

Suchmaschinen- Optimierung

Für Webentwickler

Auf einen Blick

Vorwort	11
1 Suchen im Web	13
2 Die Anatomie des World Wide Web	33
3 Architektur von Suchmaschinen	65
4 Gewichtung und Relevanz	111
5 Suchprozess	141
6 Onpage-Optimierung	167
7 Offpage-Optimierung	239
8 Spam	269
9 Aufnahme in die Suchmaschine	291
10 Monitoring und Controlling	309
A Literaturverzeichnis	325
B Quellen	327
C Abbildungsverzeichnis	331
Index	333

Inhalt

Vorwort	11
1 Suchen im Web	13
1.1 Webkataloge	15
1.1.1 Auswahl der Rubrik	17
1.1.2 Die Titelwahl	17
1.1.3 Vorsicht beim Beschreibungstext	19
1.1.4 Stichwörter mit Sorgfalt wählen	20
1.1.5 Häufige Fehler	20
1.1.6 Submit-Tools	20
1.2 Suchmaschinen	21
1.2.1 User-Interface	22
1.2.2 Hürden	23
1.2.3 Funktionen und Komponenten	24
1.3 Metasuchmaschinen	26
1.3.1 Formale Kriterien	27
1.3.2 Einsatzgebiet	28
1.3.3 Operatoren	29
1.3.4 Präsentation der Suchergebnisse	29
2 Die Anatomie des World Wide Web	33
2.1 Exkurs in HTML	34
2.1.1 HTML-Dokumentstruktur	35
2.1.2 Tags	36
2.1.3 Meta-Tags	38
2.1.4 Sonstige Meta-Tags	45
2.1.5 Cascading Style Sheets	47
2.2 Trägermedium Internet	49
2.2.1 Das Client-Server-Prinzip	50
2.2.2 TCP/IP	52
2.2.3 Adressierung der Hosts	53
2.2.4 Funktion und Aufbau eines URL	53
2.3 HTTP	55
2.3.1 Request	58
2.3.2 Response	61
2.3.3 HTTP live erleben	63

3 Architektur von Suchmaschinen 65

3.1	Dokumentgewinnung mit dem Webcrawler-System	66
3.1.1	Dokumentenindex	67
3.1.2	Scheduler	68
3.1.3	Crawler	70
3.1.4	Storeserver	72
3.1.5	Repository	77
3.2	Datenaufbereitung und Dokumentanalyse	78
3.2.1	Datenaufbereitung durch den Parser	82
3.2.2	Datennormalisierung	84
3.2.3	Wortidentifikation durch den Tokenizer	85
3.2.4	Identifikation der natürlichen Sprache	87
3.2.5	Grundformreduzierung durch Word Stemming	90
3.2.6	Mehrwortgruppenidentifikation	94
3.2.7	Stoppwörter	95
3.2.8	Keyword-Extrahierung	97
3.2.9	URL-Verarbeitung	101
3.3	Datenstruktur	101
3.3.1	Hitlist	102
3.3.2	Direkter Index	105
3.3.3	Invertierter Index	107
3.3.4	Verteilte Datenstruktur	108

4 Gewichtung und Relevanz 111

4.1	Statistische Modelle	113
4.1.1	Boolesches Retrieval	113
4.1.2	Fuzzy-Logik	114
4.1.3	Vektorraummodell	115
4.1.4	Relative Worthäufigkeit (TF)	118
4.1.5	Inverse Dokumenthäufigkeit (IDF)	119
4.1.6	Bedeutung der Lage und Auszeichnung eines Terms	120
4.1.7	Betrachtung des URL	121
4.2	Page-Rank	121
4.2.1	Link-Popularity	122
4.2.2	Das Page-Rank-Konzept und der Random Surfer	123
4.2.3	Page-Rank-Formel	124
4.2.4	Ein Beispiel zur Page-Rank-Berechnung	125
4.2.5	Effekte des Page-Rank	127
4.2.6	Der intelligente Surfer und weitere Einflussfaktoren	129
4.2.7	Bad-Rank	131
4.3	Click-Popularity	133
4.4	Cluster-Verfahren	136
4.4.1	Cluster-Verfahren im Einsatz	137
4.4.2	Vivisimo – ein Pionier	138
4.4.3	Single-Pass-Methode	139

5 Suchprozess 141

5.1 Arbeitsschritte des Query-Prozessors	142
5.1.1 Tokenizing	142
5.1.2 Parsing	142
5.1.3 Stoppwörter und Stemming	143
5.1.4 Erzeugung der Query	143
5.1.5 Verwendung eines Thesaurus	144
5.1.6 Matching und Gewichtung	144
5.1.7 Darstellung der Trefferliste	145
5.2 Suchoperatoren	146
5.2.1 Boolesche Ausdrücke	146
5.2.2 Phrasen	148
5.2.3 Wortabstand	148
5.2.4 Trunkierung	149
5.3 Erweiterte Suchmöglichkeiten	149
5.3.1 Sprachfilter	151
5.3.2 Positionierung	152
5.3.3 Aktualität	152
5.3.4 Domainfilter	152
5.3.5 Dateityp	153
5.3.6 Sonstige Suchmöglichkeiten	153
5.4 Nutzerverhalten im Web	154
5.4.1 Suchaktivitäten	155
5.4.2 Suchmodi	157
5.4.3 Welche Suchmaschine wird genutzt?	159
5.4.4 Was wird gesucht?	162

6 Onpage-Optimierung 167

6.1 Entwicklung eines Konzepts als erster Schritt	168
6.1.1 Zielgruppe und Zielsetzung	169
6.1.2 Durchführung	170
6.1.3 Spezielle Situation bei einem Relaunch	170
6.2 Strukturelle Vorbereitungen	172
6.2.1 Gültiges HTML	172
6.2.2 Einsatz von CSS	175
6.2.3 Seitenstruktur	176
6.2.4 Navigation	180
6.2.5 Frames	182
6.2.6 Die Startseite	189
6.2.7 Dateityp und dynamische Seiten	191
6.3 Schlüsselwort-Strategien	196
6.3.1 Erstes Brainstorming	199
6.3.2 Logbücher nutzen	200
6.3.3 Mitbewerber analysieren	200

6.3.4	Das Umfeld: Freunde, Kollegen und Bekannte	202
6.3.5	IDF überprüfen	203
6.3.6	Erste Bereinigung	204
6.3.7	Liste erweitern	205
6.3.8	Eigenschaften der Schlüsselwörter	208
6.3.9	Falsche orthografische Schreibweise	210
6.3.10	Getrennt oder zusammen?	211
6.3.11	Wortkombinationen und Wortnähe	212
6.3.12	Liste bereinigen	214
6.3.13	Finale Auswahl	215
6.4	Optimierung durch Tags	216
6.4.1	Title	216
6.4.2	Fließtext und die Keyword-Dichte	219
6.4.3	Aufzählungen	222
6.4.4	Texthervorhebungen	223
6.4.5	Überschriften	225
6.4.6	Links und Anchor-Text	226
6.4.7	Tabellen	229
6.4.8	Bilder und Image-Maps	231
6.4.9	Phantom-Pixel	232
6.4.10	Comment	233
6.4.11	Form und Input	234
6.4.12	Noscript	234
6.4.13	Iframe	235
6.5	PDF-Dokumente optimieren	237

7 Offpage-Optimierung 239

7.1	Webserver und Restriktionen	239
7.1.1	Webhosting	239
7.1.2	Restriktionen	241
7.2	Domainname und Verzeichnisse	242
7.2.1	Domainname	242
7.2.2	Verzeichnis- und Dateinamen	244
7.2.3	Verzeichnistiefe und Aktualität	246
7.3	Sitestructur	249
7.3.1	Redirects korrekt umsetzen	250
7.3.2	Deep Web	252
7.3.3	Seiten ausschließen (robots.txt)	255
7.4	Link-Popularity erhöhen	257
7.4.1	Interne Verlinkung optimieren	258
7.4.2	Das KAKADU-Prinzip	258
7.4.3	Qualitätskriterien potenzieller Linkpartner	260
7.4.4	An andere Webautoren herantreten	261
7.4.5	Eingehende Links erzielen	262
7.4.6	Link-Farmen und Google-Bomben	265
7.5	Click-Popularity erhöhen	266

8 Spam 269

8.1	Keyword-Stuffing	270
8.2	Unsichtbare und kleine Texte	272
8.3	Hidden-Links	278
8.4	Meta-Spam	279
8.5	Doorway-Pages	280
8.6	Cloaking	283
8.7	Bait-And-Switch	285
8.8	Domain-Dubletten	286
8.9	Page-Jacking	288
8.10	Sonstige Spammethoden	289

9 Aufnahme in die Suchmaschine 291

9.1	Suchmaschinen-Kooperationen	291
9.2	Die Anmeldung	293
9.2.1	Manuelle Anmeldung	295
9.2.2	Automatische Anmeldung	298
9.2.3	Aufnahmedauer	299
9.3	Kostenpflichtige Leistungen	301
9.3.1	Payed-Inclusion-Programme	302
9.3.2	Pay-Per-Click (PPC)	304

10 Monitoring und Controlling 309

10.1	Server-Monitoring	310
10.2	Logfile-Analyse	313
10.2.1	Anfragen pro Tag und Monat	315
10.2.2	Herkunftsland der Besucher	317
10.2.3	Seitenbesuche	317
10.2.4	Herkunft der Besucher	318
10.2.5	Besuche über Suchmaschinen	319
10.2.6	Suchbegriffe	320
10.2.7	Sonstige Informationen	321
10.3	Rank-Monitoring	322

A	Literaturverzeichnis	325
B	Quellen	327
C	Abbildungsverzeichnis	331
	Index	333

Vorwort

Sie glauben, dass sich niemand für Ihre Website interessiert? Nein, Ihr Webangebot wird einfach nicht gefunden. Die besten Sites der Welt wären nichts wert, gäbe es nicht den Generalschlüssel zum Web – die Suchmaschinen. Über 70 Prozent der Webuser starten ihre Online-Sitzung mit der Eingabe von Suchbegriffen in eine Suchmaschine. Um aus dem dichten Wald der unzähligen Webseiten herauszuragen, muss jedoch mehr getan werden, als bloß eine Site im Web zu veröffentlichen.

Dabei ist es eine Wissenschaft für sich, Webseiten für Suchmaschinen zu optimieren. Doch keine Sorge, in diesem Buch wird es nicht um unverständliche Theorie gehen, sondern um klare und verständliche Tatsachen. Erkenntnisse aus der Wissenschaft werden mit der Praxis und jahrelanger Erfahrung vermischt. Auf diese Weise wird die Welt der Suchmaschinen Schritt für Schritt erklärt, so dass Einsteiger wie Fortgeschrittene ihr Wissen über die Kunst der Suchmaschinen-Optimierung behutsam aufbauen und erweitern können. Insbesondere die fortgeschrittenen Leser werden mit Sicherheit neue Aspekte und Einblicke gewinnen können.

Denn bei den Suchmaschinen geht es nicht zuletzt um eine der faszinierendsten Aufgaben, die die Menschheit in der Informationsgesellschaft zu bewältigen hat – die Wiedergewinnung von verlorenen Informationen. Die Größe des World Wide Web steigt exponentiell an. Tausende von Webautoren veröffentlichen täglich neue Informationen, die der gesamten Welt zur Verfügung stehen. Doch niemand ist fähig, diese enorme Flut an Texten zu selektieren und Sinnvolles von Sinnleerem zu trennen. Nach welchen Kriterien sollte dies auch geschehen?

Die Wissenschaft des Information Retrieval, der Wiedergewinnung von Informationen, versucht automatische Verfahren zu entwickeln, damit selbstständige Programme Ordnung in den riesigen Datenbestand bringen. Nur mit Hilfe von Suchmaschinen, die unaufhörlich neue Ressourcen erfassen und auswerten, kann das Web überhaupt erschlossen werden. Sie schließen den Kreis zwischen dem Wissen des Webautors, seiner Website und dem interessierten Surfer.

Dabei soll es in diesem Buch in erster Linie um die Suchmaschinen im World Wide Web gehen. Doch auch für Suchdienste im Intranet und anderswo können die Ausführungen durchaus Gültigkeit beanspruchen. Denn die Web-Technologien von Google und Co. werden auch in zahllosen Unternehmensnetzwerken eingesetzt.

Nachfolgend steht nicht die reine Vermittlung von Fakten zum Thema Suchmaschinen-Optimierung im Vordergrund. Die rasante Entwicklung auf dem Markt der Suchdienste führt schnell dazu, dass Kochrezepte zur Optimierung beinahe schon dann veraltet sind, wenn sie veröffentlicht werden. Daher können Sie sich mit diesem Buch das notwendige Grundlagenwissen aneignen, das Sie langfristig dazu befähigt, eigenständig Optimierungsmaßnahmen am Puls der Zeit durchzuführen. Ferner erwerben Sie quasi nebenbei die Kompetenz, die umfangreichen Dienstleistungen und Kommentare in diversen Foren und sonstigen Publikationen fachkundig zu bewerten und einzuordnen.

Daher erhalten Sie zunächst eine Einführung in die Grundlagen des World Wide Web. Das Wissen über die Möglichkeiten und Begrenzungen des neuen Mediums hilft anschließend, die Funktionsweise von Information-Retrieval-Systemen, zu denen vor allem Suchmaschinen gehören, zu verstehen. Hier sollen Antworten auf die Frage im Mittelpunkt stehen, wie die Erfassung und Verarbeitung von Webseiten durch Suchmaschinen geschieht. Mit diesem Wissen gewappnet, sind Sie eigentlich schon in der Lage, selbstständig Optimierungen durchzuführen. Um jedoch den Einstieg zu erleichtern, wird ein weiterer Schwerpunkt auf die Optimierung als solche gelegt. Profitieren Sie in der zweiten Hälfte des Buches von den dargestellten Strategien und Vorgehensweisen, und lernen Sie Stolperfallen kennen, um nicht selbst deren Opfer zu werden. Mit diesem Know-how und Grundschatz von Erfahrungswerten werden Sie in der Lage sein, selbstständig in Sachen Suchmaschinen-Optimierung zu agieren und sich auf dem Laufenden zu halten.

Ich wünsche Ihnen viel Spaß beim Lesen dieses Buches und natürlich ebenso viel Erfolg bei der Optimierung Ihrer Webseiten. Zuvor möchte ich mich bei meinem Lektor, Herrn Stephan Mattescheck vom Galileo Press Verlag, für die vertrauensvolle Zusammenarbeit bedanken. Dieses Buch ist meinen Eltern, Verene und Peter Erlhofer, gewidmet, die mir über Jahre hinweg die Möglichkeit gaben, nicht nur mein Wissen über Suchmaschinen aufzubauen.

Über Ihre Anregungen und Kommentare würde ich mich sehr freuen:
erlhofer@mindshape.de.

Trier, den 28.02.2005
Sebastian Erlhofer

1 Suchen im Web

»Die Anzahl an Dokumenten im Index wächst stetig in beträchtlichem Ausmaß, jedoch nicht die Fähigkeit des Benutzers, diese auch anzuschauen. [...] Das Ziel des Suchens ist es, qualitativ hochwertige Suchergebnisse effizient anzubieten.«

– Sergey Brin, Lawrence Page (Erfinder von Google)

Das Internet enthält die gigantischste Informationsmenge, die der Mensch je geschaffen hat. Mechanismen zum schnellen und effektiven Auffinden von Informationen sind damit von zentraler Bedeutung geworden.

Ein Inhaltsverzeichnis, das alle Dokumente des World Wide Web enthält, gibt es leider nicht. Das ist aufgrund der dezentralen Struktur des Internets auch nicht möglich. Seit Entstehung des Webs 1991 haben sich daher verschiedene Strukturen entwickelt, um die Informationsflut zu bändigen und den Suchenden schnell ans Ziel zu führen.

Die Suchhilfen im Internet haben unterschiedliche Ausrichtungen und Ansätze. Für die Anbieter im Online-Sektor ist es unerlässlich zu wissen, über welche Wege Besucher auf ihr Angebot gelangen können und wie diese Mechanismen funktionieren, um noch effektiver die Besucherströme zum eigenen Vorteil lenken zu können.

Dabei gibt es zentrale Unterscheidungskriterien, nämlich wie Suchdienste ihren Datenbestand aufbauen, verwalten und aktualisieren. Die wichtigsten und gleichsam meistgenutzten Suchhilfen kann man in zwei Grundtypen unterteilen:

- ▶ **Webkataloge** sind verzeichnisbasierte Suchhilfen. Als Nachkommen der unorganisierten Linklisten, die zu Beginn des Webs noch genügten, werden Webkataloge mit ihrer komplexen Verzeichnisstruktur der heutigen Netzgröße gerecht. Sie sind nicht nur ein eigenständiges Recherchemittel, sondern spielen besonders im Hinblick auf die Suchmaschinen-Optimierung eine tragende Rolle.
- ▶ **Suchmaschinen** stellen den zweiten Grundtyp dar. Sie sind indexbasierte Softwareprogramme, die automatisch das World Wide Web durchsuchen und somit ihren Datenbestand stetig und selbstständig erweitern. Suchmaschinen stellen heutzutage das Kernelement der Recherche im World Wide Web dar. Eine repräsentative Studie der Bertelsmann-Stiftung Ende 2003

ergab, dass 91 Prozent aller deutschen Internetsurfer zumindest gelegentlich Suchmaschinen nutzen.

Neben den beiden Grundtypen gibt es weitere Formen von Suchhilfen im Web. Im Gegensatz zu den zentral organisierten Webkatalogen und Suchmaschinen verwalten verteilte Suchdienste die Informationen dezentral.

► **Metasuchmaschinen**

Sie scheinen auf den ersten Blick wie Suchmaschinen zu funktionieren, jedoch haben Metasuchmaschinen keinen eigenen Datenbestand. Stattdessen setzen sie bei der Suche auf den Datenbestand von Suchmaschinen und präsentieren daraus ihre eigene Ergebnisliste. Wie später deutlich werden wird, haben Metasuchdienste charakteristische Vor- und Nachteile, die je nach Suchziel abzuwägen sind.

► **P2P-Netzwerke**

Peer-To-Peer-(P2P-)Netzwerke sind mittlerweile weltweit als Medium für den Austausch von Musik- und Videodateien bekannt. Allerdings gibt es nicht nur P2P-Share-Dienste, sondern auch die weniger bekannten P2P-Suchmaschinen, die auf dem Prinzip von Napster, Kazaa und Konsorten basieren. Dabei legen die einzelnen Benutzer (peers) einen Ordner auf Ihrem Rechner mit Verweisen auf Ressourcen im Web an und beschreiben und bewerten diese im besten Fall zusätzlich. Suchanfragen werden dann über die einzelnen Peers geleitet und gesammelt. Ende 2002 erschien der berühmteste Vertreter dieser Klasse, das kostenpflichtige P2P-Programm OpenCola (das, nebenbei erwähnt, nicht in Zusammenhang mit dem weltweit bekannten Getränkeunternehmen steht). Allerdings konnte sich das Peer-Prinzip zur Suche im Web bis heute nicht durchsetzen, es handelt sich dabei vielmehr um eine Randerscheinung ohne Relevanz für den Web-Alltag.

► **Payed-Listing**

Ganz im Gegensatz dazu stehen die Recherchemöglichkeiten dieser Kategorie. Streng genommen handelt es sich hierbei eigentlich gar nicht um Suchmaschinen oder dergleichen, sondern vielmehr um Anbieter, bei denen Ranking-Positionen für Geld gekauft werden können. Dabei bekommt der Meistbietende für ein Stichwort den höchsten Platz. Auf das genaue Verfahren und weitere Zusammenhänge gehe ich eigens in Abschnitt 9.3, *Kostenpflichtige Leistungen*, näher ein.

Neben diesen Formen kann man sonstige Erscheinungen im Web unter der dritten Klasse, den **spezialisierten Suchmaschinen**, zusammenfassen. Darunter fallen die fortschrittlichen Bildsuchen, die nicht nur nach Stichwortübereinstimmung im Dateinamen suchen wie etwa Yahoo oder Google, sondern auch

Bilder finden, die Ähnlichkeit mit einem vorgegebenen Master-Bild haben und gewisse visuelle Kriterien erfüllen. Auch Googles Voice-Search, bei der über das Telefon Suchbegriffe übergeben werden können, ist dieser spezialisierten Klasse zuzuordnen. Beim letzteren Beispiel ist die Abgrenzung zur normalen Google-Suchmaschine jedoch etwas unklar, die Suchergebnisse werden nämlich traditionell auf dem Computer angezeigt.

Die gesamte Einteilung ist notwendigerweise mehr als prototypische Darstellung zu verstehen. Der stark wachsende Suchmaschinenmarkt und die Kommerzialisierung der Recherche im Web haben dazu beigetragen, dass die Grenzen fließend sind. So bieten alle großen Suchmaschinen-Betreiber mittlerweile zusätzlich eigene bzw. fremde Webkataloge auf ihren Webseiten an und beziehen diese mit in ihre Suche ein.

Im folgenden Abschnitt erhalten Sie einen Überblick über die wichtigsten hier genannten Recherchemöglichkeiten und lernen ihre jeweiligen Stärken und Schwächen kennen.

Vielleicht ist es zuvor an dieser Stelle passend, einige grundlegende Begriffe zu definieren. So ist in diesem Buch, wenn von **Website** gesprochen wird, der gesamte Webauftritt **gemeint**. Dagegen meint der Begriff **Webseite** lediglich ein einzelnes Dokument innerhalb der gesamten Struktur. Die **Homepage** entspricht dabei der Einstiegsseite einer **Website**.

1.1 Webkataloge

Zu Beginn müssen sicherlich die Webkataloge stehen. Denn sie haben einen entscheidenden Vorteil: Die Anmeldung setzt kein tieferes technisches Verständnis oder gar eine eigene Seitenoptimierung voraus. Daher empfiehlt sich der Eintrag in einen Webkatalog immer als erster Schritt vor der eigentlichen Suchmaschinen-Optimierung.

Ein Webkatalog, häufig auch Webverzeichnis oder Webdirectory genannt, ist im Grunde genommen eine Website mit thematisch geordneten Linklisten. Diese Listen sind hierarchisch in einzelne Rubriken gegliedert. Der Suchende gelangt so immer vom Allgemeinen zum Speziellen, bis er den Themenkomplex seines Interesses gefunden hat. Dabei unterstützen Querverlinkungen zusätzlich die Suche, um mehrdeutige Themengebiete über verschiedene Wege zu erschließen und »Verirrte« wieder auf den richtigen Pfad zu bringen. Den Endpunkt stellt eine Auflistung von Verweisen auf einzelne Webseiten dar.

Ist eine Anmeldung zur Aufnahme in eine Suchmaschine gar nicht oder nur sehr knapp nötig, muss bei Webkatalogen hingegen jeder einzelne Link ma-

nuell angemeldet werden. Das Anmelden ist hier jedoch im Sinne eines Vorschlags zu verstehen. Die meist umfangreichen Anmelde­daten werden an einen zuständigen Redakteur geleitet, der dann letztendlich entscheidet, ob und wie der Eintrag aufgenommen wird.

Der Redakteur versieht jeden URL mit einem Titel und einem knappen Beschreibungstext, der sich an den bei der Eintragung gemachten Vorschlag anlehnt. Die verfügbare Datenmenge umfasst daher meist nur den Link auf eine Webseite, in der Regel die Homepage, und einen kurzen Beschreibungstext.

Das Besondere an Webkatalogen ist deren redaktionelle Erstellung ohne Zuhilfenahme von Programm­routinen. Die Redakteure, neudeutsch auch Editors genannt, sind für die Pflege des Datenbestands zuständig. Genau das macht die besondere Qualität von Webkatalogen aus, denn jeder Eintrag ist von einem Mitarbeiter vor der Aufnahme gesichtet und als geeignet bewertet worden. Somit zählen im Vergleich zu den automatisierten Suchmaschinen neben dem faktischen Inhalt auch algorithmisch nicht erfassbare Faktoren wie die passende, seriöse Gestaltung oder die inhaltliche Qualität des Angebots.

Wie die einzelnen Einträge innerhalb eines Ressorts gegliedert und sortiert werden, ist von Webkatalog zu Webkatalog unterschiedlich. Klar unterscheiden lassen sich das gewichtete Verfahren und das ungewichtete Verfahren. Bei ersterem ordnet der Redakteur manuell eine Gewichtung, sprich Listenposition, zu. Dazu gibt es organisationsinterne Regelungen und nicht zuletzt die freie Meinung des Mitarbeiters. Bei Katalogen wie Allesklar [1] oder Excite [2] findet man dieses Vorgehen. Bei dem ungewichteten Verfahren wird der Datenbestand alphabetisch oder nach Datum sortiert. Bekannte Vertreter dieser Methode sind unter anderem Yahoo [3], Open Directory [4], Web.de [5] oder Bellnet [6].

Befürworter der Webkataloge nennen klar den Kernvorteil: Mit der intellektuellen Bewertung steigt die Präzision von Suchergebnissen im Vergleich zu indexbasierten Suchmaschinen. Kritiker halten dagegen, dass von Menschenhand erstellte Linklisten dem rasanten Wachstum des Webs nicht standhalten können. Einer Hand voll Redakteuren steht eine Meute von Webautoren gegenüber.

Damit haben, wie die derzeitige Web-Situation beweist, paradoxerweise beide Seiten recht. Jedoch wird oft ein wichtiges Kriterium außer Acht gelassen.

Das Beispiel des Fahrradhändlers Krause soll dies verdeutlichen. Er bietet auf seiner gewerblichen Website »Krause-Rad« zusätzlich Tipps und Tricks rund um die Pflege der Drahtesel an und möchte diese gerne in Webkatalogen anmel-

den. Leider hat er bislang noch nicht ausreichend Zeit gefunden, all sein Wissen auf der Seite zu präsentieren, so dass bis dato nur eine Hand voll Merksätze auf der Seite zu finden sind und diese somit einen recht mageren Eindruck macht.

Es ist klar, was passieren wird: Es wird bei einem kurzen Besuch des Redakteurs auf der Seite bleiben. Die Aufnahme von unfertigen, im Aufbau befindlichen Seiten in Webkataloge ist nahezu unmöglich. Im Gegensatz dazu hätte eine Suchmaschine die unfertige Webseite nach ihrem Programmschema wahrscheinlich aufgenommen. Des Weiteren achten Redakteure gerade in stark gefüllten Ressorts auf besonders hohe Qualität und Relevanz der Angebote. Häufig decken kleine Angebote nur das ab, was bereits mit einer umfassenden Website aufgenommen wurde. Die Wahl des geeigneten Suchdienstes hängt offensichtlich von der Suchanforderung ab. Webkataloge wie Suchmaschinen haben ihre Stärken.

1.1.1 Auswahl der Rubrik

Die Eintragung spielt bei Webkatalogen eine zentrale Rolle. Dabei kann ein falsch ausgefülltes Anmeldeformular selbst bei einer noch so guten Website zur Ablehnung führen.

Ein häufig gemachter Fehler ist eine falsche oder ungenaue Auswahl der Rubrik. Die Stärke der Webkataloge ist ihre feine Gliederung nach thematischen Kriterien, so dass besonders komfortabel Themengebiete erschlossen werden können, bei denen der Suchende keine passenden Stichwörter zur Suche parat hat. Geschieht die Zuordnung zu den Rubriken zu grob und ungenau, würden sich sehr bald unübersichtliche Listen mit undifferenzierten Themengebieten bilden und das Prinzip des Webkatalogs ad absurdum führen.

So würde etwa die (fertige) Webseite von Krause-Rad bei dem Open Directory Project (DMOZ) sicherlich unter `World > Deutsch > Wissen` zur Ablehnung führen, falls der zuständige Redakteur sich nicht selbst die Arbeit macht, seinen Kollegen von `World > Deutsch > Sport > Radsport > Verzeichnisse und Portale` den Vorschlag mitzuteilen. Es empfiehlt sich daher, vor der Anmeldung die Verzeichnisstruktur gründlich zu recherchieren und nach dem Grundsatz zu handeln, es dem Redakteur so einfach wie möglich zu machen.

1.1.2 Die Titelwahl

Eine weitere Hürde stellt der vorzuschlagende Titel und Beschreibungstext dar. Der Titel sollte knapp und aussagekräftig gewählt sein. Es ist ratsam, den Eigennamen der Website, wie etwa den Firmen- oder Vereinsnamen, mit in den Titel

zu übernehmen. Der Name sollte dann auch aus Platzgründen nicht mehr in dem darauf folgenden Beschreibungstext wiederholt werden.

Die Sortierung der Verweise innerhalb der Listen wird, wie Sie gesehen haben, unterschiedlich gehandhabt. Das Gedränge um die besten oberen Plätze ist gerade bei den überwiegend alphabetisch sortierten Listen groß. Lassen Sie sich dennoch nicht dazu hinreißen, Sonderzeichen oder Ziffern voranzustellen (@Radtipps, !Werksverkauf, 5-fach-billig), nur um möglichst weit oben zu stehen. Solche Titel werden in der Regel von den Redakteuren bereinigt oder führen zur Ablehnung des gesamten Eintrags.

Einen etwas anderen Weg der Datensortierung beschreiten Altavista und Bellnet mit ihren Webverzeichnissen. Dort werden an erster Stelle als **Sponsored Links** bezeichnete Einträge positioniert (siehe Abschnitt 9.3, *Kostenpflichtige Leistungen*). Diese stammen in beiden Fällen von dem Payed-Placement-Anbieter Overture. Erst nachstehend erscheinen dann die regulären Webseiten aus dem eigenen Datenbestand. Altavista baut hier sogar auf die Datenbestände des Open Directory Projects auf. Die Verweise werden jedoch nicht eins zu eins übernommen, sondern manuell von Redakteuren bewertet, was zu einer eigenen Listensortierung führt.

bellnet
Internetverzeichnis

Neue Suche:

Suchergebnis für **fahrrad**
Overture Suchergebnisse:
[Overture: Sponsored Links](#)

powered by **overture**

Klicken Sie hier und inserieren Sie mit Overture, dem Marktführer in Pay-for-Performance™ Suche.

- 1. Tolle Fahrräder bei OBI@OTTO**
Bei OBI@OTTO erhalten Sie Fahrräder und Zubehör in 1A-Qualität. Mach dir die Welt, wie sie dir gefällt. Ideen und Shopping rund um Hobby, Haus und Garten. ([Overture](#))
www.obialotto.de
- 2. Ihr Fahrrad von Quelle**
Ob Rennrad oder Mountain-Bike - Quelle hat die optimalen Räder für Anfänger und Profis. Qualität für Freizeit und Sport zu überraschenden Preisen. ([Overture](#))
www.quelle.de
- 3. Fahrrad - eBay Angebote zum Thema Fahrrad**
Sie suchen Fahrräder, Helme und, Bekleidung? Nutzen Sie eBay, den weltweiten Online-Marktplatz. 3... 2... 1... meins. ([Overture](#))
www.ebay.de

Weitere Suchergebnisse:

- 1. Fahrradartikel - Shops und Informationen**
Ballsport. Angelsport. Sportgeräte. Sportkleidung "Fahrradartikel" - ... Fahrradartikel" - Hier finden sie zu ihrer Anfrage ausgesuchte Empfehlungsadressen, derleistungsfähigsten Anbieter aus ... zahl von Tipps und Infos rund um das Thema "Fahrradartikel" ...
www.fahrrad-online.net
- 2. fahrrad.de - Hochwertiges Fahrrad und Mountainbike mit Preisvorteil | fahrrad.de - Fahrräder und Mountainbikes**
Sie suchen ein hochwertiges Fahrrad oder Mountainbike, aber mit Preisvorteil ? Bei fahrrad.de sind Sie richtig ! Das fahrrad.de Angebot: Fahrräder, Mountainbikes u.v.m.... Auf fahrrad.de kommen ausschliesslich Markenfahrräder bester Qualität zum Verkauf, welche sich ...
www.fahrrad.de

Abbildung 11 Sponsored Links bei Bellnet (gekürzte Darstellung)

1.1.3 Vorsicht beim Beschreibungstext

Der Titel allein ist selten aussagekräftig genug, um den entscheidenden Klick für sich zu gewinnen. Der Beschreibungstext soll dem Suchenden weiterführende Informationen bieten und über das zu erwartende Angebot aufklären.

Erfahrungsgemäß bestehen die Fehler beim Beschreibungstext überwiegend in unnötiger Prahlerei und übertriebenem Gebrauch von Großschreibung und Ausrufezeichen. Sie sollten auch allzu werbende und nichts aussagende Sätze vermeiden.

*Hier erfährt man alles über FAHRRÄDER, viele TIPPS UND TRICKS!
Besuchen Sie uns jetzt!!!!*

Dieser Satz lädt nicht gerade dazu ein, den Link zu klicken, finden Sie nicht auch? Das könnte man wesentlich eleganter lösen:

Umfangreiche Tipps und Tricks zur Wartung, Reinigung, Pflege und zum Ausbau für Rennrad, Mountainbike und andere Radtypen.

Der Text liefert nüchtern und objektiv echte Informationen über den zu erwartenden Inhalt.

Sorgfältig geschriebene Texte sind außerordentlich wichtig für die Aufnahme des Eintrags. Gerade bei Ressorts mit erhöhtem Aufkommen machen sich Redakteure nicht immer die Arbeit, die gelieferten Texte umzuschreiben, und lehnen die Anmeldung daher schneller ab.

Erfahrungsgemäß liegt die optimale Textlänge zwischen 15 und 25 Wörtern. Der Suchende, der die Liste mit Einträgen durchschaut, überfliegt die einzelnen Beiträge nur flüchtig. Diesen Zustand bezeichnet man auch als Scannen bzw. Scanning [7]. Sobald ein passendes Stichwort gefunden wurde, wird die entsprechende Textstelle intensiver gelesen. Ist jetzt die Aussage des Textes nicht mit wenigen Blicken zu erfassen, wirkt der Text auf den Leser nicht genügend informativ und zu langwierig. Die subjektiven »Kosten« stehen in diesem Moment nicht im passenden Verhältnis zu dem potenziellen Nutzen. Die Wahrscheinlichkeit, dass bei anderen Einträgen weiter gesucht wird, ist enorm hoch, und Sie haben einen potenziellen Besucher verloren. Kaum jemand ist so ungeduldig wie ein suchender Internet-Surfer.

Es hat sich bewährt, die Texte zur Eintragung nicht spontan in das Webformular zu tippen, sondern offline in einem Texteditor den Titel und die Beschreibung bedacht auszuformulieren und dann per Copy&Paste einzufügen. Das hat nebenbei den Vorteil, dass Sie die Texte auch über einen größeren Zeitraum mehrmals verwenden können, wenn Sie das Textdokument abspeichern.

1.1.4 Stichwörter mit Sorgfalt wählen

Der interessant und informativ gestaltete Beschreibungstext sollte darüber hinaus wichtige Schlüsselwörter enthalten, damit bei einer Stichwortsuche im Katalog eine gute Trefferchance besteht.

Neben der hierarchischen Verzeichnisstruktur stellen die meisten Webkataloge eine Stichwortsuche zur Verfügung, um Besuchern gewünschte Informationen schneller zugänglich zu machen. Im Gegensatz zu Suchmaschinen ist hier die Grundlage der Suche der in der Datenbank vorliegende Titel und Beschreibungstext, nicht der Inhalt der Webseite selbst. Umso wichtiger ist demzufolge die Wahl passender Stichwörter für die Beschreibung. Zudem ist zu beachten, dass die Stichwörter immer Substantive sind, kaum jemand gibt Verben in Suchformulare ein.

1.1.5 Häufige Fehler

Es gibt eine Hand voll mehr oder minder prominenter Fehler, die neben unfertigen Seiten immer wieder zur Ablehnung in Webkatalogen führen.

Während die Beachtung der jeweiligen Eintragsregeln selbstverständlich ist, werden immer noch häufig Seiten mit nicht funktionierenden Links (sog. broken links) vorgeschlagen. Dass die Redakteure alles andere als begeistert darauf reagieren, ist verständlich. Daneben sind störende animierte Grafiken, sinnfrei platzierte Musik und fehlende oder schwer zu findende Skip-Funktionen zum Überspringen von Flash-Intros häufige Ablehnungsgründe.

Je nach Redaktionsstruktur fallen sogar gleiche Eintragungen auf, die jedoch unter unterschiedlichen Domains gemacht sind. Wird ein solches Vorgehen von einem Mitarbeiter als absichtlicher Täuschungsversuch erkannt, führt dies in der Regel zum sofortigen Entfernen aller Einträge.

1.1.6 Submit-Tools

Im Web sind immer wieder Dienste oder Tools zu finden, die anbieten, die Eintragung in Webkataloge zu übernehmen. Hier sollten Sie jedoch gesunde Skepsis an den Tag legen. Meistens gibt es nur eine Chance, eine Website anzumelden. Daher sollte man die auch sinnvoll nutzen und wenigstens die großen deutschen Webkataloge wie Web.de, Open Directory Project und Yahoo per Hand eintragen.

Die Gefahr besteht vor allem darin, dass der Anmeldeverlauf sich verändert hat und die dadurch veraltete Software nicht mehr kompatibel ist und zu Fehlern im Anmeldevorgang führt. Ferner legen immer mehr Webkataloge besonderen

Wert auf die explizite Zustimmung ihrer Richtlinien bei dem Absenden der Daten – daher lehnen sie automatische Übermittlungen aus Prinzip ab. Wie dabei die serverseitige Erkennung solcher Submit-Tools funktioniert, wird in Abschnitt 2.3, *HTTP*, näher erläutert.

1.2 Suchmaschinen

Suchmaschinen sind umfangreiche Computerprogramme, mit denen man im Web systematisch suchen kann. Im Gegensatz zu den Webkatalogen, die nur einen sehr begrenzten Umfang an Websites erfassen, können die einmal programmierten Suchmaschinen selbstständig theoretisch das gesamte Web erfassen. Daraus leitet sich auch das wichtigste Merkmal einer Suchmaschine ab, nämlich das automatische Sammeln und Auswerten von Webseiten.

Hinzu kommt, dass das Wachstum eines Webkatalogs nicht mit dem rasanten Wachstum des World Wide Web mithalten kann. An einem durchschnittlichen Tag werden allein in Deutschland bei der verantwortlichen Organisation DENIC über 5000 neue Domains angemeldet. Vorsichtige Schätzungen gehen von einer Verdopplung der Webauftritte weltweit innerhalb von sechs Monaten aus.

Angesichts dieser gigantischen Informationsmenge ist es unumgänglich, Software einzusetzen, um auch weitflächig gezielt und effektiv nach Informationen suchen zu können.

- ▶ In den letzten Jahren sind Suchmaschinen wie Pilze aus dem Boden geschossen. Dagegen werden jedoch nur einige wenige Marktführer wirklich genutzt. So ergab eine repräsentative Umfrage der Bertelsmann-Stiftung Ende 2003, dass zwar 91 Prozent der deutschen Internetnutzer Suchmaschinen nutzen, dass allerdings die Verteilung auf die einzelnen Suchmaschinen sehr unterschiedlich ist. So gaben 70 Prozent der Befragten an, Google als Hauptsuchmaschine zu nutzen. Weit abgeschlagen auf dem zweiten Platz mit 10 Prozent stand Yahoo, auf dem dritten Platz mit nur noch 5 Prozent Lycos. Die Zahlen bestätigen weitestgehend die Ergebnisse, die bei dem Counterdienst Webhits [8] vorzufinden sind. Dabei muss erwähnt werden, dass bei Webhits die Daten nicht auf einer Befragung basieren, sondern auf automatischer Auswertung zahlreicher Browserdaten.
- ▶ Der Aufruf von Suchmaschinen geschieht gerade bei Dial-In-Providern noch oftmals automatisch direkt beim Browserstart, so dass im Vergleich zur direkten Befragung beispielsweise MSN höhere Ergebnisse erzielen konnte. Außerdem leitet der Internet Explorer von Microsoft nicht beantwortbare Domainanfragen in der Standardeinstellung an den hauseigenen MSN-Suchdienst weiter.

1.2.1 User-Interface

Im Alltagsgebrauch wird mit dem Begriff **Suchmaschine** meist nur die Webseite eines Suchdienst-Anbieters bezeichnet. Dass es sich hierbei nur um die Spitze des Eisberges handelt, wird im Folgenden deutlich zu sehen sein. Zunächst betrachten wir jedoch einmal das Offensichtliche, das User-Interface.

Auf der Startseite eines Suchmaschinen-Betreibers befindet sich eine Eingabemaske für Suchanfragen. Zwischen oftmals vorhandener Werbung, Links und Themenblöcken findet sich ein Eingabefeld, um einen oder mehrere Suchbegriffe einzugeben (siehe Abbildung 1.2). Die meisten Suchmaschinen bieten zusätzlich eine erweiterte Eingabemaske an, die für erfahrene Benutzer mehr Optionen bereitstellt. Alle gängigen Suchmaschinen heutzutage gestatten es, Suchbegriffe logisch zueinander in Verbindung zu setzen. Die meistgenutzten Hilfen aus der zugrunde liegenden booleschen Algebra sind dabei die AND- und OR-Verknüpfung (Näheres siehe Abschnitt 5.2.1, *Boolesche Ausdrücke*).

The image shows the Yahoo! Deutschland homepage. At the top, there are icons for Handy/SMS, Dating, Mail, and a Help button. The main logo is 'YAHOO! DEUTSCHLAND'. Below it, there are links for 'Assistent', 'Personalisieren', and 'Messenger'. A banner for 'Zuma' is visible. The search bar is prominent, with a 'Suche' button and options for 'Erweiterte Suche' and 'Einstellungen'. Below the search bar, there are navigation tabs for 'Web', 'Bilder', 'Verzeichnis', and 'Nachrichten'. The page is divided into several sections: 'Neu!' with a link to 'besten Konzerte'; 'Marktplatz' with links for 'Auto', 'Domains', 'Immobilien', 'Jobs', 'Reisen', 'Shopping', 'Finanzen', 'Nachrichten', 'Routenplaner', 'Sport', 'Wetter', 'Unterhaltung', 'Horoskope', 'Kino', 'Lotto', 'Musik', 'Spiele', 'Style', 'TV'; 'Organisieren' with links for 'Adressbuch', 'Fotos', 'Kalender', 'Mappe', 'Mein Yahoo!'; 'Kommunizieren' with links for 'Chat', 'Dating', 'DSL', 'GeoCities', 'Groups', 'Grußkarten', 'Handy/SMS', 'Mail', 'Messenger', and 'Alle Services'; 'Yahoo! als Startseite' and 'Toolbar mit Pop-Up Blocker'; 'Yahoo! Finanzen' with a news snippet about 'Volker Tietz exklusiv: DAX auf dem Weg zur 4000'; 'Yahoo! Sport' with a news snippet about 'Bundesliga: Alle Tore, alle Pleiten'; 'Mein Organizer' with a link to 'Anmelden'; 'Aktuelle Nachrichten' with a list of news items; and a 'HEIBOY' trailer advertisement.

Abbildung 1.2 Startseite von Yahoo

Hat der Benutzer eine erfolgreiche Suchanfrage gestellt, werden ihm eine oder mehrere Ergebnisseiten passend zu seinen Suchbegriffen angezeigt. Eine Ergebnisseite enthält dabei bei den großen Suchmaschinen (Google, Altavista,

Lycos, AllTheWeb, Fireball) in der Regel zehn Resultate. Eine Ausnahme stellt hierbei Yahoo dar, die Ergebnisliste umfasst dort 20 Einträge. Jedes einzelne gefundene Dokument ist bei allen Suchmaschinen mit einem verlinkten Titel, Beschreibungstext und weiteren anbieterspezifischen Diensten versehen (siehe Abbildung 1.3).

YAHOO! Suche
DEUTSCHLAND

Ihre Suche: [Erweiterte Web-Suche](#)
[Einstellungen](#)

Sie suchen: Seiten auf Deutsch weltweit

Web **Bilder** **Verzeichnis** **Nachrichten**

TOP 20 WEB-SITES von ca. 269,000

Diese Suche war beschränkt auf deutschsprachige Seiten. Für weitere Suchergebnisse versuchen Sie eine Suche [weltweit](#).

- Tipps + Tricks**

Tipps + Tricks rund um Fahrräder und Wandern: An dieser Stelle erhalten Sie regelmäßig **Tipps und Tricks**, die sich mit de... Lassen Sie sich überraschen. ... Das betrifft insbesondere die Lage, die Bezeichnung und den Verlauf der **Radwege** und **Rad**...
www.elberadeltag.de/kat_tipptricks.htm - 30k - [Im Cache](#)
- Radfreunde Rodalb Tipps & Tricks**

... **Tipps & Tricks**. Hilfreich für all diejenigen, die regelmäßig mit Ihrem Bike auf Tour... helm, natürlich auf dem Kopf und nicht... Verletzungen schützen! ...
www.juergenpaul.com/radfreunde_rodalb/site05.htm - 11k - [Im Cache](#)
- Tipps & Tricks - Vorderradschwinge**

Die Doppelschwinge (Serienmäßig ab Bj. 1957) ermöglicht eine besseres Fahrverhalten. Sie ist der einseitigen Schwinge von... Roller nachträglich auf die Doppelschwinge umgerüstet. Tipp 2
www.duerkopp.de/fronstsuspension.htm - 3k - [Im Cache](#)
- H@llenradspor.de - Die Ergebnisplattform der Zeitschrift HALLENRADSPORT**

Rad, **Kunstrad**, **Radspor**, **Hallenradspor**, **Saalsport**, **indoor**, **cycle** ... **Tipps & Tricks**. Auf dieser Seite finden Sie nützliche... den **Hallenradspor**. ...
www.hallenradspor.de/wsc66120906/Tipps_Tricks.htm - 60k - [Im Cache](#)
- Radwandern - Tipps & Tricks Test und Preisvergleich**

Ciao bietet Ihnen ein großes Forum um Ihre Gedanken und Meinungen zu **Radwandern - Tipps & Tricks** mit anderen auszut...
- **Tipps & Tricks**> **Radspor** - **Tipps & Tricks**. **Radwandern - Tipps & Tricks** ... 4321 Bewertungen. **Radwandern - Tipps &** ...
www.ciao.com/Radwandern-Tipps-Tricks-1183468-23k-Im-Cache-Weiter-Ergebnisse-von-dieser-Seite

Abbildung 1.3 Ausschnitt aus der Ergebnisliste von Yahoo

Geordnet werden die einzelnen Links nach angenommener Relevanz. Der erste Link passt nach Meinung der Suchmaschinen-Betreiber am besten, der letzte am wenigsten. Die einzelnen Suchmaschinen unterscheiden sich heutzutage weniger in ihren Programmstrukturen zum Anlegen des Datenbestandes, sondern vielmehr in der Anwendung der einzelnen Algorithmen und der dadurch entstehenden Gewichtung. Vergleicht man die Ergebnislisten verschiedener Suchmaschinen mit gleichen Suchbegriffen, ist dieser Unterschied deutlich zu erkennen.

1.2.2 Hürden

Typischerweise ergeben sich für den Benutzer von Suchmaschinen eine Reihe von Hürden, die es zu überwinden gilt.

So wird bei der überwiegenden Anzahl von Suchanfragen mit nur wenigen Stichwörtern eine Unmenge an Ergebnisseiten angezeigt, die sich der Benutzer im Grunde genommen gar nicht alle ansehen kann. So liefert zum Beispiel die Suche nach `tipps tricks rad` bei Yahoo ca. 158.000 Ergebnisse, und Google begnügt sich mit 46.500 Seiten. Berücksichtigt man zusätzlich, dass die wenigsten Benutzer überhaupt die zweite Ergebnisseite betrachten, zeigt sich die enorme Bedeutung effizienter Algorithmen für die Sortierung der Treffer – und nicht zuletzt die Bedeutung des Wissens, wie Suchmaschinen funktionieren und wie Seiten optimiert werden können.

Die endlos erscheinende Menge an dargebotenen Verweisen zwingt den Benutzer, sich auf das Relevanzurteil der Suchmaschine zu verlassen. Allerdings weiß jeder, der einmal eine Suchmaschine genutzt hat, dass sich hinter dem obersten Treffer nicht immer das Gewünschte befindet.

Gelegentlich trifft man obendrein auf tote Links. Die entsprechende Seite gibt es nicht mehr, sie ist temporär nicht erreichbar, oder der Inhalt hat sich geändert. Vorsichtige Schätzungen diesbezüglich gehen davon aus, dass in Suchmaschinen ein Blindanteil von zehn bis fünfzehn Prozent vorhanden ist.

Ein ganz anders geartetes Problem stellt die zunehmende Kommerzialisierung von Suchmaschinen dar. Die ersten Suchergebnisse sind nicht mehr zwingend die am besten geeigneten, sondern die am besten bezahlten. Derzeit sind diese zum Glück des Benutzers noch gesondert ausgezeichnet.

1.2.3 Funktionen und Komponenten

Im Gegensatz zur weit verbreiteten Meinung sind die dargestellten Suchergebnisse im Browser keinesfalls Live-Ergebnisse. Wenn eine Suchanfrage verarbeitet wird, sind zuvor bereits zahlreiche Systemkomponenten im Einsatz gewesen, um die möglichen Trefferdokumente zu verarbeiten. Wie bereits angesprochen wurde, ist die Weboberfläche einer Suchmaschine nur ein kleiner Teil dessen, was notwendig ist, um letztendlich eine brauchbare Ergebnisliste auf Suchanfragen liefern zu können. Typischerweise kann man einer Suchmaschine drei Funktionen zuschreiben. Dabei wird jede Funktion von einer Kernkomponente abgedeckt.

1. Datengewinnung

Bevor Daten ausgewertet werden können, müssen diese logischerweise zunächst einmal beschafft, gesichtet und in ein geeignetes Format konvertiert werden. Dafür ist das Webcrawler-System, das gelegentlich auch als Webrobot-System bezeichnet wird, zuständig. Seine Hauptaufgabe besteht

im Sammeln von Dokumenten aus dem Web. Dazu ruft das Webcrawler-System eine Seite nach der anderen auf und lädt diese herunter.

Das Webcrawler-System ist eine Zusammenstellung aus einzelnen Unterkomponenten und ist zusätzlich für die Überprüfung der Existenz und Veränderung von bereits im Datenbestand vorhandenen Dokumenten verantwortlich. Nur durch regelmäßige Vergleiche zwischen dem eigenen Datenbestand und dem Webangebot kann eine Aktualität gewährleistet werden.

2. Datenanalyse und -verwaltung

Nachdem die Dokumente lokal vorliegen, baut die nächste Komponente der Suchmaschine eine durchsuchbare Datenstruktur auf. Diese Komponente basiert auf einem so genannten **Information-Retrieval-System** (IR-System). Wie der Begriff *retrieval* (zu Deutsch *Wiedergewinnung*) sagt, sind Informationen in großen Datenbeständen zunächst verloren gegangen und müssen erst wiedergewonnen werden. Auf das Web übertragen bedeutet das eine schier unbegrenzte Anzahl von Texten, die für den Computer vorerst nichts anders darstellen als eine Aneinanderreihung von Buchstabenkombinationen. Um die Daten untersuchen zu können, werden die verfügbaren Dokumente in eine zur Verarbeitung günstige Form umgewandelt. Diese auf das Wesentliche reduzierten Texte bezeichnet man als **Dokumentenrepräsentation**. Sie stellen die Grundlage dar, anhand derer das Information Retrieval System automatisch nach bestimmten Methoden Werte vergibt, die auch als Gewichte bezeichnet werden. Jedes Dokument besitzt somit einen festgelegten Relevanzwert aufgrund seines Gewichts. Dieser gilt immer in Bezug auf ein bestimmtes Schlagwort. Findet das IR-System mehrere benutzbare Schlagwörter, wird für jedes ein eigenes Gewicht errechnet. Die Zuordnung der Schlagwörter, auch Deskriptoren genannt, bezeichnet man in diesem Zusammenhang als **Indexierung**.

3. Verarbeiten von Suchanfragen

Während das gesamte bisher erwähnte System Tag und Nacht daran arbeitet, die Datenbasis zu erweitern und aktualisieren, stellt der **Query-Prozessor** oder auch **Searcher** die Funktionalität dar, die man gemeinhin von einer Suchmaschine erwartet. Der Query-Prozessor stellt, wie zu Beginn des Abschnitts gezeigt, über das Webinterface die Schnittstelle zum Benutzer dar. Anhand der gegebenen Stichwörter wird aus dem Index des IR-Systems eine gewichtete, also sortierte Liste von Einträgen erzeugt. Diese Liste reichert der Query-Prozessor mit weiteren Informationen aus dem Datenbestand wie etwa dem Datum der Indexierung an. Abschließend wird eine Listenansicht für den Benutzer bereitgestellt, die im Browser als bekannte Ergebnisliste angezeigt wird.

Mit dem Wissen über die drei Kernkomponenten lässt sich auch endgültig erklären, weshalb es keineswegs Zufall ist, dass Suchmaschinen bei Eingabe gleicher Stichwörter teilweise gravierend unterschiedliche Ergebnisse anzeigen. Denn schon bei der Datenerfassung unterscheiden sich die Methoden der Suchmaschinen, da jede Suchmaschine andere Websites in unterschiedlicher Tiefe aufnimmt. Bei der Dokumentauswertung hängt das errechnete Gewicht vom Umfang der ausgewerteten Textpassagen ab. Wurden früher nur die ersten Passagen des Textes auf Webseiten beachtet, findet heute bei allen großen Suchmaschinen der gesamte Text einer Seite Beachtung. Unterschiede gibt es insbesondere noch bei der Beachtung von unsichtbaren Texten, Bildinformationen und HTML-Kommentaren. Zu guter Letzt wirkt sich im dritten Schritt die Wahl der Ranking-Algorithmen und deren Feinabstimmung auf das Suchergebnis aus.

Das beschriebene, vollkommen automatisierte Verfahren setzt ein striktes Regelwerk voraus. Genau hier liegt der Vorteil für den Webseiten-Anbieter. Der Redakteur eines Webkatalogs entscheidet nach mehr oder weniger freien Mustern über die Aufnahme und die Bewertung eines Eintrags. Die Suchmaschine hingegen behandelt jede Seite gleich. Kennt man die Faktoren, die eine hohe Gewichtung und Relevanzeinschätzung bewirken, kann man diese optimal ausnutzen, um eigene Seiten zu optimieren und Spitzenpositionen zu erreichen. Die Idee der Suchmaschinen-Optimierung ist geboren.

Jedoch wissen natürlich auch die Suchmaschinen-Betreiber um diese Schwäche und halten daher ihre Algorithmen und Feineinstellungen geheim. Sie verändern diese regelmäßig nicht nur zur reinen Verbesserung der Suchmaschine, sondern auch, um zu verhindern, dass eine gezielte Optimierung zu hundertprozentigem Erfolg führt.

Im Fortlauf dieses Buches werden daher die grundsätzlichen Verfahren und Methoden, die Suchmaschinen einsetzen, näher erläutert werden. Mit der genauen Kenntnis können Sie dann Websites eigenständig optimieren und sind in der Lage, auf Veränderungen kompetent zu reagieren.

Dabei folge ich an Stellen, an denen eine Differenzierung der einzelnen Komponenten nicht zwingend erforderlich ist, der Einfachheit halber dem alltäglichen Sprachgebrauch und spreche im Allgemeinen von »Suchmaschine«.

1.3 Metasuchmaschinen

Metasuchmaschinen erlauben die gleichzeitige Suche bei mehreren anderen Suchdiensten von einer einzigen Webseite aus. Die Metasuchdienste zeichnen sich dadurch aus, dass sie keinen eigenen Datenbestand besitzen, sondern nur

über ihre eigene Benutzeroberfläche via HTTP-Request auf die Webseiten anderer Suchmaschinen-Anbieter zugreifen. Die Suchanfrage wird also parallel weitergeleitet, und die zurückgelieferten Ergebnislisten der angesprochenen Suchmaschinen werden gesammelt und für die eigene Listenaufstellung verwertet.

Das Ablaufschema bei einer Suchanfrage durch den Benutzer ist dabei prinzipiell immer gleich:

1. Eingabe der Stichwörter in das Webinterface der Metasuchmaschine durch den Benutzer
2. Konvertieren der Suche für die jeweiligen Suchmaschinen
3. Paralleles Absenden der Suche per HTTP-Request und Warten auf Antwort
4. Einsammeln der HTML-Ergebnislisten und Konvertieren in weiterverarbeitbare Daten
5. Analyse der Listen, Entfernen von Dubletten und Anwendung eigener Kriterien zur Erzeugung eines Rankings
6. Darstellen der eigenen Ergebnisliste

Fälschlicherweise werden oftmals auch Webseiten mit Schnittstellen zu Suchmaschinen als Metasuchdienste bezeichnet. Bei diesen so genannten **All-In-One-Formularen** handelt es sich lediglich um die Auslagerung des Suchmaschinen-Textfeldes zur Eingabe von Stichwörtern. Die Verarbeitung der Suchbegriffe und Darstellung der Ergebnisliste übernimmt jedoch wiederum der entsprechende Suchdienst. Damit erhoffen sich die Suchmaschinen-Anbieter einerseits höhere Nutzerraten und bieten andererseits den Website-Betreibern unter anderem die Möglichkeit, die Suchergebnisse nur auf die spezielle Website einzuschränken.

1.3.1 Formale Kriterien

Aufgrund einiger Unsicherheiten wurden bereits im Juli 1998 bei einer Tagung der **Internet Society** in Genf [9] klare formale Kriterien vorgeschlagen, anhand derer eine Metasuchmaschine definiert werden kann. Dabei müssen sechs der insgesamt sieben Kriterien auf einen Suchdienst zutreffen, damit er als Metasuchdienst bezeichnet werden kann.

1. Parallele Suche

Die Metasuchmaschine muss wirklich parallel suchen, es darf sich nicht um ein All-In-One-Formular handeln.

2. Ergebnis-Merging

Die Ergebnisse müssen zusammengeführt und in einem einheitlichen Format dargestellt werden.

3. Dubletten-Eliminierung

Gleiche, mehrfach vorhandene Treffer müssen erkannt und entfernt werden.

4. AND- und OR-Operatoren

Für logische Operationen müssen mindestens die Operatoren AND und OR zur Verfügung stehen und an die abzufragenden Suchmaschinen weitergeleitet werden.

5. Kein Informationsverlust

Bietet ein Suchdienst eine Kurzbeschreibung einer Fundstelle an, muss diese übernommen werden.

6. Search Engine Hiding

Die spezifischen Eigenschaften der Quell-Suchmaschinen dürfen für Anwender keine Rolle spielen.

7. Vollständige Suche

Die Metasuchmaschine soll so lange in den Trefferlisten der abzufragenden Suchdienste suchen, bis diese keine Treffer mehr liefern.

1.3.2 Einsatzgebiet

Metasuchdienste eignen sich insbesondere für spezielle Informationsbedürfnisse, bei denen die einzelnen Suchmaschinen nur wenige Treffer aufweisen. Ferner zeigen Umfragen, dass die Zahl der bekannten und benutzten Suchmaschinen relativ gering ist. So kann der Einsatz einer Metasuchmaschine, die dem Benutzer unbekanntere Suchmaschinen nutzt, eine erstaunliche Anzahl von Ergebnissen bieten. Da Metasuchmaschinen in der Regel stets aktuell gehalten sind, werden auch neue Suchmaschinen oder ganz spezielle Datenbanken überwiegend schnell aufgenommen, die sonst noch gar nicht bekannt oder verbreitet sind.

Meist werden spezielle Metasuchmaschinen von bestimmten Nutzerkreisen eingesetzt, die oftmals sehr fachspezifische Anfragen haben. Schätzungen gehen davon aus, dass heute selbst die großen Suchmaschinen nur ein Drittel des gesamten World Wide Web erfassen. Dabei liegt das Hauptaugenmerk eher auf Themen von allgemeinem Interesse. Da die Suchmaschinen-Betreiber nach unterschiedlichen Kriterien das Web erschließen, erlaubt die Nutzung von Metasuchdiensten die Zusammenschließung dieser verschiedenen Bereiche.

Dies ist gerade bei den angesprochenen fachspezifischen Themen von besonderem Interesse. So bietet der Metasuchdienst OmniMedicalSearch [10] eine Suche in 32 medizinischen Suchmaschinen an. Dabei kann der Benutzer als spezielles Feature zusätzlich wählen, ob er Treffer für medizinische Profis (MedPro Search) oder für Anfänger (Basic Search) angezeigt haben möchte.

1.3.3 Operatoren

Ein charakteristischer Nachteil der Metasuchdienste ist die Beschränkung bei der Suchanfragenformulierung. Hier muss auf Operatoren zur logischen Verknüpfung der Stichwörter weitestgehend verzichtet werden. Diese Beschränkungen ergeben sich aus dem heterogenen Umfeld. Nicht jede Suchmaschine beherrscht die Verwendung von Operatoren im gleichen Maße, so dass oftmals der kleinste gemeinsame Nenner gerade einmal die AND- und OR-Verknüpfung ist. Nicht selten ist auch die Anfrageart bzw. die Schreibweise der Operatoren zu unterschiedlich.

Nun kam man vor einiger Zeit auf den Gedanken, die Suchanweisung für jeden Suchdienst so umzuformatieren, dass die Anfrage möglichst passend übersetzt wird. Das war eine sehr gute Idee, jedoch stößt man dabei recht schnell auf ein inhaltliches Problem. So soll beispielsweise die Suche nach Haus AND Garten Dokumente zurückliefern, die sowohl den Begriff »Haus« als auch den Begriff »Garten« enthalten. Unterstützt eine Suchmaschine diesen Operator nicht, wird als Stichwortsuche Haus Garten übermittelt. Daraufhin wird die entsprechende Suchmaschine alle Dokumente zurückliefern, in denen entweder »Haus« oder »Garten« vorkommt – also nicht nur ausschließlich Dokumente, in denen beide Begriffe gemeinsam auftreten. Aber genau das hat der Benutzer mit dem AND-Operator bezweckt.

Je komplexer die Operatoren sind, die in den Anfragen benutzt werden, desto wahrscheinlicher werden solche oder ähnliche Phänomene. Ein kleiner Ausweg bleibt den Metasuchmaschinen allerdings. Bei Anfragen mit Operatoren werden nur noch die Suchmaschinen abgefragt, die die verlangte Funktionalität besitzen. Dies führt zwar auf der einen Seite zu einer Reduktion von potenziellen Treffern, insgesamt fallen die Suchanfragen aber qualitativ höher aus. Ein Beispiel für eine solche Metasuchmaschine ist ixquick [11].

Nichtsdestotrotz ist eine komplette Annäherung der Metasuchmaschinen-Schnittstelle an die Schnittstelle einer einzelnen Suchmaschine kaum zu erreichen. Das wird besonders bei Betrachtung der erweiterten Suchfunktion deutlich. Diese beschränkt sich in der Regel auf die Auswahl der zu benutzenden Quell-Suchmaschinen. Eine Auswahl nach Dateiformat, die Einschränkung auf einzelne Domains etc. ist in der Regel nicht möglich.

1.3.4 Präsentation der Suchergebnisse

Das zentrale Problem der Metasuchmaschinen ist die Gewichtung der Verweise von den verschiedenen Suchmaschinen. Die Ranking-Algorithmen sind nicht bis ins Detail bekannt und somit auch nicht miteinander vergleichbar. Da bleibt

eigentlich nur, so scheint es zumindest, die Ergebnisse nach Suchdiensten zu gruppieren, um ein echtes Abbild der ursprünglichen Rankings zu bekommen.

Die Realität sieht allerdings wie immer etwas anders aus. Bei Metasuchdiensten werden diverse Formen der Darstellung von Suchergebnissen genutzt: Ein häufig angewandtes Verfahren ist die Übernahme der Relevanzbeurteilung. Die Positionen der einzelnen Einträge werden aus den Ergebnislisten der benutzten Suchmaschinen ermittelt, und anschließend stellt der Metasuchdienst aufgrund dieser Werte die Treffer fusioniert dar. Für Duplikate wird normalerweise ein durchschnittlicher Ranking-Wert aus den einzelnen Positionen errechnet. Allerdings funktioniert die Duplikat-Erkennung nur auf Basis der URL. Sind zwei Seiten inhaltlich gleich, aber unter unterschiedlichen URLs zu erreichen, so wird dies nicht erkannt und beide Einträge werden gelistet.

Dieses Verfahren ist offensichtlich nicht optimal, da die unterschiedlichen Suchdienste wie erwähnt sehr heterogene Verfahren zur Relevanzermittlung einsetzen. Zudem liefern die Suchmaschinen nicht die gleiche Anzahl an Einträgen zurück, so dass die Anteile einer Suchmaschine höher oder niedriger sind als die anderer. Ferner ist die Qualität der Suchergebnisse keineswegs vergleichbar. Die Ergebnisse dieses fusionierten Verfahrens können folglich nicht als vergleichbar angesehen werden. Dennoch setzen Suchmaschinen wie MetaCrawler [12] oder MetaEureka [13] diese Technik (noch) ein.

Das fortschrittlichere Verfahren übernimmt nur die Suchergebnisse, beachtet die Ranking-Position des zuliefernden Suchdienstes jedoch nicht. Unabhängig von der Quelle wird das Relevanzurteil mittels der Worthäufigkeit in Bezug auf die Stichwörter selbst berechnet. Die Basis dazu stellen die mitgelieferten Angaben zu jedem Eintrag, wie der Titel, URL und die Kurzbeschreibung, dar. MetaGer [14] ist ein Vertreter dieser Gattung (wobei anzumerken ist, dass diese Metasuchmaschine an der Universität Hannover ständig weiterentwickelt wird und einen enormen Funktionsumfang bietet).

Noch einen Schritt weiter geht das experimentelle System des NEC Research Institute mit dem Namen Inquirus [15]. Dieser Metasuchdienst verlässt sich nicht auf die Angaben der abgefragten Suchmaschinen, sondern lädt jedes Zieldokument herunter und berechnet auf der Basis dieser Originaldaten einen eigenen Ranking-Wert. Die Einträge aus den Suchmaschinen dienen quasi nur noch als verkleinerte, bereits vorsortierte Auswahl von Websites aus dem Netz. Dabei können mit diesem Verfahren auch tote Links und Duplikate erkannt werden. Ein großer Nachteil wird jedoch sofort ersichtlich, wenn man die Zeitdauer bedenkt, die solche Anfragen benötigen. Bereits die Parallelabfrage normaler Metasuchmaschinen nimmt schon wesentlich mehr Zeit in Anspruch als das Benutzen einer einzelnen Suchmaschine. Die Untersuchung der einzelnen

Trefferseiten würde ungemein länger benötigen. Vielleicht ist dieses Konzept daher noch nicht umgesetzt worden. Es existiert zwar neben wissenschaftlichen Veröffentlichungen [16] auch eine statische Ansicht des Prototyps [17], jedoch noch völlig ohne Funktion.

Bei der großen Anzahl von Metasuchmaschinen versuchen einige Anbieter, ganz eigene Wege zu gehen, um sich aus der Masse hervorzuheben. So steht bei Vivísimo [18] die Clustertechnik bei der Präsentation der Ergebnisse im Vordergrund. Dabei wird versucht, die gefundenen Treffer so in Gruppen anzuordnen, dass der Benutzer bei Auswahl eines thematischen Blocks nur noch für ihn themenrelevante Links erhält (siehe Abbildung 1.4).

The screenshot shows the Vivísimo search engine interface. At the top, there are navigation links: company | products | solutions | customers | demos | partners | press. A search bar contains the query 'fahrrad' and a dropdown menu is set to 'the Web'. A 'Search' button is visible, along with links for 'Advanced' and 'Help!'. Below the search bar, there are links for 'Refer us to a friend' and 'NEW Toolbar or MiniBar'.

The main content area is titled 'Clustered Results' and shows a list of search results for the query 'fahrrad'. The results are clustered into 17 documents. The first cluster is 'Cluster **Fahrradzubehör** contains 17 documents.' The results are listed as follows:

- 1. Fahrradzubehör - Biker's Barbecue** [new window] [frame] [preview]
Online-Shop mit nützlicher und unterhaltsamer Checkliste die verrät, welches **Fahrradzubehör** Sie
URL: www.bikersbarbecue.com/fahrradzubehoer.html - [show in clusters](#)
Sources: Lycos 8, MSN 9
- 2. Fahrrad OnlineShop | Fahrradzubehör | Fahrradteile | Beleuchtung** [...] [new window] [frame] [preview]
Wir sind seit 1985 Ihr Fahrrad Shop für **Fahrräder** und Fahrradteile/**Fahrradzubehör**. OnlineShop
URL: www.fahrrad-richter.de - [show in clusters](#)
Sources: Lycos 12, MSN 13, Open Directory 38
- 3. Fahrradartikel - Shops und Informationen** [new window] [frame] [preview]
Fahrradzubehör Sportkleidung
URL: www.fahrrad-online.net - [show in clusters](#)
Sources: Lycos 1, MSN 1
- 4. Fahrradzubehör Fahrraeder, Rennraeder, Mountainbikes Zu Guenstigen Preisen Fahrraeder**
Fahrgeschäfte verkauf **fahrrad** verkauf kettler **fahrräder** herrentreckingrad der bicycle aber fahrräder
crossrad bmx kataloge mountainbikepage fahrrad helmel; **fahrräder** rennräder das ... **Fahrradzubehör**
fahrrad ...
URL: www.manzanoarquitectos.com - [show in clusters](#)
Sources: MSN 20
- 5. Zweirad Müller** [new window] [frame] [preview]
Das **Fahrrad**-Programm des Fachhändlers und ausgewähltes **Fahrrad-Zubehör** wird mit Beschreibungen
Trekkingräder, Cityräder, Crossräder, Mountainbikes, Rennräder und Kinderräder verschiedener Marken
URL: www.zweiradmueller-ka.de - [show in clusters](#)
Sources: Open Directory 21

On the left side, there is a 'Clustered Results' sidebar with a tree view showing the following categories and counts:

- fahrrad (195)
 - Mountainbike (24)
 - Bike (27)
 - Fahrradzubehör (17)
 - Sport (13)
 - Zubehör (13)
 - ADFC, Allgemeiner Deutscher Fahrrad-Club (9)
 - Radsport (8)
 - Fahrradtour (10)
 - Fahrradtouren (3)
 - Europa (2)
 - Marions Chile-Fahrradtour (2)
 - Fahrrad routenplaner (2)
 - Other Topics (2)
 - Biker, Barbecue (6)
 - Zweirad, Müller (6)
 - More

At the bottom left, there is a 'Find in clusters:' section with an input field for 'Enter Keywords' and a search button.

Abbildung 1.4 Clustering bei der Metasuchmaschine Vivísimo

Das Besondere an der Clustering-Methode bei Vivísimo ist die »On-the-fly«-Generierung der Cluster. Aus den noch unsortierten Suchergebnissen werden automatisch thematische Gruppen generiert, und alle Treffer werden möglichst passend eingeordnet.

5 Suchprozess

»Tendenziell nutzten erfahrene Webuser jeweils nur eine Suchmaschine. Bei mangelhaften Ergebnissen wechselten sie nicht etwa das Recherche-Instrument, sondern die Begriffe. Dabei blieben nur 21 % der Bemühungen ohne Erfolg, während auf satte 50 % aller Anfragen zwischen 1 und 103 Antworten folgten. Immerhin 10 % lieferten bis zu 1.000 Suchergebnisse.«

– aus der Studie »Webzapping und seine Folgen« [54]

Im vorigen Abschnitt habe ich ausführlich beschrieben, mithilfe welcher Methoden und Modelle die Relevanz eines Dokuments anhand von Suchbegriffen bestimmt werden kann. Im Folgenden betrachten wir abschließend die bislang noch ausstehende Komponente der Suchmaschine, den **Query-Prozessor**. Während das gesamte übrige System kontinuierlich daran arbeitet, die Datenstruktur zu erweitern und zu aktualisieren, stellt der Query-Prozessor die Funktionalität zur Verfügung, die im Allgemeinen von einer Suchmaschine erwartet wird. Er verarbeitet die Eingaben des Nutzers und liefert Ergebnislisten, die nach Relevanz der Dokumente auf die Anfrage geordnet sind.

Die auch als **Searcher** bezeichnete Komponente stellt für den Benutzer das Interface bzw. Frontend zum Information-Retrieval-System dar. An dieser Stelle laufen alle Fäden zusammen; der Query-Prozessor vereint alle Funktionen des gesamten Systems. Ein Schlüsselkriterium ist hierbei die Bearbeitungsgeschwindigkeit. Man könnte annehmen, dass die Qualität der Suchergebnisse umso höher ist, je länger der Benutzer auf die Ergebnisliste warten muss. Dies trifft sicherlich bis zu einer gewissen Grenze zu. Die Konzeption eines Query-Prozessors bzw. des gesamten Systems erfordert einen Spagat bei der Architekturplanung. Schnelligkeit muss gegen Qualität abgewogen werden. Suchmaschinen im Web entscheiden sich zumeist zu Gunsten der schnellen Bearbeitung. Eine andere Entscheidung ist in Anbetracht der immensen Anzahl an zu bewertenden Dokumenten de facto nicht möglich.

Die Suchanfrage wird in den meisten Fällen von den Benutzern in ein einzeiliges Textfeld eingeben und anschließend als Zeichenkette an den Query-Prozessor übermittelt. Dabei lassen sich die Arbeitsschritte in drei Bereiche gliedern. Nach der Erfassung und Verarbeitung der Suchanfrage findet eine Relevanzbewertung der Dokumente anhand der bereits vorgestellten Gewichtungsmodelle statt. Anschließend wird dem Benutzer als Antwort auf seine Anfrage eine Trefferliste präsentiert.

5.1 Arbeitsschritte des Query-Prozessors

Die Bearbeitung der Suchanfrage (Query Processing) ähnelt in vielerlei Hinsicht der Normalisierung des Datenbestandes. Dies scheint auch logisch, da in beiden Fällen von Menschenhand geschriebene Texte in ein einheitliches, verarbeitbares und vergleichbares Format umgewandelt werden müssen. Nur so können gesuchte Dokumente und Suchanfragen miteinander verglichen werden. Diesen Vorgang bezeichnet man auch als **Matching**. Dabei werden die Stichwörter aus der Suchanfrage mit den Einträgen aus dem invertierten Index verglichen.

Allerdings sind von der Eingabe in die Suchmaske bis zum Matching bestimmte Schritte zwingend erforderlich. Wie sehen diese im Einzelnen aus?

5.1.1 Tokenizing

Nachdem der Benutzer die Suchanfrage eingeben hat und der Browser den Inhalt des Formularfeldes mittels HTTP an den Query-Prozessor gesendet hat, müssen einzelne Elemente, die als Tokens bezeichnet werden, aus dem Zeichenstrom identifiziert werden. Das betrifft natürlich einerseits die reinen stichwortbasierten Suchmaschinen wie Google, Yahoo und so weiter, andererseits aber auch die natürlichsprachigen Systeme (NLP, natural language processing systems) wie beispielsweise AskJeeves. Letztere sind darum bemüht, komplexe Anfragen wie beispielsweise »Wie wird das Wetter morgen?« sinnvoll zu beantworten, um eine benutzerfreundlichere Suche zu ermöglichen.

5.1.2 Parsing

Da die Suchanfragen der Benutzer oftmals nicht nur reine Stichwörter, sondern auch spezielle Operatoren enthalten, muss jedes einzelne Token aus dem vorigen Schritt auf seine Funktion geprüft werden. Die Operatoren werden anhand einer Liste von reservierten Zeichen und Begriffen erkannt. So können Anführungszeichen, boolesche Operatoren wie AND und OR und sonstige spezielle Funktionen wie etwa die Einschränkung bei Google, nur nach PDF-Dateien zu suchen (`filetyp:pdf`), erkannt werden.

Im Falle der natürlichsprachigen Systeme werden solche Operatoren implizit erkannt. Dabei wird die Suchanfrage einer Sprachanalyse unterzogen, die Kriterien anhand von Präpositionen, Konjunktionen und der Wortreihenfolge bewertet und daraus logische Zusammenhänge zwischen den Begriffen generiert.

5.1.3 Stoppwörter und Stemming

Sofern bereits während der Dokumentnormalisierung eine Stoppwortliste angewandt wurde, müssen freilich auch die Suchbegriffe auf Stoppwörter untersucht werden. Natürlich könnte man die Stoppwörter auch in der Suchanfrage belassen. Ein Ergebnis würde ohnehin nicht erzielt werden, da keine Entsprechung zu den Stoppwörtern im Index gefunden würde. Allerdings kostet das Erkennen und Entfernen der Stoppwörter im Vorhinein weniger Rechenzeit als die erfolglose Suche im Index. Des Weiteren würden vorhandene Stoppwörter in der Stichwortliste die Ergebnisse verschiedener Algorithmen verzerren. So steht beispielsweise das Wort »Neuseeland« in der reinen Suchanfrage »Urlaub machen in Neuseeland« an Position vier. Die Stoppworteliminierung würde ein anderes Ergebnis liefern (Urlaub Neuseeland), das mit hoher Wahrscheinlichkeit eine bessere Voraussetzung für das Matching darstellt.

Google wendet zum Beispiel, wie ich bereits in Abschnitt 3.2.7, *Stoppwörter*, dargestellt habe, eine Stoppworteliminierung erst ab einer bestimmten Anzahl von Suchbegriffen an. Adäquat dazu findet ein Stemming der Suchbegriffe auch nur dann statt, wenn dies bereits während der Dokumentverarbeitung im Information Retrieval System durchgeführt wurde. In diesem Fall ist ein Stemming sogar unumgänglich, da die Begriffe beider Textmengen – das Dokument und die Suchanfrage – auf einen gemeinsamen Stamm reduziert werden müssen, um überhaupt ein Matching erfolgreich durchführen zu können.

Führt ein Information-Retrieval-System weder eine Stoppwortprüfung noch Stemming durch, werden diese beiden Schritte natürlich übersprungen.

5.1.4 Erzeugung der Query

Die bisherigen Schritte dienen der Normalisierung der Suchanfrage. Man kann diesen Prozess auch als eine Form der Übersetzung sehen. Das wird besonders bei den natürlichsprachigen Anfragen deutlich. Hier wird eine umgangssprachliche Frage »Wie lang ist der Nil?« in ein Format übersetzt, anhand dessen der Query-Prozessor ein Matching durchführen kann. Durch Stoppwortbetrachtung und Stemming erhält man das Paar »lang nil« (wobei lang beispielsweise auch die gestemimte Form des Wortes Länge darstellen würde).

Um das Matching durchzuführen, fehlt lediglich die Relation zwischen den erhaltenen Suchbegriffen. Dazu werden die extrahierten Operatoren aus dem zweiten Schritt genutzt. Dadurch entsteht ein systemspezifisches Format, das die Repräsentation der ursprünglichen Suchanfrage darstellt.

An diesem Punkt übernehmen die meisten Suchmaschinen die Repräsentation der Suchanfrage und führen das Matching mit dem invertierten Index durch.

5.1.5 Verwendung eines Thesaurus

Bei weitergehenden Entwicklungen zeichnet sich die Verwendung eines Thesaurus ab. Eine speziell erweiterte Datenstruktur enthält ein Wortnetz, dessen Einträge miteinander verbunden sind und so in sinnvoller Relation zueinander stehen. So lassen sich beispielsweise Synonyme, Abkürzungen oder Ober- und Unterbegriffe zu einzelnen Termen bestimmen. Der Thesaurus ist ein effektives Hilfsmittel zur Sacherschließung.

Diese Erkenntnis kann man sehr gut in den Query-Verarbeitungsprozess mit einbeziehen. Oftmals nutzen Surfer eine Suchmaschine, um zu einem gewissen Themengebiet mehr zu erfahren. Infolgedessen ist das Wissen über spezielle Begriffe meist eher dünn gesät und die Suche nicht selten wenig erfolgreich oder sehr mühsam. Eine Suchmaschine, die auf Wunsch gleichzeitig alle möglichen Synonyme und verwandte Ober- und Unterbegriffe in die Suche mit einbezieht, kann hier wahre Wunder bewirken.

5.1.6 Matching und Gewichtung

Der Normalisierungsprozess ist spätestens an dieser Stelle abgeschlossen. Das Matching kann nun durchgeführt werden. Das grundsätzliche Vorgehen wurde bereits eingangs kurz vorgeangrissen.

Im Vorlauf des Matchings werden zunächst die Begriffe der Anfrage-Repräsentation in die entsprechenden WordIDs übersetzt. Anschließend werden die grundsätzlich in Frage kommenden Dokumente bestimmt. Dazu wird die Word ID anhand des invertierten Indexes durchsucht. Das Ergebnis dieser Suche ist eine Auswahl an Dokumenten, die den gesuchten Begriff enthalten.

Handelt es sich um eine Suchanfrage mit mehreren Begriffen, muss die Bedeutung der Relation zwischen den jeweiligen Begriffen berücksichtigt werden. Bei einem AND-Operator zwischen zwei Begriffen würden beispielsweise nur solche Dokumente bei der Suche im invertierten Index herausgefiltert werden, die auch tatsächlich beide Begriffe enthalten.

Anhand der Hitlist der gefundenen Einträge, die wichtige Werte über die Wortposition, Formatierung, Häufigkeit usw. enthält, werden die weiteren Berechnungen durchgeführt. Dazu werden an dieser Stelle die im vorigen Abschnitt vorgestellten statistischen Gewichtungsmodele eingesetzt. Oftmals werden zusätzlich auch die Linkstrukturen beispielsweise mittels Page-Rank ermittelt. Durch dieses Verfahren erhält jedes Dokument eine Gewichtung, die die Rele-

vanz auf die Suchanfrage ausdrückt. Diese Gewichtung bezieht dann je nach Auswahl der verwendeten Algorithmen einerseits die Art und Weise des Auftretens der Begriffe mit ein. Andererseits werden makrostrukturelle Verlinkungen betrachtet, und nicht zuletzt wird das einzelne Dokument in Relation zu den anderen in Frage kommenden Dokumenten gesetzt.

5.1.7 Darstellung der Trefferliste

Diese Relation zu anderen Dokumenten zeigt sich in der Listenposition innerhalb der Trefferliste. Je weiter oben ein Dokument anzutreffen ist, umso höher ist die vom System angenommene Ähnlichkeit zu der Suchanfrage. Die Seite auf Platz eins ist somit die angeblich ähnlichste und damit die optimal passende Ressource zur Anfrage.

Die Darstellung der Trefferliste stellt den letzten Schritt des Query-Prozessors dar. Der Benutzer erhält diese als Antwort auf seine Anfrage und muss nun anhand der präsentierten Informationen zu jedem Treffer entscheiden, auf welcher Seite er seinen Wissensdurst befriedigen kann.

Die meisten Suchmaschinen stellen hier noch die Möglichkeit zur Verfügung, das Suchergebnis zu verfeinern. Das geschieht allerdings in unterschiedlichem Ausmaß. Bei den meisten Betreibern erfolgt dies leider nur sehr rudimentär. Die Suchanfrage wird einfach nochmals in dem Suchfeld über der Ergebnisliste angezeigt und kann somit verfeinert werden. Die meisten Suchmaschinen übergeben die vorige Suche auch auf Wunsch zur Verfeinerung an die erweiterte Suchfunktion. Nur bei HotBot sieht es diesbezüglich nicht ganz so rosig aus, der Benutzer muss die Suchanfrage erneut formulieren. Bei Lycos erhält der Webuser erst gar nicht die Möglichkeit, über einen Link zur erweiterten Suche zu gelangen.

Eine Besonderheit bietet AllTheWeb nach manchen Suchanfragen. Es werden bis zu zehn weitere Begriffe zur gegebenen Anfrage vorgestellt, die der Benutzer durch Klick auf ein Plus oder Minus zur Suche hinzufügen (AND) bzw. ausschließen (NOT) kann. Abbildung 5.1 zeigt beispielhaft eine Liste der vorgeschlagenen Begriffe zu der Suche nach dem Begriff *wok*.

Eine breite Anwendung findet diese Verfeinerungsmöglichkeit jedoch wahrscheinlich nicht, da sich die vorgeschlagenen Begriffe oftmals nicht sonderlich gut eignen, um die Treffermenge sinnvoll zu erweitern bzw. zu reduzieren.



Abbildung 5.1 Verfeinerte Suche bei AllTheWeb

5.2 Suchoperatoren

Die wachsende Anzahl von Webseiten im Index der Suchmaschinen zwingt den Benutzer, immer genauere Suchanfragen zu stellen. Die Eingabe eines einzelnen Begriffs liefert heutzutage oftmals einen undurchdringbaren Wald an Treffern, bei dem selbst die Ergebnisse an erster Stelle selten dem Wunsch des Benutzers entsprechen. In diesem Zusammenhang spricht man auch von der **Practical Precision**. Diese bezeichnet die Leistung der Suchmaschine, speziell auf der ersten und zweiten Ergebnisseite eine hohe Precision-Rate zu erzielen. Der Hintergrund dieser Überlegung ist, dass sich die wenigsten Benutzer weiter hinten liegende Seiten der Trefferliste anschauen und dass sich somit die tatsächliche Präzision nicht auf die gesamte Trefferliste, sondern nur auf die meistbeachteten Treffer beziehen sollte.

Um das Gesuchte näher einzugrenzen und die Precision von der Benutzerseite aus erhöhen zu können, bedarf es einer mächtigen Anfragesprache. Diese ermöglicht es, einzelne Stichwörter in logische Verbindung zueinander zu stellen und Begriffe mit Attributen zu versehen. Dabei beherrscht heutzutage jede Suchmaschine im Web diese Grundfunktionalität, die man durchaus als gemeinsamen Standard definieren könnte.

5.2.1 Boolesche Ausdrücke

Die meisten Suchenden sind sich gar nicht bewusst, dass sie bei einer Suchanfrage mit mehr als einem Begriff bereits von der booleschen Logik Gebrauch machen. Das ist für die Suchmaschinen-Optimierung jedoch eine enorm wichtige Erkenntnis. Denn werden in dem Suchfeld beispielsweise zwei Suchbegriffe direkt hintereinander ohne Zuhilfenahme von Operatoren eingegeben, setzen alle großen Suchmaschinen automatisch ein AND dazwischen ein. Der Benutzer bemerkt dies natürlich nicht, weil er intuitiv eine AND-Verknüpfung beabsichtigt, wenn er zwei Begriffe eingibt. Erfahrene Webuser benutzen die ausgefeilte Technik der verschiedenen Operatoren, um die Anfrage im Voraus

zu präzisieren. Und auch bei weniger erfahrenen Benutzern ist zunehmend zu beobachten, dass die booleschen Ausdrücke immer häufiger Verwendung finden, nachdem ein erster »Blindschuss« mit einem Begriff nicht das gewollte Ergebnis erzielt hat. Der Boom an Praxisbüchern, die das Geheimnis der effektiven Suche im Web zu lüften versprechen, sind sicherlich als ein Zeichen für diesen Trend zu sehen.

An dieser Stelle soll nur ein kurzer Überblick über die Operatoren und deren Abkürzungen gegeben werden, um Ihr Wissen abzurunden und an späterer Stelle darauf zurückgreifen zu können.

► **AND (+)**

Jeder Begriff muss mindestens einmal im Suchergebnis enthalten sein. Als Abkürzung kann auch das Pluszeichen (+) direkt vor das zu verknüpfende Wort gestellt werden.

```
hausboot AND neuseeland
hausboot +neuseeland
```

► **OR (|)**

Hier muss nur einer der beiden Begriffe in einem Dokument vorhanden sein, damit das Dokument mit in die Treffermenge aufgenommen wird. Wie alle Operatoren kann auch dieser mittels Klammern kombiniert werden. Dabei gilt die aus der Schulmathematik bekannte Regel, dass ein Term von innen nach außen anhand der Klammern verarbeitet wird. So würde die Anfrage eines Benutzers, der an dem Erwerb eines neuen Autos interessiert ist und sich über Kauf- und Leasingangebote informieren möchte, wie folgt aussehen:

```
auto AND (kauf OR leasing)
```

Dabei würden sowohl alle Dokumente mit den beiden Begriffen `auto kauf` wie auch `auto leasing` in Frage kommen. Dokumente mit dem Thema `auto mieten` würden nicht in Frage kommen.

► **NOT (-)**

Um gewisse Themen bei der Suche auszugrenzen, kann der negative Operator verwendet werden. Möchten Sie beispielsweise alle Computermessen angezeigt haben, wollen aber die im Web stark präsente CeBIT ausschließen, würden Sie einen der folgenden Ausdrücke verwenden:

```
Computermesse NOT cebit
Computermesse -cebit
```

5.2.2 Phrasen

Durch Verwendung von Anführungszeichen können mehrere Wörter zu Ausdrücken vereint werden. Manche Begriffe lassen sich nicht in einem Wort fassen, sondern es bedarf der genauen Anordnung mehrerer Wörter. Dies trifft insbesondere auf die Kombination von Vor- und Nachnamen zu. Aber auch bei den Ausdrücken »Bundesrepublik Deutschland« oder »Universität Freiburg« ist eine Phrasensuche sinnvoll. Oftmals können auch Zitate auf diese Art leichter gefunden werden.

Die Phrasen werden dann genau in der vorgegebenen Anordnung und Schreibweise gesucht. Mit der Eingabe `Universität Freiburg` ohne Anführungszeichen erhält man eine Ergebnismenge mit Seiten, die zwar beide Wörter enthalten, aber nicht zwingend in der Reihenfolge direkt hintereinander. Die Phrasensuche ist neben den booleschen Operatoren eine weit verbreitete Methode.

5.2.3 Wortabstand

Kann oder will man den Abstand zwischen zwei Begriffen nicht genau definieren, stehen verschiedene Ausdrücke zur Verfügung, die angeben, wie nahe beieinander die Wörter in etwa stehen dürfen.

► ADJ

Dieser Operator bedingt, dass beide Begriffe direkt nebeneinander stehen müssen. Dieser Operator ist der Phrasensuche sehr ähnlich, nur dass es hierbei nicht auf die Reihenfolge der Begriffe ankommt.

► NEAR (~)

Um die gewünschte Nähe zweier Begriffe auszudrücken, kann dieser Operator genutzt werden. Zwischen den beiden Begriffen dürfen nicht mehr als zehn andere Begriffe stehen, ansonsten wird eine Seite nicht in die Ergebnismenge mit aufgenommen.

`Schneewittchen NEAR Zwerge`

Im Beispiel würde ein Dokument mit dem Satz »Schneewittchen und die sieben Zwerge« gefunden werden, andere Seiten, die diese Begriffe mit größerem Wortabstand enthalten, hingehen nicht.

► FAR

Dieser Operator ist das Gegenstück zum vorigen. Beide Begriffe müssen in einem Dokument vorkommen und dürfen nicht nahe beieinander stehen.

Bislang unterstützen nur wenige Suchmaschinen diese Funktionen. Altavista und Fireball sind neben Lycos genau genommen derzeit die einzigen Betreiber.

Aber auch diese unterstützen die Funktion nur in den Detailsuchen bzw. erweiterten Suchen. Die geringe Verbreitung liegt sicherlich an der höheren Anforderung an den Benutzer. Dieser muss eine abstrakte Vorstellung haben, wie die von ihm gesuchten Begriffe zueinander stehen. Dies ist jedoch in der Regel nicht der Fall.

5.2.4 Trunkierung

Ein Stern (*) wird als Platzhalter, ein so genanntes Wildcard, einem Begriff voran- oder nachgestellt. Der Query-Prozessor interpretiert diesen und sucht nicht nur ausschließlich nach dem angegebenen Begriff, sondern auch nach den entsprechend erweiterten Begriffen. So werden bei der Suche nach `haus*` nicht nur Dokumente mit dem Begriff `haus` angezeigt, sondern auch beispielsweise Seiten mit den Wörtern `hausmann`, `haushalt`, `hausboot` und so weiter. Entsprechendes gilt für ein vorangestelltes Wildcard.

Google, Lycos, Altavista und Yahoo wenden die Trunkierung bereits automatisch an. Will man dies verhindern, ist nicht selten ein glückliches Händchen bei der Suche nach der Funktionsbeschreibung der Suchmaschine nötig. Diese Beschreibungen sind oftmals nicht sofort auffindbar oder recht undurchsichtig. Generell sollte die Phrasensuche mit einem Begriff jedoch Abhilfe schaffen, damit keine Trunkierung stattfindet. Bei Google erreicht man dies durch Eingabe eines Begriffs in dem Format `[+begriff]`.

Bei Fireball ist bei einem Vergleich zwischen der Verwendung eines trunkierten und eines nicht-trunkierten Begriffs eine unterschiedliche Trefferliste festzustellen. Jedoch weicht diese nicht in vielen Treffern ab, da der ursprüngliche Begriff ohne Trunkierung bereits hohe Ranking-Werte erreicht. Dieses Phänomen lässt sich auch bei den oben genannten Anbietern beobachten. Die Anwendung der Wildcards führt daher selten zum gewünschten Ziel, die Ergebnismenge zu einem Begriff weiter zu fassen, da die Ranking-Kriterien relativ konstant bleiben.

5.3 Erweiterte Suchmöglichkeiten

Um die spezifischen Eigenschaften der Anfragesprache einer Suchmaschine, die nicht zu einem Quasi-Standard zusammengefasst werden können, komfortabel nutzen zu können, stellen alle Suchmaschinenbetreiber eine erweiterte Suche zur Verfügung.

Google in Erscheinung treten. Für die Bildersuche ist bei Google ein spezieller Webcrawler zuständig, der den Namen Google-Image trägt. Um diesen an der Indexierung der unter dem Verzeichnis `galerie` befindlichen Bildergalerie zu hindern, muss eine spezielle Ausnahmeregel eingebunden werden:

```
User-agent: *  
Disallow: /css/  
User-agent: Google-Image  
Disallow: /galerie/
```

Listing 7.3 Bilder bei Google nicht indexieren lassen

Allen Webcrawlern wird hier die Erfassung des Verzeichnisses `css` untersagt. Dem Bild-Crawler von Google wird außerdem das Verzeichnis `galerie` nicht zur Indexierung freigegeben. Um eine komplette Website von der Erfassung auszuschließen, würde folgende Zeile nach der entsprechenden `User-agent`-Definition eingebunden werden:

```
Disallow: /
```

Beträfe diese Zeile den Webcrawler Google-Image, so würden keine Grafiken oder Bilder der gesamten Website erfasst werden. Beachten Sie jedoch, dass beim Setzen eines Sterns in Kombination mit diesem generellen Erfassungsverbot kein einziger Webcrawler Ihre Seite mehr erfassen wird. Weiterführende Informationen zu dem Robots Exclusion Protocol finden Sie neben der oben genannten Adresse auch wiederum schön aufbereitet bei Selfhtml [97].

In einigen Fällen möchte man eine Ressource aus dem Index entfernen, nachdem sie bereits erfasst wurde. Dazu ist der entsprechende Eintrag in die `robots.txt` natürlich Pflicht. Bei dem nächsten Besuch des Webcrawlers wird die betroffene Seite aus dem Index entfernt. Dies kann allerdings mehrere Wochen dauern. In einigen Fällen kann nicht so lange gewartet werden. Daher machen die Suchmaschinen-Betreiber das Angebot, die Entfernung der Seite aus dem Index manuell vorzunehmen. Meist muss dies per E-Mail beantragt werden. Google bietet diesbezüglich allerdings ein automatisiertes Verfahren an. Nach Anmeldung über die E-Mail-Adresse kann man bestimmte Daten aus dem Index entfernen lassen [98].

7.4 Link-Popularity erhöhen

Ein wesentlicher Bereich, der nur schwer durch eigenes Zutun optimiert werden kann, ist die Link-Popularity. Aus diesem Grund setzt Google besonders auf sein Page-Rank-Verfahren. Doch auch hier kann ein Webautor gewisse Off-page-Optimierungen ansetzen, um eine höhere Link-Popularity zu erreichen.

Das Ziel der Bemühungen ist es natürlich, möglichst gewinnbringend eingehende Links auf die eigenen Seiten zu erzielen, um die eigene Link-Popularity zu erhöhen und sich somit weiter oben in den Ergebnislisten positionieren zu können. Hierbei handelt es sich um einen mehrstufigen Prozess, der grundlegende Vorbedingungen fordert. Beides soll im Folgenden beschrieben werden.

7.4.1 Interne Verlinkung optimieren

Der erste Schritt zur Erhöhung der Link-Popularity fällt streng genommen nicht in den Bereich der Offpage-Optimierung, muss aber dennoch an dieser Stelle genannt werden. Denn die Optimierung der internen Verlinkung geht selbstverständlich jeglicher anderen Optimierung voraus.

Bei der Behandlung des Page-Rank-Algorithmus wurden bereits einige typische Phänomene erläutert. Diese sollen hier nicht wiederholt werden. Vielmehr werde ich auf zwei generelle Verhaltensweisen eingehen, die aufgrund der genannten Phänomene zu bestimmten optimierenden Handlungen führen. Zum einen ist es bei Seiten mit einer Vielzahl von externen Verweisen empfehlenswert, dass ebenfalls einige Links innerhalb der eigenen Site bleiben. So verteilt man den Verlust der Link-Popularity möglichst nicht nur auf die ausgehenden Links allein, sondern behält auch etwas auf den eigenen Seiten zurück.

Des Weiteren sollte man bei Gelegenheit darauf achten, dass Seiten mit vielen externen Verweisen eine möglichst geringe Link-Popularity erzielen. So ist der Verlust des Link-Popularity-Wertes auf die Website insgesamt gesehen nicht so hoch. Die Voraussetzung für einen geringen Wert ist, dass eingehende Links nicht unbedingt auf diese Seite verweisen, sondern auf andere Seiten der Webpräsenz mit weniger ausgehenden Links.

Diese Überlegungen basieren auf dem mathematischen Funktionsmodell des Page-Rank-Algorithmus. Natürlich sind die genannten Richtlinien idealtypisch, denn oftmals sind sie in der Praxis kaum in ausreichender Konsequenz durchzuführen. Ein Grund dafür ist, dass relevanter Inhalt immer von anderen als wertvoll erachtet wird und daher verlinkt wird.

7.4.2 Das KAKADU-Prinzip

Bei der Optimierung der Link-Popularity muss der Fokus weiterhin zunächst nach innen gerichtet bleiben. Bemüht sich ein Webautor um eingehende Links, ist die Relevanz des Angebots natürlich entscheidend. Unbedeutende Inhalte regen andere Webautoren in den seltensten Fällen dazu an, einen Link auf das Angebot zu setzen. Die Link-Popularity bleibt folglich niedrig. Es lassen sich

gewisse Faktoren formulieren, die erfahrungsgemäß erfüllt sein müssen, um eine gute Link-Popularity zu erzielen.

Beim KAKADU-Prinzip steht ein bestimmter Inhaltstyp für jeden Buchstaben des Wortes. Vereint man alle miteinander auf einer einzigen Website, ist ein gehöriges Interesse von außen und somit die Grundlage für eine gute Link-Popularity gesichert.

► **Kostenlose Informationen**

Ob Tipps oder Tricks zu bestimmten Themen gegeben werden, Neuigkeiten aus einer Branche, lokale Nachrichten oder praxisbezogene Ratschläge – solange hochqualitative Informationen kostenlos sind, werden sie gerne angenommen. Ein besonderes Beispiel sind hier die so genannten Tutorials, die mit einer praktischen Schritt-für-Schritt-Anleitung zur Lösung von Problemen vornehmlich aus dem EDV-Bereich beitragen.

► **Aktuelles**

Egal, welche Informationen oder Inhalte angeboten werden, die Aktualität spielt eine entscheidende Rolle. Optimal ist selbstverständlich Brandaktuelles. Dies beinhaltet demnach oft den Faktor der Exklusivität, denn neue Themen und Inhalte sind selten weit verbreitet.

► **Künstlerisches**

Der Mensch lebt nicht nur vom Brot allein – ebenso wenig von seiner Arbeit. Der Bedarf an Videos, Musik und Grafiken aus dem Web ist in den letzten Jahren enorm angestiegen. Ein Beispiel sind die begehrten Portale, die zur Verschönerung des Arbeitsplatzes eine unerdenkliche Vielfalt an Desktop-Hintergründen und Bildschirmschonern anbieten. Ferner machen sich auch viele Webautoren selbst auf die Suche nach Bildmaterial und Grafiken für die eigene Webpräsenz. Der Markt an künstlerischen Werken im Netz ist breit gefächert.

► **Außergewöhnliches**

Je seltener ein bestimmtes Angebot zu finden ist, desto stärker konzentriert sich der Besucherstrom auf die vorhandenen Websites. Ob es sich dabei um besondere Informationen handelt, um eine außergewöhnliche Dienstleistung oder eine hervorragende Idee für das Thema einer Website ist dabei unerheblich. Man kann Nutzer mit einem sensationellen Preisangebot ebenso begeistern wie mit einem außergewöhnlichen Online-Spiel. Der Phantasie des Content-Anbieters sind keine Grenzen gesetzt.

► **Downloads**

Mit immer schnelleren Bandbreiten erhöht sich die Zahl und Größe der Dateien, die aus dem Netz heruntergeladen werden. Der Boom der Tausch-

börsen und Download-Portale, auf denen man Freeware, Shareware oder sonstige Inhalte erhält, zeigt die Stärke dieses Faktors.

► **Unerlaubtes**

Nicht zuletzt ist das Verbotene auch im Netz reizvoll. Dabei ist in erster Linie nicht das Anbieten von illegalen Inhalten gemeint, auch wenn diese unbestreitbar einen großen Anziehungseffekt aufweisen. Vielmehr ist der Bruch gesellschaftlich anerkannter Normen gemeint. Die Spanne ist auch hier sehr groß und führt von Bildern, die die Privatsphäre von Prominenten aufdecken, bis hin zu der Anleitung zum Bau einer Kartoffelkanone. Oftmals gerät ein solches Angebot in eine rechtliche Grauzone. Man denke nur an die zahlreichen Seiten, auf denen Seriennummern und Programme zum Freischalten von Shareware (Cracks) zu finden sind.

7.4.3 Qualitätskriterien potenzieller Linkpartner

Der angebotene Inhalt auf der eigenen Website stellt natürlich ein wichtiges, allerdings nicht alleiniges Qualitätskriterium dar, wenn es um die Optimierung der Link-Popularity geht. Ein Charakteristikum der Link-Popularität besteht darin, dass Hunderte von eingehenden Links von Webseiten mit eher geringem Wert nicht unbedingt einen so großen Effekt haben wie wenige Links von Webseiten mit einer eigenen hohen Link-Popularity. Die Qualität der zukünftigen Linkpartner muss daher sichergestellt werden, um den Aufwand zu rechtfertigen.

Behalten Sie die nachfolgenden Punkte ständig im Hinterkopf, so haben Sie sich eine sichere Basis für weitere Schritte geschaffen. Denn diese Grundlagen stellen die notwendige Voraussetzung dafür dar, dass eingehende Verweise auch wirklich den gewünschten Effekt, nämlich die Erhöhung der eigenen Link-Popularity, erzielen.

► **Link-Popularität prüfen**

Wenn Sie möchten, dass jemand auf Ihre Site verlinkt, sollten Sie zunächst dessen Link-Popularity überprüfen. Denn nach dem Prinzip der Vererbung kann ein Partner Sie bei Ihrem Vorhaben nur dann voranbringen, wenn dieser selbst über ausreichend hohe Werte verfügt.

Ein sehr umfangreiches Online-Tool zur Bestimmung der eigenen Link-Popularity im Vergleich zu anderen Sites bietet Marketleap auf seiner Website an [99]. Natürlich sind die Toolbars der einzelnen Suchdienste ebenfalls ein guter Anhaltspunkt, wenn auch die zugrunde liegenden Daten teilweise veraltet sind. Überprüfen Sie im Vorhinein die potenziellen Linkpartner auf deren Wert. Dabei sollte besonderes Augenmerk auf Googles Page-Rank gelegt werden, da dieser Anbieter derzeit mit Abstand marktführend ist.

Liegt der Page-Rank einer anvisierten Seite nicht mindestens um ein oder zwei Punkte höher als ihr eigener, lohnt sich eine Verlinkung nicht unbedingt im Vergleich zu dem Aufwand. Sites mit einem PR-Wert von fünf oder sechs sind im Allgemeinen gut geeignet.

► **Themenkreise wahren**

Achten Sie bei der Suche nach Möglichkeiten zur Platzierung eingehender Links darauf, dass die Partnerseiten ein möglichst ähnliches Themengebiet abdecken. Dass die thematische Verwandtschaft bei der Link-Popularity Berücksichtigung findet, wurde bereits mehrfach erwähnt. Im Sinne eines Community-Gedankens sollten Sie daher vorwiegend auf solche Seiten setzen, die sich innerhalb dieser Gemeinschaft befinden. So wird eine Seite, die sich mit Backrezepten befasst, ein höheres Gewicht erhalten, wenn sie von einer anderen Koch- oder Backseite verlinkt wird. Ein Verweis von einer gleichwertigen Webseite eines Autohauses bringt demnach weniger Punkte ein.

► **Exklusivität**

Je weniger Verweise eine Seite nach außen besitzt, desto stärker wirkt jeder einzelne nach außen hin. Die optimale Partnerseite besitzt daher wenige Links zu anderen Anbietern, sondern am besten ausschließlich zu Ihrem Webangebot.

► **Suchbegriffe im Linktext**

Die Bedeutung des Linktextes spielt auch in diesem Kontext eine wichtige Rolle. Im optimalen Fall enthält der eingehende Linktext von einer anderen Website nämlich die passenden Schlüsselwörter Ihrer Seite. Natürlich ist die Beeinflussung anderer Content-Anbieter nicht immer so einfach. Im nächsten Abschnitt gebe ich daher einige Tipps, wie man die Art der Verlinkung von außen zumindest ein wenig steuern kann.

7.4.4 An andere Webautoren herantreten

Sie haben soeben erfahren, dass bedeutender Inhalt die Grundlage für eine Verlinkung überhaupt ist. Außerdem kennen Sie die vier Gütekriterien für optimale Linkpartner. Wenn diese Punkte beherzigt und umgesetzt sind, kann man zum nächsten Schritt zur Optimierung der Link-Popularity kommen. Hier stellt sich die Frage, wie man an andere Webautoren mit der Bitte um einen Verweis herantritt, um eine möglichst positive Reaktion zu erhalten.

Eine Verlinkung auf die eigenen Seiten geschieht meist unkontrolliert durch andere. Im Sinne der oben genannten Punkte wäre es hingegen wünschenswert, ein wenig Einfluss auf die Platzierung und Art des eingehenden Verweises zu besitzen. Selbst wenn sich andere Autoren natürlich nicht gern beein-

flussen lassen, kann man unterschwellig bestimmte Informationen übermitteln. Damit erhöht man die Wahrscheinlichkeit, dass eingehende Verweise optimal gestaltet sind. Zunächst setzt man wieder innerhalb der eigenen Webpräsenz an. Eine Seite nach dem Motto »Verweisen Sie auf uns« bietet erste Möglichkeiten zur Kontrolle eingehender Links. Bieten Sie auf dieser Seite ausgewählte URLs von Seiten der eigenen Webpräsenz an, die Sie gern verlinkt haben möchten. Bieten Sie dabei explizit den HTML-Code an, so dass andere Autoren diesen nur noch kopieren müssen. Oftmals werden auch Logos oder sogar Banner angeboten, die auf fremden Seiten platziert werden können. Bieten Sie in diesem Falle verschiedene Größen der Grafiken an. Außerdem sollte auch hier der HTML-Code zum schnellen Einbinden zur Verfügung gestellt werden.

Das Anbieten eines fertigen HTML-Codes hat den Vorteil, dass viele Autoren diesen ohne Veränderung auf ihren Seiten übernehmen. Vergessen Sie in diesem Zusammenhang nicht die Kriterien der Onpage-Optimierung, insbesondere des Linktextes und der Bilder. Oftmals wird man als Webautor selbst andere Anbieter per E-Mail kontaktieren. Dabei sollte man nicht automatisch auf die soeben angesprochene Seite verweisen. Kommen Sie dem anderen aktiv entgegen, und bitten Sie um die Platzierung eines Verweises. Dabei liefert man am besten den entsprechenden HTML-Code in der E-Mail gleich mit, um dem Gegenüber die Sucharbeit abzunehmen.

Gelegentlich erfährt man von anderen, dass ein Link auf die eigene Seite positioniert wurde. Meist ist dies mit der Bitte um eine Rückverlinkung verbunden. Überprüfen Sie unabhängig davon die Gestaltung des eingehenden Links. Verweist er auf die gewünschte Seite? Enthält er die passenden Schlüsselbegriffe? Falls nicht, melden Sie sich möglichst zeitnah bei dem Autor. Die Wahrscheinlichkeit, dass er den Verweis ändert, ist erfahrungsgemäß höher, wenn nicht bereits Wochen oder Monate seit der Platzierung vergangen sind.

Natürlich kann man auch selbst nach Verweisen suchen. Nutzen Sie dazu die entsprechenden Funktionen der Suchmaschinen. Scheuen Sie sich auch hier nicht davor, einem anderen Content-Anbieter, der einen Verweis auf Ihre Seite gesetzt hat, eine Verbesserung dieses Verweises vorzuschlagen.

7.4.5 Eingehende Links erzielen

In der Regel muss ein Webautor selbst aktiv werden, um eine nennenswerte Verlinkung auf seine Seiten zu erzielen. Der Faktor Zeit tut natürlich auch sein Übriges dazu. Wenn der Inhalt für viele andere Anbieter relevant erscheint, werden mit der Zeit immer mehr Verweise auf die Site zeigen.

Doch natürlich will man versuchen, möglichst schnell und gezielt Links zu platzieren, um die Link-Popularity zu erhöhen. Die erste Adresse sind hier die renommierten Webkataloge. Ein Eintrag im Open Directory Project oder Yahoo wird beispielsweise von Google sehr hoch bewertet. Die Aufnahme in derartige Webkataloge wurde bereits zu Beginn besprochen. Nicht weniger Beachtung sollte man anderen Katalogen schenken. So gibt es häufig auf Gemeindeportalen eine Liste ausgewählter Links, oder auch spezielle Themen-Webkataloge stellen ein weites Feld dar.

Insbesondere bei nicht-kommerziellen Angeboten bieten sich natürlich Freunde und Bekannte als eine der ersten Anlaufstellen an. Oftmals besitzen diese eine eigene Website. In letzter Zeit ist es auch weltweit zur Mode geworden, ein so genanntes **Weblog** zu führen. Dabei handelt es sich um eine Seite, auf der ein Autor periodisch Kommentare, Berichte oder sonstige Beiträge zu einem bestimmten Thema veröffentlicht. Neue Einträge stehen dabei immer an oberster Stelle. Die behandelten Themen sind dabei breit gefächert und reichen von persönlichen Tagebüchern bis hin zur kritischen Betrachtung einzelner Journalisten. Diese Spezialform von Weblogs nennt man **Watchblogs**. Auch die Aktivitäten der Suchmaschinen werden in solchen Blogs beobachtet. Dazu braucht man lediglich die Stichwörter `blog suchmaschinen` in eine Suchmaschine einzugeben und bekommt unzählige Treffer. Die Blogger-Community ist enorm und wächst zusehends. Die Anmeldung, um ein eigenes Blog zu führen, ist meist kostenlos. Google selbst kaufte 2002 einen solchen Blog-Anbieter auf [100]. Die Page-Rank-Werte dort sind teilweise erstaunlich hoch. Daher sollten Sie unter Ihren Freunden und Bekannten nach Bloggern und privaten Websites fragen und sie um die Platzierung eines Verweises bitten. Oder eröffnen Sie ein eigenes Blog und verlinken Sie auf Ihre eigene Site. Für Suchmaschinen gilt dies natürlich auch als unabhängige Empfehlung wie jede andere. Sie bemerken nicht, dass es sich bei dem Blog-Autor und dem Website-Autor um die gleiche Person handelt.

Ein Punkt ist jedoch hierbei zu beachten. So versuchen die großen Suchmaschinen-Betreiber Google, Yahoo und MSN gegen so genannten **Kommentarspam** vorzugehen [101]. Damit ist das Posten von URLs in Kommentaren zu Blog-Einträgen oder in Gästebüchern gemeint, um die Link-Popularity künstlich in die Höhe zu treiben. Die technische Umsetzung soll dabei über das Einbinden des Attributs `rel="nofollow"` in den Verweis-Tag geschehen. Im Januar 2005 erklärten sich bereits mehrere Blog-Community-Anbieter zu einer Kooperation bereit und sicherten das Einbinden des Attributs zu. Ohne diese Einbindung würde die Erkennung ungemein schwieriger ausfallen. Daher kann insbesondere bei Blogs außerhalb dieser Communities durchaus noch der ein oder andere Verweis gewinnbringend platziert werden.

In anderen Kontexten kommen natürlich auch Kollegen oder Angestellte als Linkpartner in Frage. Arbeitnehmer erwähnen ihre Arbeitsstelle ohnehin meistens auf privaten Webseiten. Ist Ihnen bekannt, dass jemand sich rege an der Diskussion in Online-Foren beteiligt? Diese bieten nach einem Login die Möglichkeit, eine Signatur zu definieren, die bei jedem Posting automatisch an den Beitrag angehängt wird. Weshalb nicht den Versuch wagen, Angestellte oder Kollegen zu bitten, eine Signatur mit Verweis auf Ihre Website anzulegen? Zugegeben, meistens werden Sie wahrscheinlich nur eine verhaltene Ausrede zu hören bekommen. Aber vielleicht passt ein Forum thematisch zu den von Ihnen angebotenen Inhalten und Sie haben Erfolg mit Ihrem Aufruf.

Natürlich sollten Sie selbst innerhalb von Foren eine Signatur nutzen, die auf Ihre Website verweist. Achten Sie davon unabhängig darauf, dass der Domainname in einer solchen Signatur vollständig ist, d.h. <http://www.domain.de> und nicht www.domain.de. Erfahrungsgemäß fällt einigen Suchmaschinen eine Auswertung mit einem vollständigen URL leichter.

Neben einzelnen Personen eignen sich oftmals Webseiten von Organisationen oder Firmen als Linkpartner. So liegt es sicherlich nahe, dass ein Webautor mit seiner Seite über Taubenzucht bei verschiedenen Taubenzuchtvereinen um einen Verweis bittet. Kein Verein wird dies ablehnen, wenn der Autor das KAKADU-Prinzip beherzigt hat.

Im gewerblichen Bereich bestehen vielfältige Formen von geschäftlichen Beziehungen. Oftmals findet man bei Herstellern von Bauteilen aller Art Verweise auf die weiterverarbeitende Industrie oder umgekehrt. Dies trifft für Firmen-Beziehungen (Business-to-Business, B2B) ebenso zu wie für direkte Beziehungen zum Endkunden (Business-to-Customer, B2C). So wird eine Agentur für Webdesign im Impressum des Kunden stets einen Verweis auf die eigene Seite platzieren. Dies erhöht nicht nur die Link-Popularity, sondern führt zusätzlich potenzielle Kunden, denen die Aufmachung einer Seite gut gefällt, zum richtigen Ziel.

Im Prinzip sind dem Webautor bei der Suche nach Personen oder Organisationen zur Platzierung keine Grenzen gesetzt. Solange man auf die genannten Qualitätskriterien achtet, ist jeder eingehende Link ein Gewinn.

Selbst auf dem klassischen Weg der Werbung erreichten viele Webautoren bereits ihr Ziel. Überall werden Newsletter an eine Vielzahl von Interessierten geschickt. Oftmals sind die Newsletter-Autoren dankbar für Tipps und Hinweise auf gute Quellen. Der Verweis auf Ihre Website im Postfach tausender Benutzer kann natürlich nicht von den Suchmaschinen in die Berechnung der

Link-Popularity mit einfließen. Allerdings werden die Newsletter in der Regel im Web archiviert und sind damit auch für Webcrawler zugänglich.

Selbst klassische Öffentlichkeitsarbeit in Offline-Medien kann manchmal zum Erfolg führen. So werden Verweise, die beispielsweise in Zeitschriften vorkommen, oftmals auch auf der zugehörigen Website veröffentlicht, um den Lesern das Abtippen der URLs zu ersparen. Fallen Ihnen keine potenziellen Linkpartner mehr ein, können Sie die Suchmaschinen selbst benutzen, um an weitere zu gelangen. Untersuchen Sie dabei auch Ihre Mitbewerber. Dazu stellen Google und andere eine Link-Analyse zur Verfügung.

link: www.domain.de

Dieses Kommando veranlasst beispielsweise Google, alle erfassten Seiten aufzulisten, die auf www.domain.de verweisen. Schaut man sich die Art des Verweises und den Kontext an, kann es dann und wann durchaus möglich sein, sich dort ebenfalls mit einem Verweis auf die eigenen Seiten zu platzieren.

Schließlich kann man auch nach der Phrase "add url" suchen. Man erhält eine Unmenge an Einträgen, von denen auf jedem einzelnen die Eintragung eines Verweises möglich ist. Fügt man der Phrasensuche zusätzlich ein Stichwort bei, erhält man eventuell thematisch verwandte Seiten, auf denen man problemlos einen Verweis setzen lassen kann.

7.4.6 Link-Farmen und Google-Bomben

Allerdings sollte man bei der Suche nach Linkpartnern stets die Qualitätskriterien im Hinterkopf behalten. Als die Link-Popularity noch in den Anfängen steckte, bildeten sich schnell lange Listen mit unzähligen Verweisen. Diese Listen waren dabei keineswegs thematisch sortiert oder in irgendeiner Weise gepflegt wie die URL-Datenbank des Open Directory. Ein Netz solcher Seiten schloss sich zusammen und führte automatisierte Austauschprogramme für Links ein. Vor solchen Link-Farmen (Link-Farms) sollte man sich als Webautor hüten. Ein dort enthaltener Verweis kann unter Umständen bereits als Spamversuch gewertet werden, da Suchmaschinen in diesen Linkansammlungen eine Gefährdung des Link-Popularity-Prinzips sehen. Eine negative Auswirkung über die Bad-Rank-Systematik ist ebenfalls wahrscheinlich. Außerdem muss bei der Eintragung oftmals die eigene E-Mail-Adresse angegeben werden. Wenige Monate später wird man dann selbst zum Spamopfer durch unerwünschte Werbe-E-Mails.

Der pushende Effekt von Link-Farmen ist mittlerweile nicht mehr nennenswert. Eine viel effektivere Methode entstand in den letzten Jahren im Zusammenhang mit Online-Communities. Der Begriff Google-Bombing [102] hat sich

hier für einen gezielten Missbrauch der Link-Popularity eingebürgert. Dabei wird ein vereinbarter Linktext von allen Mitgliedern einer Community gesetzt. Dies führt natürlich bei entsprechender Größe der Community zu einer enormen Anzahl an Links, die wiederum zu einer hohen Bewertung der Seite, auf die verwiesen wird, durch die Suchmaschinen führen. Besonders beliebt wurden Google-Bomben im Jahre 2003 bei Gegnern des amerikanischen Präsidenten George W. Bush. Tausende von Bloggern und Website-Betreibern setzten einen Link auf Bushs Seite mit dem Linktext "miserable failure" (dt. jämmerlicher Versager). Kurze Zeit später war die Website des Präsidenten bei einer entsprechenden Suchanfrage auf Position Eins. Gegen eine solch gezielte Manipulation können Suchmaschinen nur durch Sperrung einzelner Webseiten vorgehen, was angesichts der enormen Anzahl unmöglich erscheint. Dieser Effekt kann natürlich auch im positiven Sinne ausgenutzt werden, wenn man über entsprechende Kontakte innerhalb einer großen Online-Community verfügt.

7.5 Click-Popularity erhöhen

Das zweite Ranking-Verfahren, das auf den hypertextuellen Eigenschaften des Webs aufbaut, ist deutlich schwieriger zu beeinflussen als die Link-Popularity. Außerdem besitzt die Click-Popularity kaum noch eine Bedeutung. Sie wird zwar von diversen Suchmaschinen eingesetzt, dient aber allem Anschein nach eher zur Überprüfung der Qualität der eigenen Ergebnislisten.

Eine Optimierung ist daher schwierig, weil kaum Parameter verändert werden können, die zu einem besseren Ranking führen. Der Suchende muss durch den Eintrag in der Ergebnisliste davon überzeugt werden, auf diesen einen betreffenden Link zu klicken und auf keinen anderen. Dazu steht dem Webautor von Seiten der Onpage-Optimierung nur der Title-Tag zur Verfügung. Dieser wird angezeigt und bei Beachtung der Parameter wie der Länge auch entsprechend unverändert angezeigt. Die dazugehörige Beschreibung wird allerdings selten aus dem Description-Meta-Tag entnommen. Vielmehr arbeiten Suchmaschinen mit eigenen Snippets aus dem Dokumentenkörper.

Wie kann man, abgesehen von einem ansprechenden und animierenden Titel, die Click-Popularity erhöhen? Bei einem Klick wird ein interner Zähler um eins erhöht. Um zu verhindern, dass ein Webautor selbst unzählige Klicks auf seine Einträge tätigt, wird entweder ein Cookie eingesetzt oder die IP-Adresse des Clients für eine Zeitspanne notiert. Ist bereits ein entsprechender Cookie vorhanden oder dieselbe IP-Adresse im Sperrfilter, so führt der Klick nicht mehr zu einer Erhöhung des Zählers. Der Einsatz von Cookies findet dabei meist nur unterstützend zu dem IP-Filter statt, da Cookies auf der lokalen Festplatte des

Suchdienst	Technologie	Katalog
AllTheWeb	FAST	ODP
Altavista	Inktomi	Looksmart
AOL	Google	ODP
Fireball	FAST	Allesklar
Google	Google	ODP
HotBot	Inktomi	ODP
Lycos	FAST	Allesklar
MSN	MSN Search	Looksmart
T-Online	Google	Allesklar
Web.de	SmartSearch [110]	Web.de
Yahoo	Inktomi	Yahoo

Tabelle 9.1 Überblick über die Suchtechnologien

Dabei liefern die Webkataloge mittlerweile ebenso wichtige Daten für die Relevanzbewertung aus dem jeweiligen Datenbestand. Die Bewertungen von Redakteuren haben daher oftmals einen großen Einfluss auf das endgültige Ranking bei Suchmaschinen.

Es sollte noch erwähnt werden, dass die gleiche zugrunde liegende Suchtechnologie nicht zwangsläufig bedeutet, dass die Suchergebnisse bei gleichen Anfragen identisch sind. Jede Suchmaschine stellt natürlich durch eigene Relevanzberechnungen mehr oder weniger unterschiedliche Ergebnisse auf gleiche Suchanfragen dar. Die Suchtechnologie bezeichnet, vereinfacht gesagt, den »Kern« des Information-Retrieval-Systems. Wie das Ranking in dem Mix von Bewertungskriterien entsteht, ist nach wie vor überwiegend suchmaschinen-spezifisch.

9.2 Die Anmeldung

Vor der Anmeldung stellt sich zunächst die Frage, welche Suchmaschinen denn überhaupt relevant sind. Bei der enormen Breite des Angebots an Suchmaschinen bieten sich natürlich zunächst jene an, die auch von den Webusern am häufigsten und intensivsten genutzt werden.

Bei der Betrachtung des Besucherverhaltens wurde bereits festgestellt, dass Google mit Abstand die meisten Nutzerzahlen vorzuweisen hat. Mit einem

gehörigen Abstand folgt die Konkurrenz. Bei einer Anmeldung sind daher insbesondere folgende Suchdienste zu berücksichtigen:

- ▶ Google
- ▶ Yahoo
- ▶ MSN
- ▶ AskJeeves
- ▶ Lycos

Nur aufgrund der Marktführerschaft von Google sollte man jedoch nicht den Fehler begehen, sich lediglich auf diesen einen Betreiber zu konzentrieren. Dass Google kurz nach Einführung den damaligen Marktführer Altavista in einem unglaublichen Tempo überholte, kann in anderer Form durchaus erneut geschehen. Auf dem Markt der Suchmaschinen zählen vor allem die Qualität der zurückgelieferten Suchergebnisse, die Geschwindigkeit und die Benutzerfreundlichkeit. Derzeit ist Google in diesen Punkten unangefochten die Nummer Eins. Jedoch zeigen Studien [111] erste Anzeichen eines leichten Rückgangs. Insbesondere bei der Qualität der Ergebnisse hat die Konkurrenz aufgeholt. Derzeit ist es jedoch unwahrscheinlich, dass ein anderer Anbieter Google vom Thron stürzen wird. Denn das Unternehmen bemüht sich kontinuierlich, sein Angebot und die Qualität der Suche zu erhöhen.

Jedoch wird letztlich die Suchmaschine die größten Nutzerzahlen verzeichnen können, die die genannten Punkte am besten erfüllt. Denn Suchmaschinen können keine Kunden binden, indem sie eine virtuelle Gemeinschaft aufbauen, die den Wert des Produkts an sich erhöht. Von diesem Effekt leben beispielsweise Amazon und eBay. Diese Unternehmen besitzen aufgrund der ausgeprägten Community eine nahezu unverdrängbare, marktbeherrschende Stellung. Denn durch die großen Nutzerzahlen profitieren andere Nutzer wiederum – sei es durch Buchrezensionen oder die große Auswahl an Versteigerungsangeboten.

Ein Webautor tut aus diesem Grund gut daran, sich bei allen großen Suchmaschinen anzumelden. An diesem Punkt muss man allerdings erwähnen, dass sich das Lager in zwei Gruppen teilt. Die einen plädieren dafür, die Anmeldung auf wenige wichtige Suchmaschinen zu beschränken. Andere wiederum verfahren nach dem Motto »viel hilft viel« und tragen ihre Site bei allen möglichen Anbietern ein.

Prinzipiell spricht natürlich nichts dagegen, die Website weithin bekannt zu machen. Allerdings ist mit der groß angelegten Anmeldung bei vielen Suchmaschinen ein gehöriger Zeit- und Kostenaufwand verbunden. Primär sollte man

sich daher zunächst auf die wenigen wichtigen konzentrieren und bei Gelegenheit die kleineren Suchmaschinen bedienen. Diese werden oftmals ohnehin aus dem Datenbestand eines größeren Suchdienstes gefüttert, wie eine Tabelle im nächsten Abschnitt zeigen wird.

Natürlich ist die Anmeldung bei den wichtigen Webkatalogen Open-Directory, Yahoo und Allesklar ebenso wichtig. Das Verfahren diesbezüglich wurde bereits eingangs genau erläutert. Bei der Auswahl der passenden Suchmaschinen für die Anmeldung wird neben den erwähnten Faktoren nicht selten die spezielle Art der Optimierung erwähnt. Vor dem Hintergrund, dass man eine Website besonders für eine einzelne Suchmaschine und deren Ranking-Verfahren optimiert, macht natürlich eine breite Anmeldung weniger Sinn. Denn nur bei dieser einen betreffenden Suchmaschine kommt es zu besonders guten Ergebnissen.

So verständlich dieser Gedanke ist, vernachlässigt er jedoch die Tatsache, dass eine wirklich zielgenaue Optimierung für eine Suchmaschine im Prinzip gar nicht möglich ist. Denn nicht umsonst werden die exakten Bewertungskriterien von den Betreibern geheim gehalten. Eine gezielte Optimierung für eine Suchmaschine muss demnach nach dem Trail-and-error-Verfahren ablaufen. Oft kostet dies so viel Zeit, dass die Suchmaschine in der Zwischenzeit bereits veränderte Bewertungskriterien einsetzt. Häufig wird dies nicht einmal sofort bemerkt, da die Aktualisierung der Bewertung einer betreffenden Site sich beispielsweise aufgrund einer geringen Wiederbesuchsfrequenz verzögert. Die Content-Anbieter verlieren daher regelmäßig den Wettlauf mit den Suchmaschinen, wenn eine spezielle Optimierung auf einen Anbieter angestrebt wird.

Die Strategie sollte vielmehr von der anderen Seite her aufgerollt werden. In Anbetracht der Top-Five-Suchmaschinen muss sich ein Webautor vor Optimierungsbeginn im Netz kundig machen, welche aktuellen Verfahren speziell bei diesen betreffenden Suchmaschinen gute Effekte erzielt haben, und diese dann breit bei dem Optimierungsprozess anwenden. Erfahrungsgemäß haben sich die in den vorigen Abschnitt vorgestellten Methoden bei den großen Suchmaschinen als sehr ergiebig erwiesen.

9.2.1 Manuelle Anmeldung

Zur direkten Anmeldung stellen Suchmaschinen, sofern eine direkte Anmeldung überhaupt möglich ist, Webformulare bereit. Gelegentlich muss man sich jedoch zunächst registrieren, um Zugriff auf diese zu bekommen. Das ist etwa bei Yahoo der Fall. Bei manchen Anbietern ist dieses Formular gleich über einen Verweis auf der Startseite zu finden. Leider ist dies nicht überall derart

pragmatisch gelöst, so dass man sich des Öfteren zunächst über einzelne Seiten regelrecht zum Ziel durchkämpfen muss.

Ist man schließlich zu dem Anmelde-Formular vorgedrungen, stellt sich die Frage, welcher URL angemeldet werden soll. Da der Webcrawler aufgrund dieser Angabe weitere Links der Site sammeln soll, eignen sich zwei Seiten für diesen Zweck besonders gut. Sind alle Punkte der bisherigen Optimierungsstrategie berücksichtigt, sollte man nämlich die Homepage und die Sitemap anmelden. Die Homepage stellt den Ausgangspunkt der Struktur dar, von dem aus alle weiteren Seiten erreichbar sein sollten. Die Sitemap wiederum bietet die direkten Verweise auf alle verfügbaren Seiten der Webpräsenz.

Reichen diese beiden Seiten? In den allermeisten Fällen ist die Frage hier mit einem klaren »ja« zu beantworten. Ich möchte nochmals betonen, dass die manuelle Anmeldung bei Seiten mit einer entsprechenden Anzahl eingehender Verweise von außen ohnehin nur unterstützend wirkt. Die Webcrawler erfassen überwiegend selbstständig das Web. Dies zeigt sich auch in den Möglichkeiten, die dem Webautor zur kostenlosen Anmeldung bleiben. Es lässt sich beobachten, dass der Trend weg von der Anmeldung eines URLs hin zu bezahlten Einträgen geht. Darauf werde ich später explizit eingehen. Ein anderer Faktor ist die Verzahnung der einzelnen Suchdienste. Die folgende Tabelle zeigt eindrucksvoll einen Ausschnitt der sehr eingeschränkten Möglichkeit zur manuellen Anmeldung.

Suchmaschine	Anmeldung
AllTheWeb	Über Yahoo
Altavista	Über Yahoo
AOL	Über Google
Fireball	Über Lycos
Google	Unbegrenzt
HotBot	Über Lycos
Lycos	Nur Katalogeintrag
MSN	Nur gegen Bezahlung (Overture)
T-Online	Über Google
Web.de	Nur gegen Bezahlung
Yahoo	Yahoo

Tabelle 9.2 Übersicht über die Möglichkeiten der manuellen Anmeldung

Es wird deutlich, dass die manuelle Anmeldung mittlerweile nur noch sehr eingeschränkt möglich ist. Für die wenigen Möglichkeiten zur Eintragung gelten jedoch gewisse Erfahrungswerte, die einen URL-Eintrag beeinflussen können.

Der wichtigste Punkt ist dabei das bereits angesprochene Oversubmitting. Übermittelt man in kurzer Zeit zu viele URLs der gleichen Site, kann dieses Verhalten als Spam interpretiert werden. Damit bewirkt man natürlich genau das Gegenteil. Ein Haufen vorgeschlagener URLs der gleichen Domain führt nämlich im Extremfall zum Ignorieren der Einträge oder der Domain insgesamt. Dabei besitzt natürlich jeder Anbieter eine andere Obergrenze für die tolerierte Anzahl an URLs. Erfahrungsgemäß sollte man von einer Webpräsenz nicht mehr als fünf Seiten pro Tag und Suchmaschine anmelden. Google bildet in diesem Punkt bislang noch eine Ausnahme. Der Betreiber schreibt ausdrücklich auf seinen Seiten, dass ein Oversubmitting keinerlei Einfluss auf eine Bearbeitung hat. Ein übermäßiges Eintragen der URLs zeigt allerdings hier wie anderswo wenige Effekte, sondern kostet lediglich wertvolle Zeit.

Das mag sicherlich damit zusammenhängen, dass das Übermitteln eines URLs an eine Suchmaschine wie auch bei den Webkatalogen nicht als Anmeldung im eigentlichen Sinne anzusehen ist. Vielmehr handelt es sich um einen Vorschlag, die betreffende Seite zu besuchen. Der übermittelte URL wird zusätzlich vor der Speicherung in der URL-Datenbank anhand einiger Filter auf seine Qualität untersucht. Dabei wird zunächst sichergestellt, dass es sich um einen syntaktisch richtigen URL handelt. Daneben führen einige Anbieter auch Tests durch, ob der URL überhaupt erreichbar ist. Dies erfolgt mittels eines kurzen HTTP-Requests ohne Content-Übermittlung. Über den zurückgelieferten Statuscode sind die Erreichbarkeit des Servers und die Situierung der Ressource erschließbar. Bei einer aufgefundenen permanenten Weiterleitung wird beispielsweise der entsprechend neue URL genutzt – oder bei restriktiverem Vorgehen der offensichtlich nicht aktuelle URL einfach entfernt. Daneben können bereits an dieser Stelle Ausschlusskriterien wie eine Überschreitung der maximal zulässigen Verzeichnistiefe, die Entfernung von Sonderzeichen oder ähnliche Filter durchlaufen werden.

Letztendlich muss eine Ressource, wie im Abschnitt über die Grundlagen des Information Retrieval beschrieben, alle Filter durchlaufen, um endgültig in die URL-Datenbank aufgenommen zu werden. Nur dann erfolgt schließlich auch der Besuch des Webcrawlers.

Neben diesen Eigenschaften, die sich auf die Ressource an sich beziehen, spielen teilweise spezielle Faktoren eine entscheidende Rolle. So ist bei Google die ausreichende Anzahl an eingehenden Verweisen ein Kriterium für eine Aufnahme. Praktisch gesehen wird demnach ein weiterer Filter hinzugefügt. Der

Datenbestand wird auf Verweise des betreffenden URLs untersucht. Sind hier nicht genügend Verweise vorhanden, gelangt die Ressource nicht durch den Filter und wird entfernt. Damit stellt Google sicher, dass nur Seiten aufgenommen werden, die nicht lose im Web stehen und eine gewisse Bedeutung in Form von eingehenden Verweisen im Sinne einer Empfehlung besitzen.

Meist wird bei der Anmeldung zusätzlich eine kurze Beschreibung verlangt. Deren Bedeutung ist bislang jedoch nicht klar. Man sollte auf jeden Fall darauf achten, dass der mitgelieferte Text auch tatsächlich beschreibenden Charakter besitzt und bei einer eventuellen Betrachtung durch einen Menschen informativ ist. So empfiehlt sich eine knappe Beschreibung der vorgeschlagenen Ressource. Nach dem Absenden des Formulars erhält der Nutzer meist eine Bestätigung, dass die Anfrage bearbeitet wird. Dies ist wie erwähnt jedoch nicht im Sinne einer Aufnahme in die URL-Datenbank zu verstehen. Es wird lediglich mitgeteilt, dass die Übertragung der Daten erfolgreich war. Der Vorschlag muss den vorgenannten Filterkriterien genügen.

9.2.2 Automatische Anmeldung

Neben der manuellen Eintragung gibt es natürlich auch die Möglichkeit, diese Aufgabe durch den Einsatz von Software zu lösen. Dabei existieren neben Online-Tools auch viele eigenständige Programme, die zunächst installiert werden müssen. Der Markt ist hier sehr unübersichtlich, und die Spanne reicht von Freeware über Shareware bis hin zu professionellen Lösungen.

Dabei wird durchgängig versprochen, dass die Software eine Website in meist über 1.000 Suchmaschinen quasi mit einem Klick automatisch einträgt. Bei dem Einsatz solcher Eintragedienste treten jedoch immer wieder bestimmte Probleme auf.

Grundsätzlich zeigt die Erfahrung, dass sich bei den genannten Top-Five-Suchdiensten auf jeden Fall der manuelle Eintrag lohnt. Damit hat man in der Regel ohnehin weit über 85 Prozent an relevanten Suchdiensten abgedeckt. Insbesondere sollte man hier nicht blind auf die automatisierten Prozesse vertrauen. Die Gefahr ist zu groß, dass aus irgendwelchen Gründen die Eintragung auf negative Reaktionen seitens der Suchmaschinen-Betreiber stößt. Insbesondere die großen Betreiber reagieren teilweise empfindlich auf die Verwendung von Eintragesoftware.

Natürlich schadet es nichts, in möglichst vielen Suchmaschinen und Katalogen aufzutreten. Bei über 1.000 kleineren Anbietern ist es auch meist unerheblich, wenn einige Eintragungen misslingen. Generell sollte man jedoch bei der Aus-

wahl eines Eintragedienstes oder einer entsprechenden Software auf folgende Qualitätskriterien achten:

► **Aktualität**

Die Anmeldesoftware simuliert den Versand der entsprechenden Daten über das eigentliche Webformular. Dies ist natürlich bei jedem Anbieter anders gestaltet, und auch der URL zur Verarbeitung der übermittelten Daten ist entsprechend unterschiedlich. Dabei ändern sich die Formulare und damit auch der URL zur Datenübermittlung mit der Zeit. Setzt man hier eine veraltete Software ein, die die Daten zur Übermittlung nicht mehr in einen korrekten URL einpasst, wird die automatische Anfrage abgelehnt. Aus diesem Grund sollten hochwertigere Produkte die Möglichkeit eines Online-Updates bieten.

► **Kennung**

Einige Suchmaschinen begrüßen diese Form der automatischen Anmeldung keineswegs. Aus diesem Grund werten sie die User-Agent-Zeile aus dem HTTP-Request bei der Eintragung aus. Im Falle der manuellen Anmeldung handelt es sich hierbei um eine gültige Browser-Kennung. Manche Tools senden allerdings eine eigene, spezifische Programmkennung mit. In diesem Fall ist der Versuch einer automatisierten Eintragung natürlich leicht zu entlarven. Daher sollte man darauf achten, dass das verwendete Produkt eine gültige Browser-Kennung vortäuscht.

► **Differenziertheit**

Insbesondere bei weniger professionellen Produkten kann oftmals nicht genügend Datenmaterial als Grundlage eingegeben werden. Viele Suchdienste möchten eine E-Mail-Adresse oder Ähnliches mitgeliefert bekommen. Einfache Tools bieten erstaunlicherweise nicht einmal die Möglichkeit, derartige Daten im Voraus einzugeben. Aber auch die Abfassung von Beschreibungstexten in unterschiedlichen Sprachen sollte möglich sein. Denn insbesondere bei Katalogen wird sehr auf das Einhalten solcher Punkte geachtet.

9.2.3 Aufnahmedauer

Einmal eingetragen, wartet man nicht selten eine gehörige Zeit, bis sich die Website in den Ergebnislisten der Suchmaschinen zeigt. Die gefühlte Zeit ist gleichwohl wesentlich länger, denn verständlicherweise sehnt sich ein Webautor danach, das Ergebnis seiner Optimierungsmühen zu sehen. Hat man es bereits beim ersten Versuch auf einen ansehnlichen Platz geschafft?

Die Spanne von der Anmeldung bis zum Erscheinen ist dabei je nach Anbieter unterschiedlich lang. Sie ist jedoch nicht nur abhängig von der Aktivität der

Webcrawler. Wie bei der Darstellung der Funktionsweise von Information-Retrieval-Systemen deutlich geworden ist, durchläuft eine Ressource etliche Stufen, bis sie endgültig im Index eingelagert wird. Die erste beobachtbare Reaktion auf eine Anmeldung ist zunächst natürlich der Besuch des Webcrawlers. Diesen kann man im Logbuch des Webservers feststellen.

Google setzt für die Neuerfassung von Ressourcen einen eigenen Typ von Webcrawler ein. Der so genannte Fresh-Bot von Google ist, wie im Abschnitt zu Cloaking gezeigt wurde, durch einen gesonderten IP-Adressbereich identifizierbar. Allerdings tauchen die Seiten in wenigen Fällen unmittelbar nach einem Crawler-Besuch im Index auf. Insbesondere bei großen Suchmaschinen kann es sich hier um mehrere Wochen handeln. Dies liegt zum einen natürlich an dem enormen Aufkommen an zu bewältigenden Daten, zum anderen aber auch an dem Abgleich des Indexes zwischen den verschiedenen Rechenzentren.

Vor Mai 2004 bezeichnete man dieses Phänomen bei Google als den Google-Dance. Einmal im Monat wurde der Index mit den neu erfassten und aktualisierten Dokumenten neu berechnet und auf alle Server überspielt. Dabei änderten sich natürlich die Ranking-Positionen im unterschiedlichen Ausmaß. Der Abgleich dauerte teilweise mehrere Tage, so dass man in dieser Zeit auf verschiedenen Servern unterschiedliche Ranking-Ergebnisse erhielt. Die Ergebnisse hüpfen quasi hin und her – daher auch der Name **Google-Dance**. Seit Mai 2004 trennte Google die Erfassung von der Bewertung und dem Überspielen des Indexes, so dass diese seitdem zeitlich getrennt voneinander stattfinden. Die Page-Rank-Berechnung scheint sich dabei auf einen ein- bis dreimonatigen Rhythmus einzupendeln. Dieser unter dem Namen **Backlink-Update** bekannte Prozess wird auf der Seite von Sebastian Karpp [112] beobachtet. Dabei ist zu beachten, dass wieder einmal mehr die individuelle Bedeutung der Ressourcen eine Rolle spielt, die über die eingehenden Verweise von außen bestimmt wird. Ein Online-Tool erlaubt den direkten Vergleich der Top-10-Treffer auf drei verschiedenen Google-Servern [113]. Hierbei wird zusätzlich eine automatische Benachrichtigung per E-Mail angeboten.

Aufgrund des Google-Dance, der in ähnlicher Form natürlich bei allen anderen Suchmaschinen dieser Größe auftritt, kommt es daher selbst bei Vorhandensein der Daten in den Datenbeständen zu einer Verzögerung der Anzeige.

Generell liegt die Spanne ab der Anmeldung bzw. dem Besuch des Webcrawlers bis zum Erscheinen im Index zwischen 24 Stunden und vier bis acht Wochen oder länger. Aus den Angaben der Suchanbieter und Erfahrungswerten kann man die Daten in folgender Tabelle zusammenstellen. Aufgrund der unterschiedlichen Faktoren dienen diese jedoch lediglich zur groben Orientierung.

Suchmaschine	Erstanmeldung	Nach Wiederbesuch
Altavista	1–2 Wochen	< 3 Tage
Dino	1 Woche	< 3 Tage
Fireball	1 Tag	Mehrere Wochen
Google	6–12 Wochen	Je nach Relevanz
Lycos	< 4 Wochen	< 4 Wochen
MSN	< 3 Wochen	Unbekannt
Yahoo	6–12 Wochen	8 Wochen

Tabelle 9.3 Ungefähre Anmeldedauer bei Suchmaschinen

Man erkennt hierbei teilweise recht deutlich, ob die Präferenz einer Suchmaschine eher auf der Neuerfassung von Webseiten liegt wie etwa bei Fireball oder ob die Pflege des bereits erfassten Datenbestandes im Vordergrund steht.

9.3 Kostenpflichtige Leistungen

Vor allem die bedeutenden Suchmaschinen sind kein Dienst an der Menschheit, sondern vielmehr Unternehmen, die vor allem ein primäres Ziel verfolgen. Und das ist nicht die Informationserschließung des Web, sondern banal ausgedrückt die Steigerung des eigenen Gewinns.

Suchmaschinen-Betreiber verfolgen dabei unterschiedliche Strategien. Zum einen lässt sich ein beliebtes Prinzip erkennen, das auch in der Offline-Welt häufig zum Erfolg geführt hat. Zunächst bietet man eine Dienstleistung kostenlos an und wartet, bis die Nutzer sich daran gewöhnt haben und sie in ihren Alltagsgebrauch eingebunden haben. Ab dem Zeitpunkt, wenn dieser Gewöhnungseffekt einsetzt und Betreiber mit Verlässlichkeit regelmäßige Benutzer vorweisen können, werden die Dienste und Serviceleistungen Stück für Stück in kostenpflichtige Angebote umgewandelt. Dabei bleiben oftmals die alten Services in einer abgespeckten Form erhalten, jedoch im Schatten der bevorzugten kostenpflichtigen Angebote.

Die Unternehmensgruppen, die die Suchdienste anbieten, sind hierbei keine Ausnahme. So verschwindet zusehends die Möglichkeit, URLs kostenlos anzumelden. Stattdessen wachsen Programme aus dem Boden, die gegen einen finanziellen Aufwand eine bessere und schnellere Erfassung ermöglichen und weitere Neuerungen beinhalten. Die bislang kostenlosen Dienste leiden natürlich darunter, weil sie zum einen durch Veraltung einem sicheren Tod entge-

Index

A

Administrator 240
AdSense 307
AdWords 307
Agentur 168
Ähnlichkeitsbestimmung 116
Aktualität 246, 247
Algebra, boolesche 22
Allesklar 16
AllTheWeb 23, 145
Alt-Attribut 227, 231
Altavista 22, 40, 194, 305
Amazon 294
Anchor 101, 182
Anchor-Text 226
Apache 195
API 323
ARPANET 51
ASP 180, 194
Aufnahmedauer 299
Auswertung
 aggregierte 315
 Fehlercodes 321
 Herkunftsländer 317
 pro Tag 315
 Seitenbesuche 317
 Tools 314

B

Backlink-Update 300
Bad-Rank 132
Banner-Blindness 277
Bellnet 16
Berners-Lee 33
Besucherverhalten 318
Bilder 182, 231
Bildhöhe 233
Black List 77, 96
Blickverlauf 217
Breadcrumb 177
Breitensuche 247
Brin, Sergey 65, 123
broken links 20
Brotkrumen 177
Browser-Weiche 284

Brückenseite 280
Buchdruck 34

C

Caching 46
Cascading Style Sheets → CSS 47
CERN 33
Checksumme 68, 287
Click-Popularity 133
 erhöhen 266
 Güte 136
 IP-Sperre 134
 Snippets 135
Client 50
Client-Server-Prinzip 50
Cluster 69
Clustering 110, 136
 conceptual 139
 Google 137
 Single-Pass 139
 Teoma 138
 thematisch 139
Clustervalidität 98
Community 294
Conflation 91
Constraints 79
Content-Management-System 54, 192
Cookies 134, 154, 266
Cosinus 116
Cronjob 249
CSS
 Datei 175
 Einbindung 175
 Formatierung 48
 Formatierungsregel 48
 Hervorhebung 223
 korrektes 172
 Text-Hiding 276
 Trennung 47
 Tutorials 47
CSSZenGarden 175

D

Dangling Pages 177
Data Manipulation Language 80

- Dateinamen 244
- Dateisystem, invertiertes 102
- Dateityp 153
- Datenaufbereitung 78
- Datenbestand
 - Aktualität 69
 - Grundlage 111
 - Integrität 195
 - Normalisierung 142
- Datennormalisierung 84, 243, 278, 284
- Datenstruktur, direkte 105
- Deep Web 252, 253
- DENIC 21
- Deskriptor 97, 120
- Direkte Datei 106
- DNS 53, 71
- DNS-Balancing 109
- DocID 67, 77, 106, 108
- Dokumentanalyse 78
- Dokumentenindex
 - Aufgaben 67
 - Inhalt 68
 - Zuordnung 106
- Dokumenttyp 75
- Domain 240
 - Domainfilter 152
 - Domainname 242
 - Top Level Domain (TLD) 54
- Domain Name Service → DNS 53
- DOS-Attacken 311
- Dublettenerkennung 75

E

- eBay 294
- Editor 16
- Eingabemaske 22
- Eisbergeffekt 129
- E-Mail 49
- Embedded Link 179, 182
- Ergebnisseite 22
- erweiterte Suche 149
- Esporting 305
- Excel 153
- Excite 16
- Express-Inclusion 303

F

- Fakten Retrieval 79

- Fehlercode 179
- Fireball 23, 40, 104, 149, 152
- Flash
 - Auswertungsmechanismen 72
 - Intro 20, 190
 - Navigation 180
 - Skip-Funktion 20
- Fließtext 192, 219, 223
- Frames 182
 - Frameset 183, 184
 - Konzeption 183
 - Noframes 185
 - Prinzip 183
 - Problem 188
 - Target-Attribut 184
- FTP 49, 241
- Fuzzy-Logik 114

G

- Geocities 240
- Gesponsorte Treffer 305
- Gewichtungsvalidität 98
- Google 22, 46, 104, 171
 - Google-Image 257
 - Videosuche 191
 - Voice-Search 15
- Googlebot 46
- Google-Dance 300
- googlen 161
- Großschreibung 104
- Gutenberg 34

H

- Hashwert 248
- Hidden-Markov-Modelle 90
- Hitlist 102, 108
- Homepage 15, 247
- Host 50
- Hotbot 303
- htaccess 195, 252
- HTML
 - Auszeichnungssprache 35
 - Body 36
 - Container-Tags 36
 - Container-Tags mit Zusatz 37
 - Dokumentkopf 35
 - Dokumentkörper 36
 - Dokumentstruktur 35

- dynamisch 180, 193
- Empty-Tags 36
- fehlerfreies 84
- Formular 234
- gültiges 172
- Head 35
- Prüfung 174
- SGML-Abkömmling 34
- Standard 34
- statisch 192
- Tag 35, 36
- HTTP
 - 404 - Not Found 62
 - Abkürzung 55
 - Ablaufschema 56
 - Accept 60
 - Aufbau, schematischer 57
 - If-Modified-Since 59
 - If-None-Match 59
 - Kommunikation 56
 - Methoden 58
 - Monitoring 310
 - Request 56, 58
 - Response 56, 61, 71, 311
 - Response-Codes 61
 - Statusbereiche 62
 - Statuscodes 61, 73
 - User-Agent 60
 - Versionen 56
- Huffman-Code 105
- I**
 - ICQ 49, 219
 - Image-Map 231
 - Inbound-Links 123
 - Index 178
 - direkter 105
 - invertierter 107
 - Metapher 78
 - Indexer 82
 - Indexierung 25
 - kontrollierte 109
 - unkontrollierte 109
 - Informatik 80, 105, 110
 - Information Retrieval
 - Definition 25
 - Unterschied 79
 - Wissenschaft 11
 - Information-Pages 281
 - Information-Retrieval-System
 - Aufgabe 78
 - Dokumentenrepräsentation 25
 - Herausforderung 111
 - Informationswiedergewinnung 323
 - optimale Werte 92
 - Informationsbedürfnisse 155
 - Inktomi 41, 42, 194, 256
 - Inquirus 30
 - Internet 291
 - Aktualität 292
 - Client-Server-Prinzip 50
 - Protocol (IP) 52
 - Service Provider (ISP) 316
 - Society 27
 - Trägermedium 49
 - Internet Protocol → TCP/IP 52
 - Inverse Dokumenthäufigkeit 119
 - invertierte Pyramide 220
 - IP-Sperrung 241
 - IRC 49
 - ISO-OSI-Modell 50
 - IVW 233
- J**
 - JavaScript 187
 - Jugendschutz 95
- K**
 - KAKADU-Prinzip 259, 264, 279
 - Keyword
 - Analyse 201
 - Dichte 221, 281
 - Extrahierung 97
 - Häufigkeit 281
 - Prominenz 281
 - Stuffing 218, 221
 - Kleinschreibung 103
 - Konzept 168
- L**
 - Lexikon 90, 106
 - Link-Farm 123, 265
 - Link-Popularity 318
 - Definition 122
 - erhöhen 257
 - Exklusivität 261

- interne Verlinkung 258
- Konzept 123
- Links erzielen 262
- Page-Rank 123
- prüfen 260
- Qualitätskriterien 260
- Suchbegriffe 261
- Themenkreise 261
- Linkstrukturen 249
- Linux 71, 245
- Listing-Enhancement 304
- Livesuche 164
- Location List → Hitlist 102
- Logbuch 200
- Logfile
 - Analyse 60, 241, 313
 - Format 314
- Logik
 - AND 147
 - boolsche 146
 - NOT 147
 - OR 147
- Luhn 100
- Lycos 23, 152, 191, 194

M

- Managed-Server 240
- Markenamt 202
- Matching 142
- MD5 68
- Mehrwortgruppenidentifikation 94
- Metacrawler 30, 207
- MetaEureka 30
- Metainformationen → Meta-Tags 38
- Metasuchdienst → Metasuchma-
schinen 27
- Metasuchmaschinen 14, 26
 - Ablaufschema 27
 - All-In-One 27
 - Clustertechnik 31
 - Einsatzgebiet 28
 - Formale Kriterien 27
 - Kurzbeschreibung 30
 - Nutzerkreise 28
 - Operatoren 29
 - Präsentation 29
 - Ranking 30

- Schnittstelle 29
- Zusammenschließung 28
- Meta-Tags 38
 - Audience 46
 - Author 46
 - Content-Type 45
 - Copyright 46
 - Date 46
 - Description 39, 40
 - Dublin Core 46
 - Expires 44
 - Keywords 41
 - Language 44, 89
 - Mehrfachnennungen 42
 - Meta-Spam 279
 - Missbrauch 38
 - Optimierung 39
 - PICS 47
 - Publisher 46
 - Refresh 45
 - Revisit-After 44
 - robots 42, 255
- Mime-Typ 75
- Monitoring
 - DNS 312
 - Rank 322
 - Server 310
- MSN 95, 303

N

- Navigation 180
- Navigationsleisten 179
- Netzwerke
 - Peer-To-Peer (P2P) 14
- Newsletter 264
- Nutzerverhalten
 - im Web 154
 - Suchmaschinen 159
 - Suchmodus 157
 - Suchverfeinerung 213
 - Surfen 157
- Nutzungsverhalten
 - Logfile 320
 - Rückschlüsse 318
 - von Suchmaschinen 165

O

- Online-Medien 49
- Open Directory 16, 17, 131, 263
- OpenCola 14
- Optimierung
 - Breitband 161
 - Durchführung 170
 - Meta-Tags 38
 - Offpage 170
 - Onpage 170
 - PDF-Dokumente 237
 - Relaunch 170
 - Tags 216
- Organigramm 189
- Oversubmitting 297
- Overture 18, 205, 303, 307

P

- Page Impression 316
- Page, Lawrence 65, 123
- Page-Jacking 288
- Page-Rank 65, 122
 - Aktualität 130
 - Bad-Rank 131
 - Beispiel 125
 - Dämpfungsfaktor 124
 - Distanz 130
 - Effekte 127
 - Formel 124
 - Intelligente Surfer 129
 - Iterationen 127
 - Problem 125
 - Random-Surfer-Modell 128
 - Startwert 125
 - Subject-Specific Popularity 131
- Parser 82
 - Alphabet 86
 - Datenaufbereitung 82
 - Fehlertoleranz 84
 - JavaScript 289
 - Prozess 83
- Payed-Listing 14, 304
- Payed-Placement 304, 305
- Pay-Per-Click 302, 304
- PDF
 - Adobe 153
 - E-Paper 237
 - Optimieren 237

Plattformunabhängigkeit 237

- Webcrawler 71
- Perl 180, 311
- Personalisierung 154, 161
- Pfad 54, 156, 249
- Phantom-Pixel 232, 278
- PHP 180, 194
- Phrasen 77, 148
- Phrasensuche 148, 149, 265
- Plugin 180
- Polyseme 88
- Popup 181
- Port 52, 54
- Powerpoint 153
- Precision
 - Definition 92
 - Mehrwortgruppen 212
 - Practical Precision 146
 - Verminderung 284
- Protokollumstellung 312
- Proximity 121, 217
- Proxy 267

Q

- QualiGo 305
- Query-Prozessor
 - Arbeitsschritte 142
 - Funktionalität 141
 - Matching 144
 - Parsing 142
 - Query 143
 - Searcher 141
 - Stemming 143
 - Stoppworte 143
 - Stoppworteliminierung 143
 - Tokenizing 142
 - Trefferliste 145
 - Wildcard 149

R

- Random Surfer 123
- Recall
 - Definition 92
 - Stemming 93
 - Stoppwort-Eliminierung 96
- Redaktionssystem → Content-Management-System 192
- Redirect 62, 250

- Referenzmodell → ISO-OSI-Modell
 - 50
- Referer 318, 320
- Relative Worthäufigkeit 118
- Relaunch 170, 249
- Relevanzbewertung 112, 221, 239, 293
- Repository 77, 107
- Retrieval
 - Aufgabenbereiche 78
 - boolsches 113
 - Fakten Retrieval 79
 - Fuzzy 114
 - Information Retrieval 81
 - Ziel 253
- RGB-Farbraum 273
- Robots Exclusion Protocol 43, 255, 256
- robots.txt 43, 255
- Root-Server 240
- Router 52

- S**
- Satzstruktur 99
- Scanning 19, 223
- Scheduler
 - Definition 68
 - DNS 71
 - Richtlinien 69
 - URL 70
- Schichten 51
- Schlüsselwort
 - Aussagekraft 196
 - Bereinigung 204
 - Brainstorming 199
 - Definition 97
 - IDF 203
 - in URL 246
 - Liste 201
 - Mitbewerber 200
 - Neue Rechtschreibung 210
 - Numerus 208
 - Schreibweise 210
 - Sonderzeichen 209
 - Strategie 196
 - Übernahme 202
 - Wahl 197
 - Wortfolge 217
 - Wortkombination 212, 213
- Schlüsselwort → keyword 197
- Scriptsprachen
 - Clientseitige 180
 - Navigationsmenüs 180
 - Probleme 180
 - Serverseitige 180
- Searcher 51
- Searcher → Query-Prozessor 25
- Seitenstruktur 176
 - Dubletten 287
 - Kriterien 176
 - Planung 176
 - Suchmöglichkeiten 154
- Semantik 99
- SessionID 194, 254
- SGML 34
- Sitemap 128, 178, 179, 186, 250
- Sitestructur 187, 247, 249
- Snippets 40, 41, 46, 266
- Sonderzeichen 18, 45, 76, 85, 172, 211
- Sorter 108
- Spam 269
 - Bait-And-Switch 285
 - Cloaking 283
 - Domain Dubletten 286
 - Doorway-Page 280
 - E-Mail 269
 - Hidden-Links 278
 - IP-Delivering 284, 285
 - Keyword-Stuffing 270
 - Oversubmitting 290
 - Popups 289
 - Spammeldung 290
 - Text-Hiding 272
 - Text-Smalling 274
- Spiegelseiten 287
- Sponsored Links
 - Bellnet 18
 - Payed-Placement 18
- Sprachen
 - asiatische 85
 - Identifikation 88
 - natürliche 85
 - Sprachfilter 151
- Sprachenerkennung 89, 90
- SQL 80
- Stamm 91
- Startseite 181, 189
- Statusbereich → HTTP 61

- Stemming 91
 - Affix-Removal 93
 - Ähnlichkeitsberechnung 93
 - Lexikon 93
 - Look-Up 93
 - Porter 93
- Stichwortvalidität 98, 99
- Stickiness 134
- Stoppworteleminierung 116
- Stoppwörter 95
- Stoppwortliste 95
- Studie
 - Bertelsmann Stiftung 13
 - DoubleClick 159
 - Forrester Research 160
 - Leseverhalten im Web 220
 - Vividence 161
- Suchaktivität
 - starting 156
- Suchaktivitäten 155
 - Browsing 156
 - Chaining 156
 - Differentiating 156
 - Extracting 157
 - Monitoring 157
 - Starting 155
- Suchbegriffe
 - Beliebtheit 163
 - Logfile 320
 - Nähe 218
 - optimal 162
 - Top 10 162
- Suchmaschinen 21
 - Anmeldung 293
 - Architektur 65
 - Besuche 319
 - Blindanteil 24
 - Datenanalyse 25
 - Datengewinnung 24
 - Definition 13
 - Eingabemaske 22
 - Ergebnisse 26
 - Hürden 23
 - Kernkomponenten 24, 26
 - Kommerzialisierung 24
 - Kooperationen 291
 - Nutzung 21
 - Query-Prozessor 25

- Resultate 23
- Suchtechnologien 293
- User-Interface 22
- Wachstum 21
- Suchmodus
 - Conditioned Viewing 157
 - Formal Search 158
 - Informal Search 158
 - Undirected Viewing 157, 190
 - Untersuchung 158
- Suchoperator 146

T

- Tabellen 229
 - CSS 229
 - Strukturierung 229
 - stukturierte 79
 - Table-Tag 229
 - Table-Trick 230
 - vercodet 106
- Tags
 - Aufzählungen 222
 - Comment 233
 - Div 276
 - Embedded 181
 - Iframe 235
 - Links 226
 - Noscript 234
 - Paragraph 222
 - proprietär 174
 - Title 216
 - Title-Attribut 228
 - Überschriften 225
- TCP/IP 52
 - Adressierung 53
 - IPv6 53
- Term Frequency 118
- Terme 120
 - Art 120
 - Klassen 121
 - Lage 120
 - Leitsatz 120
- Textanalyse 99
- Texthervorhebung 223
- Text-Link 179, 181
- Thesaurus 95, 144, 196
- Tiefensuche 247
- TLD 244

Tokenisierung 85
Tokenizer 85
Tooltip 227
Trigger 79
Trunkierung 149
Trunkierungsoperator 91
Tutorials 259

U

Überschriften 225
Umlaute 90, 104
Uniform Resource Identifiers 55
Uniform Resource Locator 53
URL
 Analyse 121
 Anmeldung 297
 Beispiel 55
 Filter 76
 Parameter 55, 195
 Pfadangabe 54
 Schrägstriche 54
 Überprüfung 76
 URI 55
 URN 55
URL-Datenbank 250
URL-Resolver 101
Usability 128, 168, 202
User-Agent 36
User-Tracking 133, 253, 254, 313

V

Vektorraummodell 115
Vermarktung 169
Verweildauer 134
Verzeichnisname 245
Verzeichnistiefe 246
Visits 316
Vivisimo 31
Vollindexierung 96, 97
Vorbereitungen, strukturelle 172

W

W3C 33
Watchblog 263
Web.de 16, 150, 152
Web-Assziator 207
Webcrawler-System 66
 Crawler 70

Datenspeicher-Module 66
Protokoll-Module 66
Storeserver 72
Verarbeitungs-Module 66
Webdirectory → Webkataloge 15
Webhits 21
Webhosting 239
Webkataloge 13, 15
 Beschreibungstext 19
 Häufige Fehler 20
 Redakteure 16
 Rubrik 17
 Submit-Tools 20
 Textlänge 19
 Titelwahl 17
 URL 40
Weblog 263
Webseite 15
Webserver 50
Website 15
Webspace 77, 240
Webverzeichnis → Webkataloge 15
White List 109
Wildcard 149
Word 153
WordID 106, 144
Wording 182
Wordtracker 207
Wortabstand 148
 ADJ 148
 FAR 148
 NEAR 148
Wortgruppen 85
Wortseparator 85
WWW
 Anatomie 33
 Dynamik 309
 Größe 11
 Wachstumsprozess 33
WYSIWYG 35

X

XML 34

Y

Yahoo 16, 40, 98, 135

Z

Zentroiden 140

Zielgruppe 169

Zielsetzung 169

Z-Index 278

Zipfsche Gesetz 100

Zugriffszeit 106