Survey on web log data in teams of Web Usage Mining

***Mrudang D. Pandya, **Prof. Kiran R Amin**
*(U.V.PATEL COLLAGE OF ENGINEERING,GANPAT UNIVERSITY, Ganpat Vidyanagar,Mehsana-Gozaria HighwayMehsana - 384012, GUJARAT, INDIA)
**(U.V.PATEL COLLAGE OF ENGINEERING, HEAD OF THE DEPARTMENT OF COMPUTER ENGINEERING, Ganpat Vidyanagar,Mehsana-Gozaria HighwayMehsana - 384012, GUJARAT, INDIA)

ABSTRACT – User clicks on Link or gone through any web site at that time particular user's browsing data or browsing actions are captured by server on server log files. In this paper, it shown how browing détails stores in web logs and also mentions different types of fields in web logs. By knowing this kind of field we can easily work with our web usage mining applications. By extraction méthods we can use only those fields which is being used in our applications.

Keywords –Web Usage mining, web log fields ,referrer field.

INTRODUCTION

Many of the Web servers offer the option to store log files or log data in either the common log format or a proprietary format. The common log file format is supported by the maximum of analysis tools but the information about each server transaction is fixed. In lots of cases it is desirable to record more information. Sites sensitive to personal data issues may wish to omit the recording of certain data. In addition ambiguities arise in analyzing the common log file format since field separator characters may in some cases occur within fields. The extended log file format is designed to meet the following needs:

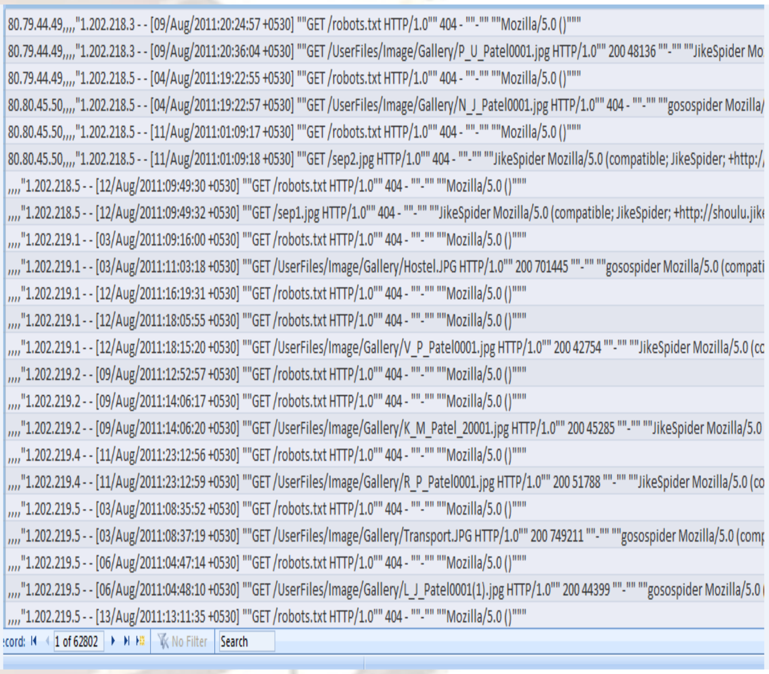
- It should permit control over the data recorded.
- Format support needs of proxies, clients and servers in a common format
- It provide robust handling of character escaping issues
- Allow exchange of demographic data.
- Allow summary data to be expressed.

The log file format described permits customized logfiles to be recorded in a format readable by generic analysis tools. A header specifying the data types recorded is written out at the start of each log.

Basically the data which available for the web usage mining analysis is called as server log files or web logs. Whenever user send any request at that time server gives automatically response and server that particular request response thing stored in server log files in a textual format. This data in the form of single-line transaction record. And data appended to an ASCII text file on the web server.

Now we are going to understand what actually in to the server log files or web logs. We understand all the different fields which stores in the server log files by user single click.

Let's see one example of server log data. And see how actually web logs or server log files look like.



```
80.79.44.49,,,,"1.202.218.3 - - [09/Aug/2011:20:24:57 +0530] ""GET /robots.txt HTTP/1.0"" 404 - ""-"" ""Mozilla/5.0 ()"""
80.79.44.49,,,,"1.202.218.3 - - [09/Aug/2011:20:36:04 +0530] ""GET /UserFiles/Image/Gallery/P_U_Patel0001.jpg HTTP/1.0"" 200 48136 ""-"" ""JikeSpider Mo
80.79.44.49,,,,"1.202.218.5 - - [04/Aug/2011:19:22:55 +0530] ""GET /robots.txt HTTP/1.0"" 404 - ""-"" ""Mozilla/5.0 ()"""
80.80.45.50,,,,"1.202.218.5 - - [04/Aug/2011:19:22:57 +0530] ""GET /UserFiles/Image/Gallery/N_J_Patel0001.jpg HTTP/1.0"" 404 - ""-"" ""gosospider Mozilla/
80.80.45.50,,,,"1.202.218.5 - - [11/Aug/2011:01:09:17 +0530] ""GET /robots.txt HTTP/1.0"" 404 - ""-"" ""Mozilla/5.0 ()"""
80.80.45.50,,,,"1.202.218.5 - - [11/Aug/2011:01:09:18 +0530] ""GET /sep2.jpg HTTP/1.0"" 404 - ""-"" ""JikeSpider Mozilla/5.0 (compatible; JikeSpider; +http:/
,,,,"1.202.218.5 - - [12/Aug/2011:09:49:30 +0530] ""GET /robots.txt HTTP/1.0"" 404 - ""-"" ""Mozilla/5.0 ()"""
,,,,"1.202.218.5 - - [12/Aug/2011:09:49:32 +0530] ""GET /sep1.jpg HTTP/1.0"" 404 - ""-"" ""JikeSpider Mozilla/5.0 (compatible; JikeSpider; +http://shoulu.jike
,,,,"1.202.219.1 - - [03/Aug/2011:09:16:00 +0530] ""GET /robots.txt HTTP/1.0"" 404 - ""-"" ""Mozilla/5.0 ()"""
,,,,"1.202.219.1 - - [03/Aug/2011:11:03:18 +0530] ""GET /UserFiles/Image/Gallery/Hostel.JPG HTTP/1.0"" 200 701445 ""-"" ""gosospider Mozilla/5.0 (compati
,,,,"1.202.219.1 - - [12/Aug/2011:16:19:31 +0530] ""GET /robots.txt HTTP/1.0"" 404 - ""-"" ""Mozilla/5.0 ()"""
,,,,"1.202.219.1 - - [12/Aug/2011:18:05:55 +0530] ""GET /robots.txt HTTP/1.0"" 404 - ""-"" ""Mozilla/5.0 ()"""
,,,,"1.202.219.1 - - [12/Aug/2011:18:15:20 +0530] ""GET /UserFiles/Image/Gallery/V_P_Patel0001.jpg HTTP/1.0"" 200 42754 ""-"" ""JikeSpider Mozilla/5.0 (cc
,,,,"1.202.219.2 - - [09/Aug/2011:12:52:57 +0530] ""GET /robots.txt HTTP/1.0"" 404 - ""-"" ""Mozilla/5.0 ()"""
,,,,"1.202.219.2 - - [09/Aug/2011:14:06:17 +0530] ""GET /robots.txt HTTP/1.0"" 404 - ""-"" ""Mozilla/5.0 ()"""
,,,,"1.202.219.2 - - [09/Aug/2011:14:06:20 +0530] ""GET /UserFiles/Image/Gallery/K_M_Patel_20001.jpg HTTP/1.0"" 200 45285 ""-"" ""JikeSpider Mozilla/5.0
,,,,"1.202.219.4 - - [11/Aug/2011:23:12:56 +0530] ""GET /robots.txt HTTP/1.0"" 404 - ""-"" ""Mozilla/5.0 ()"""
,,,,"1.202.219.4 - - [11/Aug/2011:23:12:59 +0530] ""GET /UserFiles/Image/Gallery/R_P_Patel0001.jpg HTTP/1.0"" 200 51788 ""-"" ""JikeSpider Mozilla/5.0 (co
,,,,"1.202.219.5 - - [03/Aug/2011:08:35:52 +0530] ""GET /robots.txt HTTP/1.0"" 404 - ""-"" ""Mozilla/5.0 ()"""
,,,,"1.202.219.5 - - [03/Aug/2011:08:37:19 +0530] ""GET /UserFiles/Image/Gallery/Transport.JPG HTTP/1.0"" 200 749211 ""-"" ""gosospider Mozilla/5.0 (comp
,,,,"1.202.219.5 - - [06/Aug/2011:04:47:14 +0530] ""GET /robots.txt HTTP/1.0"" 404 - ""-"" ""Mozilla/5.0 ()"""
,,,,"1.202.219.5 - - [06/Aug/2011:04:48:10 +0530] ""GET /UserFiles/Image/Gallery/L_J_Patel0001(1).jpg HTTP/1.0"" 200 44399 ""-"" ""gosospider Mozilla/5.0 (
,,,,"1.202.219.5 - - [13/Aug/2011:13:11:35 +0530] ""GET /robots.txt HTTP/1.0"" 404 - ""-"" ""Mozilla/5.0 ()"""
```

Figure 1. Simple Web Log Data

So now we focus on the each and every different fields of server log data or log files.

There are many fields in server log files or web logs. They are Remote host field, Date/Time field, HTTP Request field, Status code field, Transfer volume(Bytes) field, etc.. Now we are going to understand each and every fields in detail to understand what actually that particular field stores in it.

There are mainly three types of log formats available. They are:

1. Common Log Format (CLF)
2. Extension of Common Log Format (ECLF)
3. Microsoft IIS Log Format

By understanding this kind of field we can use particular field as per our web mining application required. As example we are focusing on user web browsing data at that time we do not require Identification field and status code field. At that time we can extract that particular fields and apply that extracted data on experimental tool. So we can manage that particular data easily and also reduce complexity. And we can also minimize the calculation time.In this way we must have to understand about all this server log data elements.

I. COMMON LOG FORMAT:

Web log data come in various formats, which changes depending on the configuration of the web server. It also called as CLF(Common Log Format). And it include the following seven fields:

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

A typical configuration for the access log might look as follows.

```
LogFormat "%h %l %u %t \"%r\" %>s %b"common

CustomLog logs/access_log common
```

2.1. Remote host field:

This field consists of the Internet IP address of the remote host making the request, such as“1.202.219.4”. if there is DNs name is available at that time we can get also the DNS name instead of Internet IP. To obtain the domain name of the remote host rather than the IP address ,the server must submit a request, using the Internet domain name system (DNS) to resolve (i.e., translate) the IP address into a host name. Since humans prefer to work with domain names and computers are most efficient with IP addresses, the DNS system provides an important interface between humans and computers.DNS example is google.com,yahoo.com, etc..

2.2. Identification field:

This field mainly used to store identity information provided by the client only if the web server is performing an identity check . However, this field is seldom used because the identification information is provided in plain text rather than in a securely encrypted form. Therefore, this field usually contains a hyphen, indicating a null value.[1]

2.3. Authuser field:

This particular field mainly used to store the authenticated client user name, if it is required. The authuser field was designed to contain the authenticated user name information that client needs to provide to gain access to directories that are password protected. If no such information is provided, the field defaults to a hyphen.[1]

2.4. Date/time field:

Some web log uses the following specialized date/time field format:“[DD:HH:MM:SS],” where DD represents the day of the month and HH:MM:SS represents the 24-hour time, given in EDT. In this above particular data set, the DD portion represents the day in August, 2011 that the web log entry was made. However, it is more common for the date/time field to follow the following format: “DD/Mon/YYYY:HH:MM:SS offset,” where the offset is a positive or negative constant indicating in hours how far ahead of or behind the local server is from Greenwich Mean Tim (GMT). For example, a date/time field of “09/august/2011:03:27:00 -0500” indicates that a request was made to a server at 3:27 a.m. on August 9, 2011, and the server is 5 hours behind GMT.

2.5. HTTP request:

This particular HTTP request field consists of all the information that the client’s browser has requested from the web server for any particular task. This whole HTTP request field is contained within quotation marks. This field may be partitioned into four areas:

- The request method,
- The uniform resource identifier (URI),
- The header, and
- The protocol.

2.5.1. The Request Method:

The request method can be GET, PUT, HEAD, POST etc.. The most common request method is GET. This request represents a request to retrieve data that are identified by the URI.

2.5.2. The uniform resource identifier (URI):

URI(Uniform resource identifies) contain the page or document name and the directory path requested by the client browser. By the use of URI we can know or analyze visitor requests for pages and files using web usage miner. Example of URI: http://www.w3.org/Icons/www/w3c_main.gif.

2.5.3. The header:

HTTP request field also contains the header section. This field contains optional information concerning the browser’s request. For example this information can be used by the web usage miner to determine which keywords are being used by visitors I search engines that point to your site.

2.5.4. The protocol:

The HTTP request field also includes the protocol section, which indicates which version of the Hypertext Transfer Protocol (HTTP) is being used by the client’s browser. Then, based on the relative frequency of newer protocol versions (e.g., HTTP/1.1), the web developer

2.6. Status code field:

This field is related to the user requests in teams of successful or not successful. Not all requests are successful. The status code field provides a three-digit response from the web server to client’s browser. In this there are some types are available. In this if the request was successful or not successful ,if there was any error occurs, it also contain which type of error occurred. In this code of the form “2xx” indicate a success, and code of the form “4xx” indicates an error. A sample of the possible status codes that a web server could send described below:

If there is Successful transmission (200 series):
It Indicates that the request from the client wareceived, understood, and completed. There are some extensions available they are:

- 200: success
- 201: created
- 202: accepted
- 204: no content

If there is Redirection (300 series):
It indicates that further action is required to complete the client’s request. There are some extensions available they are:

- 301: moved permanently
- 302: moved temporarily
- 303: not modified
- 304: use cached document

If there is Client error (400 series):
it indicates that the client’s request cannot be fulfilled, due to incorrect syntax or a missing file. There are some extensions available they are:

- 400: bad request
- 401: unauthorized
- 403: forbidden
- 404: not found

If there is Server error (500 series)
It Indicates that the web server failed to fulfill what was apparently a valid request. There are some extensions available they are:

- 500: internal server error
- 501: not implemented
- 502: bad gateway
- 503: service unavailable

2.7. Transfer volume field:

This field indicates as the name suggest size of the file in bytes, sent by the web server to the client’s browser.

```
141.243.1.172 [29:23:53:25] “GET /Software.html HTTP/1.0” 200 1497
```

In above web log example 1497 is indicated transfer volume in bytes. This field generated only when GET requests that have been completed successfully (Status = 200) will have a positive value in the transfer volume field. If status is not equals to 200 at that time the field will consist of a hyphen or a value of zero. Status code field is useful for helping to monitor the network traffic, the load carried by the network throughout the 24-hour cycle.

III. Extension Of Common Log Format (ECLF) :

The extension of CLF is variation of the CLF. It is formed by adding two additional fields onto the end of record, that particular field are the referrer field and the user agent field. Both the common log format was created by NCSA (National Center for Supercomputing Applications).

3.1. Referrer Field

The referral log contains a corresponding entry for each entry in the common log.

The fields in the Referral log are:

Date: Time Referrer
The following is an example of a record in a referral log:

```
[10/Oct/2011:21:15:05 +0500]
"http://www.gmail.com/index.html"
```

The following is a description of the fields in the Referral log:

Date: Time Timezone
([10/Oct/2011:21:15:05 +0500] in the example)
The date and time stamp of HTTP request. The date and time of an entry logged in the referral log corresponds to the resource access entry in the common log. As a result, the date and time of corresponding records from each of these logs will be the same. The syntax of the date stamp is identical to the date stamp in the common log. -referrer ("http://www.gmail.com/index.html" in the example)

The referrer is the URL of the HTTP resource that referred the user to the resource requested. For example, if a user is browsing a Web page such as http://www.gmail.com/index.html and the user clicks on a link to a secondary page, then the initial page has referred

Prefix	Meaning
s-	Server actions.
c-	Client actions.
Cs-	Client-to-server actions.
Sc-	Server-to-client actions.

the user to the secondary page. The entry in the referral log for the secondary page will list the URL of the first page(http://www.gmail.com/index.html) as its referral.

The referrer field describe URL of the previous site visited by the client, which jointed to the current page. For images, the referrer is the web page on which the image is to be displayed. The referrer field contains important information for marketing purposes, since it can track how people found your site.[1]

3.2. User agent field

The user agent field provides information about the client's browser, the browser version, and the client's operating system. Importantly, this field can also contain information regarding bots, such as web crawlers. Web developers can use this information to block certain sections of the Web site from these web crawlers, in the interests of preserving bandwidth. Further, this field allows the web usage miner to determine whether a human or a bot has accessed the site, and thereby to omit the bot's visit from analysis, on the assumption that the developers are interested in the behavior of human visitors[1].

Let's take one excellent example to understand ECLF:

Consider the following example of an extended common log format (ECLF). For privacy purposes, the URL has been partly masked.

```
139.1xx.120.126 -- smithj [28/OCT/2011:20:27:32
-5000] ``GET /Default.htm HTTP/1.1" 200
1270 ``http:/www.w3schools.com/"
``Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.0
)"
```

- Remote host: 149.1xx.120.116
- Identification: –
- Authuser: smithj
- Date/time: [28/OCT/2004:20:27:32 -5000]
- Request: "GET /Default.htm HTTP/1.1"
- Status code: 200
- Transfer volume: 1270
- Referrer:"http:/www.dataminingconsultant.com/"
- Useragent:"Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.0)"

IV. MICROSOFT IIS LOG FORMAT:

This is the third log format. In this, there are more field included then the common log format and extended common log format. For example, the elapsed processing time is included, along with the bytes sent by the client to the server; also, the time recorded is local time. Note that web server administrators need not choose any of these formats; they are free to specify which fields they believe are most appropriate for their purposes.

Table 4.1

In table 4.1 and 4.2 mentioned all possible fields that are included in the Microsoft IIS Log Format. The field name its appears as and its description is shown in tabular format.

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.
Elapsed time	-	the elapsed processing time is included, along with the bytes sent by the client to the server
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.

Table 4.2

EXAMPLE:

1998-11-19 22:48:39 206.175.82.5 - 208.201.133.173 GET /global/images/navlineboards.gif - 200 540 324 157 HTTP/1.0 Mozilla/4.0+(compatible;+MSIE+4.01;+Windows+95) USERID=CustomerA;+IMPID=01234 http://www.loganalyzer.net

V. CONCLUSION

Web usage mining techniques apply on web log data. Whatever the data on web logs we have to know each and every field of web logs. By this paper you can know what is in web logs and you can apply only those fields in your web mining techniques. You can also improve your mining results. There are mainly three types of web logs format is mentioned in this paper. So you can use or study any of the format as per your application needs.

REFERENCES

Books

[1] Data Mining the Web: *Uncovering Patterns in Web Content, Structure, and Usage*. John Wiley & Sons 2007.

Proceedings Papers:

[2] By Zdravko Markov and Daniel T. LarosePeter Chapman, Julian Clinton, Randy Kerber, Thomas Khabaza, Thomas Reinart,Colin Shearer, and Rudiger Wirth, *CRISP–DM Step-by-Step Data Mining Guide*, 2000,http://www.crisp-dm.org/.

[3] Daniel Larose, *Discovering Knowledge in Data:An Introduction to Data Mining*, Wiley,Hoboken, NJ, 2005.

[4] B. Mobasher, N. Jain, E. Han, J. Srivastava,"*Web mining:Pattern discovery from world wide web transactions*",Technical Report TR 96-050, University of Minnesota, Dept.of Computer Science, Minneapolis, 1996.

[5] R. Cooley, B. Mobasher, J. Srivastava,"*Grouping web page references into transactions for mining world wide web browsing patterns*", Technical Report TR 97-021, University of Minnesota, Dept. of Computer Science, Minneapolis,1997.

[6] Arshi Shamsi, Rahul Nayak, Pankaj Pratap Singh, Mahesh Kumar Tiwari: *Web Usage Mining by Data Preprocessing*.IIMT Engineering College, Meerut, UP, India