# *Large Scale Analytics And Information Retrival for Web Log Analysis*

**[1]Jyoti B.Ghegade, [2]Neha M.Gujar, [3]Pooja Parsewar, [4]Abhishek Malusare, [5]Prof. Swapna S. Awate**

***Department of Computer Engineering, Zeal College Of Engineering, Pune, India.***

***Abstract: -*** *In proposed system we are developing system which will provides the analysis of users. The purpose of this system is, analysis of user session, user identification, user location and Data can be obtained from databases and displayed in the form of analytical charts on the webpage. This can be used for web log analysis. The pattern analysis can be done with the help of Hadoop server.*

***Keywords: -*** *Web log mining, user identification, user session, user location.*

## I. INTRODUCTION

The web log analyzer is mainly used for analyzing users activities. The Users whole analysis can be done on the website can be done with the help of web log analyzer. The website owners used it for analysis of their users. Web log analyzer used to describe the users behavior on the website. The analysis of user can be done by different aspects. The analysis of user session, user identification, user location can be done in this paper t. It focuses on the user-interface side , so the data itself needs to be gathered by another program. With the help of the algorithm we can provide the JavaScript code for analysis. Data can be obtained from databases and displayed in the form of analytical charts on the webpage. This can be used for web log analysis. The pattern analysis can be done with the help of Hadoop server.Reports are usually generated immediately, but data extracted from the log files can alternatively be stored in a database, allowing various reports to be generated on demand. Web log analyser used for, Number of visits and number of unique visitiors, Visit duration and last visits, Authenticated users, and last authenticated visits. Web log analyzer mainly used for , Days of week and rush hours, Domains/countries of host's visitors, Hosts list, Number of page views , Most viewed, entry, and exit pages , File types , OS used , Browsers used can be identified. With the help of the algorithm we can provide the JavaScript code for analysis. Data can be obtained from databases and displayed in the form of analytical charts on the webpage. This can be used for web log analysis. The pattern analysis can be done with the help of Hadoop server.

## II. LITERATURE SURVEY

In proposed system we are developing system which will provides the analysis of users.The purpose of this system is, analysis of user session, user identification, user location and Data can be obtained from databases and displayed in the form of analytical charts on the webpage**.** This can be used for web log analysis. The pattern analysis can be done with the help of Hadoop server.

Predicting user behavior through Sessions using the Web log mining-In this paper, the method introduces to extract the user sessions from the given log files. Initially, each user is identified according to his/her IP address specified in the log file and corresponding user sessions are extracted. Two types of logs ie., 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. Client side logs can capture accurate, comprehensive usage data for usability analysis.

An opinion mining approach for web user identification and clients behavior analysis –In this paper, initially, each user is identified according to his/her IP address specified in the log file and corresponding user sessions are extracted. Two types of logs ie., server-side logs and client-side logs are commonly used for web usage and usability analysis.

Research on path completion technique in web usage mining -In this paper, they perform experiments on real data, and the evaluations show that our method is effective in identifying both the behavior of scanners and attack sequences in web logs.. Server-side logs can be automatically generated by web servers, with each entry corresponding to a user request. Client

side logs can capture accurate, comprehensive usage data for usability analysis.

## III. SYSTEM ARCHITECTUR

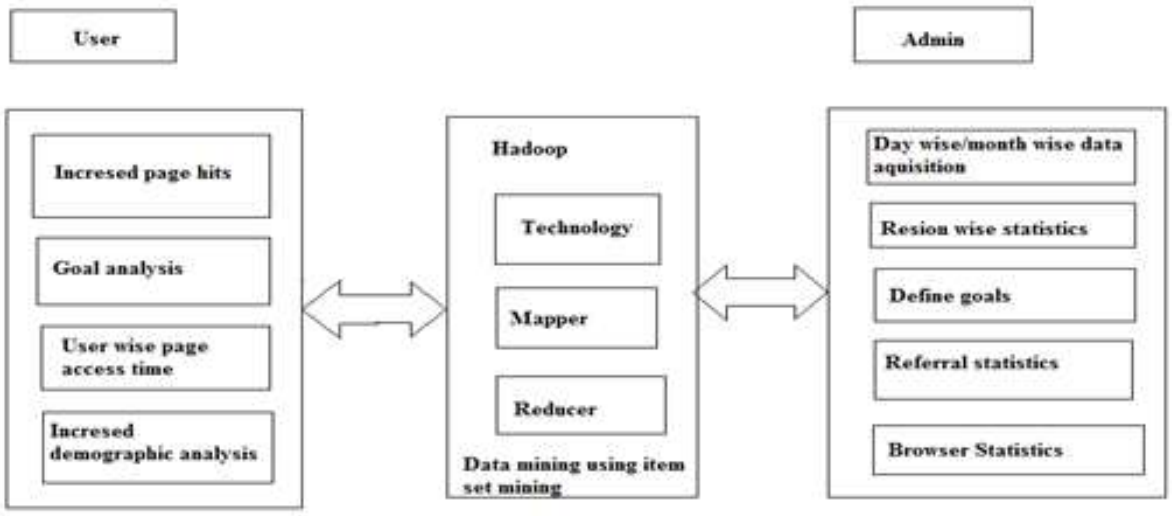


**Fig 1. System Architecture**

In above diagram, the working of web log analyzer is given. The functions of User and Admin are given briefly in the figure. With the help of hadoop sever the pattern can be analyzed. The location, browser, session of the user can be identified with the help of log analyzer. We can compare the users with the help of their analysis. Statistics can help the website owner how to develop his website through user's perspective.

Our system consists of following modules:

### 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.

1. Define page hierarchy
2. Define Goal Name
3. Behavior
4. Bounce Rate
5. Pages/session
6. Avg Session Duration
7. 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

1. Refereed from
2. Total Hits
3. Total New Users %
4. Bounce Rate
5. Goal Conversion

**5. Browser Statistics & Common Error Statistics Report**

1. Browser Name
2. Total Hits
3. Total New Users %
4. Bounce Rate
5. Goal Conversion

## IV.ALGORITHMS

**1. Aproiri Algorithm**

This algorithm is used to find out the frequency of data into the database. In our application we use to find out which pages are mostly visited. 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.

Consider an example:

1 user is viewing pages item m1, m2, m3, m4.

2nd user is viewing pages item m1, m2, m4.

3rd user is viewing pages item m1, m2.

4th user is viewing pages item m2, m3, m4

5th user is viewing pages item m2, m3

6th and 7th user is viewing pages item m3, m4 and m2, m4 respectively

So the combination got by the apriori is as follow

- 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** |

- All the item sets of size 1 have a support of at least 3, so they are all frequent.

- 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** |

- 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:

| Menu Items | Frequency |
|---|---|
| **m2,m3,m4** | **2** |

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}.

## V.CONCLUSION

In this paper we developing a system which will help for website analytics. Web log analyzer used to describe the users behavior on the website .The purpose of this system is, analysis of user session, user identification, user location and Data can be obtained from databases and displayed in the form of analytical charts on the webpage**.** This can be used for web log analysis. The pattern analysis can be done with the help of Hadoop server.

## REFERENCES

[1] "Mining Frequent Attack Sequence in Web Logs", Hue Sun, Xinhua Sun(B), and Hao

Chen,College of Computer Science and Electronic Engineering, Hunan University, Changsha, China.

[2]"Dziczkowski, G., Wegrzyn-Wolska, K., Bougueroua, L.: An opinion mining approach for web user identification and clients behavior analysis. In: 2013 Fifth International Conference on Computational Aspects of Social Networks (CASoN)",pp. 79–84. IEEE (2013).

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

[4] He, J.: Mining users potential interested in personalized information recommendation service. J. Mod. Inf. (2013).

[5] Li, Y., Feng, B.Q., Mao, Q.: Research on path completion technique in web usage mining. In: International Symposium on Computer Science and Computational Technology, pp. 554–559. IEEE (2008).