

„Vortrag Suchmaschinen und Data Mining“

Andreas Armbruster

Student Wirtschaftswissenschaften (7. Semester)

andreas.armbruster@mathematik.uni-ulm.de

30. Januar 2004

DATA MINING

Data Mining ist der Oberbegriff für Verfahren zur automatischen inhaltlichen Analyse und Erschließung großer Mengen von numerischen Daten.

Solche Verfahren werden von Suchdienstbetreibern für die Inhaltsbewertung von Webseiten verwendet und somit für das Ranking der Suchmaschinen-Positionen -> Suchmaschinen-Algorithmen.

Nicht zu verwechseln mit Web Mining, was sich mit der Analyse von Weblogs und Userbewegungen beschäftigt.

Anwendung von Data Mining auf Webseiten

Erstmalige Beschreibung der Gründer (Lawrence Page & Sergey Brin) von Google 1998.

Beschreibung ausgehend von einem statistischen Verfahren, den so genannten Assoziationsregeln, mit dem Einkaufsgewohnheiten in Supermärkten erforscht wurden.

Klassische Beispiel : Wechselbeziehung zwischen gekauftem Bier und gekauften Windeln.

Darstellung von Zusammenhang beim Kauf eines Produktes A mit dem Kauf eines Produktes B durch Verbindungslinien zwischen A und B -> bei manchen Produkt-Kombinationen stärkere Linien, bei anderen dünnere.

Verbindungslinien von Produktkombinationen

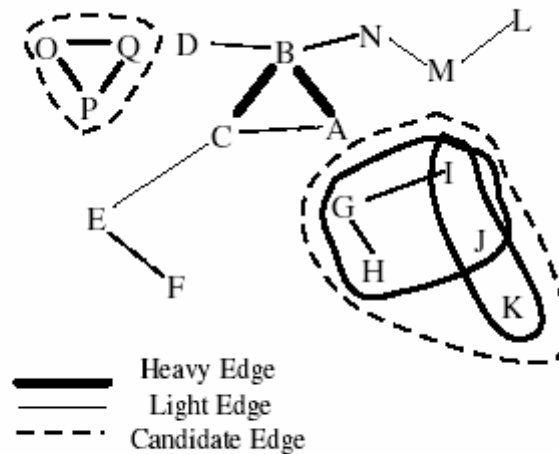


Illustration of the Heavy Edge Property

Sergey Brin & Lawrence Page, 23.2.1998

Heavy Edge: starke Verbindungslinien, Light Edge: schwache Verbindungslinien, Candidate Edge: Linie um Produktkombinationen

Assoziationsregeln

Durch Berechnungen und Vergleiche zwischen den Produkt-Kombinationen entstanden die so genannten Verknüpfungs-/Assoziationsregeln. Assoziationsregeln beschreiben also Korrelationen von gemeinsam auftretenden Dingen.

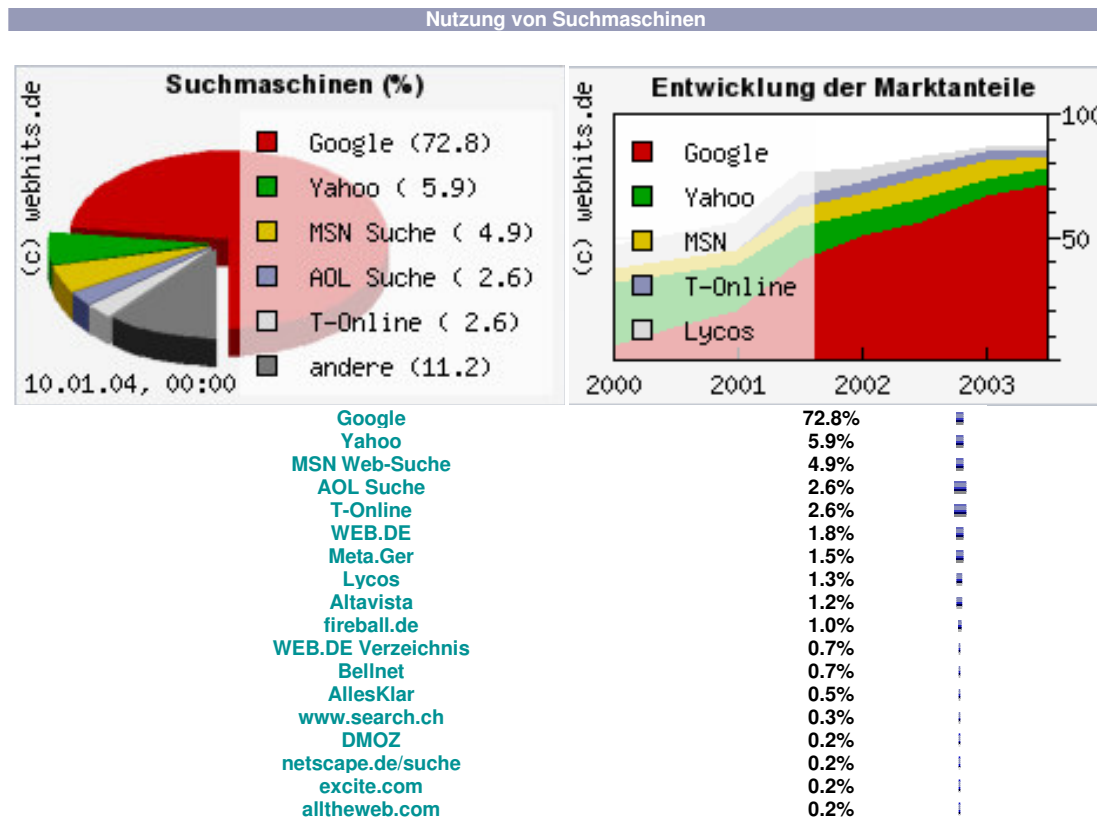
Für Assoziationsregeln sind folgende Parameter relevant:

- **Konfidenz** der Regel, d.h. Stärke der Korrelation (z.B. „in 45% der Fälle“)
- **Support** der Regel, d.h. Häufigkeit des gemeinsamen Auftretens (z.B. „in 2% aller Transaktionen“)

(Quelle: <http://www.aifb.uni-karlsruhe.de/Lehre/Winter2002-03/kdd/download/VII-3-Assoziationsregeln.pdf>, aufgerufen im Januar 2004)

Suchmaschinen-Nutzung in Deutschland

Quelle: webhits.de Stand: 10.1.2004



Datenmengen in den Suchmaschinen

Name / Betreiber	Datenbasis	Verzeichnisgröße	Erscheinungsdatum	URL / Besonderheiten
Abacho Abacho AG	Abacho Meta-Suche: Google, Excite, Yahoo	ca. 100 Millionen Web-Seiten Deutsch-Internationales Verzeichnis	März 2000	http://www.abacho.de
Alltheweb	FAST	ca. 3,2 Mrd. Web-Seiten	2000	http://www.alltheweb.com
Altavista	Altavista, Katalogteil: Looksmart	ca. 1.5 Mrd. Web-Seiten	1995	http://www.altavista.com
Ask Jeeves ASK	Teoma			http://www.ask.com
Fireball Lycos	Fireball	ca. 20 Millionen deutschsprachige Web-Seiten	Juni 1997	http://www.fireball.de
Google	Google	ca. 3.3 Mrd. Web-Seiten	September 1999	http://www.google.com weltweit täglich 150 Millionen Zugriffe
Lycos Bertelsmann	FAST		Herbst 1996	1.400 Millionen Seitenaufrufe (Europa)
Mirago	Mirago	ca. 100 Millionen Dokumente	März 2003	Katalogteil integriert und zusätzliche Sektoren-Suche (Bereichssuche) möglich
Openfind	Openfind	ca. 3,0 Mrd. Webseiten		starke Ausprägung auf den asiatischen Raum (koreanisches Verzeichnis)
QualiGO Suchtreffer AG	QualiGO	15 Millionen Web-Seiten	September 2000	http://www.qualigo.de ca. 15 Millionen Suchabfragen pro Monat
Teoma ASK	Teoma	ca. 500 Millionen indizierte URL'S	April 2000	http://www.teoma.com ca. 17 Millionen User
Wisnut Looksmart	Wisnut	ca. 1.600 Millionen Web-Seiten	2001	http://www.wisnut.com

Quelle: thom-online.de (Stand September 2003)

Wie erhalten Suchmaschinen ihre Daten?

Suchmaschinen sammeln Ihre Daten mit spezieller Software, den Robots, die ihre Informationen von den Webservern erhalten, bei denen die Webseiten abgelegt sind.

Über Hyperlinks erfahren die Robots, wo die nächsten Seiten sind, deren Inhalte auf die Anfragen der Robots an die Suchmaschine übermittelt werden.

Ein Robot stellt also lediglich Anfragen, die ihm in Form übermittelter Daten beantwortet werden. Hier verhalten sich Robots ähnlich wie Browser.

Eintragsnamen von Robots

- Googlebot/2.1 (+<http://www.googlebot.com/bot.html>)
- Scooter/3.3_SF
- Spider.TerraNautic.net - v:1.04
- Mozilla/5.0 (Slurp/cat; slurp@inktomi.com;
<http://www.inktomi.com/slurp.html>)
- teomaagent [crawler-admin@teoma.com]

Beispiel robots.txt

```
User-agent: Robot1  
Disallow: /logfiles/  
Disallow: /temp/  
Disallow: /news/
```

```
User-agent: *  
Disallow: /logfiles/  
Disallow: /temp/
```

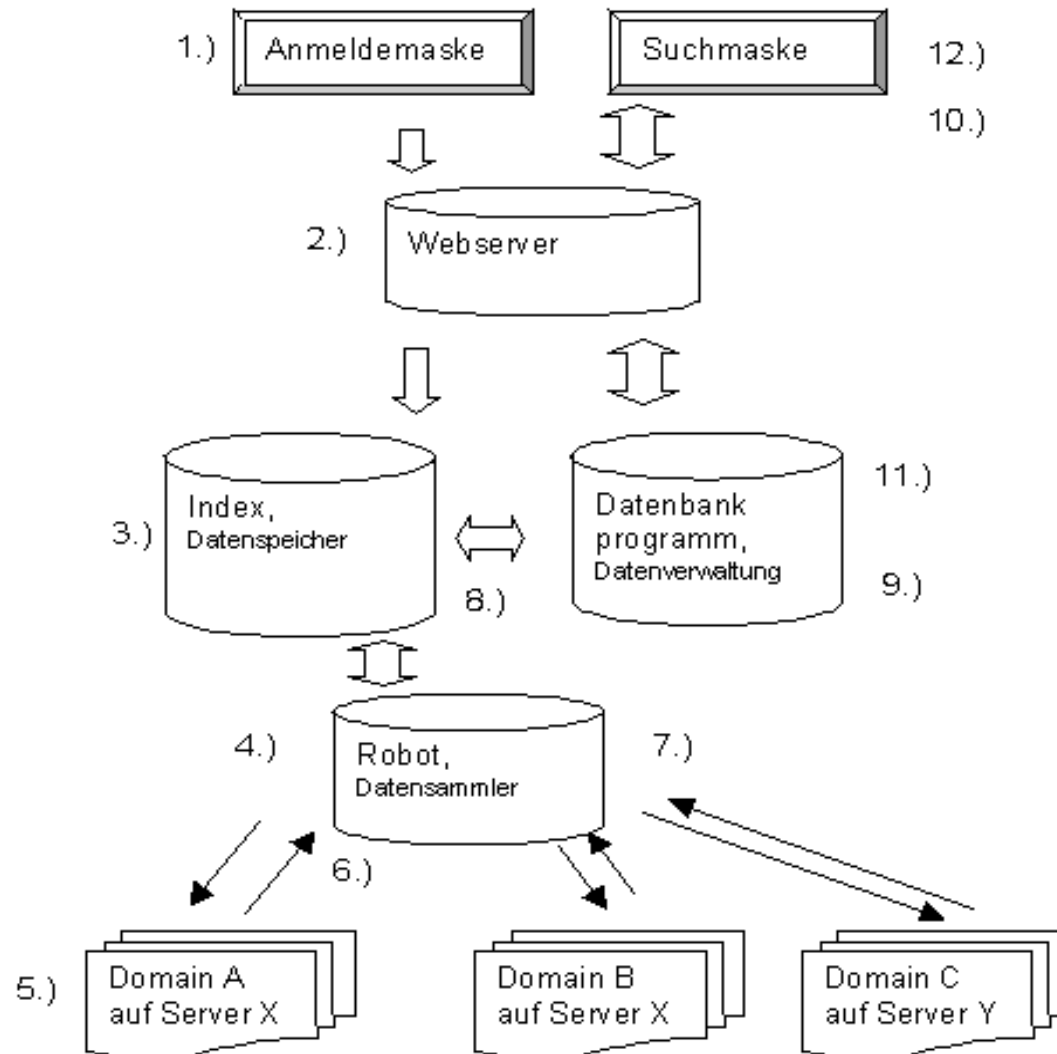
Mit User-agent gibt man die zu behandelnden Robots an, mit Disallow kann man bestimmte Verzeichnisse, Dateien usw. ausschließen. In diesem Beispiel werden für alle Robots die Verzeichnisse /logfiles/ und /temp/ ausgeschlossen und für den Robot „Robot1“ auch das Verzeichnis /news/ .

Erfassung der Webseiten

Bei Suchmaschinen gibt es 2 gängige Verfahren um Webseiten zu erfassen:

- manuelle Anmeldung des HTML-Dokuments auf der Anmeldeseite der Suchmaschine
- automatisches Verfolgen durch die Suchprogramme der Links von angemeldeten Seiten, nachdem sie von der Software erfasst und ausgewertet wurden

Indexierung und Abfrage - Suchmaschinen



(Quelle: www.at-web.de/suchmaschinen/suchmaschinen-robots.htm)

Anmerkungen zu den Punkten 1-12

- Web-Dokumente werden nie beim Absenden der Suchanfrage direkt bei der Suchmaschinenabfrage aufgerufen, sondern nur das "Abbild", in der Datenbank der Suchmaschine
- Robots erledigen lediglich die Abfragen der Seiten und geben diese Seiten zur Indexierung und Speicherung an den Server weiter
- Nicht alle Suchmaschinen erfassen die gesamten Dokumente. Oft werden nur Teile davon abgespeichert um den Speicherplatz rationell zu belegen: Meta-Tags, Titel, die ersten Zeilen, zugehörige URL
- Eine Suchmaschine deckt in der Regel nur einen kleinen Teil des Internets ab und kann deshalb immer nur Informationen liefern, die sie auch in Ihrer Datenbank gespeichert hat.

Suchmaschinen-Kategorien

- **Volltextsuchmaschinen:** u.a. sind Volltextsuche und Phrasensuche möglich, Beispiele: Google, Alltheweb/Fast, Altavista
- Speichern v. Meta-Daten als **Verschlagwortung:** Suche in Verschlagwortung. Beispiele: MetaGer.de, MetaSpinner.de, Kartoo.com
- Speichern von **Wort-Statistiken:** Stichwortsuche -> Web-Kataloge wie Yahoo.de, DMOZ.org, WEB.de

Ranking-Algorithmus

Die wichtigsten Ranking-Kriterien:

- Relative Häufigkeit des Suchbegriffes
- Titel
- Überschriften h1, h2, ...
- URL
- Meta Keywords und Description (Ausnahme Google)
- ALT-Attribut des IMG-Tags
- Links (``) (sowohl Ziel als auch Beschreibung)
- Hervorhebungen: strong, underlined, italic

LinkPopularity

Die meisten Suchmaschinen bewerten die Anzahl externer Links auf eine bestimmte Seite. Dahinter steckt der Gedanke, dass nur qualitativ wertvolle Seiten viele Links von anderen Seiten erhalten.

Die genauen Details des LinkPopularity-Algorithmus sind von Suchmaschine zu Suchmaschine verschieden.

Häufig wird auch die Relevanz der externen Links mitbewertet -> Links von themennahen Seiten werden höher bewertet als von themenfremden.

Diesbezüglich wird auch oft bewertet, ob im Linktext der Suchbegriff vorkommt, der das Hauptthema der verwiesenen Seite enthält.

Benutzung von LinkPopularity

Von folgenden Suchmaschinen ist bekannt, dass sie die Anzahl (und evtl. Relevanz) externer Links im Ranking-Algorithmus benutzen:

- AllTheWeb (FAST, Overture)
- Altavista.com (Overture)
- AOL (Google)
- Fireball
- Google
- HotBot (Inktomi)
- Lycos (FAST)
- MSN (Inktomi)
- Netscape (Google)
- Teoma (Ask Jeeves)
- Wisenut (LookSmart)
- Yahoo! (Google)

PageRank-Verfahren

Google spezielles Ranking-Verfahren, welches für die hohe Qualität der Ergebnisse sorgt. Damit wird die Wichtigkeit einer Webseite bewertet.

Wie bei der LinkPopularity wird hier ausschließlich die Verlinkung zwischen Webseiten und deren Bedeutung beurteilt.

Das PageRanking beurteilt die Wertung aller Links die auf eine Seite zeigen und beurteilt die Seite nach dem Gesamtwert aller Verweise.

Nicht nur die Startseite, sondern jede einzelne indexierte Seite ist nach dem PageRank Verfahren bewertet.

Im PageRank Verfahren werden sämtliche Links beurteilt, interne und externe Links. Hier besteht ein wichtiger Unterschied zur LinkPopularity, die nur Links von anderen Sites berücksichtigt.

Google-Toolbar

Aktuelles PageRank	Wert in der Toolbar
0,00000001 bis 5	1
6 bis 25	2
25 bis 125	3
126 bis 625	4
626 bis 3 125	5
3 126 bis 15 625	6
15 626 bis 78 125	7
78 126 bis 390 625	8
390 626 bis 1 953 125	9
ab 1 953 126	10

Quelle: www.at-web.de (aufgerufen im Dezember 2003)

Das PageRanking wird nicht linear bewertet, jedoch für die Toolbar in eine lineare Bewertung umgesetzt.

Nachteile des PageRank & LinkPopularity Konzepts

Die Benutzung der LinkPopularity als dominierendes Element im Ranking-Algorithmus führt zu einer Bevorzugung bereits bekannter Websites - oder, noch problematischer, von großen potenten Netzwerken.

Durch das PageRank-Verfahren können jedoch neuere Seiten, die einen Link (=Empfehlung) von wenigen, aber sehr gut bewerteten, Seiten bekommen, besser gerankt werden. Aber auch hier ist es für eine finanzträchtige Seite leichter an solche Links zu kommen, als für weniger starke -> blühender Handel mit Linkverkäufen und Linkaustauschen.

Suchmaschinen verlieren dadurch eine zentrale Eigenschaft: Die Unabhängigkeit der Suchergebnisse von der Marketingpotenz der Sitebetreiber.

Warum Suchmaschinen-Spamming?

Durch Verlinkungen mit Partnerprogramm-Links von tradedoubler.com, zanox.de, partnerprogramme.de oder anderen Affiliate-Programmen lässt sich Geld verdienen durch Provisionen oder Bezahlungen per Klick.

-> Existenz von zahlreichen, für das Suchmaschinen-Ranking optimierte, Spam-Seiten, die speziell darauf ausgerichtet sind möglichst viele Visits über Suchmaschinen zu bekommen und diese sofort wieder über irgendwelche Affiliate-Links, die Provisionen bringen, weiterleiten.

Ein richtiger Themenbezug wird quasi nur „vorgetäuscht“, es geht nur um die sofortige bezahlte Weiterverlinkung.

Spamming-Methoden

- der allzu häufige Einsatz eines Begriffes -> zu hohe relative Häufigkeit
- ein Textteil wird in der Hintergrundfarbe geschrieben und ist somit in einem normalen Browser nicht sichtbar, die Suchmaschinen hingegen lesen nur den Quelltext und sehen die "versteckten" Worte
- Schlüsselbegriffe werden innerhalb von HTML-Kommentaren notiert.
- in die Meta Keywords werden Begriffe sehr häufig wiederholt
- Cloaking - Verfahren, das dem Crawler einer Suchmaschine eine andere Seite liefert als einem normalen Nutzer -> gut optimierter HTML-Code
- Doorwaypages – Brückenseiten. Für die Suchmaschinen optimierte, aber für den User sinnlose Seiten -> sofortige Weiterleitung

Spamming-Beispiel

Abfrage bei Google mit dem Begriff „Lastminute Urlaub“

[Last Minute Urlaub buchen](#)

Lastminute Urlaub Buchen: Das Reise Portal mit **Lastminute** und Flug Angeboten, **Lastminute** Pauschalreisen, Individualreisen, Schnäppchen **Urlaub** und Direkt ...

www.lastminute-urlaub-buchen.de/ - 13k - [Im Cache](#) - [Ähnliche Seiten](#)

[Lastminute Urlaub Last Minute Reise](#)

... **Lastminute Urlaub** Angebote präsentiert Ihnen: ... Norwegen, USA... Direkt zu allen Skireisen. **Lastminute Urlaub**. Lastminutereisen und ...

lastminute-urlaub-angebote.de/ - 30k - [Im Cache](#) - [Ähnliche Seiten](#)

[Lastminute Reisen - Last Minute Urlaub](#)

Lastminute Urlaub. Hier bei **lastminute**-abwechslung.de bekommen Sie Last Minute Flüge zum kleinen Preis. Sie planen einen **Lastminute Urlaub**? ...

www.lastminute-abwechslung.de/ - 6k - [Im Cache](#) - [Ähnliche Seiten](#)

[Spontan-Urlaub.de - preiswert Lastminute Reiseangebote online ...](#)

... **Lastminute** Reiseangebote, Pauschalreisen und mehr... Bei Spontan-**Urlaub**.de finden Sie den passenden **Lastminute-Urlaub** und weitere aktuelle Reiseangebote. ...

www.spontan-urlaub.de/ - 16k - [Im Cache](#) - [Ähnliche Seiten](#)

[Lastminute-Urlaub online buchen](#)

... **Lastminute-Urlaub**. ... Leider kann ihr Browser keine eingebetteten Frames darstellen, deshalb finden Sie hier nicht die **Lastminute-Urlaub**-Angebote. ...

www.spontan-urlaub.de/lastminute_urlaub.php - 12k - [Im Cache](#) - [Ähnliche Seiten](#)

Spamming-Beispiel

Lastminute Urlaub **Flugreisen** Lastminute - **Schnäppchenpreise**

Lastminute Urlaub Reisen Flugreisen und vieles mehr. ... Die günstigsten Angebote für **Lastminute Urlaub**, Pauschalreisen, Flugreisen und Unterkünfte. ...

www.urlaubsreiseboerse.de/ - 8k - **Im Cache** - **Ähnliche Seiten**

lastminute urlaub **reisen buchen** - last minute

lastminute urlaub reisen buchen mit billig fluege ! fernreisen, kreuzfahrten und mehr via last minute reisen in den **urlaub**. **Lastminute Urlaub** reisen last minute. ...

www.lastminute-urlaub.ws/ - 6k - **Im Cache** - **Ähnliche Seiten**

Lastminute Urlaub

Lastminute Urlaub. Startseite.

www.tohit.de/ - 2k - **Im Cache** - **Ähnliche Seiten**

Last Minute Urlaub **buchen**

Urlaub buchen in den schönsten Hotels weltweit: ... Für die eigene Anreise mit dem Auto klicken Sie bitte den Button Autoreisen. Last Minute **Urlaub** buchen. ...

www.prohotel.de/ - 20k - **Im Cache** - **Ähnliche Seiten**

Urlaub, **Last-Minute Reisen billig buchen**

... und Ferienwohnungen weltweit, **Lastminute** und Sonderangebote in Spanien, auf den Balearen, den Kanarischen Inseln. Top Angebote für **Urlaub** in ganz Europa zum ...

www.lastminutos.de/ - 29k - **Im Cache** - **Ähnliche Seiten**

Spamming-Bekämpfung

Alle wichtigen Suchmaschinen setzen Filter ein und erkennen viele der vorgestellten Methoden.

-> deutlich schlechteres Ranking oder gar Verbannung der kompletten Website aus der Suchmaschine.

Wie letzte Folie zeigt, sind im Index von Suchmaschinen, insbesondere bei Google, immer noch mehr als genug Spam-Seiten erfolgreich gelistet.

Zum einen bestrafen nicht alle Suchmaschinen alle Spammingtechniken, zum anderen funktionieren die Filter auch nicht perfekt. Suchmaschinen-Spammer lassen sich auch immer neuere Techniken einfallen oder bestehende Techniken werden verfeinert.

Ausblick

Michael Schmitt, Software Engineer von Google, erklärte am 12.11. auf der Search Engines Strategies Conference (SESC), dass die automatisierte Bekämpfung von Spam "höchste Priorität für Google" hat.

Das Spam-Problem ist laut Schmitt in Deutschland weitaus größer als in den USA und deswegen vernachlässigt worden.

Auch in Zukunft wird kein manueller Eingriff stattfinden -> automatische Spamfilter.

Neue Anti-Spam-Algorithmen mit Schwerpunkt "Analyse der Verlinkung" sind in Entwicklung.
