IJCSE International Journal of Computer Sciences and Engineering Open Access

Review Paper Vol.7, Special Issue.1, Jan 2019 E-ISSN: 2347-2693

A New Approach to Retrieve the Structured Data from a Big Data Set using R

Sumanta Ray

Department of Computer Science, The University of Burdwan

[*]*Corresponding Author: raysumanta1@gmail.com Tel.: 9732160638*

Available online at: www.ijcseonline.org

Abstract- The present research paper computes a step-by-step approach to retrieve the requisite data from big data set, which is heterogeneous in nature. This is done by using R language. Also the process of cleaning, updating, sorting and merging dataset is been explained in this paper. We feel that the comprehensive idea of this research work will be helpful for the researchers, working in the area of big data analysis. Our main goal is to understand the R packages and commands used to analyze big data.

Keywords- R language, SPSS, Big Data, R Packages, R Sql Statements, Classification Tree.

I. INTRODUCTION

R is a open source language and widely used in statistical computing. Section 2 and Section 3 gives a detailed outline regarding Big data and R respectively. Section 5 and Section 6 describes R commands and outputs. Finally we conclude the relevance of this research paper and summarize some of the future research directions in this area.

II. BIG DATA

Big data refers to huge amount of structured and unstructured data stored in different format such as text, video, audio , images and many other formats. It has different types and different structures. In maximum cases big datasets are heterogeneous in nature. Big data has three well known attributes. Known as 3 V's, these are i) Volume, ii) Variety, iii)Velocity. First attributes implies the volume of the dataset is so large, stored in the form of Terabyte, Petabyte, Exabyte or Zettabyte. Variety implies different forms such as images, text, numeric, audio, video etc., Velocity means it is very fast growing data. Nowadays big data is used almost every area, such as social networking site, banking, healthcare, transportation, public and private industries etc.. It is also used in the field of public policy.

III. R: AN OVERVIEW

R is a dedicated tool for big data analysis. R is generally used for statistical data analysis and it is also known as statistical computing language. Portability, efficiency, memory management are some of the important features of R language. Computer language R was first written and developed by Robert Gentleman and Rose Ihaka of Statistics Department of the University of Auckland. It was developed from one language, known as S language. This research paper will focus on how to retrieve the structured and unstructured data from a big dataset using R programming language. R is an open source programming language, freely available and very dominant to analyze the big data. R language has a broad application area in statistical environment.

Here in this paper we particularly concentrate the application of R to retrieve and analyze the required data. This language has a huge number of packages, known as R packages and that is also movable from one system to another system. R also has a large number of background information, documentation and so many resources. If we concentrate on documentation part we will find one syntax "help", through which we can easily understand the nature of different R commands. The technique to process large data sets using R language is also shown in this research paper.

IV. DATA COLLECTION

Nowadays, large structured and unstructured data sources are easily accessible in the official website of different public and private organization. Maximum number of data sources are unstructured means heterogeneous in nature.

Structured data sets are easily accessible, searchable, stored in a relational database, such as spreadsheet, Word Processing etc.. Unstructured data sets are audio, video and image files. SQL is one example of unstructured data set. Structured data sets are easier than unstructured data sets.

V. METHODOLOGY

R is a well known programming language used to process big data. R programming language has so many packages with multiple functions. First we have to install the appropriate packages associated with the commands. We can easily import excel file, SPSS file, STATA file etc. to the R script and in this scenario concerned R packages are required, such as xlsReadwrite package is required to import excel files, RSQLITE, RMYSQL, and SQLDF packages are required to write a SQL query in R. To install the packages we have to write the following commands.

```
>install.packages(“RMySQL”)
>library(“RMySQL”)
```

R can import .csv files using the following command.

```
>A<-read.csv(“c:/desktop/filename.csv”, header=T)
```

R can also import .dta, .sav, .xls files using different commands.

Now we consider two big datasets and perform the following jobs which is assured step wise.

Step 1: We have two large datasets A1 and A2. A1 has 45 number of columns and 807000 number of rows. A2 has 20 number of columns and 1250575 number of rows.

Step 2: We sort two datasets A1 and A2 i.e. we considered first 500000 rows for both the datasets. Now we have A1 dataset contains 45 number of columns and 500000 number of rows and A2 dataset contains 20 number of columns and 500000 number of rows. Sorting the datasets has been done by using SPSS.

Step 3: Now we import two datasets A1 and A2 into R Studio. To import the datasets “foreign” package and “file.choose()”, “read.spss()” commands are used.

Step 4: Then we merge two datasets A1 and A2 by using one common field “serial”. (A1 and A2 both have one common field/column, named “serial”). Here “merge” command is used.

Step 5: Now we have one dataset A3, from A3 we can retrieve the required data. To do that we have to write the accurate SQL command. To write the SQL command in R “SQLDF” package is to be installed. After installing the “SQLDF” package, we can write the different query to retrieve the required data.

For example, if we want to show the complete dataset A3, we have to write the following syntax.

```
>x<-sqldf(“SELECT * FROM A3”)
```

R language supports all the SQL commands generally used to retrieve the required data.

Step 6: After getting the required data we can apply different classification method using R language, such as decision tree, SVM, ANN etc. Here also different R packages and different R commands are required for different classification model. “RPART”, “RATTLE” are well known packages and “pred”, “plot” are well known commands to design a decision tree.

Step 7: R data miner is another feature of R language. We can manage different data field by using R Data miner. Different classification tree can be drawn by changing different fields. If the classification tree is not fit for the concerned problem then the concept of pruning may be implemented.

To open a large .csv file in R, some big packages such as “bigmemory”, “biganalytics”, “biglm”, “bigsplines” and “attach.big.matrix” command are required. We can easily find out the number of rows and number of columns of a big matrix by using “dim” command. The following group of syntax are used to open big data set in R.

```
>install.packages(“bigmemory”)
>install.packages(“biganalytics”)
>install.packages(“biglm”)
>install.packages(“bigsplines”)
>library(“bigmemory”)
>library(“biganalytics”)
>library(“biglm”)

>library(“bigsplines”)
>newfile<-read.big.matrix      (“c:/desktop/bigfilename.csv”,
header=T)
>dim(newfile)
```

VI. CONCLUSION AND FUTURE WORK

In this paper it has been clearly observed the R programming language is working efficiently to process and analyze big data. Apart from R, Python, Openbugs are also well known programming languages to process big data, This paper gives a detailed outline to retrieve the structured data from a big data set. R language can be used in interdisciplinary area. Researchers from various disciplines may exercise R language when big data is compulsory.

VII. R RESOURCES

i) http://www.r-project.org
ii) https://www.tutorialspoint.com/r
iii) https://www.statmethods.net/r-tutorial
iv) http://www.r-tutor.com/r-introduction
v) https://www.w3schools.in/r

REFERENCES

[1] V. Krotov, *“Research Note: Scraping Financial Data from the Web using the R Language”*, Journal of emerging technologies in Accounting, 2018, Vol. 15, No. 1, pp. 169-181.

[2] R. Ihaka, R. Gentleman, “ *R: A Language for Data Analysis and Graphics*”, Journal of computational and graphical statistics, 1996, Vol. 5, No. 3, pp. 299-314.

[3] S. Hill, R.Scott,” *Developing an Approach to Harvesting, Cleaning, and Analyzing Data from Twitter Using R”*. Information Systems Education Journal(ISEDJ), 2017, 15(3), pp. 42-54.

[4] N. J. Horton, E. R. Brown, L. Qian, "*Use of R as a Toolbox for Mathematical Statistical Exploration"*, Taylor & Francis Ltd. On behalf of American Statistical Association, 2004, Vol. 58, No. 4, pp. 343-357.

[5] S. Ray, A. C. Mondal, "*A Brief Description of Software Tools used for Big Data Analysis in Medical Education systems Evaluation"*. IMS Manthan, The Journal of Innovation, Publishing India Group.

[6] I.Mergal,"*Big Data in Public Affairs Education*", Journal of Public Affairs Education, 2016, Vol. 22, No. 2, pp. 231-248.

Authors Profile

Sumanta Ray received his Ph.D. degree from Burdwan University in 2018. He obtained his Bachelor's degree in Computer Applications from The University of Burdwan and Master's degree in Computer Applications from West Bengal University of Technology in 2005 and 2008 respectively. He has about 7 years of teaching experience. He worked in different Institutions. His research interest is in the related areas of Big data Analysis and Soft Computing.