

**Seminararbeit  
„Suchmaschinen und Data Mining“**

**Andreas Armbruster**  
**Student Wirtschaftswissenschaften (7. Semester)**  
[andreas.armbruster@mathematik.uni-ulm.de](mailto:andreas.armbruster@mathematik.uni-ulm.de)

**21. Januar 2004**

## Inhaltsverzeichnis

<b>1. Grundlegendes zu Suchmaschinen &amp; Data Mining</b>	<b>3</b>
<b>2. Suchmaschinen-Überblick</b>	
2.1. Suchmaschinennutzung in Deutschland	5
2.2. Anzahl der Suchmaschinen-Abfragen	5
2.3. Datenmengen in den Suchmaschinen	6
<b>3. Wie erhalten Suchmaschinen Ihre Daten?</b>	<b>7</b>
<b>4. Suchmaschinen-Algorithmen</b>	
4.1. Ranking-Algorithmus	12
4.2. LinkPopularity	15
4.3. PageRank-Verfahren	17
<b>5. Suchmaschinen-Spamming</b>	
5.1. Spamming-Methoden	19
5.2. Spamming-Beispiel	20
5.3. Bekämpfung	21
5.4. Ausblick über zukünftige Algorithmen zur Spam-Bekämpfung	21
<b>Literaturverzeichnis</b>	<b>22</b>

## Abbildungsverzeichnis

1.1 Verbindungslinien von Produktkombinationen Quelle: <a href="http://www-db.stanford.edu/~sergey/ddm.ps">http://www-db.stanford.edu/~sergey/ddm.ps</a>	3
2.1 Suchmaschinen-Marktanteile in Deutschland Quelle: <a href="http://www.webhits.de">http://www.webhits.de</a>	5
2.2 Datenmengen in den Suchmaschinen Quelle: <a href="http://www.thom-online.de">http://www.thom-online.de</a>	6
3.1. Indexierung und Abfrage von Suchmaschinen Quelle: <a href="http://www.at-web.de/suchmaschinen/suchmaschinen-robots.htm">http://www.at-web.de/suchmaschinen/suchmaschinen-robots.htm</a>	8
4.1 Umsetzung der Google-Pagerankwerte in die Google-Toolbar Quelle: <a href="http://www.at-web.de">www.at-web.de</a>	17
5.1 Spam-Beispiel in Google Quelle: <a href="http://www.google.de">www.google.de</a>	20

# Kapitel 1

## Grundlegendes zu Suchmaschinen & Data Mining

Quelle: polarluft.de (aufgerufen im Dezember 2003)

**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.

Die Gründer (Lawrence Page & Sergey Brin) des Suchdienstes Google beschrieben 1998 an der Stanford University ( <http://www-db.stanford.edu/~sergey/ddm.ps> , *Dynamic Data Mining: Exploring Large Rule Spaces By Sampling* ) erstmals die Anwendung von Data Mining auf Webseiten ausgehend von einem statistischen Verfahren, den so genannten Assoziationsregeln, mit dem Einkaufsgewohnheiten in Supermärkten erforscht wurden.

Das klassische Beispiel ist die Wechselbeziehung zwischen gekauftem Bier und gekauften Windeln. Man nahm dabei die Kassen-Abrechnungen und erstellte daraus Listen von Produkten, die jeweils ein Kunde in seinem Einkaufswagen gehabt hatte (Warenkorbanalyse). Danach stellte man den Zusammenhang beim Kauf eines Produktes A mit dem Kauf eines Produktes B in einem Schaubild durch eine Verbindungslinie zwischen A und B dar, und verfuhr mit allen anderen Produkten genauso (siehe Abbildung 1.1).

So entstanden bei manchen Produkt-Kombinationen stärkere Linien, bei anderen dünnere. Durch Berechnungen und Vergleiche zwischen den Produkt-Kombinationen entstanden die so genannten Verknüpfungs-/Assoziationsregeln. Assoziationsregeln beschreiben also Korrelationen von gemeinsam auftretenden Dingen.

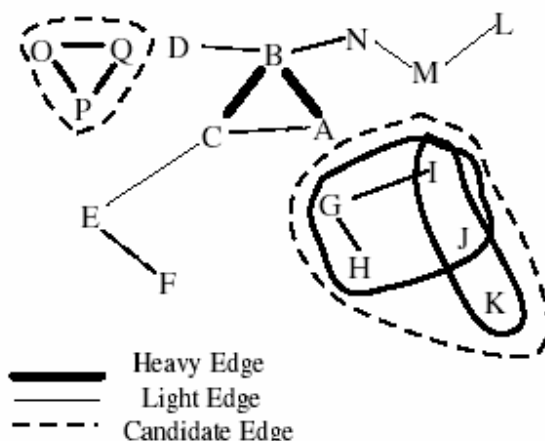


Illustration of the Heavy Edge Property

Abbildung 1.1. Verbindungslinien von Produktkombinationen  
Sergey Brin & Lawrence Page, 23.2.1998

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

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)

Um Aussagen über zusammen auf einer Webseite vorkommende Wörter zu erhalten, übertrug man dieses Modell und setzte anstelle der Produkt-Kombinationen Webseiten, anstelle der Produkte die Wörter innerhalb der Webseiten. Man begegnete dabei Schwierigkeiten wie der mit der Anzahl der Wörter exponentiell wachsende Anstieg der möglichen Verknüpfungsregeln, und damit, überhaupt erst die aussagekräftigen Verknüpfungsregeln für gemeinsam vorkommende Wörter herauszufinden.

## Kapitel 2

### Suchmaschinen-Überblick

#### 2.1. Suchmaschinen-Nutzung in Deutschland

Quelle: webhits.de Stand: 10.1.2004

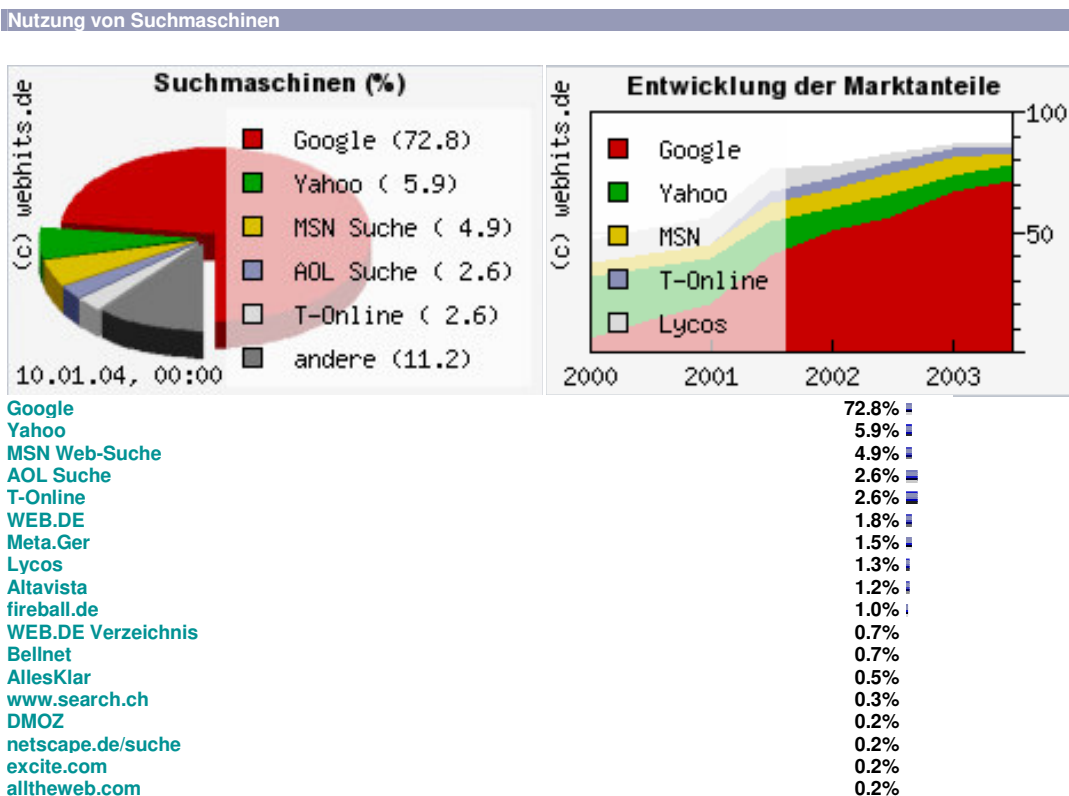


Abbildung 2.1. Suchmaschinen-Marktanteile in Deutschland

Deutlich erkennbar ist die absolute Marktführung von Google. Verstärkt wird diese Marktposition noch dadurch, dass Yahoo (derzeit noch), AOL und viele kleine Suchdienste die Daten von Google nutzen.

#### 2.2. Anzahl der Suchmaschinen-Abfragen

Google: Rund 10.000 Workstations (Power-PCs) sind auf fünf Standorte an der West- und Ostküste der USA verteilt, die über 150 Millionen Abfragen pro Tag bearbeiten. Näheres siehe Abbildung 2.2. .

## 2.3. Datenmengen in den Suchmaschinen

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

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	<a href="http://www.abacho.de">http://www.abacho.de</a>
Alltheweb Overture	FAST	ca. 3,2 Mrd. Web-Seiten	2000	<a href="http://www.alltheweb.com">http://www.alltheweb.com</a>
Altavista Overture	Altavista  Katalogteil: Looksmart	ca. 1.5 Mrd. Web-Seiten	1995	<a href="http://www.altavista.com">http://www.altavista.com</a>
Ask Jeeves ASK	Teoma			<a href="http://www.ask.com">http://www.ask.com</a> Daten stammen aus verschiedenen Quellen. Bietet auch eine Kindersuchmaschine an: <a href="http://www.ajkids.com">http://www.ajkids.com</a>
Fireball Lycos	Fireball	ca. 20 Millionen deutschsprachige Web-Seiten	Juni 1997	<a href="http://www.fireball.de">http://www.fireball.de</a>
Google Google	Google	ca. 3.3 Mrd. Web-Seiten	September 1999	<a href="http://www.google.com">http://www.google.com</a> 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	<a href="http://www.qualigo.de">http://www.qualigo.de</a> ca. 15 Millionen Suchabfragen pro Monat
Teoma ASK	Teoma	ca. 500 Millionen indizierte URL'S	April 2000	<a href="http://www.teoma.com">http://www.teoma.com</a> ca. 17 Millionen User
Wisenut Looksmart	Wisenut	ca. 1.600 Millionen Web-Seiten	2001	<a href="http://www.wisenut.com">http://www.wisenut.com</a>

Abbildung 2.2. Datenmengen in den Suchmaschinen

Da das WWW immer noch rasend wächst gibt es über die Abdeckungsgrade nur grobe Schätzungen und dürften bei Google zwischen 10-20% liegen (mehrere Quellen, teilweise ältere).

## Kapitel 3

### Wie erhalten Suchmaschinen Ihre Daten?

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

Suchmaschinen sammeln Ihre Daten mit spezieller Software, den Robots, die ihre Informationen von den Webservern erhalten, bei denen die Webseiten abgelegt sind. Im Gegensatz dazu erhalten Kataloge wie Yahoo oder Web.de ihre Informationen indem die angemeldeten Seiten von Menschen angesehen und beurteilt werden.

Eigentlich sind Namen wie **Crawler** (Kriecher), **Spider** (Spinne) oder **Worm** (Wurm) irreführend, weil sie die Vorstellung wecken, dass diese durch das Web wandern, sich dort die Seiten durchlesen, über Hyperlinks weiterwandern, die nächste Seite lesen, usw. . Tatsächlich geschieht dieser Prozess komplett auf dem Rechner der Suchmaschine im Prozess des Spiders.

Über die Hyperlinks erfahren die Robots, wo die nächsten Seiten sind, deren Inhalte auf die Anfragen der Robots an die Suchmaschine übermittelt werden. Ein Robot wandert nicht zwischen den Seiten herum, sondern er stellt lediglich Anfragen, die ihm in Form übermittelter Daten beantwortet werden. Hier verhalten sich Robots ähnlich wie Browser.

Die hier erwähnten Robots sind den Suchmaschinen zugeordnet. Es gibt noch eine Reihe weiterer Anwendungen die mit Robot-ähnlichen Programmen arbeiten. Etwa Link-Checker oder Programme, die ganz bestimmte Informationen aus Webseiten lesen. Alle Robots werden vom Webserver in die Logfiles der Domain eingetragen.

Hier eine kurze Liste von Eintragsnamen:

- 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](mailto:slurp@inktomi.com); <http://www.inktomi.com/slurp.html>)
- teomaagent [crawler-admin@teoma.com]

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

In der folgenden Darstellung ist der prinzipielle Weg dargestellt, wie eine Datei in die Datenbank einer Suchmaschine aufgenommen wird (Punkte 1-9).

Die Punkte 10-12 beschreiben die Abfrage der Suchmaschinen-Datenbank.

Die Darstellung hat sehr allgemeinen Charakter und soll lediglich zum allgemeinen Verständnis für die verschiedenen Elemente der Suchmaschine beitragen. Sie kann natürlich von Suchmaschine zu Suchmaschine sehr variieren. Die ausführliche Darstellung der Google-

Suchmaschine und ihrer Bestandteile würde wesentlich komplexer ausfallen.

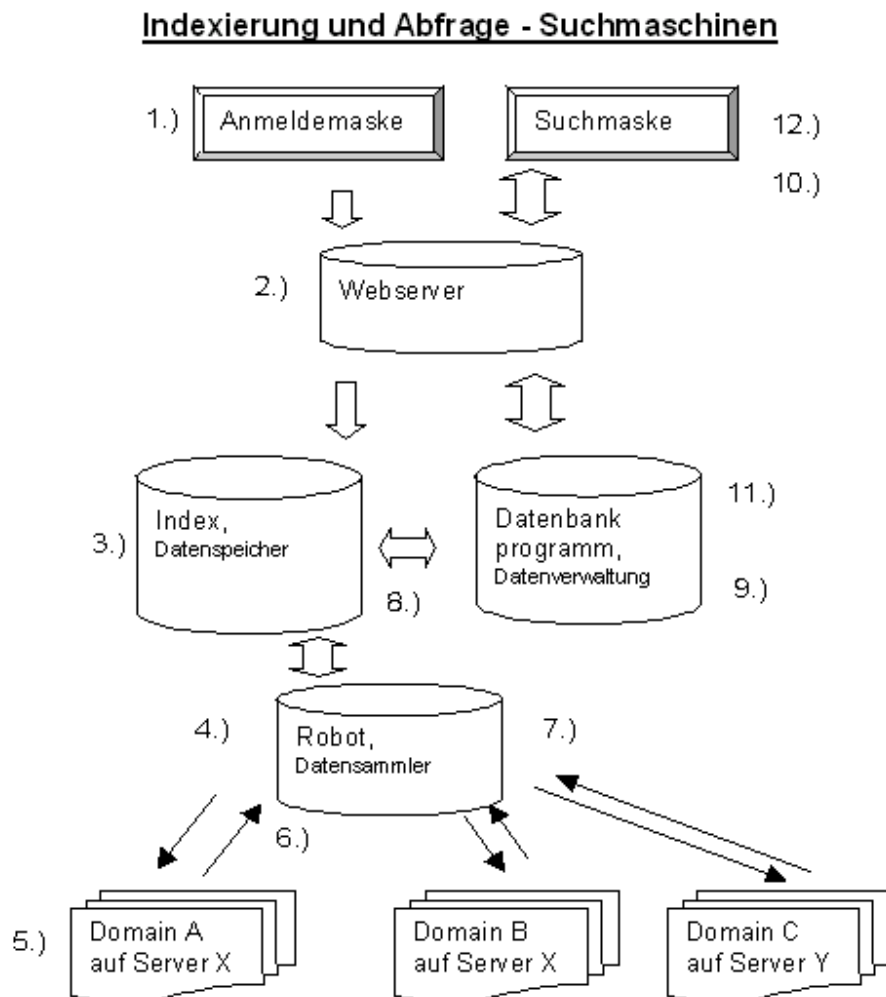


Abbildung 3.1. Indexierung und Abfrage von Suchmaschinen  
(Quelle: [www.at-web.de/suchmaschinen/suchmaschinen-robots.htm](http://www.at-web.de/suchmaschinen/suchmaschinen-robots.htm))

Erläuterung des oben dargestellten Prinzips:

1.) In der Anmeldemaske der Suchmaschine werden HTML-Dokumente angemeldet. In der Regel ist das die Startseite einer Webpräsenz. Es können natürlich auch einzelne Unterseiten angemeldet werden, wenn diese schnell in die Datenbank aufgenommen werden sollen. Die Anmeldemaske und Suchmaske gehören zum Webserver der Suchmaschine, welche auch die Ergebnislisten anzeigt.

2.) Der Webserver der Suchmaschine dient als Schnittstelle zum Datenbankprogramm.

3.) Die URL wird in den Indexer aufgenommen. Diverse Filter gegen Spam und gegen unsaubere Inhalte sorgen für aufbereitete Daten. Gesperrte Adressen werden ebenfalls ausgefiltert. Einige Suchmaschinen setzen außerdem auf redaktionelle Auslese. In der Praxis gibt es allerdings mit dem Filtern Probleme, wie man leicht feststellen kann.

Bei vielen Suchmaschinen dauert der Aufnahmevorgang nur wenige Tage. Leistungsfähige Suchmaschinen übertragen die angemeldeten Seiten innerhalb von 24 Stunden in die Datenbank. Bevor die Daten der angemeldeten Seite in die Datenbank übertragen werden können, muss ein Abfragezyklus bis Punkt 8 durchlaufen werden. In der Regel wird sofort bei der Anmeldung überprüft, ob die URL überhaupt vorhanden, d.h. erreichbar, ist.

4.) Der Datensammler wird oft als Robot, Spider oder Crawler bezeichnet. Der Robot ist ein Programm das Anfragen in das WorldWideWeb an die Server des jeweiligen Hosters sendet. Es sorgt für die Aufnahme der Daten aus dem WorldWideWeb in die Suchmaschine. Er browsst die Links ab und parst die Webseiten, um die Daten für den Index herauszulesen. Startpunkt für Robots sind in der Regel URL-Listen, die dadurch entstehen, dass die Startseite einer Webpräsenz angemeldet wird. Erweitert werden diese Listen durch Links, die beim Erfassen der Webseiten ermittelt werden.

Da die großen Suchmaschinen täglich Millionen von Dokumenten spidersn, müssen die Robots sehr schnell arbeiten.

Es kann vorkommen, dass viele der abzufragenden Dokumente auf einem Server (ein oder mehrere Rechner die viele Domains beherbergen) eines Internet-Provider gespeichert sind. Daher sollte ein Robot sehr viele Abfragen parallel ausführen können. Ein ungeschriebenes Gesetz besagt, dass die Abfragen dem Server nicht mehr als ein Prozent der Systemressourcen abverlangen sollten, damit die Antwortzeiten des Servers für die normalen Browserabfragen auf einem vertraglichen Niveau bleiben.

5.) Die Darstellung der Domains ist rein symbolisch, da sich die unglaubliche Vielzahl der abzufragenden Domains nicht anschaulich darstellen lässt. Der Webserver beantwortet die Abfrage. Mit der speziellen Datei robots.txt und den robots Meta-Tags kann jeder Betreiber einer Website bestimmen welche Robots auf seine Seiten zugreifen dürfen.

Beispiel:

```
User-agent: Robot1  
Disallow: /logiles/  
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/ .

Ob sich die Robots tatsächlich daran halten, hängt von ihrer Programmierung ab. Die Praxis zeigt, dass es Robots gibt die derartige Festlegungen nicht beachten. Google hält sich z.B. jedoch strikt an Angaben in der robots.txt. Es lassen sich auch einzelne Verzeichnisse ( /privat) ausschließen. Werden Dateien oder Verzeichnisse mit einem Zugriffsschutz (Passwort) versehen, können diese auch nicht erfasst werden.

6.) Die angeforderte(n) Datei(en) wird/werden übertragen. In der Regel werden für eine Webseite mehrere Dateien übertragen, da ein HTML-Dokument meist noch weitere Elemente wie Bilder, Grafiken oder Sound enthält. Wird eine Web-Seite nicht gefunden, erhält der Index die Mitteilung, dass die Datei nicht mehr vorhanden ist. Diese wird dann, bei einigen Suchmaschinen, aus dem Index gelöscht.

Nach diesem Prinzip können Dateien manuell aus der Datenbank der Suchmaschine entfernt werden. Sie melden die URL wo sich die Datei befand. Die Suchmaschine erhält vom Robot die Information „Datei nicht gefunden (Fehlermeldung 404)“. Alle URL's mit diesem Statuscode werden bei den meisten Suchmaschinen nach ein bis zwei Tagen automatisch aus dem Index gelöscht.

7.) Die Daten werden vom Robot erfasst und zum Index weitergeleitet. Aus den Daten, die der Robot auf die Anfrage hin erhält, parst er die Links und alle für die Suchmaschinen wichtigen Informationen. Er sendet weitere Anfragen an den Server und erhält als Antwort wiederum neue Daten. Dieser Vorgang läuft solange bis alle Daten abgefragt wurden, oder eine Zeitbeschränkung den Vorgang beendet.

8.) Das Indizierungsprogramm wertet die ihm übermittelten Daten aus. Es werden nur die Daten aufgenommen, die in der Datenbank eine entsprechende Rubrik bekommen. Der Inhalt eines Index wird durch die Sitebetreiber gestaltet. Mit dem Seiteninhalt und der Gestaltung von Titel und Metatags sorgen sie dafür, wie die Inhalte dargestellt werden.

Der unsichtbare Teil einer Webseite enthält Meta-Tags, die von einigen Suchmaschinen ausgewertet werden, von manchen jedoch auch nicht (insbesondere Google). Diese Suchmaschinen gehen davon aus, dass Meta-Tags nur die Ergebnisse verfälschen (Spamming).

Im Datenbestand von Fireball enthalten rund 19% der Webseiten den Meta-Tag keywords. Bei Fireball gibt es diesbezüglich sogar einen umfangreichen Meta-Tag Generator, der es ungeübten Nutzern das Erstellen von Meta-Tags ermöglicht.

Der Umfang der aufzunehmenden Daten ist bei jeder Suchmaschine unterschiedlich. Einige Volltextsuchmaschinen wie Google werten den gesamten Text einer Webseite aus, andere begnügen sich mit den ersten Sätzen einer Webseite. Google indiziert z.B. auch Nicht-HTML-Dokumente wie PDF- und Powerpoint-Dokumente. Auch wird sogar versucht, Bilder zu indizieren (-> Google-Bildersuche).

In folgende Kategorien der Datenspeicherung können Suchmaschinen eingeteilt werden:

- **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

Alle großen und wichtigen „Suchmaschinen“ arbeiten als Volltextsuchmaschinen.

Für die Wortstatistiken wird nicht jedes einzelne Wort abgelegt, sondern jedes Schlüsselwort wird nur einmal abgelegt, mit der Information wie oft es in der Webseite an bestimmten Stellen (Überschrift, Metatags, Text, Position im Text) enthalten ist.

Links aus den Webseiten werden extrahiert und als Suchaufgabe für den Robot zusammengestellt. Alle Links einer Webseite werden weiterverfolgt. Deshalb genügt es normalerweise, eine Seite der Domain anzumelden. Die weiteren Seiten werden dann beim nächsten Aktualisierungslauf erfasst.

Der Zeitraum bis zum Erfassen aller Seiten ist sehr unterschiedlich. Bei Google dauert es in der Regel 3-5 Wochen.

9.) Das Datenbankprogramm verwaltet die Daten der Datenbank. Es verarbeitet die Anfragen der Benutzer und bereitet die Daten für die Anzeige im Webserver auf.

10.) Wenn eine Suchanfrage in die Eingabe-Maske eingegeben wird, wird die Anfrage in eine computerverwertbare Form umgewandelt und an die Datenbank übermittelt.

Je genauer die Abfrage formuliert ist, desto kleiner und präziser wird die Treffermenge. In der einfachen Suche ist die Phrasensuche oder eine mehrfache UND-Verknüpfung nützlich. Die erweiterte Suche (oder Profisuche) stellt wesentlich umfangreichere Möglichkeiten wie die Nutzung boolescher Operatoren zur Verfügung.

11.) Das Datenbankprogramm stellt gemäß der Anfrage die Suchergebnisse mit den Daten aus dem Index zusammen. Dabei werden die jeweiligen Rankingkriterien wie z.B. das Page-Rank-Verfahren bei Google berücksichtigt.

12.) Nun erfolgt die Ausgabe der Ergebnisse. Die Darstellungsform hängt stark vom Suchsystem ab. In der Regel werden Dokumententitel sowie Inhalte der Metatags bzw. die ersten Zeilen des Dokumentes dargestellt. Bei Google wird die Umgebung des gesuchten Begriffes dargestellt. Die Ergebnisse werden in einer von der Software festgelegten Reihenfolge präsentiert (Ranking). Bei einigen Suchmaschinen lassen sich die Ergebnisse (z.B. nur Dateien von dieser Domain) nachträglich sortieren. In der erweiterten Suche bieten die meisten Suchmaschine Möglichkeiten, das Ranking nach entsprechenden Gesichtspunkten zu beeinflussen.

Mit dieser Darstellung soll deutlich werden:

- Web-Dokumente werden nie beim Absenden der Suchanfrage direkt bei der Suchmaschinenabfrage aufgerufen, sondern nur das "Abbild", die zu diesem Dokument gespeicherten Daten in der Datenbank der Suchmaschine
- Robots erledigen lediglich die Abfragen der Seiten und geben diese Seiten zur Indexierung und Speicherung an den Server weiter
- nur wenige Suchmaschinen erfassen die gesamten Dokumente. Meistens 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. Meta-Sucher, die mehrere Suchmaschinen abfragen, erfassen etwas größere Teile des Webs, liefern dafür aber ungenauere Ergebnisse.

## Kapitel 4

### Suchmaschinen-Algorithmen

#### 4.1. Ranking-Algorithmus

*Quelle: Suchmaschinentricks.de (aufgerufen im Dezember 2003)*

Mit mehreren Millionen, bei Google ca. 3 Milliarden (siehe Abbildung 2.2) gespeicherter Webseiten ist es leicht nachvollziehbar, dass Suchmaschinen für die meisten Suchanfragen Tausende von Ergebnissen liefern. Ein Verfahren, die vielen gefundenen Seiten in eine sinnvolle Reihenfolge zu bringen, nennt man **Ranking Algorithmus**.

Dabei bedeutet sinnvoll, dass die Ergebnisse, die am besten zur Suchanfrage passen, möglichst weit vorne erscheinen sollen, also nach Relevanz sortiert wird.

Diese Aufgabe bereitet mehrere Schwierigkeiten, da Suchanfragen sehr oft nicht eindeutig sind. Es ist schwierig zu wissen, welche Ergebnisse jemand erwartet, der einfach nach dem Begriff "download" sucht, oder nach mehrdeutigen Begriffen wie "Bank", sofern er nicht explizit die "Semantik" angibt und diese auch in den Webseiten abgelegt ist.

Abgesehen davon ist es problematisch, die Relevanz eines Textes (Bilder und andere Multimediaelemente werden von Suchmaschinen komplett ignoriert) nur nach den darin vorkommenden Schlüsselbegriffen zu bewerten. So kann eine Schiller-Biographie durchaus das entscheidende Wort "Schiller" nur ganz selten benutzen; um Wiederholungen zu vermeiden, werden Synonyme eingesetzt wie Dichter oder Schriftsteller oder Abkürzungen wie „S.“

Eine Suchmaschine muss bestmöglich erkennen, ob ein entsprechender Text, der vielfach das Wort Schiller einsetzt, vollständig über Schiller handelt, oder nur einen sehr geringeren bis gar keinen Bezug (z.B. dieser Text) zu Schiller hat.

Der wesentliche Trick für die Suchmaschinen besteht darin, nicht nur die Anzahl und relative Häufigkeit der einzelnen Wörter zu berücksichtigen, sondern auch die Position. HTML ist viel mehr eine Methode Text zu strukturieren als Webseiten zu gestalten. Deshalb eignen sich die entsprechenden HTML-Tags dazu, sie für eine Relevanzbestimmung zu benutzen.

#### **Die wichtigsten Ranking-Kriterien**

Suchmaschinen bewerten Seiten als besonders relevant, wenn der entsprechende Suchbegriff z.B. im Titel oder innerhalb einer Überschrift vorkommt. Ebenso werden Begriffe, die im Text weiter oben stehen, als besonders relevant erachtet. Für Suchanfragen, die zwei oder mehr Worte enthalten, etwa "Schiller Biographie", ist es wichtig, dass beide Suchbegriffe möglichst nah im Text bzw. im Titel stehen. Begriffe, die tief in verschachtelten Tabellen stehen, werden tendenziell eher schlechter bewertet.

Es spielt aber auch die Häufigkeit eine Rolle. Zumeist ist die relative Häufigkeit, die oft auch als Dichte bezeichnet wird, wichtiger als die absolute Anzahl. Ein relativ kurzer Text, in dem mehrmals (z.B. 4-5mal) das Wort "Schiller" vorkommt, wird demnach als relevanter für den Suchbegriff "Schiller" erachtet, als ein sehr langer Text mit z.B. 10-15 Vorkommen. Aus diesem Grunde haben auch kurze Seiten häufig bessere Positionen.

Auch das Vorkommen innerhalb der Meta Tags "Keyword" oder "Description" kann eine höhere Relevanz, damit also ein besseres Ranking bedeuten, wobei die Bedeutung von Meta-Tags immer unwichtiger wird und von Google z.B. gar nicht beachtet werden. Meta-Tags sind Zusatzangaben im Head-Bereich der HTML-Datei. Neben den Überschriften (<h1>-<h6>) werden oft auch andere HTML-Auszeichnungsmöglichkeiten, wie etwa <strong>, <u>, <i>, berücksichtigt.

Wichtig für die meisten Suchmaschinen ist auch die URL der Seite. Damit ist nicht nur die Domain gemeint, sondern auch der Pfad und der Dateiname auf dem Webserver.

Die oben angesprochene Schiller-Biographie wird von den meisten Suchmaschinen besser gerankt, wenn man sie z.B. "biographie.htm" benennt und sie in ein Verzeichnis namens "Schiller" ablegt. Die URL hat dann folgende Form:

<http://www.domain.de/Schiller/biographie.htm>

Zuletzt eignen sich ebenfalls die URL oder die Beschreibung von Links (also der a-Tag) und der Beschreibungstext für Grafiken (alt-Attribut des img-Tags) als zusätzliche Relevanzkriterien. Und einige Suchmaschinen bewerten offenbar auch das letzte Änderungsdatum, wobei "frische" Seiten besser bewertet werden.

### **Zusammenfassung der Ranking-Kriterien**

Die wichtigsten Stellen innerhalb der HTML-Datei für die Schlüsselbegriffe:

- Titel
- Überschriften h1, h2, ...
- URL
- Meta Keywords und Description (Ausnahme Google)
- ALT-Attribut des IMG-Tags
- Links (<a href=""></a>) (sowohl Ziel als auch Beschreibung)
- Hervorhebungen: strong, underlined, italic

Die Begriffe sollten weit oben stehen im Text, zusammengehörige Suchbegriffe stehen auch im Quelltext nahe zusammen. Eine relative Häufigkeit der Begriffe von 5 bis 8 Prozent scheint sinnvoll.

### **Ausblick**

Die hier vorgestellten Relevanzkriterien würden sehr gut funktionieren, wenn alle Autoren von Webseiten Ihre Seiten ehrlich mit Titel usw. beschreiben würden. Spätestens seit sich im Web sehr viel Geld verdienen lässt (z.B. durch Verlinkungen zu Partnerprogrammen wie [tradedoubler.com](http://tradedoubler.com), [zanox.de](http://zanox.de) oder [partnerprogramme.de](http://partnerprogramme.de)) haben die Suchmaschinen erhebliche Probleme. Einerseits versuchen Spammer, Ihre Site auch mit unehrlichen Mitteln zu optimieren, andererseits kümmern sich viele Webautoren gar nicht darum, im Web gefunden zu werden. Denn oft finden sich gute Seiten im Netz, die weder einen passenden Titel, Meta Tags usw. benutzen.

Deshalb gehen Suchmaschinen vermehrt dazu über, auch externe Informationen im Rankingmechanismus zu berücksichtigen. Eine zentrale Rolle dabei nimmt derzeit das Zählen externer Links ein: LinkPopularity; bei der Suchmaschine Google spielt ebenfalls das PageRank Verfahren eine große Rolle. Eine weitere Methode ist der DirectHit. Hier wird gezählt, wie oft

Suchmaschinenbenutzer auf ein bestimmtes Suchergebnis klicken. Und ein neuer Weg ist die Domain-Indizierung ("theme based indexing"), bei der nicht mehr die Relevanz einer einzelnen HTML-Seite bewertet wird, sondern eine Domain als Ganzes wird zur Bewertung herangezogen.

In den nächsten 2 Punkten werden die Linkpopularity und das PageRanking-Verfahren näher vorgestellt.

## **4.2. LinkPopularity**

Quelle: Suchmaschinentricks.de (aufgerufen im Dezember 2003)

### **Externe Links zählen**

Die meisten Suchmaschinen bewerten die Anzahl externer Links auf eine bestimmte Seite als Qualitätsmerkmal. Dahinter steckt der Gedanke, dass nur qualitativ wertvolle Seiten viele Links von anderen Seiten erhalten. Somit wird das Urteil vieler Menschen im Internet mit ins Ranking der einzelnen Suchmaschinen integriert.

Die genauen Details des LinkPopularity-Algorithmus sind von Suchmaschine zu Suchmaschine verschieden. Die meisten Seiten unterscheiden noch zwischen Link- und DomainPopularity. Bei der LinkPopularity werden alle Links, d.h. auch mehrere Links von einer Seite und deren Unterseiten, gezählt, bei der DomainPopularity wird die Anzahl der verschiedenen Domains gezählt, die auf eine bestimmte Seite verlinken.

Ein weiterer Ansatz ist die Relevanz der externen Links mitzubewerten. Das bedeutet, dass ein Link von einer in der Suchmaschine als sehr relevant bewertete Seite mehr wiegt, als ein Link einer Seite zu einem komplett anderen Thema.

### **LinkPopularity messen und steigern**

Einige Suchmaschinen bieten die Möglichkeit, die Anzahl externer Links auf eine Seite zu bestimmen. Die Suchanfrage bei Google ist hier „link:http://www.domain.de“ . Hier werden aber nur Links von Seiten ab PageRank (siehe Kapitel 4.3.) angezeigt. „link:http://www.yahoo.com“ bringt hier z.B. 660.000 Findungen.

### **Webkataloge - der Turbo für LinkPopularity**

Entscheidend für den langfristigen Erfolg des LinkPopularity Konzepts aus Sicht der Suchmaschinenbetreiber ist es, relevante Links höher zu bewerten als Links von irgendwelchen beliebigen Sites. Deshalb werden Links, die von bekannten Webkatalogen wie Yahoo oder dem Open Directory Project kommen, besonders hoch bewertet. Aus diesem Grund ist es für Seitenbetreiber sehr wichtig in die großen Webkataloge zu kommen.

### **Link-Text ist sehr wichtig**

Es wird zusätzlich bewertet, ob im Linktext der Suchbegriff vorkommt, der das Hauptthema der verwiesenen Seite enthält.

Beschäftigt sich eine Webseite mit dem Umweltschutz, dann sollte auch im Linktext das Wort Umweltschutz erscheinen. Ein Verweis mit dem Linktext "noch eine super Seite" meint es zwar gut, trifft aber nicht den Inhalt.

Ein Linktext "Umweltschutz für die Welt" ist in Sachen Relevanz besser. Jedoch noch besser ist, wenn nur das Wort "Umweltschutz" im Linktext enthalten ist. Dann erkennen Suchmaschinen eine 100%tige Relevanz

Wird im obigen Beispiel nur die Passage "für die Welt" verlinkt, hat die Zielseite, die über Umweltschutz berichtet weniger davon, im Sinne der Linkpopularität.

Die Relevanz trifft für interne und externe Links zu. Interne relevante Verweise erhöhen nur ihre eigene Relevanz für Suchmaschinen. Von externen Verweisen, profitiert auch die Zielseite.

Die Anzahl, wie oft der Suchbegriff im Linktext erscheinen sollte, ist für einzelne Suchmaschinen unterschiedlich.

### **Diese Suchmaschinen benutzen 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)

## **4.3. PageRank-Verfahren**

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

Die Suchmaschine Google ist bekannt für die hohe Qualität ihrer Ergebnisse. Eine wichtige Rolle spielt das für Google spezielle PageRank Verfahren, welches nachfolgend erklärt wird.

Das PageRank Verfahren ist nicht zu verwechseln mit der LinkPopularity, obwohl es ausschließlich die Verlinkung zwischen Webseiten und deren Bedeutung beurteilt. Eine Seite, die jedoch eine sehr hohe Linkpopularity besitzt, dürfte ebenfalls ein hohes PageRanking besitzen. Umgekehrt ist dies nicht unbedingt der Fall.

Die nachfolgenden Erläuterungen basieren auf den Erfahrungen mehrerer Suchmaschinen-Optimierer und sind nicht von den Betreibern der Suchmaschine Google bestätigt worden. Diese Beschreibung dürfte jedoch ziemlich realitätsnah sein.

Das PageRank Verfahren ist die Methode, mit der Google die Wichtigkeit einer Webseite bewertet.

Das Ranking wird von Google wie folgt benutzt:

- 1.) Alle Seiten finden, die zum Suchbegriff passen
- 2.) Ranking entsprechend den Seitenfaktoren, also dem Vorkommen des Suchbegriff in Seitentitel, Seitentext,...
- 3.) Berücksichtigung der Linktexte der externen Seiten
- 4.) Regulieren der Ergebnisse nach dem PageRank Verfahren und der LinkPopularity

Die Bedeutung der LinkPopularity hat abgenommen, da immer mehr versucht wird, mit Linkfarmen und ähnlichen Maßnahmen, die LinkPopularity zu beeinflussen. Sie verliert damit immer mehr Ihre ursprüngliche Bedeutung, nämlich dass ein Link von einer anderen Seite als Empfehlung anzusehen ist.

### **Grundlagen zum PageRank**

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, oft als Homepage bezeichnet, sondern jede einzelne Seite, die von Google indexiert wurde, 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.

### **Feststellung des PageRanks**

Mit der Google-Toolbar (<http://toolbar.google.com/intl/de/>) lässt sich sehr leicht feststellen, welchen PageRank-Wert eine Seite bekommen hat. Ein grüner Statusbalken zeigt den PageRank zwischen 0 und 10. Wird der Mauszeiger über den Statusbalken gehalten, erscheint die aktuelle PageRank Zahl.

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

Folgende Tabelle zeigt die Umsetzung der Werte:

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

Abbildung 4.1. Umsetzung der Google-PageRankwerte in die Google-Toolbar

Man erkennt schnell, dass es für Seiten mit zunehmendem Wert immer schwieriger wird, die nächsthöhere Stufe zu erreichen.

Die Werte der Toolbar sind nicht immer genau, sie beruhen mitunter auf einer Schätzung.

Die Google-Toolbar ist zwar nicht genau, aber die beste Möglichkeit etwas über das derzeitige PageRanking einer Seite zu erfahren. Das PageRanking wird von Google in eher unregelmäßigen Abständen (2-4 Wochen) neu berechnet.

### Weitere grundlegende Erkenntnisse

1. Je mehr Seiten eine Website aufweist, je höher wird der anfängliche PageRank.
2. Je weniger Links eine einzelne Webseite aufweist, je höher ist deren Wert für den PageRank der Seite auf die sie verweist
3. Aus 2. folgert, dass eine Sitemap nicht effektiv für den PageRank ist, da der weiterzugebende Wert, unter sehr vielen Links aufgeteilt werden muss.
4. Es sind nicht zwangsläufig Seiten mit dem höchsten PageRank, die den höchsten Feedback-Effekt bringen.  
Ein Link von einer Seite mit dem PageRank 3 kann besser sein als von einer Seite mit PageRank 6. Das trifft zu, wenn die Seite mit PageRank 3 insgesamt wesentlich weniger Links aufweist.

### **Die Nachteile des neuen PageRanks & Linkpopularity-Konzepts**

Das LinkPopularity Konzept und das PageRank-Verfahren hat auch einige Nachteile und wird von Sitebetreibern immer mehr manipuliert. Vor 1-2 Jahren erzielte Google mit diesen Methoden hervorragende Suchergebnisse. Dies lag daran, dass der Großteil der Links bisher tatsächlich als eine Art Empfehlung gesetzt wurden. 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. Im Netz blüht derzeit der Handel mit Linkverkäufen und Linkaustauschen.

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

## Kapitel 5

### Suchmaschinen-Spamming

#### 5.1. Spamming-Methoden

*Quelle: Suchmaschinentricks.de (aufgerufen im Dezember 2003)*

Alle Suchmaschinen benutzen Spamfilter, die ständig verfeinert und runderneuert werden.

Bekannte Spammingtechniken sind:

- 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, die sie in der Suche dann mit berücksichtigen (sollen).
- auf einer Seite werden mehrere title-Tags benutzt. Nur der erste soll von der Suchmaschine angezeigt werden, die restlichen dienen dazu, weitere Schlüsselbegriffe an prominenter Stelle unterzubringen.
- Schlüsselbegriffe werden innerhalb von HTML-Kommentaren notiert.
- in die Meta Keywords werden Begriffe aufgenommen, die nichts mit der Seite zu tun haben
- im Titel oder den Meta Keywords wird ein einzelner Begriff sehr oft (5-10 mal) wiederholt
- Weiterleitungen mit dem Meta Refresh Tag, per Javascript, CGI oder dynamischer Seiten (PHP, ASP) sollen den Suchmaschinen andere Inhalte präsentieren, als dem menschlichen Surfer.
- Cloaking - Verfahren, das dem Crawler einer Suchmaschine, erkannt über die User-Agent-Abfrage, eine andere Seite liefert als einem normalen Nutzer. Damit wird der Suchmaschine ein gut optimierter HTML-Code ohne Spielereien wie Flash oder JavaScript gezeigt, während den Usern die schön gestaltete Seite mit allen modernen Gimmicks angeboten wird.
- Doorwaypages - hoch optimierte Seiten, deren alleiniger Zweck es ist, bei den Suchmaschinen angemeldet zu sein und dort gut platziert zu werden. Diese Seiten haben keinen Inhalt, sondern wiederholen lediglich mehrfach den Begriff, für den sie optimiert wurden. Daher sind diese Seiten im Prinzip für den Nutzer einer Seite sinnlos und werden deshalb auch nicht von der eigentlichen Website aus verlinkt und sind nur über Suchmaschinen zugänglich.

## 5.2. Spamming-Beispiel

Eine Abfrage bei Google mit dem Suchbegriff „lastminute Urlaub“ zeigt, dass ein Großteil der ersten Seiten Spam-Seiten sind, die die speziell darauf ausgerichtet sind möglichst viele Visits über Suchmaschinen zu bekommen und diese sofort wieder über irgendwelche Affiliate-Links, die Provisionen bringen, weiterleiten.

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

### 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/](http://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/](http://www.lastminute-urlaub.ws/) - 6k - **Im Cache** - **Ähnliche Seiten**

### Lastminute Urlaub

**Lastminute Urlaub**. Startseite.  
[www.tohit.de/](http://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/](http://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/](http://www.lastminutos.de/) - 29k - **Im Cache** - **Ähnliche Seiten**

Abbildung 5.1. Spam-Beispiel in Google

### 5.3. Bekämpfung

Viele der in 5.1. vorgestellten Versuche können von den Suchmaschinen erkannt werden und führen zu einem deutlich schlechteren Ranking oder gar zur Verbannung der kompletten Website aus der Suchmaschine. Letztendlich 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.

### 5.4. Ausblick über zukünftige Algorithmen zur Spam-Bekämpfung

#### **SESC: Neue Anti-Spam-Algorithmen auf Google kommen "in Kürze" 12.11.2003**

*Quelle: suchmaschinentricks.de*

Die automatisierte Bekämpfung von Spam habe "höchste Priorität für Google", erklärte Michael Schmitt, Software Engineer von Google, auf der Search Engines Strategies Conference (SESC), die Montag und Dienstag in München stattfand.

Der aus Deutschland stammende Schmitt, der bei Google für die Bereiche Google News und Google Groups verantwortlich ist, machte deutlich, dass Google inzwischen die besondere Spamproblematik in Deutschland erkannte habe. "Wir sind eine amerikanische Company, wir sitzen alle in Kalifornien und hatten deshalb Google.de nicht im Blickfeld." Dadurch verbreitete sich im deutschen Teil des Index Spam sehr intensiv, ohne dass Google darauf reagiert hätte. "Auf Google.com ist Spam nicht so sehr das Problem; es gibt zwar dort auch Spam, aber weitaus weniger." Deshalb habe man sich in Kalifornien nicht so intensiv Gedanken über intensivere Anti-Spam-Methoden gemacht.

Dies sei, nicht zuletzt aufgrund verschiedener Presseberichte in Deutschland, nun anders. In Kürze werde Google neue und automatisierte Anti-Spam-Funktionen live schalten, erklärte der Software-Engineer. Allerdings werde auch weiterhin kein manueller Eingriff stattfinden: "Wir sind eine Technology Company, das ist nicht unser Stil", meinte Schmitt auf entsprechende Nachfragen. Er räumte aber ein, dass zumindest "vorübergehend auch eine manuelle Bekämpfung" denkbar sei, falls die technische Lösung doch noch länger dauern sollte.

Darüber hinaus versprach Schmitt, dass sich Google bemühen wird, den Kontakt zu Webmastern und vor allem zur SMO-Szene in Deutschland zu verbessern. Ein "deutscher GoogleGuy" wäre dabei nach Schmitts Ansicht die ideale Lösung, allerdings sei dies ein Zeitproblem, da GoogleGuy kaum mehr zu seinen eigentlichen Entwickleraufgaben käme. GoogleGuy ist ein Mitarbeiter von Google, der im US-amerikanischen WebmasterWorld-Forum regelmäßig mitliest und -schreibt und so den Kontakt zur SMO-Szene aufrecht erhält.

Welche Maßnahmen Google im Detail zur Spamabwehr einführen wird, wollte Schmitt nicht darstellen. Es wurde aber deutlich, dass ein Schwerpunkt dabei die Analyse der Verlinkung darstellen dürfte; client-seitige Weiterleitungen über JavaScript hingegen, so konnte man Schmitts Äußerungen interpretieren, dürften aber auch von den neuen Algorithmen nicht erkannt werden.

---

Man darf gespannt sein, wie sich das Spamming bei Google weiterentwickeln wird. Auf jeden Fall hat Google dieses Problem erkannt und dürfte nun intensiver dagegen vorgehen.

## Literaturverzeichnis

- [1] <http://www.polarluft.de>
- [2] <http://www-db.stanford.edu/~sergey/ddm.ps>
- [3] <http://www.aifb.uni-karlsruhe.de/Lehre/Winter2002-03/kdd/download/VII-3-Assoziationsregeln.pdf%20>
- [4] <http://www.webhits.de>
- [5] <http://www.thom-online.de>
- [6] <http://www.at-web.de>
- [7] <http://www.suchmaschinentricks.de>
- [8] <http://www.google.de>