The Journal of Open Source Software

datastructures: An R package for organisation and storage of data

Simon Dirmeier[1]

1 Department of Biosystems Science and Engineering, ETH Zurich

DOI: 10.21105/joss.00910

Software

- Review
- Repository
- Archive

Submitted: 20 August 2018
Published: 27 August 2018

Summary

In computer science, data structures are objects for organisation and storage of data. They are often a key for efficient computation. So far the statistical computing language R (R Core Team 2018) only supports simple data structures such as lists, matrices and data tables, whereas other data structures such as priority queues, stacks or hashmaps are not supported. These data structures are essential in many computer science and statistics problems, like graph algorithms, string analysis or scheduling. Single packages, for example `hashmap` (Russell 2017) or `dequer` (Schmidt 2017), have already been proposed, although none of them support a wider range of data structures with a unified interface.

In order to access a richer repertoire of data types, and thus enabling efficient computation of various problems, we propose the R package datastructures. The package uses Boost and STL data types in a C++ backend and exports these to R using the Rcpp (Eddelbuettel and François 2011) module system.

So far `datastructures` supports three different groups of data types. **Heaps** Fibonacci and binomial heaps can be used for priority queue operations, such as in (a version of) Dijkstra's algorithm for finding shortest paths. **Maps** Hash-, bi- and multimaps are associative arrays that establish key-value relationships. **Deques** Stacks and queues are linked lists supporting the LIFO or FIFO principle, respectively.

We used an object-oriented approach where every data structure is exported using `S4` classes. For easier maintainance and extensibility we heavily rely ond polymorphism and inheritance of the implemented class system.

The package `datastructures` aims to bridge the gap between statistical and *classical* programming languages, where usage of these data structures are fairly common. In addition, in the near future we will add support of formats such as *suffix arrays* and *suffix trees*, *red-black-* and *B*-trees.

Examples

The following example shows a use case of a `hashmap` for an artificial data set.

```
> library(datastructures)


> hm <- hashmap("integer")
>   keys <- 1:2
```

```
>   values <- list(
     environment(),
     data.frame(A=rbeta(3, .5, .5), B=rgamma(3, 1)))
>   hm[keys] <- values

> hm[2L]
[[1]]
          A         B
1 0.5649580 1.8091666
2 0.2313472 0.4522518
3 0.8533336 3.8463516
```

References

Eddelbuettel, Dirk, and Romain François. 2011. "Rcpp: Seamless R and C++ Integration." *Journal of Statistical Software* 40 (8):1–18. https://doi.org/10.18637/jss.v040.i08.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Russell, Nathan. 2017. *Hashmap: The Faster Hash Map*. https://CRAN.R-project.org/package=hashmap.

Schmidt, Drew. 2017. *dequer: Stacks, Queues, and Deques for R*. https://cran.r-project.org/package=dequer.