

**Detaillierter Blick in eine  
Link-Datenbank**

Erfahrungen, Einblicke,  
Experimente

03/13/10

Fischerländer

- Erfahrungen mit dem Link-Graph der Suchmaschine Neomo
- Link-Datenbank
- Link-Algorithmen in Theorie und Praxis
  - HITS
  - Rank-and-File
  - Co-citation

- Im öffentlichen Blickfeld seit der Einführung des Google PageRank.
- Allerdings bereits viel früher im Einsatz:  
*„Lycos uses a best-first search based on the popularity heuristic. We define popularity as the number of external Web servers with at least one link.“*  
(„Lycos: Design choices in an Internet search service“, erschienen Januar 1997)
- Link-Analyse ist das A und O jeder guten Websuchmaschine

- NEOMO
  - März 2004: Prototyp online gestellt
  - Dezember 2004: Firmengründung in Köln (Joint-Venture mit zwei Telefonbuchverlagen)
  - April 2005: öffentliche Beta-Version
  - Mai 2005: wichtiger Geschäftspartner der Verlage erklärt, dass er mit dem Projekt unglücklich ist
  - 2005/2006: Neuausrichtung Suchlösungen für Geschäftskunden
  - Frühjahr 2007: Ende des Projekts, Verkauf der vorhandenen Assets

- Indexgröße: ca. 200 Mio. gecrawlte Seiten
- Unsere Link-Analyse:
  - „LinkValue“ =  $\text{LOG}(\text{IP-Popularity})$   
Diese leicht zu ermittelnde Größe stimmte in verblüffender Genauigkeit mit den bestbewerteten deutschen PageRank-Seiten überein: bahn.de, spiegel.de, google.de.  
Allerdings war diese einfache Berechnung anfällig für (unbewusste) Manipulationen.
  - „Seed-Distance“ = Link-Abstand von den Seed-URLs  
Je weiter weg eine URL vom manuell ausgewählten Seed liegt, um so unwichtiger ist. Algorithmus ähnelte damit dem TrustRank.
  - Linktexte = Pro IP nur ein Linktext berücksichtigt.

- Meist wurden die Toptreffer für allgemeine Anfragen gut gefunden.
- „hotel“
  - Link-Clique von Hotels im Bayerischen Wald weit vorne.
  - Effekt konnte durch „Seed-Distance“ etwas gemildert werden.
- „aldi“
  - Bug im Crawler: Im Linktext auch nachfolgende Wörter enthalten.
  - Ergebnis: 1. Aldi.de, 2. Lidl.de, 3. Norma.de

- Test-Datenbank
  - ca. 6 Mio. gecrawlte URLs
  - ca. 12 Mio. bekannte URLs
  - gut 40 Mio. externe Links inkl. Linktext
  - Seed-URLs: Open Directory
  - Crawl-Tiefe: 1 (ODP-URLs plus von dort direkt verlinkt)
- Datenbank ist also relativ klein ...

- nofollow: 3.829.272 von 41.732.405 = 9,2 Prozent (Laut seomoz/Linkscape: 3 Prozent / Stand Juni 2009)
- Seiten mit den meisten nofollow-Links:
  - [www.mister-wong.de](http://www.mister-wong.de)
  - [www.google.com](http://www.google.com)
  - [del.icio.us](http://del.icio.us)
  - [yigg.de](http://yigg.de)
  - [www.linkarena.com](http://www.linkarena.com)
  - [www.webnews.de](http://www.webnews.de)
  - [www.yahoo.com](http://www.yahoo.com)
  - [digg.com](http://digg.com)
  - [www.icio.de](http://www.icio.de)
  - [www.folkd.com](http://www.folkd.com)

- Hyperlink Induced Topic Search
- HITS ist ein rekursiver Algorithmus, der zu einer Query die besten *Hubs* und *Authorities* ermittelt.
  - Textsuche => ROOT SET
  - Dazu alle In- und Outlinks => BASE SET (nur externe Links)
  - Rekursive Berechnung:
    - $\text{Hubscore}(\text{URL}) = \text{Summe}(\text{Authorityscores}(\text{Outlinks}))$
    - $\text{Authscore}(\text{URL}) = \text{Summe}(\text{Hubscores}(\text{Inlinks}))$

- Authorities

- [news.perlfoundation.org](http://news.perlfoundation.org)
- [www.perl-community.de](http://www.perl-community.de)
- [perl-nachrichten.de](http://perl-nachrichten.de)
- [www.perl-workshop.de](http://www.perl-workshop.de)
- [perl-magazin.de/?issue=8;action=show\\_issue](http://perl-magazin.de/?issue=8;action=show_issue)

- Hubs

- [reeneb-perlblog.blogspot.com](http://reeneb-perlblog.blogspot.com)
- [www.perlfoundation.org/perl5/index.cgi?blogs](http://www.perlfoundation.org/perl5/index.cgi?blogs)
- [www.fabiani.net/links.shtml](http://www.fabiani.net/links.shtml)
- [de.wikipedia.org/wiki/Perl\\_\(Programmiersprache\)](http://de.wikipedia.org/wiki/Perl_(Programmiersprache))
- [perl-howto.de/cgi-bin/mt/mt-search.pl?tag=2038&blog\\_id=1](http://perl-howto.de/cgi-bin/mt/mt-search.pl?tag=2038&blog_id=1)

- Authorities:
  - [www.biathlonworld.com](http://www.biathlonworld.com)
  - [www.biathlon-online.de](http://www.biathlon-online.de)
  - [www.biathlon2b.com](http://www.biathlon2b.com)
  - [www.quotenmeter.de/index.php?newsid=17918](http://www.quotenmeter.de/index.php?newsid=17918)
  - [www.dwdl.de/article/news\\_13755,00.html](http://www.dwdl.de/article/news_13755,00.html)
- Hubs:
  - [de.wikipedia.org/wiki/Biathlon](http://de.wikipedia.org/wiki/Biathlon)
  - [hsv-ski.asv-ski.de/content.php?folder=749](http://hsv-ski.asv-ski.de/content.php?folder=749)
  - [magdalena-neuner.de/22-Links.htm](http://magdalena-neuner.de/22-Links.htm)
  - [forum.biathlon-online.de](http://forum.biathlon-online.de)
  - [www.markhausen.de/sectionview193.html](http://www.markhausen.de/sectionview193.html)

- Authorities:
  - [rezepte-pur.lecker.de](http://rezepte-pur.lecker.de)
  - [hobbykoch.blog.de](http://hobbykoch.blog.de)
  - [www.zottarella.de/de/blog/](http://www.zottarella.de/de/blog/)
  - [www.larissatoday.de](http://www.larissatoday.de)
  - [www.wunderweib.de/kochenundgenuss/rezeptesuchen/rubrik-rezeptesuchen/](http://www.wunderweib.de/kochenundgenuss/rezeptesuchen/rubrik-rezeptesuchen/)
- Hubs:
  - [woche-heute.wunderweib.de](http://woche-heute.wunderweib.de)
  - [freizeitwoche.wunderweib.de](http://freizeitwoche.wunderweib.de)
  - [alles-fuer-die-frau.wunderweib.de](http://alles-fuer-die-frau.wunderweib.de)
  - [tina.wunderweib.de](http://tina.wunderweib.de)
  - [auf-einen-blick.wunderweib.de](http://auf-einen-blick.wunderweib.de)

- Vorteile
  - Liefert Hubs und Authorities
    - zwei Scores pro URL
    - ermöglicht vielfältigere Ergebnislisten
  - Erweiterung: disjunkte Themen erkennen
- Nachteile
  - Reagiert sensibel auf Link-Cliquen
  - Anfällig für Manipulationen
  - Feintuning sehr wichtig
  - Rekursive Berechnung ist aufwendig

- Rank-and-File ist eine Vereinfachung von HITS
  - Textsuche => ROOT SET
  - Erweitern um Outlinks, in deren Nähe die Query vorkommt
  - Zähle, wie häufig jeder Outlink vorkommt

- [www.perl.com](http://www.perl.com)
- [www.perl.org](http://www.perl.org)
- [www.vitinh.de](http://www.vitinh.de)
- [www.perl-community.de](http://www.perl-community.de)
- [www.perl-workshop.de](http://www.perl-workshop.de)
- [www.kalyxo.de/category/technologie-und-das-web/perl-p...](http://www.kalyxo.de/category/technologie-und-das-web/perl-p...)
- [board.perl-community.de](http://board.perl-community.de)
- [renee-perlblog.blogspot.com](http://renee-perlblog.blogspot.com)
- [www.perl-workshop.de/de/2009/](http://www.perl-workshop.de/de/2009/)
- [strawberryperl.com](http://strawberryperl.com)

- [www.biathlon-online.de](http://www.biathlon-online.de)
- [olympia.ard.de/olympia/sportarten/biathlon/biathlonst...](http://olympia.ard.de/olympia/sportarten/biathlon/biathlonst...)
- [www.spiegel.de/sport/wintersport/0,1518,680655,00.htm](http://www.spiegel.de/sport/wintersport/0,1518,680655,00.htm)
- [www.biathlonworld.com](http://www.biathlonworld.com)
- [www.biathlon-antholz.it](http://www.biathlon-antholz.it)
- [www.spiegel.de/fotostrecke/fotostrecke-52338.html](http://www.spiegel.de/fotostrecke/fotostrecke-52338.html)
- [www.biathlon-ruhpoling.de](http://www.biathlon-ruhpoling.de)
- [www.biathlon-obertilliach.com](http://www.biathlon-obertilliach.com)
- [www.focus.de/sport/olympia-2010/top-meldung/biathlon-...](http://www.focus.de/sport/olympia-2010/top-meldung/biathlon-...)
- [www.dosb.de/de/start/details/news/biathlon\\_maenner\\_er...](http://www.dosb.de/de/start/details/news/biathlon_maenner_er...)

- [www.rezepte-nachkochen.de](http://www.rezepte-nachkochen.de)
- [www.rezepte-und-tipps.de](http://www.rezepte-und-tipps.de)
- [www.allgaeuer-rezepte.de](http://www.allgaeuer-rezepte.de)
- [www.rezepte-cocktails.de](http://www.rezepte-cocktails.de)
- [www.chefkoch.de](http://www.chefkoch.de)
- [www.vegetarische-rezepte.com](http://www.vegetarische-rezepte.com)
- [www.rezepte.li](http://www.rezepte.li)
- [www.rezepte.net](http://www.rezepte.net)
- [seo.de/3009/seo-rezepte-fur-brotchen-und-andere-backl...](http://seo.de/3009/seo-rezepte-fur-brotchen-und-andere-backl...)
- [rezepte.nit.at](http://rezepte.nit.at)

- [www.adobe.de/products/acrobat/readstep2.html](http://www.adobe.de/products/acrobat/readstep2.html)
- [www.adobe.com/de/products/acrobat/readstep2.html](http://www.adobe.com/de/products/acrobat/readstep2.html)
- [www.adobe.com/shockwave/download/index.cgi?Lang=Germa...](http://www.adobe.com/shockwave/download/index.cgi?Lang=Germa...)
- [get.adobe.com/de/flashplayer/](http://get.adobe.com/de/flashplayer/)
- [www.adobe.de/products/acrobat/readstep.html](http://www.adobe.de/products/acrobat/readstep.html)
- [get.adobe.com/de/reader/](http://get.adobe.com/de/reader/)
- [www.macromedia.com/go/getflashplayer](http://www.macromedia.com/go/getflashplayer)
- [www.adobe.com/go/getflashplayer](http://www.adobe.com/go/getflashplayer)
- [www.adobe.com/products/acrobat/readstep2.html](http://www.adobe.com/products/acrobat/readstep2.html)
- [www.adobe.com/shockwave/download/download.cgi?P1\\_Prod...](http://www.adobe.com/shockwave/download/download.cgi?P1_Prod...)

- Vorteile:
  - einfache und schnelle Berechnung (keine Rekursionen)
  - überraschend gute Ergebnisse mit wenig Daten
  - Beeinflussung durch Linkmanipulationen relativ einfach erkennbar

- Co-citation
  - Seite U verlinkt sowohl auf V als auch auf W. Dann nennt man V und W koziert.
  - Seiten, die häufig koziert werden, haben offenbar eine Gemeinsamkeit.

- [www.sistrix.de](http://www.sistrix.de)
- [www.abakus-internet-marketing.de](http://www.abakus-internet-marketing.de)
- [wordpress.org](http://wordpress.org)
- [www.mediadonis.net](http://www.mediadonis.net)
- [www.sistrix.de/news/](http://www.sistrix.de/news/)
- [www.inhouse-seo.de](http://www.inhouse-seo.de)
- [www.seo-campixx.de](http://www.seo-campixx.de)
- [www.seonauten.com](http://www.seonauten.com)
- [www.seomoz.org](http://www.seomoz.org)
- [www.seodeluxe.de](http://www.seodeluxe.de)

- [www.google.de](http://www.google.de)
- [www.yahoo.de](http://www.yahoo.de)
- [www.lycos.de](http://www.lycos.de)
- [www.fireball.de](http://www.fireball.de)
- [www.mister-wong.de](http://www.mister-wong.de)
- [www.web.de](http://www.web.de)
- [www.altavista.de](http://www.altavista.de)
- [www.google.com](http://www.google.com)
- [www.yahoo.com](http://www.yahoo.com)
- [www.alltheweb.com](http://www.alltheweb.com)

- [www.guenstiger.de](http://www.guenstiger.de)
- [www.preissuchmaschine.de](http://www.preissuchmaschine.de)
- [www.evendi.de](http://www.evendi.de)
- [www.ebay.de](http://www.ebay.de)
- [www.kelkoo.de](http://www.kelkoo.de)
- [www.kostenlos.de](http://www.kostenlos.de)
- [www.geizkragen.de](http://www.geizkragen.de)
- [www.ideal.de](http://www.ideal.de)
- [www.billiger.de](http://www.billiger.de)
- [www.google.de](http://www.google.de)

- [www.idealo.de](http://www.idealo.de)
- [www.welt.de](http://www.welt.de)
- [www.autobild.de](http://www.autobild.de)
- [www.abendblatt.de](http://www.abendblatt.de)
- [www.morgenpost.de](http://www.morgenpost.de)
- [www.sportbild.de](http://www.sportbild.de)
- [www.hoerzu.de](http://www.hoerzu.de)
- [www.bild.de](http://www.bild.de)
- [www.rollingstone.de](http://www.rollingstone.de)
- [www.immonet.de](http://www.immonet.de)

**Vielen Dank!**