A Survey on Framework for Improved Web Data Clustering Using Language Processing Technique

Ms. Aashwini T Thakare
Asst. Professor CSE, GNIEM
GNIEM Nagpur, Maharashtra, India
e- mail: ashwinithakre25@gmail.com

Prof. M. S. Chaudhari
HOD, CSE, PBCOE
PBCOE Nagpur, Maharashtra, India
e-mail: manojchaudhary2@gmail.com

Abstract: Now a day, World Wide Web becomes very popular and interactive for transferring of information. It is a massive repository of web pages and links. It provides information about vast area for the internet user. The web is huge, diverse and active and thus increases the scalability, multimedia data & temporal matters. The growth of the web has outcome in a huge amount of information that is now freely offered for user access. Since due to tremendous usage, the log files are growing at a faster rate & the size is becoming huge. Preprocessing plays a vital role in efficient mining process as log data is normally noisy and indistinct. Reconstruction of session and paths are completed by appending missing pages in preprocessing. Additionally, the transactions which illustrate the behavior of users are constructed exactly in preprocessing by calculating the Reference Length of user access by means of byte rate, the clustering task the ability to capture the uncertainty among web user's navigation performance.

Keywords: *WWW, Web Usage Mining, Preprocessing, Clustering.*

*****

1. Introduction

In this internet era, web sites on the internet are useful source of information in everyday life. Therefore there is an enormous development of World Wide Web in its volume of traffic and size and complexity of web sites. World Wide Web is a huge repository of web pages and links. It provides abundance of information for the Internet users. The growth of web is tremendous as approximately one million pages are added daily. 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 data mining is the application of data mining techniques in web data. Web Usage Mining applies mining techniques in log data to extract the behavior of users which is used in various applications like personalized services, adaptive web sites, customer profiling, and pre-fetching, creating attractive web sites etc. Data mining is defined as the automatic extraction of unknown, useful and understandable patterns from large database. To increase the performance of web sites better web site design, web server activities are changed as per users' interests. Web mining is the application of data mining techniques to automatically retrieve, extract and evaluate information for knowledge discovery from web documents and services. The objects of Web mining are vast, heterogeneous and distributing documents. The logistic structure of Web is a graph structured by documents and hyperlinks, the mining results may be on Web contents or Web structures. Web mining is divided into three types. They are Web content mining, Web structure mining and Web usage mining. Web Content Mining deals with the discovery of useful information from the web contents or data or documents or services. Web Structure Mining mines the structure of hyperlinks within the web itself. Structure represents the graph of the link in a site or between the sites. Web Usage Mining mines the log data stored in the web server.

2. Problem Definition

The first step is to define the time frame for which the log of data is collected and worked upon. This question is imperative to the size of the dataset. The size of the access log of 1 day is more than 250 -500 MB. It is difficult to obtain any useful pattern sequence by analyzing a single day's access log. Therefore, the main problem faced is quantity of time frame for which the data should be processed to produce realistic statistical analysis and discover pattern and results that would do justice to the project and produce reasonable results.

3. Literature Survey

J. Vellingiri, S.Chenthur Pandian presented a survey paper on Web Usage Mining, Web usage mining is the type of data mining process for discovering the usage patterns from web information for the purpose of understanding and better provide the requirements of web-based applications. Web usage mining involves of three phases, namely, preprocessing, pattern discovery and pattern analysis. There are different techniques available for web usage mining with its own advantages and disadvantages.

Hemanshu Rana, Mayank Patel presented a survey paper on Web Log Analysis Using Clustering Techniques, Web usage mining is the area of web mining which deals with the extraction of interesting knowledge from web log information produced by web servers. Web usage mining techniques can be applied for web log analysis. The first one is simple K-means, second K-means using Neural Network concept and Self Organization Map (SOM).

V.Chitraa, Dr. Antony Selvdoss Davamani survey on Preprocessing Methods for Web Usage Data, the objects of Web mining is vast, heterogeneous and distributing documents. Web mining is divided into three types. They are Web content mining, Web structure mining and Web usage mining. Web Content Mining deals with the discovery of useful information from the web contents or data or documents or services. Web Structure Mining mines the structure of hyperlinks within the web itself. Structure represents the graph of the link in a site or between the sites. Web Usage Mining mines the log data stored in the web server.

4. Praposed Work

4.1 Web Usage Mining:

Web usage mining also known as web log mining is the application of data mining techniques on large web log repositories to discover useful knowledge about user's behavioral patterns and website usage statistics that can be used for various website design tasks. The main source of data for web usage mining consists of textual logs collected by numerous web servers all around the world. There are four stages in web usage mining.

1. Data Collection*:* users log data is collected from various sources like server side, client side, and proxy servers and so on.

2. Preprocessing: Performs a series of processing of web log file covering data cleaning, user identification, session identification, path completion and transaction identification.

3. Pattern discovery: Application of various data mining techniques to processed data like statistical analysis, association, clustering, and pattern matching and so on.

4. Pattern analysis: Once patterns were discovered from web logs, uninteresting rules are filtered out. Analysis is done using knowledge query mechanism such as SQL or data cubes to perform OLAP operations.

All the four stages are depicted through the following figure.

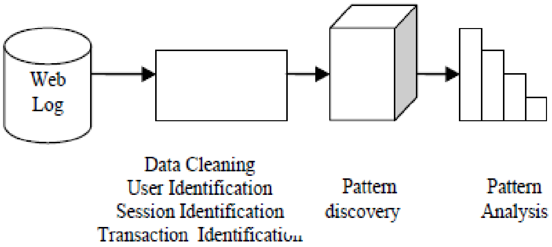

Figure1. Phases of web usage mining

A. Data Collection

Data Collection is the first step in web usage mining process. It consists of gathering the relevant web data. Data source can be collected at the server-side, client-side, proxy servers, or obtain from an organization's database, which contains business data or consolidated Web data.

Server level collection collects client requests and stored in the server as web logs. Web server logs are plain text that is independent from server platform. Most of the web servers follow common log format as " ip address username password date/timestamp url version status-code bytes-sent". Some servers follow extended log format along with referrer and user agent.

Client Side Collection is advantageous than server side since it overcomes both the caching and session identification problems. Browsers are modified to record the browsing behaviors. Remote agents like Java Applets are used to collect user browsing information. Java applets may generate some additional overhead especially when they are loaded for the first time. But users are to be convinced to use modified browser.

Proxy level collection is the data collected from intermediate server between browsers and web servers. Proxy caching is used to reduce the loading time of a Web page experienced by users as well as the network traffic load at the server and client sides. Access log from proxy servers are of same format as web server log and it records the web page request and response for the server. Proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers.

B. Data Preprocessing

The information available in the web is heterogeneous and unstructured. Therefore, the preprocessing phase is a prerequisite for discovering patterns. The goal of preprocessing is to transform the raw click stream data into a set of user profiles.

1. Data Cleaning

Data Cleaning is a process of removing irrelevant items such as jpeg, gif files or sound files and references due to spider navigations. Improved data quality improves the analysis on it. The Http protocol requires a separate connection for every request from the web server. If a user request to view a particular page along with server log entries graphics and scripts are download in addition to the HTML file.

2. User Identification

User Identification of individual users who access a web site is an important step in web usage mining. Various methods are to be followed for identification of users. The simplest method is to assign different user id to different IP address. But in Proxy servers many users are sharing the same address and same user uses many browsers. An Extended Log Format overcomes this problem by referrer information, and a user agent. 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 user, then the user is identified as a new user in the same address. Caching problem can be rectified by assigning a short expiration time to HTML pages enforcing the browser to retrieve every page from the server.

3. Session Identification

A user session can be defined as a set of pages visited by the same user within the duration of one particular visit to a web-site. A user may have a single or multiple sessions during a period. Once a user was identified, the click stream of each user is portioned into logical clusters. The method of portioning into sessions is called as Sessionization or Session Reconstruction. There are three methods in session reconstruction. Two methods depend on time and one on navigation in web topology.

1. Time Oriented Heuristics: The simplest methods are time oriented in which one method based on total session time and the 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 to 24 hours while 30 minutes is the default timeout by R.Cooley. The second method depends on page stay time which is calculated with the difference between two timestamps. If it exceeds 10 minutes then 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, content size of web pages are not considered.

2. Navigation-Oriented Heuristics: uses web topology in graph format. It considers webpage connectivity, however it is not necessary to have hyperlink between two consecutive page requests. If a web page is not connected with previously visited page in a session, then it is considered as a different session. Cooley proposed a referrer based heuristics on the basis of navigation in which referrer URL of a page should exists in the same session. If no referrer is found then it is a first page of a new session.

4. 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.

a. ***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 a 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 a auxiliary reference.

b. 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.

c. ***Time Window:*** A time window transaction is framed from triplets of ip address, 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.

C. Pattern Discovery and Analysis

Once user transactions have been identified, a variety of data mining techniques are performed for pattern discovery in web usage mining. These methods represent the approaches that often appear in the data mining literature such as discovery of association rules and sequential patterns and clustering and classification etc.

4.2. Clustering Techniques

1.Density –Based Algorithm: It start by searching for core object , and they are growing the clusters based on these cores and by looking for objects which are in a neighborhood within a radius[2] the advantage of these types of algorithms is that they can identify arbitrary form of clusters and it can filter out the noise. DBSCAN and OPTICS are density based algorithms.

2. Grid Based Algorithm: This algorithm uses a hierarchical grid structure to decompose the object space into finite number of cells. For every cell statistical information is accumulated about the objects and the clustering is achieved on these cells. The advantages of this approach are the fast processing time that is in commonly independent of the number of data objects. Grid-based algorithms are CLIQUE, STING and Wawe Cluster.

3. Model –based algorithm utilize different distribution models for the clusters which should be verified during the clustering algorithm. A model based clustering is MCLUST.

4.Fuzzy algorithm deduce that refusal hard cluster exits on the set of objects ,but one object can also be assigned to more than one cluster .The best known clustering algorithm is FCM.

5. Conclusion

A data preprocessing treatment system for web usage mining has been analyzed and implemented for log data. It has undergone various steps such as data cleaning, user identification, session identification and clustering. Dissimilar from usual implementations records are cleaned effectively by removing robot entries. By considering the byte transfer rate the reference length is computed. Apart from using Maximal Forward Reference (MFR) and Reference Length (RL) algorithm. Time Window (TW) concept is also combined to find content pages. This preprocessing step is used to give a reliable input foe data mining tasks. Web Personalization method and introduce successful clustering techniques.

REFERENCES

[1] By J Vellingiri, S.Chenthur Pandian,” A Survey on Web Usage Mining” , Global Journal of Computer Science and Technology, Volume 11 Issue 4 Version 1.0 March 2011.

[2] Hemanshu Rana, Mayank Patel ,” A Study of Web Log Analysis Using Clustering Techniques”, International Journal of Innovative Research in Computer and Communication Engineering ,Vol. 1, Issue 4, June 2013.

[3] V.Chitraa, Dr. Antony Selvdoss Davamani,” A Survey on Preprocessing Methods for Web Usage Data”, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010.

[4] Tasawar Hussain, Dr. Sohail Asghar, Simon Fong,” A Hierarchical Cluster Based Preprocessing Methodology for Web Usage Mining”.

[5] Vijayashri Losarwar, Dr. Madhuri Joshi,” Data Preprocessing in Web Usage Mining”, International Conference on Artificial Intelligence and Embedded Systems (ICAIES'2012) July 15-16, 2012 Singapore.

[6] M.Rathamani, Dr. P.Sivaprakasam,” Cloud Mining: Web usage mining and user behavior analysis using fuzzy C-means clustering” , IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727 Volume 7, Issue 2 (Nov-Dec. 2012), PP 09-15.

[7] Abhishek Mathur, Trapti Agrawal,” A Survey: Access Patterns Mining Techniques and ACO”, International Journal of Engineering and Advanced Technology (IJEAT) ISSN: 2249 – 8958, Volume-2, Issue-5, June 2013.

[8] B. Madasamy, Dr. J. Jebmalar Tamilselvi “General Web Knowledge Mining Framework”, International Journal on Computer Science and Engineering (IJCSE) ISSN : 0975-3397, Vol. 4 No. 10 Oct 2012,PP 1744-1750.

[9] Dhanasekaran.K, Rajeswari.R,” A Research-oriented Survey and Current Status on Feature Extraction, Ontology Construction towards Natural Language Processing”, IJCSI International Journal of Computer Science Issues, Vol. 9, Issue 3, No 1, May 2012.

[10] Shaily G. Langhnoja, Mehul P. Barot, Darshak B. Mehta,” Web Usage Mining to Discover Visitor Group with Common Behavior Using DBSCAN Clustering Algorithm”, International Journal of Engineering and Innovative Technology (IJEIT) Volume 2, Issue 7, January 2013.

[11] Richa Chourasia, Prof. Preeti Choudhary,” A Survey On Web Log Pre-Processing And Evidence Preservation For Web Mining”, International Journal of Innovative Research in Technology & Science (IJIRTS) , ISSN:2321-1156.

[12] Vishal Gupta, Gurpreet S. Lehal ,” A Survey of Text Mining Techniques and Applications”, Journal Of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009.

[13] B.Uma Maheshswari, Dr.P.Sumathi, ”A New Clustering & Preprocessing for Web Log Mining”, 2014 World Congress on Computing and Communication Technologies.