FP Growth Algorithm for Preprocessing Web Log Data for Web Mining

Serin J
Research Scholar, Research & Development center,
Bharathiar University, Coimbatore, Tamil Nadu, India.
Email:serin.j@gmail.com
Dr.R. Lawrance
Director, Department of Computer Applications,
Ayya Nadar Janaki Ammal College, Sivakasi, Tamil Nadu, India
Email:lawrancer@yahoo.com

Abstract- Today, the www has produce huge amount of data storage. Hence the all users of activities will be stored as a log file. The log file shows the interest on the particular website. With a wide usage of internet, the log file size is growing rapidly. Web mining is the process of extracting information from web data. The raw data won't expose the users' accessing pattern. So, the preprocessing plays a major role in web mining. Web Usage Mining is the very important in data mining to extract and analyze the usage pattern of users using the server log file. The quality of the input decides the quality of the output. Preprocessing is the noteworthy process before mining the interesting information from data. In this paper it has been proposed and implemented the preprocessing techniques to data cleaning, user identification and session identification.

Keywords- log file, preprocessing, user identification, session identification.

1. INTRODUCTION

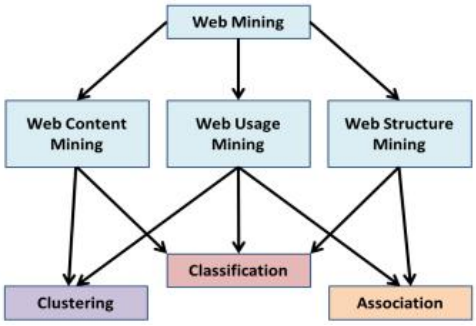
Data Mining is the process of extracting the information from the huge amount of data. It discovers the patterns and relationships exist among the data using various techniques like classification, clustering and association rule mining.

Web mining is the application of data mining techniques based on web data. It can be divided into three domains based on the mining methods such as web content mining, web structure mining and web usage mining.

Web usage mining, also called as Web Log Mining, is the process of mining interesting Patterns in access log files. It helps to uncover the users' behavior who accessing the particular web site through the server log files [1].

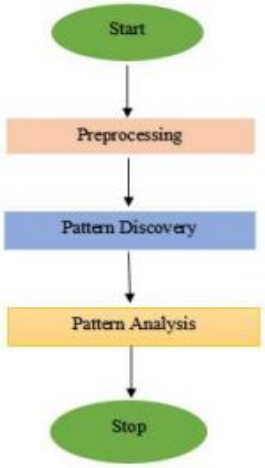
In this paper it has proposed to preprocess the web log data, the raw log file is transform into user sessions. First, the log file is cleaned by removing irrelevant request and error code. In session identification steps, each user page references are divided into session.

This paper is organized as follows section 2 discussed the related works of this research, section 3 discusses the methodology, section 4 shows the experimental details and section 5 the conclusion.



Web usage mining is used to discover interesting usage patterns from web data, in order to understand the needs of web-based applications. It is the third type of web mining and also the application of data mining techniques. The web content and structure mining mines the primary data on the web but web usage mining utilize the secondary data derived from the interactions of the users. Web usage mining analysis results of user interactions with a web server, including weblogs, click streams, and database transactions at a web site of a group of related sites Web Usage Mining consists of a three phase process

a) Pre-processing / Data Preparation
b) Pattern Discovery
c) Pattern Analysis



A. Preprocessing

Data preprocessing is the essential process and must be performed prior to applying data mining algorithms to the data sources. The aim of data preprocessing is identifying the unique users, user sessions and transactions are presented in this paper.

B. Pattern Discovery

The second phase of web usage mining is pattern discovery. In this phase, patterns are discovered from preprocessed data by using some data mining methods like association, clustering and statistical analysis and so on.

C. Pattern Analysis

This is the last phase in the Web Usage Mining process. In this phase patterns are analyzed to extract the useful information from result of pattern discovery data by using knowledge query mechanisms such as SQL or data cubes to perform OLAP operations.

2. RELATED WORKS

In order to do the survey various algorithms have been studied. The researcher must know the basic knowledge of data filters, user identification techniques and session identification techniques. Preprocessing is the significant and time consuming process. The work is deals with the conversion of raw server log file into formatted user session. Data preprocessing is used to process the raw data to prepare it for another processing procedure.

Raiyani et al, proposed a method Distinct User Identification (DUI) to identify the users and help to optimize the performance of preprocessing technique [1].

Maideen et al, presented and discussed a framework called MS Log cleaner. It removes the unnecessary entries from the log file. It removes the noise data in log file based on the user requirement [2].

Castellano, G et al, developed a tool is known as Log Data Preprocessor (LODAP) to preprocess the server log file. They have cleaned the data and find out the session and remove the unnecessary resources from the server log files [3].

Sumathi, C.P et al, discussed that how the log files are preprocessed. They have done data cleaning, user identification and session identification process [4].

Pamutha, T et al, developed a method on preprocessing the log file. They focused on the session identification process. Also they produced the statistical information like total unique IPs, total unique pages, total sessions, Session length and the frequency visited pages.

Langhnoja, S.G et al, identified the user and session of each user after the cleaning process. They cluster the session file after the session identification [5].

Deepa A, et al, they have discussed the log file preprocessing and implemented. They have used filtering technique to remove least requested resources [6].

Valera, M. et al, they have presented data cleaning and distinct user identification technique to enhance the pre-processing steps of web log usage data in data mining [7].

3. METHODOLOGY

Data preprocessing is required stage, which can done by data collection, data cleaning, user identification, and session identification. The goal of preprocessing is to improve the quality and accuracy of data.

A. Data format

The web log data have been collected from E-commerce website. Initially the data size is 1000 records.

29.173.67 – [2017-02-22 21:45:23 30.75.225.192] - - - - 80 GET /_includes/follow/follow_us.php – 200 100 HTTP/1.1 - curl/7.9.8(i686-pc-linux-gnu)libcurl7.9.8(OpenSSL0.9.6b)(ipv6enabled)http://www.bing.com/search?q=SIEM&src=IE-SearchBox&FORM=IE11SR

Fig.1 Data Format of web log file

The fields of the IIS log file are, Client IP address, user name, date, time, service and instance, server name, server IP address, time taken, client bytes sent, server bytes sent, status code, windows status code and request type.

Table 1: Web Log Data Fields

Fields	Values
Date	2017-02-22
Time	00:00:10
IP	30.75.225.192

User Name	- (Mentioned as "-" as user name is not available)
Computer Name	w3svc3
Server IP	30.75.225.192
Server Port	80
Method Used	GET
CS-URI-STEM	/_includes/follow/follow_us.php
CS-URI-Query	search?q=siem&src=ie-searchbox&form=ie11sr
Status	200
CS-Version	http:/1.1
CS-Host	www.bing.com
CS-Referrer	http://www.bing.com/search?q=SIEM&src=IE-SearchBox&FORM=IE11SR

1. Data Cleaning

Data cleaning is the first step of preprocessing phase. It removes the irrelevant request from the log file. It eliminates the graphic file such as (gif, jpg, jpeg, mov, png, mp3, and swf), failure status code, unwanted method and spider navigation. Status code is a value that indicates the success or failure of user request. 100 series indicates that the request is in processing. 200 series indicates that the request is successfully processed. 400 series indicates that the client error and 500 errors indicate the server error. Remove the method except GET and Post.

Web robots is also known as Web crawlers or Web spiders, which are programs that automatically search and download complete Web sites by following every hyperlink on every page within the site in order to update the index of search engine. Requests created by Web robots are not considered usage data and, consequently, have to be removed. All records containing the name "robots.txt" in the requested resource name (URL) are identified and straightly removed [3].

Phase 1 - Data Cleaning Algorithm

Begin
Step1: Read Log Record from Web Server Log File.
Step2: If status code is 200 and method is GET then
Step3: Retrieve the page extension
Step4: Remove fields where '.js" or '.jpg' or '.jpeg' or '.gif' or '.png' or 'robot.txt' or '.css' or audio or video formats
Step5: Retrieve the url page
Step6: Go to step 2.
Step7: Stop, If End of File.
End

3. User Identification

The goal of user identification is to find out whom accessing the website and what type of pages. It creates logical clusters of pages for every user. The unique users can be identified by using IP address and user agent. If the IP address is different, requests are from different users. If it is same, check the user agent. If agent is different, the request is from different user. Otherwise both request from same user. User identification means identifying individual users by using their IP address. There are following rules to identify unique users:

1) If there is new IP address then there is a new user;
2) If the IP Address is same but the visited date is different then there is also a new user.

In this paper, 172 unique users are identified.

Phase 2 – User Identification

Begin
Step1: Read Log Record from Web Server Log File.
Step2: Compare the IP Address and User Agent
Step3: Identify same user when both IP address of first log entry and second log entry are same
Step4: Identify different user when IP address of the first and second log entry are different
Step5: Go to step 2.
Step6: Stop, If End of File.
End

The data cleaning is an important process of data preprocessing. It is used to remove the irrelevant and duplicate records from the log file that are not required for mining. After cleaning is done, the data set will be reduced from 1000 to 910.

4. Session Identification

After the user identification, the next step is to identify the session in the log file. The session identification is the process of splitting the user into the group of pages according to the time interval. That is for every user find out the session [8]. There are 2 heuristics method to find out the session. Such as: time oriented and structure oriented. In this paper, it has been applied time oriented heuristics. When the duration is exceeds the threshold value then starts the new session. The default threshold value is 30 minutes [9].

Phase 3 – Session Identification

Begin
Step1: Read Session File.
Step2: For all page P do
Step2: Compare the time between page request and check new user or not
Step3: Assume new session where time between page request exceeds certain time limit or Referred page is null or new user
Step4: Assume different session, where a new user there is is new session.
Step5: Go to step 2.
Step6: Stop, If End of File.
End

4. FP GROWTH FOR WEB MINING

The FP Growth algorithm is used to perform item set mining. Item set mining means to find frequent patterns in web log data. This algorithm has three traversal approaches like top-down, bottom-up and hybrid. Each item is stored together with its id list and compute the support of an item set by using the intersection based approach [10].

4.1 FP stands for frequent pattern

In the first pass, the algorithm counts occurrence of items in the dataset, and stores them to 'header table'. In the second pass, it builds the FP-tree structure by inserting instances. Items in each instance have to be sorted by descending order of their frequency in the dataset, so that the tree can be processed quickly. Items in each instance that do not meet minimum coverage threshold are discarded. If many instances share most frequent items, FP-tree provides high compression close to tree root.

Recursive processing of this compressed version of main dataset grows large item sets directly, instead of generating candidate items and testing them against the entire database. Growth starts from the bottom of the header table by finding all instances matching given condition. New tree is created, with counts projected from the original tree corresponding to the set of instances that are conditional on the attribute, with each node getting sum of its children counts. Recursive growth ends when no individual items conditional on the attribute meet minimum support threshold, and processing continues on the remaining header items of the original FP-tree.

Algorithm 2: FP-Growth

Input: A database DB, represented by FP-tree constructed according to Algorithm 1and a minimum support threshold

Output: The complete set of frequent patterns.

Method:

Call FP-growth(FP-tree, null).

Procedure FP-growth(Tree, a)

{

(1) if Tree contains a single prefix path then

{

(2) let P be the single prefix-path part of Tree;

(3) let Q be the multipath part with the top branching node replaced by a null root;

(4) for each combination (denoted as ß) of the nodes in the path P do

(5) generate pattern ß ∪ a with support = minimum support of nodes in ß;

(6) let freq pattern set(P) be the set of patterns so generated;

}

(7) else let Q be Tree;

(8) for each item ai in Q do { // Mining multipath FP-tree

(9) generate pattern ß = ai ∪ a with support = ai .support;

(10) construct ß's conditional pattern-base and then ß's conditional FP-tree Tree ß;

(11) if Tree ß ≠ Ø then

(12) call FP-growth(Tree ß , ß);

(13) let freq pattern set(Q) be the set of patterns so generated;

}

(14) return(freq pattern set(P) ∪ freq pattern set(Q) ∪ (freq pattern set(P) × freq pattern set(Q)))

}

The goal of the paper is to mine the frequent link from the web log file by using the FP Growth algorithm. Frequent link mining also used to find information like set of pages repeatedly accessed together by web users. The administrator can modify the website according to the mining result.

5.EXPERIMENTAL RESULTS

In this research work, it has been used web log dataset. The data set rows is 9000 web log data. It is in IIS log file format. This work has been implemented and in java for experimental results.

Table 2: Log File

S.No	Log Records
1.	2017-02-22 21:45:23 30.75.225.192 - - - - 80 GET /_includes/follow/follow_us.php – 200 100 HTTP/1.1 - curl/7.9.8(i686-pc-linux-gnu)libcurl7.9.8(OpenSSL0.9.6b)(ipv6enabled) http://www.bing.com/search?q=SIEM&src=IE-SearchBox&FORM=IE11S
2.	2017-02-22 22:13:23 70.69.152.165 - - - - 80 GET /blog/index.php – 200 100 HTTP/1.1 - Mozilla/5.0(Linux;U;Android4.0.4;en-ca;SGH-I757MBuild/IMM76D)AppleWebKit/534.30(KHTML,likeGecko)Version/4.0Mobile-Safari/534.30
..	…

Table 2 represents the raw log data file, the raw data has been cleaning using data cleaning algorithm which are explained in the methodology section. Data cleaning is the first step of preprocessing phase.

It removes the irrelevant request from the log file. It eliminates the graphic file such as .gif, .jpg, .jpeg, .mov, .png, .mp3, and .swf and robot.txt, failure status code, unwanted method and spider navigation. Table 3 shows the cleaned data.

Table 3: After the Data Cleaning

IP Address	Method	Path	Status	Protocol
30.75.225.192	GET	/_includes/follow/follow_us.php	200	http/1.1
70.69.152.165	GET	/blog/index.php	200	http/1.1
34.87.4.6	GET	/shopping/cart/confirm.jsp	200	http/1.1
147.106.118.104	GET	/shopping/cart/checkout.jsp	200	http/1.1
5.35.225.115	GET	/_includes/follow/follow_us.php	200	http/1.1
128.196.108.201	GET	/_includes/follow/follow_us.php	200	http/1.1
70.69.152.165	GET	/_includes/follow/follow_us.php	200	http/1.1

The second phase of preprocessing is user identification is to find out whom accessing the website and what type of pages. The unique users can be identified by using IP address and user agent. If the IP address is different, requests are from different users. If it is same, check the user agent. If agent is different, the request is from different user. Otherwise both request from same user. Table 4 represents the user identification process. The user identification gives the total number of unique users visited the website. Here the total number of unique visitors is 7.

Table 4: User Identification

SNO	User Name (IP_Address | User_Agent)
1.	30.75.225.192
2.	70.69.152.165
3.	34.87.4.6
4.	147.106.118.104
5.	5.35.225.115
6.	128.196.108.201
7.	70.69.152.165
...	.....

After the user identification phase, next to identify the session in the log file using session identification algorithm. The session identification is the process of splitting the user into the group of pages according to the time interval. This phase gives us the total number of sessions created by various users. Table 5 represents the session identified table, where, pages in session 3 are viewed consecutively. In session 4 the first page has no referrer hence it is treated as a different session. In Session 5, the page has a referrer but the time duration between the previous page and current page exceeds 30 Minutes hence it is also treated as a new session.

Table 5: Session Identification

Session ID	User who used the session	Pages Accessed during the session
1.	30.75.225.192	/_includes/follow/follow_us.php
2.	70.69.152.165	/blog/index.php
3.	34.87.4.6	/shopping/cart/confirm.jsp
4.	147.106.118.104	/shopping/cart/checkout.jsp
5.	5.35.225.115	/_includes/follow/follow_us.php
6.	128.196.108.201	/_includes/follow/follow_us.php
7.	70.69.152.165	/_includes/follow/follow_us.php
...	....	....

6. CONCLUSION

Data Preprocessing is one of the important tasks before applying mining algorithms. It converts the raw log file into user session. In this research it has been implemented server log file. Through this preprocessing phase number of accessed users can be easily identified. Through session identification process to determine the most frequently accessed web page and least frequently accessed page also relationship among web pages. The preprocessed data will be used to discover the access pattern efficiently. After cleaning the data using Data Cleaning method, the number of records was reduced and the file size was significantly reduced and gives a reduction. Web Usage Mining is one of the parts of web mining and extracts the web users' behavior from log files. Preprocessing is done by data cleansing, user identification, session identification. Data cleaning is used to reduce the size of web log file and also improves the quality of contents in the log file. Then the FP growth algorithm is applied in preprocessed data for mining the frequent link from web log data. The FP Growth algorithm has high performance compared to Apriori. Finally the Discovered Patterns can be analyzed using Pattern Analysis methods and interesting patterns can be obtained which can be used to personalize websites and improve design of web pages. In future the preprocessed log data will be applied to various data mining techniques to discover the usage pattern and user behavior using data mining.

REFERENCES

[1] Raiyani, ashwin g., and sheetal s. Pandya. "Discovering User Identification Mining Technique for Preprocessed Web Log Data."

[2] Maideen, C. M., & Palanivel, M. K. MS Log Cleaner: A framework to discover efficient use of web service.

[3] Castellano, G., Fanelli, A. M., and Torsello, M. A. 2007. Log data preparation for mining web usage patterns. In IADIS InternationalConference AppliedComputing (pp. 371-378).

[4] Sumathi, C. P., et al., 2011. An Overview of Preprocessing Of Web Log Files For Web Usage Mining. Journal of Theoretical and Applied Information Technology, ISSN: 1992-8645.

[5] Langhnoja, S. G., Barot, M. P., & Mehta, D. B. (2013). Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery.International Journal.

[6] A Deepa and P Raajan. Article: An Efficient Preprocessing Methodology of Log File for Web Usage Mining. IJCA Proceedings on National Conference on Research Issues in Image Analysis and Mining Intelligence NCRIIAMI 2015(2):15-16, June 2015.

[7] Valera, M. and Parmar, S., A Step up in Data Cleaning and User identification of Preprocessing on Web Usage data.

[8] Sumathi C. P, Padmaja Valli. R, Santhanam . T, An Overview of preprocessing of web Log files for web usage Mining, 2011.

[9] Rajashree Shettar, International Journal of Engineering Science and Advanced Technology, 2014.

[10] Sudheer Reddy. K, Partha Sarandhi Varma, Kantha, International journal of Computer Theory and Engineering, 2014.