Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 4, Issue, 12, pp. 279-281, December, 2012

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

ISSN: 0975-833X

RESEARCH ARTICLE

DATA MINING TECHNIQUES FOR PREPROCESSING PROCESS IN WEB LOG MINING

[1]Pappu Rajan*, A. and [2]Victor, S. P.

[1]Department of Computer Science and Research Center , St. Xavier's College(Autonomous), Palayamkottai, Tirunelveli , Tamil Nadu, India
[2]Department of Computer Science and Director of Computer Science and Research Center , St. Xavier's College(Autonomous), Palayamkottai, Tirunelveli , Tamil Nadu, India

ARTICLE INFO

Article History:
Received 19th September, 2012
Received in revised form
20th October, 2012
Accepted 19th November, 2012
Published online 18th December, 2012

Key words:
Web mining,
Web log mining,
Data preprocessing.

ABSTRACT

Web applications are increasing at an enormous speed and its users are increasing at exponential speed. The evolutionary changes in technology have made it possible to capture the users' essence and interactions with web applications through web server log file. This is a study on web log mining application Web log files store data related to the use of a web site. Analyzing these data in details is therefore this more complicated task for improving the user browsing experience. Preprocessing being preliminary and essential step to data cleaning, data filtering, path completion, user identification, session identification. In this paper deals with a introductory idea about the web log mining and data preprocessing in the web data.

INTRODUCTION

Web mining is one of the major and important sub division of data mining. Already we have web usage analyzer, site maintainers, pre fetched systems. Most of the data mining techniques are applied on contents, structures and log files of web sites, which mainly used to improve the performance, web personalization etc., Web log analysis is mainly focuses web log file and their structures it is one the key field of web log mining system. So web log analysis is used to record users' browsing information on web servers. All the time user visiting profile information can store into separate the web log file as and when it generate a new log record. What types of information can be stored and what is the usage of it, why we need this type information? In this paper can give to answering these questions. Web log mining deals users' behavior, user name, access time, search web blog information which are stored in separate log, that is called log record. The system mines web log records to discover user access patterns of all web pages.

The analysis of web log records can find regular user, potential customer in the business environment, enhance quality of services and performance of the same in the web server. The web log file consists of three different part. First one is Client Log File, which mainly deals user behavior like authentic information about the users also mainly shows relationship between user and web site. [3][4][6] Second one is Proxy Log File which deals user access data in the log file. This most complicated task to identifying user and their behavior because the user mayn't be using individual login, in that case mostly many to many relationship is possible. In the web log analyzing very most challenging task is identifying or capturing user behavior. Third one is Server Log File. A server log is a log file automatically created and maintained by a server and their activity performed by it. Web server log file mainly keeping information about a history of web request. The common server log file types are Access Log, Agent Log, Error Log, and Referrer Log. General Format of the log files are ASCII text format. Other different types of format also available these are Common Log File Format (NCSA); Extended Log Format (W3C); and IIS Log Format (Microsoft). The main purpose of the different file type are access Log File is mainly maintains all the clicks, hits and accesses of the web site use. User web browser details, operating system details are maintain and keep into the Agent log file. Error log file used to store the errors of the web site. Referrer log file using to maintain information about the referrer.[5][6]

***Corresponding author:** ap_rajan2001@yahoo.com

The Log File Structure

Table 1. Log File Structure

The log file is maintain the following information.

IP address	The computer's IP address
Date	Date and time of the request
Request	The request resources
Status	The Status code of the HTTP
Size	The content length
User agent	The user agent information

Log and user name	Log and user name information

Each and every request the above information's are keeping track and store into the log file.

The Format of the Error Log

Typical format of the error log file are given below:

Wed Oct 11 14:32:52 2000] [error] [client 127.0.0.1] client denied by server configuration: /export/home/live/ap/ htdocs/test

In the above format there are four entries giving information about the format. The first item in the log entry is the date and time of the message. The second entry listing all errors info being reported when system displays errors. The Log Level directive is used to control the types of errors that are sent to the error log by restricting the severity level. The third entry gives the IP address of the client that generated the error. Beyond that is the message itself, which in this case indicates that the server has been configured to deny the client access. The server reports the file-system path of the requested document. [7]

Data Preprocessing

Web Usage Mining (WUM) is one of the application of data mining techniques to discover usage patterns from global Web data. WUM can be having data preprocessing, pattern discovery and pattern analysis.[1] In the preprocessing phase raw Web logs needed to be cleaned, analyzed and then converted in to pattern mining process. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The data is recorded in the server logs system, such as the user IP address browser, viewing time, etc, These are available to identify users and sessions. Some time server logs analyze is not accurate and reliable. So the system also support cookies and sessions. The server logs authentication must be formalized with standard set of format and it should be updated to capturing user access data. Most of the preprocessing techniques leads less quality so we need to improve the quality of preprocessed data and their algorithms. Some new technique is essential to analyze the log file. The Basic Process of web Log Mining can be concentrate Data Preparation, Data Mining, Pattern Analysis. Fig. 1 explains the data preprocessing system in web log analysis. Using any one of the technique, which can be solving the problems and ultimately data mustcan be converted into knowledge.

In all systems are finding different system for extracting exact data from huge data which can be essential needed to develop better system in data mining. In the service oriented architecture is deployed in a run time execution environment. Here Preprocessing execution log can create a logs, we can be identify the better system to identify the pattern. We described the survey of literature on web log preprocessing. We found that web log system are having these essential steps to get exact pattern. The steps are data cleaning, data filtering, path completion, user identification, session identification, cluster of web session, data visualization. The above preprocessing techniques must be standardize and needed to be updated the technology. Web log mining used to enhance server performance, Improve web site navigation, improve the design of web applications, improve the multidimensional web log analysis, identifying web access association or pattern analysis. In the pattern analysis we have to analyze web cashing, pre fetching, swapping frequently used predefined reports should be included the following different type of reports : summary report of hits, list of top requested URLs, referred , list of common browsers, hit per time, error report.

Data Cleaning

Data cleansing, data cleaning, or data scrubbing is the process of detecting and correcting or removing corrupt or inaccurate records from a record set, table, or database. If we find irrelevant data we use technique to identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this noise data. The process of data cleaning can have data auditing, workflow specification, execution, post processing. Every time access log file gets information about their accessibility which can have minimum of 10 to 20 mega bytes. With out data cleaning it gets large size which may leads system gets very slow also that will affect accuracy of user access pattern. These are the common steps required to cleaning data : 1. Read the data from data base/ log file 2. Identify the irrelevant data / inaccurate data /user agent Modify the data if required 3. Clean the data using technique

User Identification

User identification is the process of identifying an end-user who is browsing via the Vital Security system. In that system deals authorization, authorization, auditing. Authorization is applying the correct policy to the end-user for example Security, Logging, and HTTPS policies etc.,[2] Authorization is deciding whether a user is authorized to browse via the system Auditing is Tracing end-user activity through logs, that is, recording / logging transactions with details for future viewing and analyzing activities performed by the user. There are different method can be used to identify the users:1. Through user's Registration which is applicable to only registered user. 2. Using cookie to identify the repeated user access 3.Track and analyze the IP address which is related with IP range and URL Lists. IP range allows system administrators to set rules according to the source IP ranges of the end-users. URL Lists are System administrators can set different rules, based on the URL of the request. This can allow the administrator to configure the system in such a way that it will perform authentication only for specific URLs or bypass authentication based on the URL. There are different researcher doing their research to identify user's behavior, interest and user's pattern etc.,

DATA FILTERING

A data filter is a group of criteria that segments a subscriber list or data extension. The data filter segmentation is based on subscriber attribute values or measures user create from behavioral data. The challenge of data filtering is find a structured method to filter data from errors and noise, Present the methods of filtering, so that they can be implemented in an arbitrary language and applied to filter general data. Find and present a universal method for evaluation of the performance of one filter or for comparing the performance of one filter to

another. There are different types of data can be used to preprocessing so this step also most challenging one.

Session Identification

All data that needs to be available to the application across different requests within the same session is called session state or session state data. After the user identification, which is to decompose each user's visit sequence over a period of time. The following are the important point to be remember to handle the session. Sending the state information back to the client. With the next request the current state is transmitted to the server again. Keeping the necessary data structures on the server. A session was created. A session was terminated. A session for the received identifier does not exist. The peer's IP address changed within the session. The session identification will be concentrate the following factors:total number of session , time of initialization , time of termination , total no of visit .find the new user to initialize.

Path Completion

Caching is a well known strategy for improving the performance of web based system. The heart of this system is page replacement policy, which selects the pages to be replaced in a cache when a request arrives. Any how each page request has been stored in the cache. Cache mechanism some time lost the data because when user click back button, hyperlink between the current request page and a next page can be store in the cache. So Implement a back operation, the page should be added. Here path identification of each request, length of the path are identified and analyzed to improve the system.

Transaction Identification

This is mainly used to group user session and the reference length, maximum length are important factors for transaction identification. The forward navigation pattern and backward navigation also suitable for this system. Each and every transaction can be divided into small one, then it will be identified uniquely.

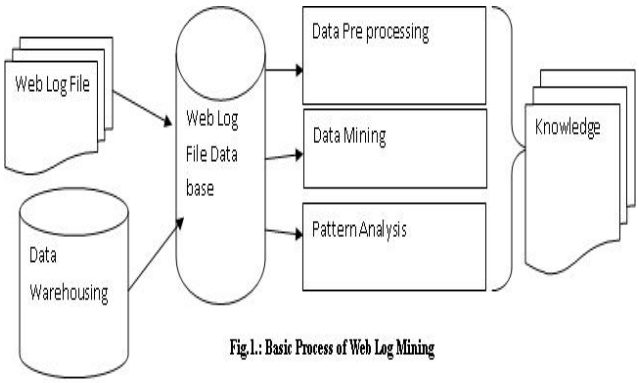

Fig.1.: Basic Process of Web Log Mining

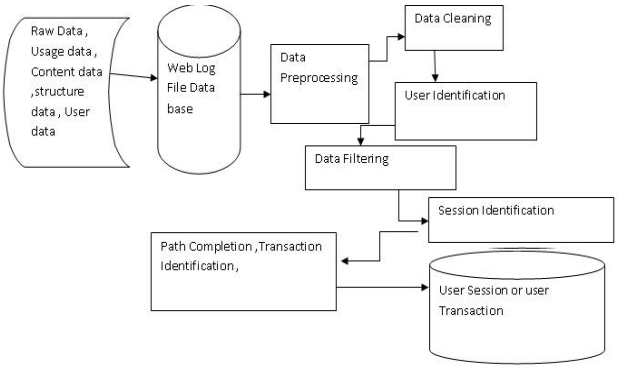

Fig.2. The Process of Data Preprocessing in Web log Mining

Conclusions

Web log mining is a relatively new research area, which has a broad development and application. However, still has many problems to be solved and in depth studied. We under working how to improve the efficiency of data preprocessing algorithm with solution is the next step to be taken into a account. We can do research in Some new techniques can provide the user with the opportunity to analyze the log file at different level of abstraction such as data cleaning, user's behavior, user sessions.

REFERENCES

[1] Domingues, M. Aurlio Jorge, Carlos, Leal. J. Paul. Machado, and Prdro, "A data warehouse for web intelligence, "in Proceeding of the workshop for web intelligence," in the 13th Portuguese Conference on Arti cial Intelligence EPIA by Springer, 2007

[2] Suneetha, K. R. and D. R. Krishnamoorthi (2009). "Identifying User Behavior by Analyzing Web Server Access Log File." IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

[3] Tanasa, D., & Trousse, B. (2004). Advanced Data Preprocessing for Intersites Web Usage Mining. IEEE Intelligent Systems 19(2) 59-65 ISSN 1094-7167

[4] X. Zhang, J. Edwards, J. Harding (2007) Personalised online sales using web usage data mining, Computers in Industry 58: 772–782.

[5] Areerat Songwattana, "Mining Web logs for Prediction in Prefetching and Caching," IEEE International Conference on Convergence and Hybrid Information Technology, 2008, pp. 1006-1011.

[6] A.Pappu Rajan, S.P.Victor," Features and Challenges of web mining systems in emerging technology, "International Journal of Current Research, Vol.4, Issue, 07,pp.066-070, July, 2012, ISSN : 0975-833X

[7] The Apache Software Foundation,"Log Files," http://httpd.apache.org/docs/1.3/logs.html"

****