Preprocessing phase for University Website Access Domain

Nirali Honest, Dr. Bankim Patel, Dr. Atul Patel

Abstract—In the current era websites are growing in size which adds complexity to website design. It becomes necessary to understand who are the users of site, what they are interested in accessing from the website, how much time the users are spending on the website and what is the accessing frequency of users during regular days and during special days. To understand these behaviors we consider a University Website Access Domain (UWAD), where regular days means the operational things that are part of routine activities, like checking the syllabus, checking the faculties details, checking the courses offered, etc. by special days we mean events like admission process, recruitment process, Workshop/Seminar details, etc. The access pattern of users during the regular days is different compared to the access patterns during the special events. Considering these we try to form the preprocessing phase for UWAD. The website is formed of the pages designed using Content Management Systems(CMS), and the preprocessing phase includes data cleaning, user and session identification, determining site structure, mapping of page number and name, path completion, and creation of academic events.

Index Terms— Academic Events, Data Cleaning ,Data Preprocessing,Log Files , Pages designed using CMS, Site Map ,Web Usage Mining.

IJSER

1 INTRODUCTION

IN Web usage mining (WUM)the three steps of process are not coordinated to create a coherent and unique process, so ,there are three main points for generation of this concept,

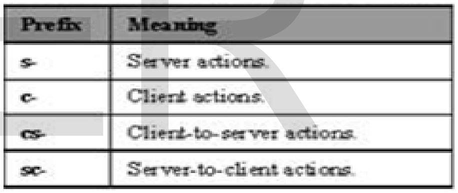
1. Web usage mining process can be molded according to the specific goal w.r.t mining.
2. There is no support for generation of reports for particular events; you need to remember the interval of the event for generating the report of the event. (i.e. specify date intervals every time for generating the reports.)
3. The websites designed by Content Management System (CMS), Master Page concept ,where web pages are given unique page numbers that are fetched from database, the report generation for per page frequency is not supported by certain tools, and if supported the page name cannot be known if it is generated by the ID number.

The paper focuses with the problems and discuss the preprocessing phase for the university website, where pages are designed using the CMS and each page has a page number rather than name. The paper includes the details of preprocessing phase for the given domain.

2 TYPES OF LOG FILES

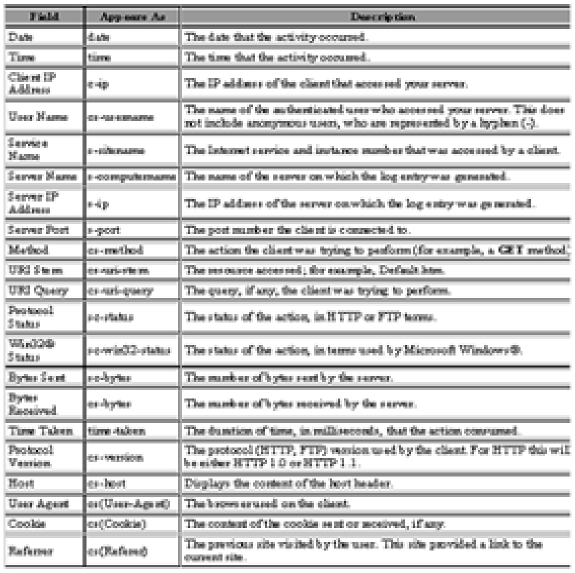
Information about internet user is stored in different raw log files. Web Server logs are plain text (ASCII) files, that is independent from the server platform. There are some distinctions between server software, but traditionally there are four types of server logs[1]:

TABLE 1
W3C EXTENDED LOGGING FIELD DEFINITION



Prefix	Meaning
s-	Server actions.
c-	Client actions.
cs-	Client-to-server actions.
sc-	Server-to-client actions.

TABLE 2
LOG FILE FIELD DESCRIPTION



Field	Appears As	Description
Date	date	The date that the activity occurred.
Time	time	The time that the activity occurred.
Client IP Address	c-ip	The IP address of the client that accessed your server.
User Name	cs-username	The name of the authenticated user who accessed your server. This does not include anonymous users, who are represented by a hyphen (-).
Service Name	s-sitename	The Internet service and instance number that was accessed by a client.
Server Name	s-computername	The name of the server on which the log entry was generated.
Server IP Address	s-ip	The IP address of the server on which the log entry was generated.
Server Port	s-port	The port number the client is connected to.
Method	cs-method	The action the client was trying to perform (for example, a GET method;
URI Stem	cs-uri-stem	The resource accessed; for example, Default.htm.
URI Query	cs-uri-query	The query, if any, the client was trying to perform.
Protocol Status	sc-status	The status of the action, in HTTP or FTP terms.
Win32® Status	sc-win32-status	The status of the action, in terms used by Microsoft Windows®.
Bytes Sent	sc-bytes	The number of bytes sent by the server.
Bytes Received	cs-bytes	The number of bytes received by the server.
Time Taken	time-taken	The duration of time, in milliseconds, that the action consumed.
Protocol Version	cs-version	The protocol (HTTP, FTP) version used by the client. For HTTP this will be either HTTP 1.0 or HTTP 1.1.
Host	cs-host	Displays the content of the host header.
User Agent	cs(User-Agent)	The browser used on the client.
Cookie	cs(Cookie)	The content of the cookie sent or received, if any.
Referrer	cs(Referer)	The previous site visited by the user. This site provided a link to the current site.

- Nirali Honest is working as an Assistant Professor at Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, CHARUSAT, Changa. E-mail: niralihonest.mca@ecchanga.ac.in
- Dr. Bankim Patel is working as an Director at Shrimad Rajchandra Institute of Management and Computer Application, Uka Tarsadia University Bardoli. E-mail:bankim_patel@srimca.edu.in
- Dr. Atul Patel is working as an Pricipal at Smt. Chandaben Mohanbhai Patel Institute of Computer Applications, CHARUSAT,Changa,. E-mail: atulpatel.mca@ecchanga.ac.in

(i) Transfer Log (ii) Error Log (iii)Agent Log (iv) 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. Table1 shows the prefix indication used in the W3C extended log file format and Table 2 shows the basic fields used in the W3C extended log file.

3 Concept of WUM and UWAD

A university website is designed as a set of pages that constite the Institute details, news and announcements, Listing of major actions like admission, recruitment , technical and non-technical events, workshops and seminars, etc. By extracting access patterns of users behavior, various types of analysis can be developed over the given period which includes the listing of errors occurred while accessing the website, per page frequency, hit ratio for the given interval, maximum referred page, minimum referred page, first page accessed by the user, last page accessed by the user, navigation path of the user, hit ratio for special events like admission process, convocation, recruitment, etc.The UWAD can certainly benefit from insights provided by WUM, but it is necessary to incorporate specific concepts and procedures from the context in study. According to Zaiane et al. [2][3], this transposition of characteristics to a new domain is not trivial, requiring that techniques and tools originally developed for the electronic commerce context (e.g. [4]) be customizable for adjusting to requirements of a new domain. For instance, they stress that the evaluation of Web-based learning environments using WUM techniques must be, at least partially, executable by domain-related people (e.g. instructors, site designers), who in general are not familiar with data mining techniques. UWAD concept is designed to know the user behavior and generate required patterns for the access of website during regular academic days and special academic days. By special days we mean the patterns generated for particular events like admission, recruitment, convocation, workshop detailing, etc. We provide the facility to create events and generate the patterns for regular days as well as patterns specific to particular events.

4 Preprocessing phase

It is necessary to perform a data preparation to convert the raw data for further process. The data taken as input are web server logs, site file and academic calendar of a given University. The general architecture of the preprocessing phase for UWAD is show in figure 1

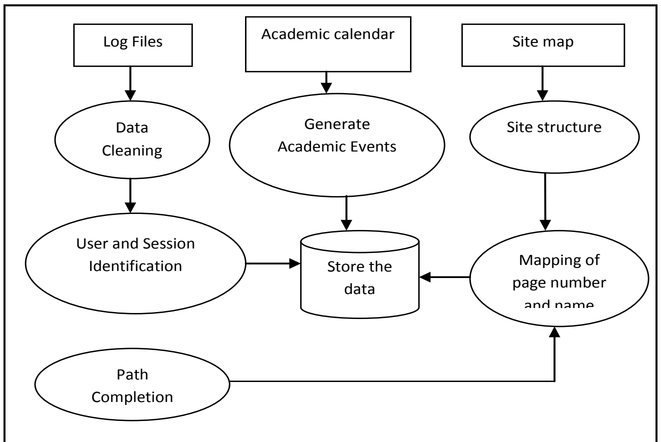


Fig. 1 Preprocessing phase for the UWAD

Purpose of preprocessing phase is to reduce the size of log file, to increase the quality of the available data, and to filter the required data in the required format. Steps of preprocessing phase are as below,

- Collect the raw log file.
- Clean the file as per requirement.
- Identify users and sessions
- Know the site map and Design site structure.
- Map the id number of page with page name.
- Complete the path traversed by user in a given session
- Collect the academic calendar of a given University and generate Academic events for it.

4.1 Detailing of Preprocessing Steps

4.1.1 Data Cleaning

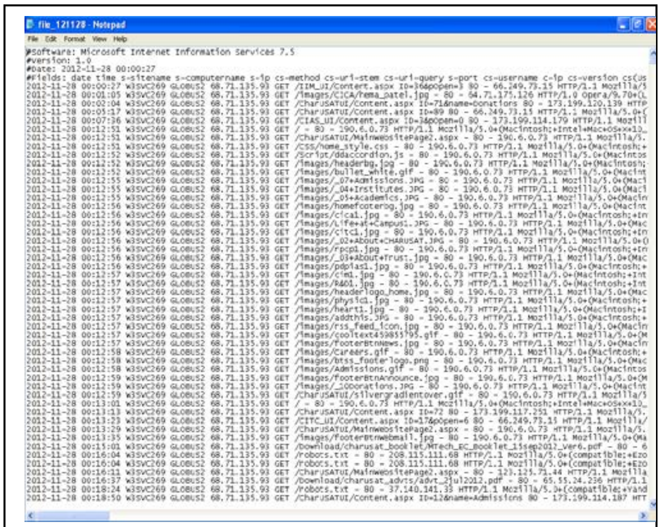
Data cleaning means eliminate the irrelevant information (information which is not useful for the further process, or even misleading or faulty) from the original Web log file for further processing.

The steps of algorithm are as below,

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 include 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.

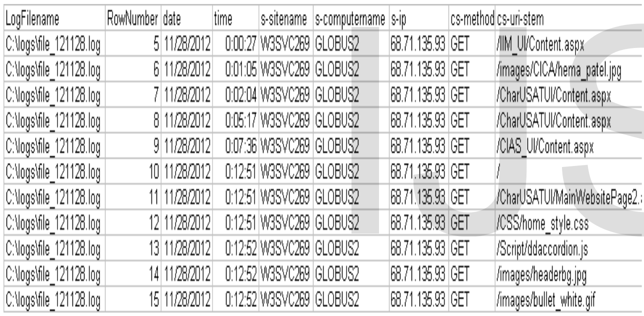
IJSER

Figure 2 shows the raw log file and Figure 3 shows the log file after it has been parsed.



```
#Software: Microsoft Internet Information Services 7.5
#Version: 1.0
#Date: 2012-11-28 00:00:27
#Fields: date time s-sitename s-computername s-ip cs-method cs-uri-stem cs-uri-query s-port cs-username c-ip cs-version cs(Us
2012-11-28 00:00:27 W3SVC269 GLOBUS2 68.71.135.93 GET /IIM_UI/Content.aspx ID=36&pOpen=3 80 - 66.249.73.15 HTTP/1.1 Mozilla/5
2012-11-28 00:01:05 W3SVC269 GLOBUS2 68.71.135.93 GET /images/CICA/hema_patel.jpg - 80 - 64.71.175.126 HTTP/1.0 Opera/9.70+(L
2012-11-28 00:02:04 W3SVC269 GLOBUS2 68.71.135.93 GET /CharUSATUI/Content.aspx ID=71&name=Donations 80 - 173.199.120.139 HTTP
2012-11-28 00:05:17 W3SVC269 GLOBUS2 68.71.135.93 GET /CharUSATUI/Content.aspx ID=89 80 - 66.249.73.15 HTTP/1.1 Mozilla/5.0+(
2012-11-28 00:07:36 W3SVC269 GLOBUS2 68.71.135.93 GET /CIAS_UI/Content.aspx ID=3&pOpen=0 80 - 173.199.114.179 HTTP/1.1 Mozill
2012-11-28 00:12:51 W3SVC269 GLOBUS2 68.71.135.93 GET / - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;+Intel+Mac+OS+X+10_
2012-11-28 00:12:51 W3SVC269 GLOBUS2 68.71.135.93 GET /CharUSATUI/MainWebsitePage2.aspx - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.
2012-11-28 00:12:51 W3SVC269 GLOBUS2 68.71.135.93 GET /CSS/home_style.css - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;+
2012-11-28 00:12:52 W3SVC269 GLOBUS2 68.71.135.93 GET /Script/ddaccordion.js - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintos
2012-11-28 00:12:52 W3SVC269 GLOBUS2 68.71.135.93 GET /images/headerbg.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;
2012-11-28 00:12:52 W3SVC269 GLOBUS2 68.71.135.93 GET /images/bullet_white.gif - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macint
2012-11-28 00:12:55 W3SVC269 GLOBUS2 68.71.135.93 GET /images/_07+Admissions.JPG - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Maci
2012-11-28 00:12:55 W3SVC269 GLOBUS2 68.71.135.93 GET /images/_04+Institutes.JPG - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Maci
2012-11-28 00:12:55 W3SVC269 GLOBUS2 68.71.135.93 GET /images/_05+Academics.JPG - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macin
2012-11-28 00:12:56 W3SVC269 GLOBUS2 68.71.135.93 GET /images/homefooterbg.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macint
2012-11-28 00:12:56 W3SVC269 GLOBUS2 68.71.135.93 GET /images/cica1.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;+In
2012-11-28 00:12:56 W3SVC269 GLOBUS2 68.71.135.93 GET /images/Life+at+Campus1.JPG - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Mac
2012-11-28 00:12:56 W3SVC269 GLOBUS2 68.71.135.93 GET /images/citc1.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;+In
2012-11-28 00:12:56 W3SVC269 GLOBUS2 68.71.135.93 GET /images/_02+About+CHARUSAT.JPG - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(
2012-11-28 00:12:56 W3SVC269 GLOBUS2 68.71.135.93 GET /images/rpcp1.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;+In
2012-11-28 00:12:56 W3SVC269 GLOBUS2 68.71.135.93 GET /images/_03+About+Trust.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Mac
2012-11-28 00:12:56 W3SVC269 GLOBUS2 68.71.135.93 GET /images/pdpias1.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;+
2012-11-28 00:12:57 W3SVC269 GLOBUS2 68.71.135.93 GET /images/cim1.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;+Int
2012-11-28 00:12:57 W3SVC269 GLOBUS2 68.71.135.93 GET /images/R&D1.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;+Int
2012-11-28 00:12:57 W3SVC269 GLOBUS2 68.71.135.93 GET /images/headerlogo_home.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Mac
2012-11-28 00:12:57 W3SVC269 GLOBUS2 68.71.135.93 GET /images/physic1.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;+
2012-11-28 00:12:57 W3SVC269 GLOBUS2 68.71.135.93 GET /images/heart1.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;+I
2012-11-28 00:12:57 W3SVC269 GLOBUS2 68.71.135.93 GET /images/addthis.JPG - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;+
2012-11-28 00:12:57 W3SVC269 GLOBUS2 68.71.135.93 GET /images/rss_feed_icon.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macin
2012-11-28 00:12:57 W3SVC269 GLOBUS2 68.71.135.93 GET /images/cooltext459855795.gif - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(M
2012-11-28 00:12:57 W3SVC269 GLOBUS2 68.71.135.93 GET /images/footerBtnNews.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macin
2012-11-28 00:12:58 W3SVC269 GLOBUS2 68.71.135.93 GET /images/Careers.gif - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;+
2012-11-28 00:12:58 W3SVC269 GLOBUS2 68.71.135.93 GET /images/btss_footerlogo.png - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Mac
2012-11-28 00:12:58 W3SVC269 GLOBUS2 68.71.135.93 GET /images/Admissions.gif - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintos
2012-11-28 00:12:59 W3SVC269 GLOBUS2 68.71.135.93 GET /images/footerBtnAnnounce.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(M
2012-11-28 00:12:59 W3SVC269 GLOBUS2 68.71.135.93 GET /images/_10Donations.JPG - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macint
2012-11-28 00:12:59 W3SVC269 GLOBUS2 68.71.135.93 GET /CharUSATUI/silvergradientover.gif - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5
2012-11-28 00:13:01 W3SVC269 GLOBUS2 68.71.135.93 GET / - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Macintosh;+Intel+Mac+OS+X+10_
2012-11-28 00:13:13 W3SVC269 GLOBUS2 68.71.135.93 GET /CharUSATUI/Content.aspx ID=72 80 - 173.199.117.251 HTTP/1.1 Mozilla/5.
2012-11-28 00:13:23 W3SVC269 GLOBUS2 68.71.135.93 GET /CITC_UI/Content.aspx ID=17&pOpen=6 80 - 66.249.73.15 HTTP/1.1 Mozilla/
2012-11-28 00:13:29 W3SVC269 GLOBUS2 68.71.135.93 GET /CharUSATUI/MainWebsitePage2.aspx - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.
2012-11-28 00:13:35 W3SVC269 GLOBUS2 68.71.135.93 GET /images/footerBtnWebmail.jpg - 80 - 190.6.0.73 HTTP/1.1 Mozilla/5.0+(Ma
2012-11-28 00:15:01 W3SVC269 GLOBUS2 68.71.135.93 GET /Download/charusat_booklet/MTech_EC_Booklet_15sep2012_ver6.pdf - 80 - 6
2012-11-28 00:16:04 W3SVC269 GLOBUS2 68.71.135.93 GET /robots.txt - 80 - 208.115.111.68 HTTP/1.1 Mozilla/5.0+(compatible;+Ezo
2012-11-28 00:16:04 W3SVC269 GLOBUS2 68.71.135.93 GET /robots.txt - 80 - 208.115.111.68 HTTP/1.1 Mozilla/5.0+(compatible;+Ezo
2012-11-28 00:16:11 W3SVC269 GLOBUS2 68.71.135.93 GET /CharUSATUI/MainWebsitePage2.aspx - 80 - 123.125.71.44 HTTP/1.1 Mozilla
2012-11-28 00:16:37 W3SVC269 GLOBUS2 68.71.135.93 GET /Download/charusat_advts/advt_2jul2012.pdf - 80 - 65.55.24.236 HTTP/1.1
2012-11-28 00:18:24 W3SVC269 GLOBUS2 68.71.135.93 GET /robots.txt - 80 - 37.140.141.33 HTTP/1.1 Mozilla/5.0+(compatible;+Yand
2012-11-28 00:18:50 W3SVC269 GLOBUS2 68.71.135.93 GET /CharUSATUI/Content.aspx ID=12&name=Admissions 80 - 173.199.114.187 HTT
```

Fig. 2 Snapshot of log file



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

Fig. 3 Snapshot after Parsing a log file

After Data cleaning the summary of Preprocessing stage is show in table 3.

TABLE 3

RESULT ANALYSIS OF PROPOSED CLEANING PROCESS

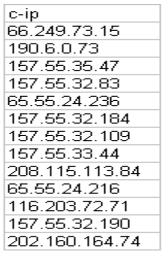
Stage of Preprocessing	No. of Web Objects Retrieved
Initial size of file	12886 KB (12.5 MB)
Size of file after cleaning	608 KB
Before Cleaning	26380
After Cleaning	2451
Total time taken for	2.31 minutes

4.1.2 User Identification

User identification means identifying each user accessing Web site, whose goal is to mine every user's access characteristic. We have taken the combination of IP address browser and OS details to identify users. The steps are as below.

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.

Figure 4 shows the unique users.



c-ip
66.249.73.15
190.6.0.73
157.55.35.47
157.55.32.83
65.55.24.236
157.55.32.184
157.55.32.109
157.55.33.44
208.115.113.84
65.55.24.216
116.203.72.71
157.55.32.190
202.160.164.74

Fig. 4 Snapshot of Unique users

4.1.3 Session Identification

After identification of users the next step is to identify sessions, a session is a series of actions performed by a user while accessing the website. The goal of session identification is to find pages accessed during the sessions, which further helps in knowing the users access pattern and frequency path. Reconstruction of user session is challenging as the HTTP protocol is stateless and connectionless.[5]

There are two heuristics available for session identification, time oriented and navigation oriented. In Time Oriented Heuristics, one method is to know total session time and other method is to know single page stay time. The set of pages visited by a specific user at a specific time is called page viewing time. It varies from 25.5 minutes [6] to 24 hours [7] while default time is 30 minutes by R.Cooley [8]. 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. In Navigation Oriented Heuristics, it considers webpage connectivity. Current work follows the time oriented heuristics. We have used a threshold of 30 minutes as a default timeout, based on empirical data.

Steps for session identification are as below,

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.

After user and session identification summary of preprocessing pahse is shown in table 4.

TABLE 4
RESULT ANALYSIS AFTER USER AND SESSION IDENTIFICATION

Stage of Preprocessing	No. of Web Objects Retrieved
No. of Entries	2197
No. of Unique IP Addresses	305
No. of Users	352
No. of sessions	365

4.1.4 Site Map and Site Structure

In order to complete path, we must know the Web site's topological structure beforehand. There are two ways to solve this problem. One way is that the hyperlink relation between Web pages can be found by hand and then be transformed into the hyperlink relation database. For this way, the amount of work is huge. The other ways is that an automatic search engine algorithm, like hyperlink relation between Web pages and form the Web site's topological structure. The website considered for the work has structure given in figure 5 and the page linking is shown in figure 6.

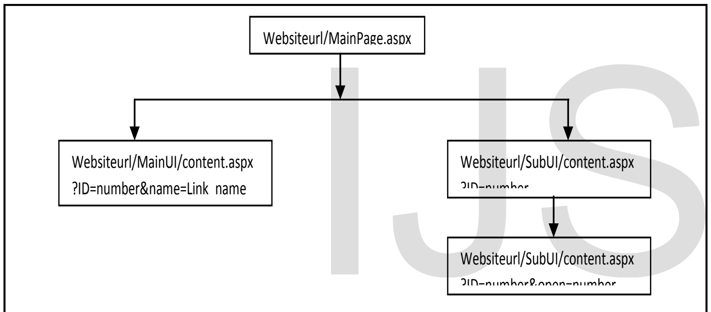

Websiteurl/MainPage.aspx
Websiteurl/MainUI/content.aspx
?ID=number&name=Link name
Websiteurl/SubUI/content.aspx
?ID=number
Websiteurl/SubUI/content.aspx
?ID=number&open=number

Fig. 5 General Structure of website designed using CMS

IJSER

Site structure creation includes following steps,

a. Generate the site map.
b. Input site map and separate the fields.
c. The fields include domain, ui, id, name, key and pOpen.
d. Add the details

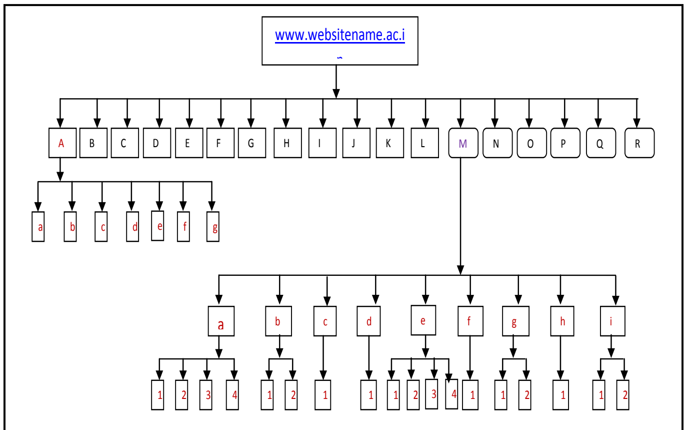


Fig. 6 Linked pages of website

4.1.5 Mapping of Page ID with Page Name

The website assumed for the work is designed using CMS, so the URL formed while accessing the website are of the below **form /folder_UI/ContentPage.aspx?ID=10** so it doesn't give meaning about which page is accessed and it is difficult to perform the path completion based on the ID numbers generated. So before we proceed for path completion, it is necessary to perform mapping of Page ID with Page name, below steps are considered for mapping.

a. Read the URI stem. (cs-uri-stem)
b. Read the URI query. (cs-uri-query)
   - Get the parameters(ID, name, pOpen and key.)
c. Link it with the name and reform the URL
d. Store it.

4.1.6 Path Completion

Due to proxy servers and cached versions of the pages used by the client using 'Back', the sessions identified have many missed pages [9][10]. So path completion step is carried out to identify missing pages. Below steps are performed for path completion, the heuristic assumes that if a page is requested that is not directly linked to the previous page accessed by the same user, the referrer log can be referred to see from which page the request came. Hence each session reflects the full path, including the pages that have been backtracked. After the path completion phase, the user's session file results in paths consisting of a collection of page references including repeat page accesses made by a user, as shown in figure 7.

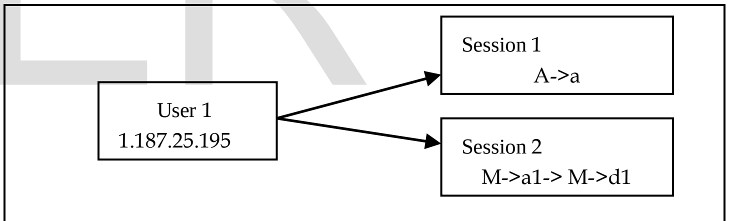


Fig. 7 Pages traversed by the user in particular session

4.1.7 Specifying Academic events

In any University, there are lot many events that may occur during the particular academic year. An event is a special occurrence of operation that occurs for a finite time. In the university academic events can be considered as Recruitment, Admission, Display of results, Announcement of workshop, etc. During these academic events the access of website is different than the regular access. So in this paper we try to show the insertion of academic events to be specified, so that the patterns of access during these events can be analyzed and compared to the normal access of website. Following details are stored for generating academic events,

a. Specify the Academic year
b. Specify the Name of Institute
c. Specify the Name of event
d. Specify the Intervals of the event i.e start and end date.
e. Store these details.

5 Conclusion

In order to derive more specific patterns and results pertaining to a University environment we tend to create this preprocessing phase. Apart from it we focus the problems raised because of pages designed using CMS and try to give solution for mapping of pages ID with page Name, which would help the administrator to analyze the patters more specifically. The next work is to apply and analyse patterns for the cleaned and formatted data.

Acknowledgment

The authors would like to thank Charotar University of Science and Technology (CHARUSAT) for providing the necessary resources for accomplishing the work.

References

[1] Mohd Helmy Abd Wahab, Mohd Norzali Haji Mohd, et. Al, "Data Preprocessing on Web Server Logs for Generalized Association Rules Mining Algorithm" , World Academy of Science, Engineering and Technology, 2008.

[2] O. R. Zaïane, Web Usage Mining for a Better Web-Based Learning Environment. Department of Computing Science University of Alberta Edmonton, Alberta, Canada, 2001.

[3] O. R. Zaïane and J. Luo, Towards Evaluating Learners' Behaviour in a Web-Based Distance Learning Environment. In: Proceedings IEEE ICALT 2001, Madison, USA. 2001.

[4] R. Cooley, B. Mobasher, and J. Srivastava, Data Preparation for Mining World Wide Web Browsing Patterns. Journal of Knowledge and Information Systems, (1), 1999.

[5]] Chungsheng Zhang and Liyan Zhuang , "New Path Filling Method on Data Preprocessing in Web Mining ,", Computer and Information Science Journal , August 2008

[6] Catlegde L. and Pitkow J., "Characterising browsing behaviours in the world wide Web,", Computer Networks and ISDN systems, 1995. 23

[7] Spilipoulou M.and Mobasher B, Berendt B.,"A framework for the Evaluation of Session Reconstruction Heuristics in Web Usage Analysis," INFORMS Journal on Computing Spring ,2003.

[8] Robert.Cooley,Bamshed Mobasher, and Jaideep Srinivastava, "Web mining:Information and Pattern Discovery on the World Wide Web,",In International conference on Tools with Artificial Intelligence, pages 558-567, Newport Beach, IEEE,1997.

[9] Yan LI , Boqin FENG, Qinjiao MAO, "Research on Path Completion Technique in Web Usage Mining", International Symposium on Computer Science and Computational Technology, 2008.

[10] Yan LI ,Bo-qin FENG, "The Construction of Transactions for Web Usage Mining", International Conference on Computational Intelligence and Natural Computing, 2009.

IJSER