Big data in R

Nazia Gill and Fentaw Abegaz

University of Groningen
JBI of Mathematics and Computer Science

April 15, 2013

Outline

Big Data: Manage it, don't drown in it

Big Data

- Data sets that grow so large and complex that they become awkward to work with using on-hand data base management tools.
- Difficulties may include capture, storage , search , sharing, analytics and visualizing
- Why big data is important?

Challenge of Big Data

- Making decision based on too much information that is not properly managed can be just as dangerous as making decision on too little
- Big data is not only the size of data,analysis time and computer memory.

Memory Limitation

- One of the main problems when dealing with large data set in R is memory limitations.
- Therefore, one cannot store larger data into memory.
- It is impracticable to handle data that is larger than the available RAM for it drastically slows down performance.

1. Working with standard software programs like SAS and R for extremely large data sets are poorly suited and can cause problem to analyzing it .
2. Netflix Prize data consists on 99072112 rating from 480189 users for 17770 movies- in short billion pieces of data.
3. R is very well suited for the development of new methodology but does not handle massive data sets.

Two important commands lines prior to R

1. Chossing tools that can shrink the problems.
2. Fine-tuning R to handle massive data files.

First Pass: Subset the Data

- Data files are often much larger than we need . A better approach is to remove the excess data from the data file before loading in to R.

The Bite Sized- Chunk Approach to Big Data

- Principle : Its impossible to eat a big steak in one bite, we cut our steak into smaller pieces and eat it in one after another.
- Big data rely on same principle. If we need all the data .We break the data in to chunks and fit within the allocated memory.
- Different operations possible,thousands rows a time , or only a few columns.

- Fortunately, there are a handful of packages that have facilitate the use of big data in R and they work by automating and simplifying the bite -sized data approach.
- Generally, they allow most of the data to stay in the working directory, this means that the data do not have to be loaded into the memory.
- They creates an R object with in the memory that acts like a matrix object, but in reality its just a way for you to efficiently access different parts of the data file.

Biglm

- Some algorithms have been designed specifically to handle massive data sets .
- Biglm: implements an iterative algorithm for linear regression that process the data in chunks. However, such solutions are

  not always possible, implementation of different algorithm for a certain type of analysis simply to support different data sized that seems inefficient.

If you have money , buy a bigger computer,If you haven't use memory mapping

- Purchasing more RAM is an option,so moving from a32 bit to 64 bit version of R can alleviate some problems.
- On windows,2GB and 4GB limit for 32 bit and 64 bit respectively.

Bigmemory

- Multi- gigabyte data sets are the challenge and frustrate R users even on well -equiped hard ware.
- The C programming language allows quick , memory efficient operations on massive objects , but not suitable to handle interactive data exploration .
- The package bigmemory bridge the gap between R and C.
- The data sets may also be file-backed , to easily manage and analyze data sets larger than the available RAM.
- The features of big memory projects open the door for power and memory efficient parallel analyses on massive data sets.

Thank you.