ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Review the Steps of Server Log Data Processing for Web Usage Mining

R.Rooba, Dr.V.Valli Mayil

Assistant Professor, Dept. of Computer Technology and Information Technology, Kongu Arts and Science College, Nanjanapuram, Erode, Tamilnadu, India

Associate Professor and Head, Dept. of Computer Science and Applications, Periyar Maniammai University, Thanjavur, Tamilnadu, India

ABSTRACT: World Wide Web has been growing as a dominant platform for retrieving information and discovering knowledge from web data. Users' accesses are recorded in web logs. Because of the tremendous usage of web, the web log files are growing at a faster rate and the size is becoming huge. Web mining is the application of data mining techniques in web data. Web Usage Mining applies mining techniques in web log data to extract the behavior of users. Web usage mining consists of three phases preprocessing, pattern discovery and pattern analysis. Web log data is usually noisy and ambiguous and preprocessing is an important process before mining. For discovering patterns sessions are to be constructed efficiently. This paper reviews existing work done in the preprocessing stage.

KEYWORDS: Web Server Log, Data Cleaning, Preprocessing, User Identification, Session identification, Path Completion, Transaction Identification.

I. INTRODUCTION

World Wide Web is a huge, interconnected, semi-structured, widely distributed, highly heterogeneous and hypertext information repository. Web mining technologies are the proper solutions for knowledge discovery on the Web. Web mining is the application of data mining techniques to discover patterns from the Web. Web mining can be classified into three different types, which are Web content mining, Web structure mining and Web usage mining. Web content mining is the process of extracting and integration of useful data, information and knowledge from Web page contents. Web structure mining is the process of using graph theory to analyze the node and connection structure of a web site. Web Usage Mining is a part of web mining which deals with the extraction of interesting knowledge from log files produced by web server [1].

Web Usage Mining consists of three major steps: Preprocessing, Pattern Discovery and Pattern Analysis. Preprocessing remove noisy, incomplete and inconsistent data from the web server log file. Pattern Discovery process extracting information from preprocessed data. Several techniques such as data mining, machine learning are used to extract patterns. Pattern Analysis is the final stage of Web Usage Mining. This step extracting interesting patterns from the discovered patterns.

This paper mainly concentrated on the survey of Web Log Data Preprocessing methods described by different researchers. This paper is organized as follows. Section II provides sources, types and formats of Web log file. Section III explainssteps in preprocessing methods and the techniques followed by various researchers. Section IV compares preprocessing methods followed by various researchers and Section V lists the future work Section VI concludes this paper.

II. WEB LOG FILE

Web log files are files that contain information about website visitor activity. Log files are created by web servers automatically. Each time a visitor requests any file (page, image, etc.) from the site information of request is appended to a current log file. Most log files have text format and each log entry is saved as a line of text.

A. Location of web log file:

Web log file is located in three different locations.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

- Web server logs: Web log files provide most accurate and complete usage of data to web server. The log file do not record cached pages visited. Data of log files are sensitive, personal information so web server keeps them closed.
- Web proxy server: Web proxy server takes HTTP request from user, gives them to web server, then result passed to web server and return to user. Client send request to web server via proxy server.
  The two disadvantages are: Proxy-server construction is a difficult task. Advanced network programming, such as TCP/IP, is required for this construction. The request interception is limited.
- Client browser: Log file can reside in client's browser window itself. HTTP cookies used for client browser. These HTTP cookies are pieces of information generated by a web server and stored in user's computer, ready for future access.

B. Types of web log file

There are four types of server logs.

- Access log file: Data of all incoming request and information about client of server. Access log records all requests that are processed by server.
- Error log file: list of internal error. Whenever an error is occurred, the page is being requested by client to web server the entry is made in error log.Access and error logs are mostly used, but agent and referrer log may or may not enable at server.
- Agent log file: Information about user's browser, browser version.
- Referrer log file: This file provides information about link and redirects visitor to site.

C. Web log file format

Web server log files are the primary data sources used in web usage mining. These are plain text (ASCII) files. Web server log files store click stream data which can be useful for mining purposes. The data is stored as a result of user's interaction with a website. Web Server log files are records of web server activities. They give information about users file requests to a web server and the server response to those requests.

There are three kinds of log file formats to record log files.

i). Common Log Format (CLF)

This is the most common and standardized text format of a web server log file. This can be produced by several web servers and read by variety of log analysis programs.

Sample CLF Log format:

127.0.0.1 - john [12/nov/2011:12:53:46-0700] "GET/apache_pb.gif HTTP/1.0" ǁ 200 2326

Each part of this log entry is described below :

127.0.0.1 : IP address of the client (remote host) which made the request to the server.
- :hyphen in the output indicates that the requested piece of information is not available.
john : user id of the person requesting the document as determined by HTTP authentication.
[12/nov/2011:12:53:46-0700] :The time which the server finishes processing the request.
"GET/apache_pb.gif HTTP/1.0" : The request line from the client is given in double quotes.
GET :the method used by the client.
/apache_pb.gif : the resource requested by the client.
HTTP/1.0 : The protocol used by the client.
200 : status code that the server sends back to the client.
2326 : size of the object returned to the client, not including the response headers.

i). W3C (World Wide Web Consortium) Extended log file format (EClF):

This is the default log file format used by IIS. It uses ASCII text format and the time recorded as UTC (Greenwich Mean Time).

Sample ECLF Log format:

217.13.12.211 - - [11/Nov/2011:05:45:26 -0500] "GET /meta_tags.htm HTTP/1.0" 200 26150 "http://www.google.com/search?q=meta+and+tag" [Mozilla/11.0(compatible; MSIE 8.0: Windows XP; DigExt)]

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Two new fields Referrer URL and User Agent field are added in the ECLF format.
"http://www.google.com/search?q=meta+and+tag " **:** Referrer URL Field

[Mozilla/11.0(compatible; MSIE 8.0: Windows XP; DigExt)] : Agent field it describes Borwser and operating system versions

ii). Microsoft IIS (Internet Information Services) log file format

Microsoft IIS log file format is a non-customizable ASCII format used to record more information than the NCSA Common format but less than the W3C format. It uses comma to separate fields and uses the local time. It includes the user's IP address, user name, request date and time, Service status code and number of bytes received, the elapsed time, the number of bytes sent, the action and the target file.

SampleIIS log file format:
192.168.114.201,--,11/25/2011,9:45:25, wsvc2,4504,3223,11/25/2011,23:58:11,MSFTPSVC,GET,Salesdeptlogo.gif

iii) NCSA (National Centre for Supercomputing Applications) Log file format

This is a fixed (non-customizable) ASCII format. It does not support FTP sites. Since the entries are small with this format, the storage space required for logging is less compared to other formats. It logs the basic information about user requests such as remote host name, user name, date, time, request type, HTTP status code, and the number of bytes sent by the server. It records the time byusing the local time and fields are separated by spaces.

Sample NCSA log file format:
172.21.13.45-REDMOND/sam[11/11/2011:25:28:06-0800]"GETscripts/iisadmin/sm.dll?http/serv HTTP/1.0" 200 3401

III. DATA PREPROCESSING

Web log data preprocessing is a complex process. The aim of data preprocessing is to select essential features by removing irrelevant records and finally transform raw data into sessions. It consists of following steps:

i) Data Cleaning
ii) User Identification
iii) Session Identification
iv) Path Completion
v) Transaction Identification

Various research works are carried out in this preprocessing area, for grouping sessions and transactions, which is used to discover user behavior patterns.

A. Data Cleaning

The process of data cleaning is removal of irrelevant data. Analyzing the huge amounts of records in server logs is a cumbersome activity. So initial cleaning is necessary. If a user requests a specific page from server entries like gif, JPEG, etc., are also downloaded which are not useful for further analysis are eliminated. The records with failed status code are also eliminated from logs. Automated programs like web robots, spiders and crawlers are also to be removed from log files. Thus various researcher's [2],[3],[4],[5],[8],[12],[14],[15] preprocessing includes the following steps

i). If the status code of each and every record is less than 200 and greater than 299 then those records are removed.
ii). The URL field is checked for its extension filename. If the filename has gif, jpg, JPEG, CSS, etc., they are removed.
iii). The records which request robots.txt are removed and if the time taken is very small like less than 2 seconds are considered as automated programs traversal and they are also removed. [5].

Ravindra Gupta and Prateek Gupta [6] have introduced application specific data cleaning method. They introduced new step before data cleaning called Customization. The Customization algorithm takes user choice for normal, multimedia, graphics or e-commerce applications. This method read raw web log file and remove logs according to user selection to make intermediate file, which consists only user application specific log data.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

Priyanka Patil and Ujwala Patil [13] introduced field extraction algorithm prior to the data cleaning process. This algorithm separating the fields from the single line of the log file.

B. User Identification

The next important and complex step is unique user identification. The complexity is due to the local cache and proxy servers. To overcome this cookies are used. But users may disable cookies.[7]Another solution is to collect registration data from users. But users neglect to give their information due to privacy concerns. So majority of records does not contain any information in the user-id and authentication fields. The fields which are used by the various researcher [2], [3], [4],[5],[12],[13]to find unique users and sessions are

i) IP address
ii) User agent
iii) Referrer URL

Users are identified by using these fields are as follows.

- If two records has different IP address they are distinguished as two different users else if both IP address are same then User agent field is checked.
- If the browser and operating system information in user agent field is different in two records then they are identified as different users.

Vijayashri Losarwar and Dr.Madhuri Joshi [8] has described in their work the unique user is the combination of IP address, Agent and Operating System. They have created Log in Table. Attributes in this table are Date, User ID, Client IP, In-time and Out-time. A unique numbers is given to all pages visited by users. From that table they identified the user.

T. Revathi , M. Mohana Rao, Ch. S. Sasanka [14] used referrer URL and site topology to identify the user. If the IP address of a user is same as previous entry and user agent is different then the user is assumed as a new user. If both IP address and user agent are same then referrer URL and site topology is checked. If the requested page is not directly reachable from any of the pages visited by the use0 , then the user is identified as a new user in the same address.

B. Session Identification

After users are identified the next step is identification of sessions. A session is a sequence of activities made by one user during one visit to the site. The goal of session identification is to divide the page accesses of each user into individual sessions.

After identifying users if both IP address and Browser OS are same the Referrer URL field is checked. If URL in the referrer URL field in current record is not accessed previously or if Referrer URL field is empty then it is considered as a new user session. Reconstruction of accurate user sessions from server access logs is a challenge task since the access log protocol[HTTP protocol] is stateless and connectionless [9]. There are three heuristics available to identify sessions from users. Two are based on time and one based on the navigation of users through the web pages.

Time Oriented Heuristics: The simplest methods are time oriented in which one method based on total session time and other based on 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 [10] to 24 hours [11] while default time is 30 minutes by R.Cooley [7]. 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.

Time based methods are not reliable because users may involve in some other activities after opening the web page and factors such as busy communication line, loading time of components in web page and contents in web page are not considered.

Navigation Oriented Heuristics: This method uses web topology in graph format. It considers webpage connectivity, however it is not necessary to have hyperlink between consecutive page requests.

Both methods are used by different researcher for effective reconstruction of sessions.

The researchers [2],[3],[13] combined Referrer URL based method and Time oriented heuristics to identify the sessions. A User Session Set is obtained from the Web Access Log Set by following rules such as different users are distinguished by different IP address. If the IP addresses are same, the different browsers or operating systems indicate different users and if the IP addresses are same, the different browsers and operationsystems are same, the

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

referrer information is taken into account. The Referrer URL field is checked and a new user session is identified if the URL in the Referrer URL field has never been accessed before, or there is a large interval between the access time of this record and the previous one if the Referrer URL field is empty. If the sessions identified by the previous step contain more than one visit by the same user at different time, the time-oriented heuristics is then used to divide the different visits into different user sessions.

The researchers [4],[5],[8],[12],[14],[15] have presented the session identification process of all the pages accessed by a user using threshold time.. It is assumed that the user has started a new session if the time between two pages requests exceeds the given time limit. They used 30 minutes as default threshold timeout to identified the session.

D. Path Completion

There are chances of missing pages after constructing transactions due to proxy servers and caching problems. So missing pages are added as follows: The page request is checked whether it is directly linked to the last page or not. If there is no link with last page check the recent history. If the log record is available in recent history then it is clear that "back" button is used for caching until the page has been reached. If the referrer log is not clear, the site topology can be used for the same effect. If many pages are linked to the requested page, the closest page is the source of new request and so that page is added to the session. There are three approaches in this regard.

Reference Length approach: This approach is based on the assumption that the amount of time a user spends on a page correlates to whether the page is an auxiliary page or content page for that user. It is expected that the time spent on auxiliary page is small and content page is more. A reference length can be calculated that estimates the cut off between auxiliary and content references. The length of each reference is estimated by taking the difference between the time of the next reference and the current reference. But the last reference has no next reference. So this approach assumes the last one is always an auxiliary reference.

Maximal Forward Reference:A transaction is considered as the set of pages from the visited page until there is a backward reference. Forward reference pages are considered as content pages and the path is taken as index pages. A new transaction is considered when a backward reference is made.

Time Window:A time window transaction is framed from triplets of ipaddress, user identification, and time length of each webpage up to a limit called time window. If time window is large, each transaction will contain all the page references for each user. Time window method is also used as a merge approach in conjunction with one of the previous methods.All the above three methods are used for Path completion as well as Transaction Identification.

The researchers [2],[3],[4] have described their path completion technique using the Reference length method and Referrer URL field.The researcher [15] have presented their path completion technique using Referrer URL and Site Topology technique.

E. Transaction Identification

The goal of transactions identification is to create meaningful clusters of references for each user. Transaction identification is done by merges or divides approaches. To find out the user's travel pattern and user's interests, two kinds of transactions are defined. i.e., travel path transactions and content only transactions. The travel path is a combination of auxiliary and content pages accessed by a user. The content only transactions are only content pages which are used in mining to discover user's interest and cluster users visiting the same web site.

The researchers [2],[3],[4],[14],[15] have presented their transaction identification process using the combination Reference Length method,

Maximal Forward Reference method and Time window method to achieve effective transaction. Maximal forward reference method is used to identify Travel Path set for each user id.

Reference Length method is used to distinguish the content page and the auxiliary page.

Time window method is used to identify whether the last page is the content page or auxillary page. Generally the last page is content page , but some time it may the auxillary page. To avoid this confusion Time Window method is used. The time difference between the first page and last page access is calculated. If the difference is less than the cut off time it is considered as a auxillary page or as content page.

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

IV. COMPARISON OF PREPROCESSING METHODS

Author	Cleaning	User Identfication	Session Identificati on	Path Completion	Transaction Identification
V.Chitraa, Dr.Antony selvadoss Davamani	Removed status code < 200 and >299, .jpg,.gif, robots.txt	IP Address, User Agent, Referrer URL	Referrer URL & Default time 30 Minutes	Reference Length & Referrer URL	Reference Length, Maximal Forward reference, Time window
J.Velliangiri and S.Chenthur Pandian	Removed status code < 200 and >299, .jpg,.gif, robots.txt	IP Address, User Agent, Referrer URL	Referrer URL & Default time30 Minutes	Reference Length & Referrer URL	Reference Length, Maximal Forward reference, Time window
C.P..SumathiR,Padma javalli and T.Santhanam	Removed status code <200 and >299, .jpg,.gif,etc robots.txt	IP Address, User Agent, Referrer URL	Default Time 30 Minutes	Referrer URL only	Reference Length, Maximal Forward reference, Time window
G.T. Raju and P.S.Satyanarayana	Removed status code < 200 and >299, .jpg,.gif,etc robots.txt	IP Address, User Agent, Referrer URL	Default Time 30 Minutes	Relational Database table containing Session and visits are created	
Ravindra Gupta and Pradeep Gupta	Customization step introduced	-	-	-	-
Vijayashri Losarwar and Dr.Madhuri Joshi	Removed status code <200 and >299, .jpg,.gif,etc robots.txt	Login Detail Table Created	Default Time 30 Minutes	-	-
Thanakorn Pamutha, Siriporn Chimphlee,	Removed status code < 200 and >299, .jpg,.gif,etc robots.txt	IP Address	Default Time 30 Minutes	-	-
Priyanka Patil and Ujwala Patil	Field Extaction Introduced before Cleaning	IP Address, User Agent, Referrer URL	Referrer URL & Default time 30 Minutes	-	-
T.Revathi M.Mohana, ChS.Sasanks	Removed status code < 200 and >299, .jpg,.gif,etc robots.txt	IP Address, User Agent, Referrer URL & Site Topology	Default Time 30 Minutes	Data integration and transformation is performed using data mining techniques	
Naga Lakshmi, Raja Sekhara Rao	Removed status code <200 and >299, .jpg,.gif,etc robots.txt	IP Address, User Agent, Referrer URL	Default Time 30 Minutes	Referrer URL and Site Topology	Merge and Divide Approach

VI. CONCLUSION

Web sites are one of the most important tools for advertisements in international area for universities and other foundation. The quality of a website can be evaluated by analyzing user accesses of the website. To know the quality of a web site user accesses are to be evaluated by web usage mining. The results of mining can be used to improve the website design and increase satisfaction which helps in various applications. Log files are the best source to know user behavior. But the raw log files contains unnecessary details like image access, failed entries etc., which will affect the accuracy of pattern discovery and analysis. So preprocessing stage is an important work in mining to make efficient pattern analysis. To get accurate mining results user's session details are to be known. The survey was performed on a selection of web usage methodologies in preprocessing proposed by research community. More

ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 11, November 2015

concentration is done on preprocessing stages like session identification and path completion and we have presented various works done by different researchers. Our research in future is to create more efficient session reconstructions through Semantic analysis and gives more accurate patterns for analysis of users.

VII. FUTURE PLAN

There are a number of issues in preprocessing of log data. The methods discussed above were explained by various researchers. Log includes entries of document traversal, file retrieval and unsuccessful web events among many others that are organized according to the date and time. It is important to eliminate the irrelevant data. So cleaning is done to speed up analysis as it reduces the number of records and increases the quality of the results in the analysis stage. More research can be done in preprocessing stages to clean raw log files, and to identify users and to construct accurate sessions. In future we have planned to preprocess the web log file semantically which will clean the log effectively and produce necessary information for pattern analysis process.

REFERENCES

[1] Robert.Cooley,Bamshed Mobasher and Jaideep Srinivastava,"Data Preparation for Mining World Wide Web Browsing Patterns", Journal of knowledge and Information Systems,1999.
[2] V.Chitraa and Dr. Antony Selvadoss Davamani " An Efficient Path Completion Technique for Web log mining", IEEE International Conference on Computing Intelligence and Computing Research, 2010.
[3] J.Velliangiri and S.Chenthur Pandian " A Novel Technique for Web log Mining with Better Data Cleaning and Transaction identification", Journal of Computer Science, 2011.
[4] C.P..Sumathi, R,Padmajavalli and T.Santhanam, " An Overview of Preprocessing of Web Log Files For Web usage Mining", Journal of Theoretical and Applied Information Technology, 2011.
[5] G.T. Raju and P.S.Satyanarayana " Knowledge Discovery from Web usage Data: Complete Preprocessing methodology", International Journal of Computer Sciecne and Network Secutiry, 2008.
[6] Ravidra Gupta and Prateek Gupta, " Application Specific Web Log Preprocessing" , International Journal of Computer Technology and Applications", 2012.
[7] Robert.Cooley, Bamshed Mobasher and Jaideep Srinivastava, "Data Preparation for Mining World Wide Web Browsing Patterns ", Journal of knowledge and Information Systems,1999.
[8] Vijayashri Losarwar and Dr.Madhuri Joshi," Data Preprocessing in web Usage Mining", International Conference on Artificial Intelligence and Embedded Systems", 2012.
[9] Chungsheng Zhang and Liyan Zhuang , "New Path Filling Method on Data Preprocessing in Web Mining", Computer and Information Science Journal , August 2008.
[10] Catlegde L. and Pitkow J., "Characterising browsing behaviours in the world wide Web", Computer Networks and ISDN systems, 1995.
[11] 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.
[12]Thanakorn Pamutha, Siriporn Chimphlee, Chom Kimpan1, and Parinya Sanguansat,"Data Preprocessing on Web Server Log Files for Mining Users Access Patterns",International Journal of Research and Reviews in Wireless Communications (IJRRWC) Vol. 2, No. 2, June 2012
[13]Priyanka Patil and Ujwala Patil , "Preprocessing of web server log file for web mining", World Journal of Science and Technology, 2012.
[14] T. Revathi , M. Mohana Rao, Ch. S. Sasanka and K. Jayanth Kumar, B. Uday Kiran, "An Enhanced Pre-Processing Research Framework for Web Log Data", International Journal of Advanced Research in Computer Science and Software Engineering , Volume 2, Issue 3, March 2012 ISSN: 2277 128X.
[15] Naga Lakshmi, Raja Sekhara Rao , Sai Satyanarayana Reddy, "An Overview of Preprocessing on Web Log Data for Web Usage Analysis",International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-2, Issue-4, March 2013 .

BIOGRAPHY

Ms.R.Rooba is an Assistant Professor in the Department of Computer Technology and Information Technology at Kongu Arts and Science College, Erode , Tamilnadu, India. She is having 12.6 years of teaching experience. She is doing her Ph.D in the area of Semantic Web Technology at Bharathiar University, Coimbatore, Tamilnadu. Her research interests are Data Mining, Semantic Web Mining.

Dr.V.Valli Mayil is a Head and Associate Professor in the Department of Computer Science and Applications at Periyar Maniammai University,Thnajavur, Tamilnadu, India. She is having 18.6 years of experience in teaching and 12 years experience in research. Currently she is guiding 8 Ph.D Scholars. Her area of interest includes Web Mining, Semantic Web, Big data and Cloud Computing.