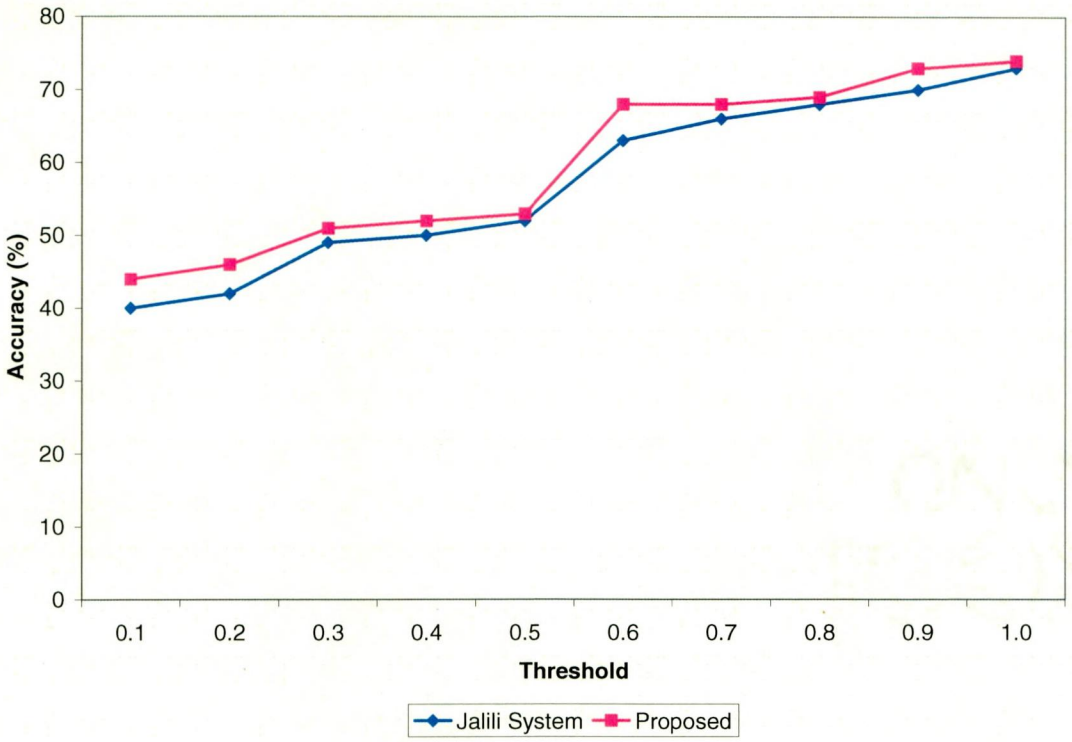


Figure 4.13: Comparison in terms of Accuracy

From the results, it is apparent that the proposed ant-based clustering when combined with LCS produces better result when compared to the existing system, which uses graph partitioning clustering algorithm and LCS algorithm. The result shows a trend that when the threshold value increases, the accuracy also increases and the maximum accuracy achieved by the proposed system is 74%. On average the proposed systems shows 4.18% efficiency positive increase when compared with the existing system.

4.5. CHAPTER SUMMARY

From the experimental data presented it can be concluded that the proposed algorithm discovers the navigation pattern of the user in an efficient manner. The results obtained satisfy the requirements and proves that the implementation of the proposed system will improve the navigation experience of the user in a positive manner. The results obtained are summarized and concluded in the next chapter, Summary and Conclusion.