

Visualise a web site with tag clouds generated by R

Sigbert Klinke^{1,2}

¹ Institute for Statistics and Econometrics, School of Business and Economics, Humboldt-Universität zu Berlin

² Business and Human Resource Education, Dept. of Law and Economics, Johannes-Gutenberg-Universität Mainz



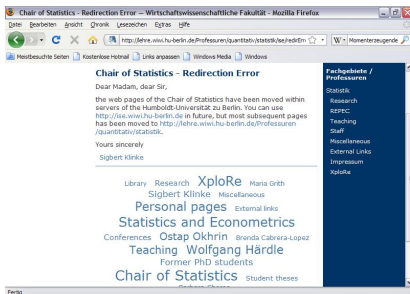
useR! 2009

Session: Textmining

08-10 Jul 2009, Rennes, France

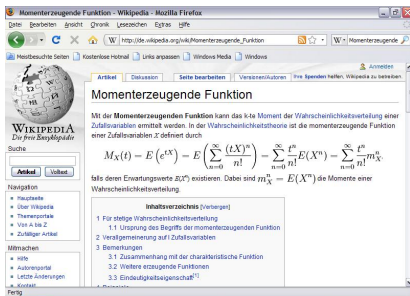


Problem: Redirection of web users



- Changes to web site structure produces errors on access
- How can we redirect the users to a large number of pages?
- Solution: Use a tag cloud where the size of an entry corresponds to the number of visits in the past year

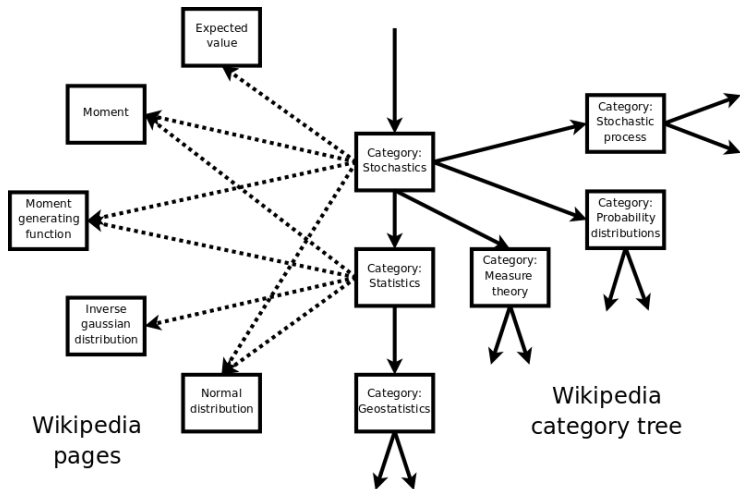
Problem: Teaching statistics



Links to *Moment*, *Wahrscheinlichkeitsverteilung*, ...

- Wikipedia is often a (starting) source for students
- Dictionary structure does not allow for an overview of a topic
- Solution: Use a tag cloud to visualise the neighbourhood of a page

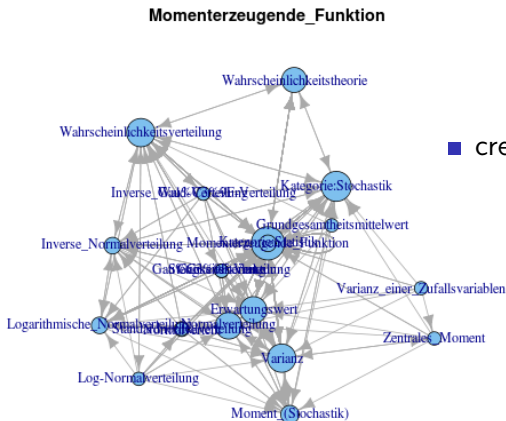
Wikipedia structure



Work flow

- PHP script crawls Wikipedia and stores the link structure
 - crawler from `http://w-shadow.com` using `cURL`
 - store in csv format: `fromPage ; toPage`
- R generates a tag cloud for each page
 - load linkstructure `read.csv`
 - build link network: `igraph` by Gabor Csardi
 - for importance compute `pagerank page.rank` (font size)
 - extract neighbourhood `graph.neighborhood` (of distance 1)
 - compute (bivariate) positions `layout.mds` (location)

igraph (layout.mds)



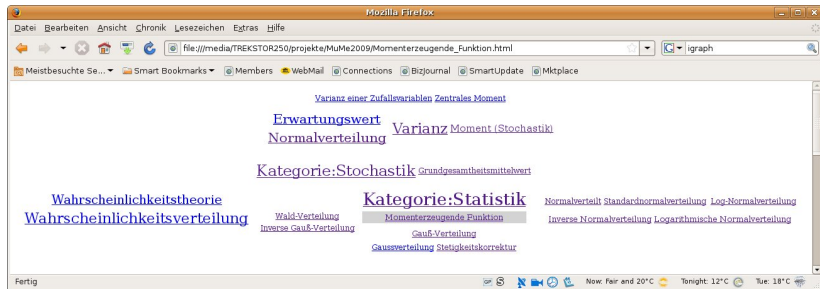
- create HTML tag clouds

- create dendrogram from positions (table-based)

- use a top/bottom - left/right approach (compact)

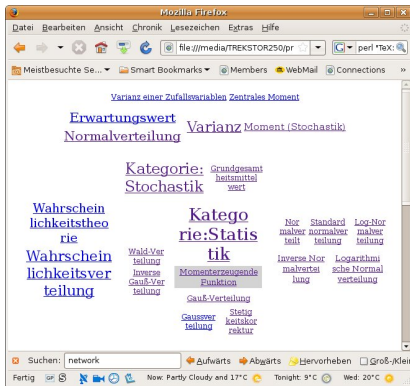
- use one dimensional MDS (onliner)

Tag cloud: table-based



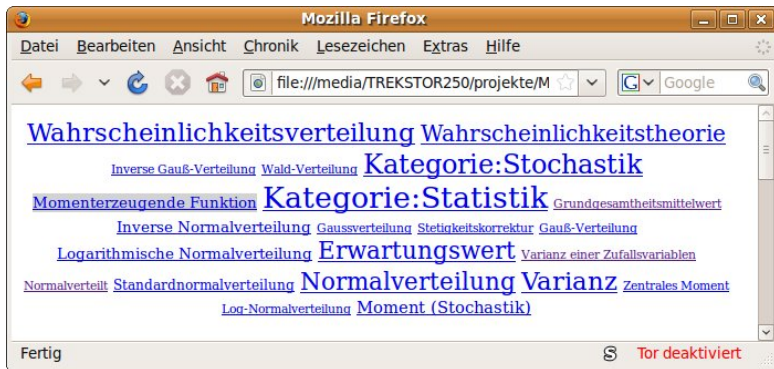
- Most page titles are long (e.g. Moment (mathematics))
- Take hyphenation into account

T_EX hyphenation



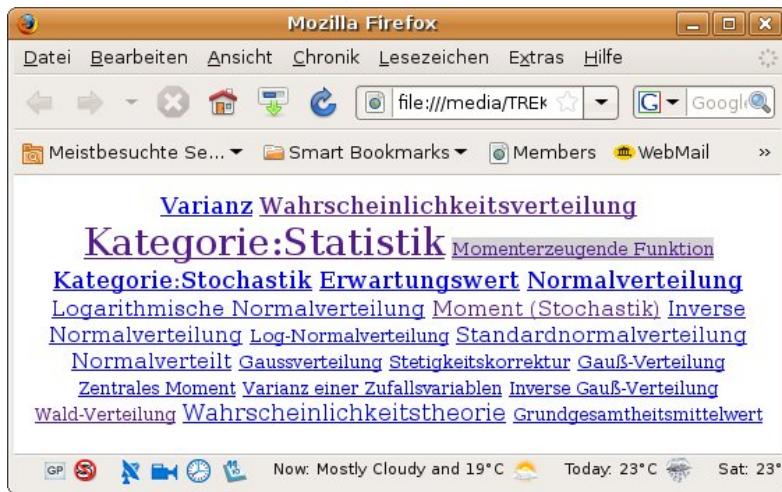
- utilise the T_EX hyphenation
- Perl program available
 - TeX : :hyphen by Jan Pazdziora
 - hyphen.pl with german hyphenation by Tilman Kranz
 - add ​ (zero width space)

Tag cloud: compact



- algorithm needs some more polishing

Tag cloud: one liner



createTagCloud parameters

<code>g</code>	igraph object
<code>graph.order</code>	size of neighbourhood (currently only 1)
<code>graph.layout</code>	layout function from igraph (layout.mds)
<code>fontsize.method</code>	method to compute the font size (page.rank.vector)
<code>fontsize.transform</code>	transformation method for font size (log10)
<code>fontsize.min</code>	font size minimum (7.5)
<code>fontsize.max</code>	font size maximum (20.5)
<code>buildHTML.method</code>	method to build tag cloud(s) (one)
<code>buildHTML.landscape</code>	landscape format (T)
<code>buildHTML.hyphenate</code>	should T _E X hyphenation be applied (TRUE)
<code>file.html</code>	name(s) of HTML/PNG file(s)
<code>file.png</code>	(vertex%i.html, vertex%i.png)
<code>no</code>	index of vertices for which tag clouds are generated (NA)
<code>...</code>	further parameters

Outlook

- Use Wikipedia XML dump instead own web crawler
- Account for redirects in Wikipedia
- Add “virtual” links
 - Analyse text (TreeTagger)
- Colour links in tag cloud (Inbound, Outbound, Bidirectional)
- Increase neighbourhood
- Add MediaWiki output
- Improve hyphenations?

Literature/Links

- Csardi, G. (2009): igraph,
<http://cran.r-project.org/web/packages/igraph>
- Kaser, O., Lemire, D. (2007): Tag-cloud Drawing: Algorithms for Cloud visualization, arXiv,
<http://arxiv.org/abs/cs/0703109>
- Kranz, T. (2009): hyphen.pl,
<http://tk-sls.de/texte/sil-ben-tren-nung.html>
- Liang, F.M. (1983): Word Hyphen-a-tion by Com-put-er, Stanford University, CA 94305, Report No. STAN-CS-83-977.
- Münz, S. et al. (2007): SELFHTML 8.1.2,
<http://de.selfhtml.org/>
- Pazdziora, J. (2002): TeX::Hyphen,
<http://search.cpan.org/dist/TeX-Hyphen>