International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

Web Log Analyser for Customer Churn Prediction

Prasadkumar Jadhav[1], Avinash Kale[2], Dipti Gore[3], Aniket Chavan[4]
[1, 2, 3, 4] *(Student, Department of Computer Engineering), MIT college of Engineering, Pune, Maharashtra, India*

Abstract: ***The Log Analyzer project provides an easy to use but powerful front end for searching, reviewing and analyzing network event data, including syslog, windows event log and many other event sources. It focuses on the user-interface side of this project, so the data itself needs to be gathered by another program.***
People are more interested in analysing log files which can offer more useful insight into web site usage. Log file in simple words is a file that documents the actions of a web server. Almost every web server has a system that records the happenings in a website right from the time a person enters a website to the time he exits.
This paper gives the system that can also analyze and track the records what is happening in a website from the right time a person enters a website till he quits. In addition what motivates the project to be implemented is the statistics generated of the referred URL and browser.
Keywords: Log Analyzer, Re-uploading Pages, Activity Statistics, Referral Statistics, Analysis of Customer Churn Prediction, Apriori Algorithm.

I. INTRODUCTION

The web is an important part of the Internet [1], and it becomes the largest publishing system in the world with its pragmatic natural attributes. With increased information sharing through network, attackers are attracted by the range and diversity of information, which causes the continued increase of attack frequency. A web may allow users to interact and collaborate with each other in a social media dialogue as creators of user-generated context in a virtual Community.

So, World Wide Web becomes more popular and user friendly for transferring information Therefore, people are more interested in analysing log files which can offer more useful insight into web site usage. Log file in simple words is a file that documents the actions of a web server. Almost every web server has a system that records the happenings in a website right from the time a person enters a website to the time he exits. The server can identify the pages, images and files requested, details of those who requested them and the number of bytes transferred.

In the proposed system the system starts with user session tracking and generates reports activities like total hit counts, new user and access date. Website owner can define goals that the user should accomplish when browsing through website.

The system provides referral statistics report to know how traffic is coming to your site and though which URL and also provides Browser statistics which gives information about the browser.

II. LITERATURE SURVEY

Hui Sun [1] gives an approach to mining frequent attack sequence based on Prefix Span. The experiments are performed on real data, and the evaluations show that the method prefixed span is effective in identifying both the behaviour of scanners and attack sequences in web logs.

G. Neelima et al.[2] proposes a system to analyze the user sessions so that admin get the information regarding the problems occurred to the users. User is identified according to his/her IP address specified in the log file. The user behavior as the time spends on a particular page can be finding.

Jiawei Yuan et al.[3] propose a practical privacy-preserving K-means clustering scheme that can be efficiently outsourced to cloud servers. Our scheme allows cloud servers to perform clustering directly over encrypted datasets, while achieving comparable computational complexity and accuracy compared with clustering over unencrypted ones. We also investigate secure integration of Map Reduce into our scheme, which makes our scheme extremely suitable for cloud computing environment. Thorough security analysis and numerical analysis carry out the performance of our scheme in terms of security and efficiency.

Yongli Ren et al.[4] presents A formalisation of the LQB graph model, a concise representation of user behaviour across the physical and cyber spaces; (2) A comprehensive analysis of the physical and cyber contextual influence on people's moving, querying, and browsing behaviours in an indoor retail space; and (3) The application of the LQB graph model to location, Web content and query Recommendation in this retail space.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.177
Volume 7 Issue V, May 2019- Available at www.ijraset.com

III. PROPOSED SYSTEM

A. *System Architecture*

People are more interested in analysing log files which can offer more useful insight into web site usage. Log file in simple words is a file that documents the actions of a web server. Almost every web server has a system that records the happenings in a website right from the time a person enters a website to the time he exits.

The proposed system can also analyse and track the records what is happening in a website from the right time a person enters a website till he quits. In addition what motivates the project to be implemented is the statistics generated of the referred URL and browser.Figure 1 gives the structural design of proposed system followed by the working.

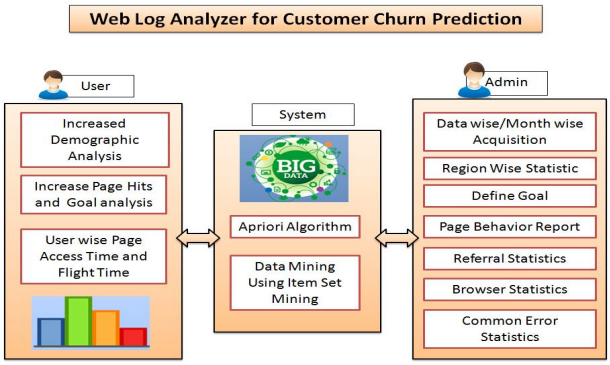

Figure 1: System Architecture

1) *Website Admin:* Admin of the Website whose log need to be generated, adds an analytics JavaScript code at the bottom of the all the WebPages And re-uploads all the pages to the hosting server. As soon as user accesses any of the webpage, the script at the bottom finds all the browser statistics and uploads it to centralized server.
2) *Analytics*

a) *User Session Tracking:* For each user session tracking, a random number is generated and stored in the browser cookies [SESSION_ID_WEBLOG]. The same number is sent in JSON format to server, server keeps track of the unique sessions based on the SESSION_ID_WEBLOG variable. Similarly machine's mac address cannot be sent in the http header, so cross browser cookie is generated for tracking machines.

b) *Activity Statistics:* User wise activity report is shown on the webpage. Each activity report should have Total hits, Total New Visitors, New Users %, access date

3) *Define Goals:* Weblog analysis is important for a company as they want to know which customer is visiting which page and what is causing the user to deviate from the site. Website owner can define goals that the user should accomplish when browsing through website. Report should have following attributes,

a) Define page hierarchy
b) Define Goal Name
c) Behavior
d) Bounce Rate
e) Pages/Session
f) Avg Session Duration
g) Goal Conversion

4) *Referral Statistics:* Site may be referred from different website, example: nowadays people use Google for browsing the website. These search engine URLs are called as referrer URLs. Referral statistics report is important to know how traffic is coming on your site. The report should have

a) Refereed from
b) Total Hits
c) Total New Users %
d) Bounce Rate
e) Goal Conversion

5) Browser Statistics & Common Error Statistics Report

a) Browser Name
b) Total Hits
c) Total New Users %
d) Bounce Rate
e) Goal Conversion

IV. ALGORITHMS USED

A. *Apriori Algorithm*

This algorithm is used to find out the frequency of data into the database. In our application we use to find out which menu items are frequently ordered. Most frequently seen page combinations will be displayed as a mostly browsed ages, so website owner will be able to understand which pages are liked by all the users.

B. *Consider an Example*

1) 1 user is viewing pages item m1, m2, m3, m4.
2) 2nd user is viewing pages item m1, m2, m4.
3) 3rd user is viewing pages item m1, m2.
4) 4th user is viewing pages item m2, m3, m4
5) 5th user is viewing pages item m2, m3
6) 6th and 7th user is viewing pages item m3, m4 and m2, m4 respectively

So the combination got by the apriori is as follow

a) The first step of Apriori is to count up the number of occurrences, called the support, of each member item separately, by scanning the database a first time. We obtain the following result

Item	Frequency
m1	3
m2	6
m3	4
m4	5

b) All the item sets of size 1 have a support of at least 3, so they are all frequent.
c) The next step is to generate a list of all pairs of the frequent items:

Items	Frequency
m1,m2	3
m1,m3	1
m1,m4	2
m2,m3	3
m2,m4	4
m3,m4	3

d) The pairs {m1, m2}, {m2, m3}, {m2, m4}, and {m3, m4} all meet or exceed the minimum support of 3, so they are frequent. The pairs {m1, m3} and {m1, m4} are not. Now, because {m1, m3} and {m1, m4} are not frequent, any larger set which contains {m1, m3} or {m1, m4} cannot be frequent. In this way, we can *prune* sets: we will now look for frequent triples in the database, but we can already exclude all the triples that contain one of these two pairs:
e) Menu Items Frequency m2,m3, m4 2
f) So {m2,m3,m4} is the best and 1st combo and 2nd , 3rd and 4th combos we will take as {m2, m4}, {m2, m3}, {m3,m4}

C. *Experimental Result*

To develop a system for analysing the total hits, new user and access date for the particular website where the user frequently or rarely visits with the help of Apriori Algorithm which helps in identifying the frequency of the visit of user to particular webpage. This system requires Apache Tomcat framework to be installed in the system with the version 7.0. We use Eclipse 3.3 Indigo and MY SQL GUI browser for the implementation and run on Intel P4 Processor machine with 256MB RAM. The Microsoft Windows XP Professional is used as an operating system. With this system we successfully analyze the frequency of hits as the user visits the page and the reason for deviation of user from a particular page or time spent on particular page.

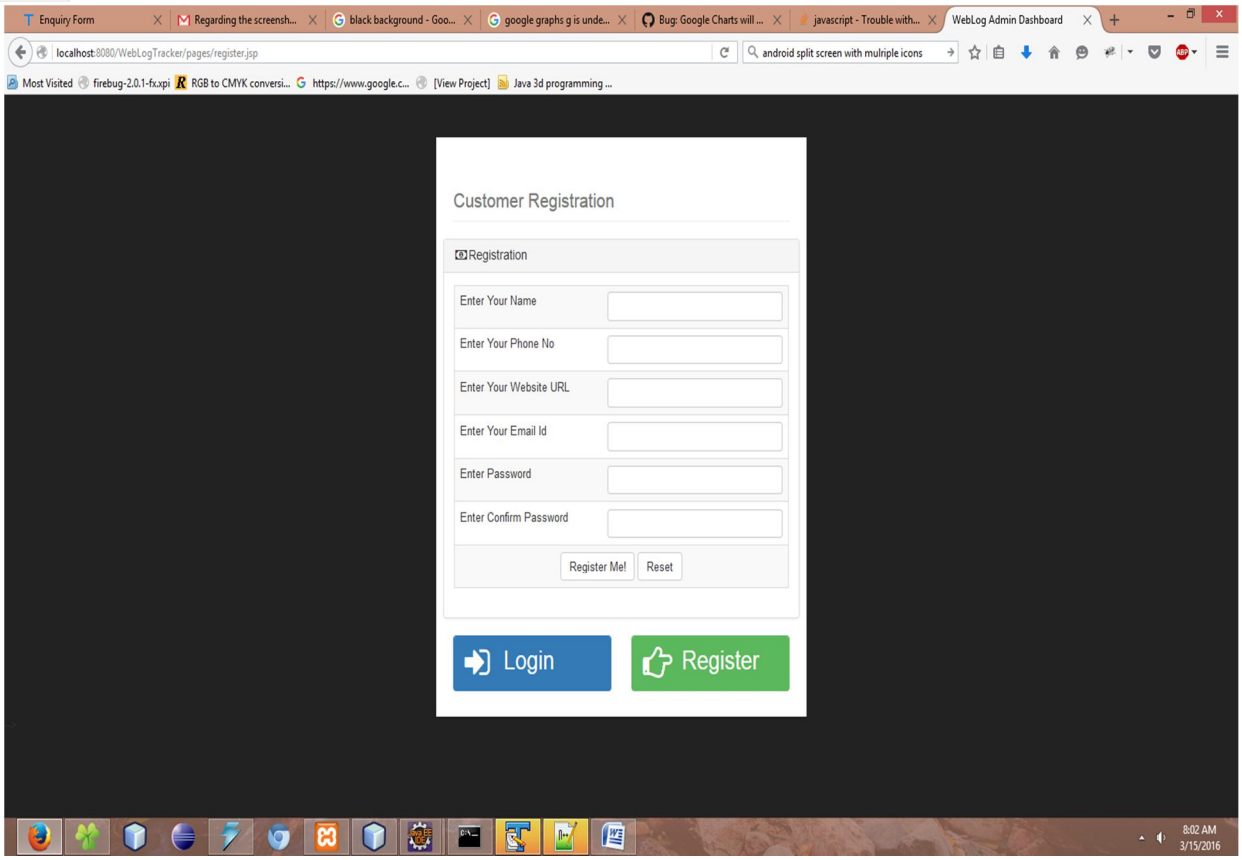

Figure 8.1: Customer Registration

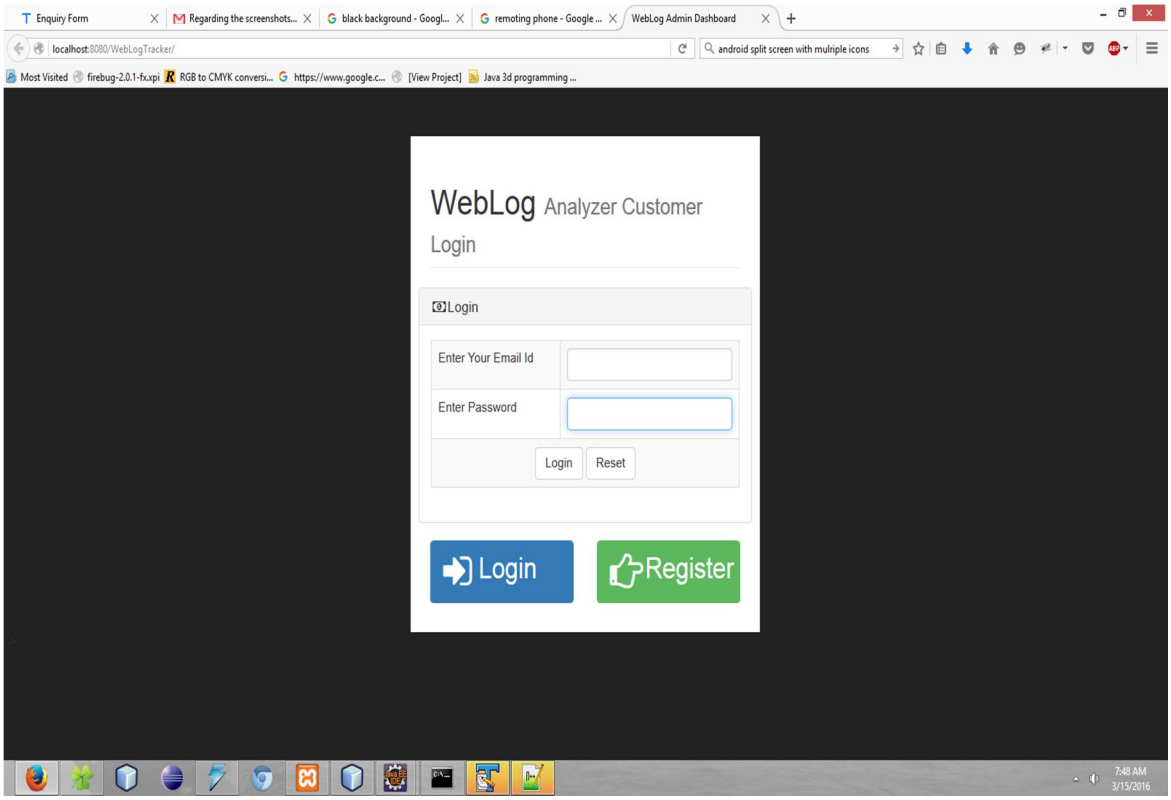

Figure 8.2: Customer LogIn

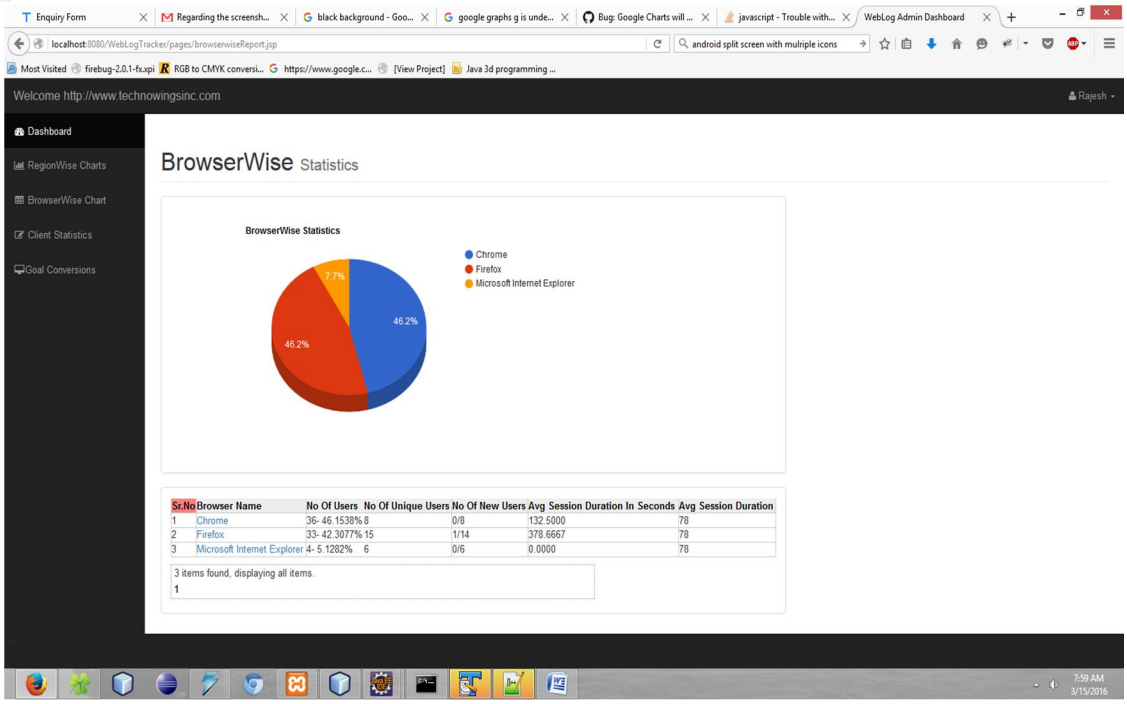

Figure 8.3 Referral Statistics

This system is very useful for better marketing decisions and to attract new viewers or website visitors.

V. CONCLUSION

Here we propose a system which can analyse the particular page events by the user. Using this we come to know that in which particular page user are interested and what is causing the user to deviate from the site. So session tracking and reports generation activities like total hit counts, new user and access date we can finds all the browser statistics.

REFERENCES

[1] LogsHui Sun, Jianhua Sun(B), and Hao Chen “Mining Frequent Attack Sequence in Weblog”, College of Computer Science and Electronic Engineering, Hunan University, Changsha, China {huisun,jhsun,haochen}@hnu.edu.cn.

[2] G. Neelima,Dr. Sireesha Rodda,"Predicting user behavior through Sessions using the Web log mining",International Conference on Advances in Human Machine Interaction (HMI - 2016).

[3] Yongli Ren, Martin Tomko, Flora Salim, Jeffrey Chan, Charles L.A. Clarke, Mark Sanderson ,“A Location-Query-Browse Graph for Contextual Recommendation”, DOI 10.1109/TKDE.2017.2766059, IEEE Transactions on Knowledge and Data Engineering.

[4] Practical Privacy-Preserving MapReduce Based K-means Clustering over Large-scale Dataset Jiawei Yuan, Member, IEEE, Yifan Tian, Student Member, IEEEIEEE Transactions on Cloud Computing VOL. 03, NO. 2, 2017.

[5] Chong Wang, Achir Kalra, Li Zhou, Cristian Borcea, Member, IEEE and Yi Chen, Member, IEEE" Probabilistic Models For Ad View ability Prediction On The Web",10.1109/TKDE.2017.2705688, IEEE Transactions on Knowledge and Data Engineering.