DOI: 10.24996.ijs.2017.58.2C.17

ISSN: 0067-2904

Identifying of User Behavior from Server Log File

Wajih Abdul Ghani Abdul Hussain
Department of Computer, College of Science, University of Baghdad, Baghdad, Iraq.

Abstract

Due to the increased of information existing on the World Wide Web (WWW), the subject of how to extract new and useful knowledge from the log file has gained big interest among researchers in data mining and knowledge discovery topics.

Web miming, which is a subset of data mining divided into three particular ways, web content mining, web structure mining, web usage mining. This paper is interested in server log file, which is belonging to the third category (web usage mining). This file will be analyzed according to the suggested algorithm to extract the behavior of the user. Knowing the behavior is coming from knowing the complete path which is taken from the specific user.

Extracting these types of knowledge required many of KDD (Knowledge Discovery in Database) steps such as preprocessing, pattern discovery, and pattern analysis. After that, the complete graph of the visited web will be drawn. The knowledge discussed in this paper, helps the web designers to improve their web site design and helps to improve their website usability and visitor's browsing experience by determining related link connections in the website.

Keywords: Data Mining, Web Mining, KDD, Web Usage Mining, Log File.

معرفة سلوك المستخدم من خلال ملف تسجيل الخادم

وجيه عبد الغني عبد الحسين
قسم علوم الحاسبات ، كلية العلوم ، جامعة بغداد، بغداد، العراق.

الخلاصة

مع تزايد المعلومات المتوفرة على الشبكة العنكبوتية (World Wide Web) فان استخلاص المعرفة من هذا الكم الهائل من البيانات اصبح محط اهتمام المحللين ضمن ابحاث تنقيب البيانات واكتشاف المعرفة.

تنقيب الويب والذي هو جزء من تنقيب البيانات ينقسم الى ثلاثة اقسام، التنقيب عن محتوى الويب، التنقيب عن هيكلية الويب، والتنقيب عن استخدام الويب. هذا البحث يُعنى بملف تسجيل الخادم والذي ينتمي الى القسم الثالث (التنقيب عن استخدام الويب).

هذا الملف سيتم تحليله بالاعتماد على خوارزمية مقترحة من اجل استخلاص سلوك المستخدم. معرفة السلوك يأتي من خلال معرفة المسار الكامل الذي اتخذه المستخدم.

استخلاص هذه الانواع من المعرفة تتطلب عدد من خطوات الـ(KDD) (عملية اكتشاف المعرفة من قواعد البيانات) مثل المعالجة الاولية، اكتشاف الانماط، تحليل الانماط. بعد ذلك سيتم رسم المخطط الكامل لصفحات الويب التي زارها المستخدم اثناء تجواله على الانترنيت.

Email: wageh_82@yahoo.com

المعرفة المكتشفة في هذا البحث تساعد مصممي صفحات الويب على تحسين تصميم صفحات الويب
بالشكل الامثل وكذلك تساعد على الاستعمال الافضل لصفحات الويب من خلال وضع مسار مباشر بين
الصفحات الاكثر ترابطاً على الويب.

Introduction

Web mining is the application of data mining techniques on the web data existed on WWW (World Wide Web). Web mining is divided into three distinct ways, web content mining, web structure mining, and web usage mining. This paper will be concerned with web usage mining, which is discovering the hidden information founded in log file. One of hidden information discovered from log file is extracted the complete path which is taken from the user during navigation time.

Structure of Server Log File

A server log file is a textual file, independent of server platform, in which a Web server enters a record whenever a user requests for a resource [1]. Server log files record activities of the server, web administrator can use these logs to monitor the server and to help troubleshooting if necessary. [2]

With log file analysis tools, it's possible to get a good idea of where visitors are coming from, how often they return, and how they navigate through a site. Using cookies enables web master to log even more detailed information about how individual users are accessing a site. The user access log has very significant information about a Web server [3]. Figure- 1 shows a sample of server log file.

```
147.91.173.31 [16/Nov/2009:00:02:24+0100] "style.css" 200 6554 "http://www.vtsns.edu.rs/"
"Mozilla/5.0(Windows;U;WindowsNT5.1;sr;rv:1.9.1.5)Gecko/20091102Firefox/3.5.5"

147.91.173.31 [16/Nov/2009:00:02:51+0100] "vesti.php" 200 3367
"http://www.vtsns.edu.rs/konsultacije.php"
"Mozilla/5.0(Windows;U;WindowsNT5.1;sr;rv:1.9.1.5)Gecko/20091102Firefox/3.5.5"
```

Figure 1- Sample of Server Log File

A log file contains a complete history of web pages accessed by users. By analyzing these logs, it is possible to discover various kinds of knowledge, which can be applied to improve the performance of Web services. [3]

A log file contains a sequence of lines containing ASCII characters. Each line may contain either a directive or an entry. Entries consist of a sequence of fields relating to a single HTTP transaction. Fields are separated by white space or comma or hash. If a field is unused in a particular entry dash "-" marks the omitted field. [4]

Traditionally there are four types of server logs:

1. Transfer Log or Access Log.
2. Agent Log.
3. Error Log.
4. Referrer Log.

The first two types of log files 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. Each HTTP protocol transaction, whether completed or not, is recorded in the logs and some transactions are recorded in more than one log. [5]

Format of Log File

Web Log File comes in various formats, which vary depending on the configuration of the web server. [5]

- W3C Extended Log File Format,
- NCSA Common Log File Format,
- IIS Log File Format.

In addition to the three available formats, custom log file format can also be configured. Besides that there are two formats, which is **Common Log File** and **Extended Common Log File** Format, The common log format (CLF or “clog”) is supported by a variety of web server applications and includes the following seven fields:

- Remote host field
- Identification field
- Authuser field
- Date/time field
- HTTP request
- Status code field
- Transfer volume field

A log file in the all formats explained above contains a sequence of lines containing ASCII characters. Each line may contain either a directive or an entry. Entries consist of a sequence of fields relating to a single HTTP transaction. Fields are separated by white space or comma or hash. If a field is unused in a particular entry dash "-" marks the omitted field. [6]

The extended common log format (ECLF) is a variation of the common log format, formed by appending two additional fields onto the end of the record, the **referrer field**, and the **user agent field** [7].

Fields of Log File

Here we will explain each field in the server log file, depending on the extended log file format which are as follows :- [7]

1) Remote Host Field

This field consists of the Internet IP address of the remote host making the request, such as "141.243.1.172". If the remote host name is available through a DNS lookup, this name is provided, such as "wpbfl2-45.gate.net."

2) Identification Field

This field is used to store identity information provided by the client only if the web server is performing an identity check.

3) Authuser Field

This field is used to store the authenticated client user name, if it is required.

4) Date/Time field

This field contains the date and time of the request from the user's browser to the web server.

5) HTTP Request Field

The HTTP request field consists of the information that the client's browser has requested from the web server. The entire HTTP request field is contained within quotation marks. Essentially, this field may be partitioned into four areas: the request method, the uniform resource identifier (URI), the header, and the protocol.

The most common request method is GET, which represents a request to retrieve data that are identified by the URI, Besides GET, other requests include HEAD, PUT, and POST. The uniform resource identifier contains the page or document name and the directory path requested by the client browser. The URI can be used by web usage miners to analyze the frequency of visitor requests for pages and files. For example, the request field "GET /Software.html HTTP/1.0," representing a request from the client browser for the web server to provide the web page Software.html. The header section contains optional information concerning the browser's request.

6) Status Code Field

Not all browser requests succeed. The status code field provides a three-digit response from the web server to the client's browser, indicating the status of the request, whether or not the request was a success, or if there was an error, which type of error occurred. Codes of the form "2xx" indicate a success, and codes of the form "4xx" indicate an error.

7) Transfer Volume (Bytes) Field

The transfer volume field indicates the size of the file (web page, graphics file, etc.), in bytes, sent by the web server to the client's browser.

8) Referrer Field

The referrer field lists the URL of the previous site visited by the client, which linked to the current page.

9) User Agent Field

The user agent field provides information about the client's browser, the browser version, and the client's operating system.

Related Work

Among the large published research works in the field of web usage mining, some of these works are introduced below:

- In 2004, Azizul Azhar bin Ramli produced useful results for analyzing the web usage pattern for University Utara Malaysia (UUM Elearning) web site, his study used the basic association rules-Apriori algorithm. The outcomes are used by web administrator to plan necessary improvement, enhancement and valuable actions to the web site [2].
- In 2009, K. R. Suneetha, R. Krishnamoorthi were concerned with the in-depth analysis of Web log data of NASA website to find information about a web site, top errors, potential visitors of the site which help system administrator and Web designer to improve their system by determining occurred systems errors, corrupted and broken links by using web usage mining [8].
- In 2011, Kiruthika M, Rahul Jadhav, Dipa, Rashmi J, Anjali Nehete, Trupti Khodkar were discussed how association rules can be used to discover patterns in web usage mining. This discussion starts with preprocessing of the given weblog, followed by clustering them and finding association rules. These rules provide knowledge that helps to improve website design, in advertising, web personalization [3].
- In 2009, Zidrina Pabarskaite in Ph.D. dissertation developed methods to improve knowledge discovery steps mining in Web log data that would reveal new opportunities to the data analyst [9].

Proposed Miner

from the client during the web browsing. This miner uses C sharp or C# programming language as a tool for solving the problems of the practical part of the paper. In addition it uses SQL Server 2005 as a data repository (database). Figure- 2 depicts the phases of this miner.

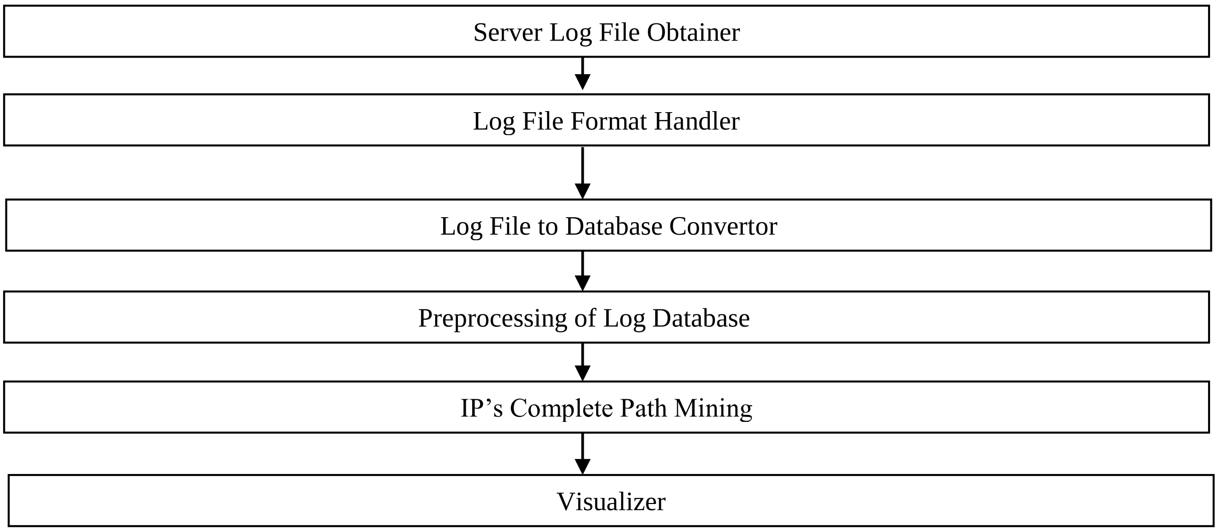

Server Log File Obtainer
Log File Format Handler
Log File to Database Convertor
Preprocessing of Log Database
IP’s Complete Path Mining
Visualizer

Figure 2- Operation flow of converting and analysis of log file

1) Server Log File Obtainer

In this miner, the log file was collected from the official website of the advanced school of technology in Novi Sad made in November 2009. It exists on the web site http://www.vtsns.edu.rs/maja. The file follows extended common log format for the analysis purposes, The raw log files consist of 12 attributes such as *Client IP, Identification, Auth User, Date & Time, Request Method, URI-Stem, Protocol Version, Status Code, Size in Bytes, Referrer, User Agent*. A sample of a single entry log file is displayed in Figure- 3.

```
147.91.173.31 - - [16/Nov/2009:00:02:51 +0100] "GET /vesti.php HTTP/1.0" 200 3367
"http://www.vtsns.edu.rs/konsultacije.php" "Mozilla/5.0 (Windows; U; Windows NT 5.1; sr;
rv:1.9.1.5) Gecko/20091102 Firefox/3.5.5"
```

Figure 3- Single entry of raw log file follows ECLF.

By matching the above attributes with the Figure- 3, we notice that:

Client IP	: 147.91.173.31
Identification	: -
Auth User	: -
Date & Time	: [16/Nov/2009:00:02:51 +0100]
Request Method	: GET
URI-Stem	: /vesti.php
Protocol Version	: HTTP/1.0
Status Code	: 200
Size in Bytes	: 3367
Referrer	: http://www.vtsns.edu.rs/konsultacije.php
User Agent	: Mozilla/5.0 (Windows; U; Windows NT 5.1; sr; rv:1.9.1.5) Gecko/20091102 Firefox/3.5.5

2) Log File Format Handler

In this block, the format of log file and separator character must be determined, indeed, there are many characters are used as a delimiter (split character) in the log file such as comma, space, hash and others. Therefore, this character should be determined to the miner.

3) Log File to Database Convertor

The selected log file should be converted to a database table with designated structure. The conversion process requires knowing of the log file format and the delimiter character which separate among the log file fields. This database contains a table for each format. Figure- 4 shows the suggested algorithm for the converting data from text file to the database.

```
Input: Log File
Output: Log Table (LT)
Begin
1. Open a DB connection
2. Create a table to store log data
3. Open Log File
4. While not end of log file
5.     Read an entry of log file
6.     Tokenize the fields depending on delimiter char.
7.     Insert all fields into the Log Table (LT)
8. End while
9. Close a DB connection and Log File
End.
```

Figure 4-Suggested algorithm for converting text file to database

4) Preprocessing of Log Database

From the technical point of view, Web usage mining is the application of data mining techniques to usage logs of large data repositories. The purpose of it is to produce results that can be used to improve and optimize the content of a site. In this phase, the critical point for successful log mining is data preprocessing as shown in Figure- 5.

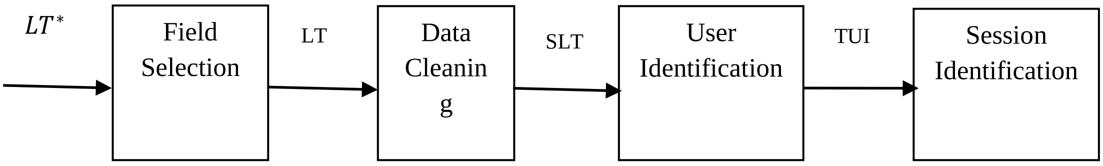


Figure 5- Preprocessing Steps in web usage mining

- **Field Selection**

The log entry, (now exist in the database), contains various fields which are not interesting in specified mining operation, for example, the miner will be concerned in the fields such as IP address, date and time of the request, requested page and referrer fields. According to rule mining, other fields may be important in other mining tasks.

- **Data Cleaning**

Data cleaning removes unnecessary items in the log file. A web site can be accessed by thousands of users. The failed requests from the client server also involve in log file(failed status code), also log file embedded objects that may not be important for the purpose of analysis, including references to style files(.css), graphics or sound files(.jpg, .gif, .mp3). Therefore, some of the entries are useless for analysis process that is cleaned from the log files. A suggested algorithm for cleaning the entries of server logs is presented in Figure- 6.

```
Input: Log Table (LT)
Output: Summarized Log Table (SLT)
Begin
For each record in LT
    Read fields (Status code, method);
    If suffix.URL_Link is not required Then
    Remove this record from LT;
    If Status code ='200'and method= 'GET' Then
       Get IP_address and URL_link;
End For;
End.
```

Figure 6-Suggested algorithm for data cleaning

- **User Identification**

After all previous steps, now we must recognize and identify each user separately from the others. This recognition will be according to the IP address. The goal of user identification algorithm Figure- 7 is to reconstruct, from the clickstream data, the actual sequence of actions performed by one user during one visit to the site. We do that to facilitate applying the mining techniques on the log file.

```
Input: Log table without user identification
 (summarize log table) (SLT)
Output: Log table with user identification (TUI)
Begin
  i = 1, session = 1;
  While (i < Count of Records in SLT) Do
     Add (Record(i)) to TUI;
     For j = ++i To Count of Records in SLT Do
        IF (Record(i).IP = Record(j).IP) Then
           Add (Record(j)) to TUI;
        Else IF (Record(j).IP $\in$ TUI.IPs) Then
           Continue;
        Else {
           Increment session;
           i = j; }
     End For;
  End While;
End.
```

Figure 7- User identification algorithm

- **Session Identification**

Session identification divides all the pages accessed by a user into different sessions. Depending on different approaches, some of these approaches are:

- **Session Identification by the Time Gap**

The most popular session identification technique uses time gap between entries. If the time gap between two pages requests made by the same user exceeds some threshold (Ex. 30 min.), a new session is cre ated. In table1 (which is proposed table), the user (which has IP address 192.168.1.20) first request given 0:12 and last request given 0:55, then the difference between them > 30 minutes. So this table must be divided into two sessions.

Table 1-proposed info. to demonstrate session iden. By time gap.

Time	Ip Address	URL	REFF	Agent
0.12	192.168.1.20	A	-	IE6;XP
0.15	192.168.1.20	B	A	IE6;XP
0.20	192.168.1.20.	C	B	IE6;XP
0.25	192.168.1.20	D	C	IE6;XP
0.35	192.168.1.20	D	C	IE6;XP
0.45	192.168.1.20	E	D	IE6;XP
0.49	192.168.1.20	F	C	IE6;XP
0.55	192.168.1.20	G	F	IE6;XP

Time	IP Address	URL	REFF	Agent
0:12	192.168.1.20	A	-	IE6;XP
0:15	192.168.1.20	B	A	IE6;XP
0:20	192.168.1.20	C	B	IE6;XP
0:25	192.168.1.20	D	C	IE6;XP
0:35	192.168.1.20	D	C	IE6;XP

Session1

Time	IP Address	URL	REFF	Agent
0:45	192.168.1.20	E	D	IE6;XP
0:49	192.168.1.20	F	C	IE6;XP
0:55	192.168.1.20	G	F	IE6;XP

Session2

- ***Session Identification by the Structure Oriented***

This type use either the static site structure or the implicit linkage structure captured in the referrer fields of the server logs, table2 shows this fact.

Table 2- proposed info. to demonstrate session iden. by structure oriented.

Time	IP	URL	Ref
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:15	1.2.3.4	A	-
1:26	1.2.3.4	F	C
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 1

Time	**IP**	**URL**	**Ref**
0:01	1.2.3.4	A	-
0:09	1.2.3.4	B	A
0:19	1.2.3.4	C	A
0:25	1.2.3.4	E	C
1:26	1.2.3.4	F	C

Time	**IP**	**URL**	**Ref**
1:15	1.2.3.4	A	-
1:30	1.2.3.4	B	A
1:36	1.2.3.4	D	B

Session 2

Once the request for F (with time stamp 1:26) is reached, there are two open sessions, namely, A B C E and A. But F is added to the first because its referrer, C, was invoked in session1. The request for B (time stamp 1:30) may potentially belong to both open sessions, since its referrer, A, is invoked both in session 1 and in session 2. In this case, it is added to the second session, since it is the most recently opened session.

5) IP's Complete Path Mining

One of the most important knowledge in log table is the mining of complete path, the way which the user takes when navigating through web site. This path will clarify the actual user behavior on the web site. Three fields must be selected from log file to achieve this type of mining; the interested fields are IP address, requested page, and referrer page. Figure -8 shows the suggested algorithm for finding complete path.

```
Input: Table contains IP Address, URL, Referrer fields (TI).
Output: Table contains complete path for each IP Address (TO).
Begin
  For i = 0 To TI.Rows.Count Do
     GET IP(i);
     IF ( IP(i) ∉ TO ) Then
        GET Page(i) , Referrer(i);
        Add Referrer(i) , "➔" , Page(i) to TO;
         j = i + 1;
        While ( j < TI.Rows.Count ) Do
           GET IP(j);
           IF ( IP(i) = IP(j) )  Then
           GET Referrer(j);
           IF ( Page(i) = Refferer(j) ) Then
              Add "➔" , Page(j) to TO;
           Else IF (Referrer(i) exist in previous entries) Then
              {Add "Back" , "➔" , Referrer(j) , "➔" , Page(j) to TO;
               j++; }
           Else
              User starts a new session;
        End While;
   End For;
End.
```

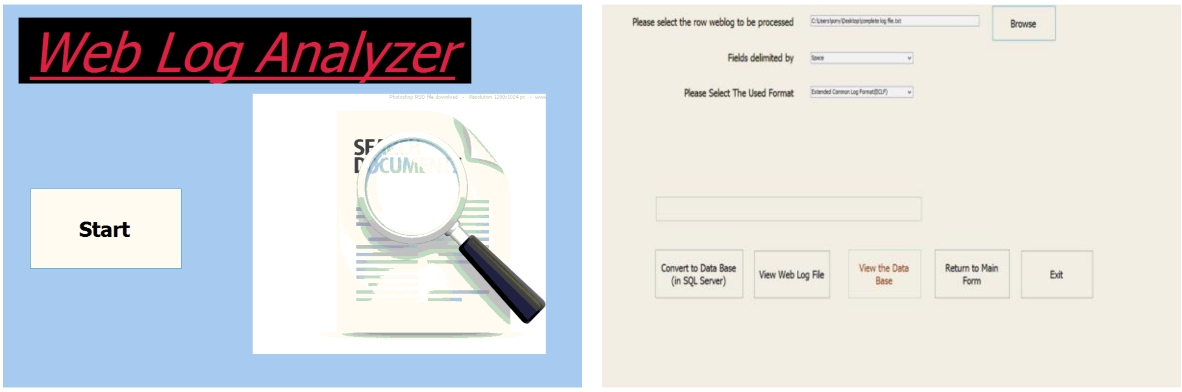
Figure 8- Suggested algorithm for finding the completing path for each IP Address.

Design and Experimental Results of Access Log File Miner

The proposed miner must be involved the following steps:

1- Obtain server log file and format handler

When we running the project, the first interface will appear, we must press *(Start)* button to show the second interface as shown in Figure- 9. In this interface, the format of log file and separator character must be determined, indeed, there are many characters used as a delimiter (split character) in log file such as comma, space, hash and others. Therefore, this character should be determined to the miner. Also, the log file format such as Common Log Format (CLF) or Extended Common Log Format (ECLF) should be determined to the miner.



Please select the row weblog to be processed: C:\Users\sony\Desktop\complete log file.txt

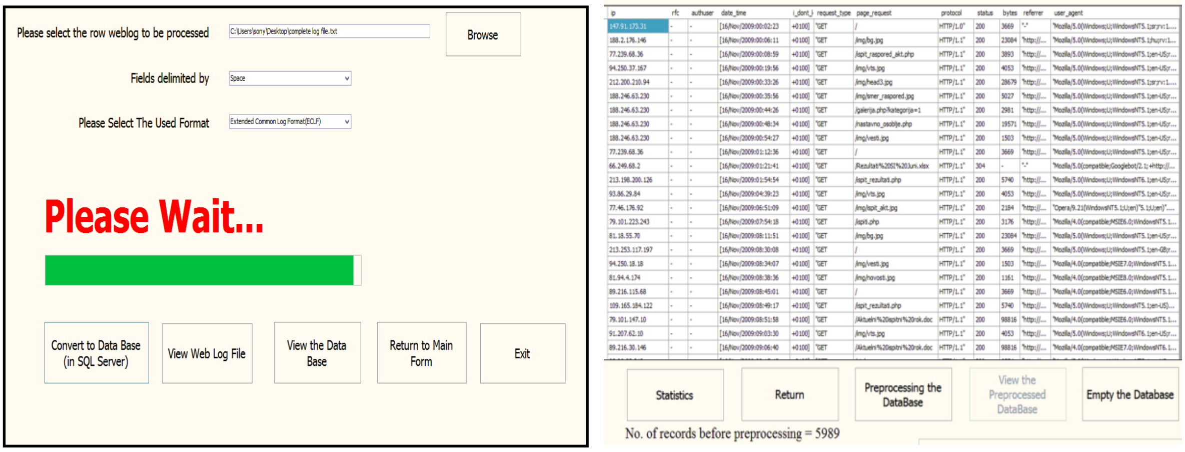
Fields delimited by: Space

Please Select The Used Format: Extended Common Log Format(ECLF)

Figure 9- the first and second interfaces of the miner.

2- The log file is then stored into a database.

After reading the server log file, the server log data will be transferred to SQL Server relational database in order to make it appropriate to apply the data mining techniques in the next phase of the process. This happen when we press on *(Convert to Data Base)* button as shown in Figure- 10.



Please select the row weblog to be processed: C:\Users\sony\Desktop\complete log file.txt

Fields delimited by: Space

Please Select The Used Format: Extended Common Log Format(ECLF)

ip	rfc	authuser	date_time	i_dont_	request_type	page_request	protocol	status	bytes	referrer	user_agent
147.91.173.31	-	-	[16/Nov/2009:00:02:23	+0100]	"GET	/	HTTP/1.0"	200	3669	"-"	"Mozilla/5.0(Windows;U;WindowsNT5.1;sr;rv:1....
188.2.176.146	-	-	[16/Nov/2009:00:06:11	+0100]	"GET	/img/bg.jpg	HTTP/1.1"	200	23084	"http://...	"Mozilla/5.0(Windows;U;WindowsNT5.1;hu;rv:1....
77.239.68.36	-	-	[16/Nov/2009:00:08:59	+0100]	"GET	/ispit_raspored_akt.php	HTTP/1.1"	200	3893	"http://...	"Mozilla/5.0(Windows;U;WindowsNT5.1;en-US;r...
94.250.37.167	-	-	[16/Nov/2009:00:19:56	+0100]	"GET	/img/vts.jpg	HTTP/1.1"	200	4053	"http://...	"Mozilla/5.0(Windows;U;WindowsNT5.1;en-US;r...
212.200.210.94	-	-	[16/Nov/2009:00:33:26	+0100]	"GET	/img/head3.jpg	HTTP/1.1"	200	28679	"http://...	"Mozilla/5.0(Windows;U;WindowsNT5.1;sr;rv:1....
188.246.63.230	-	-	[16/Nov/2009:00:35:56	+0100]	"GET	/img/smer_raspored.jpg	HTTP/1.1"	200	5027	"http://...	"Mozilla/5.0(Windows;U;WindowsNT5.1;en-US;r...
188.246.63.230	-	-	[16/Nov/2009:00:44:26	+0100]	"GET	/galerija.php?kategorija=1	HTTP/1.1"	200	2981	"http://...	"Mozilla/5.0(Windows;U;WindowsNT5.1;en-US;r...
188.246.63.230	-	-	[16/Nov/2009:00:48:34	+0100]	"GET	/nastavno_osoblje.php	HTTP/1.1"	200	19571	"http://...	"Mozilla/5.0(Windows;U;WindowsNT5.1;en-US;r...
188.246.63.230	-	-	[16/Nov/2009:00:54:27	+0100]	"GET	/img/vesti.jpg	HTTP/1.1"	200	1503	"http://...	"Mozilla/5.0(Windows;U;WindowsNT5.1;en-US;r...
77.239.68.36	-	-	[16/Nov/2009:01:12:36	+0100]	"GET	/	HTTP/1.1"	200	3669	"http://...	"Mozilla/5.0(Windows;U;WindowsNT5.1;en-US;r...
66.249.68.2	-	-	[16/Nov/2009:01:21:41	+0100]	"GET	/Rezultati%20SI%20Juni.xlsx	HTTP/1.1"	304	-	"-"	"Mozilla/5.0(compatible;Googlebot/2.1;+http://...
213.198.200.126	-	-	[16/Nov/2009:01:54:54	+0100]	"GET	/ispit_rezultati.php	HTTP/1.1"	200	5740	"http://...	"Mozilla/5.0(Windows;U;WindowsNT6.1;en-US;r...
93.86.29.84	-	-	[16/Nov/2009:04:39:23	+0100]	"GET	/img/vts.jpg	HTTP/1.1"	200	4053	"http://...	"Mozilla/5.0(Windows;U;WindowsNT5.1;en-US;r...
77.46.176.92	-	-	[16/Nov/2009:06:51:09	+0100]	"GET	/img/ispit_akt.jpg	HTTP/1.1"	200	2184	"http://...	"Opera/9.21(WindowsNT5.1;U;en)"5.1;U;en)".....
79.101.223.243	-	-	[16/Nov/2009:07:54:18	+0100]	"GET	/ispit.php	HTTP/1.1"	200	3176	"http://...	"Mozilla/4.0(compatible;MSIE6.0;WindowsNT5.1...
81.18.55.70	-	-	[16/Nov/2009:08:11:51	+0100]	"GET	/img/bg.jpg	HTTP/1.1"	200	23084	"http://...	"Mozilla/5.0(Windows;U;WindowsNT5.1;en-US;r...
213.253.117.197	-	-	[16/Nov/2009:08:30:08	+0100]	"GET	/	HTTP/1.1"	200	3669	"http://...	"Mozilla/5.0(Windows;U;WindowsNT5.1;en-GB;r...
94.250.18.18	-	-	[16/Nov/2009:08:34:07	+0100]	"GET	/img/vesti.jpg	HTTP/1.1"	200	1503	"http://...	"Mozilla/4.0(compatible;MSIE7.0;WindowsNT5.1...
81.94.4.174	-	-	[16/Nov/2009:08:38:36	+0100]	"GET	/img/novosti.jpg	HTTP/1.1"	200	1161	"http://...	"Mozilla/4.0(compatible;MSIE8.0;WindowsNT5.1...
89.216.115.68	-	-	[16/Nov/2009:08:45:01	+0100]	"GET	/	HTTP/1.1"	200	3669	"http://...	"Mozilla/4.0(compatible;MSIE6.0;WindowsNT5.1...
109.165.184.122	-	-	[16/Nov/2009:08:49:17	+0100]	"GET	/ispit_rezultati.php	HTTP/1.1"	200	5740	"http://...	"Mozilla/5.0(Windows;U;WindowsNT5.1;en-US)....
79.101.147.10	-	-	[16/Nov/2009:08:51:58	+0100]	"GET	/Aktuelni%20ispitni%20rok.doc	HTTP/1.1"	200	98816	"http://...	"Mozilla/4.0(compatible;MSIE6.0;WindowsNT5.1...
91.207.62.10	-	-	[16/Nov/2009:09:03:30	+0100]	"GET	/img/vts.jpg	HTTP/1.1"	200	4053	"http://...	"Mozilla/5.0(Windows;U;WindowsNT6.1;en-US;r...
89.216.30.146	-	-	[16/Nov/2009:09:06:40	+0100]	"GET	/Aktuelni%20ispitni%20rok.doc	HTTP/1.1"	200	98816	"http://...	"Mozilla/4.0(compatible;MSIE7.0;WindowsNT6.1...

No. of records before preprocessing = 5989

Figure 10- Log file after transferring to sql server data base.

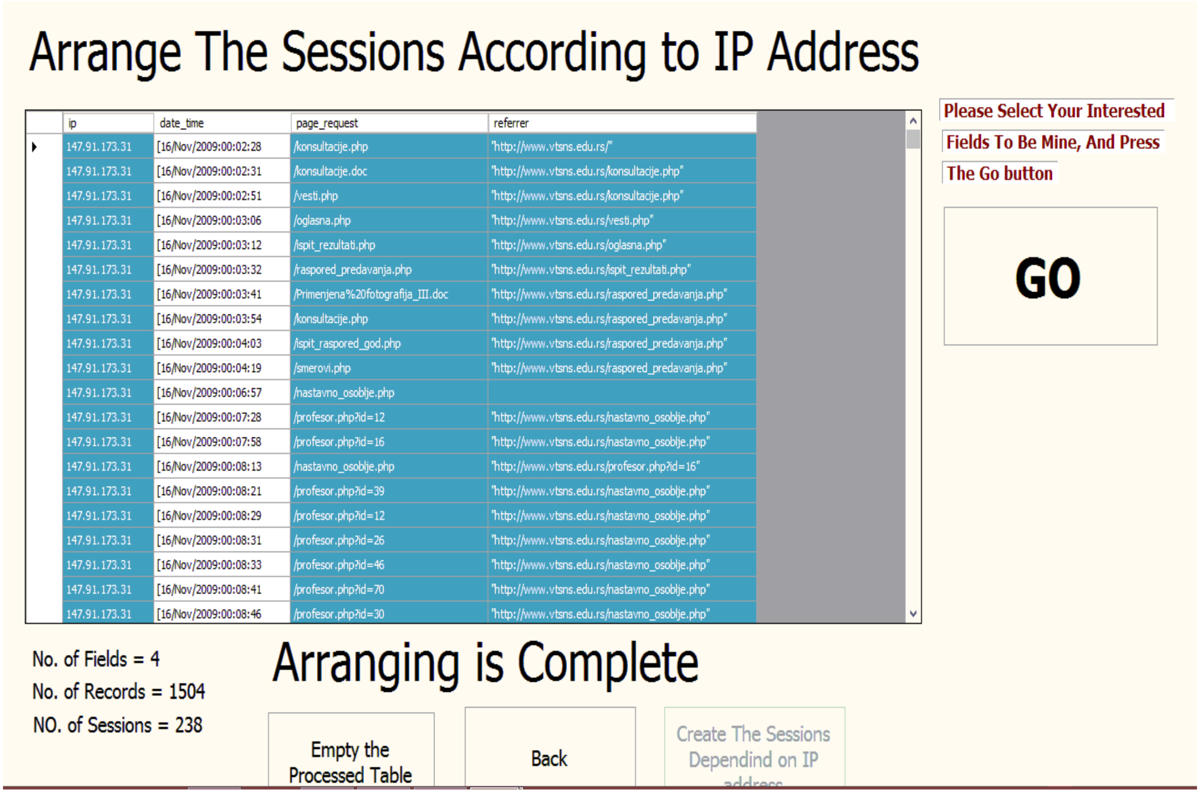
This database contains a lot of tables which be ready to receive huge amounts of data. In C Sharp, the miner can do this connection with the SQL Server database by using ADO.Net (ActiveX Data Objects). The pseudo code shown in Figure-11 will prepare connection with SQL Server database.

```
using System.Data.SqlClient
string connectionString =@"Data Source=baghdad\SQLEXPRESS;
AttachDbFilename=""Path of DB.MDF"";Integrated Security=True; Connect
Timeout=30;User Instance=True";
SqlConnection con = new SqlConnection(connectionString);
con.Open();
```

Figure 11- Pseudo code of connection with SQL server DB.

3- Preprocessing the log file

After transferring the log file to the data base, then we will press on the (*Preprocessing the DataBase*) button in order to achieve the preprocessing step which involves the following four steps (explained earlier), *Field Selection, Data Cleaning, User Identification, Session Identification***.** We notice that the count of records was decreasing from 5989 to 1504 as explain in Figure- 12.



Arrange The Sessions According to IP Address

ip	date_time	page_request	referrer
147.91.173.31	[16/Nov/2009:00:02:28	/konsultacije.php	"http://www.vtsns.edu.rs/"
147.91.173.31	[16/Nov/2009:00:02:31	/konsultacije.doc	"http://www.vtsns.edu.rs/konsultacije.php"
147.91.173.31	[16/Nov/2009:00:02:51	/vesti.php	"http://www.vtsns.edu.rs/konsultacije.php"
147.91.173.31	[16/Nov/2009:00:03:06	/oglasna.php	"http://www.vtsns.edu.rs/vesti.php"
147.91.173.31	[16/Nov/2009:00:03:12	/ispit_rezultati.php	"http://www.vtsns.edu.rs/oglasna.php"
147.91.173.31	[16/Nov/2009:00:03:32	/raspored_predavanja.php	"http://www.vtsns.edu.rs/ispit_rezultati.php"
147.91.173.31	[16/Nov/2009:00:03:41	/Primenjena%20fotografija_III.doc	"http://www.vtsns.edu.rs/raspored_predavanja.php"
147.91.173.31	[16/Nov/2009:00:03:54	/konsultacije.php	"http://www.vtsns.edu.rs/raspored_predavanja.php"
147.91.173.31	[16/Nov/2009:00:04:03	/ispit_raspored_god.php	"http://www.vtsns.edu.rs/raspored_predavanja.php"
147.91.173.31	[16/Nov/2009:00:04:19	/smerovi.php	"http://www.vtsns.edu.rs/raspored_predavanja.php"
147.91.173.31	[16/Nov/2009:00:06:57	/nastavno_osoblje.php	
147.91.173.31	[16/Nov/2009:00:07:28	/profesor.php?id=12	"http://www.vtsns.edu.rs/nastavno_osoblje.php"
147.91.173.31	[16/Nov/2009:00:07:58	/profesor.php?id=16	"http://www.vtsns.edu.rs/nastavno_osoblje.php"
147.91.173.31	[16/Nov/2009:00:08:13	/nastavno_osoblje.php	"http://www.vtsns.edu.rs/profesor.php?id=16"
147.91.173.31	[16/Nov/2009:00:08:21	/profesor.php?id=39	"http://www.vtsns.edu.rs/nastavno_osoblje.php"
147.91.173.31	[16/Nov/2009:00:08:29	/profesor.php?id=12	"http://www.vtsns.edu.rs/nastavno_osoblje.php"
147.91.173.31	[16/Nov/2009:00:08:31	/profesor.php?id=26	"http://www.vtsns.edu.rs/nastavno_osoblje.php"
147.91.173.31	[16/Nov/2009:00:08:33	/profesor.php?id=46	"http://www.vtsns.edu.rs/nastavno_osoblje.php"
147.91.173.31	[16/Nov/2009:00:08:41	/profesor.php?id=70	"http://www.vtsns.edu.rs/nastavno_osoblje.php"
147.91.173.31	[16/Nov/2009:00:08:46	/profesor.php?id=30	"http://www.vtsns.edu.rs/nastavno_osoblje.php"

Please Select Your Interested Fields To Be Mine, And Press The Go button

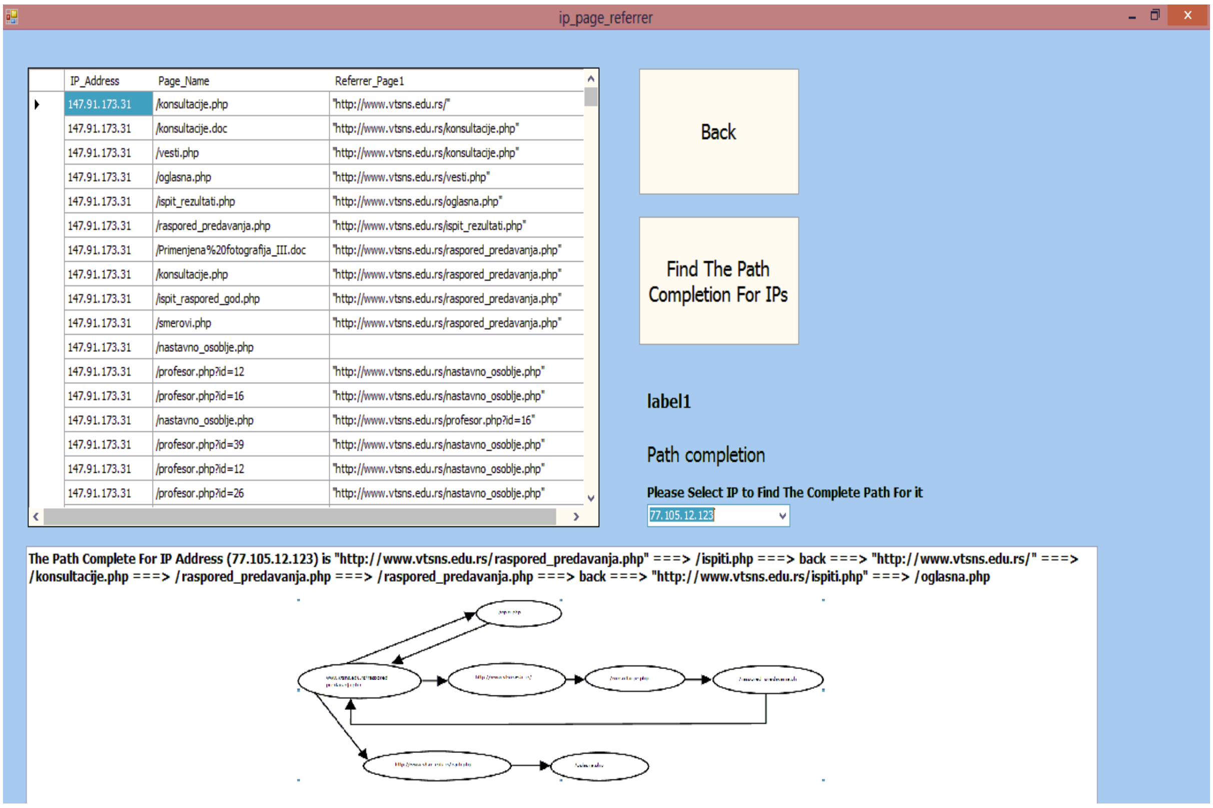
No. of Fields = 4
No. of Records = 1504
NO. of Sessions = 238

Arranging is Complete

GO
Empty the Processed Table
Back
Create The Sessions Dependind on IP address

Figure 12- The miner after preprocessing steps and selects the interested fields.

After that, we must select the interested fields to the mining process. In this paper, we selected the *IP address, page request, and referrer* fields in order to find the complete path for each user (IP address). Then press the (*GO*) button to show the path completion form Figure- 13. In this figure, we select the IP address which have short path because the limited restriction of area.



IP_Address	Page_Name	Referrer_Page1
147.91.173.31	/konsultacije.php	"http://www.vtsns.edu.rs/"
147.91.173.31	/konsultacije.doc	"http://www.vtsns.edu.rs/konsultacije.php"
147.91.173.31	/vesti.php	"http://www.vtsns.edu.rs/konsultacije.php"
147.91.173.31	/oglasna.php	"http://www.vtsns.edu.rs/vesti.php"
147.91.173.31	/ispit_rezultati.php	"http://www.vtsns.edu.rs/oglasna.php"
147.91.173.31	/raspored_predavanja.php	"http://www.vtsns.edu.rs/ispit_rezultati.php"
147.91.173.31	/Primenjena%20fotografija_III.doc	"http://www.vtsns.edu.rs/raspored_predavanja.php"
147.91.173.31	/konsultacije.php	"http://www.vtsns.edu.rs/raspored_predavanja.php"
147.91.173.31	/ispit_raspored_god.php	"http://www.vtsns.edu.rs/raspored_predavanja.php"
147.91.173.31	/smerovi.php	"http://www.vtsns.edu.rs/raspored_predavanja.php"
147.91.173.31	/nastavno_osoblje.php	
147.91.173.31	/profesor.php?id=12	"http://www.vtsns.edu.rs/nastavno_osoblje.php"
147.91.173.31	/profesor.php?id=16	"http://www.vtsns.edu.rs/nastavno_osoblje.php"
147.91.173.31	/nastavno_osoblje.php	"http://www.vtsns.edu.rs/profesor.php?id=16"
147.91.173.31	/profesor.php?id=39	"http://www.vtsns.edu.rs/nastavno_osoblje.php"
147.91.173.31	/profesor.php?id=12	"http://www.vtsns.edu.rs/nastavno_osoblje.php"
147.91.173.31	/profesor.php?id=26	"http://www.vtsns.edu.rs/nastavno_osoblje.php"

Please Select IP to Find The Complete Path For it
77.105.12.123

The Path Complete For IP Address (77.105.12.123) is "http://www.vtsns.edu.rs/raspored_predavanja.php" ===> /ispiti.php ===> back ===> "http://www.vtsns.edu.rs/" ===> /konsultacije.php ===> /raspored_predavanja.php ===> /raspored_predavanja.php ===> back ===> "http://www.vtsns.edu.rs/ispiti.php" ===> /oglasna.php

Figure 13- The path completion form.

Example to clarify the idea

Table 3- sample of log database for one IP Address

IP Address	Requested Page	Referrer Page
147.91.173.31	/konsultacije.php	"http://www.vtsns.edu.rs/"
147.91.173.31	/konsultacije.doc	"http://www.vtsns.edu.rs/konsultacije.php"
147.91.173.31	/vesti.php	"http://www.vtsns.edu.rs/konsultacije.php"
147.91.173.31	/oglasna.php	"http://www.vtsns.edu.rs/"
147.91.173.31	/ispiti.php	"http://www.vtsns.edu.rs/oglasna.php"
147.91.173.31	/raspored_predavanja.php	"http://www.vtsns.edu.rs/"
147.91.173.31	/ispiti.php	"http://www.vtsns.edu.rs/"
147.91.173.31	/ispit_raspored_god.php	"http://www.vtsns.edu.rs/ispiti.php"
147.91.173.31	/ispit_raspored_akt.php	"http://www.vtsns.edu.rs/ispiti.php"

This means, depending on table 3, the user (firstly) visits the home page (http://www.vtsns.edu.rs) and then visits the /konsultacije.php) web page. In the second entry, one can notice that the page in the previous entry (entry 0) which is /konsultacije.php) is equal or belong to the referrer of the current entry (entry 1), consequently the miner concludes that the user visits /konsultacije.doc) web page after /konsultacije.php) web page directly and so on. But in the fourth entry, the miner will notice that the previous page is not equal to the current referrer. Since the current referrer has previously been visited, then the miner concludes that the user presses Back button to that referrer.
The Complete path For IP Address (147.91.173.31) is

"http://www.vtsns.edu.rs/" ===> /konsultacije.php ===> /ispiti.php/konsultacije.doc ===> /vesti.php ===> back ===> "http://www.vtsns.edu.rs/" ===> /oglasna.php ===> /ispiti.php ===> ack ===> "http://www.vtsns.edu.rs/" ===> raspored_predavanja.php ===> back ===> http://www.vtsns.edu.rs/" ===> /ispiti.php ===> /ispit_raspored _god.php ===> / ispit_raspored_akt.php

And the complete graph for this IP address is shown in Figure- 14.

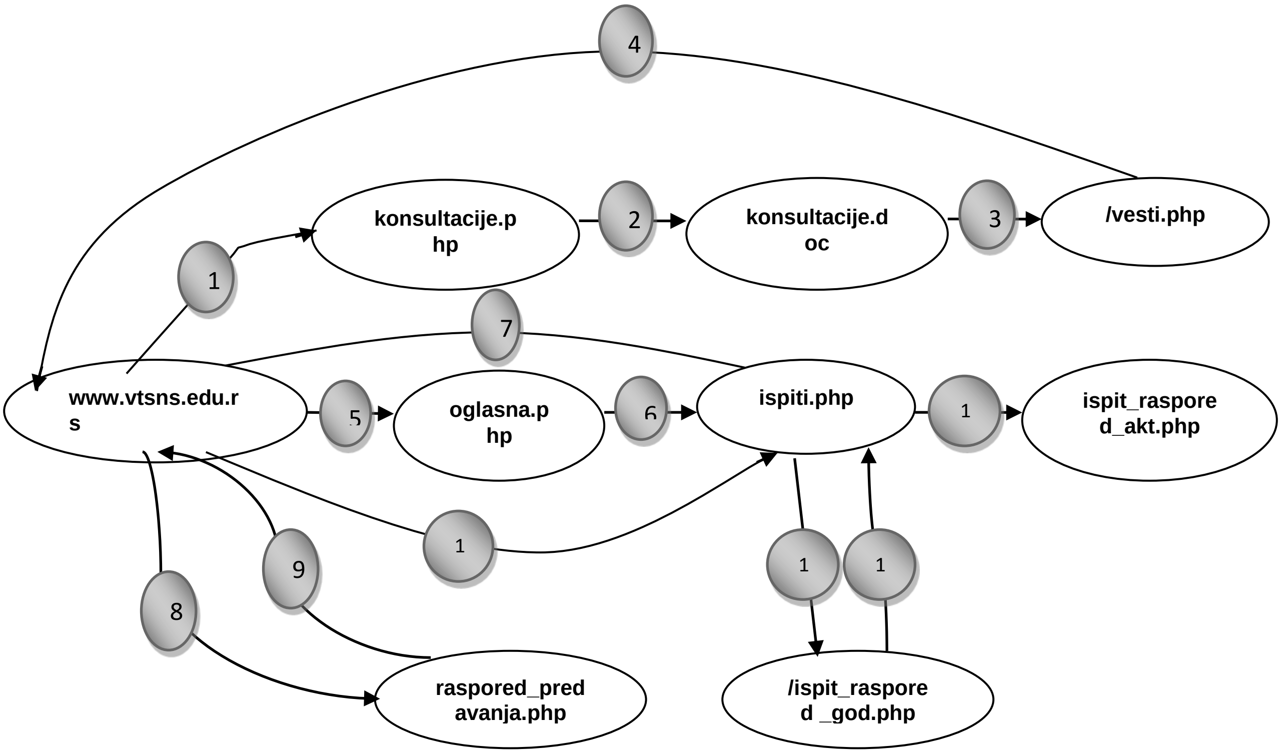


Figure 14- the complete path for IP address 147.91.173.31.

Conclusion and future work

- From the complete path, the miner can conclude a partial structure for this web site. This structure can be represented as a graph where pages represented as nodes and the hyperlinks are the relations connecting the nodes. The analyst can XORed, ANDed, ORed to find many relationships among the users. Also, graph mathematics and theories can be used to re-engineering the sites.
- Web log inspection allows improving navigation. This can manifest itself by organizing important information into the right places, managing links to other pages in the correct sequence, pre-loading frequently used pages. This can help to place the most valuable information (ex. Advertisement objects) on the frequently accessed pages.

- After standing the formats of log file, the following fact concludes that most of the time, users do not visit the home page of a website, they directly navigate to a particular page by getting the URL from search engines. This point is regarded as lacking of data related to the server log file.
- Data cleaning (step in preprocessing phase) on server log file reduced the size of database more than half. In this work, the size of data base was reduced to a quarter (from 5989 to 1504 records) but it consumed considerable amount of execution time.
- As a future work, it is preferable to enrich the server log file with a proxy log file and client log file by involving various data integration tasks to extract more novelty knowledge. Also the future work, another technique for analyzing server log data can be used like clustering, classification, etc. Also, it is possible to perform several data mining algorithms on log files coming from web servers in order to identify user behavior on a particular web site. Also, the future will be concerned with the fields in the log file that are not concerned in the interest of this work in order to obtain useful knowledge. So, according to rule mining, other fields may be important in other mining tasks.

References

1. Anand, S. and Aggarwal, R.R. **2012.** An Efficient Algorithm for Data Cleaning of Log File using File Extensions. *International Journal of Computer Application*, Thapar University, Patiala-147004 (India), **48**(8): 13-18.
2. Bin Ramli, A.A. **2004.** Web Usage Mining For UUM Learning Care Using Association Rules. M.Sc. thesis, University of Utara Malaysia.
3. Kiruthika, M., Jadhav R., Dixit D., Rashmi J., Nehete A., Trupti, Khodkar. **2011.** Pattern Discovery Using Association Rules. *International Journal of Advanced Computer Science and Applications (IJACSA)*, Navi Mumbai, India, **2**(12): 69-74.
4. Pani, K., Panigrahy, L., Sankar, V.H., Ratha, B.K., Mandal, A.K., Padhi, S.K. **2011.** Web Usage Mining: A Survey on Pattern Extraction from Web Logs, India. *International Journal of Instrumentation, Control & Automation (IJICA)*, **1**(1): 10-19.
5. Abd Wahab, M.H., Haji Mohd, M.N., Hanafi, H.F. and Mohsin, M., **2008.** Data Pre-processing on Web Server Logs for Generalized Association Rules Mining Algorithm. *World Academy of Science, Engineering and Technology*, **48**: 190-197.
6. Pani, K., Panigrahy, L., Sankar, V.H, Ratha, B.K., Mandal, A.K, Padhi, S.K. **2011.** Web Usage Mining: A Survey on Pattern Extraction from Web Logs. India, *International Journal of Instrumentation, Control & Automation (IJICA)*, **1**(1): 10-19.
7. Markov, Z. and Larose, T. **2007.** *Data Mining The Web*. WILEY- INTERSCIENCE, Central Connecticut State University New Britain, CT.
8. Sumathi, C.P., Valli, R.P, Santhanam, T. **2011.** An Overview of Preprocessing of Web Log Files For Web Usage Mining. Tamil Nadu state, India, *Journal of Theoretical and Applied Information Technology*. **34**(1): 88-95.
9. Pabarskaite, Z. **2009.** Enhancements of Pre- Processing, Analysis, and Presentation Techniques in Web Log Mining. Ph.D. dissertation, Vilnius Gediminas Technical University.