International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 9, Issue 4, April 2020||

Enhancement of Application Performance by Understanding the User Behaviour Analysis

Anitha E[1], Muthukumar T[2], Haran Shankar C A[3], Selva Santhosh M[4]

Assistant Professor, Dept. of CSE, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India[1]

U.G Student, Dept. of CSE, Bannari Amman Institute of Technology, Sathyamangalam, Tamilnadu, India[2,3,4]

ABSTRACT: In Retailer online applications, clients are generating reports to identify the sales for a particular period. The report contains sales details of a particular product or a seller. Every time the client has to access the database to generate the same kind of reports to know about weekly or monthly sales of their markets or products. In this system, we are predicting the day, the client will generate the reports and suggest some frequent reports on that day to the client. We proposed a methodology for characterizing and identifying user behaviours. For analysis, the real time data from retailer online application and clustering algorithms is used to group the users that share similar behavioural patterns. The day will be predicted, when the user generates the next report by using the Logistic regression classification algorithm. The predicted reports will be stored in cache. By predicting frequent reports to the user, increases the system performance and improves the user experience.

KEYWORDS: User Behaviour Analysis, Logistic Regression Algorithm, Clustering Algorithms, Application Performance, Data Exploration, Usage Pattern Identification.

I. INTRODUCTION

The internet is very well developed in the past few years, the number of users on the internet are increasing rapidly and the content is also increasing on the other side. Nowadays online transactions are done often. In these environments users interact with the site and with other users through a series of multiple interfaces. For manufacturers and retailers, if they know what customers buy then it will help them to grow their business faster. In today's world, customers and salesmen hold the key to hidden success and growth. To help the manufacturers and retailers, "how the customer thinks and buys", we designed a prediction system. The identification of different classes of user behaviour has the potential to improve application performance in online platform. Different types of patterns can be observed for different types of user groups. After data collected, we use advanced statistics and data science to ensure our data accurately reflects each market, so clients can use it to understand the purchasing behaviours that explain sales.

II. LITERATURE REVIEW

Mimi Zhang et al. [1] proposed that the User behaviour analysis helps enterprises better understand the user's preferences, developing the value of users and ultimately bringing more benefits to enterprises. The development of computer and cloud computing technology has promoted the production of big data. The annual national viewing data volume reaches 3.2 PB. Some enterprises through the full use of the data and digging in the battle to win. The reason which Amazon can win in the book industry is that it deeply mines and analyses the vast information of user behaviour. User behaviour analysis helps Amazon learn more about the user's preference and provide more targeted services.

Zubi Z. S. and Riani M.S.E. [2] discusses the use of web mining techniques to classify the web page's type according to user visits. This classification helps to understand the web user behaviour. The classification and association rule techniques for discovering the interesting information from browsing patterns.

Avnet Saluja et al. [3] in their work is user future request prediction using web log records and user information. The purpose of the effort is to provide a benchmark for evaluating various methods used in the past, a present and which can be used in a future to minimize the search time of a user on the network.

Enrique Frias-Martnez et al [4] proposed that the Prediction accuracy depends on the parameter n, the distance in clicks between the antecedent and the consequent. If n is large the prediction accuracy suffers.

III. DATA COLLECTION

A. *DATA DESCRIPTION*

This dataset contains the data of usage of auser. The four factors affecting system performance are Application processes, Userbehavior, Type of content and hardware/software systems. These details are taken from the real time system. Application log details are current process details, pendingprocesses, Applicationerrors, no of users

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 9, Issue 4, April 2020||

accessing that process. Userbehavior details are Time of login, location of the user, How long was the session, Frequency of login, Type of content accessed, UserType. Content based details are How much data was accessed, Type of data, data Selection methodology, Objectsize, Time required to access that content, Complexity of content. Hardware system details are CPU Usage, MemoryUsage, NetworkFailure, HardwareFailure, System log details, Data losses.

B. *DATA ANALYSIS*

This report has the details of date, userid, reqid, path, value, cache, cache hit and other basic things that are required to understand the user better. Each user has generated many reports and each report has various paths and values. Their execution and error in displaying some results may be due to various factors. In this report we are trying to analyse the data given and relate the data in every possible way to get meaningful insights. Then we would suggest a better way to minimize the errors and then get the effective functioning for the system All unique users accessing the reports in the period of month. Each user can generate any number of reports. So we need to find on which days of the week the users are generating the reports and thereby we can allocate the resources to get the effective performance accordingly.

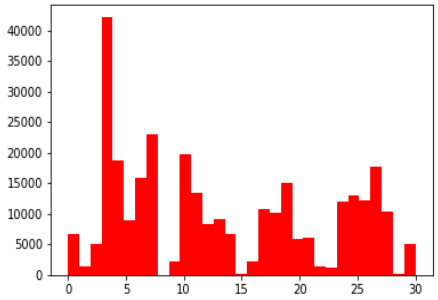

Fig 3.1 Users VS Days

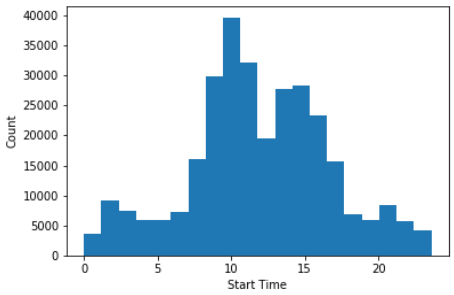

Fig 3.2 Time vs number of reports

Mostly the users are not accessing the reports during the weekends. OnMondays reports accessed are more. From the above inference you were able to know on which days the user is generating more reports. Now we need to know during which time of the day the user is generating morereports. By knowing this we can allocate more assistance during the period and reduce the assistance when there is less traffic. Reports generated are highest between 8am to 5pm as it is theworking hour.

In total there are 6 dimensions involved in the generation of various reports. All dimensions will we can understand which dimension is the most accessed one. Fourth dimension has the highest number of reports and the fifth dimension has the least number of reports generated.

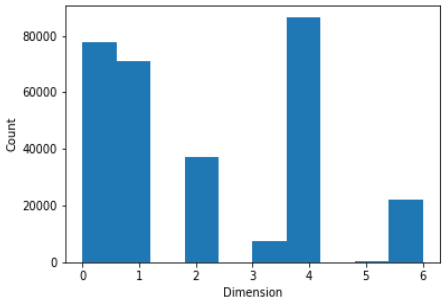

Fig 3.3 Dimension Frequency

To get the better efficiency of the system we need to know the number of unique reports generated based on the path and value for a particular user. If we are able to do so we can find the repeatedly accessed reports and therefore pre-cache them for faster access. Wednesday has more number of unique reports generated. Even when the users accessing on Sunday are less it has more unique reports. Even when the users accessing on Monday are high but the number of unique reports generated are less on Monday.i.e., similar unique reports are generated repeatedly on Monday.so if we are able to store it on Sunday itself the performance can be improved.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 9, Issue 4, April 2020||

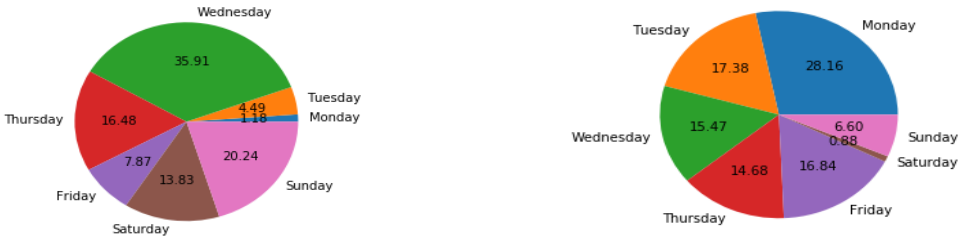

Fig 3.4 No of unique reports and users

We need to know data requested on various days. If we do it, we will be able to make the users comfortable in accessing reports faster. Sunday has the highest number of data cells and Monday has the least data cell count. Wednesdays have the reports that took more time to generate the report than the others. Mondays overall time taken is the least than other days.

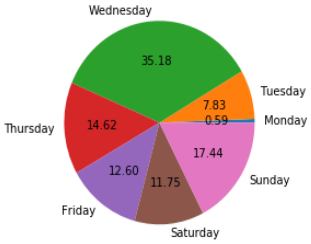

Fig 3.5 Data cell count vs days

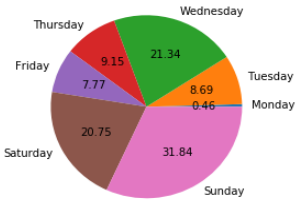

Fig 3.6 Running time vs days

If we know the success and failure on a particular day we can determine which day is successful in generating the report. Then we can work accordingly in increasing the success percentage. Highest success and failure are on Monday because Monday has more number of reports. Saturday has least success and failure because of the least number of reports but the success rate is less when compared to others while excluding the number of reports

day	success	failure	s_percent	f_percent
1	64693	20255	76	24
2	48462	6852	87	13
3	37522	7027	82	18
4	41944	6912	85	15
5	44507	8334	84	16
6	2242	992	69	31
7	13336	2519	84	16

Table 1 Success and failure reports

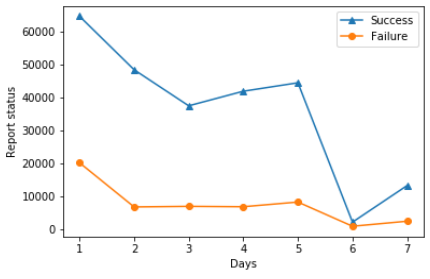

Fig3.7 No of success and failure reports

If we know the success and failure at a particular time we can determine which day is successful in generating the report. Then we can work accordingly in increasing the success percentage. Highest success and least failure occurs at time division 1 and 5 as these are not crowded hours of the day. All the remaining regions are equally distributed.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 9, Issue 4, April 2020||

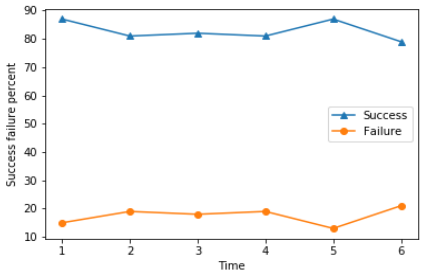

Fig 3.8 No of success and failure reports

time	success	failure	s_percent	f_percent
1	19155	2758	87	15
2	22331	5100	81	19
3	90772	20041	82	18
4	67300	15622	81	19
5	38088	5429	87	13
6	15060	3941	79	21

Table 2 Success and failure reports

For the effective analysis, here the total data is divided into three regions low, medium and high based on the data cell count. On the starting Monday of the month there are more small size reports generated as the day's progress the number of high data reports are increasing.

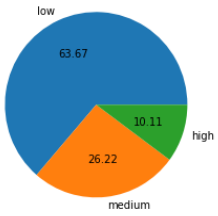

Fig 3.9 Data cell count range

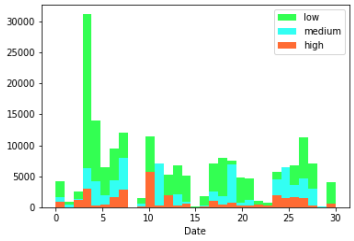

Fig 3.10 Data cell count range vs days

So to understand the reports better we need to know how the cache behaves in various situations. In this, error occurring is a case. Whenthe error occurs the cache does not get stored. i.e. the cache pct becomes zero.

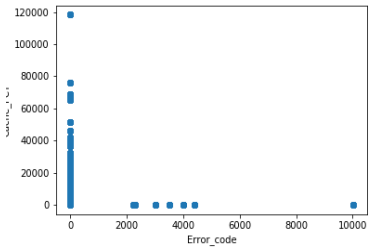

Fig 3.11 Error Code Vs Cache_PCT

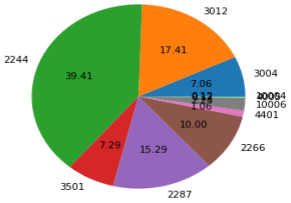

Fig 3.12 Error codes

If we know how many times a path is used, then we could store the data of the particular even before the request by the user and thereby improving the efficiency. The first ten paths have a large number of values and the paths between 125-175 have large number of values.so if we are able to store these paths in the cache then we could utilise it for effective functioning. Later on we try to predict the repeated paths for a user this would help the better system performance.

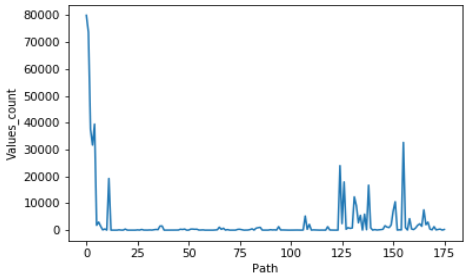

Fig 3.13 Path vs Values count

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 9, Issue 4, April 2020||

Generally, cache refers to the amount of data that is stored in the cache memory. If we have more cache for a particular website or web app they can be accessed in less time. The cached data may not be efficient and cannot be stored for a long time. So we need to analyse the cache and determine various aspects of the cache. Here the total data is divided into three regions based on the amount of cache pct.

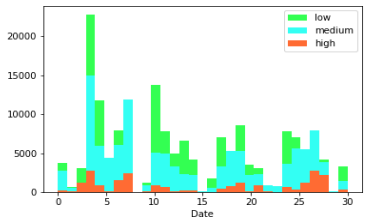

Fig 3.14 Cache analysis

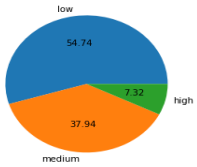

Fig 3.15 Cache Based on days

To understand the users better let us plot the reports having various cache sizes. If we plot in such a way, we will be able to know which days have large sized cached reports and thereby we can conclude some pre caching is done or not. Low cache reports are more on Mondays and the high cache reports are increasing gradually. So if we are able to cache these reports on Sunday we can effectively use the cache. We need to know the cache hit for a particular report and the total duration taken for it. If we do so then we can understand the efficiency of the cached data and can be able to determine which reports have efficient cache and which all do not have it.

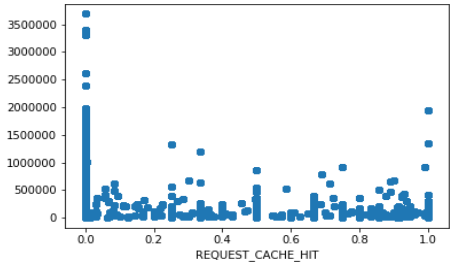

Fig 3.16 Cache hit vs Total duration

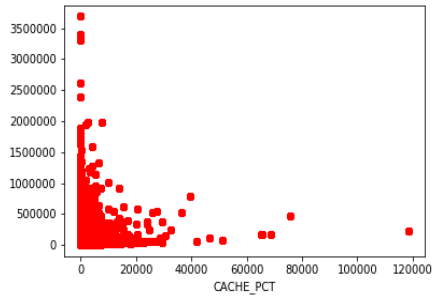

Fig 3.17 Cache PCT vs Total Duration

When the cache hit is high generally the time taken will be low.so here we have some cases even when the cache hit is low we have less time this is because the data cell count requested in those is low. In some cases, when the cache hit is high the time taken is also high this is because every file in cache may be in the requested data cell but every data cell may not be in the cache or there may be many inter layers that fetches unwanted data and stores in cache pct.Cache pct refers to the data cached. If we plot this, we can understand which reports have efficient cache. When the cache pct is low the total duration is high and vice versa. In some cases, even when the cache pct is present it will take more time this happens when there no cache hit or less amount of cache hit.

Data cell count determines how many data cells are present. Ingeneral, we expect the reports with large data cell count take more time to generate. But it cannot be concluded at a stretch, we need to understand various factors involving taking more time. So if we plot count vs total time and count vs cache we can understand we can get some insights from it**.**

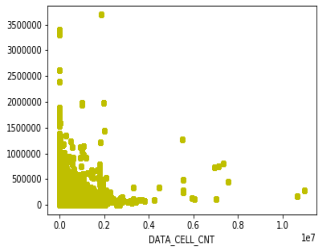

Fig 3.18 Data_cell_count vs Total Duration

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 9, Issue 4, April 2020||

In general, when the data cell count increases the time duration should increase but in some cases it doesn't occur because of the presence of cache. There are some cases where even the data cell count is low, the time is high, one of its reasons may be there is no cache present or the data cell has internal calls. If the data cell count is less it has more chance to present in the cache.

IV. CLUSTERING METHODLOGY

This part explains the methodology for clustering the users. A clustering algorithm is used and that allot users to groups through a distance measure that is analysed based on the values of a normalized vector representation of the users. The normalized vector representation of the users and the clustering procedure is detailed in the following sections [5]. In order to group users that share similar behavioural pattern-means [7] is used as the clustering algorithm and the Euclidean distance 4 as the distance measure. Briefly, this algorithm selects K points in space to be the initial guess of the K centroids 5. Remaining points are then allocated to the nearest centroid. The entire methodology is repeated until no points switch cluster assignment or a number of iterations is performed. In addition to the feature vectors, this algorithm also requires the number of clusters to be created (K) as input. A question then arises: How many clusters should we choose? In [7] the authors suggest that this question can be answered by examining the variation of two metrics: the intracluster distance (average distance between each cluster point and its centroid) and the intercluster distance (average distance between centroids), both characterized by their Coefficient of Variation (CV). The aim is to minimize intracluster CV while maximizing the intercluster CV. The ratio between the intracluster CV and the intercluster CV, denoted by β CV, can help us define the value of K. Varying the number of clusters yields different values for β CV. The symptom for K would be finest when β CV becomes relatively stable.

Algorithm 1 - Clustering Identification Algorithm.

1: K ⇐ 2;
2: repeat
3: K ⇐ K + 1;
4: run K-means algorithm;
5: for (each cluster k returned by K-means) do
6:C k ⇐ centroid of cluster k;
7: if (∃ k, x | d (C k, C x) < T) then
8: merge users from clusters k and x;
9: end if
10: end for
11: until (d (C i, C n) < T) ∧ (d (C n, C j) < T), ∃ i, j, n, where i ≠ j, j ≠ n, i ≠ n, {i, j, n} ∈ [1, K];
12: manually analyze the features and associate the cluster centroids to user behaviours

V.CLASSIFICATION METHODOLOGY

By using clustering methodology, users are grouped into four clusters. After grouping, Logistic regression algorithm is used to classify users into different categories. Using this algorithm, the next report generation day is predicted. For this classification there are four main features that are used. Those features are RFM Clusters, Days between the last three purchases, Mean and Standard Deviation. FM means Recency, Frequency and Monetary values. Recency calculated by most recent purchase date of each customer and see how many days they are inactive for. To create frequency clusters, we need to find total number orders for each customer.agg () method is used find the mean and standard deviation of the difference between purchases in days.

Algorithm – Logistic Regression

Logistic regression is a statistical method for predicting binary classes. The output of the target feature is dichotomous in nature. Dichotomous means there are only two possible classes.

Linear Regression Equation:

$$y = \beta 0 + \beta 1X1 + \beta 2X2 + \ldots\ldots + \beta nXn$$

Where, y is dependent variable and x1, x2 ... and Xn are explanatory variables.

Sigmoid Function: $P = 1/1 + e^{-y}$

Apply Sigmoid function on linear regression:

$$p = 1/1 + e^{-(\beta 0 + \beta 1X1 + \beta 2X2 \ldots\ldots \beta nXn)}$$

Our feature variables are applied in this algorithm and next report

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | Impact Factor: 7.194 |

||Volume 9, Issue 4, April 2020||

generation day predicted. The accuracy of the model can be calculated using cross validation and accuracy of the algorithm boosted by XGBoost algorithm.

VI. Conclusion and Future Work

Analysing the entire data, we have come with the insights that help us to understand the user better. As some users access the same path many times we can predict which path will be used by those users in future. The present caching system is manually done. So if we cache the predicted paths beforehand we can improve the system performance. When the data cell count is high we have error in generating the reports. This issue prevails on few days of the week. So to overcome it, we can increase the hosts to perform the task. Thus effective cache management and data handling can be obtained using this prediction model.

References

[1].Mimi Zhang, Yan Wang, JianPing Chai.Review of User Behavior Analysis Based on Big Data: Method and Application,2015.
[2].Z.S. Zubi, M.S. Riani, "Applying web mining application for user behaviour understanding", Recent Advances in Image, Audio and Signal Processing.
[3].A. Saluja, B. Gour, and L. Singh., "Web Usage Mining Approaches for User's Request Prediction: A Survey", IJCSIT-International Journal of Computer Science and Information Technologies, Vol. 6(3), 2015.
[4].Enrique Frias-Martnez, Vijay Karamcheti, A Customizable Behaviour model for Temporal Prediction of Web User Sequences, in WEBKDD 2002,LNAI 2703, pp 66-85,2003.
[5].Marcelo Maia, Jussara Almeida, Virgílio Almeida .Identifying User Behavior in Online Social Networks.January 2008.
[6]. Douglas Cirqueire, Markus Hofer, Dietmar Nedbal, Markus Helfert, Marija Bezbradica.Customer Purchase Behavior Prediction in E-commerce: Current Tasks, Applications and Methodologies ,2008
[7].A. Jain, M. Murty, and P. Flynn. Data Clustering: A Review. ACM Computing Surveys, 1999.
[8].Muqaddas Gull and Arshi Pervaiz.Customer Behavior Analysis Towards Online Shopping using Data Mining, April 2018.
[9].Jia Li,Comput. Center, Anshan Normal Univ., Anshan, China Research of Analysis of User Behavior Based on Web Log,2013.
[10].Atta-ur-Rahman, Sujata Dash, Ashish Kr. Luhach, Naveen Chilamkurti, Seungmin Baek & Yunyoung Nam.A Neuro-fuzzy approach for user behaviour classification and prediction,2019.