International Journal of Latest Trends in Engineering and Technology
Vol.(7)Issue(3), pp. 211-216
DOI: http://dx.doi.org/10.21172/1.73.030
e-ISSN:2278-621X

Performance Improvement of Web Server through Log files cleaning

Akshay Gupta[1], Prof. M.A.Rizvi[2]

Abstract- Mining the web data is one of the most challenging tasks in the area of data mining and management for research scholars because there is huge heterogeneous, less structured data available on the web. Web server's logs represent actual usage. Such data have been used for usage-based testing and quality assurance and also for understanding user behavior and guiding user interface design. By analyzing these logs, Web workload was characterized and used to suggest performance enhancements for Web servers. Data preparation techniques and algorithms can be used to process the raw Web server logs, and then mining can be performed to discover users' visitation patterns for further usability analysis. In this research proposal it is proposed to extract actual user behavior from Web server logs, capture user behavior and also proposed a method which can filter out plenty of irrelevant log values based on the complex prefix of their uniform resource locator. This work will improve data quality of web logs by further filtering out more URL requests comparing with traditional data cleaning methods and analyze the results.

Keywords – Usage mining, FP-Tree, K-Means, Data Cleaning, Data Preparation& Pre-processing, Pattern discovery.

I. INTRODUCTION

Web has recently become a powerful platform for not only retrieving information but also discovering knowledge from web data. The concept of discovering useful pattern on the data has been given a verity of names like data mining, knowledge extraction, information discovery and data pattern processing. Web server maintains web log files. Log file are located in different location like web server, web proxy server and client browser.

Web server log file is a simple plain text file which record information about each user. Log file contain information about user name, IP address, date, time, bytes transferred, access request. A Web log is a file to which the Web server writes information each time a user requests a resource from that particular site. When user submit request to a web server that activity are recorded in web log file. When user submit request to a web server that activity are recorded in web log file. Log file used for debugging purpose. Analyzing log file are used to detecting attacks on web.

There are some issues with server log data, including unique user identification and caching. Typically, each unique IP address in a server log may represent one or more unique users. Pages loaded from

[1] *Computer Engineering &Application National Institute of Technical Teachers' Training and Research, Bhopal (MP), INDIA*
[2] *Computer Engineering &Application National Institute of Technical Teachers' Training and Research, Bhopal (MP), INDIA*

client- or proxy-side cache will not be recorded in the server log. These issues make server log data sometimes incomplete or inaccurate. To alleviate these problems, we identified unique users through a combination of IP addresses, user agents, and referrers and used site topology and referrer information along with temporal information to infer missing references. Our method captures user interactions at the site and page levels and thus cannot reveal user experiences at the page element level.

II. PROPOSED ALGORITHM

A. *Extraction of Log Files-*

Server side logs and client side logs are used for Web usage and usability analysis.

Server-side logs can be automatically generated by Web servers, with each entry corresponding to a user request. By analyzing these logs, Web workload was characterized and used to suggest performance enhancements for Internet Web servers. Because of the vastly uneven Web traffic, massive user population, and diverse usage environment, coverage-based testing is insufficient to ensure the quality of Web applications.

Therefore, server-side logs have been used to construct Web usage models for usage-based Web testing or to automatically generate test cases accordingly to improve test efficiency.

Server logs have also been used by organizations to learn about the usability of their products. For example, search queries can be extracted from server logs to discover user information needs for usability task analysis.

Logs can provide insight into real users performing actual tasks in natural working conditions versus in an artificial setting of a lab. Logs also represent the activities of many users over a long period of time versus the small sample of users in a short time span in typical lab testing.

Data preparation techniques and algorithms can be used to process the raw Web server logs, and then mining can be performed to discover users' visitation patterns for further usability analysis.

Organizations can mine server-side logs to predict users' behavior and context to satisfy users' need. Users' revisitiation patterns can be discovered by mining server logs to develop guidelines for browser history mechanism that can be used to reduce users' cognitive and physical effort.

Steps for Extraction of Log Files:

1. *Data Preparation and Pre-processing:* In this, we investigate log files for better use. We investigate the log files to prepare useful log data for preprocessing.

2. *Data cleaning:* In this, removing extraneous references to style files, graphics, or sound files that may not be important for the purpose of analysis.

3. *User identification:* In this, IP, user agent, and referrer fields to identify unique users

4. *User session identification.*

5. *Path completion*: In this, missing references can often be heuristically inferred from the knowledge of site topology and referrer information, along with temporal information from server logs.

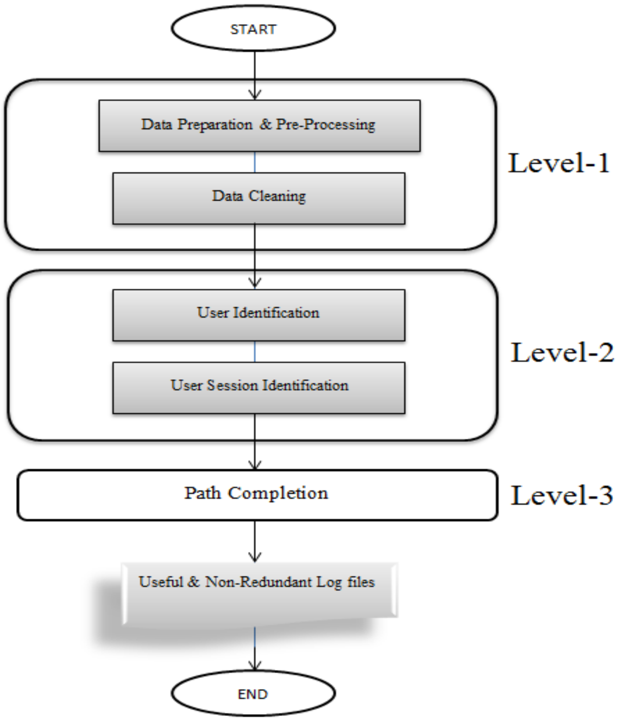
START
Data Preparation & Pre-Processing
Data Cleaning
Level-1
User Identification
User Session Identification
Level-2
Path Completion
Level-3
Useful & Non-Redundant Log files
END

Fig.1 Steps for Extraction of Log Files

B. Algorithm –

The proposed work extract actual user behavior from Web server logs, capture anticipated user behavior with the help of cognitive user models, and perform a comparison between the two. This deviation analysis would help identify some navigation related usability problems. Correcting these problems would lead to better functional convenience as characterized by both better effectiveness (higher task completion rate) and efficiency (less time for given tasks).

This focuses on identifying navigation related problems as characterized by an inability to complete certain tasks or excessive time to complete them. The proposed method identify navigation related usability problems by comparing Web usage patterns extracted from server logs against anticipated usage represented in some cognitive user models. Usability engineers often use server logs to analyze users' behavior and understand how users perform specific tasks to improve their experience.

Web server logs are main data source. Each entry in a log contains the IP address of the originating host, the timestamp, the requested Web page, the referrer, the user agent and other data. Typically, the raw data need to be preprocessed and converted into user sessions and transactions to extract usage patterns.
For implementation of proposed system we will use Java programming language. For web server logs we will use Apache web server.

i. Algorithm for Data Preparation and Pre-processing & Data cleaning

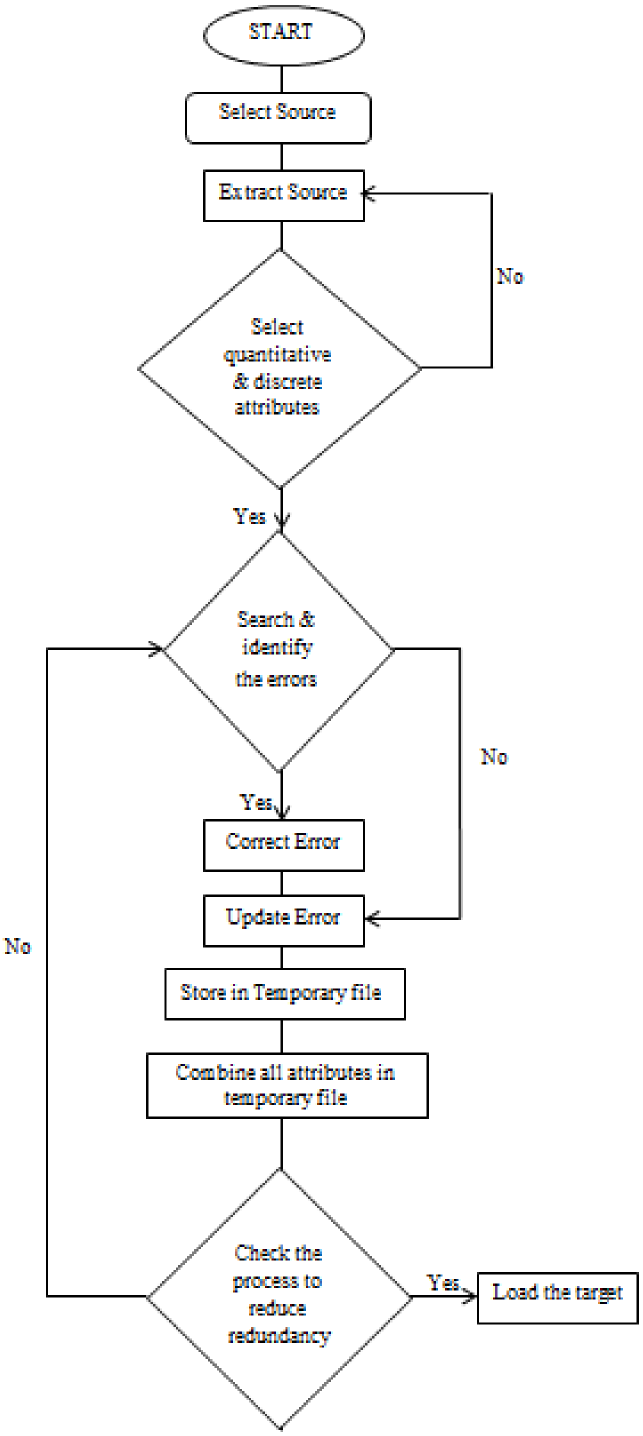
START
Select Source
Extract Source
No
Select quantitative & discrete attributes
Yes
Search & identify the errors
No
Yes
Correct Error
Update Error
No
Store in Temporary file
Combine all attributes in temporary file
Check the process to reduce redundancy
Yes
Load the target

Fig.2 Algorithm for Data Preparation and Pre-processing & Data cleaning which is shown in Fig. 1 of Level1.

ii. *Algorithm to find the usage mining and pattern discovery in web sites.*

Through this algorithm we analyze users' behavior by mining the database and understand how users perform specific tasks to improve their experience through the pattern matching.

Firstly we have to create FP-Tree

```
Scan the database files
Collect F the set of frequent item set and support of each frequent item set
Create the root of FP-Tree and label it as NULL
For
    each transaction in database
    Select the frequent item set in trans and sort them according to the FList
    Insert frequent item set into tree
End for
End
```

After creating the FP tree it is possible to discover the pattern through this algorithm.

Assumption: tree ->FP-Tree, a-> node

```
If tree contains a single path then
    Let P be the single prefix path part of tree
    Let Q be the multipath part with the top branching  node replaced by NULL root
    For
        each combination B of nodes in the path P do
        Generate pattern B union a with minimum     support of nodes in B
        Let freq_pattern_set(P) be set of pattern so generated
    End For
Else
    Let Q be tree
    For each item in ai in Q do
    Generate pattern B union ai with minimum support=ai
Construct B's conditional pattern base and then B's conditional FP-Tree Tree B
If Tree B not NULL then
    Call FP-growth(Tree B,B)
Let freq_pattern_set(Q) be set of pattern so generated
Return freq_pattern_set(P) U freq_pattern_set(Q) U freq_pattern_set(P) x freq_pattern_set(Q)
End
```

IV.CONCLUSION

We conclude the description about area of research about actual usage patterns that can be extracted from Web server logs routinely recorded for operational websites by first processing the log data to identify users, user sessions, and user task-oriented transactions and then applying a usage mining algorithm to discover patterns among actual usage paths. By applying this method on web mining user will get the content efficiently which reduces the size of log files & also get the certain pattern of searching. Hashing techniques is also to be used along with K-means searching algorithm which makes the server response time faster.

REFERENCES

[1] Ruili Geng, Member, IEEE, and Jeff Tian, Member, IEEE "Improving Web Navigation Usability by Comparing Actual and Anticipated Usage" IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 45, NO. 1, FEBRUARY 2015, pp-84-95.

[2] C. Kallepalli and J. Tian, “Measuring and modeling usage and reliability for statistical Web testing” IEEE Trans. Softw. Engin., vol. 27, no. 11, pp. 1023–1036, Nov. 2001.

[3] Hongzhou Sha, Tingwen Liu, Peng Qin, Yong Sun, Qingyun Liu, “EPLogCleaner: Improving Data Quality of Enterprise Proxy Logs for Efficient Web Usage Mining” Elsevier International Conference on Information Technology and Quantitative Management, 2013, pp- 812-819.

[4] D. Tanasa, B. Trousse, “Advanced Data Preprocessing for Intersites Web Usage Mining” IEEE Intelligent Systems 19 (2) (2004) 59–65.

[5] Y. Zhang, L. Dai, Z. Zhou, “A New Perspective of Web Usage Mining: Using Enterprise Proxy Log” in: Proceedings of the 2010 International Conference on Web Information Systems and Mining (WISM), Vol. 1, IEEE, 2010, pp. 38–42.

[6] N. Tyagi, A. Solanki, S. Tyagi, “An Algorithmic Approach to Data Preprocessing in Web Usage Mining” International Journal of Information Technology and Knowledge Management 2 (2) (2010) 279–283.

[7] G. Castellano, A. Fanelli, M. Torsello, “LODAP: A Log Data Preprocessor for Mining Web Browsing Patterns” in: Proceedings of the 6th Conference on 6th WSEAS Int. Conf. on Artificial Intelligence, Knowledge Engineering and Data Bases, 2007, pp. 12–17.

[8] Tec-Ed, “Assessing web site usability from server log files,” White Paper, Tec-Ed, 1999.