International Journal of Engineering Research in Computer Science and Engineering (IJERCSE) Vol 2, Issue 10, October 2015

Big data analysis for Information and Pattern discovery – A case study with reference to jordyblue.com

R. Divyavathi,
Research Scholar,
CMR University

I. INTRODUCTION

The data is growing faster compared to the computation speeds. The source of the growing data could be web, mobile etc. The big data cannot be handled by a single machine and even a single machine cannot store all the data. Hence the big data solution lies in distributing data over large clusters. Few examples of Big data are the Face book's log, Google web index etc. The hardware required for Big Data are lot of hard drives and CPUs, serves.

Big data as the term denotes if it is used properly has a lot of benefits to the business, science etc. The term Big data refers to a huge amount of information management and analysis using technologies and these technologies exceed the efficiency of technologies used for traditional analysis. Although each data can be independently managed and searched, the challenge now is how the organization is makes use of all the types of data. When there is a large amount of data, to make use of the data effectively and efficiently is the challenge of big data.

II. INFORMATION AND PATTERN DISCOVERY

Information is being generated in large numbers in larger organizations where the need to extract value from unstructured information arises. Some of the examples of such processing are the sentimental analysis, financial analysis, social networking analysis etc.

Information is often used along with pattern discovery where the analytical process of the information takes place. The pattern usually covers the analytical processes which are associated with the information. The information pattern discovery is an offline process that takes the information churns it and creates an analytical model.

In this paper, the requirements of the web analytics team of www.jordyblue.com to understand the patterns observed in their website.

III. DATA MANAGEMENT EVOLUTION

Each and every innovation in data management is a fresh start and it is totally disconnected from the past. All the new concepts of data management happens to build on the past data management. Data management has to be viewed as a holistic approach towards data which includes advances in technology in various ways. Hence the various methods the way the data is managed has given new opportunities [1].

As all these technologies factors join together, the way the data is managed is transformed. Big data is the latest trend to overcome all these. Big data is defined as any kind of data source that has at least three characteristics – Volume, Velocity and Variety [1]

Big data is very important because helps the organizations to do many operations with data at the right speed, time and insights.

IV. DEFINING BIG DATA

Given below are few definitions by leading experts and consulting companies:

- The IDC definition of Big Data : "A new generation of technologies and architectures designed to economically extract value from very large volumes of a wide variety of data by enabling high-velocity capture, discovery and/or analysis [2]
- A simple definition by Jason Bloomberg: "Big Data: a massive volume of both structured and unstructured data that is so large that it's difficult to process using traditional database and software techniques." This is also in accordance with the definition given by Jim Gray in his seminal book [3].
- The data that is taken has to be cleansed and documented. The big data that is generated has to be analyzed not in the traditional warehouses. Hence there is a need of entirely new framework to deal with big data.

The main challenges identified in handling big data are:

1) Systems to handle large amount of data
2) Presenting the big data in a way that could be easily understood for the business.

3) How to add value to the business.

V. RESEARCH OBJECTIVES

1. To estimate the descriptive of the variables representing the user profile visiting the website of jordyblue.com
2. To analyze the relationship between the unique page view and no. of visitors.
3. To assess the factors which results into the exit of the visitors

VI. METHODOLOGY

Descriptive and Empirical research is carried out during the study. The data is taken through the Google AdWords of the company website and it is coded in R software. Correlation and ANOVA was carried out to find out the relationship between the variables. The sample size for the analysis is 5000.

VII. KEYWORDS

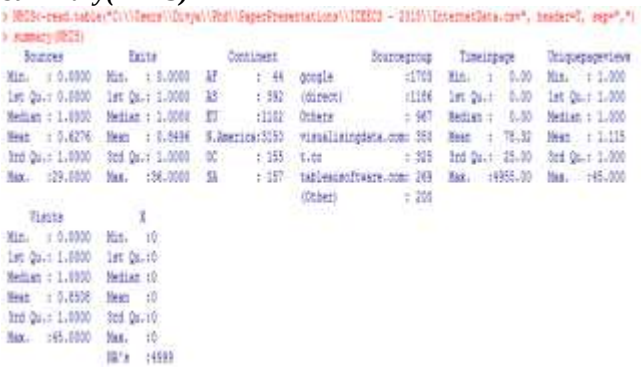
- Bounces variable represents the percentage of visitors who enter the site and leave the site.
- Exits variable represents the percentage of visitors to a site who actively click away to a different site from a specific page.
- Continent variable shows the access of the site from which continent.
- Source group variable shows from where the visitor has accessed the site.
- Time on page variable shows the duration the user has spent on that particular page of the website.
- Unique page view variable represents the number of sessions during which that page was viewed one or more times.
- Visits variable counts all visitors, no matter how many times the same visitor may have been to your site.

VIII. ANALYSIS AND DISCUSSION

To begin with the researcher intends to estimate the descriptive of the variables representing the user profile visiting the website of jordyblue.com

The summarization of data is done to understand how the data is distributed across the data set to know the count, maximum value, minimum value etc. In order to perform the summarization, the following code was used.

summary(NHIS)



From the result it can be observed that the numerical data includes maximum, minimum and mean data. The categorical data like continent indicates the number of times it is repeated in the dataset. The analysis shows that 44 users were from the AF continent, 392 users from AS continent, 1102 from EU continent, 3150 from N. America continent, 155 from OC continent, and 157 from SA continent. It is seen that there were maximum of 30 bounces in the site.

Further to analyze the relationship between the unique page view and no. of visitors, Correlation technique was used. Unique page view represents the number of sessions during which that page was viewed one or more times. A visit is how many times the same visitor has been to the site. Hence the analysis is done whether the unique page view value depends on visits. To do this the following hypothesis was proposed.

H_0: - The uniquepageviews and the visits are not related

H_1:- The uniquepageviews and the visits are related

To do the same the following code was used.

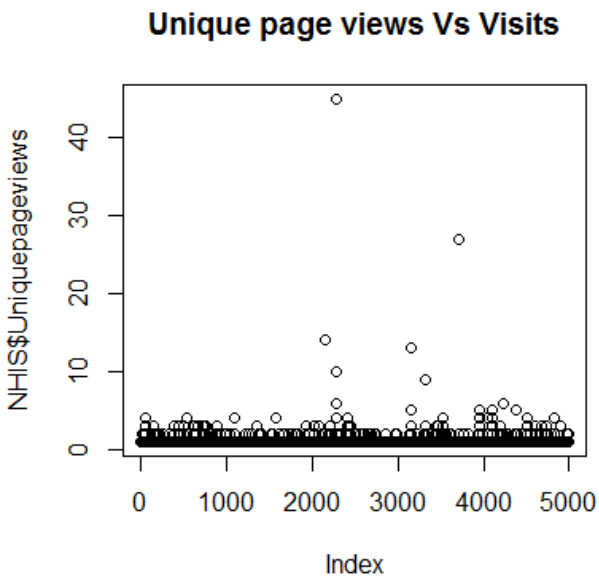
```
x<-NHIS$Iniquepageviews
y<-NHIS$Visits
cor(x,y)
```

From the correlation result it is observed that the correlation value is 0.881. Since the correlation value is between 0.5 and 0.9, it can be concluded that unique page view and the visits are strongly related.

Plotting the content of the dataset provides information about the relationships between all the columns. In this analysis plot generates a scatterplot of the unique page views and the visits.

To do the same the following code was used.

The resulting plot is shown below.



In the above graph it is seen that each point corresponds to the Unique page view and the visits. It indicates that the unique page view is related to the visits.

Continuing to the analysis of the different variables, it can be seen from the results of the descriptive study it is seen that the mean of the time of visit for the user is 78.32. In order to assess the factors which results into the exit of the visitors ANOVA was performed.

Here the probable factor to be found out as why the user leaves the website for a session and goes to another one. Analysis is done to find out whether the exit from the web page is dependent on the other variables. In order to carry on this analysis the following hypothesis was proposed.

H0:- The exit is not dependent on the Timeinpage, Continent, Sourcegroup, Bounces, Uniquepageviews, visits.

H1:- The exit is dependent on the Timeinpage, Continent, Sourcegroup, Bounces, Uniquepageviews, visits.

To carry out this analysis the following code was used.

Ano<-
aov(Exits~Timeinpage+Continent+Sourcegroup+Bounces+Uniquepageviews+Visits, data=NHIS)

Summary(ano)

```
> plot(NHIS$Uniquepageviews, NHIS$visits, main="Unique page views Vs Visits")
> Ano<-aov(Exits~Timeinpage+Continent+Sourcegroup+Bounces+Uniquepageviews+Visits, data=NHIS)
> summary(Ano)
                Df Sum Sq Mean Sq  F value  Pr(>F)
Timeinpage       1    0.4     0.4    3.138 0.07653 .
Continent        5   12.3     2.5   20.032 < 2e-16 ***
Sourcegroup      7   50.7     7.2   59.055 < 2e-16 ***
Bounces          1 2853.3  2853.3 23277.168 < 2e-16 ***
Uniquepageviews  1  350.4   350.4  2858.695 < 2e-16 ***
Visits           1    1.1     1.1    8.774 0.00307 **
Residuals     4953  610.8     0.1
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the table it is observed that p value for continent, source group, bounces and uniquepageview is 2e-16, visits is 0.00307 which is less than 0.05. p value for Timeinpage is 0.076. It can be interpreted that the exit from the website is dependent on the source group, bounces, unique page view, visits and continent whereas it is not dependent on Timeinpage.

This is desirable because the exit due to source group may be due the connectivity issues, the exit from the bounce is an exit by itself, the exit due to uniquepageview may be due to the expiring of the session and the exit due to visits is less significant. The exit is not dependent on the timeinpage. The Tukey test is done on Continent and source group variables to see where the difference lies.

The TukeyHSD test for Continent variable is

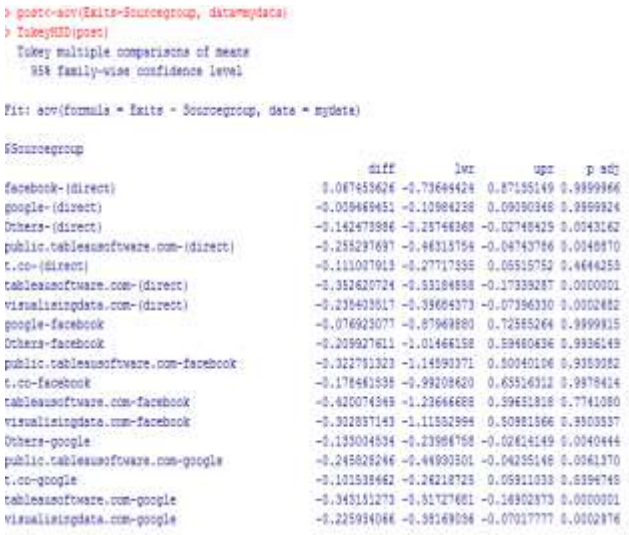
```
> post<-aov(Exits~Continent, data=mydata)
> TukeyHSD(post)
  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = Exits ~ Continent, data = mydata)

$Continent
                     diff          lwr        upr     p adj
AS-AF          0.21103896 -0.187788722 0.60986664 0.6588120
EU-AF          0.09767365 -0.287970364 0.48331767 0.9793128
N.America-AF   0.19246753 -0.188332752 0.57326782 0.7019105
OC-AF          0.22785924 -0.200635778 0.65635425 0.6539663
SA-AF          0.04429647 -0.383594666 0.47218760 0.9997011
EU-AS         -0.11336531 -0.260885985 0.03415537 0.2421220
N.America-AS  -0.01857143 -0.152921342 0.11577848 0.9987779
OC-AS          0.01682028 -0.221190271 0.25483082 0.9999545
SA-AS         -0.16674249 -0.403664136 0.07017915 0.3386789
N.America-EU   0.09479388  0.007000459 0.18258730 0.0255433
OC-EU          0.13018559 -0.085004577 0.34537575 0.5153555
SA-EU         -0.05337718 -0.267362348 0.16060798 0.9806691
OC-N.America   0.03539171 -0.170992297 0.24177571 0.9965703
SA-N.America  -0.14817106 -0.353298343 0.05695621 0.3091045
SA-OC         -0.18356277 -0.467598545 0.10047301 0.4384296
```

From the analysis it is seen that the p value for N.America-EU are 0.0255433 smaller than the significant level 0.05. This suggests that the exit N.America is significantly different from EU, whereas it is seen from the table that for the other pairs the p values are greater that 0.05 and hence the exit from each other continent is not significantly different from the others.

The TurkeyHSD test for the sourcegroup is



From the analysis it is seen that the p-value for others-(direct), public.tableausoftware.com-(direct), tableausoftware.com-(direct), visualisinput.com-(direct), others-google, public.tableausoftware.com-google,tableausoftware.com-google is less than 0.05. This suggests that the exit from these source is significantly differently from the others.

IX. FINDINGS

1. The analysis shows that more no. of users are from the N.America continent and the no. of bounces for the site is 30.
2. It is also observed that the unique page view and the visits are strongly related ie the no. of sessions during which the page was viewed is connected with the visits.
3. The exit from the website is dependent on the source group, bounces, unique page view and visits and continent

REFERENCES

[1] Big Data for Dummies, Alan nugent, Fern Halper, Judith Hurwitz and Marcia Kaufman

[2] http://www.emc.com/collateral/analyst-reports/idc-extracting-value-from-chaos-ar.pdf.

[3] The Fourth Paradigm: Data-Intensive Scientific Discovery. Edited by Hey, T. , Tansley, S. and Tolle, K.. Microsoft Corporation, October 2009. ISBN 978-0-9825442-0-4.

[4] HaiDong Meng et.al, "Research and Implementation of Clustering Algorithm for Arbitrary Clusters", Proceeding of International Conference on Computer Science and Software Engineering, 978-0-7695-3336-0/08,2008.p.255-258

[5] Han and Kamber, " Data mining: concepts and techniques", Morgan Kaufmann publishers, Second edtion,2006, ISBN: 978-1-55860-901-3

[6] Pang-ning tan et.al , " Introduction to Data mining", Pearson Education, Inc.,2006, ISBN: 978-81-317-1472-0

[7] Data Science and Big Data Analytics : Discovering, Analyzing, Visualizing and Presenting Data, EMC2