ISSN:0975-9646

Detecting Users Behavior from Web Access Logs with Automated Log Analyzer Tool

Deepti sahu
M.Tech student
ITM,Gwalior

Shweta Meena
Asst.Prof
ITM, Gwalior

Abstract:**Internet is a vast source of information or we can say that internet work as a mine of information and we try to mine information from these types of mines according to need. As the amount of website pages keeps on growing the web gives the data mineworkers simply the elements for extracting data. Keeping in mind the end goal to coddle this developing need an uncommon term called Web mining was coined.**

When a user interacts with the web, the interaction information of the user store in a special type of repository called web logs. Special type of information like user name, IP addresses Session information etc. and from these logs we can extract information to find out navigational pattern, to extract data which helps in search engine optimization etc.In this paper we present overview of web usage mining and some methods to Detect Users Behavior from Web Logs.

Key words:Web log , Web usage mining ,Web server log files

INTRODUCTION

A revolution has been observed in the way people work on the internet. People are making use of this important tool for disseminating their ideas, conducting business and most important entertaining themselves. Data on the web is rapidly increasing day by day [1].

A. *Web mining*

Web mining one of the types of data mining is used to extract web data from web pages.

Data mining comprises with the structured data form while the web mining comprises with the unstructured data form and semi structured data form . Web mining is classified in three groups i.e. web usage mining, web structure mining and web content mining to extract web data.

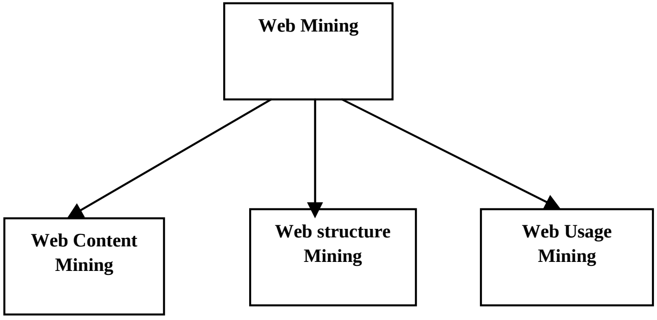

Fig.1 classification of web mining

A.1 .Web content mining

Web content mining means to extract content or Text from the web, It is also known as Text mining or we can say that Content mining is the scanning and mining of images,Text and graphs of a Web page to determine the content relevent to the search query. Text mining is directed toward specific information provided by the customer search information in search engines.

A.2 Web usage mining

Web use mining is the technique to record the activities of the customers while they are skimming and investigating through the Web. The fundamental point of comprehension the route inclination of the guests is to improve the nature of electronic trade administrations (e-business), to customize the Web entrances , to personalize the Web portals or to improve the Web structure and Web server performance. Web Usage mining also categorized in three types depending on the kind of usage data considered into: Application Server Data ,Web Server Data, and Application Level data .

A.3 Web Structure mining

Web structure mining deals with the linking structure of the WebPages and used to fetch information from these linking structures[8].This process use graph theory to analyze the node structure and connection structure of a web site. web structure mining can be classified as two types:

1. Extracting patterns from hyperlinks in the web: a hyperlink is a structural component that connects the web page to a different location.

2. Mining the document structure: analysis of the tree-like structure of page structures to describe HTML or XML tag usage.

B. *Web logs*

The Web Server data is actually the user logs that are generated on the Web Server. Logs enable the analyst to track and analyze the behaviors of user's who visit the website[1]. Web logs are work as a container, contain the user interaction information with the web means to stores the click information by the user in a website. And this information is useful for mining.

Weblogs are the plaintext (ASCII) files contain information about user IP Address, User Name, URL that Referred, Access Request, Time Stamp, error codes etc. and generally reside in the web servers. Server logs are delegated Transfer log, Error Log ,Agent log, and Referrer Log

B.1 Error log

At whatever point a slip is accomplished while the page is, undoubtedly requested by the client to the web server the area is made in the omission log.. First entry in error log for the date and time of message, second entry is for lists the severances of the error being reported.

Level directive is also used for controlling the types of errors that are sent to the error log by restricting

the severity level. The third entry of the provide the information about the IP address of the client that generate the error message. By then next entry is the message itself, which for this circumstance shows that the server has been intended to deny the client access.

B.2 Access Log

The server access log records all requests that are processed by the server. The location and content of the access log are control by Custom Log directive. Custom Log directive is used to put the requests to the server. A log format is specified, and the information of login might be optionally be made condition on solicitation qualities using environment variables Log Format directive can be used to simplify the selection of the contents of the logs.

LOCATION OF WEB LOGS FILES

A Web log is a file to which the Web server writes information each time a user requests a web site from that particular server. A log file can be placed in three different places:[4]

- Web Servers
- Web proxy Servers
- Client browsers

A. ***Web Server Log files***

Logs files that store the activities of the user according to their web access pattern and that resides on the web server are called as web server log files.

B***. Web Proxy Server Log files***

A Proxy server is an intermediate server that reside between the client and the Web server. If the Web server gets a request from the client via the proxy server then the entries to the log file will be the information of the proxy server and not of the original user. These web proxy servers maintain a separate log file for gathering the information of the user.[4]

C. ***Client Browsers Log files***

Log files that can be exist in the client's browser window itself. Special types of softwares can be downloaded by the users to their browser window.

WHY LOG ANALYSIS IS REQUIRED?

For analyzing the data stored in the server logs one need to preprocess it. Sub steps of preprocessing are Cleaning, Session identification and identification of transactions. And for the data cleaning various algorithms have been proposed. Apart from making use of cleaning algorithms and then applying various mining algorithms to the cleansed data, yet another way by which we can extract useful information from this log data is by making use of automated log analysis tools. Web access patterns mined from Web logs are interesting and useful knowledge in practice [5].And the analysis of these pattern is required for finding navigational pattern of the user means the pages are visited frequently by the user ,From which browser is being used by the user who access the website , types of errors users get etc.

STEPS INVOLVE IN WEB USAGE MINING

1. Data collection – In this step Web log documents are gathered which stays informed regarding visits of every last one of guests
2. Data Integration – Coordinate different log records into a solitary document
3. Data preprocessing – cleaning and organizing information to get ready for example or pattern extraction
4. Pattern extraction – Extracting interesting patterns or examples
5. Pattern analysis and visualization – analyze the concentrated example or pattern
6. Pattern applications – Apply the pattern in real world problems

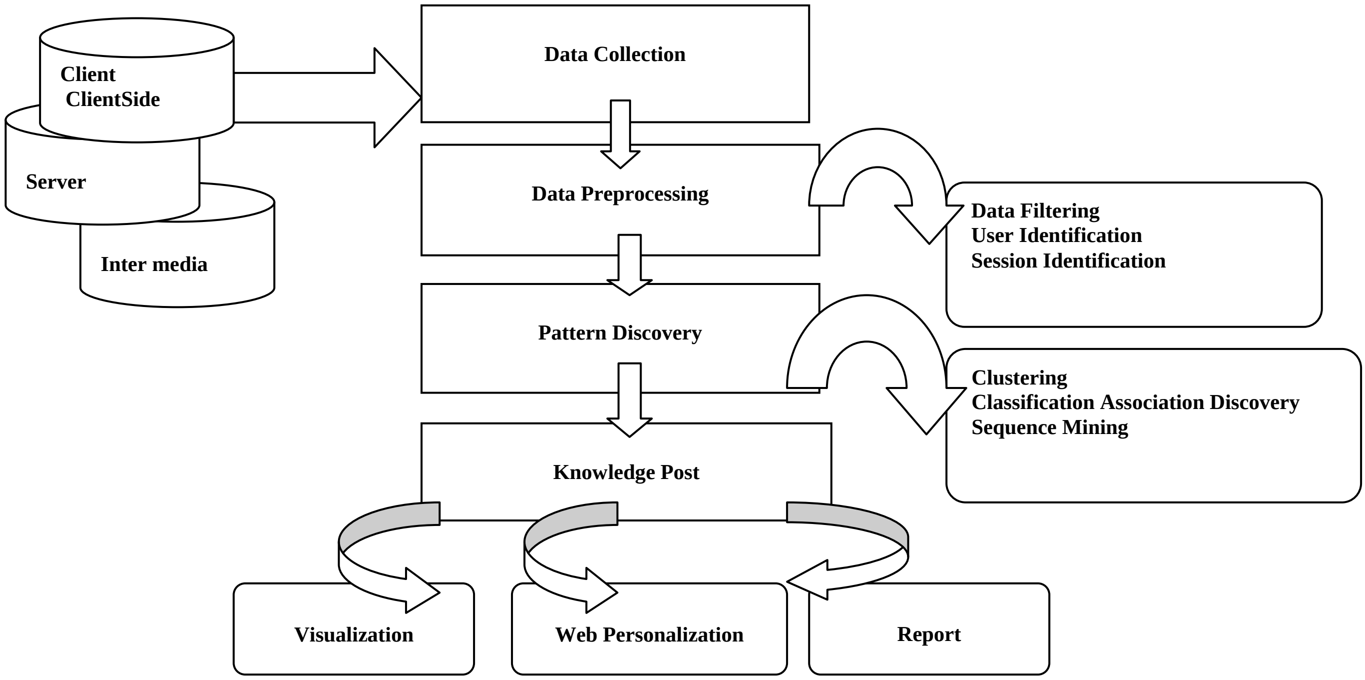

Fig.2 Steps involve in Web mining

RELATED WORKS:

1) Neha goel,c.k. jha [1] Proposed a work in which log analyzer tool called Web Log Expert used for determining the behavior of clients who access a astrology website. It additionally gives a relative study between a couple of log analyzer tools available.

Web log Expert is a freeware and comes in four editions: Enterprise, Professional, Standard and Lite. It is a fast and powerful access log analyzer. The installation is quite easy and GUI provided by the tool is highly user friendly. This tool gives the data about our site's guests: movement detail, accessed files, ways through the site, data about referring pages, browsers, search engines, working frameworks, and then some. It delivers simple to-peruse reports as diagrams and tables.

2) **Ida Mele[1]** gives a methodology to enhancing search engine performance through static caching of search results, and helping users to discover interesting web pages by recommending news articles and blog entries. A query covering approach was used to search the web pages from cache and web logs and searching time, recall and precision was calculated on behalf of that.

3)L. K. J. Grace, V. Maheswari, D.Nagamalai [4] This paper gives a definite dialog about log records, their configurations, their creation, access techniques, their utilization, different calculations utilized and the extra parameters that might be utilized within the log documents which thus offers route to a successful mining. It additionally gives the thought of making an expanded log document and taking in the client conduct.

4)Jian Pei, Jiawei Han, B. Mortazavi-asl & Hua Zhu studied the problem of mining access pattern from web logs efficiently. A data structure called web access pattern tree or WAP tree in short is developed for efficient mining of access pattern from the web logs.The web access pattern tree stores highly compressed , critical information for access pattern mining and facilitates the development of new algorithm for mining access pattern in large dataset of web logs.

5)S.K. Pani, , L. Panigrahy, V.H.Sankar, Bikram Keshari Ratha, A.K.Mandal, S.K.Padhi present a survey paper in which they studied about approaches attempts to extract knowledge from Web, generate some useful result from that knowledge and apply the result to certain real world problems like analyzing uers behavior ,find out navigational pattern etc. and related work on pattern extraction.

ANALYSIS OF PROBLEM

When the Internet time builds, sharing of resource also increases and this prompts create a automated method to rank each one web content asset. Distinctive search engine utilizes diverse strategies to rank search results for the user query. Web content mining enhances the seeking process and gives important data by dispensing with the repetitive and insignificant substance as indicated by client inquiries. Then again, some cases queries may not precisely speak to users' particular information needs since numerous uncertain queries may cover a expansive point and diverse users may need to get data on distinctive perspectives when they submit the same query. For example, when the query like the sun is submitted to a search engine, some users need to search the homepage of a United Kingdom daily paper, while some of the users need to take the natural knowledge of the sun. Therefore, it is essential to capture different user search goals in information retrieval. The induction and investigation of client inquiry objectives can have a ton of points of interest in enhancing web search tool importance and client experience.

PROPOSED METHODOLOGY

Streamlining of search-engine performance is clearly of imperativeness, given that a typical search engine accepts consistently large number of queries, and clients expect low reaction times. The induction and investigation of user search objective can have a lot of advantages in improving search engine relevance and user experience which can be achieved by web usage mining while web content mining and eliminate persistence problem. In this proposed method we concentrates on join together approach of web usage mining and web content mining and weighted technique to mine the web content catering to the user needs. And also work on a new approach is acquainted to rank the important pages focused on the content and keywords instead of keyword and page ranking gave by search engines. Based on the user request, search engine results are discovered.

Proposed methodology is based on the joint approach of web content mining and web usage mining.

This methodology proposed weighted technique to recover the web content catering to the user needs.

(1) User Request- Request of a user is processed by the search engine to obtain the results. Search results are extracted and sent for preprocessing.

(2) Pre- Processing- Pre- Processing is an critical step in content based mining. Real world data tend to be noisy, incomplete and inconsistent. Information pre-processing system can enhance the quality of the data, thereby helping to improve the accuracy and exactness and proficiency mining process. Data preprocessing is an vital step in the knowledge discovery process, since quality choices must be focused around quality information. All user query, keyword and content words are pre processed to remove evacuate words.

(3) Parameters Calculation- Frequency term and occurrence positions are calculated. The calculations depend on the user query.

(4) Calculation of the Page Relevance- After pre-processing the user query is checked with the related words (synonyms). Every result of the keywords and content words are compared by full word matching. If a match is found then a point is awarded to each words based on their position using weighted technique. Finally all matched keywords and contents words are summarized and normalized so that the total must be less than or equal to 1.

At last, the normalized value of each result is sorted in descending order to get the most relevant content for the user query. Then Reordered results are sent back to the user so that the top most page is more relevant for the user query.

REFERENCES:

[1] Neha goel,c.k. jha” Analyzing Users Behavior from Web Access Logs using Automated Log Analyzer Tool” International Journal of Computer Applications (0975 –8887) Vol 62–No.2, January 2013.

[2] Ida Mele” Web Usage Mining for Enhancing Search-Result Delivery and Helping Users to Find Interesting Web Content” ACM 978-1-4503-1869-3/13/02

[3] L. K. J. Grace, V. Maheswari, D.Nagamalai*:* Analysis of Web Logs & Web User In Web Mining, in IJNSA, Vol. 3, No. 1, Jan 2011.

[4] Jian Pei, Jiawei Han, B. Mortazavi-asl & Hua Zhu: Mining Access Patterns Efficiently from Web Logs.

[5] S. K. Pani, , L. Panigrahy, V.H.Sankar, Bikram Keshari Ratha, A.K.Mandal, S.K.PadhiInternational Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011

[6] http://www.google.com/analytics/

[7] AWStats log file analyzer 7.1 Documentation , LogAnalyzerComparison:http://awstats.sourceforge.net/docs/awstats_compare.html

[8] Neeraj Raheja and V.K.Katiyar “Efficient web data extraction using clustering approach in web usage mining ” IJCSI International Journal of Computer Science Issues, Vol. 11, Issue 1, No 2, January 2014