

Michael Glöggler

Suchmaschinen im Internet

Funktionsweisen, Ranking Methoden,
Top Positionen

Mit 47 Abbildungen und 25 Tabellen

Die Reihe **Xpert.press** des Springer-Verlags vermittelt Professionals in den Bereichen Betriebs- und Informationssysteme, Software Engineering und Programmiersprachen aktuell und kompetent relevantes Fachwissen über Technologien und Produkte zur Entwicklung und Anwendung moderner Informations-technologien.



Michael Glöggler
Via delle Vergini 9
53100 Siena
Italien
gloeggler@top-webpromotion.de



2003. 2596

Bibliografische Information Der Deutschen Bibliothek
Die Deutsche Bibliothek verzeichnet diese Publikation
in der Deutschen Nationalbibliografie; detaillierte bibliografische
Daten sind im Internet über <<http://dnb.ddb.de>> abrufbar.

ISBN 3-540-00212-X Springer-Verlag Berlin Heidelberg New York

Dieses Werk ist urheberrechtlich geschützt. Die dadurch begründeten Rechte, insbesondere die der Übersetzung, des Nachdrucks, des Vortrags, der Entnahme von Abbildungen und Tabellen, der Funksendung, der Mikroverfilmung oder der Vervielfältigung auf anderen Wegen und der Speicherung in Datenverarbeitungsanlagen bleiben, auch bei nur auszugsweiser Vervielfältigung, vorbehalten. Eine Vervielfältigung dieses Werkes oder von Teilen dieses ist auch im Einzelfall nur in den Grenzen der gesetzlichen Bestimmungen des Urheberrechtsgesetzes der Bundesrepublik Deutschland vom 9. September 1965 in der jeweils geltenden Fassung zulässig. Sie ist grundsätzlich vergütungspflichtig. Zuwiderhandlungen unterliegen den Strafbestimmungen des Urheberrechtsgesetzes.

Springer-Verlag Berlin Heidelberg New York

ein Unternehmen der BertelsmannSpringer Science+Business Media GmbH

<http://www.springer.de>

© Springer-Verlag Berlin Heidelberg 2003

Printed in Germany

Die Weitergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, dass solche Namen im Sinne der Warenzeichen- und Markenschutzgesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürfen. Text und Abbildungen wurden mit größter Sorgfalt erarbeitet. Verlag und Autor können jedoch für eventuell verbliebene fehlerhafte Angaben und deren Folgen weder eine juristische Verantwortung noch irgendeine Haftung übernehmen.

Satz: Computer to film von pdf Daten der Firma perform, Heidelberg
Umschlaggestaltung: Kinkellopka Werbeagentur, Heidelberg
Gedruckt auf säurefreiem Papier 33/3142PS - 5 4 3 2 1 0

Vorwort

Das Internet hat mit der Kommerzialisierung des World Wide Webs Mitte der neunziger Jahre ein Wachstum erreicht, das bisher von keinem anderen Medium erzielt wurde. Die Hauptursache der rapiden Entwicklung weltweit liegt in der Besonderheit des World Wide Webs, einer relativ einfachen und kostengünstigen Erstellung, Verwaltung und Veränderung von Informationsinhalten sowie deren globalen Publizierung und Zugriffsmöglichkeit. Wesentlich unterstützt wurde das Wachstum durch Erweiterungen der ursprünglich rein Text basierten Informationsangebote durch Bild-, Audio- und Videodateien.

Während Mitte der neunziger Jahre von der Mehrzahl aller Unternehmen das World Wide Web als nationaler oder globaler Marktplatz in seiner zukünftigen Bedeutung bei weitem noch nicht erkannt wurde, zählt das Internet heute für viele Branchen als eines der wichtigsten Medien und ist fester Bestandteil des Marketing-Mix geworden. Betrachtet man die Nutzerzahlen im Jahre 2002 alleine nur für Deutschland mit circa 30 Millionen Usern, wird das Potential des Marktplatzes Internet überaus deutlich.

Eine sehr wichtige Rolle im elektronischen Markt nehmen die Suchmaschinen und Webkataloge ein. Sie stehen als Mittler zwischen Informationsanbietern und Informationssuchenden. Entsprechend aktueller Studien beginnt zu 90 Prozent eine Suche im Internet nach Informationen, Produkten oder Dienstleistungen über die Suchhilfen. Dieser Umstand macht es für Content-Anbieter unabdingbar, dass sie möglichst optimal im Datenbestand der Suchhilfen vertreten sein müssen, da sie ansonsten nicht gefunden werden. Optimal vertreten bedeutet, dass das Informationsangebot eines Content-Anbieters über nur einige wenige Suchworte schnell von Usern gefunden wird.

Diesem Anspruch steht jedoch die enorme Menge an Informationsangeboten gegenüber, die es erschweren schnell gefunden zu werden. Darüber hinaus besteht nach wie vor ein erheblicher Mangel an Kenntnissen wie Websites gestaltet werden müssen, um bei den Suchmaschinen eine relevante Rangposition zu erreichen.

Ziel dieses Buches ist es, einfach und ohne viel Fachterminologie in die Technologien und Funktionsweisen von Suchmaschinen im Internet einzuführen. Dabei werden alle erforderlichen Informationen, Verfahren und Methoden so umfassend dargestellt, dass es jedem gelingt Websites auf den vorderen Rängen bei den Suchmaschinen zu positionieren. Optimal positionieren bedeutet in diesem Zusammenhang, eine Website technisch und inhaltlich so zu gestalten, dass sie

von den Suchmaschinen in ihren Datenbestand aufgenommen und unter bestimmten Suchwörtern eine möglichst hohe Rangposition in den Suchergebnissen erzielt. Ein umfangreiches Optimierungskonzept (Kap. 6 bis 11) zeigt auf einfache Weise, wie Websites für Suchmaschinen aufbereitet werden sollen, welche Fehler man nicht begehen darf und wie eine Indexierung dauerhaft überprüft werden kann.

Zum besseren Verständnis welche Arten von Suchhilfen im Internet existieren und wie ihre technischen Unterschiede Einfluss auf das Ranking haben, werden die verschiedenen Typen in Kapitel 1 kurz zueinander abgegrenzt. Für all diejenigen die mit den Grundlagen des Hypermedia nicht vertraut sind, geht Kapitel 2 auf diejenigen Bereiche und Parameter ein, die zur Optimierung der Rangposition von Websites Einfluss haben.

Kapitel 3 beschäftigt sich sehr umfangreich mit dem technischen Hintergrund der Suchmaschinen. Ziel ist es, durch eine leicht verständliche Erklärung der Funktionsweisen von Suchmaschinen grundlegende Kenntnisse zu vermitteln. Mit diesem Wissen ist es möglich, die unterschiedlichen Verfahren der Relevanzbewertung zu verstehen, bewerten und die Methoden zur Optimierung von Websites sicher anwenden zu können. Im Mittelpunkt stehen dabei die Methoden wie Suchmaschinen Dokumente im Internet finden und wie sie geeignete Parameter zur Bewertung von Dokumenten gewinnen können. Kapitel 4 verdeutlicht sehr gut, auf welchen Bewertungsverfahren Suchmaschinen aktuell entscheiden, welches Dokument die höchste Relevanz zu einer Suchanfrage hat und deshalb auf Position eins im Suchergebnis erscheint. Diese Gewichtsmodelle haben direkten Einfluss auf die Optimierung von Websites und werden deshalb ausführlich erläutert.

Wie Suchmaschinen ihre Suchergebnissen aufbauen und welche Suchmethoden sie Anwendern zur Verfügung stellen, wird in Kapitel 5 erklärt. Speziell die Art wie Anwender bei den Suchmaschinen suchen können und wie sie tatsächlich Suchanfragen stellen, hat Auswirkung auf die inhaltliche Gestaltung von Websites.

Aus den Systematiken des Hypermedia sowie den Funktionsweisen der Suchmaschinen lassen sich Handlungsanweisungen ableiten, wie Websites am besten gestaltet werden müssen. In Kapitel 6 werden dazu alle Methoden die auf ein Dokument angewendet werden sollten, umfangreich und leicht verständlich dargestellt. Kapitel 7 beschreibt fortführend zahlreiche Verfahren die sich aus dem Hypermedia ergeben und gleichfalls zur Website Optimierung empfohlen werden.

Eine Einschränkung der Indexierung stellen die Spam-Regeln der Suchmaschinen dar. Spam sind unerlaubte Verfahren, die zum Ausschluss aus dem Datenbestand bei Suchmaschinen führen können. Um nicht Gefahr zu laufen Spam zu betreiben, wird in Kapitel 8 detailliert auf das Thema Spam eingegangen und dazu einige Tipps und Tricks gezeigt, um die Spamererkennung zu umgehen. Kapitel 9 beschreibt die verschiedenen Verfahren, um in den Datenbestand von Suchmaschinen aufgenommen zu werden, als auch mögliche Fehler eines Content-Anbieters, die dies verhindern können.

Der käufliche Erwerb von Rangpositionen stellt mittlerweile eine von nahezu allen Suchmaschinen angebotene Methode dar, schnell auf prominenten Positionen zu erscheinen. Für verschiedene Content-Anbieter stellt dies durchaus eine Alternative zur technisch basierten Website Optimierung dar und wird aus diesem Grunde umfassend in Kapitel 10 behandelt.

Die technische Überwachung der Funktionsfähigkeit der eigenen Website sowie eine permanente Analyse der Server Log Files schaffen die Voraussetzung für ein dauerhaft gutes Ranking bei Suchmaschinen. In Kapitel 11 werden die wichtigsten Web Site Monitoring und Web File Analysen mit Handlungsempfehlungen vorgestellt.

Nach dem Lesen dieses Buchs verfügt der Leser über sehr gute Kenntnisse des technischen Systemumfeldes von Suchmaschinen im Internet, deren Funktionsweisen sowie den Kriterien der Relevanzbildung. Es wird ein umfassendes Spezialwissen bereitgestellt das es ermöglicht, eigene Websites technisch und inhaltlich so zu optimieren, dass eine optimale Rangposition erzielt wird. Hierzu werden alle anzuwendenden Methoden und Verfahren umfassend und leicht verständlich erklärt und dargestellt. Zahlreiche Abbildungen, Tabellen und eine Vielzahl an weiterführenden Links ins Internet unterstützen ein optimales Verständnis. Bei genauer Betrachtung und Einsatz der in diesem Buch beschriebenen Systematiken ist es folglich jedem möglich, eine Website auf Seite 1 bei allen relevanten Suchmaschinen zu positionieren.

Viel Spaß beim Lesen und noch mehr Erfolg bei der Optimierung von Websites!

Michael Glöggler

Inhaltsverzeichnis

1	Überblick Suchdienste im World Wide Web.....	1
1.1	Webkataloge	1
1.2	Suchmaschinen.....	1
1.3	Metasuchmaschinen.....	1
1.4	Payed Listing Suchmaschinen.....	1
2	Systemumfeld der Suchmaschinen im WWW	1
2.1	Client-Server-Prinzip bei Suchmaschinen.....	1
2.2	Das HTTP-Protokoll	1
2.2.1	HTTP-Protokollinformationen.....	1
2.2.2	Server-Statusinformationen	1
2.3	Funktion und Aufbau eines URL.....	1
2.4	Hypertextdokumente.....	2
2.4.1	HTML-Dokumententitel.....	2
2.4.2	HTML-Metainformationen	2
2.4.3	HTML-Dokumentenkörper	2
3	Funktionsweisen von Suchmaschinen	2
3.1	Webrobots und die Erfassung des WWW	2
3.1.1	Arbeitsweisen von Webrobots im Überblick.....	2
3.1.2	Erfassung des WWW durch Webrobots	3
3.1.3	Loader und URL-Datenbank.....	3
3.1.4	Der Checker.....	3
3.2	Datenaufbereitung und Analyse	3
3.2.1	Information Retrieval Systeme	3
3.2.2	Verfahren der Datenaufbereitung und Analyse im Überblick	4
3.2.3	Datennormalisierung	4
3.2.4	Wortidentifikation.....	4
3.2.5	Sprachidentifikation.....	4
3.2.6	Word Stemming.....	5
3.2.7	Deskriptorengewinnung	5
3.2.8	Stoppwortliste und Black List.....	5
3.3	Datenstrukturen der Information Retrieval Systeme.....	5
3.3.1	Besonderheit der Datenstrukturen von IR-Systemen	5
3.3.2	Direkte Dateien.....	6
3.3.3	Invertierte Dateien.....	6
3.3.4	Indexdatei.....	6

4 Relevanz und Gewichtsmodelle	67
4.1 Zusammenhang von Relevanz und Rangbildung.....	67
4.2 Effektivität von Suchmaschinen.....	70
4.3 Statistische Gewichtsmodelle.....	73
4.3.1 Das Vektorraummodell.....	73
4.3.2 Die relative Worthäufigkeit (TF-Algorithmus).....	76
4.3.3 Die inverse Dokumentenhäufigkeit (TTF-Algorithmus).....	77
4.3.4 Bedeutung der Lage eines Keywords.....	79
4.4 Hypermedia basierte Gewichtsmodelle.....	80
4.4.1 PageRank von Google.....	81
4.4.2 Systematik der Click Popularity.....	87
4.5 Cluster-Verfahren.....	90
5 Suchprozess und Suchformen	95
5.1 Der Query Processor – Suchtool der Suchmaschine.....	95
5.2. Suchoperatoren der Suchmaschinen.....	98
5.3. Expertensuche bei den Suchmaschinen.....	101
5.4. Ergebnislisten der Suchmaschinen.....	106
5.4.1 Suchergebnisliste von Fireball.de.....	107
5.4.2 Suchergebnisliste von Lycos.de.....	109
5.4.3 Suchergebnisliste von Altavista.de.....	110
5.4.4 Suchergebnisliste von Google.de.....	112
6 On the Page Methoden der Optimierung	115
6.1. Priorität der Website Optimierung und Erfordernis eines Konzepts.....	116
6.2. Wahl des Dokumententyps.....	118
6.3. Framesets.....	122
6.4. Keywords und Keyword-Strategie.....	125
6.5. Groß- und Kleinschreibung, Pluralform und Sonderzeichen.....	128
6.6. Der Dokumententitel.....	131
6.7. Die Bedeutung der Meta-Tags.....	137
6.8. Einfluss von Keyword-Dichte.....	144
6.9. Lage von Keywords.....	146
6.10. Keywords in ALT-Tags und Kommentaren.....	149
6.11. Textauszeichnung, Textgröße und Überschriften.....	151
6.12. Link-Strukturen und Verzeichnistiefe.....	153
6.13. Sprachen und Länderindexe.....	156
6.14. Valides HTML.....	160
6.15. Vermeidbare Fehler.....	163
7 Off the Page Methoden der Optimierung	169
7.1. Domain-Name und Bezeichnung der Verzeichnisse.....	169
7.2. Aktualität und Änderungsfrequenz.....	171
7.3. Serververfügbarkeit, URL-Anzahl und IP-Sperrung.....	175
7.4. Link Popularity und Link Analyse.....	178
7.5. Click Popularity und Dokumententitel.....	184

8 Spam – unzulässige Maßnahmen	187
8.1. Keyword Stuffing.....	189
8.2. Blindtexte und Schriftgröße.....	191
8.3. Meta-Informationen ohne inhaltlichen Bezug.....	194
8.4. Doorway Pages.....	195
8.5. IP-Delivering und Cloaking.....	197
8.6. Unsichtbare Hyperlinks.....	200
8.7. Erkennung von Dubletten.....	202
8.8. Spam durch Link Farms.....	204
9 Aufnahme in den Datenbestand	207
9.1. Relevante Suchmaschinen für die Indexierung.....	207
9.2. Suchmaschinen und Kooperationen.....	212
9.3. Anmeldung bei Suchmaschinen.....	216
9.4. Aufnahme gegen Bezahlung.....	220
9.5. Filterketten der Webrobot Systeme.....	229
9.6. Überprüfung der Indexierung.....	232
10 Payed Placement – Ranking gegen Bezahlung	235
10.1. Payed Placement Programme bei Google.....	235
10.2. Trusted Feed-Programm von Altavista.....	238
10.3. Index Connect von Inktomi.....	240
10.4. Overture – Payed Placement.....	242
10.5. Esporting – Payed Placement.....	248
10.6. QualiGo – Payed Placement.....	251
11 Monitoring und Controlling	257
11.1. Website-Monitoring.....	257
11.2. Log File Analyse.....	261
11.3. Rank Monitoring.....	269
Anhang	273
Literaturverzeichnis.....	273
Sachwortverzeichnis.....	275

Abbildungsverzeichnis

Abb. 1.1. Altavista.de Suchergebnisseiste – verkürzte Darstellung	2
Abb. 2.1. Client-Server-Prinzip bei den Suchmaschinen	14
Abb. 2.2. Struktur von HTML-Dokumenten	21
Abb. 3.1. Komponenten eines Webrobot-Systems	29
Abb. 3.2. Struktur eines invertierten Dateisystems	42
Abb. 3.3. Prozesse der Datennormalisierung und Dokumentenanalyse	45
Abb. 3.4. Zusammenhang von Worthäufigkeit und Wortrelevanz	55
Abb. 3.5. Invertiertes Dateisystem – schematische Darstellung	59
Abb. 3.6. Beispiel eines b-tree organisierten invertierten Dateisystems	65
Abb. 4.1. Vergleich binärer Vektor – gewichteter Vektor	74
Abb. 4.2. Vektorrepräsentation im Vektorraummodell	75
Abb. 4.3. Dokumentengewichtung bei Page Rank-Ausgangssituation	84
Abb. 4.4. Dokumentengewichtung PageRank – erste Kalkulation	85
Abb. 4.5. Dokumentengewichtung PageRank – zweite Kalkulation	86
Abb. 4.6. Klickzählung bei Fireball.de – Stand 1.2.2002	89
Abb. 4.7. Klickzählung bei Alltheweb.com – Stand 1.11.2002	89
Abb. 4.8. Cluster-Suche bei Google	91
Abb. 4.9. Cluster-Suche bei Teoma	92
Abb. 5.1. Erfassung von Booleschen Suchanfragen durch Serverprotokollierung	99
Abb. 5.2. Expertensuche bei Alltheweb.com – Teil 1, Stand 1.11.2002	101
Abb. 5.3. Expertensuche bei Alltheweb.com – Teil 2, Stand 1.11.2002	103
Abb. 5.4. Suchergebnisseiste Fireball.de – verkürzte Darstellung	108
Abb. 5.5. Suchergebnisseiste Lycos.de – verkürzte Darstellung	110
Abb. 5.6. Suchergebnisseiste Altavista.de – verkürzte Darstellung	111
Abb. 5.7. Suchergebnisseiste Google.de – verkürzte Darstellung	112
Abb. 6.1. Darstellung von Framesets ohne DESCRIPTION-Tag bei Altavista	124
Abb. 6.2. Darstellung des Title-Tags im Browser	131
Abb. 6.3. Darstellung eines Titels in der Favoritenliste	132
Abb. 6.4. Lage von Text in einer HTML-Tabelle	147
Abb. 6.5. Beispiel einer Robots.txt-Datei	167
Abb. 7.1. Datumsangabe der Indexierung bei Altavista	172
Abb. 7.2. Datumsangabe der Indexierung bei Fireball – verkürzte Darstellung	173
Abb. 7.3. Google Toolbar mit PageRank Information – verkürzte Darstellung	179
Abb. 7.4. Zirkelbezug bei Link Popularity	181
Abb. 9.1. Darstellung der anfragenden Suchmaschinen mittels Server Logs	209
Abb. 9.2. Listing Enhancement bei Altavista	223

Abb. 10.1.	Listing bei Overture	243
Abb. 10.2.	Darstellung des Paved Placement von Overture bei Altavista	246
Abb. 10.3.	Espotting Listing bei Yahoo	250
Abb. 10.4.	C-P-C Preis Darstellung bei QualiGo	254
Abb. 11.1.	Monatliche Seitenabrufe – Balkendiagramm	262
Abb. 11.2.	Monatliche Seitenabrufe, Besucheranzahl und Anfragen	263
Abb. 11.3.	Anfrageverteilung über Wochentage-Balkendiagramm	263
Abb. 11.4.	Tägliche Seitenabrufe und Anfragen	264
Abb. 11.5.	Anfragen nach Webseiten	265
Abb. 11.6.	Bericht der Top 50 Suchbegriffe	266
Abb. 11.7.	Verweisende Websites	267
Abb. 11.8.	Anfragen nach Suchmaschinen	268

Tabellenverzeichnis

Tabelle 1.1.	Portal-Partner von Overture in Deutschland	10
Tabelle 1.2.	Portal-Partner von Overture in USA	10
Tabelle 1.3.	Portal-Partner von Espotting in Deutschland	10
Tabelle 3.1.	Darstellung einer Datenbanktabelle.....	40
Tabelle 6.1.	Suchmaschinen und Meta-Tags.....	138
Tabelle 6.2.	Zulässige Textdateiformate bei Suchmaschinen.....	164
Tabelle 9.1.	Nutzerzahlen von Suchportalen in Deutschland	208
Tabelle 9.2.	Nutzerzahlen von Suchportalen in der Schweiz	208
Tabelle 9.3.	Kooperationen von Suchmaschinen und Webkatalogen in Deutschland.....	213
Tabelle 9.4.	Kooperationen von Suchmaschinen und Webkatalogen in USA ..	213
Tabelle 9.5.	Restriktionen bei der URL-Anmeldung bei Suchmaschinen	216
Tabelle 9.6.	Restriktionen bei der URL-Anmeldung bei Webkatalogen.....	216
Tabelle 7.7.	Landesindexe bei Altavista.....	222
Tabelle 9.8.	PartnerSite – Partner von FAST.....	223
Tabelle 9.9.	Portal-Partner von FAST	225
Tabelle 9.10.	Portal-Partner von Inktomi	226
Tabelle 10.1.	Portal-Partner von Google für AdWords Werbung	238
Tabelle 10.2.	Portal-Partner von Overture in Deutschland	245
Tabelle 10.3.	Portal-Partner von Overture in USA.....	245
Tabelle 10.4.	Rangposition bei Overture Portal-Partnern.....	246
Tabelle 10.5.	Partner von Espotting in Deutschland.....	248
Tabelle 10.6.	Mindestkosten für die Aufnahme bei QualiGo	252
Tabelle 10.7.	QualiGo Partner in der Schweiz.....	253
Tabelle 10.8.	QualiGo Partner in Österreich	253
Tabelle 10.9.	QualiGo Partner in Deutschland	253