Dynamic Documents with R and knitr

Gene Dan

github.com/genedan
genedan.com

February 3, 2014

Literate Programming

- Literate programming, conceived by Don Knuth, is a programming paradigm that combines numeric output with documentation.

- Literate programming can be implemented via markup language (HTML, LaTeX, Markdown, etc.), which describes the logic and output of a program written in a general-purpose langauge (C++, Python, Perl, etc.). The program code is embedded within the markup language in a text file. This file can then be compiled into a variety of different formats (.pdf, web pages, slideshows, etc.) for visual display.

- Today I'll demonstrate one implemntation of literate programming, using LaTeX as our markup language to describe programs written in R. **knitr**, an R package written by Yihui Xie, is used to translate and compile the code.

LaTeX

LaTeX is a markup language that is popular amongst academics and technical writers for its ability to display mathematical notation:

$$DB = \frac{1}{n}\sum_{i=1}^{n}\max_{i\neq j}\left(\frac{\sigma_i+\sigma_j}{d(c_i,c_j)}\right)$$

$$D = \min_{1\leq i\leq n}\left\{\min_{1\leq j\leq n, i\neq j}\left\{\frac{d(i,j)}{\max_{1\leq k\leq n} d'(k)}\right\}\right\}$$

Advantages

- LaTeXmakes it easy for people to convey complex mathematcial notation via the internet or in print.
- This allows people to collaborate over great distances via websites like Stack Exchange or Mathoverflow. This rapidly increases the pace of innovation and discovery.

Disadvantages

- LaTeX is just a markup language. While it's great at displaying text, it cannot perform complex calculations (well maybe it can if you're extremely creative).
- This limitation causes inexperienced users to produce research output using another language like R, then copying and pasting source code and images from an IDE into a .tex file before compiling. This method, in addition to being inefficient, breaks the dependencies between the output and its source.
- This in turn makes it harder for the reasearcher to effectively communicate their findings and make their reasearch understandable and reproducible by their peers.

R

R is an open source statistical programming language.

Advantages

- R is designed so that its objects closely resemble the mathematical objects that statisticians use. This makes it easy for statisticians to both learn the language and carry out their research on the computer.
- R's open source platform makes it easy for anyone to debug and contribute.

Disadvantages

- R has limited capability for formatting and exporting output.
- This makes creating books, documents, websites, slide shows, and other presentations labor-intensive and error-prone.
- Fortunately, we can use knitr to address these shortcomings.

knitr

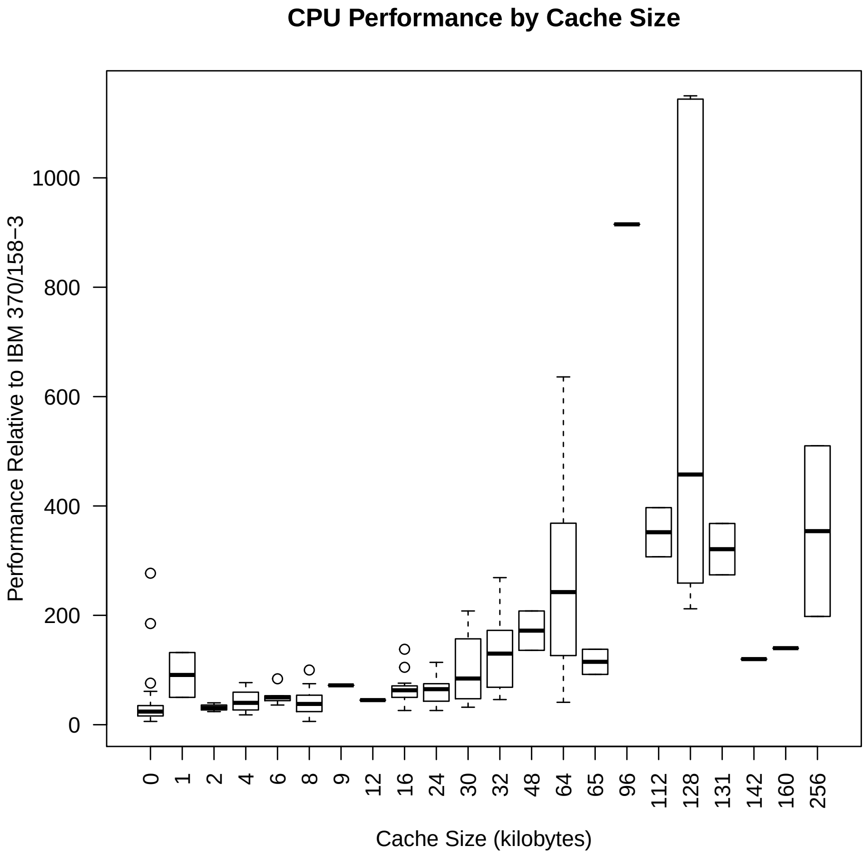
knitr is an R package developed by Yihui Xie that integrates LaTeX and R to produce dynamic documents. R is embedded within a LaTeX document via sections called **code chunks**:

```
library(MASS)
head(cpus[, c(1, 5, 8)])

##              name cach perf
## 1  ADVISOR 32/60  256  198
## 2  AMDAHL 470V/7   32  269
## 3  AMDAHL 470/7A   32  220
## 4 AMDAHL 470V/7B   32  172
## 5 AMDAHL 470V/7C   32  132
## 6  AMDAHL 470V/8   64  318
```

Code chunks that appear later in the document can reference objects that appear in previous code chunks. Here, we produce a figure that references cpus from the package MASS:

```r
plot(factor(cpus$cach), cpus$perf, las=2,
  main="CPU Performance by Cache Size",
  xlab="Cache Size (kilobytes)",
  ylab="Performance Relative to IBM 370/158-3")
```



Additional Languages

- R is not the only language! Other languages, such as Python, Perl, Ruby, and Bash can be embedded in code chunks. These chunks will be evaluated at compilation with output printed onto the document.
- In addition to LaTeX, knitr supports Markdown and HTML as markup languages.

Additional Formats

- PDFs are not the only documents that can be produced with knitr. You can use knitr to generate webpages and slide shows.

Additional Resources

Reproducible Research with R and RStudio - Christopher Gandrud

This is an excellent text on project management. Not only does it cover dynamic document generation via knitr, but also project integration and version control with git and github.

Dynamic Documents with R knitr - Yihui Xie

This is a book written by the creator of knitr. Covers both basic syntax and in-depth package configuration.