A Super Fast 'RSiteSearch'

by Spencer Graves and Sundar Dorai-Raj

The RSiteSearch package provides a means to quickly and flexibly search the help pages of contributed packages, finding functions and datasets in seconds or a few minutes that could not be found in hours or days by any other means we know.

The results are returned in a data.frame of class RSiteSearch. Other R functions can then be used to quickly find what you want among possibly hundreds or thousands of hits.

Two examples are considered below: First we find a dataset containing a variable Petal.Length. Second, we find packages with spline capabilities, including looking for a function named spline.

Petal.Length

For example, a document discussing R provides an example using a variable Petal.Length from a famous Fisher data set but without naming the dataset nor where it can be found nor even if it exists in R.

```
> help.search('Petal.Length')
No help files found ...
```

The following retrieved 2 pages of up to 20 hits each:

```
> PL <- RSiteSearch.function('Petal.Length')
```

The summary.RSiteSearch method returns the number of hits, max(Score), and sum(Score) by Package:

```
> summary(PL)

Total number of hits: 23
Number of links downloaded: 23

Packages with at least 1 hit
using search pattern 'Petal.Length':
          Count MaxScore TotalScore
yaImpute      8        1          8
<...>
datasets      1        2          2
<...>
```

One of the listed packages is datasets. Since it's part of the default R distribution, we decide to look there first. We can select that row of PL just like we would select a row from any other data.frame:

```
> PL[PL$Package=='datasets', 'Function']
[1] iris
```

Problem solved in less than a minute!

spline

A couple of years ago, I decided I wanted to learn more about splines. I started my literature search as follows:

```
RSiteSearch('spline')
```

While preparing this manuscript, this command identified 1526 documents. That is too much, so I restricted it to functions:

```
RSiteSearch('spline', 'fun')
```

This identified only 631. That's an improvement over 1526 but is too much. To get a quick overview of these 631, we can proceed as follows:

```
splinePacs <- RSiteSearch.function('spline')
```

This downloaded a summary of the 200 highest-scoring help pages in the 'RSiteSearch' data base in roughly 5-10 seconds, depending on the speed of the Internet connection. To get all 631 hits, increase maxPages:

```
splineAll <- RSiteSearch.function('spline',
                                  maxPages=999)
```

To find a function named spline from this, we can proceed as follows:

```
selSpl <- (splineAll[,'Function']=='spline')
splineAll[selSpl, ]
```

This has 0 rows, because there is no help page named spline.

We can expand this to include any help page containing spline in the name using grep:

```
> fns <- tolower(splineAll[,'Function'])
> select <- grep('spline', fns)
> splineAll[select, c(1, 4, 5, 7)]
  Count Package             Function Score
31   34  assist              lspline     1
35   30     fda create.bspline.basis    48
<...>
```

This identified 66 help pages, the first of which is 'lspline' in the 'assist' package. The RSiteSearch engine assigned it a Score of 1. Evidently, that search engine found only minimal evidence of its relevance to the requested search string. It appeared at the top of this list, because the assist package had 34 help pages identifed as potentially relevant to that search string.

To establish priorities among different packages for further study, it might be nice to have a Pareto of the 10 packages with the most help pages relevant to our serach string. We can get this as follows:

```
> spSm <- attr(splineAll,'PackageSummary')
> spSm[1:10,'Count']
    assist     fda            gss       mgcv
        34      30             25         22
      VGAM kernlab DierckxSpline bayesSurv
        17      17             16         16
smoothSurv splines
        15      14
```

To obtain a similar Pareto by 'TotalScore' requires a little more effort:

```
> o <- rev(order(spSm[, 'TotalScore']))
> splineSum[o, ][1:10, ]
              Count MaxScore TotalScore
gss              25       35        448
splines          14       45        354
fda              30       48        275
<...>
```

This analysis gave us in seconds a very informative overview of `spline` capabilities in contributed R packages in a way that can help establish priorities for further study of the different packages and functions.

HTML

Example of use of HTML.RSiteSearch???

Summary

In sum, we have found the `RSiteSearch.function` in the `RSiteSearch` package to be a very quick and efficient method for finding things in contributed packages when the obvious alternatives would have taken more effort.