Böööhze Links

Wie Suchmaschinen die Gumminudeln in der Linksuppe finden…

Prof. Dr. Mario Fischer
Fakultät für Informatik & Wirtschaftsinformatik
Schwerpunkt Electronic Commerce
Hochschule für angewandte Wissenschaft
University of Applied Sciences
Würzburg

7.-8. März 2009 - SEO-Campixx Berlin

Alle hier gezeigten Webseiten/Screenshots sind zufällig zur Demo ausgewählt.
Sie haben keinerlei Bezug zu den Aussagen der Folien!

Über welche Dimensionen sprechen wir?

Zumindest bei

- Index: Mehrere hundert Milliarden Dokumente
- eine Billion Links
- ca. 200 Datencenter weltweit, ca. 500.000 Server und 10 Exabyte Speichervolumen
- Speicherung aller gesetzten (und entfernten) Links
- Eine Suchanfrage beschäftigt mehrere hundert Server, braucht > 10 Mrd. Prozessorzylen und fragt > 100 MB Daten ab
  Antwortzeit: 0,2 Sekunden – incl. Realtime-Filter
- ca. 15.000 Suchabfragen weltweit / Sekunde
- Datenverarbeitung führt Google etwa zehn Mal effizienter / günstiger als andere IT-Unternehmen aus
- (interne) PageRank-Berechnung wichtiger Sites mehrmals pro Tag (näherungsweise)
- Geschätzter Zeitraum, bis Maschinen verstehen, was Menschen meinen…?

Warum werden Links gesetzt?

- Hinweis auf guten Content
- Gefälligkeit / Geschäftsbeziehungen
- Linktausch
- Linkkauf
- Affiliate
- Zum Behumsen des Rankings

Wie kann *gut* und *schlecht* (bööööhze) unterschieden werden?

- Hinweis auf guten Content
- Gefälligkeit / Geschäftsbeziehungen
- Linktausch
- Linkkauf
- Affiliate
- Zum Behumsen des Rankings

Easy to detect: Linktausch

A linkt auf B

B linkt auf A

Schwieriga to detect: alle anderen Link-Gründe

Kernfrage: Welches sind die Links, die wegen gutem Content gesetzt wurden?

Unnütze Links: Herkömmliche Methoden

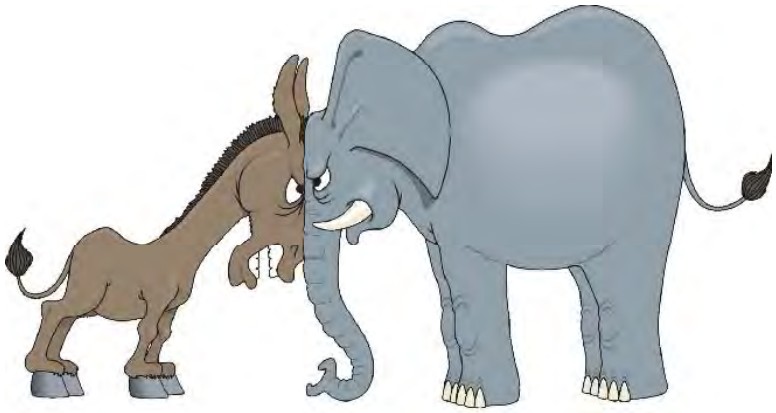
Kernfrage:
Wer ist mit wem verwand?

- Gleicher IP D-Block
- Gleicher Admin C
- Gleiche Faxnummer
- Identische Google-Accounts (Webmastertools, Analytics)
- Gleicher Adsense-Code
- Gleiche Adresse im Impressum
- …

Unnütze Links: Weitere Methoden

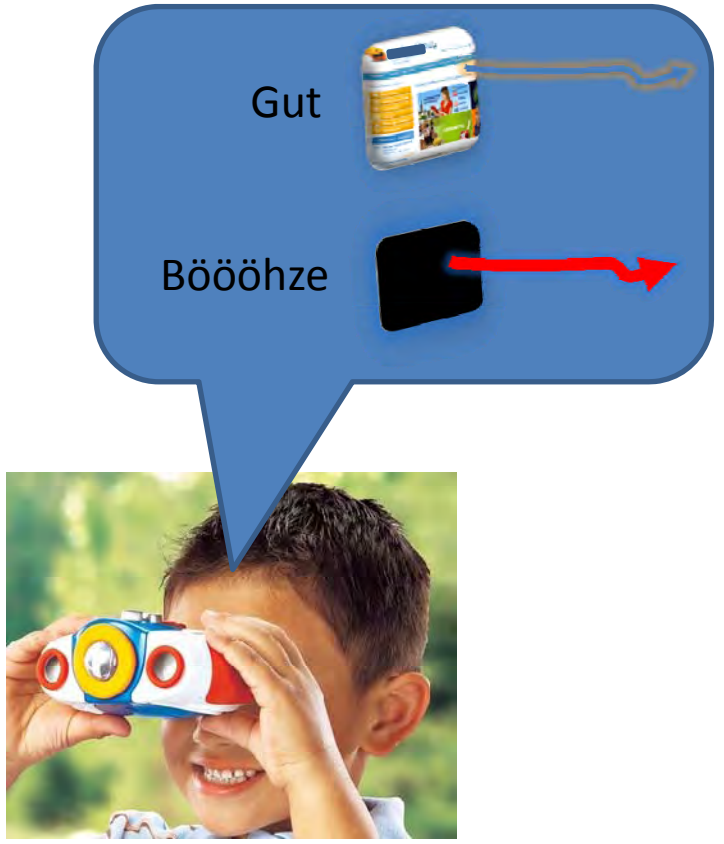
- Aufstellen von sog. Blacklists ✗
- Einsatz von Heuristiken ✗
- Nachfolgende Datenanalyse ✓

Problem:
So leicht wie es sich anhört,
ist es bei weitem nicht…



Maschinelles Lernen

- Mustererkennung
  - Automatisches Training
  - Gesteuertes Training



Maschinelles Lernen: Mustererkennung Beispiel

Maschinelles Lernen: Mustererkennung Beispiel

Im Prüfset von Domains stellt der selbstlernende Algorithmus z. B. fest:

		PenaltyPoints
1	Über 2 Ausgehende Links am Seitenende. Auf >75% aller Seiten einer Domain	8
2	>60% aller eingehenden Links kommen aus einem D-Class-Netz	3
3	Die Linkmuster (extern) von >50% der anlinkenden Domains sind zu 80% identisch	5
4	>40% der anlinkenden Domains sind direkt untereinander verlinkt	3
5	>60% der anlinkenden Seiten haben andere Keywordtopics als die verlinkten Seiten	7
6	Es fehlen Links von vertrauenswürdigen Domains	7
7	Am Linknetzwerk hängen Sites mit bestimmten Keywords wie z. B. „Suchmaschinenoptimierung“	5

38

Entscheidungstabelle:

ab 20 De-Ranking
ab 30 P4 Penalty
ab 40 tschüss

→ *Test, Feinjustierung, Collateral-Prüfung*

Alle Zahlen beispielhaft! Um Gottes willen!

Maschinelles Lernen: Erkennungsraten

Wie gut sind die Maschinen heute?

Wahrscheinlichkeit, eine Linkfarm zu entdecken: **95,1%**

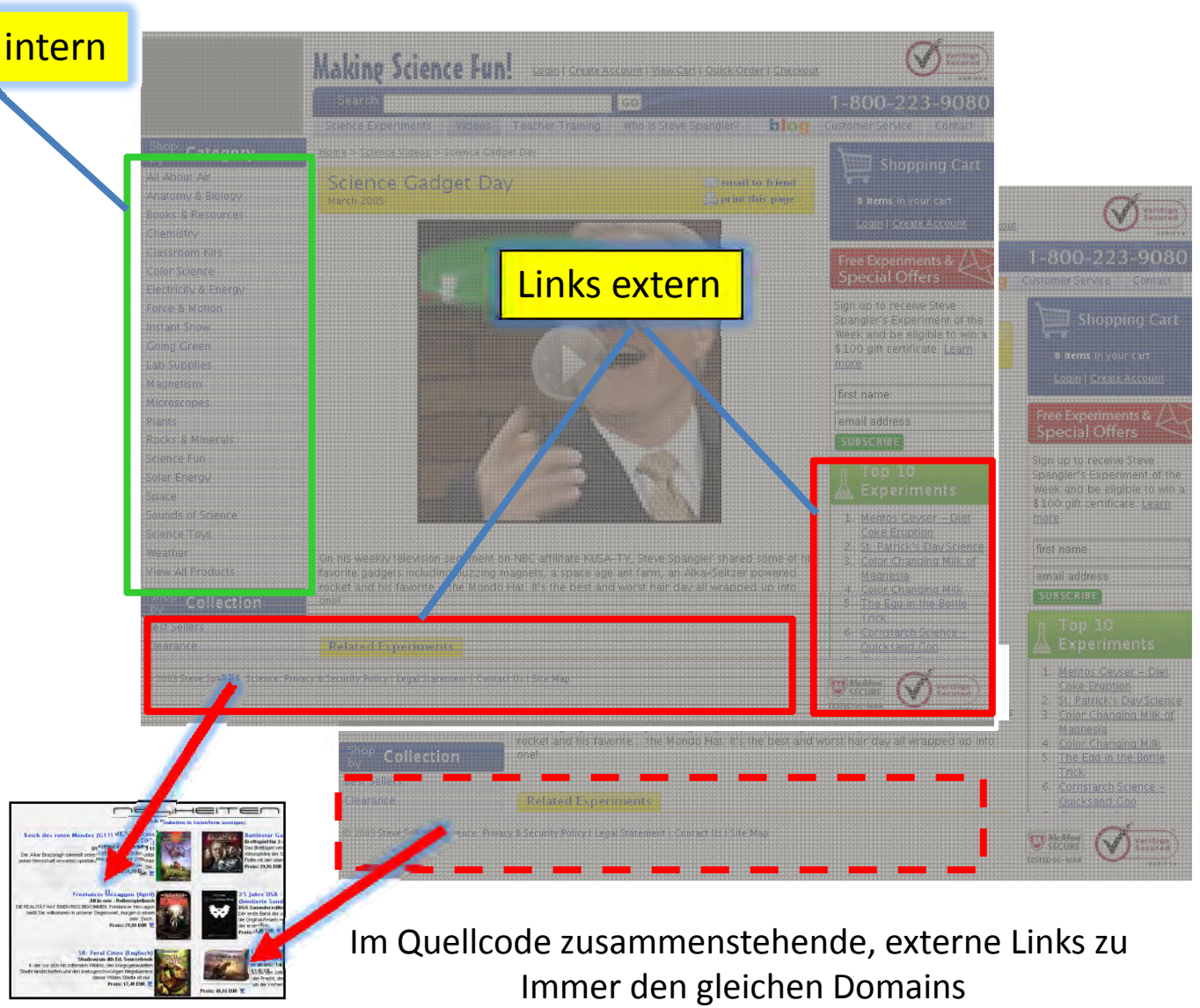
Wahrscheinlichkeit, einen Spam-Linkhub (Kern) zu entdecken: **~80%**

Maschinelles Lernen: Mustererkennung Beispiel

Sites erkennen, die Links verkaufen:

Mögliche Kriterien:

- Linkverhältnis intern/extern Domain und/oder Seite
- Wechselnde Linktexte bei gleichem Linkziel
- Linktextwechsel über die Zeit
- Verhältnis Deeplinks zu Startseitenlinks
- Anzeichen von PR-Sculpting
- Fehlen von Domainnamen in den Linktexten
- Externe Links ohne inhaltlichen Zusammenhang zur Site
- Unterschiedliche Topics der extern verlinkten Sites
- Werthaltige/“verdächtige“ Keywords auf den Linktexten
- Alter des Linkbestehens
- Linkwechsel in bestimmten Intervallen (z. B. monatlich)
- Keine/wenig Intext-Links
- Nicht natürlich wirkende Geschwindigkeit beim Linkaufbau



Im Quellcode zusammenstehende, externe Links zu Immer den gleichen Domains

Maschinelles Lernen: Mustererkennung Beispiel

Entscheidungsbaum des C4.5 Algorithmus von Brian D. Davidson:

```
FromPath contains tilde = 0:
|   From contains Simple gTLD = 0:
|   |   Domains are same = 0: -1 (35.0)
|   |   Domains are same = 1: 1 (30.0/1.0)
|   From contains Simple gTLD = 1:
|   |   FromPage has > 200 links = 1: 1 (812.0)
|   |   FromPage has > 200 links = 0:
|   |   |   Same DNS servers = 0:
|   |   |   |   Same contact email = 1: 1 (5.0)
|   |   |   |   Same contact email = 0:
|   |   |   |   |   Pages share > 10% links = 0: -1 (62.0/7.0)
|   |   |   |   |   Pages share > 10% links = 1: 1 (3.0/1.0)
|   |   |   Same DNS servers = 1:
|   |   |   |   Complete hostnames are same = 1: 1 (374.0)
|   |   |   |   Complete hostnames are same = 0:
|   |   |   |   |   Domains are same = 0: 1 (68.0/1.0)
|   |   |   |   |   Domains are same = 1:
|   |   |   |   |   |   Pages share > 20% links = 1: 1 (18.0)
|   |   |   |   |   |   Pages share > 20% links = 0:
|   |   |   |   |   |   |   FromPage has > 100 links = 1: 1 (14.0)
|   |   |   |   |   |   |   FromPage has > 100 links = 0:
|   |   |   |   |   |   |   |   FromPage has <= 5 links = 1: 1 (2.0)
|   |   |   |   |   |   |   |   FromPage has <= 5 links = 0:
|   |   |   |   |   |   |   |   |   1 path components are same = 1: -1 (5.0/2.0)
|   |   |   |   |   |   |   |   |   1 path components are same = 0:
|   |   |   |   |   |   |   |   |   |   FromPage has <= 10 links = 0: 1 (21.0/6.0)
|   |   |   |   |   |   |   |   |   |   FromPage has <= 10 links = 1: -1 (5.0/2.0)
FromPath contains tilde = 1:
|   1 path components are same = 0: -1 (56.0)
|   1 path components are same = 1: 1 (26.0)
```

Quelle: www.cse.lehigh.edu/~brian/pubs/2000/aaaiws/aaai2000ws.pdf, S. 3

Maschinelles Lernen: Mustererkennung Beispiel

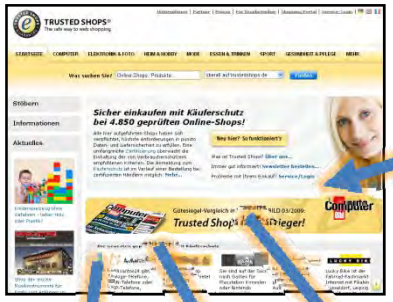
Spammer sind Menschen und Menschen machen Fehler: Footprints

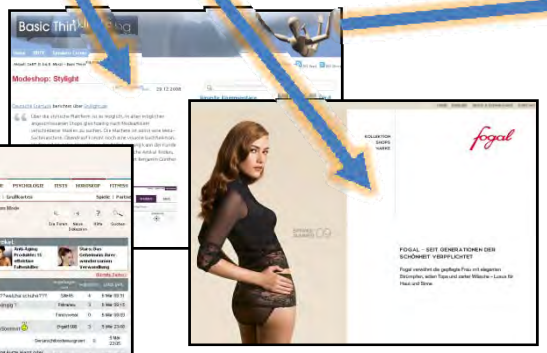
- Identische Titel/Description
- Vergessene Comments im Quelltext
- Im Pfad stehen gleiche Begriffe wie /user, /home /~ oder ähnliche, identische Zeichenketten (gleiche Software!)
- Gleiche E-Mail-Adresslinks
- Domain A hat > x% identische Ausgangslinks wie Domain B

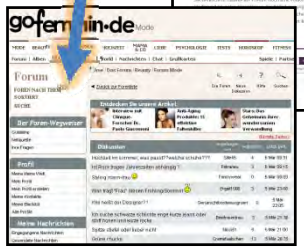
```
<div id=showdiv style='visibility:hidden;position:absolute;' onclick='showDiv();'>
<!-- insert your hidden text here. Do not forget to <H1></H1> your keywords -->
<h1>Music box</h1> - A mechanical music device, can be disc music box or cylinder
<!-- end insert here -->
```

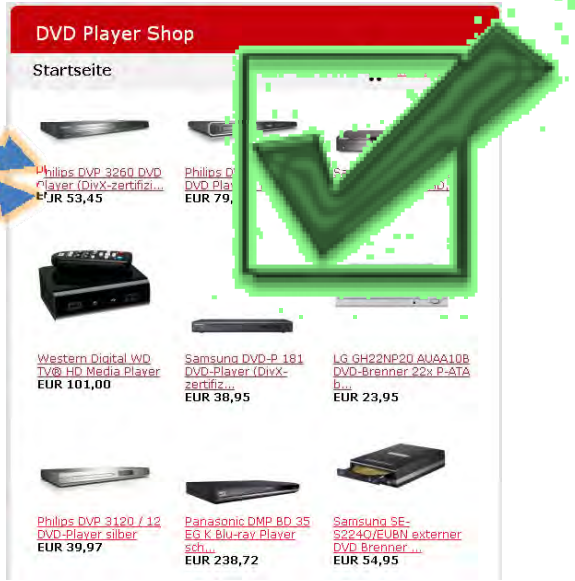
Maschinelles Lernen: Positivmuster erkennen

Set vertrauenswürdiger Sites









Inbound-Linkmuster:

- Trustrank 6,4
- 1.250 Links, davon
  - 920 diff. Domains
  - 633 diff. IP-Class
  - 540 mit Domainnamen

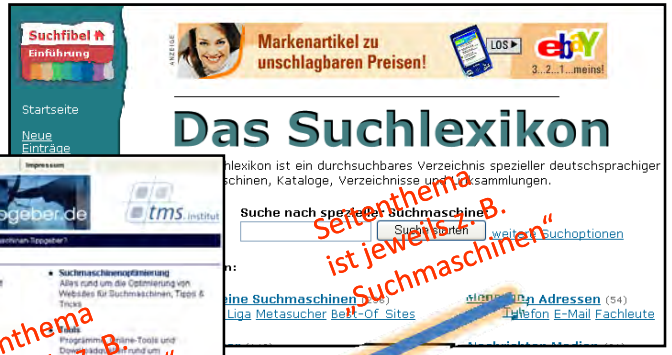
Maschinelles Lernen: Semantisch: Hilltop / HITS

Identifikation von Expertenseiten (Hubs) und Autoritäten (Authorities)

ab PR 4

„Experten“

müssen in den Top 200 sein und zu mindesten 5 **unabhängigen***, themenrelevanten Seiten verweisen



ab PR 4

Seitenthema ist jeweils z.B. „Suchmaschinen“

Seitenthema ist jeweils z.B. „Suchmaschinen“

Seitenthema ist jeweils z.B. „Suchmaschinen“

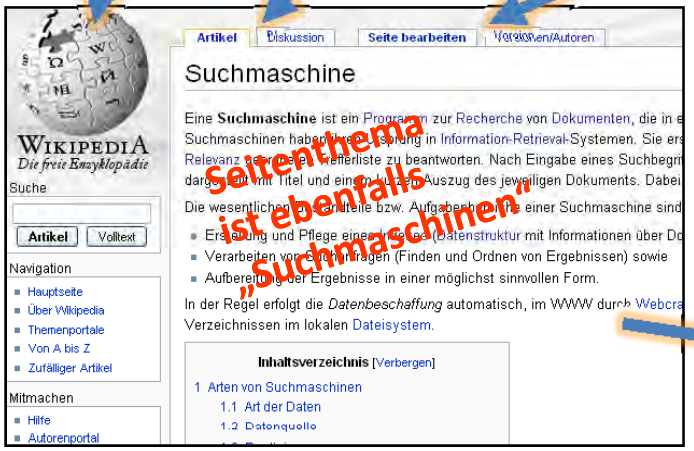
ab PR 4

„Autorität“

>2 **unabhängige*** „Experten“-Links

*nach der IP-Nummer und Domainnamen!

Kandidat für die Top 10



Seitenthema ist ebenfalls „Suchmaschinen“

Berechnung eines **Relevanz-wertes**, der vom Verlinkungsgrad der Expertenseiten abhängt.

mögl. wenig Links nach außen

Derzeit wahrscheinlich nur bei allgemeinen, werthaltigen Begriffen einsetzbar, da rechenintensiv.

Hilltop: Entwickelt von Krishna Bharat und George A. Mihaila, Google hat das Patent erworben ;-)

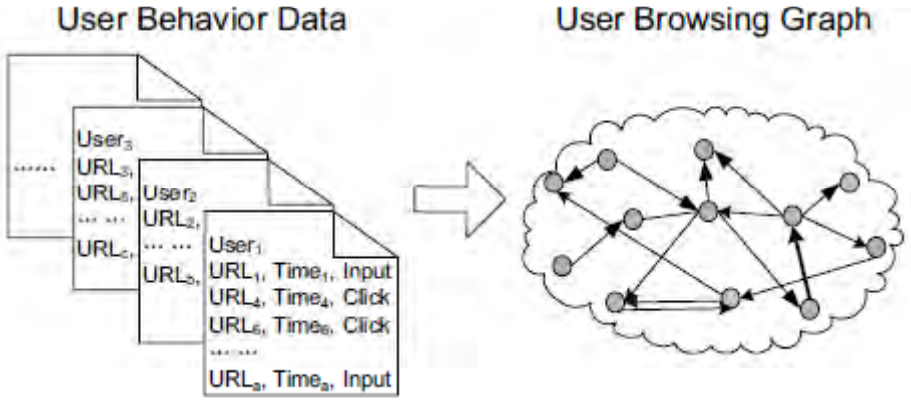
Quality-Rank / Vertrauenswürdigkeit: Beispiel

Hinweis: Bei all diesen Werkzeugen:
Werte immer mit Vorsicht / Skepsis betrachten

Guckst Du: www.linkvoodoo.com/score_site.php

BrowseRank / Wo klicken die Menschen?

- Von Microsoft Research veröffentlicht
- Soll die Schwächen das Linkgraphen-basierten PageRanks beheben
- Bewegungen zwischen und Verweildauer auf Webseiten der Nutzer werden -anonym- via Toolbars und Phone-Home (?) erhoben – auch URL-Eingabe oder Linkklick



Quelle: http://research.microsoft.com/en-us/people/tyliu/fp032-liu.pdf, S. 2

URL	TIME	TYPE
http://aaa.bbb.com/	2007-04-12, 21:33:05	INPUT
http://aaa.bbb.com/1.htm	2007-04-12, 21:34:11	CLICK
http://ccc.ddd.org/index.htm	2007-04-12, 21:34:52	CLICK
http://eee.fff.edu/	2007-04-12, 21:39:03	INPUT
...	...	...

Quelle: http://research.microsoft.com/en-us/people/tyliu/fp032-liu.pdf, S. 3

$$\pi_i = \frac{\frac{\tilde{\pi}_i}{q_{ii}}}{\sum_{j=1}^{N} \frac{\tilde{\pi}_j}{q_{jj}}} \qquad -\frac{q_{ij}}{q_{ii}} = \begin{cases} \alpha \frac{\tilde{w}_{ij}}{\sum_{k=1}^{N+1} \tilde{w}_{ik}} + (1-\alpha)\sigma_j, & i \in V, j \in \tilde{V} \\ \sigma_j, & i = N+1, j \in V \end{cases}$$

- Problematisch: Verstärkung von Trampelpfaden, Neue (interessante) Seiten/Sites werden in der Suchmaschine schlecht gefunden, Signifikanz „kleiner“ Sites möglicherweise zu niedrig
  - -> Nur wo viele sind bzw. schon waren, ist es gut?
  - -> Braucht man zum Finden von Powersites überhaupt eine SuMa?
  - -> Wie PageRank auch manipulierbar, z. B. über Botnetze

* Formel hat keinen Bezug zur Abbildung, sieht aber schön kompliziert aus!

BrowseRank / Wo klicken die Menschen?

Unterschiede PageRank / TrustRank / BrowseRank in der Bewertung

No.	PageRank	TrustRank	BrowseRank
1	adobe.com	adobe.com	*myspace.com*
2	passport.com	yahoo.com	msn.com
3	msn.com	google.com	yahoo.com
4	microsoft.com	msn.com	*youtube.com*
5	yahoo.com	microsoft.com	live.com
6	google.com	passport.net	*facebook.com*
7	mapquest.com	ufindus.com	google.com
8	miibeian.gov.cn	*sourceforge.net*	ebay.com
9	w3.org	*myspace.com*	*hi5.com*
10	godaddy.com	*wikipedia.org*	*bebo.com*
11	statcounter.com	phpbb.com	*orkut.com*
12	apple.com	yahoo.co.jp	aol.com
13	live.com	ebay.com	*friendster.com*
14	xbox.com	nifty.com	*craigslist.org*
15	passport.com	mapquest.com	google.co.th
16	*sourceforge.net*	cafepress.com	microsoft.com
17	amazon.com	apple.com	*comcast.net*
18	paypal.com	infoseek.co.jp	*wikipedia.org*
19	aol.com	miibeian.gov.cn	*pogo.com*
20	*blogger.com*	*youtube.com*	*photobucket.com*

Quelle: http://research.microsoft.com/en-us/people/tyliu/fp032-liu.pdf, S. 6

Wissenschaftliche Quellen

- Zum Hilltop-Algorithmus: http://ftp.cs.toronto.edu/pub/reports/csrg/405/hilltop.html
- Zum HITS-Algorithmus, Univ. of Missouri-Columbia: http://www2002.org/CDROM/refereed/643/
- Deutsche Seminararbeit zum HITS-Algorithmus: http://www-dbs.informatik.uni-heidelberg.de/teaching/ss2005/seminar/Ausarbeitung/1-2.pdf
- Weiterentwicklung von HITS zum ACR-Algorithmus (mehr Themenbezug der Links + Gewichtung): http://www.cs.cornell.edu/home/kleinber/www98-arc.pdf
- Standford-Publikation zu ACR: http://theory.stanford.edu/people/raghavan/www7/181.html
- Zum Trust-Rank-Algorithmus, Univ. Stanford: http://www.vldb.org/conf/2004/RS15P3.PDF
- Zum BrowseRank von Microsoft: http://research.microsoft.com/en-us/people/tyliu/fp032-liu.pdf
- PageRank-Patent von 2006 mit Nutzerverhalten: http://tinyurl.com/y9mnts
- Linknachbarschaften C4.5-Algo: http://www.cse.lehigh.edu/~brian/pubs/2000/aaaiws/aaai2000ws.pdf
- Linknachbarschaften am unstimmigen Spachmodell erkennen: http://tinyurl.com/bty6f4
- Zufallsgesteuert Linkfarmen finden: http://airweb.cse.lehigh.edu/2007/papers/paper_116.pdf
- Spamseiten über Content-Analyse finden: http://research.microsoft.com/pubs/65140/www2006.pdf
- Linkanalyse zur Spamerkennung (Univ. Rom und Yahoo!): http://tinyurl.com/cambcv

Bissi Eigenmarketing: Blog auf www.Website-Boosting.de und
http://twitter.com/mariofischer