Reshape

Flexible data restructuring with R

Hadley Wickham
Statistics, Iowa State University

Introduction

- What is reshaping?

- Why it's not easy at the moment

- Example

- Details

- Future work

Aggregating

- Reduce big table to small table

- (must lose information)

- Each cell in the new table corresponds to multiple cells in the old table

Reshaping

- Like aggregating, but each new cell corresponds to one old cell

- Useful when investigating relationships between different aspects of your data (and especially when using lattice graphics)

- Similar to transposing a matrix (but what happens to ragged data?)

ftable table xtabs
tapply by merge
aggregate match
split reshape mapply

Which one do I use?

Motivation

- Many different tools in R

- Tend to be rather specialised/limited

- Can be difficult to figure out which one to use

- Reshape handles many different needs within one framework

Example

Conceptual framework

- Id vs. measured variables
  - random variables vs. their indices
  - categorical vs. continuous
- Gives a more flexible data format
- Deals with very ragged data
- Missing values implicit

Subject	Age	Height	Weight
John	20	1.95	100
John	21	1.96	NA

Subject	Age	Variable	Value
John	20	Height	1.95
John	20	Weight	100
John	21	Height	1.96

Mechanics

- Specifying the output format

- Adding margins

- Functions that return multiple values

- Row and column names

Output specification

- How do you specify what output you want?

- I've used a formula type interface

- Formatting output

- What are other alternatives?

Margins

- Multiple levels of aggregation at the same time

- Useful for summaries

- (Pivot table inspired)

Row & column names

- Explicit vs implicit

- Most inbuilt functions store implicitly (frustrating when trying to plot!)

- Reshape stores explicitly (but makes it easy to get rid of them)

Efficiency

- Size (limited by memory)
  - Multiple copies of data
  - 150,000 x 5
- Speed
  - merge is slow
  - RDMS?

Future

- Aggregate function on data frame

- Non-numeric data and summaries

- Built-in graphical display of output

- Larger data/database integration

http://had.co.nz/reshape