PREPROCESSING OF WEB LOGS

Ms. Dipa Dixit
Lecturer
Fr.CRIT, Vashi

Ms. M Kiruthika
Assistant Professor
Fr.CRIT, Vashi.

Abstract-**Today's real world databases are highly susceptible to noisy, missing and inconsistent data due to their typically huge size data and their origin from multiple, heterogeneous sources. Hence, pre-processing of data is necessary to help improve the quality of data and consequently the mining results. There are number of data pre-processing techniques. In this paper, we would like to discuss two different approaches for data pre-processing one based on XML and other based on text file. But the basic algorithm and steps involved in pre-processing are considered same for both the approaches.**

Keywords: Pre-processing; Pattern Discovery; User Navigation Patter; weblogs

I INTRODUCTION

The World wide Web has become one of the most important media to store, share and distribute information .At present, Google is indexing more than 8 billion Web pages[1]. The rapid expansion of the Web has provided a great opportunity to study user and system behavior by exploring Web access logs. The WWW is serving as a huge widely distributed global information service center for technical information, news, advertisement, e-commerce and other information service. This makes information retrieval process difficult. Most users may not have good knowledge of the structure of the information network, and may easily get bored by taking many access hops and losing impatience when waiting for the information. These challenges will have been solved efficiently by Web mining, which is the application of data mining technologies. Web mining [2] that discovers and extracts interesting knowledge/patterns from Web is classified into three types as Web Structure Mining that focuses on hyperlink structure, Web Contents Mining that focuses on page contents as well as Web Usage Mining that focuses on Web logs. In this paper, we are concerned about Web Usage Mining (WUM), also named as Web log mining. The process of WUM[3] includes three phases shown in Fig. 1: data preprocessing, pattern discovery, and pattern analysis. Data Preprocessing basically consists of data cleaning, data integration, transformation and data reduction. Pattern discovery deals with extracting knowledge from preprocessed data. Some of the techniques used in Pattern discovery are Association rules, Classification, Clustering etc. Pattern Analysis filters out uninteresting rules or patterns from the set found in the pattern discovery phase. Basic Web Log Mining structure is given below in figure1.

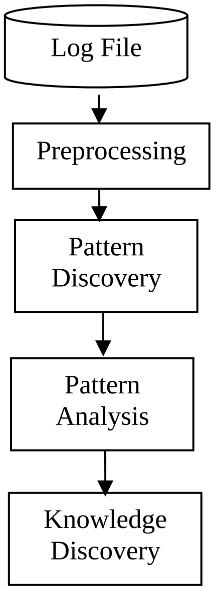


Figure 1: Basic Web Log Mining Structure

In this paper, we would be focusing on two approaches of data pre-processing.

II RELATED WORK

In this section, we introduce some related work in data preprocessing. In the recent years, there has been much research on Web usage mining [9], [10], [11], [12]. However, data preprocessing has received far less attention than it deserves. Methods for user identification, session identification, and path completion, are presented in [13]. In another work [14] the authors compared time-based and referrer-based heuristics for visit reconstruction. They found out that a heuristic's appropriateness depends on the design of the Web site (i.e. whether the site is frame-based or frame-free) and on the length of the visits (the referrer-based heuristic performs better for shorter visits). In [15] Marquardt et al. addressed the application of WUM in the e-learning area with a focus on the preprocessing phase. In this context, they are defined the notion of visit from the e-learning point of view. Moreover, in the same paper, the authors have presented several data preparation techniques to identify Web users, i.e, the path completion and the use of site topology. To identify the user sessions, a fixed time period, say thirty minutes [6], is used to be the threshold between two sessions. Zaïane et.al[16] have applied various traditional data mining techniques to Internet log files in order to find different types of patterns, which can be harnessed as electronic commerce decision support knowledge. The pre-processed data can be finally loaded into Knowledge Base and can be further used for analysis with help of data mining techniques like classification, clustering and association mining.

III BLOCK DIAGRAM FOR DATA PRE-PROCESSING

Steps involved in data pre-processing are shown with the help of block diagram in Figure 2 below:

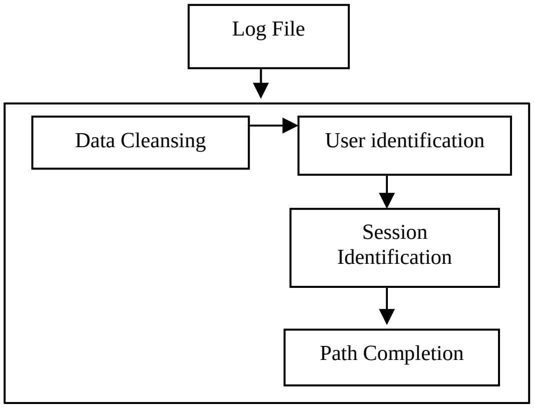


Figure 2: Block diagram for data Pre-processing

A. Log File

Log File is the input to pre-processing block. A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site. Data preprocessing a web usage mining model (Web-Log preprocessing) aims to reformat the original web logs to identify user's access sessions. The Web server usually registers all users' access activities of the website as Web server logs. Due to different server setting parameters, there are many types of web logs, but typically the log files share the same basic information, such as: client IP address, request time, requested URL, HTTP status code, referrer, etc.. Data sets consisting of web log records for 5446 users were collected from De Paul University website. Web log consist of 17 attributes with the data values in the form of records.
The following is a fragment from the IIS server logs:

date time c-ip cs-username s-sitename s-computername s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status time-taken cs-version cs-host cs(User-Agent) cs(Referer)

```
2002-04-01 00:00:10 1cust62.tnt40.chi5.da.uu.net - w3svc3
bach bach.cs.depaul.edu 80 get /courses/syllabus.asp
course=323-21-603&q=3&y=2002&id=671 200 156 http/1.1
www.cs.depaul.edu
mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+
4.90;+msn+6.1;+msnbmsft;+msnmen-us;+msnc21)
http://www.cs.depaul.edu/courses/syllabilist.asp
depaul.edu/courses/syllabilist.asp
```

B Basic Algorithm And Steps Involved In Preprocessing

Steps involved in data pre-processing are explained below

1 **Data Cleansing**: Irrelevant records are eliminated during data cleansing. Since target of web usage mining is to get traversal pattern, following two kinds of records are unnecessary and should be removed :

a. The records of graphics, video and format information. The records having filenames suffixes of GIF, JPEG, CSS and so on, which can be found in cs_uri_stem field of record.

b. The records with failed HTTP status codes. By examining the status field of every record in the web log, the record with status code over 299 and under 200 are removed**.**

2 **User and Session Identification**: The task of user and session identification is to find out the different user sessions from the original web access log. A referrer-based method is used for identifying sessions. The different IP addresses distinguish different users.

a. If the IP addresses are same, the different browsers and operation systems indicate different users which can be found by client IP address and user agent who gives information of user's browsers and operating system.

b. If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The ReferURI (cs_referer) is checked, a new user session is identified if the URL in the ReferURI - field hasn't been accessed previously, or there is a large interval (usually more than 30 minutes) between the accessing time of this record.

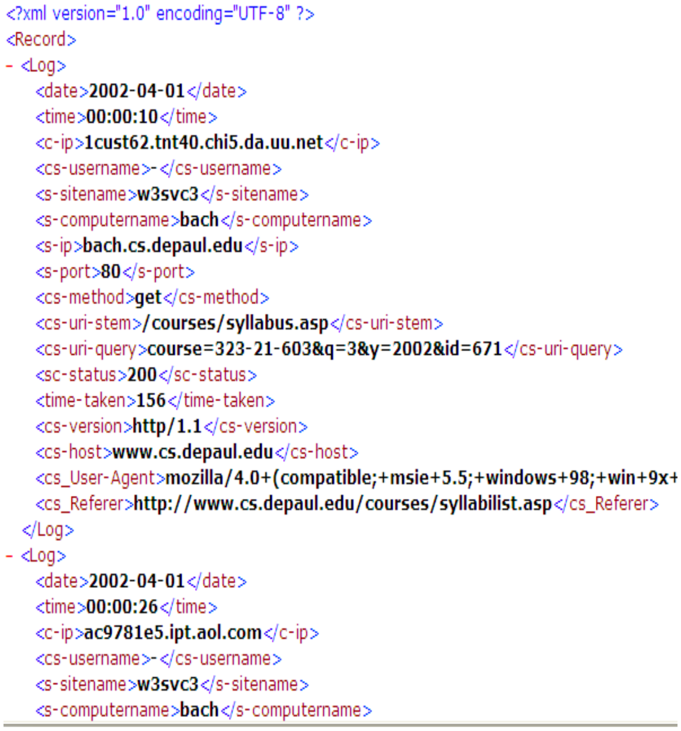
3 **Path Completion:** Path Completion should be used acquiring the complete user access path. The incomplete access path of every user session is recognized based on user session identification. If in a start of user session, Referrer as well URI has data value, delete value of Referrer by adding '-'. Web log preprocessing helps in removal of unwanted click-streams from the log file and also reduces the size of original file by 40-50%.

IV APPROACHES FOR DATA PRE-PROCESSING

A. *Pre-Processing Using Xml*

Data Pre-processing can be done using XML (Extended Markup Language). XML provides a structure to the records which are present in web logs. Hence, understanding of web logs becomes easier. Steps involved in pre-processing using above approach are:

a. Logs recorded in the web log which is a text file are converted into DOM tree structure using XML parsers which is shown below:



```
<?xml version="1.0" encoding="UTF-8" ?>
<Record>
- <Log>
    <date>2002-04-01</date>
    <time>00:00:10</time>
    <c-ip>1cust62.tnt40.chi5.da.uu.net</c-ip>
    <cs-username>-</cs-username>
    <s-sitename>w3svc3</s-sitename>
    <s-computername>bach</s-computername>
    <s-ip>bach.cs.depaul.edu</s-ip>
    <s-port>80</s-port>
    <cs-method>get</cs-method>
    <cs-uri-stem>/courses/syllabus.asp</cs-uri-stem>
    <cs-uri-query>course=323-21-603&q=3&y=2002&id=671</cs-uri-query>
    <sc-status>200</sc-status>
    <time-taken>156</time-taken>
    <cs-version>http/1.1</cs-version>
    <cs-host>www.cs.depaul.edu</cs-host>
    <cs_User-Agent>mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+
    <cs_Referer>http://www.cs.depaul.edu/courses/syllabilist.asp</cs_Referer>
  </Log>
- <Log>
    <date>2002-04-01</date>
    <time>00:00:26</time>
    <c-ip>ac9781e5.ipt.aol.com</c-ip>
    <cs-username>-</cs-username>
    <s-sitename>w3svc3</s-sitename>
    <s-computername>bach</s-computername>
```

Figure 3: DOM tree structure of web log file (Log.txt)

Apply cleansing process ie; remove the records having **status code** above 299 and below 200.Also remove the records in which the attribute **cs_uri_stem** has extensions like css, jpeg, gif.

b. Next step is user identification and session identification is same as given basic algorithm of data pre-processing.

c. Finally, the path completion helps to complete and format the paths in user session, so that these paths can be further used for analysis.

d. After the above steps, transfer the records which are present in XML file into Knowledge base.

B. *Pre-Processing Using Text File*

Data pre-processing is applied on records which are present in the web log file (text file).Steps for preprocessing are:

1) Logs recorded in the web log which are in unprocessed form are shown below:

```
2002-04-01 00:00:10 1cust62.tnt40.chi5.da.uu.net -
w3svc3 bach bach.cs.depaul.edu 80 get
/courses/syllabus.asp course=323-21-
603&q=3&y=2002&id=671 200 156 http/1.1
www.cs.depaul.edu
mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+
9x+4.90;+msn+6.1;+msnbmsft;+msnmen-us;+msnc21)
http://www.cs.depaul.edu/courses/syllabilist.asp
depaul.edu/courses/syllabilist.asp

2002-04-01 00:00:26 ac9781e5.ipt.aol.com - w3svc3 bach
l.edu 80 get /advising/default.asp - 200 16 http/1.1
www.cs.depaul.edu
mozilla/4.0+(compatible;+msie+5.0;+msnia;+windows+9
8;+digext)
http://www.cs.depaul.edu/news/news.asp?theid=573
```

2) Before applying cleansing process, attributes in the text file needs to be separated using delimiter as space. These spaces help in identifying exact position of attributes/fields. Hence, remove the records having status code above 299 and below 200 and records in which the attribute has extensions like css, jpeg, gif.

3) Steps 3 & 4 are same as in above approach.

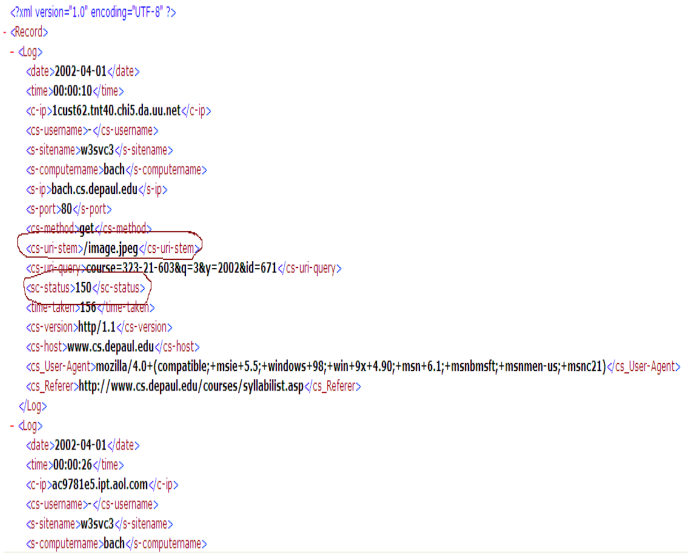
4) After the above steps, transfer the records which are present in text file into Knowledge base.

V RESULTS

Data preprocessing was implemented using two different approaches.

A. Step wise results for XML approach are discussed below:

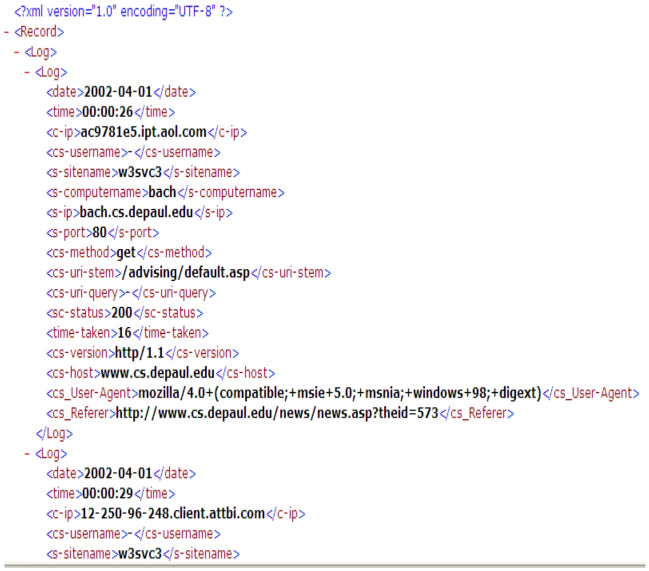
1) When XML approach was applied on unprocessed text file of size 21KB having 500 records, the size of the file increased by 30% (62KB) as it got converted into DOM tree structure. Figure no 4 below shows the record encircled having status code as 150 and cs_uri_stem as .jpeg extension



```
  <?xml version="1.0" encoding="UTF-8" ?>
- <Record>
 - <Log>
    <date>2002-04-01</date>
    <time>00:00:10</time>
    <c-ip>1cust62.tnt40.chi5.da.uu.net</c-ip>
    <cs-username>-</cs-username>
    <s-sitename>w3svc3</s-sitename>
    <s-computername>bach</s-computername>
    <s-ip>bach.cs.depaul.edu</s-ip>
    <s-port>80</s-port>
    <cs-method>get</cs-method>
    <cs-uri-stem>/image.jpeg</cs-uri-stem>
    <cs-uri-query>course=323-21-603&q=3&y=2002&id=671</cs-uri-query>
    <sc-status>150</sc-status>
    <time-taken>156</time-taken>
    <cs-version>http/1.1</cs-version>
    <cs-host>www.cs.depaul.edu</cs-host>
    <cs_User-Agent>mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90;+msn+6.1;+msnbmsft;+msnmen-us;+msnc21)</cs_User-Agent>
    <cs_Referer>http://www.cs.depaul.edu/courses/syllabilist.asp</cs_Referer>
   </Log>
 - <Log>
    <date>2002-04-01</date>
    <time>00:00:26</time>
    <c-ip>ac9781e5.ipt.aol.com</c-ip>
    <cs-username>-</cs-username>
    <s-sitename>w3svc3</s-sitename>
    <s-computername>bach</s-computername>
```

Figure 4: XML file before pre-processing

2) Later, after applying pre-processing, above record is deleted from XML file , thus size of the file gets reduced by 15%(49 KB), which is shown below in figure no 5.

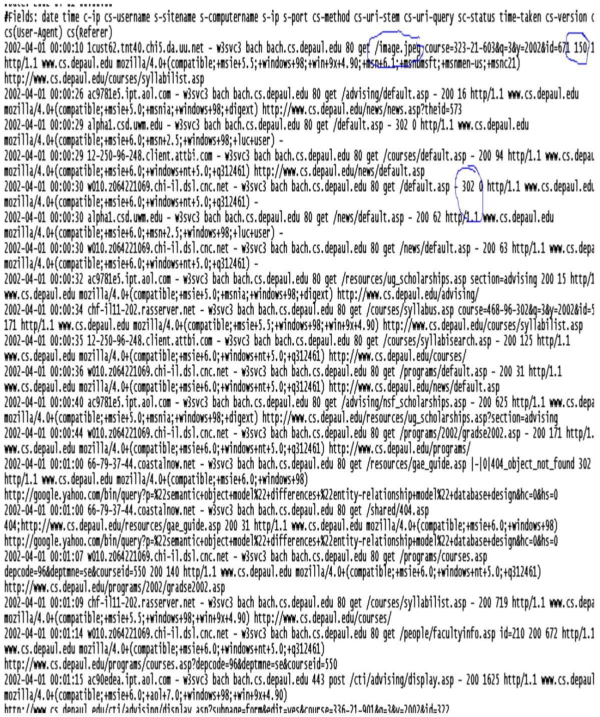


```
  <?xml version="1.0" encoding="UTF-8" ?>
- <Record>
 - <Log>
  - <Log>
     <date>2002-04-01</date>
     <time>00:00:26</time>
     <c-ip>ac9781e5.ipt.aol.com</c-ip>
     <cs-username>-</cs-username>
     <s-sitename>w3svc3</s-sitename>
     <s-computername>bach</s-computername>
     <s-ip>bach.cs.depaul.edu</s-ip>
     <s-port>80</s-port>
     <cs-method>get</cs-method>
     <cs-uri-stem>/advising/default.asp</cs-uri-stem>
     <cs-uri-query>-</cs-uri-query>
     <sc-status>200</sc-status>
     <time-taken>16</time-taken>
     <cs-version>http/1.1</cs-version>
     <cs-host>www.cs.depaul.edu</cs-host>
     <cs_User-Agent>mozilla/4.0+(compatible;+msie+5.0;+msnia;+windows+98;+digext)</cs_User-Agent>
     <cs_Referer>http://www.cs.depaul.edu/news/news.asp?theid=573</cs_Referer>
    </Log>
  - <Log>
     <date>2002-04-01</date>
     <time>00:00:29</time>
     <c-ip>12-250-96-248.client.attbi.com</c-ip>
     <cs-username>-</cs-username>
     <s-sitename>w3svc3</s-sitename>
```

Figure 5: XML file after pre-processing

B. Step wise results for Text File approach are discussed below:

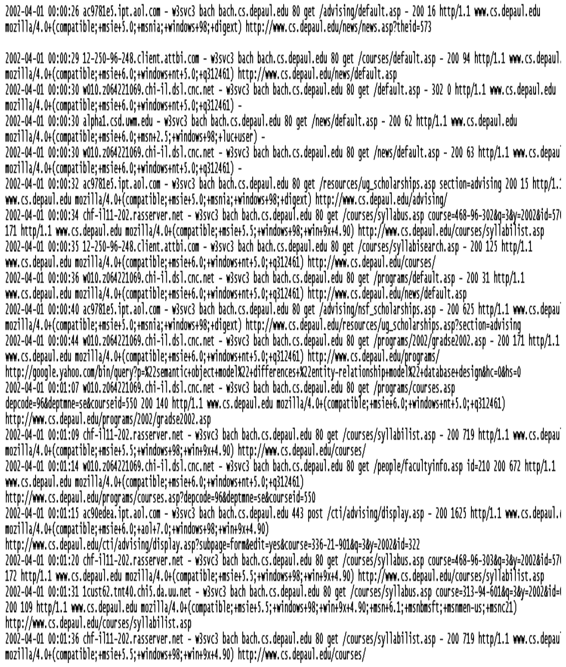
1) In text file approach, direct pre-processing algorithm was applied on unprocessed text file of size 21 KB having 500 records. Text File before pre-processing is figure no 6.



```
#Fields: date time c-ip cs-username s-sitename s-computername s-ip s-port cs-method cs-uri-stem cs-uri-query sc-status time-taken cs-version c
cs(User-Agent) cs(Referer)
2002-04-01 00:00:10 1cust62.tnt40.chi5.da.uu.net - w3svc3 bach bach.cs.depaul.edu 80 get /image.jpeg course=323-21-603&q=3&y=2002&id=671 150 1
http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90;+msn+6.1;+msnbmsft;+msnmen-us;+msnc21)
http://www.cs.depaul.edu/courses/syllabilist.asp
2002-04-01 00:00:26 ac9781e5.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /advising/default.asp - 200 16 http/1.1 www.cs.depaul.edu
mozilla/4.0+(compatible;+msie+5.0;+msnia;+windows+98;+digext) http://www.cs.depaul.edu/news/news.asp?theid=573
2002-04-01 00:00:29 alpha1.csd.uwm.edu - w3svc3 bach bach.cs.depaul.edu 80 get /default.asp - 302 0 http/1.1 www.cs.depaul.edu
mozilla/4.0+(compatible;+msie+6.0;+msn+2.5;+windows+98;+luc+user) -
2002-04-01 00:00:29 12-250-96-248.client.attbi.com - w3svc3 bach bach.cs.depaul.edu 80 get /courses/default.asp - 200 94 http/1.1 www.cs.depau
mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/news/default.asp
2002-04-01 00:00:30 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /default.asp - 302 0 http/1.1 www.cs.depaul.edu
mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) -
2002-04-01 00:00:30 alpha1.csd.uwm.edu - w3svc3 bach bach.cs.depaul.edu 80 get /news/default.asp - 200 62 http/1.1 www.cs.depaul.edu
mozilla/4.0+(compatible;+msie+6.0;+msn+2.5;+windows+98;+luc+user) -
2002-04-01 00:00:30 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /news/default.asp - 200 63 http/1.1 www.cs.depa
mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) -
2002-04-01 00:00:32 ac9781e5.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /resources/ug_scholarships.asp section=advising 200 15 http/1
www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.0;+msnia;+windows+98;+digext) http://www.cs.depaul.edu/advising/
2002-04-01 00:00:34 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=468-96-302&q=3&y=2002&id=5
171 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90) http://www.cs.depaul.edu/courses/syllabilist.asp
2002-04-01 00:00:35 12-250-96-248.client.attbi.com - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabisearch.asp - 200 125 http/1.1
www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/courses/
2002-04-01 00:00:36 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /programs/default.asp - 200 31 http/1.1
www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/news/default.asp
2002-04-01 00:00:40 ac9781e5.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /advising/nsf_scholarships.asp - 200 625 http/1.1 www.cs.depa
mozilla/4.0+(compatible;+msie+5.0;+msnia;+windows+98;+digext) http://www.cs.depaul.edu/resources/ug_scholarships.asp?section=advising
2002-04-01 00:00:44 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /programs/2002/gradse2002.asp - 200 171 http/1.
www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/programs/
2002-04-01 00:01:00 66-79-37-44.coastalnow.net - w3svc3 bach bach.cs.depaul.edu 80 get /resources/gae_guide.asp |-|0|404_object_not_found 302
http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+98)
http://google.yahoo.com/bin/query?p=%22semantic+object+model%22+differences+%22entity-relationship+model%22+database+design&hc=0&hs=0
2002-04-01 00:01:00 66-79-37-44.coastalnow.net - w3svc3 bach bach.cs.depaul.edu 80 get /shared/404.asp
404;http://www.cs.depaul.edu/resources/gae_guide.asp 200 31 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+98)
http://google.yahoo.com/bin/query?p=%22semantic+object+model%22+differences+%22entity-relationship+model%22+database+design&hc=0&hs=0
2002-04-01 00:01:07 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /programs/courses.asp
depcode=96&deptmne=se&courseid=550 200 140 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461)
http://www.cs.depaul.edu/programs/2002/gradse2002.asp
2002-04-01 00:01:09 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabilist.asp - 200 719 http/1.1 www.cs.depa
mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90) http://www.cs.depaul.edu/courses/
2002-04-01 00:01:14 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /people/facultyinfo.asp id=210 200 672 http/1.1
www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461)
http://www.cs.depaul.edu/programs/courses.asp?depcode=96&deptmne=se&courseid=550
2002-04-01 00:01:15 ac90edea.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 443 post /cti/advising/display.asp - 200 1625 http/1.1 www.cs.depaul
mozilla/4.0+(compatible;+msie+6.0;+aol+7.0;+windows+98;+win+9x+4.90)
http://www.cs.depaul.edu/cti/advising/display.asp?subpage=form&edit=yes&course=336-21-901&q=3&y=2002&id=322
```

Figure 6: Text file before pre-processing

2) After applying pre-processing algorithm, size of the original file reduced by approx.10%.(12 KB) is shown below in Figure no 7



```
2002-04-01 00:00:26 ac9781e5.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /advising/default.asp - 200 16 http/1.1 www.cs.depaul.edu
mozilla/4.0+(compatible;+msie+5.0;+msnia;+windows+98;+digext) http://www.cs.depaul.edu/news/news.asp?theid=573

2002-04-01 00:00:29 12-250-96-248.client.attbi.com - w3svc3 bach bach.cs.depaul.edu 80 get /courses/default.asp - 200 94 http/1.1 www.cs.depaul.
mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/news/default.asp
2002-04-01 00:00:30 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /default.asp - 302 0 http/1.1 www.cs.depaul.edu
mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) -
2002-04-01 00:00:30 alpha1.csd.uwm.edu - w3svc3 bach bach.cs.depaul.edu 80 get /news/default.asp - 200 62 http/1.1 www.cs.depaul.edu
mozilla/4.0+(compatible;+msie+6.0;+msn+2.5;+windows+98;+luc+user) -
2002-04-01 00:00:30 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /news/default.asp - 200 63 http/1.1 www.cs.depau
mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) -
2002-04-01 00:00:32 ac9781e5.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /resources/ug_scholarships.asp section=advising 200 15 http/1.
www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.0;+msnia;+windows+98;+digext) http://www.cs.depaul.edu/advising/
2002-04-01 00:00:34 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=468-96-302&q=3&y=2002&id=57
171 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90) http://www.cs.depaul.edu/courses/syllabilist.asp
2002-04-01 00:00:35 12-250-96-248.client.attbi.com - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabisearch.asp - 200 125 http/1.1
www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/courses/
2002-04-01 00:00:36 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /programs/default.asp - 200 31 http/1.1
www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/news/default.asp
2002-04-01 00:00:40 ac9781e5.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 80 get /advising/nsf_scholarships.asp - 200 625 http/1.1 www.cs.depau
mozilla/4.0+(compatible;+msie+5.0;+msnia;+windows+98;+digext) http://www.cs.depaul.edu/resources/ug_scholarships.asp?section=advising
2002-04-01 00:00:44 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /programs/2002/gradse2002.asp - 200 171 http/1.1
www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461) http://www.cs.depaul.edu/programs/
http://google.yahoo.com/bin/query?p=%22semantic+object+model%22+differences+%22entity-relationship+model%22+database+design&hc=0&hs=0
2002-04-01 00:01:07 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /programs/courses.asp
depcode=96&deptmne=se&courseid=550 200 140 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461)
http://www.cs.depaul.edu/programs/2002/gradse2002.asp
2002-04-01 00:01:09 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabilist.asp - 200 719 http/1.1 www.cs.depau
mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90) http://www.cs.depaul.edu/courses/
2002-04-01 00:01:14 w010.z064221069.chi-il.dsl.cnc.net - w3svc3 bach bach.cs.depaul.edu 80 get /people/facultyinfo.asp id=210 200 672 http/1.1
www.cs.depaul.edu mozilla/4.0+(compatible;+msie+6.0;+windows+nt+5.0;+q312461)
http://www.cs.depaul.edu/programs/courses.asp?depcode=96&deptmne=se&courseid=550
2002-04-01 00:01:15 ac90edea.ipt.aol.com - w3svc3 bach bach.cs.depaul.edu 443 post /cti/advising/display.asp - 200 1625 http/1.1 www.cs.depaul.
mozilla/4.0+(compatible;+msie+6.0;+aol+7.0;+windows+98;+win+9x+4.90)
http://www.cs.depaul.edu/cti/advising/display.asp?subpage=form&edit=yes&course=336-21-901&q=3&y=2002&id=322
2002-04-01 00:01:20 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=468-96-303&q=3&y=2002&id=57
172 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90) http://www.cs.depaul.edu/courses/syllabilist.asp
2002-04-01 00:01:31 1cust62.tnt40.chi5.da.uu.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabus.asp course=313-94-601&q=3&y=2002&id=
200 109 http/1.1 www.cs.depaul.edu mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90;+msn+6.1;+msnbmsft;+msnmen-us;+msnc21)
http://www.cs.depaul.edu/courses/syllabilist.asp
2002-04-01 00:01:36 chf-il11-202.rasserver.net - w3svc3 bach bach.cs.depaul.edu 80 get /courses/syllabilist.asp - 200 719 http/1.1 www.cs.depau
mozilla/4.0+(compatible;+msie+5.5;+windows+98;+win+9x+4.90) http://www.cs.depaul.edu/courses/
```

Figure 7: Text file after pre-processing

VI COMPARISON AND APPLICATION OF THESE APPROACHES

A. XML Approach

1 Since, DOM tree structure is used, preprocessing stages can be analyzed very well

2 Time taken for Conversion of records from XML to Knowledge base was 20 minutes when the above approach was used.

3 XML approach can be used when the Web log file consists of more number of attributes describing usage profile of user as in IIS Web Server having Extended Log File Format having 17 attributes as shown in section 3.1

B. Text File Approach

1 Understanding of each step of pre-processing would be difficult for user because this approach demands analysis and knowledge of how weblog looks.

2 Time taken to transfer records from text file to Knowledge Base was 10 seconds when the above approach was used.

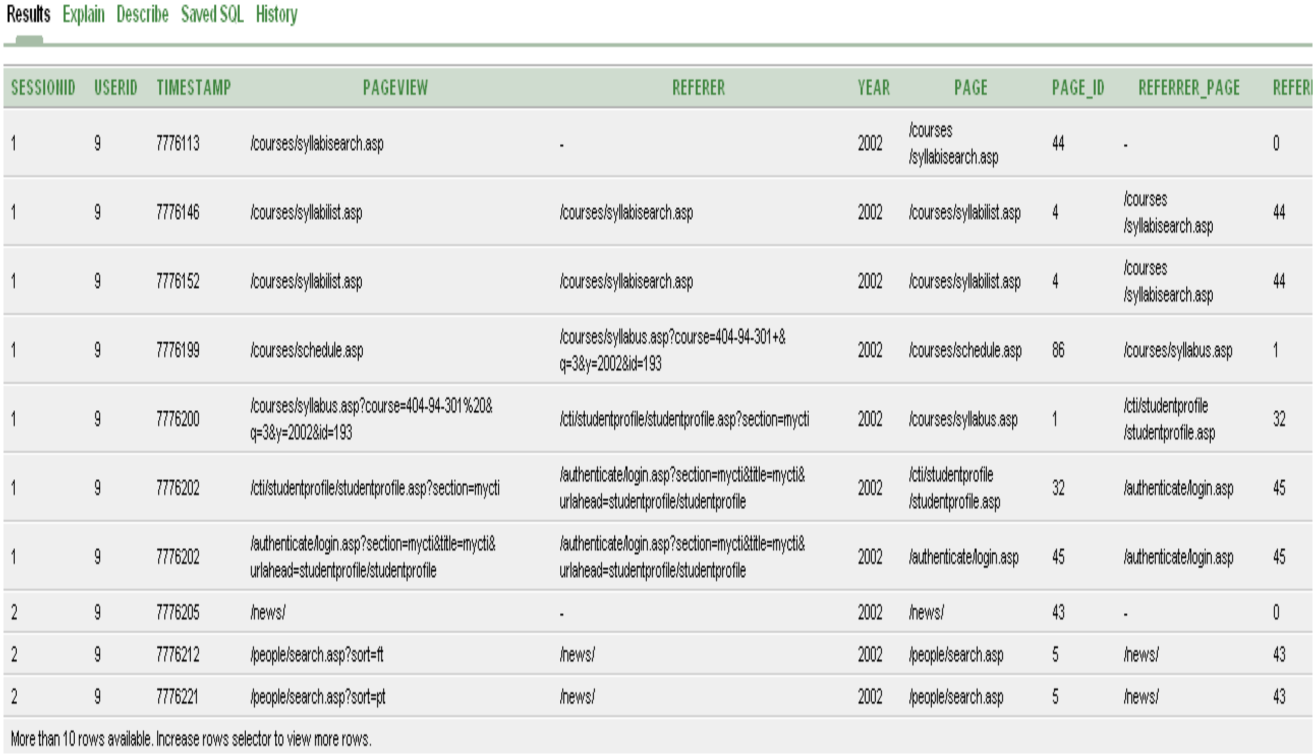
3 Text File approach can be used when the web log file consist of very few attributes describing usage profile of users ie; less than 10 as in Common Log Format as shown below

```
127.0.0.1 - frank [10/Oct/2000:13:55:36
-0700] "GET /apache_pb.gif HTTP/1.0"
200 2326
```

Or Combined Log format as shown below

```
127.0.0.1 - frank [10/Oct/2000:13:55:36
-0700] "GET /apache_pb.gif HTTP/1.0"
2002326
"http://www.example.com/start.html"
"Mozilla/4.08 [en] (Win98; I ;Nav)"
```

Records which are present in Knowledge base after pre-processing using both approaches is shown below in fig.8



SESSIONID	USERID	TIMESTAMP	PAGEVIEW	REFERER	YEAR	PAGE	PAGE_ID	REFERRER_PAGE	REFERI
1	9	7776113	/courses/syllabisearch.asp	-	2002	/courses /syllabisearch.asp	44	-	0
1	9	7776146	/courses/syllabilist.asp	/courses/syllabisearch.asp	2002	/courses/syllabilist.asp	4	/courses /syllabisearch.asp	44
1	9	7776152	/courses/syllabilist.asp	/courses/syllabisearch.asp	2002	/courses/syllabilist.asp	4	/courses /syllabisearch.asp	44
1	9	7776199	/courses/schedule.asp	/courses/syllabus.asp?course=404-94-301+& q=3&y=2002&id=193	2002	/courses/schedule.asp	86	/courses/syllabus.asp	1
1	9	7776200	/courses/syllabus.asp?course=404-94-301%20& q=3&y=2002&id=193	/cti/studentprofile/studentprofile.asp?section=mycti	2002	/courses/syllabus.asp	1	/cti/studentprofile /studentprofile.asp	32
1	9	7776202	/cti/studentprofile/studentprofile.asp?section=mycti	/authenticate/login.asp?section=mycti&title=mycti& urlahead=studentprofile/studentprofile	2002	/cti/studentprofile /studentprofile.asp	32	/authenticate/login.asp	45
1	9	7776202	/authenticate/login.asp?section=mycti&title=mycti& urlahead=studentprofile/studentprofile	/authenticate/login.asp?section=mycti&title=mycti& urlahead=studentprofile/studentprofile	2002	/authenticate/login.asp	45	/authenticate/login.asp	45
2	9	7776205	/news/	-	2002	/news/	43	-	0
2	9	7776212	/people/search.asp?sort=ft	/news/	2002	/people/search.asp	5	/news/	43
2	9	7776221	/people/search.asp?sort=pt	/news/	2002	/people/search.asp	5	/news/	43

More than 10 rows available. Increase rows selector to view more rows.

Figure 8: Records are arranged user session wise after pre-processing

VII. CONCLUSION

In this paper, we have discussed two approaches for data pre-processing. Results of both the approaches are given and also a comparison is done. Data pre-processing is an important step in the knowledge discovery process, because quality decisions are based on quality data. Hence, we have made an attempt to present these approaches in this paper.

VIII. REFERENCES

[1] Google Website. http://www.google.com.

[2] R. Kosala, H. Blockeel. “Web Mining Research: A Survey,” In SIGKDD Explorations, ACM press, 2(1): 2000, pp.1-15.

[3] R. Srikant, R. Agrawal. “Mining sequential patterns:Generalizations and performance improvements,” In 5th International Conference Extending Database Technology, Avignon, France, March1996, pp. 13-17..

[4] W.W.W Consortium the CommonLogFileformat http://www.w3.org/Daemon/User/Config/ Logging.html#common-Log file-format, (1995)

[5] W3C Extended Log File Format, Available at http://www.w3.org/TR/WD-logfile.html (1996).

[6] R. Cooley, B. Mobasher and J. Srivastava. “Data Preparation for Mining World Wide Web Browsing Patterns,” In Journal of Knowledge and Information Systems, vol. 1, no. 1, 1999. pp. 5-32.

[7] Jiawei Han and M. Kamber. “Data Mining: Concepts and Techniques,” In Morgan Kaufmann publishers, 2001[8] ZY COMPUTING-2003 ,123 Log Analyzer. San Jose USA. Available at http://www.123loganalyzer.com

[8] B. Mobasher, H. Dai, T. Luo and M. Nakagawa. “ Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization,” In proceedings of Data Mining and Knowledge Discovery,2002, pp 61-82.

[9] R. Kosala, H. and Blockeel. “Web mining research: a survey,”Inproceedings of special interest group on knowledge discovery & data mining, SIGKDD:2000 , ACM 2 (1), pp.1–15. [11] R. Kohavi and R. Parekh. “Ten supplementary analyses to improve e-commerce web sites,” In Proceedings of the Fifth WEBKDD workshop, (2003).

[10] B. Mobasher, R. Cooley and J. Srivastava. “CreatingAdaptive Web Sites through usage based clustering of URLs,” In proceedings of Knowledge and Data Engineering Exchange, 1999, Volume 1, Issue1, 1999, pp.19-25.

[11] J. Srivastava, R. Cooley, M. Deshpande and P. N. Tan.“Web usage mining: discovery and applications of usage patterns from web data,” In SIGKDD Explorations, 2002, pp. 12–23.

[12] B. Berendt, B. Mobasher, M. Nakagawa and M. Spiliopoulou. “The Impact of Site Structure and User Environment on Session reconstruction in Web Usage Analysis,” In Proceedings of the Forth WebKDD 2002 Workshop, At the ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD'2002), Edmonton, Alberta, Canada,pp.1-13.

[13] C. Marquardt, K. Becker and D. Ruiz. “A Pre-Processing Tool for Web Usage Mining in the Distance Education Domain,” In Proceedings of the International Database Engineering and Applications, IDEAS 2004, pp. 78-87.

[14] Zaïane, O.R. Xin and M. Han. “Discovering Web Access Patterns andTrends by Applying OLAP and Data Mining Technology on Web Logs,” In Proceedings of Advances in Digital Libraries Conference(1998), pp. 19-29.

[15] Kamal A. ElDahshan, Hany Maher Said Lala Kamal “Data Warehouse based Statistical Mining,” In ICGST-AIML Journal, ISSN: 1687-4846, Volume 9, Issue I, February (2009), pp.41-48.