ISSN:0975-9646

Survey on Data Preprocessing Method of Web Usage Mining.

Wasvand Chandrama[1], Prof.P.R.Devale[2], Prof. Ravindra Murumkar[3]

[1]*Research scholar of Department of Information technology,*
Bharati Vidyapeeth University College of Engineering,
Pune, Maharashtra, India.,

[2,]*Department of Information technology,*
Bharati Vidyapeeth University College of Engineering, Pune, Maharashtra, India ,

[3] *Department of Information technology, Pune University, Pune Institute of Computer Technology*
Pune, Maharashtra, India.

Abstract-By Web Usage Mining from raw data it is possible to make the relevant information which is helpful in analysis. User identification in the web is possible by Web usage Mining. User interacts with web by providing set of keywords in their query. Web usage mining is accomplished by three different processes, Data Preprocessing, Pattern Discovery and Pattern Analysis. Data Preprocessing is important to ensure the quality of web log mining. Result of Preprocessing has direct influence on the choosing of mining algorithm.

This paper discuss about the different algorithms and methods which are available for data preprocessing. For analyzing user navigational pattern the methods used are like Correlation Regulation Discovery, Sequence Pattern Recognition and Cluster pattern. The algorithm, K-mean is used for Clustering and Apriori is used for Correlation Regulation.

Keywords: Web Usage Mining, Data Preprocessing, Data Cleaning, User Identification, Pattern Discovery.

I. INTRODUCTION

There are many applications which use web usage mining for analyzing user navigational pattern. Web usage mining is technique of web mining. Data preprocessing is required and important phase in web usage mining. The data cleaning and user identification are method in Data Preprocessing. The web log file is data source to data preprocessing method. The purpose of Data cleaning is to eliminating irrelevant items. The task of User identification is to identify who access the web site and which pages are accessed in web site.

Current research is on data preprocessing methods which are data cleaning and user identification. Different technique are provided for data cleaning but still there are problems remain in data collection and accuracy metric of user identification. This paper provides review on algorithm and different technique used in data preprocessing that are used for web usage mining.

Data preprocessing is used to clean the data so that when it provide to the pattern discovery it will identify the technique which will used to discover the users navigational pattern and after processing it will passes that to pattern analysis so that it will take only relevant pattern and removing irrelevant pattern.

Data mining is the data-driven techniques to discover patterns in large volumes of raw data. Web mining can be referred as the extension of the data mining techniques to web data. Web mining has three different phases that include – web content mining, web usage mining and web structure mining of web data. Mining the content involves extracting the useful information from content of the web document. Web structure mining studies the structure information from the web and it perform the structure mining on hyperlink level. Web Usage Mining (WUM) is discovery of meaningful patterns from data generated by client-server transactions on one or more Web localities. WUM is about identifying user browsing patterns over WWW, with the aid of knowledge acquired from web logs. The outcomes of the WUM can be used in web personalization, improving the performance of the system, modification of the site, business intelligence, usage characterization etc.

A. Web usage mining:

Web usage mining is performing in three steps – Data preprocessing, pattern discovery and pattern analysis. Results of the pattern discovery directly influenced the quality of the data processing. Good data sources discover quality patterns and also improve the WUM algorithm. Hence, data preprocessing is an important activity for the complete web usage mining processes and vital in deciding the quality of patterns. In data preprocessing, the collection of various types of data differs not only on type of data available but also the data source site, the data source size and the way it is being implemented. These steps see in detail:

The data preprocessing of Web usage mining is usually complex. Purpose of data preprocessing is to offer reliable, structural and integrated data source to pattern discovery. It consists of four processes: Data cleaning, User identification, Session identification, Path completion [4].

Pattern discovery is the key process of the Web mining, which covers the algorithms and techniques from several research areas, such as data mining, machine learning, statistics and pattern recognition. The techniques such as statistical analysis, association rules, clustering, classification, sequential pattern and dependency modeling are used to discover rules and patterns. The knowledge that can be discovered is represented in the form of rules, tables, charts, graphs, and other visual presentation forms for characterizing, comparing, predicting, or classifying data from the Web access 1og.

The final stage of the Web usage mining is pattern analysis. The aim of this process is to extract the interesting rules or patterns from the output of the pattern discovery process by eliminating the irrelative rules or patterns.

Here we focus on data preprocessing method of WUM.

B. Data Preprocessing:

Data Preprocessing include the following processes:

1. Data cleaning.
2. User identification.
3. Session identification.
4. Path completion.

1) Data cleaning

The web log file is data source to data cleaning process. The purpose of data cleaning is to eliminate irrelevant items or irrelevant record from the log file.

2) User identification:

After data cleaning, user identification is done. In user identification how many users visited the web site is to be identified. This is done with the help of IP address and agent the user has used for browsing the web site.

3) Session identification:

Session is the time between logged in and logged out. During this time user visit many pages. Session is to find the sequence of pages and trace the user activity.

4) Path completion:

There are some reasons that result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators (URL) recorded in log may be less than the real one. The purpose of path completion is to discover user's travel pattern, the missing pages in user access path should be appended.

II. **RELATED WORK:**

Huaqiang Zhou, Hongxia Gao and Han Xiao [1] focus on data preprocessing method. In which they provide five processes. Those are Data cleaning, user identification, path completion session identification and transaction identification. In transaction identification data in user session may be too large and needed to convert into smaller transaction by using segmentation algorithm to identify. In analysis of preprocessing method the Frame Page Filter and Time out Threshold Value Setting used.

Here in frame page filtering putting the frame page between two steps above of session identification and path completion. So if sub frame page of session is deleted that is the information contain in subframe is also deleted or lost. So they define that there must be site structure in which they contain frame subframe relational table and whatever you want to extract useful frame subframe then extract it from that table. Through the frame page filtration it is possible to searching of frame page and subframe page.

In Time out threshold value setting is used to identify the user session. They adopt the time out method in the process of session identification. So that the different value of requested time between two pages exceeded certain limit new session is started. To adjust these threshold value of access time for pages.

In Improve the filtration method of frame page as mentioned above applying the filtration method of frame page can effectively eliminate the influence of frame page on the log mining. Using filtration algorithm of frame page is judging the each page of each user session that is it frame or subframe and deleting subframe page. Decision tree sorting algorithm is used. It allowing quick sort of multi granularity layer to solve this problem.

In this paper it adopt ID3 algorithm in decision tree algorithm and improve the filtration method of frame page in this way the filtering efficiency can be improved. ID3 algorithm is greedy algorithm which uses divide and rule method to constructing decision tree.

In decision tree to establish the root node of tree they first check all characteristics of training data set and select feature with largest information gain and make it as root node and according to different values of this characteristic and conduct recursion on example subset of each branch. Using this algorithm the quality of data preprocessing result is improved also it increased the interest degree of the mining result in large degree. It increases efficiency of whole web log.

Theint Theint Aye [2] proposed a technique for preprocessing i.e. Field Extraction and Data Cleaning. Main task of clean the web log file and insert that processed data into a relational database so that we apply data mining technique on it.

They provide Field Extraction algorithm. Field extraction is the process of separating field from the single line of log file. Server used different character like (comma) ',' or space character which works as separator. In Field Extraction algorithm they mention to open database connection and create table to store log data file. Then open data log file and read all field contain in log file. Separate out the attribute in the string Log. Extract all fields and add into the log Table (LT).

Data Cleaning algorithm is used to eliminating irrelevant or unnecessary items in the analyzed data. Web log file also record the failed HTTP status codes and suffix. So by data cleaning inconsistency will be detected and removed to improve quality of data. Here in algorithm for data cleaning they used log table as an input which is generated after field extraction. It read each record in log table. Read the field like status code and method. If status code and method is ** then get IP address and URL link. And if suffix. URL_link = *.gif, *.jpg, *.css then remove suffix.URL_link.

In web server logs paper [3] author mention the four types of server logs. Those are Access log, Agent log, Error log and Referrer log and different log format like W3C log file format, IIS log file, NCSA log file. In data preprocessing method like data extraction, joining extracted log file, data cleaning and filtering. Here they provide methodology for extracting log file and joining log file.

Here they provide two efficient algorithms for web log i.e. extracting web log and Joining web log. These two algorithms are easy to implement and proved to be an efficient. Incorporating these two algorithms and included it in functionality of developed tool called WebIS. WebIS is software tool for retrieving and joining web log file for processing.

Here they mention extracting log file algorithm. Extracting log file from web server there are several log files are exist in one or more server. So we can extract numerous log files from various servers at a time. They provide one approach for reading web server log files. Here they mention extracting log file algorithm. In which they use two const operators to read and write. They used one SRKREC portal which focus on online education resources and provide information related to educational processes such as online tutorials question banks etc. Then set LogReader to create Server.CreateObject in which it read SRKR log. With the help of LogReader it open log file and ReadLogRecord. While not then they cal to LogReader.EndOfLogRecord. That is the end of that file.

In Joining log files means log contain several log file that gather the request from all log file into joint file. The name of server is not included in the request of log file. But we add this information in the request in order know web server name to differentiate between request made to various web server. Also have to synchronize the web server clock so for time synchronizes they provide joining algorithm. By applying data fusion and cleaning they reduce the size of log. The entry in server log contains a time stamp of traversal from a source web page to target web page.

In novel pre-processing technique for web log mining by removing global noise and web Robots paper [6] author propose a new technique for data cleaning. In which they include additional method are Elimination of Local and Global Noise and Robot Cleaning method.

In elimination of global and local noise, the global noise corresponds to unnecessary objects with huge granularities, which are no smaller than individual pages. It remove noise includes mirror sites, duplicated web pages and previous versioned web pages, etc. But still most pages have some noise such as “contact”, “company profiles”, “copyright” and other noise words.

Local noise corresponds to irrelevant items inside a web page. This noise includes banner ads, navigational guides, decoration pictures etc. these noises have to remove for better result.

In Robot Cleaning, web robot is software tool that periodically scans a web site to extract its content. It automatically follows all hyperlinks from web pages. Google also use web robot to gather all the pages from web site in order to update their search indexes. Eliminating web robot generated a log entries not only simplifies the mining task but also remove uninteresting session from log file. To identify web robot request, all records containing name “robot.txt” in requested recourse name (URL) are identified and removed. And in other crawler retrieves pages in automatic and exhaustive manner so they distinguished by high browsing speed. Therefore each different IP add and browsing speed us calculated and all request with this value more than threshold are regarded as made by robot and consequently removed.

Preprocessing phase with robot cleaning was carried out using UCI machine learning repository. The dataset used for preprocessing with robot cleaning is Anonymous Microsoft web dataset and MSNBC.com anonymous web dataset. After robot cleaning the time required for prediction of user interested pattern using initial log is just half of the time actually required.

Arvind Kumar Dangi and Sunita Sangwan proposed [4] a new approach for user identification in web usage mining. Here they provide a new method for data preprocessing in which first phase they select some website and different location that access these website. In second phase applying the java tool and method then find out IP address of that website. In the final phase combine them i.e. web link navigation + IP address of website + session of usage. This framework helps to investigate the web user usage behavior efficiently.

With the help of java language by using function and code the IP address identification and web link visited identification can done. Here they create a connection object. The Logger class is used. The java.util.logging package provides the logging capabilities. The Logger you create a actually a hierarchy of Logger, and a . (dot) in the hierarchy indicates a level in the hierarchy. So if you get a Logger for the com. example key this Logger is a child of the com Logger and the com Logger is child of the logger for empty string. Configure the main logger and this affects all its children.

Data integration means dataset that can be retrieved and added to other datasets to create greater robust and useful dataset. The data integration is used for achieving the consistency access and delivery of data across the spectrum of data subject areas.

To minimize the redundancy and dependency of the data Data makeover process is applied. It applied on data that which is collected from different sources of data. To maintain smaller volume of data integrity of the original data the data diminution is used.

E-commerce is one of the applications of WUM. The most important think of e-commerce is to understand what customer wants love and value orientation as much as possible. Here they can use personalized service means when user browses the web site. It constantly meets the each user browsing interest. So that each user feel like he or she is unique user [5].

Mahendra Pratap yadav, Mhd feeroz and Vinod Kumar Yadav prepose a new approach for customer behavior using web usage mining in E-commerce [5]. Here they use a term like correlation regulation discovery, sequence pattern recognition and cluster pattern.

In correlation regulation discovery means searching of series of pages that satisfy supportability from one time client conversion and by analyzing the correlation regulation to predict next page client will want to visit. So here for that they are using apriori algorithm to satisfy supportability.

Sequence pattern recognition makes concern with correlation associated with the time. The sequence patterns have more type and access path pattern are important. Using expanding directed tree model to observe maximum forward path and frequented visited client path.

Cluster pattern here two kind of client cluster and page cluster. In Client cluster similar behavior of client in client cluster. According to interest of client the website delivers the services. And in Page cluster the content of page are similar kind that are page cluster. So for that they using k-means algorithm.

III. CONCLUSION:

Before data cleaning, the exaction of log file and joining of log extraction file is necessary. Field Extraction algorithm is better than extraction of log file and joining of log file. With the help of anonymous Microsoft web dataset the time gets reduced to predict user interested pattern. User identification can be done with the help of IP address plus web link navigation plus session usage and combine them together for user identification and pattern discovery.
K-mean algorithm is good for analyzing customer interest.

ACKNOWLEDGEMENT:

We finished paper under support of Department Information Technology of Bharti Vidyapeeth, and Head of the Department.

REFERENCES:

[1] Huaqiang Zhou, Hongxia Gao and Han Xiao ”Rsearch on Improving method of Preprocessing in web log mining”, IEEE 2010

[2] Theint Theint Aye ” web log cleaning for mining of web usage patterns”, IEEE 2011

[3] K Sudheer Reddy, G. Partha Saradhi Varma and I Ramesh Babu “Preprocessing the web server logs an illustrative approach for effective usage mining”, ACM 2012

[4] Arvind Kumar Dangi and Sunita Sangwan, “A new approach for user identification in web usage mining Preprocessing”, IOSR-JCE may 2013

[5] Mahendra Pratap yadav, Mhd feeroz and Vinod Kumar Yadav, “Mining the customer behavior using web usage mining in E-commerce” IEEE 2012

[6] P.nithya, Dr.P sumathi “Novel Pre-Processing Technique for web log mining by removing global noise and web robots.” IEEE 2012