IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 4 Issue 1, January 2017
ISSN (Online) 2348 – 7968 | Impact Factor (2015) - 4.332
www.ijiset.com

Intelligent Web Server Navigation Usability and Web Usage Data Mining for Improving Users Satisfaction

Amit Kumar Mishra
Assistant Professor , Department of Computer Science & Engineering , W.I.T ,Darbhanga , Bihar

Abstract- By analyzing web server logs, Web workload was characterized and used to suggest performance enhancements for Internet Web servers. 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. Our propose work will improve data quality of web logs by further filtering out more URL requests comparing with traditional data cleaning methods.We propose a method which can filter out plenty of irrelevant log values based on the complex prefix of their uniform resource locator. We have collected dataset from different web sites of web server. Our proposed method improved the approach to discover web usage pattern in an application and web server usability system. Such analyses might be more meaningful and yield more interesting results for Web applications with complex structure and operation sequences.It also improved the web user's overall satisfaction. Our proposed work also improved the performance of web server. Experimental results analyzed at the trail level instead of at the individual page level.

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

I. INTRODUCTION

Web has recently become a powerful platform for retrieving useful information. It can also be used for discovering knowledge from web based data.As the World Wide Web WWW[1] becomes predominant today, ensuring and building easy-to-use Web based systems is becoming an essential competency for business survival. Because of the vastly un even Web traffic, massive user population, and diverse usage environment, coverage-based testing is insufficient to ensure the quality of Web applications. With the help of simple log cleaner filtered data is valuable and irrelevant, some keep alive link add time stamp into their URL.Mining the web data[2] is one of the most challenging tasks for the data mining and data management scholars because there are huge heterogeneous, less structured data available on the web. Web based 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 analysing these logs[3], Web workload was characterized and used to suggest performance enhancements for Internet Web servers. Data preparation[4] methods and procedures can be used to process the raw Web based server logs. The next is then mining can be achieved to determine users' visitation patterns for additional usability analysis.

Data preparation procedures and systems can be extensively used to process the raw data from Web based server logs, and then data mining can be achieved to determine users' web site visitation patterns[5] for additional usability investigation.Large organizations can mine server side based logs to forecast users' activities and context to fulfil users' need. Users' re-visitiation of web page patterns can also be discovered by mining server based logs to develop strategies for web browser history mechanism that can also be used to decrease users' physical and cognitive effort. Log data kept at Web servers represent actual usage. Such data have been used for usage-based testing and quality assurance and also for understanding user behaviour and guiding graphical user interface design. The real usage data can be mined from Web server based logs regularly recorded for many operational web sites by first processing the log based data to recognize user sessions, users, and user based task oriented transactions, and then applying a usage data mining procedure to determine patterns among actual usage paths. The objective of the research is as follows. To implement the web server usability technique with improved system with large web application.To improve the approach to discover web usage pattern in an application and web server usability system.To improve the web user's overall satisfaction.Usability is well-defined as the satisfaction, efficiency, and effectiveness with which particular users can complete precise tasks in a precise environment. Web design principles[6] are structural functional convenience, firmness, and presentational pleasure. Structural firmness relates mainly to the features that inspiration the web site performance and security.

Large organizations can mine server side based logs to forecast users' activities and context to fulfil users' need. Users' re-visitiation[7] of web page patterns can also be discovered by mining server based logs to develop strategies for web browser history mechanism that can also be used to decrease users' physical and cognitive effort. Log data kept at Web servers represent actual usage. Such data have been used for usage-based testing and quality assurance and also for

understanding user behaviour and guiding graphical user interface design. The real usage data can be mined from Web server based logs regularly recorded for many operational web sites by first processing the log based data to recognize user sessions, users, and user based task oriented transactions, and then applying a usage data mining procedure to determine patterns among actual usage paths. Web systems is becoming a core competency for business survival. Web design principles were identified to help improve users' online experience. Heuristic evaluation by experts and user-centered testing are typically used to identify usability issues and to ensure satisfactory usability. However, significant challenges exist, including accuracy of problem identification due to false alarms common in expert evaluation, unrealistic evaluation of usability due to differences between the testing environment and the actual usagev environment, increased cost due to the prolonged evolution and maintenance cycles typical for many Web applications.Paper is organized as follows. Section II provides literature survey of web mining and usage mining. Section III provides the proposed architecture and algorithm. Section IV provides implementation of the proposed work. Section V concludes the paper.

IILITERATURE SURVEY

Proposed method in[8]emphasis on identifying navigation related problems as characterized by an inability to complete certain tasks or excessive time to complete them. The suggested scheme recognize web navigation associated usability difficulties by equating Web based usage patterns extracted from web based server logs against predicted usage embodied in some intellectual user models. Path completion[9] can missing references can often be heuristically inferred from the knowledge of site topology and referrer information, along with temporal information from server based logs Ideal user interactive path (IUIP) models capture anticipated Web based usage. Proposed architecture includes three modules Usage Pattern Extraction, IUIP modelling and Usability problem identification. User pattern extraction module extract actual navigation paths from server logs and discover patterns for some typical events. In[10] recommend the first technique named EPLogCleaner that can clean out adequately unrelated data items based on the common preface of their URLs. Make an assessment of EPLogCleaner with a actual network traffic trace captured from one enterprise sed proxy. Experimental outcomes illustrate that EPLogCleaner can increase data quality of enterprise based proxy logs by further cleaning out additional than 40% URL requests relating with traditional data cleaning procedures.

[11] provides an algorithmic methodology to data item pre processing in web based usage mining. It take requirements from graphical web page content, or some other file which may be induced into web based page, or navigation duration accomplished by web spiders and robots into consideration. Nevertheless[12] conversed the significance of proxy based log, the item or data cleaning procedures used is relatively easy.

[13] discussed the classification of useless data which should be cleaned away, and divided them into two categories. One is for requests without analyzed resources such as images and multimedia files. The other is for requests generated by web robot. However, some irrelevant information such as requests with wrong status code can pass through their method and bring unnecessary calculation in the upcoming tasks.

[14] develop a tool named LODAP for the pre-processing of web log file. They provide three sub steps of pre-processing, including data cleaning, data structuring and data filtering. Data cleaning was performed based on access method, status code, multimedia objects, and request generated through robots. However, most of software updates and requests from network behaviouranalyser may survive after the sub step of data cleaning. So their method is inefficient and incomplete.

Statistical testing[15] and reliability analysis can be used effectively to assure quality for Web applications. The usage information is used to build models for statistical Web testing. The related failure information is used to measure the reliability of Web applications and the potential effectiveness of statistical Web testing.

Two types of logs, i.e., server-side logs and client-side logs, are commonly 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 analysing 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 [16]. 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.

III PROPOSED WORK

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.

Proposed Steps

Data collection: In this step we will collect the dataset from different web sites. We will collect log files containing information about Internet Protocols address of the computer, name of the user, time and date. The log file may also contain some information which is not useful for web usage and navigation.

Data Preparation and Pre-processing: The next step is to prepare data for pre-processing for data usage mining. In this step we investigate log files for better use. We investigate the log files to prepare useful log data for preprocessing.

Data cleaning: The next step is to clean the data for useful web information discovery. In this step unwanted data is removed from the web server log files. It is called data cleaning. Removing extraneous references to sound files, style files, video files, and graphics, or that may not be important for the purpose of analysis.

User identification: The next step is to identify the user of web site. The user information is present in server log file. The user information is extracted from the server log file using FP-Growth algorithm. The Internet Protocol address is used for identifying user. The user agent i.e. the software used by user can also be identified for web usage mining. The referrer fields is also used to identify unique users.

User session identification: The next step in proposed work is to identify user session. In this step we extract the used login time, page visited, and data usage and login session of the user. This information is useful to find user navigation.

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

Data cleaning: Data cleaning method is used to clean the data for useful web information discovery. Eliminating extraneous situations to video files, style files, sound files, graphics file that may not be important for the purpose of analysis. We have developed an algorithm for data cleaning. In this step unwanted data is removed from the web server log files. It is called data cleaning.

Algorithm 1: Data cleaning

Select the source file from the database.
For each source i.e. web log server
    Extract source file i.e. web server log file
    Select needed attributes from source (Selection of attributes)
    Select quantitative and discrete attribute only
    Search and identify the error instances (auditing)
    Correct the errors
    Update the correction
    Store in temporary file
Combine all attributes in temporary files needed for targets
Check the process to record redundancy
Load the target for further processing

Using FT-Growth algorithm we find the usage mining and pattern discovery in web sites.

Algorithm 2

Create FP-Tree
Scan the database files
Assemble F i.e. the set of frequent data item set and support of each frequent set
Create the root of Frequent Pattern -Tree and make it as NULL
For each transaction in database
    Select the common items in transaction and category them according to the FList
    Insert frequent item set into tree.
End for
End

FP-Growth algorithm

Tree represent FP-Tree
a represent node
If tree comprises a single route then
    Let P be the only preface route fragment of tree
    Let Q be the multiple path fragment with the topmost separating node substituted by NULL root
    For each combination B of nodes in the path P do
        Generate pattern B union a with smallest support of nodes in B
        Let frequent_patterns_sets(P) be set of values so generated
    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 TreeB
        If TreeB not NULL then
            Call FP-growth(TreeB,B)
            Let frequent_patterns_sets(Q) be set of pattern so generated
            returnfrequent_patterns_sets (P) U frequent_patterns_sets(Q) U

frequent_patterns_sets(P) x
frequent_patterns_sets(Q)
end

Algorithm 3:
Develop a table of links in the web site.
Divide the web log by visitor:
Arrange the web log file by visitor unique ID as the primary key and timestamp as the subordinate key, or
For every visitor, divide web log such that each subsequence dismisses in a target page.
For each visitor and target page, find any expected locations for that target page:
Let {P1, P2 ….. ,Pn} be the set of visited pages, where Pn is a target page.
Let B := @ denote the list of backtrack pages.
fori := 2 to n- 2 begin
if ((P_{i-1} = P_{i+1}) or (no link from Pi to P_{i+1}))
Add Pi to B. // Pi is a backtrack point.
End loop
if (B not empty)
Add (Pn, B, P_{n-1}) to (current URL, backtrack list, Actual Location) table;

IV 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. Windows operating system is considered good in the security point of view. The implementations are carried out in college lab. The server system used is Intel machine having an i3 3.0 GHz processor, 4GB of RAM and a 7,200 RPM 500 GB SATA disk. Every client computer executes the Java environment client prototype of structural design on an Intel PIV machine having a single 2.5 GHz processor, 2 GB of RAM and two 7200 RPM 500 GB SCSI disks. We have used XAMPP server which includes Apache web server, MYSQL relational database, and PHP programs. Apache web server provides web related services to Internet users. Main dataset file used in implementation is web server log file stored in Apache web server.
The different parameters used to find the usage and web navigation are as follows.
Number of web site hits: It represent 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 represent number of web file used by user. It can be measured in hours, daily, weekly and monthly basis.
Figure below represents the software developed for identifying web usability. The web server log file is opened and monitor for analyzing the user behavior.

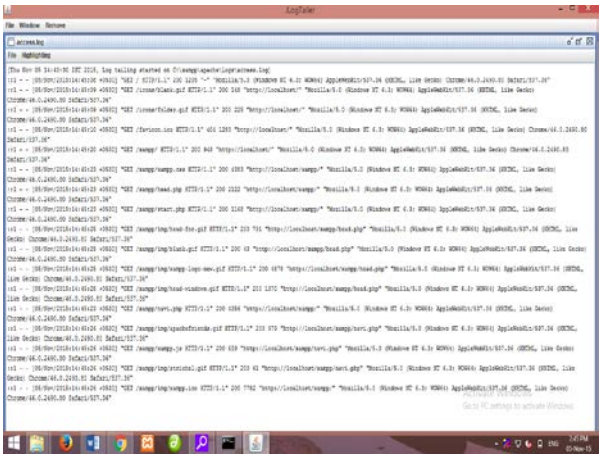

Figure 1 The figure below shows the usage of different user in Kbytes. It provides information of top 10 of 300 total sites by Kbytes.

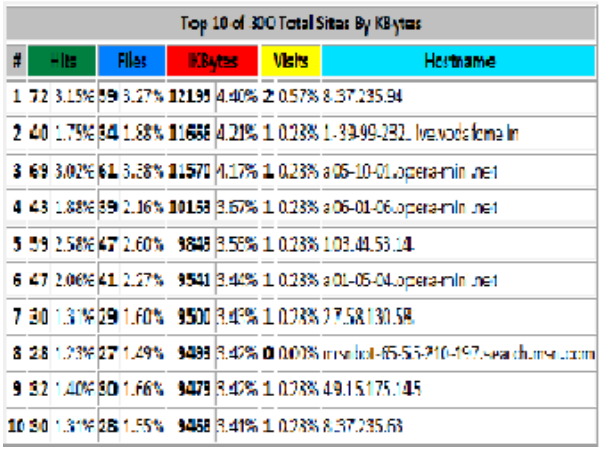

Figure 2 The figure below represent the usage summary of web sites for year 2015-2016.

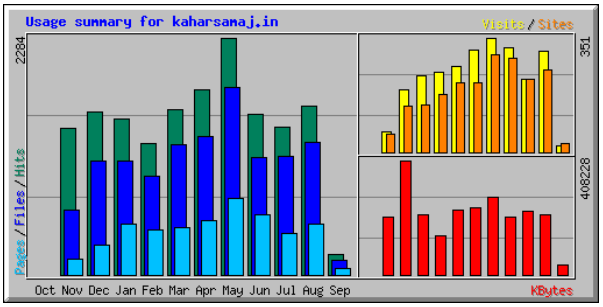

Figure 3 The figure below represent the usage summary of web sites for daily usage for May 2016.

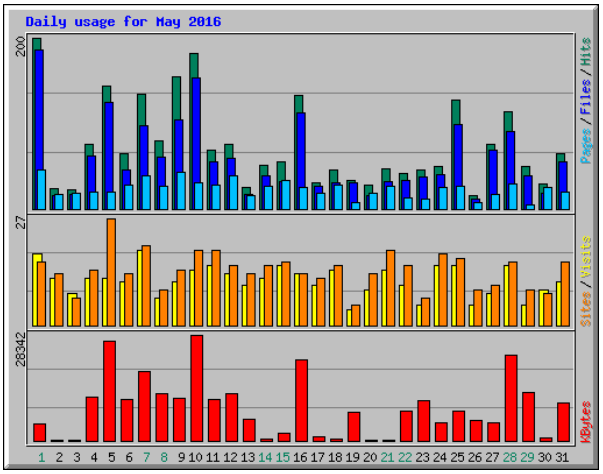

Figure 4

We have collected dataset from different web sites of web server. Our work improved the approach to discover web usage pattern in an application and web server usability system. It also improved the web user's overall satisfaction. Experimental results analysed at the trail level instead of at the individual page level. Such analyses might be more meaningful and yield more interesting results for Web applications with complex structure and operation sequences. Our proposed work improved the performance of web server. The data extracted from the server logs is used for developing user interfaces.

V. CONCLUSION

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.Web has in recent times become a dominant platform for not only retrieving information but also determining knowledge from web data. The web data is stored as unstructured as well as structured format. 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. We proposed procedures for web usage mining and navigation usability. We have implemented the web server usability technique with improved system with large web application. We have collected dataset from different web sites of web server. Our work improved the approach to discover web usage pattern in an application and web server usability system. It also improved the web user's overall satisfaction. Experimental results analyzed at the trail level instead of at the individual page level. Such analyses might be more meaningful and yield more interesting results for Web applications with complex structure and operation sequences. Our proposed work improved the performance of web server. The data extracted from the server logs is used for developing user interface. We plan to focus on designing a host supported system for the complete data leakage detection for large-scale organizations.A mischievous insider or a section of program may hack sensitive personal or organizational information from a web database. Host-based defenses such as identifying the infection onset need to be positioned in its place.

REFERENCES

[1] 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

[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] MitthiasRauterberg, AMME: an Automatic Mental Model Evaluation to analyse user behaviour traced in a finite, discrete state space, ERGONIMCS 1993,vol 36,NO 11 1369-1380

[5] 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.

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

[7] Eduardo H. CalvilloGámez, Paul Cairns, Anna L. Cox, From the Gaming Experience to the Wider User Experience, *BCS HCI* '09, 24,29

[8] 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.

[9] Tec-Ed, "Assessing web site usability from server log files," White Paper, Tec-Ed, 1999.

[10] 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.

[11] PooyaJaferian, David Botta, Fahimeh Raja, Kirstie Hawkey, Konstantin Beznosov, Guidelines for Designing IT Security Management Tools, pp 221-230
[12] Tec-Ed, Inc., Assessing Web Site Usability from Server Log Files, White Paper, 2008, pp-1-16
[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] *Ashwini R*, Enhancing Web Navigation Appropriateness by Correlating Actual and Predictable Practice, IJECS Volume 05, 2016, pp-16369-16376
[15] S. Ancy, V. Subhashini, R. Pooja Karpagam, S. Sujeethra, L. Jayachitra, An Overview of Personalized Recommendation System to Improve Web Navigation, International Journal on Recent and Innovation Trends in Computing and Communication, Volume: 4, 2016, pp-283-287
[16]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