Research on the System of Parallel Computing in R

Xianyi Cheng[1,*], Lu Xie[2], Quan Shi[1] and Ping Qu[3]
[1]School of Computer Science and Technology, Nantong University, Nantong 226019, China
[2]School of Electrical Engineering, Nantong University, Nantong 226019, China
[3]Cheento(Beijing) Information Technology Co., LTD, 100080,China
*Corresponding author

***Abstract*—With the advent of big data, the research of data analysis software become the concern of academia and industry in big data. In order to let R to adapt to the big data processing, must break through to parallel computing and memory limit restrictions. In this paper, by analyzing the Rhdoop package, bigmemory package, gputools package and so on, discuss the system of parallel computing in R.**

Keywords-big data; R language; parallel computing

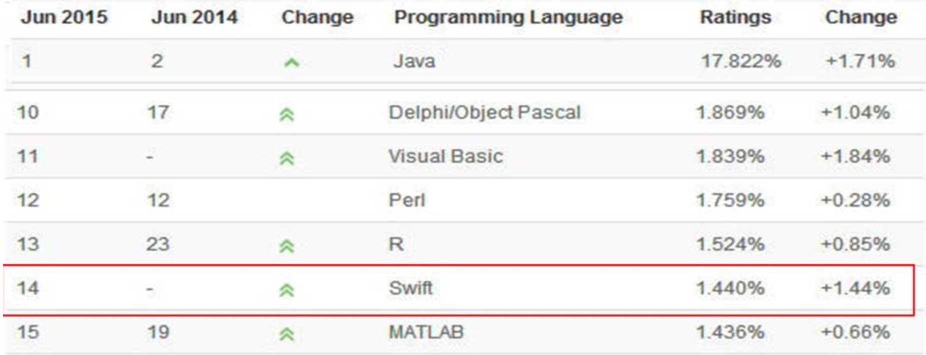
I. INTRODUCTION

R language is a kind of free software programming language and operating environment, it is mainly used for statistical analysis, drawing and data mining[1].

R industry applications include: statistical analysis, applied mathematics, econometric, financial analysis, human science, data mining, artificial intelligence, bioinformatics, geographical science, data visualization.

With the growing demand for big data processing, R application is becoming more and more widely. Table 1 is 2015 issued by the TIOBE programming language ranking list [2].It is expected to squeeze into the top 10 in 2016.

TABLE I. PROGRAMMING LANGUAGE RANKING LIST IS ISSUED BY THE TIOBE IN 2015



Jun 2015	Jun 2014	Change	Programming Language	Ratings	Change
1	2	^	Java	17.822%	+1.71%
10	17	⇈	Delphi/Object Pascal	1.869%	+1.04%
11	-	⇈	Visual Basic	1.839%	+1.84%
12	12		Perl	1.759%	+0.28%
13	23	⇈	R	1.524%	+0.85%
14	-	⇈	Swift	1.440%	+1.44%
15	19	⇈	MATLAB	1.436%	+0.66%

II. R LANGUAGE PARALLEL COMPUTING BREAKTHROUGH

A. Parallel Package

As is known to all, in the era of big data, single thread is one of the weaknesses of R language. But after R2.14, R is built in the parallel package, to strengthen the R parallel computing ability.

The parallel package is similar to lapply function, it is segmentation, calculation and integration input data .Parallel computing is used only to the different CPU to operation.So, parallel computing process can use the following ways:

(1) to start the M a dependent process, and initialized.

(2) for tasks, distribute all the data for each of the attached process.

(3)divided the task into M chunks roughly, and then sends these chunks to dependent process.

(4) wait for all dependent process to complete the computing tasks, and returns a result.

(5) also repeat 2 to 4 for other tasks.

(6) closed dependent process.

In the parallel package, in parallel way, two core function is:

```
parLapply(cl, x, FUN, ...)              #Windows
mclapply(X, FUN, ..., mc.cores)        #Linux
```

Example 1:Without the use of parallel computing, directly using the lapply (implicit cycle function, it is practical for different data using the same function) :

```
>fun <- function(x){
>return (x+1);
>}
>system.time({res <- lapply(1:5000000, fun);});
>user      system   elapsed
21.42     1.74      25.70
```

Example 2:Using the parallel package to accelerate.

```
>library(parallel)
```

```
>#Open the quad-core, specific auditing according to the auditing decisions of the machine
>cl <- makeCluster(getOption("cl.cores", 4));
>system.time({res <- parLapply(cl, 1:5000000,   fun)});
>user system elapsed
  6.54   0.34   19.95
```

B. Foreach package

Foreach package which is contributed to R of the open source community by revolution analytic company,It can make the parallel computation of R is more convenient.

foreach parameters stated below:

foreach(.combine,init,final=NULL,inorder=TRUE,multicombine=FALSE,maxcombine=if(.multicombine)100 else 2,.errorhandling=c('stop','remove','pass'),.packages =NULL,.export= NULL,.noexport=NULL, verbose= FALSE).

combine:Results after operation display mode,default is list,"c" return a vector, cbind and rbind return matrix,"+" and "*" can return after rbind "+" or "*".

init:first variable on ".combine" function .

final:Return last result.

inorder:TRUE the results of the return is the same as the original input sequence,FALSE Returns the result of the no-order (can significantly improve the efficiency of the operation).

muticombine:Set ".combine" the function of passing parameters,default is FALSE.

maxcombine:Set ".combine" the biggest parameters.

errorhandling:If there is an error in the loop, the error processing method.

packages:Specified the depends on package in the operation % dopar %

C. RHdoop Package

Hadoop infrastructure is a distributed system, users can develop distributed application, in the case of not understand the distributed low-level details. Make full use of the power of cluster computing and storage at a high speed. Hadoop implements a Distributed File System (HDFS). HDFS has a characteristic of high fault-tolerance and designed to deploy on cheaper hardware. And it provides the high transfer rate to access the application's data, suitable for those with very large data-set applications. HDFS eased the POSIX requirements that can flow in the form of access data in a file system [3].

RHadoop is a products combination Hadoop and R language, it is developed by Revolution Analytics company, and the code on the open source community[4].RHadoop contains three R package (RMR, RHDFS, rhbase), respectively is corresponding to MapReduce, HDFS, HBase in Hadoop system architecture.

III. R LANGUAGE MEMORY LIMIT BREAKTHROUGH

Table 2 describes several R package can be implemented in memory to store data [5] :

TABLE II. R PACKAGE REALIZING DATA IS STORED IN THE MEMORY

R package	description
ff	Provides a data structure stored in hard disk, but operate like in memory
bigmemory	Support the creation, storage, large-scale matrix read and manipulate. Matrix was assigned to the Shared memory or memory mapped file（memory-mapped files）
filehash	Implements a simple key - value database, in which the character string value associated with the key and the data stored in the hard disk.
ncdf, ncdf4	Provides an interface to Unidata netCDF data files.
RODBC, RMySQL, ROracle,	Can be read external relational database management system of data with these packages

Packages in table 2 can help customer service R memory limit.Besides, when you need to analyze large data sets in a limited time, use special method is also a must.

A. RODBC Package

Using R database connection (for example RMySQL), biopsy for data processing.

(1) connect to the database

channel <-odbcConnect("datasource",uid="XXX",pwd="XXX")

(2) query

This is a common type of operation that can be queried, can add or remove modify againda<-sqlQuery(channel,"select * from student")

(3)Read the table to the data frame

df<-sqlFetch(channel,"student")

B. Bigmemory package

This scheme is suitable for large-scale matrix operations. Bigmemory Family ncluding bigmemory, biglm, biganalytics, synchronicity, Bigtabulate and bigalgebra. Using steps are as follows:

(1) build 'big. The memory' object

Write large data file format ,first to establish “filebacked. Big. Matrix”

```
filebacked.big.matrix(nrow, ncol, type = options()$bigmemory)
```

Filebacked.big.matrix This method of storage backup files, and need a descriptor file;

Type refers to storage format of atomic element in big.matrix,The default is 'double'(8 byte),can be changed to 'integer'(4 byte), 'short'(2 byte) or 'char'(1 byte);

Avoid using rownames and colnames,Because it is memory. If must change, use options (bigmemory.allow.dimnames=TRUE) to set colnames, rownames.

Input x after directly after the command prompt（x is a big matrix）,return x description.Print out the matrix all content, use x[,].

(2)To big. The columns of the matrix of specific elements of the conditional filtering

There is no limit to the memory; And which is more flexible than the traditional.

```
mwhich(x, cols, vals, comps, op = 'AND')
```

x can be big matrix, but also the traditional R object;

cols:The number of column

vals:cutoff, such as c(1, 2)

comps:'eq'(==), 'neq'(!=), 'le'(<), 'lt'(<=), 'ge'(>) and 'gt'(>=)

op:'AND' or 'OR'

Can compare to 'NA','Inf' and '-Inf' directly

(3)Other functions in bigmemory

nrow: The number of rows.

ncol: The number of column

dim: The dimensions of the matrix.

is.big.matrix: Judgment of whether big matrix

as.big.matrix: Converted to big matrix

read.big.matrix: Read big matrix.

write.big.matrix: Write big matrix

is.filebacked: Judgment of whether backup the big matrix.

The enhancement of 'which' is 'mwhich' ,the enhancement of 'order' is 'morder' in base package, more distinctive is bigkmeans clustering.

IV. R LANGUAGE COMPUTING SPEED BREAKTHROUGH

The problem of R calculation speed slowly, especially when cycle. By parallel computing can improve the calculation speed, but this section was calculated acceleration under single thread.

A. Plyr Package

plyr package can be similar to the operation of the pivot table, data is divided into smaller data, some data of the split after operation, finally summarize the results of the operation. use plyr package can according to different data types, within a function at the same time to complete the split-apply-combine three steps, in order to achieve maximum efficiency and concise. plyr package is especially suitable for dealing with large data sets.

Involving matrix of the recycled, the apply function package can have significant speedup.

B. Call C

Create a new directory 'C:/D_package/work_source/R_work/convolve',and create file 'convolve.c'.

```
> setwd('C:/D_package/work_source/R_work/convolve')
> system("R CMD SHLIB convolve.c")
gcc -m32 -I"C:/PROGRA~1/R/R-31~1.2/include" -DNDEBUG -I"d:/RCompile/CRANpkg/extralibs64/local/include" -O3 -Wall -std=gnu99 -mtune=core2 -c convolve.c -o convolve.o
gcc -m32 -shared -s -static-libgcc -o convolve.dll tmp.def convolve.o
-Ld:/RCompile/CRANpkg/extralibs64/local/lib/i386-Ld:/RCompile/CRANpkg/extralibs64 /local/lib -LC:/PROGRA~1/R/R-31~1.2/bin/i386 –lR
```

C. GPU Accelerate

CUDA is a GPU a new general computing platform, will use the GPU computing threshold decreased very much. And GPU prices kept falling, and constantly improve the performance. The CUDA can be thought of as an extension to C, so the harder also is not big. Grammar is not a problem, the algorithm is the problem.

If your computer is N CARDS, general is to support CUDA operation, openGL platform is supported by A card. R use GPU computing rely mainly on gputools the package [6]:

V. SUMMARY

As a kind of resource, data has been the third largest resources outside as human and material; As an industry, a data-centric service industry has become the main way of economic transformation in the developed countries; As a science, data science also provides a boundless space for research and innovation.

R is an open source software programming language, is widely used in data mining, statistical analysis and data visualization, R combination with Hadoop bring big data processing new era.

ACKNOWLEDGMENTS

This work is supported by Natural Science Foundation of China National(No: 61340037, 61202006).

REFERENCES

[1] Viktor Mayer-Schönberger, Keneth Cukie. Big Data:A Revolution That Will Transform How We Live, Work, and Think [M].Cheers Publishing,2012.

[2] Julie Steele & Noah Lliinsky. Beautiful Visualization[M].O'Reilly Media, Inc.Publishing,2011.

[3] Winston Cbang.R Graphics Cookbook[M].O'Reilly Media, Inc.Publishing,2015.

[4] Xue Yi AND Chen Li-ping. Statistical modeling and R software [M]. Beijing: Tsinghua university press,2007.

[5] http://baike.baidu.com/view/1488597.htm

[6] http://brainarray.mbni.med.umich.edu/Brainarray/Rgpgpu/