

Big Data Analytics Using R

Sanchita Patil

MCA Department, Vivekanand Education Society's Institute of Technology, Chembur, Mumbai – 400074.

Abstract - R is an open-source data analysis environment and programming language. The process of converting data into knowledge, insight and understanding is Data analysis, which is a critical part of statistics. For the effective processing and analysis of big data, it allows users to conduct a number of tasks that are essential. R consists of numerous ready-to-use statistical modelling algorithms and machine learning which allow users to create reproducible research and develop data products. Although big data processing may be accomplished with other tools as well, it is when one steps on to the data analysis that R really stands unique, owing to the huge amount of built-in statistical formulae and third-party algorithms available.

Key Words: R, Big data, Analytics

1. INTRODUCTION

A statistical analysis package called S was developed by Bell labs in the States. Later in 1994, Ross Ihaka and Robert Gentleman wrote the first version of S at Auckland University and named it R. R is an open-source implementation of S, and differs from S largely in its command-line. For statistical analyses, R has a broad set of facilities that has been specially constructed. As a result, R is said to be a very powerful statistical programming language. The open-source nature of R indicates that, as new techniques for statistics are developed, new packages for R usually become freely available very soon after. It consists of its own inbuilt statistical algorithms – the sheer amount of machine learning algorithms and mathematical models available to users in R and third-party packages is staggering and continues to grow. R can also carry out important analyses that are difficult or next to impossible in many other such packages, including Generalized Additive Models, Linear Mixed Models and Non-Linear Models. R consists of broad range of graph-drawing tools, which makes it easy to produce standard graphs of your data. In traditional analysis, developing a statistical model takes more time than by performing the calculation by the computer. In case of Big Data this proportion is turned upside down. Big Data comes

into picture when the CPU time for the calculation takes longer than the process of designing a model. Data sets that contain up to millions of records can easily be processed with standard R. Data sets with almost one million to one billion records can also be processed in R, but requires some additional effort. Worldwide, millions of statisticians as well as data scientists use R in order to solve their most challenging problems in the field, right from quantitative marketing to computational biology. R have become the most popular language for data science and an most essential tool for analytics-driven companies such as Google, Facebook, LinkedIn and Finance .

2. BIG DATA ANALYTICS

Business value is not generated by stored data and this is true as for traditional databases, data warehouses, also for the new technologies like Hadoop for storing big data. Once the data is appropriately stored, it can be analyzed, and thus immense value can be created. In-memory analytics, in-database analytics and a variety of analysis, technologies and products have arrived that are mainly applicable to big data.

2.1 History of Analytics

The origin for understanding analytics is to explore its roots. In the 1970s, to support decision making, Decision support systems (DSS) were the first systems. DSS was used as an academic discipline and description for an application. Over time, additional decision support applications like executive information systems, dashboards, scorecards as well as OLAP (online analytical processing) became popular. Then in the 1990s, an analyst at Gartner, Howard Dresner, promoted the term business intelligence. Business intelligence (BI) is a process driven by technology for data analyzing and offering actionable information to aid corporate executives, business managers and other end users to construct more informed business decisions. The third interpretation is that analytics is the use of machine learning algorithm to analyze data. It is useful to distinguish between

three kinds of analytics as the dissimilarity have indication for the architectures and technologies used for big data analytics.

3. R's GROWTH

In 2015, IEEE had listed R at 6th position in the top 10 languages of 2015. In addition to this, as the amount of intensive data work increases, demand for tools like R for data-mining, processing and visualization will also increase.

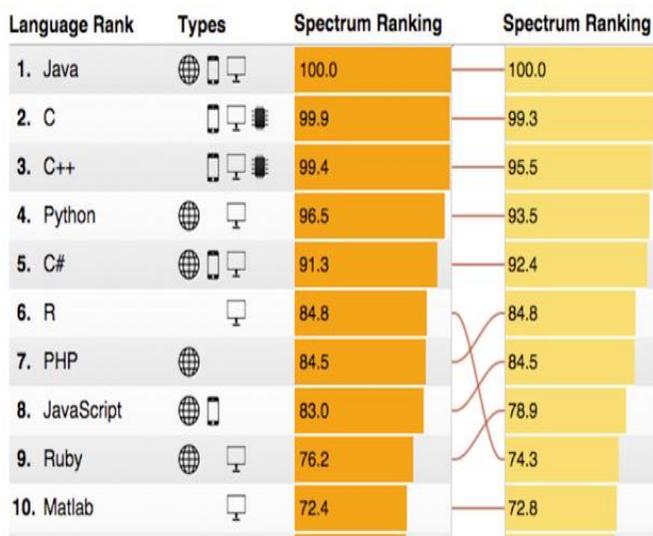


Fig -1: R's Growth [4].

3.1 R in business

R was originated as an open-source version of the S programming language in the 90s. It has gained the support of a number of companies since then, mostly R Studio and Revolution Analytics that are used to create various packages, and services related to the language. R has support from large companies that power to some of the largest relational databases in the world. Oracle, as one, has incorporated R into its offerings.

3.2 R in higher education

R is also originated in academia. Ross Ihaka and Robert Gentleman in New Zealand at the University of Auckland created it, and it's also been widely adopted in graduate programs that include intensive study of statistics. Massive open online course such as the Coursera Data Science Program also makes use of R.

3.3 R has a diverse community

The R community is diverse, along with many individuals coming from unique professional backgrounds. This list includes statisticians, business analytics, academics, scientists and professional programmers. The comprehensive R Archive Network (CRAN), maintains packages that are been created by community members that reflect this background. Packages exist in order to create maps, perform stock market analysis, engage in high-throughput genomic analysis and perform natural language processing.

3.4 R is fun

R is FUN! R has an ability to generate charts and plots in very few lines of code. Tasks that would require multiple lines of code in some other language could be accomplished in R in only a few lines of code. While it's been considered strange when you compare it with many popular languages, it includes powerful features specifically when geared towards data analysis.

4. BENEFITS OF USING R

4.1 Package ecosystem.

One of R's strongest qualities is the vastness of package ecosystem. There's a lot of functionality that's built in and that's built for statisticians.

4.2 R is extensible

R provides rich functionality for developers to build their own tools and methods for analysing data. Lots of people aroused to it from other fields such as biosciences and even humanities. People can extend it without a need to ask permission.

4.3 Free software

At the time when R first came out, the biggest advantage of it was that it was free software. Every single thing and source code about it was available to look at.

4.4 R's graphics and charting capabilities

For data manipulation and plotting the dplyr and ggplot2 packages, respectively have literally improved quality of life.

4.5 R's strong ties to academia

Any new research in the field probably has an associated R package to go. So R stays progressive. The caret package also offers a pretty smart way of doing machine learning in R

through a relatively unified API [3]. A lot of popular machine learning algorithms are implemented in R.

5. R's CHALLENGES

For all of its benefits, R has its share of shortcomings as follows-

- Memory management
- Speed
- Efficiency.

These are probably the biggest challenges R faces. Also, people coming to R from other languages might also consider R odd.

When working with very large data sets the design of the language can sometime lead to problems. Data has to be stored in physical memory. But this can become a minor issue, as nowadays computers have plenty of memory.

Abilities such as security were not built into the R language. Also, R cannot be embedded in a Web browser. You can't use it for Web-like or Internet-like apps. It was primarily next to impossible to use R as back-end server to perform calculations due to lack of security over the Web. For a long time, there was not a lot of interactivity in the language. Languages such as JavaScript still have to enter in to fill this gap. Although an analysis may be done in R, the furnishing of results might be accomplished in different language like JavaScript.

6. BIG DATA STRATEGIES IN R

Big Data can be tackle with R, using five different strategies as follows:

6.1 Sampling

If data is too large in size to be analyzed completely, its size can be reduced by means of sampling. Eventually, the question stands up whether sampling decreases the performance of a model or not. Much data is always better than little data of course. If sampling needs to be avoided it is recommendable to use another Big Data strategy. But if for some reason sampling is necessary, it still can lead to various satisfying models, especially when the sample is kind of big in total numbers, not much small in proportion to the full data set and not biased as well.

6.2 Bigger hardware

R keeps all objects in memory, but this can become a problem if the data gets too large. One of the easiest ways to deal with Big Data in R is to simply increase the machine's memory. Today, R can address to 8 TB of RAM if it runs on 64-bit machines. In many situations this is a sufficient improvement compared to about 2 GB addressable RAM on 32-bit machines.

6.3 Store objects on hard disc and to analyse it chunk wise

As an alternative, there are various packages available to avoid storing data into the memory. Instead, objects are stored on hard disc and then analyzed chunk wise. As a side effect, the chunking also leads to parallelization naturally, if the algorithms allow parallel analysis of the chunks. A negative side of this strategy is, only those algorithms and R functions can be executed that are designed explicitly to deal with datatypes that are hard disc specific.

6.4 To integrate higher performing programming languages like C++ or Java.

Another alternative is to integrate high performance programming languages. The main aim is to balance R's more refined way to deal with data on one side and the higher performance of other languages on the other.

The outsourcing of code chunks from R to another language can easily be hidden with the help of functions. Thus, proficiency is mandatory in other programming languages for the developers, but not for the users of these functions. A connection package of R and Java that is r Java is an example of this kind.

7. Scope of big data analysis using R

For statistical data analysis, R is an open source software platform. Largely because of its open source nature, R is speedily adopted by statistics departments in universities around the world, attracted by its extensible nature as a platform for academic research. [1]Free in cost surely played a role as well. And it wasn't long before researchers in data science, statistics and machine learning started to publish papers in academic journals along with R code applying their new methods. R builds this process very easily and anyone can produce an R package to CRAN that stands for

Comprehensive R Archive Network and make it available to everyone.

An excellent open-source interactive development environment has been created by R Studio for the R language, further boosting the productivity of R users everywhere. [1]

Google, Ford, Twitter, US National Weather Service, The Rockefeller Institute of Government, The Human Rights Data Analysis Group makes use of R.

8. CONCLUSION

To create a powerful and reliable statistical model, data transformation, evaluation of multiple model options, and visualizing the results are essential. This is the reason why the R language has proven so popular: its interactive language uplifts exploration, clarification and presentation. Revolution R Enterprise gives the big-data support and speed to allow the data scientist to repeat through this process quickly.

REFERENCES

- [1] <http://www.r-statistics.com/tag/hadley-wickham/>
- [2] <http://www.infoworld.com/article/2940864/application-development/r-programming-language-statistical-data-analysis.html>
- [3] <http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages>
- [4] <http://www.analytics-tools.com/2012/04/r-basics-introduction-to-r-analytics.html>
- [5] <http://blog.revolutionanalytics.com/>
- [6] <http://www.r-bloggers.com/handling-large-datasets-in-r/>
- [7] <http://www.analytics-tools.com/2012/04/r-basics-introduction-to-r-analytics.html>
- [8] <http://data.vanderbilt.edu/~hornerj/brew/userR2007.r.html>
- [9] <http://blog.ukdataservice.ac.uk/the-power-of-r-methods-for-processing-big-data/>
- [10] <http://bigdatauniversity.com/moodle/course/view.php?id=522>
- [11] <http://aisel.aisnet.org/cgi/viewcontent.cgi?article=3785&context=cais>
- [12] <http://www.revolutionanalytics.com/what-r>
- [13] <http://blog.revolutionanalytics.com/2013/12/tips-on-computing-with-big-data-in-r.html>
- [14] <http://blog.ukdataservice.ac.uk/the-power-of-r-methods-for-processing-big-data/>

- [15] <http://www.oddbms.org/blog/2013/02/on-big-data-analytics-interview-with-david-smith/>
- [16] <http://www.slideshare.net/bytemining/r-hpc>
- [17] <https://www.pluralsight.com/blog/software-development/r-programming-language>
- [18] <http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages>
- [19] <http://www.stat.yale.edu/~mjk56/temp/bigmemory-vignette.pdf>
- [20] https://rpubs.com/msundar/large_data_analysis
- [21] <http://r.cs.purdue.edu/pub/ecoop12.pdf>
- [22] <http://www.inside-r.org/why-use-r>
- [23] http://www.unt.edu/rss/R_Programming_Notes.pdf
- [24] <http://www.cyclismo.org/tutorial/R/input.html>
- [25] <http://www.rosettacode.org/wiki/Category:R>
- [26] https://cran.r-project.org/doc/contrib/Lam-IntroductionToR_LHL.pdf

BIOGRAPHIES



Sanchita Patil,
M.C.A 3rd year Student of
Vivekanand Education Society's
Institute of Technology.