A SYSTEM FOR CUSTOMER CHURN PREDICTION USING WEB LOG FILES

Mr. Aditya Narayan [1], Dr. Vikas Nandgaonkar [1] , Dr. Soumitra Das [1], Dr. Sunil Rathod [1]

[1] Department of Computer Engineering, Indira College of Engineering and Management, Pune, Maharashtra, India

Corresponding Author

Dr. Vikas Nandgaonkar,
vikas.nandgaonkar@gmail.com

DOI
10.29121/shodhkosh.v5.i1.2024.5616

Funding: This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

The World Wide Web is a significant part of the Internet which is the biggest publishing system in the world. Web analytics is increasing at a very rapid rate ever since the growth of the World Wide Web. Web analytics is the tool for the collecting, measuring, analyzing and reporting of web data and information to understand and optimize web usage.
The paper focuses on the implementation of a system that helps to track record and analyze the users as soon as users enter and leaves the system. The users time spent on a particular web page is recorded, from which the statistics is generated which helps the system to analyze the users traffic in the website along with total number of hits, total new visitors, percentage of new users, date of access.
Such type of analysis helps to find out the cause that deviate the customer from the web site, so that improvement can be carried out to attract the attention of the new visitors to the website.

Keywords: Open Web Analytics, Log File, Log Analyzer, Churn Prediction

1. INTRODUCTION

Log files are the files that contain the information about the user's activities that are accessing the web page [1]. The contents of the log files may be like access time, access date, user name, internet protocol address and universal resource locator and these log files are stored and maintained on web server. The log file that exists in the web server notes the activity of the client who accesses the web server for a web site through the browser. The web server stores all of the files necessary to display the web pages on the user's computer. Log file records information about each and every user and provide information about behavior of users. By analyzing the data from web log files different kinds of statistics can be created like information getting from the weblog will help for analysis of the user visiting the site or leaving the site from a particular page.[2,3] This improves the website to attract new viewers or website visitors and making better marketing decisions.

How to cite this article (APA): Narayan, A., D Nandgaonkar, V., Das, S., and Rathod, S. (2024). A System for Customer Churn Prediction Using Web Log Files. *ShodhKosh: Journal of Visual and Performing Arts, 5*(1), 2532–2539. doi: 10.29121/shodhkosh.v5.i1.2024.5616

World Wide Web becomes more popular and user friendly for transferring information. Therefore, people are more interested in analyzing log files which can offer more useful insight into web site usage [4]. For internet users the information presented on web has developed into an essential source.

In this paper, a method for analyzing web log is presented by using Apriori algorithm. The system analyze the total hits, new user and access date for the particular website where the user frequently or rarely visits, which helps in identifying the frequency of the visit of user to particular webpage. 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.

2. LITERATURE SURVEY

G. Neelima et al. [5] proposes a system to analyze the user sessions so that the admin gets the information regarding the problems that occurred to the users. The 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 found.

Jiawei Yuan, Yifan Tian, et al.[6] proposes a practical privacy-preserving K-means clus- tering scheme that can be efficiently outsourced to cloud servers. Our method enables cloud servers to cluster encrypted datasets directly on their servers. While achieving comparable computational complexity and accuracy compared with clustering's over unencrypted ones. We also investigate the secure integration of Map Reduce into our scheme, which makes our scheme extremely suitable for a 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.[7] presents A formalization of the LQB graph model, a concise rep- resentation of user behavior across the physical and cyber space's; (2) A comprehensive analysis of the physical and cyber contextual influence on people's moving, querying, and browsing behaviors 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.

Dr. M.Balasubramanian et al. [8] gives churn prediction in the mobile telecom system using data mining techniques for reducing customer churn and also likely to reduce error ratio.

Pradeep B et al.[9] build a model for churn prediction for a company using data mining and machine learning techniques namely logistic regression and decision trees.

Jeong & Kim [10], this study uses web session logs to profile user behavior and predict future actions on e-commerce platforms. A machine learning approach combining Random Forest and clustering achieves high accuracy. It helps identify users likely to churn, enabling businesses to take proactive engagement and retention measures.

Adhikary et al. [11], the paper compares deep learning models such as CNN and LSTM for customer churn prediction using telecom datasets. The authors demonstrate improved prediction accuracy compared to traditional models. Their approach emphasizes scalable and automated churn detection systems leveraging behavioral log data and deep feature learning.

Ahmad & Chen [12], this paper analyzes churn prediction using various machine learning models with feature reduction techniques. The authors focus on preprocessing web log data, dimensionality reduction, and ensemble methods to enhance prediction accuracy. The optimized models are applied to telecom customer datasets to demonstrate effectiveness in real-world applications.

Alotaibi & Alshamrani [13], the authors propose a hybrid model using static and dynamic user features extracted from web browser usage logs to predict customer churn. Their method enhances classification accuracy by combining time-series and demographic data, demonstrating improved performance over traditional models in web-based service environments.

Kumari & Chaurasia [14], this study investigates various machine learning algorithms—Random Forest, KNN, Naïve Bayes, XGBoost—for churn prediction in telecom, using log-based features from the Orange dataset. Results show XGBoost outperforms others with 95.6% accuracy, highlighting the importance of proper algorithm selection for churn analysis using behavioral data.

Matsumoto & Yatani [15], the paper explores churn prediction by analyzing usage logs from competitive mobile service apps. The authors propose a method that identifies disengaged users based on cross-platform app interaction data. Their findings suggest usage intensity in rival services is a strong churn predictor, aiding competitive retention strategies.

3. PROPOSED SYSTEM

People are more interested in analyzing log files which can offer more useful insight into web site usage. The proposed system helps to analyze the reason for deviation of user from a particular page or time spent on particular page. It can analyze the frequency of hits as the user visits the page. Figure 1. Shows the architecture of the proposed system.

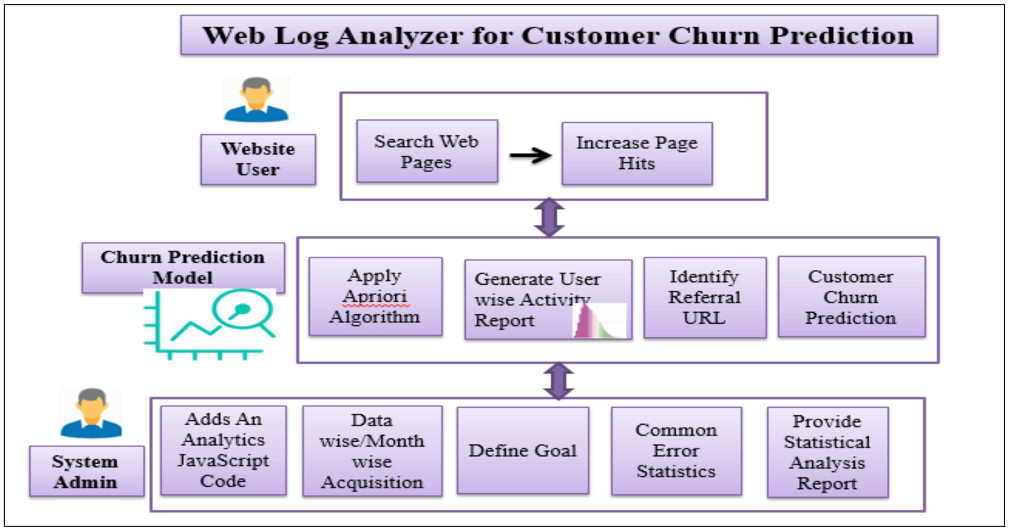

Web Log Analyzer for Customer Churn Prediction
Website User
Search Web Pages
Increase Page Hits
Churn Prediction Model
Apply Apriori Algorithm
Generate User wise Activity Report
Identify Referral URL
Customer Churn Prediction
System Admin
Adds An Analytics JavaScript Code
Data wise/Month wise Acquisition
Define Goal
Common Error Statistics
Provide Statistical Analysis Report

Figure 1: System Architecture of the proposed system

The module wise working of the proposed system is as follows:

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 Web Pages and re-uploads all the pages to the hosting server. When a visitor visits a webpage, the script at the bottom collects all browser statistics and uploads them to a central server.
2) **Analytics:** Session Tracking for Users:A random number is produced and stored in the browser cookies [session_id weblog] for each user session monitoring. The identical number is transmitted to the server in JSON format, and the server uses the session_id_weblog variable to keep track of the unique sessions. Because the MAC address of a machine cannot be transmitted in the http header, a cross-browser cookie is used to track machines.

Statistics on Activity: On the webpage, a user-by-user activity report is displayed.

Each activity report should have Total hits, Total New Visitors, New Users %, access date

3) **Define your Goals:** A company's weblog analysis is significant since it allows them to see which customers are visiting certain pages and what is prompting them to leave the site. The website owner might set goals for the user to achieve while surfing the site. Report should have following attributes
   - Define page hierarchy
   - Define goal name
   - Behavior
   - Bounce rate
   - Pages/Session
   - Average session duration
   - Goal conversion
   - Referral statistics

4) **Statistics on Referrals:** Site may be referred from different website, example: nowadays people use Google for browsing the website. Referrer URLs are the URLs that lead to a search engine. The referral statistics report is essential for understanding how your site's traffic is generated. The report should have
   - Refereed from
   - Total Hits
   - Total New Users %
   - Bounce Rate
   - Goal Conversion

5) Browser Statistics & Common Error Statistics Report
   - Browser Name
   - Total Hits
   - Total New Users %
   - Bounce Rate
   - Goal Conversion
   - Analysis of Customer Churn Prediction

6) Analysis of Customer Churn Prediction: After Analyzing the weblogs the website administrator can predict the reasons causing users leaving the website. And also make better marketing decisions and admin can improve the website to attract new viewers or website visitors. The detail data flow of the system is shown in figure 2.

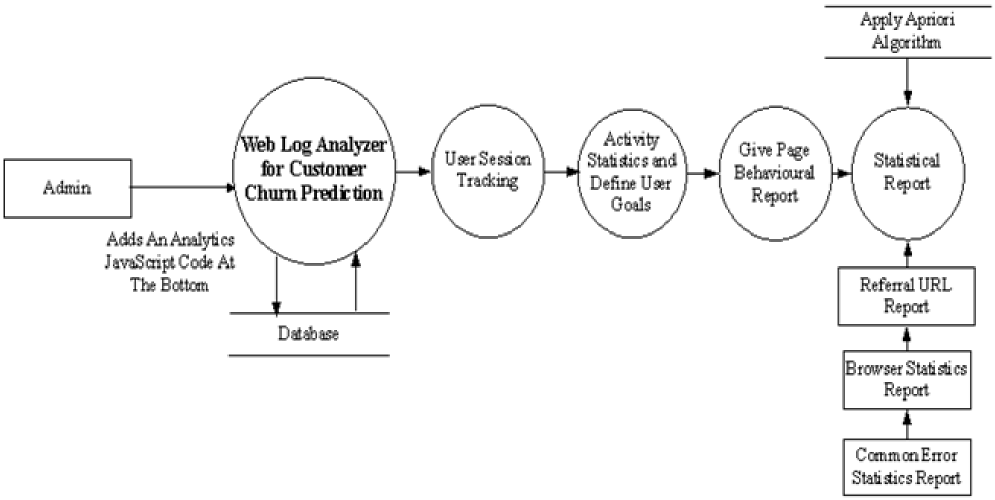

Apply Apriori Algorithm
Admin
Web Log Analyzer for Customer Churn Prediction
User Session Tracking
Activity Statistics and Define User Goals
Give Page Behavioural Report
Statistical Report
Adds An Analytics JavaScript Code At The Bottom
Database
Referral URL Report
Browser Statistics Report
Common Error Statistics Report

Figure 2: Dataflow Diagram of the proposed system

4. TECHNIQUE USED

The proposed system makes the use of an apriori algorithm for calculating maximum accessed page by the user. This algorithm is used to determine the frequency with which data is entered into the database. We use our application to figure out which menu items are most frequently ordered. The most often seen page combinations will be displayed as a most frequently browsed ages, allowing the website owner to discover which sites are popular among all users. For example consider the below users

- Customer No.1 is viewing items I1, I2, I3 and I4 in the browsed page.
- Customer No.2 is viewing items I1, I2 and I4 in the browsed page.
- Customer No.3 is viewing items I1 and I2 in the browsed page.
- Customer No.4 is viewing items I2, I3 and I4 in the browsed page.
- Customer No.5 is viewing items I2 and I3 in the browsed page.
- Customer No.6 is viewing items I3 and I4 in the browsed page.
- Customer No.7 is viewing items I2 and I4 in the browsed page.

The proposed algorithm will give the following results when executed against the item sets

- Initially the algorithm will count up the number of occurrences, called the support of each member item separately, by scanning the database. The result obtained are

Item	Frequency of occurrences
I1	3
I2	6
I3	4
I4	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:

Item	Frequency of occurrences
I1, I2	3
I1, I3	1
I1, I4	2
I2, I3	3
I2, I4	4
I3, I4	3

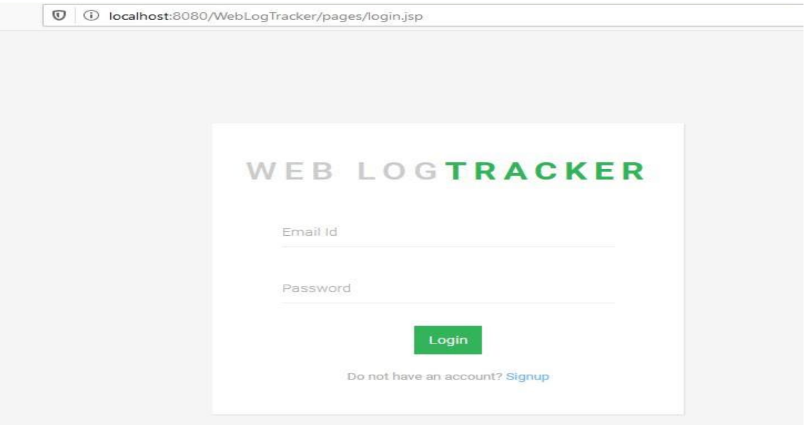
- The pairs {I1,I2}, {I2, I3}, {I2, I4}, and {I3,I4} all meet or exceed the minimum support of 3, so they are frequent. The pairs {I1, I3} and {I1,I4} are not. Now, because {I1, I3} and {I1, I4} are not frequent, any larger set which contains {I1, I3} or {I1, I4} cannot be frequent.
- We can prune sets in this way, we'll now search the database for common triples, but we can already exclude all triples that contain one of these two pairs:

Item	Frequency of occurrences
I2, I3,I4	2

So {I2,I3,I4} is the best and 1st combo and 2nd , 3rd and 4th combos we will take as {I2, I4},
{I2,I3}, {I3,I4}

5. IMPLEMENTATION DETAILS

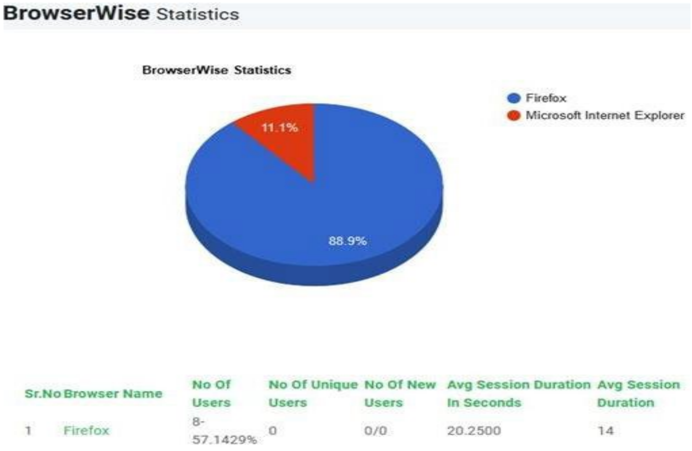
This system necessitates the installation of the Apache Tomcat framework version 7.0. For the implementation, we utilize Eclipse 3.3 Indigo and the mysql graphical user interface browser on an Intel P4 processor with 256MB Ram. As an operating system, Microsoft Windows XP Professional is used. We were able to successfully assess the frequency of hits when the user visits the page, as well as the cause for the user's deviation from a specific page or time spent on a specific page, using this method. This approach is quite effective for making better marketing selections and attracting more viewers or visitors to a website. The user must first connect in to the system using a valid email address and password.



localhost:8080/WebLogTracker/pages/login.jsp

Figure 3: Login Page

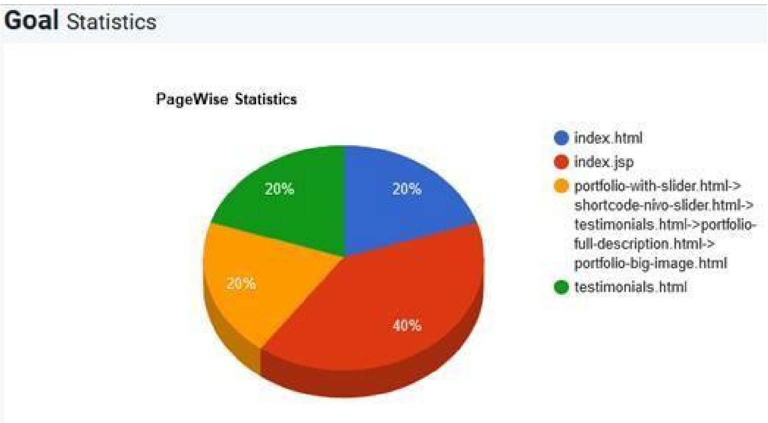
The system provides the browser wise statistic which gives information about the browser as user are now using Firefox for browsing



Sr.No	Browser Name	No Of Users	No Of Unique Users	No Of New Users	Avg Session Duration In Seconds	Avg Session Duration
1	Firefox	8- 57.1429%	0	0/0	20.2500	14

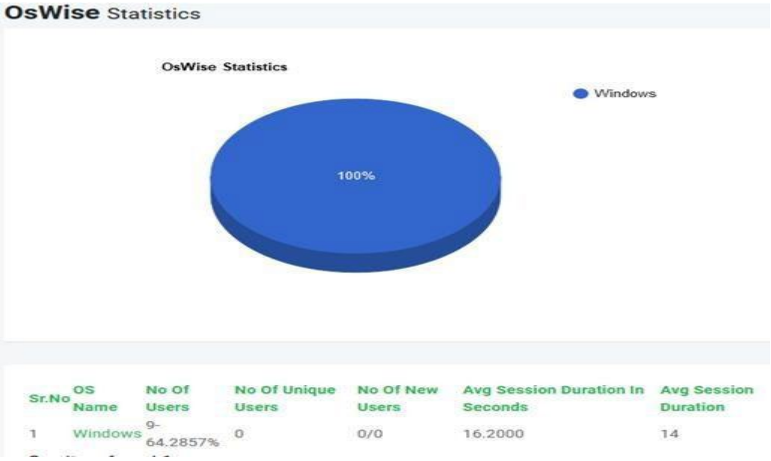
Figure 4: Browser wise Statistics

The system gives how traffic is coming to your site and though which URL



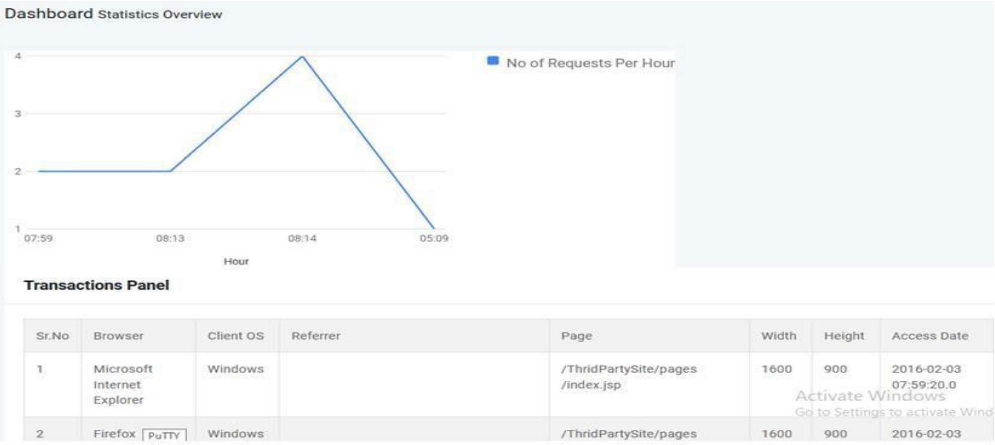
Goal Statistics
PageWise Statistics
index.html
index.jsp
portfolio-with-slider.html->shortcode-nivo-slider.html->testimonials.html->portfolio-full-description.html->portfolio-big-image.html
testimonials.html
20%
20%
20%
40%

Figure 5: Page Wise Statistics



Sr.No	OS Name	No Of Users	No Of Unique Users	No Of New Users	Avg Session Duration In Seconds	Avg Session Duration
1	Windows	9- 64.2857%	0	0/0	16.2000	14

Figure 6: OS Wise Statistics The dashboard provides the all the statistic overview.

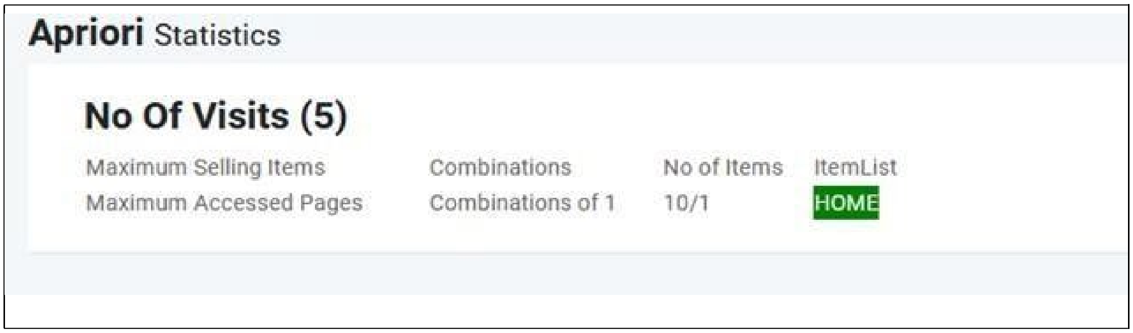


Transactions Panel

Sr.No	Browser	Client OS	Referrer	Page	Width	Height	Access Date
1	Microsoft Internet Explorer	Windows		/ThridPartySite/pages /index.jsp	1600	900	2016-02-03 07:59:20.0
2	Firefox PuTTY	Windows		/ThridPartySite/pages	1600	900	2016-02-03

Figure 7: Dashboard

The system analyses the total hits, new user and access data for the particular website where the user frequently or rarely visits with the help of the Apriori Algorithm, which helps in identifying the frequency of the visit of the user to a particular webpage



No Of Visits (5)

Maximum Selling Items	Combinations	No of Items	ItemList
Maximum Accessed Pages	Combinations of 1	10/1	HOME

Figure 8: Apriori Statistics

6. CONCLUSION

The proposed system analyze and track the records about what is happening in a website from the time a person enters a website till he quits. It helps to find out the reason for the deviation of the user from a particular page. 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.

CONFLICT OF INTERESTS

None.

ACKNOWLEDGMENTS

None.

REFERENCES

L. K. Joshila GraceV. MaheswariDhinaharan Nagamalai, "Web Log Data Analysis and Mining",International Conference on Computer Science and Information Technology CCSIT 2011: Advanced Computing pp 459-469.

Dr.R.Chinnaiyan,Dr.V.Ilango,"Analyzing the User Behaviours by Mining Web Access Log Files",International Journal of advanced studies in Computer Science and Engineering IJASCSE Volume 4, Issue 11, 2015.

Jayanti Mehra,Dr. R S Thakur,"An Effective method for Web Log Preprocessing and Page Access Frequency using Web Usage Mining" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 13, Number 2 (2018) pp. 1227-1232.

Savita Devidas Patil, "Use Of Web Log File For Web Usage Mining",International Journal of Engineering Research & Technology (IJERT),Vol. 2 Issue 4, April - 2013.

G. Neelima,Dr. Sireesha Rodda,“Predicting user behavior through Sessions using the Web log mining”,International Conference on Advances in Human Machine Interac-tion (2016).

Practical Privacy-Preserving MapReduce Based K-means clustering over Large-scale Dataset Jiawei Yuan, Member, IEEE, Yifan Tian, Student Member, IEEETrans-actions on Cloud Computing Vol. 03, No.2, 2017.

Y. Ren, M. Tomko, F. D. Salim, J. Chan, C. L. A. Clarke and M. Sanderson, "A Location-Query-Browse Graph for Contextual Recommendation," in IEEE Transactions on Knowledge and Data Engineering, vol. 30, no. 2, pp. 204-218, 1 Feb. 2018.

Dr. M.Balasubramanian *, M.Selvarani ,“Churn Prediction In Mo- Bile Telecom System Using Data Mining Techniques”, International Web Log Analyzer for Customer Churn Prediction, Journal of Scientific and Research Publications, Volume 4, Issue 4, April 2014.

Pradeep B, Sushmitha Vishwanath Rao and Swati M Puranik†,Akshay Hegde,"Analysis of Customer Churn prediction in Logistic Industry using Machine Learning", International Journal of Scientific and Research Publications, Volume 7, Issue 11, November 2017.

Jeong, S., & Kim, J. (2021). Log-based session profiling and online behavioral prediction in e-commerce websites. Information Systems Frontiers, 23(4), 995–1009

Adhikary, S., Saha, S., Saha, H. N., & Kundu, R. (2021). Customer churn prediction: A comparative study using deep learning. Procedia Computer Science, 192, 790–798

Ahmad, A., & Chen, B. (2022). Model optimization analysis of customer churn prediction using machine learning algorithms with focus on feature reductions. Expert Systems with Applications, 189, 116092.

Alotaibi, R., & Alshamrani, M. (2022). Customer churn prediction for web browsers using dynamic and static features. Expert Systems with Applications, 188, 116032.

Kumari, A., & Chaurasia, V. (2023). Customer churn analysis and prediction in telecommunication sector implementing different machine learning techniques. In Proceedings of International Conference on Advanced Computing and Intelligent Technologies (ACVAIT 2022) (pp. 187–196). Springer.

Matsumoto, M., & Yatani, K. (2023). Stay ahead of the competition: An approach for churn prediction by leveraging competitive service app usage logs. In Proceedings of the 2023 ACM International Symposium on Wearable Computers (pp. 205–208).