Visualising a web site with tag clouds generated by R

Sigbert Klinke1,2,*

1. Humboldt-Universität zu Berlin, School of Business and Economics, Institute of Statistics and Econometrics, Spandauer Strasse 1, D-10718 Berlin, Germany
2. Johannes Gutenberg University Mainz, Dept. of Law and Economics, Chair of Business and Human Resource Education, Jakob-Welder-Weg 9, D-55099 Mainz, Germany
* Contact author: sigbert@wiwi.hu-berlin.de

Keywords: igraph, network, page rank, visualisation, Wikipedia

The Wikipedia is the first source for a lot of users to gather information about a specific topic. To get an overview about a topic the user needs to follow a number of links to various pages in the Wikipedia. To visualise the link structure between pages, outbound and inbound, would help the users to cover a topic more easily.

The Wikipedia itself allows the categorisation of pages. Each page may belong to at least one category which reflects the topic and classes that are directly related to the subject of the page (Wikipedia, 2009). For example, the article about Student’s t-test belongs to the categories Statistical tests, Statistical methods and Parametric statistics. It is possible to build hierarchies of categories, for example all three categories are part of the category Statistics.

In the German Wikipedia, the category Statistics consists of approximately 500 pages and only 14 sub-categories, in the English Wikipedia the category Statistics consists of 8 pages and 54 sub-categories. It is obvious that the categories, as hand-made by user, may not provide an easy way to get an overview about a topic.

Search engines, such as Google, use, amongst other things, the link structure between pages to measure the importance and the closeness of pages. Based on all the links between pages (unidirectional: inbound, outbound and bidirectional) in one category, we generate a distance matrix for the pages. Using multidimensional metric scaling we determine the position of the page and its direct neighbours in a two-dimensional space. The page rank (Page and Brin 1998) of each page gives us the importance of each page. The R package igraph (Csardi 2008) supports the generation of the network (page positions and page importance).

For each page in the German Wikipedia, in the category Statistics a tag cloud with the page names will be generated. The position of the page names are determined by the multidimensional scaling and the font size by the page rank.

<table>
<thead>
<tr>
<th>Ähnlichkeitsmaß</th>
<th>Kategorie: Statistik</th>
<th>Wahrscheinlichkeitsrechnung</th>
</tr>
</thead>
<tbody>
<tr>
<td>Distanzmaß</td>
<td>Maßtheorie</td>
<td>Gegenwahrscheinlichkeit</td>
</tr>
<tr>
<td>Wahrscheinlichkeitsfunktion</td>
<td>Maßraum</td>
<td>Wahrscheinlichkeitsstheorie</td>
</tr>
<tr>
<td>Satz von Cramer-Wold</td>
<td>Fast-Überall</td>
<td></td>
</tr>
</tbody>
</table>

Figure 1: Tagcloud for “Maßtheorie”. Note that only links to pages which belong to the category Statistics are included in the tag cloud although many more pages link to and from the page “Maßtheorie”.

References


Csardi, G. (2008). Igraph: Routines for simple graphs, network analysis (Online; accessed 26-Feb-09), [http://cran.r-project.org/web/packages/igraph](http://cran.r-project.org/web/packages/igraph)