iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 5 **Issue:** IV **Month of publication:** April 2017

DOI: http://doi.org/10.22214/ijraset.2017.4249

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)

Analysis of Web Log Server Files of E-Commerce Websites to Study Customer Behavior Pattern

Sauhard Nandan Chaurasia[1], Parth Sharma[2], Asst. Prof. Mr. Vivek Jain[3]
[1]B.Tech(CSE), [1,3]Dept. of Computer Science and Engg., IMS Engineering College, Ghaziabad

Abstract: ***web usage mining refers to the automatic discovery and analysis of patterns in click stream and related data collected or generated as a result of user interactions with web resources on one or more websites. It consists of three phases which are data preprocessing, pattern discovery and pattern analysis. Data preprocessing involves removal of unnecessary data. In the pattern discovery phase, frequent pattern discovery algorithms are applied on raw data. In the pattern analysis phase, interesting knowledge is extracted from frequent patterns and these results are used for website modification. With the exceptional growth of available information online, especially with the increase in popularity of electronic commerce, web data mining is being paid much attention. In this paper, we propose combining of web data mining and e-commerce with the help of linear regression algorithm for obtaining frequent access patterns from the web log data and providing valuable information about the user's interest.***
Keywords: Web usage mining, linear regression algorithm, big data, web server log data

I. INTRODUCTION

In the modern era, the Internet has become very vast, dynamic and diverse and it contains unstructured data mostly. It supplies us with gigantic amount of information and also makes it complex to search for the relevant information. With the explosive growth of information in web, mining of data has become drastically important. This is where web data mining come to the rescue. Web mining is the application of data mining (searching) techniques to discover patterns from the Web[1].

E-Learning, Digital libraries, Security and Crime investigation and Customer relationships in E-Commerce and E-Business are some of the areas in which web mining can be applied[2]. Among which customer relationship stands out as the major application of Web mining. A website should be designed to entice the customers. Web mining helps in predicting the customer's behavior and the nature of his next product. This can help in improving sites performance and to recommend products and links according to his need. Visitors entering a site exhibit different behaviors. A user might just search for a particular product and not buy it or might end up in purchase. For understanding your customer better and thus improving the performance of the site, certain standards should be used like performing mining on web log data.

II. LITERATURE SURVEY

Following are the literature survey performed in this area

A. Huiping Peng in 2010 stated the use of FP-growth algorithm for processing the web log records, obtaining a set of frequent access patterns, then using the combination of browse interestingness and site topology interestingness of association rules for web mining.
B. Hao Yan et al. in 2010 proposed a two-step K-means clustering algorithm to search user groups in realistic data collected from WAN. Jiawei Han et al. in 2004 proposed a novel frequent-pattern tree structure, which is an extended prefix-tree structure for storing compressed, crucial information about frequent patterns.
C. Rakesh Agrawal and Ramakrishan Srikant in 1994 considered the problem of discovering association rules between items in a large database of sales transactions.
D. Mohdhelmy Abd Wahab et al. in 2008 describes the pre-processing techniques on IIS Web Server Logs ranging from the raw log file until before mining process can be performed.
E. C.P Sumathi et al. in 2011 presented an overview of the various steps involved in the preprocessing stage. Renata Ivancsy et al. in 2006 investigated three pattern mining approaches from the web usage mining point of view.
F. Vaibhav Kant Singh et al. in 2008 show how the different approaches achieve the objective of frequent mining
G. Dr. R. Krishnamoorthi and K.R Suneetha in 2009 have done the in-depth analysis of Web Log Data of NASA website to find

information about a web site, top errors, potential visitors of the site etc. which help system administrator and Web designer to improve their system by determining occurred systems errors, corrupted and broken links by using web using mining.

H. Mirghani. A. Eltahir, Anour F.A. Dafa-Alla et al introduced this capability of using the data mining techniques to extract information from the server logs with the stages of WUM.

I. Rahul Mishra, Abhachoubey et al discuss about the use of web page accesses from the different server logs to discover the frequent usage by the client and also from the experimental study, finds some interesting patterns through association rule mining algorithm and compares between pattern mining algorithm i.e. Apriori and FP growth algorithm.

J. K.R. Suneetha, Dr. R.K. Krishnamoorthi about the importance of collecting reliable user usage data and the way one can achieve this is by determining problems like system errors, corrupted and broken links and errors which arise in Web Surfing.

K. S.K. Pani, L. Panigrahy, V.H. Sankar, Bikram Keshari Ratha, A.K. Mandal, S.K .Padhiprovided a survey of the pattern extraction algorithms used for web usage mining.

L. R.M. Suresh, R. Padmajavalli discussed the importance of data preprocessing methods and various steps involved in getting the required content effectively.

III. SIGNIFICANCE

The major significance of this project is that an individual can search for his product on e-commerce site more easily. Predicting the behavior of the consumer boosts the performance of that site which also helps the site become more popular as it strengthens its root in web among various other sites[3].

A. The Scope of the Project is

1) To ingest the data from a web server log file into Hadoop.
2) Clean the data.
3) Transform data.
4) Form a pattern of customer behavior.
5) Create queries and reports to visualize results of analysis.

B. In General, some Usages of Applications of Web Data Mining in E-Commerce are listed below[4]-

1) Usage on enterprise resource planning.
2) Usage on management on clients' relationship.
3) Usage on management on products data.
4) Usage on assessment of commercial credit

IV. REQUIREMENTS

Information mentioned below is used by companies to indicate factors such as which gender prefers to shop online more often, which page is most frequently visited and which isn't, which product is the most popular/unpopular in different age groups, the type of errors that a user encounters whilst using the website[5].

A. Downloaded Data must be Readable

1) The dataset used in the project is a web server log file. It should be readable in text format and available as ASCII text with fields separated by commas.
2) Relevant information must be available. Following is the information mandatory and must to be available in the log file:
a) User IP -address
b) Time Stamp
c) Response
d) Product searched
e) Http request
f) Status code
g) No of Times Visited
h) Browser Name
i) Date and time of user's visit to specific pages

3) Data should exist for multiple visitors to the website
4) Data should be clean
a) Data in HDFS should have no null values for primary keys
b) Data in HDFS should have valid IP Addresses
c) Date and time information should be consistent in format eg. yyyy-dd-mm-hh:mi:ss
5) Duplicate data should be removed.
6) Allow sorting data by ip-address, date of visit and page visited
7) Allow filtering of data so that data for a given visitor may be queried
8) Reports and graphs should be available that display number of visitors per day
9) Reports and graphs should be available that display pages with most hits by date
10) Reports and graphs should be available that display repeat visits by a visitor
11) Reports and graphs should be available to depict future page visits

V. METHODOLOGY

Up to this point, we've arranged for the web server log files and we have used the Eclipse framework to load the unstructured data set into a more structured format. Afterwards, we plan to create tables of the aforementioned data set in HIVE, run descriptive analysis using Scalar and predictive analysis using Spark-R and generate a graphical visualization of the study conducted.

A. This project is Divided into following Phases[6]
1) Requirements
2) Design
3) Development and testing of programs & queries
4) Depiction and visualization of results

B. Mainly the procedure of project can be divided into following parts

Table 1: Steps of Implementation

Load log file to HDFS
Create a raw table in hive
Create a validation table in hive
Create error table in hive
Write spark code to validate log file field length and primary key
Store validated file In validated table
Store invalid record in error table
Implement Regression Analysis on validated data
Store data in data storage layer perform data Visualization and analytics over stored data get to conclusion

The graphical representation of our project is shown in Fig. 1.

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

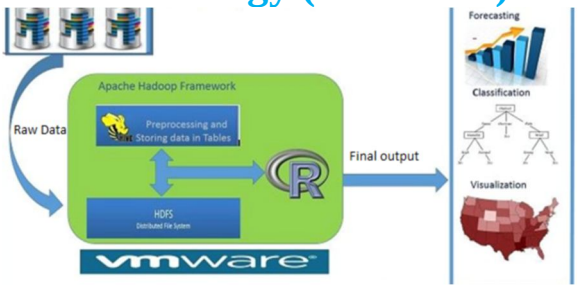


Fig 1: Graphical representation of the Project

We use Linear Regression Algorithm to calculate the precision of the output.

C. Linear Regression Algorithm

Linear regression was developed in the field of statistics and is studied as a model for understanding the relationship between input and output numerical variables, but has been borrowed by machine learning. It is both a statistical algorithm and a machine learning algorithm[7].

Linear regression is a linear model, e.g. a model that assumes a linear relationship between the input variables (x) and the single output variable (y). More specifically, that y can be calculated from a linear combination of the input variables (x).

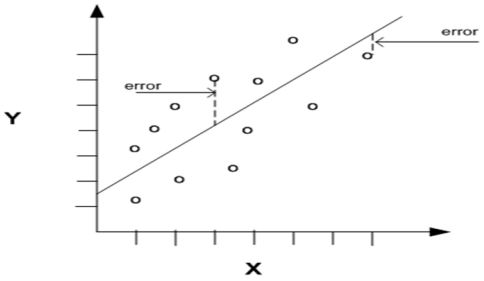


Fig. 2: Error in Linear Regression

D. Linear Regression Model Representation

The representation is a linear equation that combines a specific set of input values (x) the solution to which is the predicted output for that set of input values (y). As such, both the input values (x) and the output value are numeric[8].

The linear equation assigns one scale factor to each input value or column, called a coefficient and represented by the capital Greek letter Beta (β). One additional coefficient is also added, giving the line an additional degree of freedom (e.g. moving up and down on a two-dimensional plot) and is often called the intercept or the bias coefficient.

For example, in a simple regression problem (a single x and a single y), the form of the model would be:

$$y = \beta_0 + \beta_1 * x$$

In higher dimensions when we have more than one input (x), the line is called a plane or a hyper-plane. The representation therefore is the form of the equation and the specific values used for the coefficients (e.g. β_0 and β_1 in the above example).

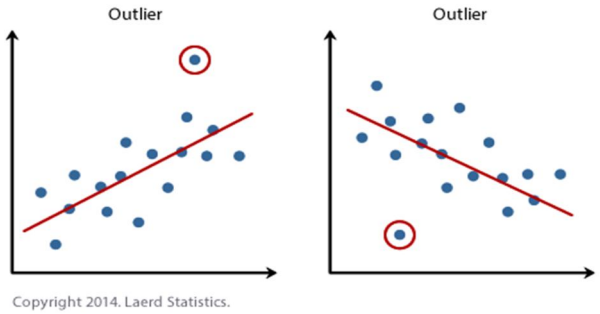


Fig. 3: Outlier Results

It is common to talk about the complexity of a regression model like linear regression. This refers to the number of coefficients used in the model.

When a coefficient becomes zero, it effectively removes the influence of the input variable on the model and therefore from the prediction made from the model (0 * x = 0). This becomes relevant if we look at regularization methods that change the learning algorithm to reduce the complexity of regression models by putting pressure on the absolute size of the coefficients, driving some to zero.

E. Algorithm[9]

1) simple-regression(D) :
2) sx := 0,sy := 0
3) for i = 1, . . . , n do
4) sx := sx + x
5) sy := sy + y
6) od
7) $\bar{x}$:= sx/n, $\bar{y}$:= sy/
8) a := 0, b := 0
9) for i = 1, . . . , n do
10) a := a + $(x_i - \bar{x})(y_i - \bar{y}$
11) b := b + $(x_i - \bar{x})$
12) o
13) β_1 := a/b
14) β_0:= $\hat{y} - \beta_1 x$
15) return (β_0, β_1) where
16) the predictor X is simple, i.e., one-dimensional (X = X_1)
17) r(x) is assumed to be linear: r(x) = $\beta_0 + \beta_1$
18) assume that the variance does not depend on x: Y = $\beta_0 + \beta_1 x +$, E(|x) = 0, V (|x) = σ^2

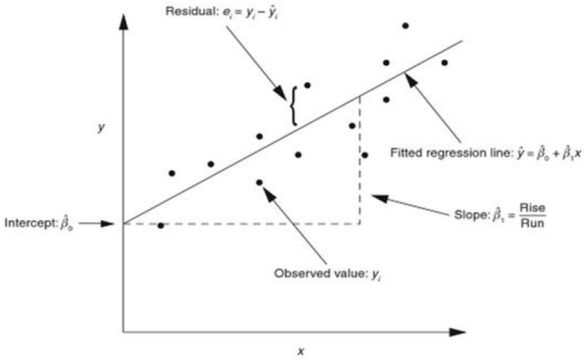


Fig. 4: Keywords of Linear Regression Algo.

VI. IMPLEMENTATION

Following are the elaborated steps performed to achieve desired results

A. Loading the Log File into HDFS

After procuring a log file from a reliable source, we load the concerned log file into HDFS. Currently this log file is unstructured and in .csv format with null values and error prevailing in it. We create a raw table where the concerned log file shall be stored and henceforth load its content into the raw table as shown in Fig.5.

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

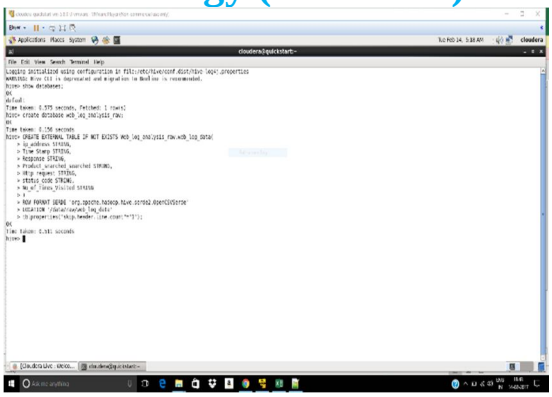


Fig. 5: Loading the log file into HDFS

B. Creation of Error and Validation Tables

With the creation of a raw table, we steadily move forward to creating an error and validation table for the concerned log file where in the contents of the raw table are segregated into error table (if they do not meet the required prerequisite parameters set by us) and validation table as shown in Fig.6 and Fig.7. The content of the validation table shall be processed further for our analysis.

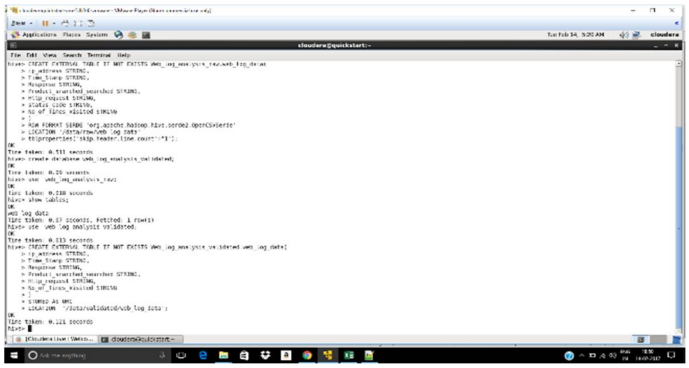


Fig. 6: Creation of Error table

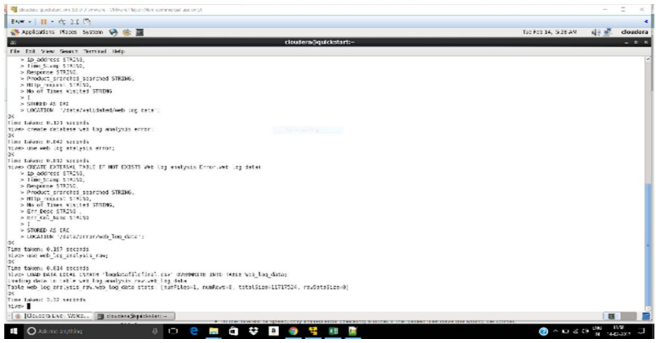


Fig. 7: Creation of validation table

C. Segregation and Storing of Data into Error and Validated Tables

Using a Spark command the concerned raw table's data is now successfully filtered into validated tables (which contains data that passes through the set parameters) and error tables (contains the data that could not pass through the parameters set to be validated). This process is shown in Fig.8 and Fig.9.

Volume 5 Issue IV, April 2017
ISSN: 2321-9653
International Journal for Research in Applied Science & Engineering Technology (IJRASET)

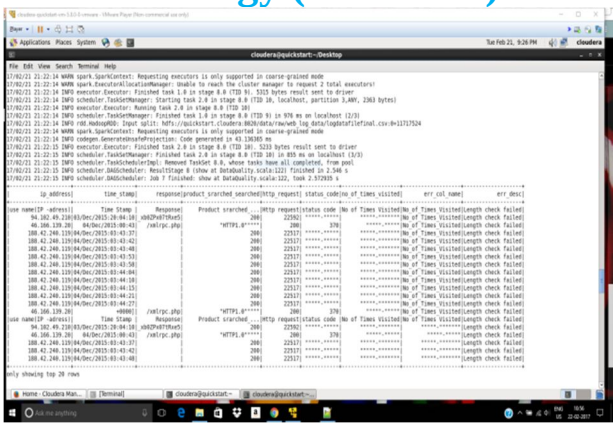


Fig. 8:Segregation of data into Error Table

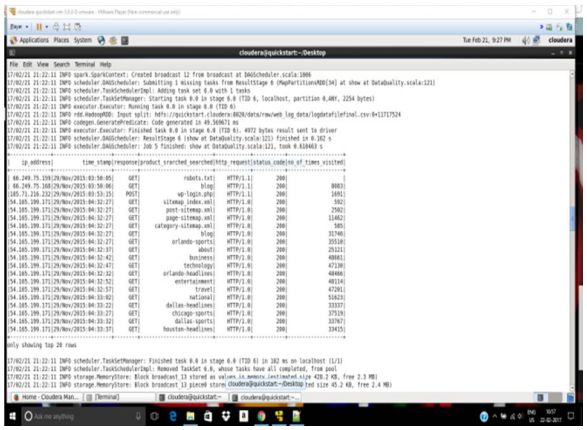


Fig.9: Segregation of data into Validated Table

D. Implement Linear Regression Algorithm on Validated data

The Linear Regression algorithm is implemented on the validated data, wherein the x values shall be the 'User-ID' of a customer and the 'No. of Times a Product is Searched' and the y value shall be the 'Product Searched'[10]. The linear regression algorithm is implemented under supervised learning format as we are aware of the results beforehand. This is shown in Fig.10, Fig.11 and Fig.12.

```
scala> val lrModel = lr.fit(trainingData)
17/03/06 23:59:12 WARN WeightedLeastSquares: regParam is zero, which might cause numerical instability and overfitting.
lrModel: org.apache.spark.ml.regression.LinearRegressionModel = linReg_0318ec6c25b6

scala> println("Coefficients: " + lrModel.coefficients)
Coefficients: [-1.4505146314661586E-14,2.1569937056727044E-4]

scala> print("Intercept: " + lrModel.intercept)
Intercept: 42336.39909810966
scala> lrModel.summary.r2
res11: Double = 0.9891887174009423
```

Fig.10: Coefficient Values

```
scala> import org.apache.spark.ml.evaluation.RegressionEvaluator
import org.apache.spark.ml.evaluation.RegressionEvaluator

scala> val evaluator = new RegressionEvaluator()
evaluator: org.apache.spark.ml.evaluation.RegressionEvaluator = regEval_90b7d0817347

scala> evaluator.setPredictionCol("prediction")
res13: evaluator.type = regEval_90b7d0817347

scala> evaluator.setLabelCol("label")
res14: evaluator.type = regEval_90b7d0817347

scala> evaluator.setMetricName("r2")
res15: evaluator.type = regEval_90b7d0817347

scala> evaluator.evaluate(predictions)
res16: Double = 0.9911889882787136
```

Fig.11: Value of R^2

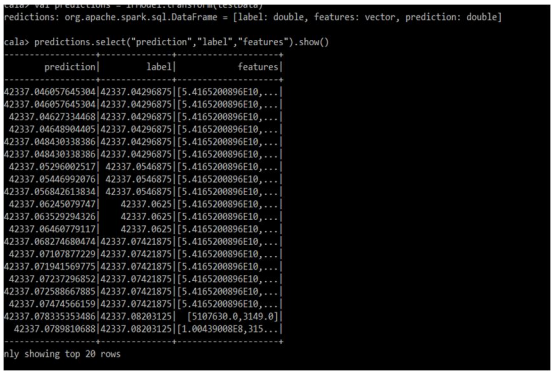


Fig.12: Prediction Values

E. Visualization of Derived Results for Conclusion

The derived result is now used to plot a graph using R which graphically represents the results in a more user-friendly and comprehensive way.

VII. CONCLUSION & FUTURE WORK

Web base data mining has emerged as a research area as the development of the Internet not only presents challenge but also provides a broad application prospects and with the consistent development of the computer, networks and communication technology, the research of Web-based data mining will be further in-depth and it will play an increasingly important role in Web site design, personalized e-commerce services and enterprises decision-making. This project works in the direction so that most frequent pages and products visited by a customer can be illustrated to the owner of the website. With the help of frequent navigated pages some important applications can be performed like page caching, site modification, page personalization, website restructuring etc.

In reference to the future scope, an intelligent data mining system of information can be developed with the combination of data mining and artificial intelligence[11].

REFERENCES

[1] Rupinder Kaur, Kamaljit Kaur, "An Improved Web Mining Technique to Fetch Web Data Using Apriori and Decision Tree" International Journal of Science and Research (IJSR) Volume 3 Issue 6, June 2014.

[2] Shaily G. Langhnoja, Mehul P. Barot, Darshak B. Mehta, "Web Usage Mining Using Association Rule Mining on Clustered Data for Pattern Discovery" International Journal of Data Mining Techniques and Applications Vol 02, Issue 01, June 2013.

[3] Mirghani. A. Eltahir , Anour F.A. Dafa- Alla "Extracting Knowledge from Web Server Logs Using Web Usage Mining" Blue Nile University ,Dmazin, Sudan, IEEE 2011

[4] Rahul Mishra, Abha Choubey "Comparative Analysis of Apriori Algorithm and Frequent Pattern Algori-thm for Frequent Pattern Mining in Web Log Data." International Journal of Computer Science and Information Technologies, Vol. 3(4), 2012

[5] S. K. Pani, , L. Panigrahy, V. H. Sankar, Bikram Keshari Ratha, A.K.Mandal, S. K. Padhi, "Web Usage Mining: A Survey on Pattern Extraction from Web Logs" International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011.

[6] K. R. Suneetha, Dr. R. Krishnamoorthi, " Identifying User Behavior by Analyzing Web Server Access Log File" IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.4, April 2009.

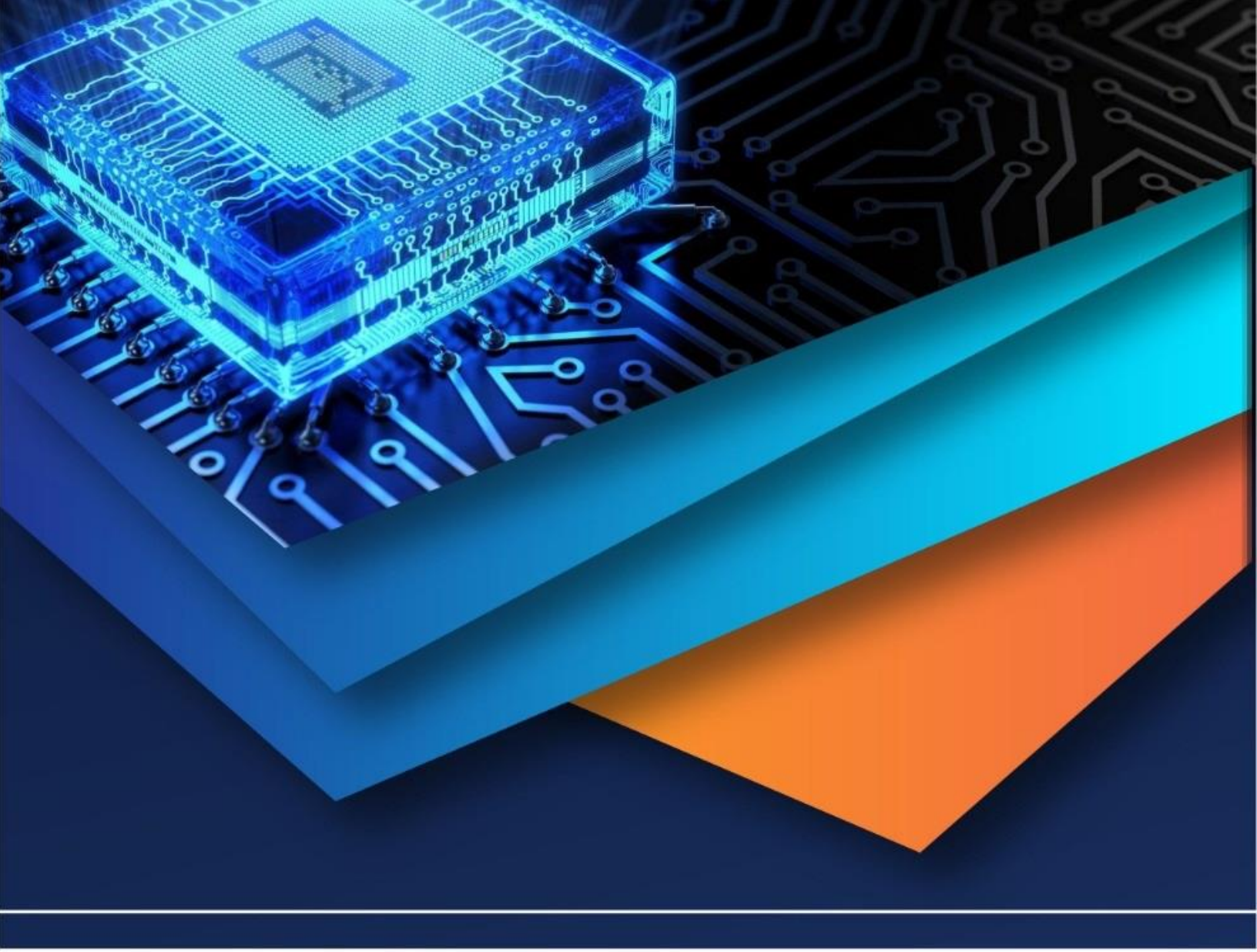
[7] Linear Regression for Machine Learning by Jason Brownlee, Machine Learning Algorithms-. http://machinelearningmastery.com/linear- regression-for-machine-learning/ as accessed on date April 15th, 2017.

[8] R.M. Suresh, R. Padmajavalli, "An Overview of Data Preprocessing in Data and Web Usage Mining", RMK Engineering College, Kavaraipettai, IEEE 2006

[9] Latika Tamrakar, S. M. Ghosh, "Identification of Frequent Navigation Pattern Using Web Usage Mining" International Journal of Advanced Research in Computer Science & Technology (IJARCST 2014) 296 Vol. 2, Issue 2, Ver. 2 (April - June 2014).

[10] Mr. Rahul Mishra, Ms. Abha Choubey, "Discovery of Frequent Patterns from Web Log Data by using FP-Growth algorithm for Web Usage Mining" International Journal of Advanced Research in Computer Science and Software Engineering , Volume 2, Issue 9, September 2012

[11] Guangcan Yu, Chuanlong Xia, Xingyue Guo, " Web Data Mining and its Application in Electronic Commerce", International School of Software Wuhan University, Wuhan, IEEE 2009



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY