International Journal of Engineering Research & Management Technology
ISSN: 2348-4039
November-2017 Volume 4, Issue-6
Email: editor@ijermt.org www.ijermt.org

AN INTEGRATED METHOD TO IMPROVE WEB USABILITYAND WEB SERVER PERFORMANCE

ABHISHEK LAVALE, RAVI HEMRAJ
Department Computer Science & Engineering
Shri Balaji Institute of Technology & Managment
Betul, RGPV University M.P. India

ABSTRACT-

In the real World Wide Web, web sites need navigation. It provides users an idea of the scope of the web site, it is a great fallback system, it decreases users' intellectual overhead. Though, noteworthy challenges occur, comprising correctness of problem identification because of false alarms collective in skilled evaluation, impracticable assessment of usability because of dissimilarities between the testing and actual usage environment, amplified cost because of the extended maintenance cycles and evolution distinctive for many Web based applications Log file comprise information about date, user name, time, IP address, access request and bytes transferred. Server based logs have similarly been used by most organizations to acquire knowledge about the usability of their valuable products. For instance, search based queries can be mined from server based logs to determine user info desires for usability task investigation. Server based Logs can also provide awareness into actual users executing real tasks in usual working circumstances versus in a simulated situation of a research lab. Usability is defined as the efficiency, satisfaction, and usefulness through which actual users can comprehensive specific tasks in a specific environment. The proposed method improves the web usability and performance of the web server. The experimental result represented that our method improves the overall user satisfaction and also help developer to develop better web site design.

Keywords-Data mining, Web Server Logs, Usage Mining, Proxy Server, Web Navigation

INTRODUCTION

Web [1] can also be used for find out information from Internet based data. The concept of discovering useful pattern on the data has been given a verity of name like knowledge extraction, data mining, information discovery and data pattern processing. Mining the web data [2] is the furthermost inspiring tasks for the data management and data mining scholars because there are less structured, huge heterogeneous data available on the web. Usability [3] is defined as the efficiency, satisfaction, and usefulness through which actual users can comprehensive specific tasks in a specific environment. Web design principles [4] are firmness, structural functional convenience, and presentational pleasure. Structural firmness relates primarily to the features that encouragement the web site security and performance. Functional convenience mentions the accessibility of suitable characteristics, such as a web site’s ease of navigation, and ease of use and that help users’ communication with the GUI interface. Presentational delight talks about the website features that motivate users’ wits. Client side logs and Server side based logs are used for Web usage [5] and usability analysis. Server side based logs can be automatically produced by Web based servers like Apache web server, with each entrance matching to a user demand. By analyzing these server logs, Web capability was considered and used to recommend performance heightening for Internet Web based servers. Server based logs have similarly been used by most organizations to acquire knowledge about the usability of their valuable products. For instance, search based queries can be mined from server based logs to determine user info desires for usability task investigation. Server based Logs can also provide awareness into actual users executing real tasks in usual working circumstances versus in a simulated situation of a research lab.

As per the World Wide Web turn out to be biggest today, building and ensuring easy-to-use Web based organizations is becoming a necessary proficiency for commercial existence. Due of the immensely un even Web data traffic, enormous user populace, and miscellaneous usage environment, coverage constructed testing is unsatisfactory to guarantee the quality of Web based applications. By means of simple log cleanser filtered data is irrelevant and valuable, some preserve alive link add period stamp into their URL. So their precede cannot be further added openly to prefix library simply by threshold. Besides the design of threshold and estimation method of precision rate is irrelevant simple. There have different data cleaning [6] method of server log is still very enlighten. Some researchers also focus on data cleaning method of proxy log and describe the difference between proxy log and server log. One thing is noticeable that Standard Filter is impossible to filter out any relevant item. Web design values were acknowledged to assistance progress users' online involvement. Experimental estimation by professionals and user-centered analysis are characteristically used to categorize usability matters and to guarantee agreeable usability. Though, noteworthy challenges occur, comprising correctness of problem identification because of false alarms collective in skilled evaluation, impracticable assessment of usability because of dissimilarities between the testing and actual usage environment, amplified cost because of the extended maintenance cycles and evolution distinctive for many Web based applications. Web based log file investigation began as a way for Information. Log[7] file comprise information about date, user name, time, IP address, access request and bytes transferred. A Website log is a text file to which the Web based server marks information to each time a user demands a resource from that certain web site. Paper is organized as follows. Section II provides related work to web mining, navigation and web usability, web server log files. Section III provides proposed algorithm of the web usability mining. Section IV implements the algorithm and provides result analysis. Section V concludes the paper.

RELATED WORK:

Proposed method in[8]prominence on recognizing celestial navigation related difficulties as regarded as by an helplessness to complete definite tasks or unnecessary period to complete them. The advised scheme distinguish website navigation accompanying usability complications by equating Website constructed usage patterns take out from website established server logs against predicted usage embodied in some intellectual user models. Usability engineers often use server logs to analyze users' behavior and appreciate how consumers accomplish definite tasks to expand their experience. The main steps in system are Web Data Preparation for mining and Pre-processing. Next is data cleaning i.e. removing extraneous graphics, references to style links, or audio files that might not be essential for the purpose of analysis? Then next is user identification step in which referrer fields, user agent, and IP address to identify unique users. Next process is user session identification. Path completion or misplaced references can habitually be heuristically contingent from the information of site topology and referrer info, sideways with time-based info from server constructed logs Perfect user collaborating path models detention projected Web based usage. Planned architecture consist of three modules IUIP modeling, Usage Pattern Extraction, and Usability problematic identification. Consumer pattern withdrawal component remove actual direction finding paths from server logs and find out patterns for nearly characteristic actions. In equivalent, IUIP models for the equal events. IUIP models are built on the perception of user behavior and can characterize estimated paths for detailed user-oriented tasks. The outcome examination employs the contrivance of test oracle. A revelation is commonly used to decide whether a test has failed or passed. Here, these models used the revelation to recognize the usability problems associated to the users' genuine navigation paths by analyzing the unconventionalities between the two.

Website server logs are foremost data source. Individually entry in a log encompasses the timestamp, the IP address of the inventing host, the referrer, the requested Website page, the user agent and other related data.

Characteristically, the raw information necessity to be pre-processed and transformed into user transactions and sessions to excerpt usage patterns.
A new technique to recognize navigation- related Web based usability difficulties based on show a relationship between Actual and Expectable usage data patterns. The tangible usage data patterns can be mined from Internet based server logs regularly recorded for functioning websites by first handling the log based data to user sessions, identify users, and consumer task focused transactions, and then smearing web usage mining procedure to determine data patterns amongst actual usage pathways. The expected website page usage, together with info about both the path and time essential for user focused tasks, is taken by perfect user interactive track models created by intellectual experts based on their reasoning of user behavior. The assessment is accomplished through the method of test oracle for examination of results and recognizing user navigation problems. The deviation web data produced from this assessment can help us find out usability related concerns and recommend corrective activities to increase web based usability. A program implemented to automate a noteworthy part of the actions involved. With the help of experimentation on a lesser service oriented web site, system identified web usability difficulties, which were cross validated by field specialists, and enumerated usability development by the lower time and effort and higher task success rate for given tasks after recommended improvements were implemented.

The [9] represent a new technique and tool for activity demonstrating through qualitative successive data item analysis. In specific, address the problem of creating a symbolic abstract illustration of an activity from an action trace. To use information engineering methods to help the analyst construct ontology of the activity, i.e., a set of hierarchical semantics and graphical symbols that supports the construction of action models. The ontology building is evolutionist, pragmatic and driven by the analyst in accordance with their modeling objectives and their research based questions. The scheme helps the analyst to define transformation guidelines to process the raw data trace into abstract traces created on the ontology. The analyst imagines the abstract traces and repeatedly tests the transformation rules, ontology, and the visualization presentation to approve the models of action. With this method, and tool found pioneering ways to represent a motor car driving movement at dissimilar levels of abstraction from activity traces composed from an instrumented means of transportation. As samples, report two new policies of track changing on motor ways that modeled and found with this approach.

[10] Defines an examination to determine the technical possibility of identifying and discovering the several situations experienced by actor or human absolutely from a trace of period stamped data values of variables. More precisely, the objective of investigation was to determine the circumstances that a human actor practiced, while performing a strategic task in a simulator based situation, the arrangement of these circumstances and their time-based interval. The significant variables that were witnessed in the trace were designated apriorithru a human. The conclusion of the procedure was matched with human or manual context utilization of the similar traces. One probable usage of such automated background finding is to assist for building self-directed strategic agents proficient of performing the similar tasks as the actor or human. As such, similarly quantitatively compared the outcomes of with the help of the COPAC-derived circumstances with those achieved with human derived context utilization in constructing self-directed strategic agents.

Modern Web based applications are fully, difficult software systems. Consequently, the development of Web based applications needs a procedurally comprehensive engineering methodology called Web Engineering. It is not very clear, however, to which level present solutions from appropriate areas, most remarkably software engineering can be use again as such for the improvement of Web applications and subsequently, if Web Engineering is really a discipline on its own. [11] Highlights the characteristics of Web based application improvement as found in current literature thus provided that a requirement for analyzing the suitability of present engineering solutions. The characteristics are characterized according to comprising the software product itself, four dimensions, its improvement, its use and advancement as a cross cutting concern.

[12] refer to the dissimilar features between proxy based log and server based log, thus deliver a data cleaning technique for enterprise based proxy. Though the assessment of proxy based log and server based log is considerable, they use plenty of experimental value as the threshold in the experimentation without theoretical or description support. It also creates the experimental outcome doubtful.

A significant aspect that influences the efficiency of security schemes within an association is the web usability of security administration tools. [13] Present a review of design strategies for such procedures. [13] Collected recommendations and guidelines related to IT security administration systems from the literature.

PROPOSED ALGORITHM

Web based server logs text files are main data source. Normally, the value or data prerequisite to be converted and pre-processed into transactions and user sessions also to mine usage patterns.

The proposed steps is described in section below.

Data gathering: The first phase is to bring together the dataset from dissimilar web sites. The web pages of different web sites and log text files encompassing information about name of the user, Internet Protocols address of the system, timestamp and date. The log text file might too comprise some information which is irrelevant to discover may have to be removed from dataset. Some irrelevant data also discarder from the web log text files irrelevant to navigation and usage.

Data Pre-processing and Preparation: The subsequent step is to get ready data for before applying processing procedure for data navigation and usage mining. In this stage the investigation of log text file is done for better applying algorithm and use. This step inspects and prepares the log text files for useful log data and preprocessing.

Data cleaning: The subsequent step is to remove unwanted data from log file. The cleaned data can easily mine suitable web information from log files. In this phase undesirable data is detached from the web based server log text files. This process is termed data cleaning. Eliminating unimportant references to style files, sound files, graphics, and video files, or other resources that might not be essential for the purpose of data analysis.

User identification: The subsequent stage is to categorize the consumer of web site. The consumer information is existing in server log text file. The user info is taken out from the server log text file using FP-Growth procedure. The machine Internet Protocol number is used for classifying consumer. The user illustrative i.e. the application used by consumer can likewise be recognized for web usage mining procedure. The referrer columns are also used to categorize unique users.

User session identification: The subsequent phase in suggested work is to recognize user session. Session represents login, logout and web site visiting activities. In this stage we eradicate the page visited, used login time, login session and data usage of the web user. This specific information is beneficial to discover user page navigation.

Path completion: In subsequent phase misplaced references can every so habitually be heuristically inferred after the information of web site referrer and topology information, alongside with temporal info from server logs text file.

Algorithm

The data cleaning procedure is given below. The input to procedure is log files and output is cleaned log file.

```
Input: Server log file
Output: Cleaned server log file
Start
Choose the input file from the web log server database.
For each input log file i.e. web site log server file
        Excerpt input source file i.e. web site log file
        Selection of attributes i.e. Choose desirable features from input source
        Select discrete and quantitative feature only
        Auditing or identify and search for the error occurrences
        Spot on and correct the log errors
        Bring up to date the correction information in file
        Temporarily store in file
        Remove all unwanted links
Associate all features in temporary text files which will needed for algorithm 2
Remove redundant and unwanted record from log files
End
```

The above algorithm removes all the redundant, irrelevant links from the log file and prepared it for further processing.

After pre-processing phase FP-Growth procedure is applied to find the pattern discovery and usage mining in web sites log files.

```
Create a table of visited links from web log file
Separate the web log file data by visitor:
  Organize the web site log file by guest unique ID as the unique value and date and timestamp as the subsidiary key
  For every single web site visitor, divide web site log such that to each next terminates in a target web page.
For every target page and page visitor, discover any anticipated places for that target web page:
Let {A1, A2 ….. , An} be the set of visited web pages, where An is a target web page.
Let B := @ symbolize the list of back down web pages.
fori := 2 to n- 2 begin
if (($A_{i-1}$ = $A_{i+1}$) or (no link from Ai to $A_{i+1}$))
       Add Ai to C.
   End loop
if (C not empty)
      Add (An, C, $A_{n-1)}$ to (backtrack list, current URL, Actual Location) table;
End
```

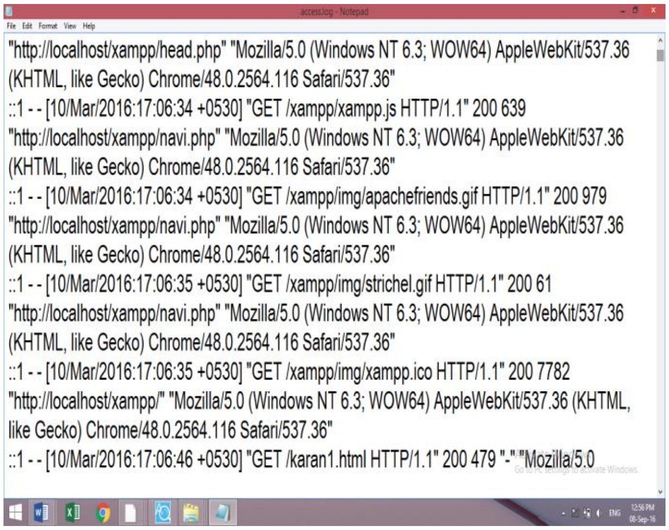
The above algorithm removes all the redundant, irrelevant links from the log file and prepared it for further processing.

After pre-processing phase FP-Growth procedure is applied to find the pattern discovery and usage mining in web sites log files.

IMPLEMENTATION:

Java platform is used for the implementation of the algorithm. The dataset used for implementation of proposed work is the web site and log files from Apache web server. We have used XAMPP server which includes Apache web server, MYSQL relational database, and PHP programs.

The server log file is represented in the figure below. The file consists of user login, logout, and machine and session detail of the user.



```
"http://localhost/xampp/head.php" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/48.0.2564.116 Safari/537.36"
::1 - - [10/Mar/2016:17:06:34 +0530] "GET /xampp/xampp.js HTTP/1.1" 200 639
"http://localhost/xampp/navi.php" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/48.0.2564.116 Safari/537.36"
::1 - - [10/Mar/2016:17:06:34 +0530] "GET /xampp/img/apachefriends.gif HTTP/1.1" 200 979
"http://localhost/xampp/navi.php" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/48.0.2564.116 Safari/537.36"
::1 - - [10/Mar/2016:17:06:35 +0530] "GET /xampp/img/strichel.gif HTTP/1.1" 200 61
"http://localhost/xampp/navi.php" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36
(KHTML, like Gecko) Chrome/48.0.2564.116 Safari/537.36"
::1 - - [10/Mar/2016:17:06:35 +0530] "GET /xampp/img/xampp.ico HTTP/1.1" 200 7782
"http://localhost/xampp/" "Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML,
like Gecko) Chrome/48.0.2564.116 Safari/537.36"
::1 - - [10/Mar/2016:17:06:46 +0530] "GET /karan1.html HTTP/1.1" 200 479 "-" "Mozilla/5.0
```

Figure 1: A server log file from web server

The server log file contents are extracted from the web server and developed program is executed to remove unwanted data for further processing.

Figure below represents the daily uses of the web page.

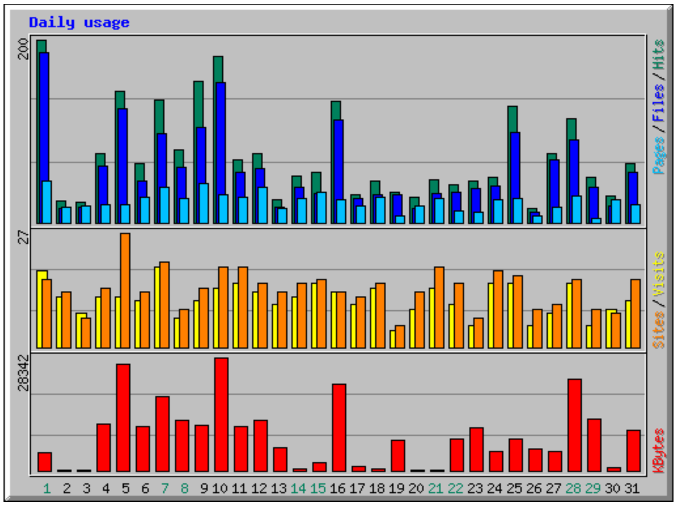


Figure 2: Daily usage for month October 2017

Figure above represents sites/visits, bytes used and page/files/hits of the web site in October month.

The different parameters used to find the usage and web navigation are as follows.

Number of web site hits: It represents the number of web sites used by user. A counter is used to count the number of web site hits. It can be measured in hours, daily, weekly and monthly basis.

Number of data used: The data used by user of a particular web site represent how much data accessed by user.

Number of web page visits: It represents how many web pages of a particular web site is visited by user.

Number of files in web site: It represents number of web file used by user. It can be measured in hours, daily, weekly and monthly basis.

Table below provides the top 7 URL of the web server.

#	Hits		Kbytes		URL
1	59	2.58%	2222	0.80%	/js/jquery.js
2	71	3.11%	1875	0.68%	/css/bootstrap.min.css
3	38	1.66%	1528	0.55%	/fonts/fontawesome-webfont.woff
4	161	7.05%	1481	0.53%	/
5	72	3.15%	780	0.28%	/css/animate.css
6	28	1.23%	743	0.27%	/js/jssor.slider.min.js
7	56	2.45%	700	0.25%	/js/bootstrap.min.js

Table 1: Top 7 of 25 Total URLs by Kbytes

The table above represents number of hits, byte used of a top 7 URL.

CONCLUSION:

Web has in recent era come to be a dominant platform for not only accessing digital information but also determining knowledge from web data. The web data is stored as unstructured as well as structured format. The logs present in Web servers represent actual usage of the web sites. The file called log file comprise information about bytes transferred, Internet Protocol address, name of the user, date, access request time. The real usage values can be mined from different Internet server logs consistently recorded for functioning websites by primary processing the log value to identify, user sessions, users, and user task-oriented transactions. The proposed algorithm improved the performance of the web server. The experimental outcome represented that the proposed algorithm is better for web usability.

REFERENCES

1. R. Cooley, B. Mobasher, and J. Srivastava,, “Data preparation for mining World Wide Web browsing patterns,” Knowl. Inf. Syst., vol. 1, no. 1, pp. 5–32, 1999.
2. M. F. Arlitt and C. L. Williamson, “Internet Web servers: Workload characterization and performance implications,” IEEE/ACMTrans. Netw., vol. 5, no. 5, pp. 631–645, Oct. 1997.
3. C. M. Barnum and S. Dragga, Usability Testing and Research. White Plains, NY, USA: Longman, Oct. 2001.
4. M. C. Burton and J. B. Walther, “The value of Web log data in use-based design and testing,” J. Computer.-Mediated Commun., vol. 6, no. 3, p. 0, 2001.
5. T. Carta, F. Patern`o, and V. F. D. Santana, “Web usability probe: A tool for supporting remote usability evaluation of web sites,” in Human-Computer Interaction—INTERACT 2011. New York, NY, USA: Springer, 2011, pp. 349–357.
6. R. Cooley, B. Mobasher, and J. Srivastava,, “Data preparation for mining World Wide Web browsing patterns,” Knowl. Inf. Syst., vol. 1, no. 1, pp. 5–32, 1999.
7. 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
8. RuiliGeng, 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
9. Pandey A.,Bansal K.K.(2014): “Performance Evaluation of TORA Protocol Using Random Waypoint Mobility Model” *International Journal of Education and Science Research Review* Vol.1(2)
10. Tiwari S.P.,Kumar S.,Bansal K.K.(2014): “A Survey of Metaheuristic Algorithms for Travelling Salesman Problem ” *International Journal Of Engineering Research & Management Technology* Vol.1(5)
11. Olivier L. Georgeon, Alain Mille, Thierry Bellet, Supporting Activity Modeling from Activity Traces, 2013, pp- 201-237
12. Viet C. Trinh and Avelino J. Gonzalez, “Discovering Contexts from Observed Human Performance”, IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS, VOL. 43, NO. 4, JULY 2013, pp-359-471
13. GertiKappel, Elke Michlmayr, Birgit Pröll, Siegfried Reich4, Werner Retschitzegger5 Web Engineering - Old wine in new bottles?, White paper, 2—3, pp-25-31
14. 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.
15. PooyaJaferian, David Botta, Fahimeh Raja, Kirstie Hawkey, Konstantin Beznosov, Guidelines for Designing IT Security Management Tools, pp 221-230