The `file.frame` Package

Bryan W. Lewis
blewis@illposed.net

October 10, 2011

1 Preface

I've been working with some data from large text files of values in columns separated by columns (CSV) format. The files are gigabytes in size and have about 20 million rows each. My computer has enough memory for R to load the data. But it takes a while and I can be pretty impatient.

Now, I don't really need the entire data set in memory. I really just need to filter the data a bit and then sample from the rows. Is there a faster way to get to the data I need?

The `file.frame` package lets me quickly and efficiently work with subsets from a text file without loading the entire file into memory. A "file frame" presents a text file as a kind of simple data frame, but directly without first loading the text file into memory. File frames lazily load data from their backing files only when required, for example by an indexing operation. They are essentially a lazy wrapper for the `read.table` function with a few extra convenience features.

There are several compelling R packages for working directly with file-backed data (sometimes called "out of core" data): The bigmemory package by Emmerson and Kane provides a memory mapped matrix object free of R indexing constraints along with a comprehensive suite of fast analysis functions. The very straightforward to use mmap package by Jeff Ryan defines a data frame-like memory mapped object. And the ff package by Adler, Oehlschleğel, et. al. defines a variety of memory mapped data frame-like objects and functions. All of these packages have really interesting features. Most of them are designed to facilitate working with objects larger than the physical RAM available on a computer.

But, my data sets fit into the RAM on my computer (RAM is really cheap)! My main problem is the bottleneck of reading the entire data set in at once. And to my knowledge, none of the available packages for R address that problem directly.

Note: don't use file frames for small data files–just read the whole data file into memory in that case. File frames seem useful for data files with millions or tens of millions of rows. For *r*eally large data sets, `bigmemory` is a good option.

2 Using file.frame package