Performance Optimisation Strategies in R: Memoisation

Sebastian Warnholz – INWT Statistics GmbH

September 28th, 2017 – BRUG

INWT Statistics

What we do!

INWT Statistics GmbH

INWT Statistics specializes in intelligent data analysis and delivers solutions in the fields of online marketing, CRM, data management and business intelligence/reporting.

Online Marketing

- Customer Journey Analysis
- Conversion Optimization
- Fraud Detection

CRM

- Customer Lifetime Value
- Customer Segmentation
- Churn Management

BI/Reporting

- Data Management
- Data Consolidation
- Dashboards

Training

Selected Customers:

CosmosDirekt. Die Versicherung. OTTO SIXT rent a car tipico Webtrekk zalando

Memoisation

Wikipedia says:

In computing, [...] memoisation is an optimization technique used primarily to speed up computer programs by storing the results of expensive function calls and returning the cached result when the same inputs occur again.

INWT Statistics

Performance Optimisation (in R)

What is there to be optimised?

- The *run-time*
- CPU usage vs. more CPUs
- Memory usage vs. more Memory
- Run-time vs. development time

My personal experience: In data analysis, *memory* is the limiting dimensions.

INWT Statistics

Performance Optimisation

Options:

- Use `utils::Rprof` and identify bottlenecks
- Use vectorisation and be weary of loops
- Have code optimisation strategies in your tool box (e.g. memoisation)
- If you have a database, use it instead of R
- Use more hardware: parallel computing, HPC
- Use different language: C++, Julia, ...
- Use different framework: Spark, H2O, ...

Some helpful packages:

- `parallel`
- `BatchJobs`
- `Rcpp`
- `sparklyr`, `h2o`
- `data.table` and `dplyr`
- many more ...

INWT Statistics

Performance Optimisation

Redefine your problem!

- Reduce the amount of data: e.g. simple random sampling and LLN
- Value quality more than quantity
- Software Alchemy (https://arxiv.org/abs/1409.5827)
- 80/20 in statistical modelling
- Don’t underestimate the cost of using a new framework!

INWT Statistics

Live demo

When to use it?

Wikipedia says:

A function can only be memoized if it is referentially transparent; that is, only if calling the function has exactly the same effect as replacing that function call with its return value. (Special case exceptions to this restriction exist, however.)

INWT Statistics

When to use it?

Wikipedia says:

A function can only be memoized if it is referentially transparent; that is, only if calling the function has exactly the same effect as replacing that function call with its return value. (Special case exceptions to this restriction exist, however.)

What I do:

- Reduce amount of queries against database
- Reduce amount of calls to external APIs
- Avoid the reloading of local data

INWT Statistics

Packages

`R.cache`

- Bengtsson 2007 – 23 versions
- 10 reverse imports
- Not working with recursive functions
- The cache memory is persistent on the file system

`memoise`

- Wickham et.al. 2010 – 4 versions
- 17 reverse imports + 4 reverse depends
- Not working with recursive functions
- More features (invalidate cache, in-memory and on-disk)

INWT Statistics

Thank You for Your Attention!

INWT Statistics GmbH

E-Mail: info@inwt-statistics.de

Internet: `www.inwt-statistics.de`

GitHub: `www.github.com/wahani`