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. 2, Issue 12, December 2014

Mining of Web Logs Using Preprocessing and Clustering

Pankaj M. Meshram, Prof. Gauri A. Chaudhary

PG Student, Dept of CSE, Y.C.C.E, Nagpur (MS), India

Asst. Professor, Dept of CT, Y.C.C.E, Nagpur (MS), India

ABSTRACT: The data pre-processing plays a major role in efficient mining process as Log data is normally noisy and indistinct. In data pre-processing method the rebuild of session and paths are going to complete by annexing lost pages. Additionally the transaction which explains the behaviour of users made accurate in pre-processing by calculating the time taken by the user to view particular page is accessed in the form of byte rate. By using web clustering various types of object can be clustered into different groups. The belief function similarity measures in algorithm include the clustering task by Dempster – Shafer's theory. The main aim of this work is to achieve pre-processing and clustering of web log and to improve the website performance.

KEYWORDS: Pre-processing; Data Cleaning; Clustering.

I. INTRODUCTION

Today, the internet is very important in order to collect the information as early as possible in this hasty life. Whenever we connect to internet web confused if we do not know the correct URL, as there are many pages for the information you want and more are adding day by day. Web mining is the mining of data i.e whatever we want whether it is chart technology, Artificial intelligence and so on, we must contain it and will be available to the user according to their need. This data is stored in the well precise pattern which makes it easy to retrieve data easily. Today the activities like web designing, creating attractive web sites are the part of web usage mining techniques.

Whenever the user requests any data, the web server begins to gather data in log file. Which contains client IP address, URL requested etc. but all the information comes with the different format is mostly issued by apache and IIS. Web usage mining consist of pre-processing in which unwanted hits of record discarded in mining process, which involves identity of user, their IP address, session identification etc. The unwanted records get cleaned during the mining called as data cleaning. Session identification can be perceived by following example. Considering the user logging to social sites uses the page for 30 minutes that means the session time will be recorded as 30 minutes, if the user hitting the other page the session time will be changed. The total number of pages hit by user creates the user session. The pattern analysis and clustering based techniques used with the help of belief function.

II. LITERATURE REVIEW

Web usage mining is the invention made for user while accessing the internet. It's about the user requiring the particular information from the web server. Web usage mining is the application involves the different usage patterns and techniques which make it easy to use and satisfy the needs of web based application. This application is also useful to know the browsing behaviour of the user at the website.

The source of recognizing health status of a system log files are used and it is also used to capture action performed in a computer system and networks. Logs are collection of log entries and each log file consists of information related to particular action that has performed in a network. Many logs also contains the records related to the computer security which are produced by many sources including operating system on servers, workstation computers, networking equipments and other security software's, such as antivirus, firewalls, intrusion detection and prevention systems and many other applications. Regular log analysis is advantageous for identifying security fraudulent activity, policy

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. 2, Issue 12, December 2014

violations and other operational problems. It is also helpful for executing forensic analysis, internal investigation recognizing, operational trends and long term problems. At standing log were used to overcome problems, but now days it is used for many operations in many organization and associations like optimizing system and network performance, recording user actions and providing data for investigating malicious activity[7].

Ling Zang[10] used an improved data preprocessing technology, in order to solve some existing problems in traditional data preprocessing technology for web log mining.

Doru Tanasa[8] whose contribution for a web users mining is appreciable an impressive theoretical procedure for processing the web logs, who invented the general methodology with their approaches their relation with concrete method and their patterns.

The web log records where processed with the help of algorithm FP-growth by the Huiping peng[9] who found the set of frequent access patterns. The browsing activity with site topology and their associate rules for web mining, he created the data which helps in the process of constructions of new sites.

The first step of pre-processing is data cleaning which is used for removal of outliers. Analysing millions and trillions of records in server logs is a cumbersome activity. If a user requests a specific page from server entries which consist of gif, JPEG, etc., are irrelevant records and removed from the logs. The records indicate that the client's request cannot be fulfilled, due to incorrect syntax or a missing file also ignored from logs. Automated programs are requested to server as per fixed time limit, if the time taken is less than 2 seconds such type of entries are eradicated.[1] The time taken by the user to view a particular page is called as Reference Length [2]. If the user plays a video on website or views a particular image, the time spent by the user will not be considered. The time removed from the web page referred as stay time. To find actual browsing time of user is a big deal.

The data transfer rate and size of page is also considered and the reference length is calculated as
RL time = RLT' - bytes_sent / c
Where RLT' [6] denotes the difference between session time and number of hits other than the particular one and bytes_sent is taken from log entry of a record and c is the data transfer rate.

The next important step is unique user identification. The User Identification [3] is done by identifying pages which are accessed and who accessed the website. The fields which are useful to find unique user IP address, referrer field, user agent field. The IP address represents the unique identification. If IP addresses of two records are same then browser information is checked. If user agent values of two records are same then they are identified as same user because user agent field provides information about the client's browsers, the browser version and the client's operating system.

A series of requests made by a single user over a certain period of time is called user session. The main aim of session identification is to separate the page accessing each user into individual sessions. These sessions are used for various predictions, classification, clustering into groups and other tasks. The referrer URL field's present record is matched with past referred record if this is not used previously and referrer URL field is empty then it is considered as a new user session. The number of pages visited by specific user called page viewing time and default time is 30 minutes [4]. The page stay time calculated based on difference between two timestamps that depend on time spend on a particular page. If the time goes above 10 minutes the second entry is understood as a new session.

The web personalization involve the different phases like data collection and pre-processing, pattern discovery and evaluation together which creates the useful information for data mining which helps in perceiving the behaviour of the user towards the web. To accomplish the purpose of the multiple data sources the flexibility is established under web personalization which is effectively discovered in an automated system. The web personalization can also be aided by data mining algorithm[5] include clustering techniques, association rule mining, sequential pattern mining. At last, for effective personalization, data mining frameworks uses the variety of channels.

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. 2, Issue 12, December 2014

III. PROPOSED WORK

In data mining there are different kinds of log data from different datasets are used and it contains a various problems occurred at the time of preparation. The main problem is not getting a reliable dataset for mining. The prepared dataset is to be constructed under transactions and users accessing behaviour will be reliable

Data pre-processing techniques improve the overall quality of the patterns. The raw data is processed in advance to get reliable session for efficient mining. It contains different domains dependent task as like data cleaning, user identification, clustering, session identification and construction of transactions. The number of visitors on a particular site, their getting the information is watched by the web server. The web server log file plays important role in web usage mining as the server logs collect all the information, their session online behavior and interest towards the specific information.

IV. CONCLUSION

The analysis and implementation of data processing system can be performed through web usage mining for log data which confines the data cleaning, user identification and clustering. The unwanted hits like robot entries get cleaned regularly. The time taken by the user on a particular page is computed under the reference length. The total number of reference length creates the session time in toto. All the steps of pre-processing are effective enough to furnish a reliable input for data mining. There are several other theories which give the useful clustering techniques. But the trivial input is the one which reads each and every second for the record.

REFERENCES

1. Istvan K. Nagy and Csaba Gaspar-Papanek “User Behaviour Analysis Based on Time Spenton Web Pages”, Web Mining Applications in E-commerce and E-Services, Studies in Computational Intelligence, Springer, 2009
2. Yan Li and Boqin FENG “The Construction of Transactions for Web Usage Mining,” International Conference on Computational Intelligence and Natural Computing, IEEE, 2009.
3. Robert.Cooley, Bamshed Mobasher and Jaideep Srinivastava, “Data Preparation for Mining World Wide Web Browsing Patterns”, International journal of knowledge and Information Systems,1999.
4. N. M. Abo El-Yazeed,“Weblog pre-processing based on partial ancestral graph technique for session construction”.
5. Bamshad Mobasher “Data Mining for Web Personalization, ” LCNS, Springer-Verleg Berlin Heidelberg, 2007.
6. B.Uma Maheswari and Dr. P.Sumathi,” A NewClustering and Pre-processing for Web Log Mining”, IEEE, 2014.
7. Muhammad Kamran Ahmed, Mukhtar Hussain and Asad Raza “An Automated User Transparent Approach to log Web URLs for Forensic Analysis” Fifth International Conference on IT Security Incident Management and IT Forensics 2009.
8. Doru Tanasa and Brigitte Trousse, ”Advanced Data Preprocessing for Intersites Web Usage Mining “, Published by the IEEE Computer Society, pp. 59-65, March/April 2004.
9. Huiping Peng, “Discovery of Interesting Association Rules Based On Web Usage Mining”, IEEE Conference, pp.272-275, 2010.
10. Ling Zheng, Hui Gui and Feng Li, “Optimized Data Preprocessing Technology For Web Log Mining”, IEEE International Conference On Computer Design and Applications (ICCDA), pp. VI-19-VI-21, 2010.