On-Site Maßnahmen für die Sichtbarkeit von Websites in Suchmaschinen

erstellt am
Fachhochschul-Studiengang
E-Business
FH OÖ, Standort Steyr

Bachelor-Arbeit I
zur Erlangung des akademischen Grades
Bachelor of Arts in Business (BA)
für wirtschaftswissenschaftliche Berufe

Eingereicht von
Michael Schachner
E-Mail: webmaster@onlinemarketing-blog.at
WWW: http://www.onlinemarketing-blog.at

Eingereicht bei: Mag. (FH) Andreas Greiner

Steyr, am 31.01.2011

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis / Glossar

AJAX	Asynchronous JavaScript and XML
CGI	Common Gateway Interface
CSS	Cascading Style Sheets
DocID	Document Identifier
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer Protocol
RIA	Rich Internet Application
SEO	Search Engine Optimization
SERP	Search Engine Results Page
URL	Uniform Resource Locator
WWW	World Wide Web
XML	Extensible Markup Language

Kurzfassung

Das World Wide Web besteht aus Webseiten, die über Hyperlinks miteinander verbunden sind. Suchmaschinen bedienen sich dieser Charakteristik, um Informationen für den Benutzer auffindbar zu machen. Die Verwendung von Suchmaschinen repräsentiert mittlerweile die zweitwichtigste Online-Aktivität unter allen Altersgruppen. Google hält einen weltweiten Marktanteil von 67 % (11/2010) und ist somit der führende Suchanbieter.

Das Bestreben von Websitebetreibern ist, dass ihre wichtigsten Webseiten in den Datenbanken von Suchmaschinen aufgenommen und diese regelmäßig besucht werden, damit die gespeicherten Informationen aktuell bleiben. Werden Webseiten von den Ergebnislisten ausgeschlossen, sind diese in Folge für Benutzer nicht existent. In der Praxis ergeben sich eine Reihe an technischer sowie struktureller Hürden, die dazu führen, dass Webseiten von Suchmaschinen nicht erreicht bzw. deren Inhalte nicht korrekt interpretiert werden können.

Die Arbeit untersucht die Fragestellung, wie die Sichtbarkeit von Websites durch On-Site Maßnahmen effektiv gestaltet werden kann. Am Beginn werden Grundlagen der Suchmaschinenindexierung vermittelt, welche Prozesse der Dokumentengewinnung, Analyse der Webseiteninhalte sowie des Aufbaus einer durchsuchbaren Datenstruktur darstellen. Danach werden Voraussetzungen im Bereich Webserver zusammen mit verschiedenen Formen des Web Hosting analysiert. In weiterer Folge werden mit technologischen Kritierien und strukturellen Entscheidungen der Informationsarchitektur die zentralen Themen hinsichtlich der Sichtbarkeit in Suchmaschinen besprochen. Sitemaps finden als ergänzende Maßnahmen für eine effektive Indexierung Beachtung.

Diese Arbeit identifiziert in der Literatur potentielle Indexierungsprobleme beim Einsatz verschiedener Technologien sowie bei der Gestaltung der Informationsarchitektur einer Website. Davon ausgehend werden Empfehlungen für effektive On-Site Maßnahmen abgeleitet. Eine hohe Verfügbarkeit des Webservers stellt die Grundvoraussetzung dar, damit Webseiten erfasst und in der Datenbank von Suchmaschinen aktuell gehalten werden. Bei der Verwendung neuer Web-Technologien schaffen Pfade über HTML Links und alternative Inhaltsbereitstellung in Textform Abhilfe. Die Hierarchie einer Website organisiert Webseiten in Kategorien und vermittelt deren Relevanz für die Suchmaschinenindexierung. Die Navigation stellt die interne Linkstruktur einer Website dar und ermöglicht die effektive Exploration von Inhalten. Abschließend bieten Sitemaps einen weiteren Zugang zu sämtlichen Webseiten, der von technischen und strukturellen Hürden unabhängig ist.

Executive Summary

The World Wide Web consists of web pages that are connected via hyperlinks. Search engines rely on this characteristic to make information accessible to users. Nowadays the utilization of search engines represents the second most important online activity among all age groups. Google holds a global market share of 67 % (11/2010) and is therefore the leading search provider.

The ambition of website owners is that their most important web pages will be included in the databases of search engines and that they are continuously revisited in order to keep the stored data current. If web pages are excluded from search results, they are non-existent to users. In practice there are a number of technical and structural barriers that lead to web pages being inaccessible for search engines or their content being incorrectly interpreted.

The thesis investigates the question of how the visibility of websites can be effectively managed by on-site activities. At the beginning fundamentals in the field of search engine indexation are provided that cover processes of document retrieval, analysis of the web content, and the creation of a searchable data structure. Thereafter critical web server requirements are analyzed along with different types of web hosting. Subsequently, technological criteria and structural decisions of the information architecture are discussed as the main issues of search engine visibility. Sitemaps are considered as additional measures for effective indexation.

This thesis identifies potential indexing problems in the literature that arise when using various technologies and designing the information architecture of a website. On that basis, recommendations for effective on-site activities are determined. A high availability of the web server represents the basic requirement in order to have web pages fetched and constantly updated in the database of search engines. When using the latest web technologies, paths in the form of HTML links and alternative content as simple text can be an appropriate workaround. The hierarchical structure of a website organizes web pages in categories and provides an indication of relevance for search engine indexation. The navigation represents the internal linking structure of a website and allows the effective exploration of content. Finally sitemaps provide access to all web pages without being dependent on technical and structural constraints.

1 Einleitung

Suchmaschinen kommt bei der Auffindung von Informationen im World Wide Web eine tragende Rolle zu. Der Aufwärtstrend im Suchmarkt hält mit einem weltweiten Jahreswachstum bei den Suchanfragen von 46 % auf mehr als 131 Milliarden weiterhin an. Der Anbieter Google nimmt als Spitzenreiter einen Marktanteil von knapp 67% ein.[1] Nach E-Mail stellt die Online-Suche die zweithäufigste Aktivität im Internet dar, wobei jede dritte Suche einen kommerziellen Hintergrund hat.[2] So recherchieren Benutzer Informationen hinsichtlich Produkte, Preise sowie Hersteller und weisen durch den aktiven Prozess der Informationsbeschaffung eine hohe Kaufbereitschaft auf.

Unternehmen haben das enorme Potential von Suchmaschinen für ihre Vertriebs- und Marketingaktivitäten für sich erkannt, weshalb die Ausgaben für Suchmaschinenmarketing einen immer größeren Anteil am Gesamtwerbebudget einnehmen. Von einem langfristigen Standpunkt betrachtet hat hierbei die Suchmaschinenoptimierung einen hohen Stellenwert, um die eigene Website für die wichtigsten Suchanfragen in den organischen Ergebnissen auf den vorderen Rängen zu positionieren.

1.1 Problemstellung

Unternehmen werden sich bewusst, dass eine hohe Sichtbarkeit in den Suchmaschinenergebnisseiten einen wesentlichen Erfolgsfaktor für die Gestaltung ihrer Online-Aktivitäten darstellt. Betreiber von Websites, die mit ihrem Angebot nicht auf den vorderen Positionen gelistet werden, sind für Benutzer in der Regel nicht existent.

Gründe dafür, dass Websites nicht oder nur mangelhaft im Index von Suchmaschinen enthalten sind, können bis auf die Planungsphase zurückgeführt werden. Fehlentscheidungen in den Bereichen Web Hosting, Web-Technologien sowie Informationsarchitektur sind nachträglich mit großem Änderungsaufwand verbunden. Für eine zielgerichtete Gestaltung der Optimierungsaktivitäten sind Kenntnisse über crawlerbasierte Suchmaschinen im Zusammenhang mit deren Umgang mit Web-Technologien und Website-Strukturen ein entscheidender Erfolgsfaktor. Auf Basis dieses Wissens kann sichergestellt werden, dass Webseiten von Suchmaschinen gefunden und für die spätere Listung in den Ergebnisseiten erfasst werden.

---

[1] Vgl. URL: http://www.comscore.com/Press_Events/Press_Releases/2010/1/Global_Search_Market_Grows_46_Percent_in_2009 [13.11.2010]

[2] Vgl. Schwarz, 2008, S. 321.

1.2 Zielsetzung

Ziel dieser Bachelor-Arbeit ist, technologische wie auch strukturelle On-Site Kriterien herauszuarbeiten, welche einen Einfluss auf die Indexierung von Websites durch crawlerbasierte Suchmaschinen haben.

Fragestellung:

Wie kann die Indexierung von Websites durch On-Site Maßnahmen effektiv gestaltet werden?

Folgende Unterfragen werden zur Zielerreichung beantwortet:

- Welche Rahmenbedingungen sind im Bereich Web Hosting von Relevanz?
- Welchen Einfluss haben technologische Kriterien auf die Sichtbarkeit von Websites in Suchmaschinen und was ist beim Einsatz verschiedener Web-Technologien zu beachten?
- Nach welchen strukturellen Gesichtspunkten sollte das Design der Informationsarchitektur erfolgen?
- Welche weiteren Maßnahmen tragen dazu bei die Indexierung von Website-Inhalten zu optimieren?

Die Beantwortung der Fragen erfolgt über die Diskussion einschlägiger Fachliteratur. Aufgrund des hohen Marktanteils von Google im deutschsprachigen Raum werden ergänzend aktuelle Stellungnahmen des Suchmaschinenanbieters zu den behandelten Themenbereichen zitiert.

1.3 Aufbau und Struktur

Das Kapitel 2 gibt einen Einblick in die Grundlagen der Suchmaschinenindexierung, welche das Auffinden von Webseiten durch Crawler sowie die anschließende Aufnahme in den Index betreffen.

Das Kapitel 3 definiert die Rahmenbedingungen einer Website im Bereich Web Hosting, indem die wesentlichen Erfolgsfaktoren hinsichtlich der Indexierung gemeinsam mit verschiedenen Hosting-Typen erläutert werden.

In Kapitel 4 werden potentielle technologische Kriterien mit potentiellen Problemen für die Indexierung diskutiert und entsprechende Empfehlungen für den Einsatz von Technologien abgegeben.

Strukturelle Kriterien der Informationsarchitektur von Websites und der Website-Navigation finden in Kapitel 5 Beachtung.

In Kapitel 6 werden Sitemaps als Maßnahme zur Optimierung der Indexierung durch Webcrawler vorgestellt.

Der Schlussteil gibt auf Basis der in den einzelnen Kapiteln behandelten Themen Erkenntnisse für Websitebetreiber hinsichtlich der effektiven Gestaltung von On-Site Maßnahmen ab.

2 Grundlagen der Suchmaschinenindexierung

2.1 Einleitung

Zur späteren Gestaltung von On-Site Maßnahmen ist es notwendig ein Verständnis für die Thematik der Suchmaschinenindexierung aufzubauen, insbesondere wie Dokumente im Web für die Verwendung in Suchmaschinen gefunden werden und in Folge der Index als Basis für Suchanfragen aufgebaut sowie laufend aktualisiert wird.

2.2 Einschränkungen und Herausforderungen

Rein theoretisch könnte das Web auf Basis von Web Information Retrieval durch Suchmaschinen in seiner Gesamtheit erfasst werden, da Dokumente über Links miteinander verknüpft sind. Aus wirtschaftlicher und technologischer Sicht sind dieser Anschauung in der Praxis klare Grenzen gesetzt.

Das sogenannte „Invisible Web“ bezeichnet den Teil der Dokumente, der für Benutzer über Suchmaschinen nicht auffindbar ist. Ein Grund liegt darin, dass aufgrund der großen Datenmengen und dem Anspruch an Qualität nicht alle Inhalte in den Index aufgenommen werden können. Außerdem ist es für Website-Betreiber möglich bewusst Sperren einzurichten, die den Zugriff auf Inhalte verwehren. Ein anderes Hindernis stellen technologische Faktoren dar, welche die Erfassung von Dokumenten ausschließen.[3]

Für das Erfassen von Webseiten und den Aufbau einer geeigneten Datenstruktur werden ressourcenintensive Prozesse notwendig. Aufgrund beschränkter Rechenkapazität müssen Suchmaschinen die Aufnahme von Webseiten in ihre Datensammlung nach ausgewählten Kriterien beschränken. Hinzu kommt der hohe Zeitaufwand für die Abwicklung dieser Prozesse. Neue Inhalte in bereits erfassten Dokumenten werden mit Verzögerung in den Suchergebnissen repräsentiert.[4]

Eine weitere Herausforderung beim Web Information Retrieval ist die fehlende Strukturierung in den Dokumenten. Zwar wird HTML als Standard verwendet, doch unterschiedliche Web-Formate, Sprachen, variierende Länge der Dokumente, Verknüpfungen über eine Hyperlink-Struktur und das Vorkommen von Dubletten, also mehrfaches Vorkommen der gleichen Dokumente, erhöhen die Komplexität der Datenverarbeitung.[5]

[3] Vgl. Lewandowski, 2005, S. 42ff.
[4] Vgl. Price/Sherman, 2001, S. 32f.
[5] Vgl. Lewandowski, 2005, S. 75f.

Die finale Leistung einer Suchmaschine resultiert aus einem Kompromiss zwischen drei Faktoren: Geschwindigkeit der Informationsbereitstellung, Präzision der bereitgestellten Dokumente und der Recall-Rate (Anteil relevanter, bereitgestellter Dokumente, gemessen an der Gesamtanzahl relevanter Dokumente). Eine Balance dieser Faktoren gestaltet sich schwierig, da die Anzahl an Dokumenten und Benutzern stetig zunimmt.[6]

2.3 Aufbau von Suchmaschinen

Benutzer verstehen unter dem Begriff Suchmaschine meist das User Interface auf der Website des jeweiligen Anbieters, auf der über ein Eingabefeld Anfragen gestellt werden können. Zur Auslieferung der Ergebnisseiten kommen im Hintergrund komplexe Systeme und Algorithmen zur Anwendung.[7] Um Verzögerungszeiten für die Bereitstellung von Suchergebnissen so gering wie möglich zu halten, sind eine Reihe an Aufgaben vorab zu erledigen. Vor der Datenauswertung sammeln Webcrawler-Systeme Dokumente und übernehmen die Überwachung von Existenz bzw. Aktualisierung bereits bekannter Dokumente. Ein Information-Retrieval-System erstellt davon ausgehend eine durchsuchbare Datenstruktur. Die finale Verarbeitung von Suchanfragen übernimmt ein sogenannter Query-Prozessor, welcher die Ergebnislisten für die Benutzer erzeugt.[8]

2.4 Dokumentengewinnung

Die Auffindung von Web-Dokumenten geschieht mittels eines Webcrawler-Systems. In Abbildung 1 werden die einzelnen Komponenten sowie deren Verbindungen zueinander veranschaulicht.

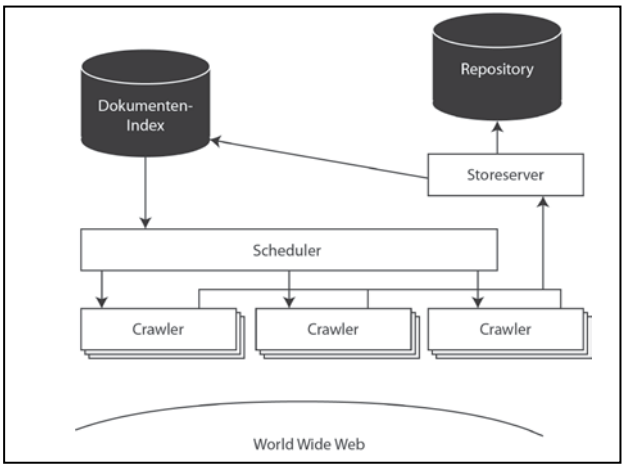

Abbildung 1: Webcrawler-System[9]

[6] Vgl. Kobayashi/Takeda, 2000, S. 149.
[7] Vgl. Ledford, 2008, S. 5.
[8] Vgl. Erlhofer, 2008, S. 31f.
[9] Vgl. ebenda S. 73.

2.4.1 Dokumenten-Index

Der Dokumenten-Index stellt eine spezielle Datenbank dar, in der Informationen zu jedem Dokument abgelegt sind. Jedes Dokument erhält einen eindeutigen Schlüssel in Form der sogenannten DocID, nach der Einträge im Index sortiert abgelegt werden. Zusätzlich enthält ein Datensatz u.a. den aktuellen Dokumentenstatus, eine Check-Summe zur Feststellung von Änderungen im Dokument sowie einen Verweis auf eine Kopie im Repository.[10]

2.4.2 Crawler

Crawler, auch Spider oder Robots genannt, sind Software Agents, die dem Auffinden und Katalogisieren von Informationen im WWW dienen.[11]

Jerkovic beschreibt den Vorgang des Crawling wie folgt:

> *„Crawling is the automated, systematic process of web document retrieval initiated by a web spider.“*[12]

Die Herausforderung des Crawling besteht darin, die Erstellung neuer, Löschung alter oder Änderung bestehender Webseiten rasch zu erkennen. So kommt es immer wieder zu Diskrepanzen zwischen tatsächlichem Inhalt auf Webseiten und den im Suchmaschinen-Index enthaltenen Informationen. Um den Index so aktuell wie möglich zu halten, priorisieren Spider Websites mit höheren Änderungsraten, bzw. qualitativ hochwertigem Content und kehren zu diesen häufiger zurück.[13]

Crawler werden vom Scheduler, der zentralen Verwaltungskomponente im Webcrawler-System, angewiesen Dokumente bestimmter URLs neu herunterzuladen oder hinsichtlich ihrer Existenz bzw. Aktualität zu untersuchen. Zur Erfassung neuer Inhalte wird zuerst die URL über DNS in eine IP-Adresse aufgelöst. Anschließend erfolgt ein HTTP Request an die IP-Adresse des Servers. Die darauf folgende HTTP Response liefert die Dokumentendaten sowie Header-Informationen, welche vom Crawler im letzten Schritt der Auftragserledigung an den Storeserver weitergeleitet werden.[14]

Um Latenzzeiten zu überbrücken, arbeitet man mehrere URLs gleichzeitig ab. Sollte eine URL nicht erreichbar sein, wird diese vorläufig ignoriert und erst zu einem späteren Zeitpunkt eine neue Anfrage gestartet.[15]

---

[10] Vgl. Erlhofer, 2008, S. 73f.
[11] Vgl. Ledford, 2008, S. 7f.
[12] Jerkovic, 2010, S. 172.
[13] Vgl. Moran/Hunt, 2009, S. 43ff.
[14] Vgl. Erlhofer, 2008, S. 76f.
[15] Vgl. Koch, 2007, S. 28.

2.4.3 Storeserver

Der Storeserver übernimmt die Sicherung der Daten, welche von den Crawlern gewonnen wurden und die Überprüfung der Datenintegrität. Der Dokumentenindex wird auf Basis der Auswertung von HTTP-Response-Headern, welche von den abgefragten Webservern zurückgegeben werden, aktualisiert. Danach kommt an den Dokumenten eine Reihe an Filtern zur Anwendung. Unbrauchbare Dokumententypen, Vorliegen von Dubletten sowie ein URL-Filter, der dynamische Dokumente und unerwünschte Wörter erkennt, beeinflussen, welche Ressourcen letztendlich im Repository abgelegt werden dürfen. Potentielle Probleme in der Verarbeitung werden somit vorzeitig eliminiert[16]

2.4.4 Repository

Das Repository legt Webseiten als lokale Kopie ab. Die einzelnen Datensätze sind über die DocID fortlaufend nummeriert und über diese auffindbar. Um Speicherplatz zu sparen, werden Daten in komprimierter Form aufbewahrt. Falls eine aktuellere Version im Vergleich zum Datenstand gefunden wurde, wird die DocID gesucht und der letzte Stand im Repository abgespeichert.[17] Eine Besonderheit ist, dass Webseiten meist einschließlich HTML-Code in diesem Datenspeicher aufbewahrt werden. Die DocID stellt dabei keine Sortierung dar, sondern gibt lediglich die Reihenfolge der Speicherung wieder.[18]

2.5 Analyse der Webseiteninhalte

Ein kritischer Faktor bei der Analyse und Verarbeitung von Webseiteninhalten stellt die mangelnde Strukturierung von Webdokumenten dar. Diese weisen zwar Parallelen zu unstrukturiertem Fließtext auf, doch enthalten Tags für die Bedeutung und das Priorisieren von Inhalten.[19]

Crawler erfassen neben reinen HTML-Seiten aber auch andere Dokumententypen, welche zur weiteren Analyse in ein Standardformat überführt werden müssen. Die Verwendung von Tags weist Suchmaschinen auf die Bedeutung der Inhalte hin. Durch Festlegung von Überschriften, Text-Formatierung und Position können gezielt Prioritäten definiert werden. Zusätzliche Informationen über eine Webseite erhält die Suchmaschine durch den Title Tag, der den Namen der Seite enthält. Sogenannte Metatags wie etwa die Description geben eine kurze Erläuterung der Seiteninhalte. Unabhängig vom Code erkennen Suchmaschinen anhand der einzelnen Wörter die Sprache, was für die spätere Auslieferung der SERPs in unterschiedlichen Ländern wesentlich ist. Überdies hinaus

[16] Vgl. Erlhofer, 2008, S. 78ff.
[17] Vgl. ebenda S. 83.
[18] Vgl. Koch, 2007, S. 29.
[19] Vgl. Lewandowski, 2005, S. 59f.

können Websites anhand des Inhalts und der Linkstruktur einem bestimmten Themengebiet zugeordnet werden.[20]

2.6 Aufbau des Suchmaschinen-Index

In der Analysephase wurden relevante Schlüsselwörter identifiziert, welche die einzelnen Webseiten beschreiben. Zur Beantwortung der Suchanfragen der Benutzer legen die Suchmaschinenanbieter eine geeignete Datenstruktur in Form des Suchindex mit Informationen zu jedem Schlüsselwort an.

Die Realisierung des Index basiert auf drei Strukturen - der Hitlist, dem direkten Index und dem indirekten Index. Die Grundlage bilden die in der Analyse gewonnenen Schlüsselwörter einer Webseite, für die eine Gewichtung in Abhängigkeit bestimmter Eigenschaften ermittelt wird. Das Resultat ist die sogenannte Hitlist, auch Location List genannt, welche Zusatzinformationen wie Position, Gesamtvorkommen und Formatierung eines Begriffs innerhalb eines Dokuments enthält. Die Daten werden in kompakter Form durch Verwendung interner Codes abgespeichert. Dazu wird der direkte Index angelegt, welcher für jedes Dokument, repräsentiert durch die DocID, die dazugehörige Hitlist für jedes Schlüsselwort ablegt.[21] Tabelle 1 zeigt ein Beispiel eines direkten Index in codierter Form. Dieser bleibt letztendlich als direkte Datei erhalten.

DocID	WordID	Hitlist
000255	004324	E34C5
000255	002343	67C33
000255	005321	DC423

Tabelle 1: Direkter Index[22]

Für Suchanfragen eignet sich die Struktur des direkten Index jedoch nicht, da der Benutzer Informationen zu bestimmten Schlüsselwörtern auf Basis einer Suchanfrage benötigt. Die Lösung ist nun die Umkehrung der direkten Datei in eine indirekte Datei mit Sortierung nach Wörtern, repräsentiert durch die WordID. Für die Beantwortung von Suchanfragen ist später sowohl der direkte als auch der indirekte Index notwendig. Zu jedem Worteintrag wird eine Liste an Dokumenten gespeichert, in denen der jeweilige Begriff enthalten ist.[23] Tabelle 2 enthält ein Beispiel für einen indirekten Index.

[20] Vgl. Moran/Hunt, 2009, S. 48ff.
[21] Vgl. Erlhofer, 2008, S. 109ff.
[22] Vgl. ebenda S. 113.
[23] Vgl. Price/Sherman, 2001, S. 29f.

WordID	DocID
004324	002345, 000234
002343	000345, 006534
005321	001325, 001234

Tabelle 2: Indirekter Index[24]

2.7 Zusammenfassung

In diesem Kapitel wurde das Wesen der Suchmaschinenindexierung in ihren Grundzügen beschrieben. Es sollte verdeutlichen, mit welchen Herausforderungen Suchmaschinen bei der Erfassung von Dokumenten konfrontiert sind und wie sie mit der hohen Diversität der Informationsstruktur umgehen.

Der erste Abschnitt erläuterte Beschränkungen, die sich durch die Eigenschaften des WWW bei der Indexierung ergeben. Es wurde gezeigt, dass aufgrund unterschiedlicher Technologien und der enormen Datenmengen eine vollständige Erfassung der Webseiten über die Hyperlinkstruktur nur in der Theorie realisierbar ist.

Im nächsten Abschnitt wurde der Aufbau von Suchmaschinen beschrieben. Diese setzen sich aus mehreren Komponenten zusammen, welche Prozesse der Dokumentengewinnung, Erstellung eines durchsuchbaren Index bis hin zur Bearbeitung von Suchanfragen abdecken.

Der Abschnitt „Dokumentengewinnung“ ging auf die Funktionsweise eines Webcrawler-Systems ein. Dieser definierte, wie Dokumente einer Website von Suchmaschinen erfasst werden und welche Vorgehensweise zur Anwendung kommt, um den Datenbestand aktuell zu halten.

Die Analyse der Webseiteninhalte schilderte, wie Informationen aus den gesammelten Dokumenten extrahiert und anhand dieser Prioritäten für bestimmte Elemente festgelegt werden. Als Ergebnis der Inhalts- und Linkstruktur wird daraus das Thema für eine gesamte Website abgebildet.

Im letzten Abschnitt wurde erklärt, wie Suchmaschinen aus den gesammelten Informationen eine effiziente Datenstruktur anlegen, die eine schnelle Auslieferung von Suchergebnissen für Benutzeranfragen möglich macht.

[24] Vgl. Erlhofer, 2008, S. 115.

3 Rahmenbedingungen Web Hosting

3.1 Einleitung

Für die Sichtbarkeit einer Website in Suchmaschinen sind optimale Rahmenbedingungen zu schaffen. Zu Beginn ist die Entscheidung für eine verlässliche Hosting-Lösung wesentlich. Diese bietet zwar nur begrenzt direkte Einflussmöglichkeiten für die Suchmaschinenindexierung, ist jedoch mit ausschlaggebend für die Aufnahme und Aktualisierung von Webseiten im Index.

Bei der Wahl des Web Hostings für eine Website müssen eine Reihe kritischer Faktoren hinsichtlich Suchmaschinen beachtet werden. Je nach Form des Web Hostings ergeben sich potentielle Risiken und Probleme. Nachfolgend werden die wichtigsten Erfolgsfaktoren sowie unterschiedliche Web Hosting Lösungen mit deren entscheidenden Charakteristika behandelt.

3.2 Erfolgsfaktoren

Im Bereich Web Hosting ergibt sich eine Reihe an Herausforderungen. Werden Entscheidungen nicht sorgfältig getroffen, kann dies im schlechtesten Fall dazu führen, dass die eigene Website nicht mehr im Index geführt wird.

3.2.1 Erreichbarkeit des Webservers

Koch hebt die Relevanz der Verfügbarkeit eines Webservers hervor, damit Webcrawler Webseiten besuchen und in den Index aufnehmen. Werden Suchmaschinen-Anfragen nicht beantwortet, besteht eine hohe Wahrscheinlichkeit, dass Dokumente unter Umständen aus dem Index gelöscht werden.[25]

In der Literatur unterstützt auch Erlhofer diese Sichtweise hinsichtlich der Indexierung. Bei Nichterreichbarkeit der gesamten Website könnte dies im schlechtesten Fall bis zur vollständigen Entfernung der einzelnen Webseiten auf Basis des Domain-Namens oder sogar der gesamten IP-Adresse führen.[26]

Nach Cutts würde eine Downtime des Webservers für wenige Tage noch nicht bewirken, dass Webseiten direkt aus dem Index fallen, da Webcrawler zu einem späteren Zeitpunkt erneut zurückkommen. Bei mehreren erfolglosen Versuchen über einen längeren Zeitraum hinweg ist eine Löschung jedoch als unvermeidbar anzusehen.[27]

---

[25] Vgl. Koch, 2007, S. 234.

[26] Vgl. Erlhofer, 2008, S. 275.

[27] Vgl. URL: http://www.youtube.com/watch?v=qXrwyTGOf1E [08.01.2011]

Die getroffenen Aussagen unterstreichen, dass eine hohe Erreichbarkeit des Webservers unabdingbar für eine Aufnahme und langfristige Aufbewahrung im Index ist.

3.2.2 Ladezeiten von Webseiten

Die Ladezeit für die Auslieferung von Webseiten an Suchmaschinenspider stellt einen Erfolgsfaktor dar, dessen Auswirkungen auf die Indexierung in der Literatur nur zu einem gewissen Maß eingegrenzt werden können.

Moran und Hunt messen niedrigen Ladezeiten von Webseiten Bedeutung bei, welche jedoch weniger ausschlaggebend als die Erreichbarkeit des Webservers an sich sei:

> *„Although your site is technically up, the pages might be displayed so slowly that the spider soon abandons the site. Few spiders will wait 10 seconds for a page*."[28]

Nach Beus haben die Webserver Performance und somit als Implikation die Ladezeiten von Webseiten einen wahrnehmbaren Einfluss auf das Crawling-Verhalten von Suchmaschinen. Bei zu hohen Antwortzeiten ist es wahrscheinlich, dass die Crawl-Frequenz abnimmt und in Folge die gesamte Domain weniger häufig besucht wird. Besonders Websites mit laufender Aktualisierung von Informationen sind davon negativ betroffen, da Webseiten nur verzögert in den Trefferlisten von Suchmaschinen aufscheinen.[29]

3.2.3 Geteilte IP-Adresse

Ein anderer Erfolgsfaktor betrifft die IP-Adresse, unter der eine Website erreichbar ist. Je nach Hosting-Typ kann diese mit anderen Website-Betreibern geteilt werden. In der Literatur wird das Thema der geteilten IP-Adressen mit einer Reihe an Schwachstellen für die Sichtbarkeit in Suchmaschinen verbunden.

Enge et al. weisen auf die Gefahr von Spam-Nachbarn bei geteilter IP-Adresse hin. Verstößt eine der Websites innerhalb der „Nachbarschaft" gegen die Richtlinien von Suchmaschinenanbietern, erfolgt mitunter der Ausschluss aller Webseiten einer bestimmten IP-Adresse.[30]

Eine weitere Gefahr identifiziert Erlhofer darin, dass manche Suchmaschinenbetreiber die Anzahl indizierter Dokumente auf Basis der IP-Adresse beschränken. Befinden sich große

---

[28] Moran/Hunt, 2009, S. 252.

[29] Vgl. URL: http://www.sistrix.com/blog/902-what-is-the-impact-of-the-webserver-speed.html [08.01.2011]

[30] Vgl. Enge et al., 2010, S. 115f.

Websites auf dem gleichen Webserver, könnte sich dies negativ auf die Repräsentation eigener Webseiten im Suchmaschinen-Index auswirken.[31]

Cutts relativiert die Abstrafungen, die in der Literatur hinsichtlich des Auftretens von Spam innerhalb geteilter IP-Adressen genannt werden. Normalerweise gibt es keine Konsequenzen für Websitebetreiber, die sich an die Richtlinien der Suchmaschinen halten, auch wenn unter der gleichen IP-Adresse Spam durch andere Anbieter entdeckt wird. Nur wenn ein starkes Ungleichgewicht an Websites mit Spam-Inhalten entsteht, könnte dies negative Folgen für andere Betreiber haben.[32]

Daraus lässt sich schließen, dass die Problematik geteilter IP-Adressen meist weniger schwer wiegt als manches Mal angenommen. Aus SEO-Perspektive sollte dennoch eine eigene IP-Adresse bevorzugt werden, da diese die vorhin erwähnten Problembereiche direkt ausschließt.

3.3 Formen von Web Hosting

Je nach Wahl des Web Hostings für eine Website können die im vorherigen Abschnitt erläuterten Erfolgsfaktoren hinsichtlich SEO unterschiedlich erfüllt werden. Nachfolgend werden die wichtigsten Typen in ihren Grundzügen besprochen.

3.3.1 Free Hosting

Free Hosting ist eine kostenlose Form des Hosting, welche mit weitreichenden Limitierungen verbunden ist. Diese betreffen etwa begrenzten Speicherplatz, beschränkte Bandbreite und mangelhafte Unterstützung von Datenbanken. Meist befinden sich unzählige Websites auf einem physikalischen Server und werden über eine einzige IP-Adresse ausgeliefert. Wird die IP mit Spam in Verbindung gebracht, könnte dies negative Auswirkungen auf die Sichtbarkeit eigener Webseiten in Suchmaschinen haben.[33]

Erlhofer schätzt Free Hosting ebenfalls als unzureichend für die effektive Indexierung von Webseiten ein, da Websites meistens in einem Unterverzeichnis abgelegt werden und unter keiner eigenen Domain erreichbar sind. Diese Adressierung ist für Suchmaschinen nicht geeignet.[34]

Free Hosting ist somit aus Gründen der technischen Limitierung nicht empfehlenswert. Auch Spam innerhalb einer IP-Adresse darf bei Free Hosting nicht außer Acht gelassen

[31] Vgl. Erlhofer, 2008, S. 276.
[32] Vgl. URL: http://www.youtube.com/watch?v=AsSwqo16C8s [08.01.2011]
[33] Vgl. Jerkovic, 2010, S. 47.
[34] Vgl. Erlhofer, 2008, S. 274.

werden, da die Qualität von Websites aufgrund des kostenlosen Services als eher niedrig einzustufen gilt.

3.3.2 Shared Hosting

Beim Shared Hosting werden ähnlich zu Free Hosting mehrere Websites auf einem Webserver abgelegt. Diese Angebote bieten meist mehr Speicherplatz, höhere Bandbreite und Unterstützung von Datenbanken. Für SEO sollte laut Jerkovic sichergestellt werden, dass man für seine Website eine eigene IP erhält. Es bestehe jedoch noch immer die Gefahr, dass sich die Belastung des Webservers durch den parallelen Betrieb mehrerer Websites negativ auf die Webserver Performance auswirkt.[35]

Cutts betont hingegen, dass eine einzigartige IP für eine Website in einer Shared Hosting Umgebung nicht unbedingt notwendig ist. Suchmaschinen wie Google würden mit virtuellen Hosts im Vergleich zu einzigartigen IP-Adressen gleichermaßen umgehen. Es gibt jedoch immer wieder Internet Service Provider, welche die Konfiguration des virtuellen Hostings inkorrekt vornehmen.[36]

Das Hauptargument gegen Shared Hosting besteht darin, dass die Webserver Performance durch die Bereitstellung mehrer Websites auf einem Webserver beeinträchtigt sein kann. Dies könnte negative Implikationen auf die Erreichbarkeit der gesamten Website bzw. die schnelle Auslieferung von einzelnen Webseiten an Suchmaschinen-Spider bedeuten.

3.3.3 Dedicated Server Hosting

Eine Website lässt sich im Idealfall auf einem eigenen Webserver ablegen, welcher naturgemäß über eine separate IP-Adresse erreichbar ist. Damit können die Voraussetzungen für die Erfüllung der im ersten Abschnitt besprochenen Erfolgsfaktoren bestmöglich geschaffen werden.

Erlhofer gibt die Empfehlung ab einen eigenen Server anzumieten, bei dem man als Betreiber uneingeschränkte Administratorrechte besitzt und somit einen größeren Handlungsspielraum für entsprechende Maßnahmen eingeräumt bekommt. Hierbei werden jedoch Kenntnisse in der Wartung und Pflege eines Webservers notwendig.[37]

Auch Beus sieht klare Vorteile für die Bereitstellung einer Website über einen eigenen Server. In der Regel können damit Performance-Probleme des Shared Hosting vermieden

[35] Vgl. Jerkovic, 2010, S. 47f.
[36] Vgl. URL: http://www.mattcutts.com/blog/myth-busting-virtual-hosts-vs-dedicated-ip-addresses/ [08.01.2011]
[37] Vgl. Erlhofer, 2008, S. 274.

werden und die Auslieferung einzelner Webseiten an Suchmaschinen geht schneller vonstatten.[38]

3.4 Zusammenfassung

Dieses Kapitel legte mit Web Hosting den Grundstein für eine suchmaschinenfreundliche Website. Es wurde gezeigt, welche Problemfaktoren auf Webserver-Ebene die Indexierung beeinträchtigen können. Einhergehend folgte die Konkretisierung verschiedener Hosting-Typen in Bezug auf ihre Eignung für die Suchmaschinenoptimierung.

Die Erfolgsfaktoren im Bereich Web Hosting befassten sich mit der Erreichbarkeit des Webservers, den Ladezeiten von Webseiten und dem Einfluss einer geteilten IP-Adresse. Die grundlegende Aussage ist, dass eine hohe Verfügbarkeit und Performance des Webservers gewährleistet sein muss, damit Anfragen von Webcrawlern in ausreichender Zeit beantwortet werden. Geteilte IP-Adressen spielen eine eher untergeordnete Rolle für die Indexierung und bringen nur in Einzelfällen bedeutsame Nachteile mit sich.

Beim Angebot an Hosting-Typen hat sich herausgestellt, dass eine Website für SEO idealerweise auf einem eigenen Webserver abgelegt wird, um Probleme von geteilten IP-Adressen und Performance-Engpässen vorweg zu vermeiden. Shared Hosting stellt, eine ordnungsgemäße Konfiguration und ein hoher Service Level durch den Hosting-Dienstleister vorausgesetzt, eine brauchbare Alternative dar. Von Free Hosting ist auf jeden Fall abzuraten.

[38] Vgl. URL: http://www.sistrix.com/blog/721-hosting-with-own-ip-address-necessary-for-seo.html [08.01.2011]

4 Technologische Kriterien

4.1 Einleitung

Technologischen Aspekten kommt im Hinblick auf das Crawling und die spätere Indexierung von Webseiten durch Suchmaschinen eine tragende Rolle zu. Web Spider sind aufgrund einer Reihe von Faktoren nicht in der Lage, alle Inhalte gleichermaßen syntaktisch und semantisch zu erfassen. Bereits in der Design-Phase sollten die grundsätzlichen technologischen Kriterien und ihre Relevanz für das Thema SEO berücksichtigt werden. Dieses Kapitel erläutert die wichtigsten Merkmale der einzelnen Technologien und gibt Empfehlungen für ihren Einsatz auf Websites ab.

4.2 Dynamische URLs

4.2.1 Beschreibung

Dynamische URLs werden in Verbindung mit Webseiten verwendet, deren HTML Code im Gegensatz zu statischen Webseiten nicht permanent als eigenständige Datei in einem Verzeichnis abgelegt wird. Die Erstellung erfolgt, sobald eine Anfrage per HTTP Request an den Server gestellt wird. Über CGI (Common Gateway Interface) ruft dann der Webserver eine Applikation zur dynamischen Erzeugung des HTML Codes auf und gibt diesen in der darauf folgenden HTTP Response zurück. Die dafür notwendigen Daten werden dazu aus einer Datenbank ausgelesen[39] Abbildung 2 enthält ein Szenario für die Generierung einer dynamischen Webseite. URLs dieser Webseiten enthalten meist mehrere Parameter und sind an der Verwendung von Hash-Werten und anderen kryptischen Zeichenfolgen zu erkennen.[40]

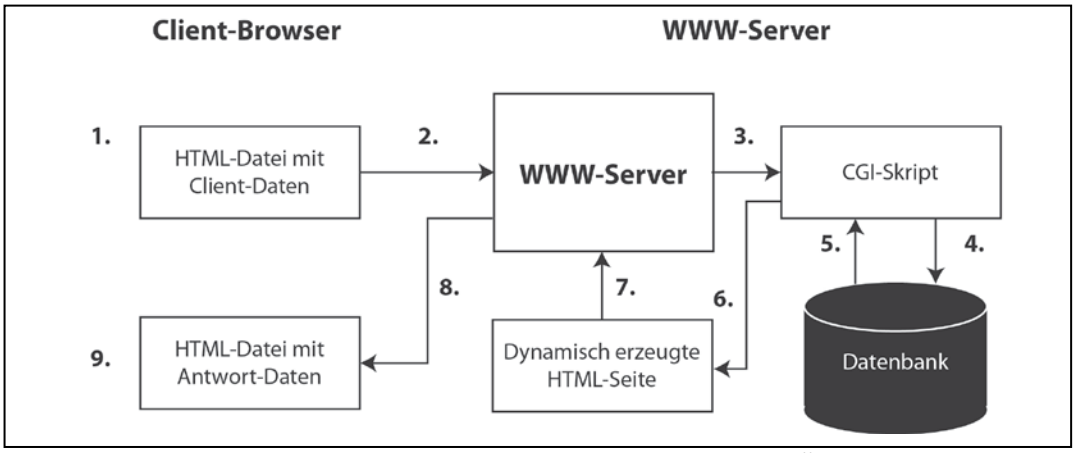

Abbildung 2: Generierung dynamischer Webseiten[41]

[39] Vgl. Moran/Hunt, 2009, S. 244f.
[40] Vgl. Schwarz, 2008, S. 356.
[41] Vgl. Eckert, 2009, S. 140.

Beispiel für eine dynamische URL:
http://www.example.com/index.php?cat=234

Neben Content-bezogenen Parametern erwähnen Moran und Hunt dynamische URLs in Bezug auf deren Einsatz mit Session IDs. Diese ermöglichen die eindeutige Identifikation von Sitzungen eines Benutzers verbunden mit dem aktuellen Anwendungszustand, um das Verhalten statistisch aufzuzeichnen.[42]

4.2.2 Problembereiche

Ein kritisches Hindernis für die Indexierung wird in der Literatur mit Session IDs in Verbindung gebracht, die URL-Varianten für einzelne Besucher erzeugen anstatt sich auf unterschiedliche Inhalte zu beziehen.

Parameter der dynamischen URLs können rein theoretisch unendlich viele Werte annehmen. Betreffen Parameter nicht den Content, sondern sind diese auf Besucher bezogen, würden Suchmaschinen Spider immer die gleiche Webseite unter von einander abweichenden URLs vorfinden. In der Folge werden unter Umständen sämtliche dieser Varianten in Form von Duplikaten in den Index aufgenommen.[43]

Des Weiteren kann der Einsatz von Session IDs je nach Umsetzung dazu führen, dass bei der erneuten Abfrage einer auf diese Weise generierten URL der eigentliche Inhalt der Webseite nicht mehr gefunden wird und eine Fehlermeldung ausgegeben wird.[44]

4.2.3 Einsatzempfehlung

Heutzutage haben dynamische Websites eine hohe Verbreitung erreicht zB für Blogs, Content Management Systeme, Social Network Services oder Online-Shops im E-Commerce, weshalb die entsprechenden URLs in der Regel von Suchmaschinen korrekt erfasst werden können.

Für die problemlose Aufnahme von dynamischen Webseiten in den Suchindex sollten URLs nach Moran und Hunt mehrere Kriterien erfüllen. Suchmaschinen würden auf die Anzahl von Parameter achten. Mehr als zwei Parameter sind somit nicht empfehlenswert. Ferner soll die Zeichenanzahl der URL 1000 nicht überschreiten. Session IDs müssen für die korrekte Erfassung unbedingt aus der URL entfernt werden[45]:

---

[42] Vgl. Jerkovic, 2010, S. 167.
[43] Vgl. Moran/Hunt, 2009, S. 246.
[44] Vgl. Erlhofer, 2008, S. 243.
[45] Vgl. Moran/Hunt, 2009, S. 246f.

Dynamische URLs enthalten für Suchmaschinen jedoch keinerlei semantische Bedeutung. Deshalb empfiehlt es sich diese durch „URL Rewriting“ in kürzere statische URLs zu transformieren, die relevante Schlüsselwörter anstatt der kryptischen Zeichenfolgen enthalten.[46] Auch Erlhofer unterstützt diesen Ansatz, da dynamische Seiten mit einer hohen Anzahl an Parametern oft in der Indexierung zur Wahrung der Datenintegrität nicht berücksichtigt werden.[47]

Google nimmt zum korrekten Umgang mit dynamischen URLs Stellung. Grundsätzlich gibt es keine Beschränkung für die Anzahl der Parameter, die verarbeitet werden. Das Ziel sollte aber sein kurze URLs zu seinen Webseiten zur Verfügung zu stellen. Entscheidet man sich für den Einsatz von URL Rewriting dürfen keine wesentlichen Parameter versteckt werden, die für Suchmaschinen wichtige Strukturinformationen enthalten. Im Zweifelsfall empfiehlt es sich deshalb eher die dynamischen URLs in ihrer Ursprungsform beizubehalten.[48]

4.3 Redirects

4.3.1 Beschreibung

Weiterleitungen, sogenannte Redirects, werden meist im Falle von Umstrukturierungen innerhalb einer Website verwendet. Werden Webseiten an einen neuen Ort verschoben, ändern sich meist deren URLs.

Jones betont die Wichtigkeit Suchmaschinen durch Weiterleitungen über den neuen Aufenthaltsort von Webseiten zu informieren:

> *„Do not expect the search engines to understand your intentions. You may consider moving a page to be a logical and intelligent decision, but the search engines initially see nothing but a missing page.”*[49]

Suchmaschinen haben im Rahmen der Indexierung die alte Version gespeichert, welche nach dem Umzug für den Spider lediglich als fehlende Webseite zu interpretieren ist. Über eine Weiterleitung muss nun der Suchmaschinen-Spider angewiesen werden, den angefragten Inhalt unter einer neuen URL abzuholen.[50]

---

[46] Vgl. King, 2008, S. 28f.
[47] Vgl. Erlhofer, 2008, S. 244.
[48] Vgl. URL: http://googlewebmastercentral.blogspot.com/2008/09/dynamic-urls-vs-static-urls.html [08.01.2011]
[49] Jones, 2008, S. 85.
[50] Vgl. Enge et al., 2010, S. 254.

4.3.2 Problembereiche

In der Literatur wird als Hauptproblem von Weiterleitungen genannt, dass es mehrere Varianten der Umsetzung gibt, welche jedoch nicht gleichermaßen von Suchmaschinen verstanden werden.

Moran und Hunt weisen darauf hin, dass Webmaster oft fälschlicherweise auf clientseitige Weiterleitungen über JavaScript oder Meta Tags zurückgreifen. Bei beiden Varianten indiziert die Suchmaschine die alte URL, während der Benutzer zur neuen Version gelangt. Aufgrund ihrer Funktionsweise wurde diese Art von Weiterleitungen von Spammern zur Täuschung von Suchmaschinen verwendet.[51] Beim JavaScript Redirect erfolgt die Weiterleitung, sobald die Webseite vom Client heruntergeladen und der Script-Code ausgeführt wurde. Der Meta Refresh Redirect hingegen wird über die Attribute http-equiv und content definiert. Das Content-Attribut enthält die URL sowie eine Verzögerungszeit, nach der weitergeleitet werden soll.[52]

Die offiziellen Google Richtlinien sprechen ebenfalls klar gegen den Einsatz von JavaScript für die Umsetzung von Redirects:

> *„Along those lines, it violates the Webmaster Guidelines to embed a link in JavaScript that redirects the user to a different page with the intent to show the user a different page than the search engine sees. When a redirect link is embedded in JavaScript, the search engine indexes the original page rather than following the link, whereas users are taken to the redirect target."*[53]

Um JavaScript für Weiterleitungen einsetzen zu können, muss laut Erlhofer der Inhalt der ursprünglichen Seite mit der Zielseite übereinstimmen. Änderungen werden dann aber jeweils auf der neuen Hauptseite sowie auf der weiterleitenden Webseite notwendig. Deshalb sind JavaScript-Weiterleitungen nicht praxisgerecht.[54]

4.3.3 Einsatzempfehlung

Generell sollten Weiterleitungen mit server-seitigen HTTP Status Codes umgesetzt werden. Wenn eine Anfrage für eine Webseite von einem Benutzer oder Webcrawler an einen Server gesendet wird, gibt der Server den entsprechenden HTTP Status Code zurück. Dieser informiert darauf über den Status der Anfrage.[55]

---

[51] Vgl. Moran/Hunt, 2009, S. 251.

[52] Vgl. Chellapilla/Maykov, 2007, S. 81f.

[53] URL: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=66355

[54] Vgl. Erlhofer, 2008, S. 287.

[55] Vgl. URL: http://www.google.com/support/webmasters/bin/answer.py?answer=40132&cbid=-1pe1wokkq7axt&src=cb&lev=%20answer [08.01.2011]

Enge et al. gehen näher darauf ein, welche HTTP Status Codes für Weiterleitungen gedacht sind und wie diese von Suchmaschinen interpretiert werden. Mittels Status Code 301 wird Suchmaschinen mitgeteilt, dass die Ressource permanent unter der neuen URL erreichbar ist. Diese erhält von Suchmaschinen den Linkwert der alten URL gutgeschrieben. Eine Alternative stellt der 302 Redirect dar, welcher für temporäre URL-Änderungen verwendet werden kann. Eine Übertragung des Linkwerts findet nicht statt.[56]

Bei einer großen Anzahl an verschobenen Webseiten ist zu beachten, nicht alle auf die Homepage, sondern auf die relevantesten Inhaltsseiten einer Website weiterzuleiten. Andernfalls könnte dies von Suchmaschinenbetreibern als Manipulationsversuch gewertet werden, um die Startseite zu stärken.[57]

4.4 Robots.txt

4.4.1 Beschreibung

Die robots.txt stellt eine simple Textdatei zur Steuerung des Zugriffs von Suchmaschinen Robots und Spidern auf eine bestimmte Domain dar. Diese Datei muss dabei im Stammverzeichnis, also der obersten Ebene, abgelegt werden. Beim Besuch eines Robots sucht er zuerst nach der robots.txt und folgt den enthaltenen Anweisungen. Da es sich beim Robots Exclusion Protocol um ein hinweisendes Protokoll handelt, besteht für Robots dadurch keine technische Hürde, welche den Zugriff auf bestimmte Dateien oder Verzeichnisse einschränken würde.[58]

Für Robot-Anweisungen unterscheidet man zwei grundlegende Statements:

- User-agent
- Disallow

User-agent legt fest, für welche Suchmaschinen Spider das darauf folgende Disallow-Statement gilt. Ein Asterisk als Repräsentant des User-agent bezieht sich auf alle Spider. Es können stattdessen auch Regeln für spezifische Spider angegeben werden. Mittels *Disallow* teilt man dem Spider mit, welche Inhalte nicht gecrawlt werden dürfen. Dies kann Dateinamen oder Verzeichnisse betreffen.[59]

[56] Vgl. Enge et al., 2010, S. 255.
[57] Vgl. URL: http://www.seomoz.org/blog/301-redirect-or-relcanonical-which-one-should-you-use [08.01.2011]
[58] Vgl. Koch, 2007, S. 115.
[59] Vgl. Moran/Hunt, 2009, S. 241.

Die nachfolgende Tabelle enthält eine Webcrawler-Übersicht der größten Suchmaschinenbetreiber, welche zur Angabe eines speziellen User-agent notwendig sind:

Google	
Googlebot	Webseiten
Googlebot-Mobile	Seiten für mobile Geräte
Googlebot-Image	Bilder für Bildersuche
Yahoo!	
Slurp	Webseiten
Yahoo-MMAudVid	Video-Dateien
Yahoo-MMCrawler	Bilder
Bing	
MSNBot	Webseiten
MSNBot-Media	Media Dateien

Tabelle 3: Übersicht Webcrawler[60]

Die robots.txt in Abbildung 3 gibt ein Beispiel für den formalen Aufbau der Datei, welche für jeden User-agent eine eigene Zeile vorsieht. Die erste Anweisung für den mit Asterisk ausgewiesenen User-agent schließt für alle Suchmaschinen das Crawling der Webseite „webseite1.html“ sowie des Verzeichnisses „/verzeichnis1/“ aus. Darunter folgt eine exklusive Anweisung für den User-agent „Googlebot“, dem Webcrawler von Google, das Verzeichnis „/verzeichnis2/“ zu ignorieren.

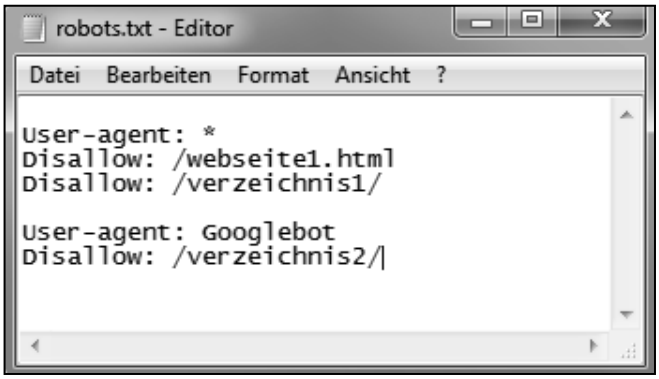

Abbildung 3: Beispiel robots.txt

Ergänzend zu robots.txt nennen Moran und Hunt Robots-Anweisungen, die auf Webseitenebene in Form von Metatags im Head-Bereich des HTML Codes festlegbar sind. Diese regeln, ob die Webseite indiziert wird und Links verfolgt werden sollen. Erfolgt ein Ausschluss bereits in der robots.txt, finden Robots Metatags keine Berücksichtigung mehr.[61]

[60] Vgl. Jerkovic, 2010, S. 178.
[61] Vgl. Moran/Hunt, 2009, S. 242.

4.4.2 Problembereiche

Probleme ergeben sich beim Einsatz von robots.txt dadurch, dass Anweisungen von Website-Betreibern nicht korrekt verstanden werden. Fehlendes Wissen führt im Extremfall dazu, dass Suchmaschinen komplett ausgesperrt werden.

Robots.txt Ausschlüsse werden häufig für Websites durchgeführt, die sich in Entwicklung befinden und nicht in den Suchergebnisseiten gelistet werden sollen. Wird nach der Fertigstellung irrtümlicherweise die entsprechende Robots-Anweisung nicht geändert, bleiben Suchmaschinen-Spider weiterhin ausgesperrt und die Website wird in Folge nicht gefunden.[62]

Manche Webmaster verbieten Webcrawlern von Suchmaschinen den Zugriff in weiten Teilen einer Website, um Server-Bandbreite zu sparen. Dies hat in der Regel negative Folgen auf die Anzahl indexierter Webseiten.[63]

4.4.3 Einsatzempfehlung

Selbst wenn keine Ausschlüsse per robots.txt in Betracht gezogen werden, ist eine Erstellung der Datei im Root-Verzeichnis empfehlenswert, welche den Zugriff auf alle Dateien und Verzeichnisse zulässt. Andernfalls werden in den Log-Dateien des Webservers zusätzliche Einträge angelegt, wenn die robots.txt nicht auffindbar ist.[64]

Robots-Anweisungen können im Umgang mit Duplicate Content auf einer Website helfen. Sind gleiche Inhalte über verschiedene Linkvarianten aufrufbar, wird der Wert eingehender Links aufgeteilt. Per robots.txt wird festgelegt, welche URL indiziert werden soll. Hinsichtlich der beschränkten Bandbreite eines Webservers ist es sinnvoll für eine Erhaltung der Performance Multimedia-Dateien und CPU-lastige Webseiten von der Indexierung auszuschließen.[65]

Jones geht näher auf Bilddateien ein, die häufig von Spezialsuchmaschinen erfasst werden. Bei einem hohen Datenumfang wirkt sich dies negativ auf die Server-Bandbreite aus. Eine Verweigerung des Webcrawler-Zugriffs auf Verzeichnisse mit einer großen Sammlung an Bildern ist deshalb ratsam.[66]

---

[62] Vgl. Moran/Hunt, 2009, S. 241.
[63] Vgl. Jerkovic, 2010, S. 176.
[64] Vgl. Koch, 2007, S. 117.
[65] Vgl. Jerkovic, 2010, S. 176.
[66] Vgl. Jones, 2008, S. 72.

Daraus folgt, dass robots.txt-Dateien für eine effektive Indexierung einer Website mit Bedacht eingesetzt werden sollten und Ausschlüsse bewusst danach festgelegt werden, welche Inhalte für Suchmaschinenergebnisseiten obsolet sind.

4.5 Web-Technologien

4.5.1 JavaScript

4.5.1.1 Beschreibung

JavaScript ist eine Skript-Sprache, die in Webseiten eingebettet werden kann. Zur Ausführung des Codes wird ein JavaScript-fähiger Browser auf der Client-Seite vorausgesetzt. Oftmals bringt man JavaScript mit dem Einsatz für Spam-Methoden in Verbindung, bei denen Suchmaschinen gezielt getäuscht werden sollen.[67]

Moran und Hunt schreiben die Popularität von JavaScript der im Rahmen von Web 2.0 aufgekommenen Interaktionsmöglichkeiten für den Benutzer zu:

> *„JavaScript is a very useful programming language that allows your Web pages to be more interactive, responding to the visitor's cursor, for example, and JavaScript also allow your Web pages to use cookies."* [68]

4.5.1.2 Problembereiche

Suchmaschinen haben JavaScript früher oft nur in simpler Form interpretiert und konnten in den Code eingebettete Links meist nicht folgen. Verschiedene Quellen aus der Literatur enthalten noch immer Hinweise, die vor einem Einsatz von JavaScript für wichtige Elemente einer Website wie der Navigation warnen, damit Inhalte bestmöglich von Suchmaschinen indiziert werden.

Laut Jerkovic stelle die Verwendung von JavaScript innerhalb einer dynamischen Navigation ein Hindernis dar, da Webcrawler enthaltene Links zu weiterführenden Webseiten nicht verfolgen und somit nicht indexieren können.[69]

In der Praxis wurde dazu festgestellt, dass bekannte Suchmaschinen wie Google und Bing Links in JavaScript-basierten Drop-Down Menüs mittlerweile indizieren können. Suchmaschinen-Spider von Yahoo! hingegen haben jedoch Probleme bei der Erfassung gezeigt.[70]

Google selbst weist daraufhin, dass eine JavaScript Navigation nicht immer korrekt verstanden wird. Hinsichtlich der Indexierung werden einfache HTML Strukturen weit besser

---

[67] Vgl. Erlhofer, 2008, S. 475.
[68] Moran/Hunt, 2009, S. 249.
[69] Vgl. Jerkovic, 2010, S. 151f.
[70] Vgl. URL: http://searchengineland.com/an-update-on-javascript-menus-and-seo-16060 [08.01.2011]

verstanden. Somit sind reine HTML Links zu bevorzugen, um die Auffindbarkeit von Webseiten zu gewährleisten.[71]

4.5.1.3 Einsatzempfehlung

Es zeigt sich, dass Suchmaschinen heutzutage mit JavaScript umgehen können. Für eine optimale Indexierbarkeit von Webseiten sollten für kritische Elemente wie der Navigation zumindest alternative Wege für Webcrawler bereitgestellt werden.

Moran und Hunt empfehlen neben in JavaScript eingebettete Links alternativen HTML Code innerhalb eines Noscript-Bereichs für Suchmaschinen einzusetzen. Als Ergänzung sollten einzelne Webseiten in einer Sitemap eingetragen werden, um optimale Voraussetzungen für ihre Indexierung zu schaffen.[72]

[71] Vgl. URL: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=81766 [08.01.2011]
[72] Vgl. Moran/Hunt, 2009, S. 243f.

4.5.2 AJAX

4.5.2.1 Beschreibung

AJAX (Asynchronous JavaScript and XML) ist eine im Zuge von Web 2.0 bekannt gewordene Technologie, welche Benutzern im Vergleich zu herkömmlichen Webseiten ein höheres Maß an Interaktivität bietet.[73]

Jerkovic beschreibt die Technologie als eine Kombination von JavaScript und XML, um bestimmte Bereiche innerhalb einer Webseite zu ändern, indem nur einzelne Fragmente anstatt der gesamten Webseite geladen werden zB direktes Einfügen von Kommentaren oder Ergänzungen von Sucheingaben.[74]

Die Web-Kommunikation mittels Einsatz von AJAX ist durch Teil-Aktualisierungen einer Webseite gekennzeichnet, wie in Abbildung 4 beschrieben. Zwischen Benutzer und Server wird eine eigene Kommunikationsebene (AJAX Communication Layer) eingefügt. Über JavaScript werden im Hintergrund Anfragen an den Webserver gesendet, um asynchrone Updates der Webseite zu erhalten. Bei traditioneller Web-Kommunikation müsste eine Webseite für Inhaltsänderungen vollständig geladen werden.[75]

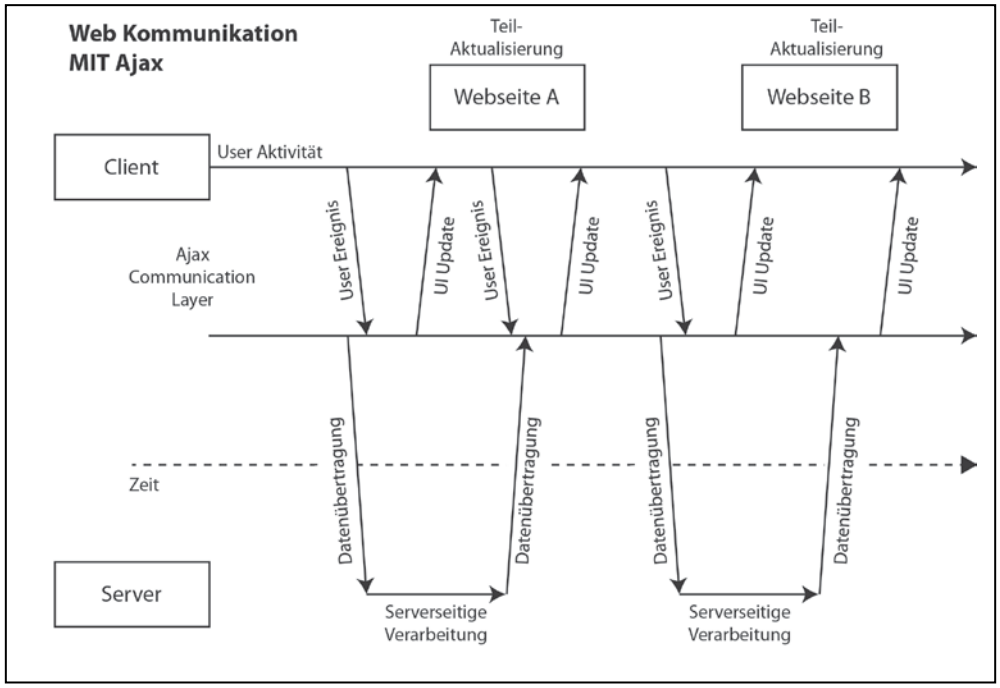

Abbildung 4: Web Kommunikation mit AJAX[76]

[73] Vgl. Erlhofer, 2008, S. 266.
[74] Vgl. Jerkovic, 2010, S. 152.
[75] Vgl. King, 2008, S. 218.
[76] Vgl. ebenda S. 219.

4.5.2.2 Problembereiche

In der Literatur identifizierte Probleme beziehen sich auf den Einsatz von AJAX ohne Optimierungsmaßnahmen für Suchmaschinen.

Das Hauptproblem beim Einsatz von AJAX basiert auf der Aktualisierung von Informationen innerhalb einer Webseite. Suchmaschinen indexieren Webseiten auf Basis einer bestimmten URL und der darauf enthaltenen Schlüsselwörter. Wird AJAX nun zur Änderung von Inhalten eingesetzt und gibt es keine alternative Bereitstellung, sind diese für Suchmaschinen nicht vollständig sichtbar. Mittels AJAX werden durch regelmäßige Updates von Informationen mehrere Versionen einer Webseite erzeugt, für die keine URLs zum Erfassen durch Suchmaschinen-Spider definiert sind.[77]

Erlhofer schreibt ergänzend, dass für Suchmaschinen nur initial gesendeter Code sichtbar ist. Ausschließlich über AJAX geladene Informationen bleiben hingegen verborgen und können ohne alternative Bereitstellung nicht indiziert werden.[78]

4.5.2.3 Einsatzempfehlung

Aufgrund der technischen Restriktionen sollte AJAX ohne erweiterte Optimierungsmaßnahmen nur dann eingesetzt werden, wenn die Indizierung von neu geladenen Informationen auf Webseiten keine große Bedeutung hat.

Hinsichtlich einer besseren Indexierung durch Suchmaschinen haben sich verschiedene Methoden entwickelt, um Inhalte für Webcrawler sichtbar zu machen.

Für Suchmaschinen sollten laut Erlhofer komplementär zu AJAX eigene Spezialseiten erstellt werden. Verweise auf diese Seiten können auf der Hauptseite in einen Noscript-Bereich eingefügt (innerhalb der Tags <noscript>, </noscript>) und später von Webcrawlern gefunden werden.[79] Diese Methode stellt somit keine direkte Lösung für die Indexierung der über AJAX nachgeladenen Informationen dar.

Google hat eine automatisierte Methode vorgestellt, welche die manuelle Bereitstellung aktualisierter Informationen aus AJAX ersetzt. Dazu generiert der Webserver ein HTML-Abbild für jede AJAX-URL. Dieses enthält sämtliche Inhalte, die nach der Ausführung von JavaScript angezeigt werden. Die Suchmaschine zieht zur Indexierung die HTML-Seite heran, liefert aber an Benutzer über die Suchergebnisseiten die entsprechende AJAX-URL aus.[80]

---

[77] Vgl. Holdener, 2008, S. 921.

[78] Vgl. Erlhofer, 2008, S. 269.

[79] Vgl. ebenda S. 269f.

[80] Vgl. URL: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=174992 [08.01.2011]

4.5.3 Flash

4.5.3.1 Beschreibung

Flash ist eine Technologie von Adobe Systems, die sich zur Umsetzung von Animationen und Benutzer-Interaktion eignet. Mittlerweile findet diese weitgehend Verbreitung und wurde bereits 2008 von ca. 98 % aller Computer mit Internetzugang unterstützt.[81] Bei Flash handelt es sich um kein Dokumenten-Format. Für die Wiedergabe dieser multimedialen Inhalte muss ein Flash-Player installiert werden.[82]

4.5.3.2 Problembereiche

Die Literatur identifiziert im Zusammenhang zwei kritische Problemfaktoren für die Sichtbarkeit von Flash in Suchmaschinen:

- Linkprobleme
- Fehlende Dokumentenstrukur

HTML ermöglicht flexible Links zu Inhalten, die in einzelne Webseiten unterteilt sind. Basieren Websites vollständig auf Flash, sind spezifische Links auf Unterseiten nicht mehr ohne Weiteres möglich. Benutzer müssen im schlechtesten Fall manuell navigieren um die gewünschten Inhalte zu erreichen.[83]

Aus SEO-Sicht bedeutet dies, dass Unterseiten in Flash nicht indiziert werden und der Wert von Links auf die Hauptseite nicht an Unterseiten innerhalb von Flash weitergegeben wird.

Enge et al. weisen im Zusammenhang mit Flash auf die problematische Dokumentenstruktur hin. Suchmaschinen machen sich in HTML Dokumenten Ankertexte, Formatierungstags, Überschriften, ALT Attribute für Bilder oder Title Tags zunutze, um die Bedeutung und Relevanz von Inhalten zu ermitteln. In Flash lässt sich eine ähnliche Strukturierung nicht realisieren. Hinzu kommt, dass aufgrund von Texteffekten Buchstabenfolgen innerhalb einzelner Wörter getrennt werden und somit der Inhalt von Suchmaschinen nicht korrekt interpretiert wird.[84]

---

[81] Vgl. Moran/Hunt, 2009, S. 254.
[82] Vgl. Koch, 2007, S. 190.
[83] Vgl. Perkins, 2009, S. 65.
[84] Vgl. Enge et al., 2010, S. 269.

Wurde Flash von Suchmaschinen einst vollständig ignoriert, indizieren mittlerweile Google und Yahoo! Textinhalte und verfolgen Links innerhalb von Flash. Bildinhalte bleiben weiterhin unberücksichtigt.[85]

4.5.3.3 Einsatzempfehlung

Grundsätzlich ist hinsichtlich einer effektiven Indexierung davon abzuraten Websites komplett auf Basis der Flash-Technologie zu erstellen. Optimierungsmöglichkeiten sind zwar vorhanden, bringen jedoch einen erheblichen Mehraufwand mit sich.

Es ist nicht vollständig geklärt, inwieweit einzelne Flash-Elemente von Suchmaschinen im Vergleich mit ihrem HTML-Äquivalent bewertet werden. Deshalb sollte für Suchmaschinen neben der Flash-Website parallel eine Alternative auf HTML-Basis erstellt werden. Weisen beide Versionen jedoch die gleichen Informationen auf, besteht die Gefahr von Duplicate Content.[86]

Für wichtige Inhalte, welche ausschließlich als Flash-Filme auf einer Webseite vorhanden sind, kann ebenfalls eine textbasierte Variante angelegt und per Link verfügbar gemacht werden. In der Robots.txt-Datei wird dann für Suchmaschinen-Spider definiert, dass nur die HTML-Version indiziert werden darf.[87]

Google äußert ebenfalls Vorbehalte, was den weitläufigen Einsatz von Flash und die Indexierung durch Suchmaschinen betrifft:

> *„Note that while Google can index the content of Flash files, other search engines may not be able to. Therefore, we recommend that you use rich-media technologies like Flash primarily for decorative purposes, and instead use HTML for content and navigation."*[88]

Um eine optimale Indexierung der Website-Inhalte zu gewährleisten, sollte Flash nur gezielt eingesetzt werden, zB wenn eine Umsetzung mittels HTML/CSS nicht möglich ist.

---

[85] Vgl. Perkins, 2009, S. 64.

[86] Vgl. ebenda S. 64f.

[87] Vgl. Koch, 2007, S. 192.

[88] URL: http://www.google.com/support/webmasters/bin/answer.py?answer=72746 [08.01.2011]

4.5.4 Frames

4.5.4.1 Beschreibung

Bei Frames handelt es sich um eine alte Technik auf Basis von HTML Code, welche verschiedene Inhaltsquellen in separaten, scrollbaren Fenstern innerhalb einer Webseite darstellen kann. Diese wird von den meisten Website-Betreibern aufgrund von Usability-Schwächen nicht mehr eingesetzt.[89]

Die Grundlage für das Layout von Frames stellt eine Frameset-Page dar, die das Browserfenster in einzelne Bereiche, den sogenannten Frames, gliedert und festlegt, welche HTML Dokumente darin dargestellt werden. Der eigentliche Body-Bereich in der Frameset-Datei bleibt somit leer. Die Navigation besteht aus einzelnen Verweisen, welche über ein target-Attribut dafür sorgt, dass die im Link enthaltene Zielseite in einem festgelegten Frame geladen wird.[90]

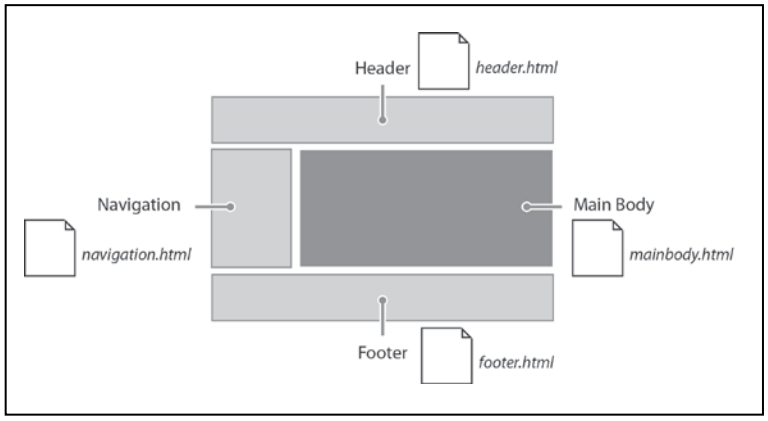

Abbildung 5: Frameset Struktur[91]

In der in Abbildung 5 angeführten Frameset-Struktur würden sich Inhalte im „Main Body"-Frame ändern. Header, Navigation sowie der Footer blieben bestehen.

Ergänzend zu den Grundelementen einer Frameset-Page führt das W3C (World Wide Web Consortium) die Verwendung eines sogenannten Noframes-Elements an. In diesem kann alternativer Inhalt für User-agents enthalten sein, die keine Frames unterstützen bzw. diese deaktiviert haben. Die Platzierung erfolgt im Dokument am Ende des Framesets.[92]

---

[89] Vgl. Moran/Hunt, 2009, S. 255.
[90] Vgl. Erlhofer, 2008, S. 233f.
[91] Vgl. Jerkovic, 2010, S. 157.
[92] Vgl. URL: http://www.w3.org/TR/REC-html40/present/frames.html [08.01.2011]

4.5.4.2 Problembereiche

Die Indexierung von Frames ist aufgrund der Inhaltsdarstellung von mehreren URLs auf einer Webseite noch immer mit Schwierigkeiten behaftet.

Jerkovic vertritt die Aussage, dass Suchmaschinen unter Umständen Inhalte eines Framesets ignorieren und nicht alle darin enthaltenen Webseiten erfassen können. Hinzu kommt, dass es keinen Sinn machen würde, alle Webseiten innerhalb der Frame-Struktur zu erfassen, da diese somit losgelöst von der Navigation angezeigt werden.[93]

Laut Erlhofer können mittlerweile die größeren Suchmaschinenbetreiber diese Strukturen verstehen und relevante Inhalte daraus gewinnen. Die Frameset-Seite an sich wird nicht gefunden, da diese keine für Suchmaschinen indizierbaren Inhalte aufweist. Ein zusätzlicher Negativeffekt ergibt sich für eingehende Links, welche ausschließlich auf die Startseite verweisen können, da nur für diese das Frameset geladen wird.[94]

In einer Stellungnahme weist Google klar daraufhin, dass Frames nicht dem konzeptuellen Modell des Webs entsprechen, bei dem eine Webseite nur eine URL darstellt. Google erkennt zwar Zusammenhänge in der Struktur, garantiert dies aber nicht.[95]

4.5.4.3 Einsatzempfehlung

Die oben genannten Probleme lassen die Schlussfolgerung zu, dass Frames von Suchmaschinen nicht vollständig interpretiert werden können und diese deshalb so schnell wie möglich in bestehenden Websites abgeschafft bzw. im Rahmen einer Neukonzipierung nicht mehr verwendet werden sollten.

Für Suchmaschinen wie auch Benutzer mit Browser ohne Frames-Unterstützung kann ein Noframes-Bereich Abhilfe schaffen. In diesen fügt man eine Beschreibung der Website ein, welche für Crawler sichtbar ist und in Folge interpretiert wird. Es sollte sichergestellt werden, dass Informationen innerhalb des Noframes-Tags mit dem Frameset übereinstimmen.[96] Enthaltene Informationen erhalten von Suchmaschinen weniger Gewichtung, da Noframes-Tags für Spam-Methoden wie Keyword-Stuffing verwendet wurden.[97]

Für das Problem der fehlenden Navigationsleiste bei direktem Aufruf einer Unterseite gibt es erweiterte Lösungsmöglichkeiten um das Frameset nachzuladen. Per JavaScript lässt sich überprüfen, ob eine Seite innerhalb eines Frames liegt. Wenn nicht, wird die

[93] Vgl. Jerkovic, 2010, S. 157.
[94] Vgl. Erlhofer, 2008, S. 234.
[95] Vgl. URL: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=34445
[96] Vgl. King, 2008, S. 51.
[97] Vgl. Koch, 2007, S. 202.

Frameset-Datei geladen und der Benutzer auf die Startseite weitergeleitet. Als Ergänzung lässt sich feststellen, von wo die Weiterleitung durchgeführt wurde und man kann in Folge die ursprünglich aufgerufene Unterseite liefern.[98]

4.5.5 Allgemeiner Hinweis zu RIAs

Rich Internet Applications (RIA) sind webbasierte Systeme für die Umsetzung komplexer Anwendungen und Interfaces. Eine wesentliche Charakteristik ist, dass der Datenaustausch zwischen Webserver und Client minimiert wird und Ebenen für Interaktion sowie Präsentation an den Client übergeben werden. Die für die Anwendung notwendigen Daten werden vom Client geladen, wobei Kommunikation zum Server nur auf Anfrage des Benutzers oder bei der Übermittlung von Daten stattfindet.[99]

Zu den bekannteren RIA-Technologien zählen Adobe Flash (siehe Kapitel 4.5.3), Java von Sun Microsystems und Microsoft Silverlight. Während Flash wie erwähnt von Suchmaschinen mittlerweile indiziert werden kann, stellen die anderen Technologien für die Indexierung ein großes Problem dar.

Google sagt zum Thema der RIAs, dass sie mittlerweile enthaltene Links und Texte in beschränkter Form extrahieren. Meist geht aber die Struktur und der jeweilige Kontext verloren. Wenn RIA Inhalte in den Index aufgenommen werden, ist es möglich, dass Inhalte oder Links fehlen.[100]

Ergänzend lassen sich Informationen in Form eines ALT-Attributs oder beschreibenden Textes einbauen, die von Spidern gelesen werden. Falls die Navigation ebenfalls in der Anwendung implementiert ist, helfen alternativ HTML Links weiter, um die einzelnen Webseiten zu erfassen.[101]

RIAs sollten angesichts der mangelhaften Indexierbarkeit nur dann eingesetzt werden, wenn die Listung einer Webseite in den Suchmaschinenergebnisseiten keine Bedeutung hat.

4.6 Zusammenfassung

In diesem Kapitel wurden technologische Kriterien erläutert, welche einen Einfluss auf die Indexierung von Webseiten durch Suchmaschinen haben. Neben der Funktionsweise einzelner Technologien wurde auf potentielle Problembereiche und Empfehlungen für einen sinnvollen Einsatz auf Websites eingegangen.

---

[98] Vgl. Erlhofer, 2008, S. 237f.

[99] Vgl. Bozon et al., 2006, S. 907.

[100] Vgl. URL: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=72746 [08.01.2011]

[101] Vgl. Koch, 2007, S. 210.

Der erste Abschnitt erklärte den Aufbau von dynamischen URLs und wie diese von Suchmaschinen interpretiert werden. Richtlinien für ihre Verwendung zeigen, wie die Indexierung von dynamischen Websites reibungslos abgewickelt werden kann.

Der zweite Abschnitt behandelte die korrekte Durchführung von Weiterleitungen. Da es unterschiedliche Möglichkeiten der Umsetzung gibt, wissen Websitebetreiber oft nicht, welche Variante Suchmaschinen verstehen.

Der nächste Abschnitt ging auf das Robots Exclusion Protocol ein, das Suchmaschinen darauf hinweist, welche Ressourcen sie indexieren dürfen. Ein inkorrekter Einsatz kann mitunter dazu führen, dass Webcrawler der Zugriff auf sämtliche Verzeichnisse verwehrt wird. Eine Anführung der wichtigsten Statements und deren Funktionen veranschaulichte, wie ein Ausschluss von Inhalten durchgeführt wird und gleichzeitig der Zugriff durch Suchmaschinen auf die wichtigsten Webseiten erhalten bleibt.

Im letzten Abschnitt wurden Web-Technologien hinsichtlich ihrer Eignung für SEO analysiert. Es wurde gezeigt, dass interaktive Technologien Suchmaschinen im Zeitalter von Web 2.0 mitunter vor technische Hürden stellen, welche mit teils beträchtlichem Optimierungsaufwand umgangen werden müssen. Im Falle der RIAs ist eine effektive Indexierung momentan nur beschränkt möglich. Außerdem sind es auch alte Technologien wie Frames, welche immer noch im Einsatz sind, aber beträchtliche Probleme für Suchmaschinen mit sich bringen.

5 Strukturelle Entscheidungen der Informationsarchitektur

5.1 Einleitung

Bei der Indexierung von Dokumenten einer Website beziehen Suchmaschinen strukturelle Kriterien mit ein. Eine der grundlegenden Entscheidungen betrifft die Organisation von Inhalten in eine logische, hierarchische Struktur, welche es Webcrawlern erlaubt die wichtigsten Webseiten zu besuchen. Ein anderes strukturelles Kriterium stellt die Website-Navigation dar. Durch diese werden die interne Linkstruktur und der Bezug von Webseiten zu verschiedenen Themen ausgedrückt.

5.2 Website-Struktur

5.2.1 Bedeutung

Suchmaschinen benötigen für die optimale Indexierung von Dokumenten innerhalb einer Website eine logische Struktur, welche sicherstellt den Zugang zu wichtigsten Inhalten so einfach wie möglich zu gestalten. Eine intuitive Architektur wird von Suchmaschinen hoch bewertet, da sie Beziehungen zwischen einzelnen Webseiten herstellt und Muster erkennen lässt.[102] Die Website-Struktur berücksichtigt die Zuweisung von Elementen zu hierarchischen Ebenen sowie eine Einordnung in relevante Gruppen durch Kategorisierung.

5.2.2 Hierarchische Struktur

Hierarchien stellen die Grundlage einer effektiven Informationsarchitektur dar. Diese sind in ihrer Struktur einfach zu erfassen und es ist jederzeit möglich, die aktuelle Position innerhalb der jeweiligen Ebenen festzustellen.[103]

Die oberste Ebene wird dabei meist von der Startseite repräsentiert, von der aus die weiteren Inhalte einer Website entdeckt werden können. Daraus ergibt sich eine Baumstruktur aus einzelnen Knoten, die durch über- und untergeordnete Ebenen gekennzeichnet ist.[104]

[102] Vgl. Enge et al., 2010, S. 191.
[103] Vgl. Morville, 1998, S. 33.
[104] Vgl. Kalbach, 2007, S. 212.

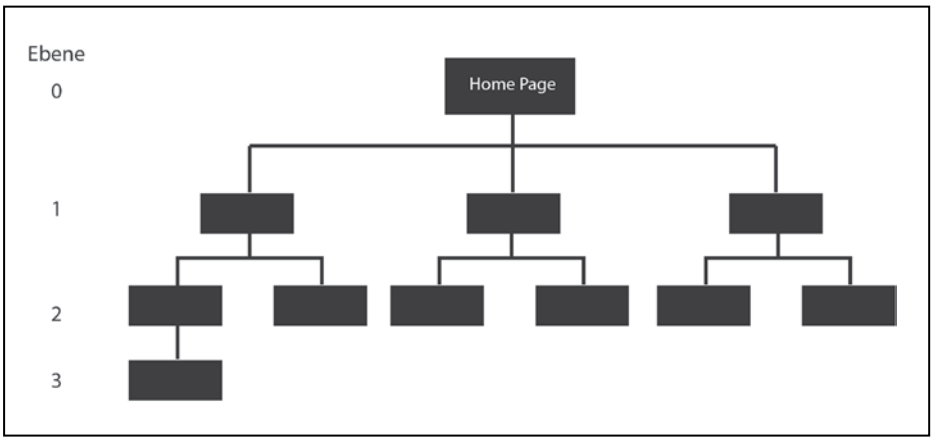

Abbildung 6: Hierarchische Struktur[105]

Für die Exploration von Website-Hierarchien wenden Suchmaschinen Methoden der Graphentheorie an. Man unterscheidet die Breiten- und Tiefensuche. Bei der Breitensuche ist das Ziel sämtliche Dokumente innerhalb einer Ebene zu erfassen, bevor bei der nächsten begonnen wird. Die bekannteren Suchmaschinen wie Google oder Yahoo! verfolgen jedoch das Prinzip der Tiefensuche, bei der Dokumente innerhalb eines Astes verfolgt werden. Sobald das Ende erreicht wurde, geht der Crawler auf die nächsthöhere Ebene zurück und nimmt einen alternativen Weg nach unten. Rein theoretisch kann somit die gesamte Website in ihrer Vollständigkeit erfasst werden.[106]

Bei der Indexierung erreichen Suchmaschinen jedoch oft nicht alle Ebenen einer Hierarchie. Aufgrund beschränkter Rechen- und Speicherressourcen könnten sonst manche Websites nicht in gleichem Umfang erfasst werden. Daraus würde eine unverhältnismäßige Gewichtung und somit Verzerrung von Suchergebnissen resultieren.[107] Nach Cutts gibt es rein theoretisch bei Google keine Beschränkung der Anzahl an Webseiten, die innerhalb einer Website besucht werden. Die Ressourcen werden hinsichtlich der Reputation und Autorität, zB nach der Anzahl eingehender Links, entsprechend ausgerichtet.[108]

[105] Vgl. Wodtke/Govella, 2009, S. 172.
[106] Vgl. Koch, 2007, S. 100.
[107] Vgl. Erlhofer, 2008, S. 281.
[108] Vgl. URL: http://www.youtube.com/watch?v=gW7AUnZvSAE [08.01.2011]

Morville beschreibt die Bedeutung der Festlegung des hierarchischen Aufbaus für Benutzer. Eine einfache Erreichbarkeit hilft auch Suchmaschinen beim Aufspüren von wichtigen Inhalten:

> *„(…) it is important to consider the balance between breadth and depth in your information hierarchy. Breadth refers to the number of options at each level of the hierarchy. Depth refers to the number of levels in the hierarchy. If a hierarchy is too narrow and deep, users have to click through an inordinate number of levels to find what they are looking for.“*[109]

Flache Hierarchien sind zu bevorzugen, da diese von Suchmaschinen einfacher erfasst werden können. Bei kleineren Websites mit bis zu 10.000 Webseiten ist eine Anzahl von drei Klicks zur Erreichung aller Inhalte umsetzbar. Bei flachen Strukturen müssen jedoch auch Einschränkungen von Suchmaschinen berücksichtigt werden. So werden von Spidern oft nicht mehr als 100 Links pro Webseite verfolgt.[110] Die Empfehlung wurde früher eingeführt, da Suchmaschinen wie zB Google Webseiten mit mehr als 100K abgeschnitten und nur teilweise indiziert haben. Dies ist heute nicht mehr der Fall. Wenn Webseiten mehr als 100 Links aufweisen, ist es trotzdem sehr wahrscheinlich, dass Webcrawler nicht allen Links folgen und in den Index aufnehmen. Zusätzlich negativ wirkt sich aus, dass der Linkwert auf eine größere Anzahl an Links aufgeteilt wird.[111]

Daraus lässt sich schließen, dass für das Design der Website-Hierarchie eine Balance zwischen Breite und Tiefe gefunden werden muss. Insbesondere die erste Ebene unterhalb der Startseite sollte sich auf die wesentlichen Bereiche einer Website konzentrieren, damit Webcrawler diese letztendlich auch besuchen und über die Website-Struktur deren Bedeutsamkeit klar erkennen.

[109] Morville, 1998, S. 34.

[110] Vgl. Enge et al., 2010, S. 194f.

[111] Vgl. URL: http://www.mattcutts.com/blog/how-many-links-per-page/ [08.01.2011]

5.2.3 Kategorisierung von Inhalten

Gleichzeitig zur Festlegung der physischen Anordnung und Verbindung von Webseiten, der Website-Hierarchie, werden Kategorien und Themen festgelegt, nach denen die Organisation von Informationen erfolgt.

Die Homepage beschreibt das Hauptthema und verweist auf Unterseiten, die zur inhaltlichen Unterstützung beitragen. Suchmaschinen gewinnen im Zuge des Crawlvorgangs eine breite Datenbasis, aus denen Muster und semantische Beziehungen extrahiert werden. Durch eine klare Kategorisierung von Informationen erkennen Suchmaschinen die Autorität einer Website hinsichtlich bestimmter Themenbereiche.[112] Die folgende Abbildung veranschaulicht, welche vielfältigen Beziehungen der Suchmaschinen-Anbieter Google zum Wort „Familienhund“ herstellt. Hinter den assoziierten Begriffen verbergen sich weitere Vernetzungen.

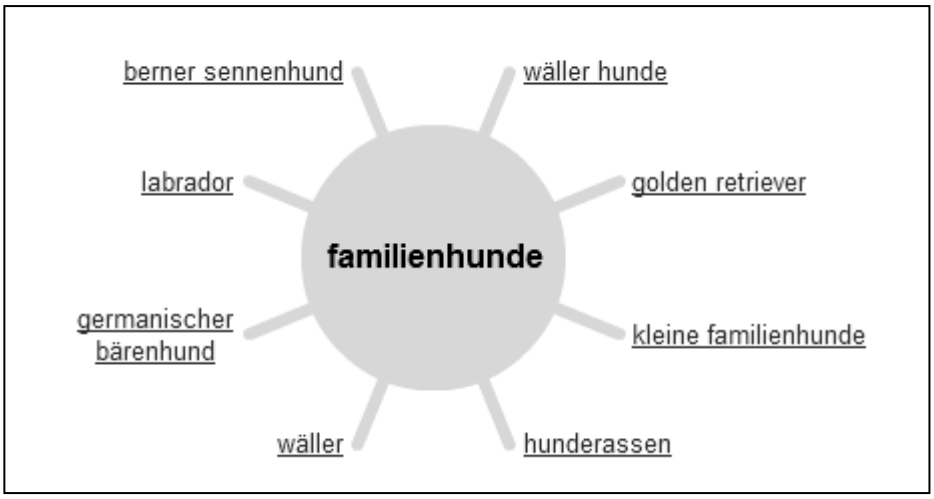

Abbildung 7: Google Wonder Wheel[113]

Wichtig ist, dass die Kategorisierung von Inhalten auf einem einheitlichen Prinzip erfolgt. Websites sind nach einem organisatorischen Schema aufgebaut. Ein Schema legt Gemeinsamkeiten einzelner Elemente fest und setzt den Rahmen für die logische Gruppierung dieser Elemente.[114]

---

[112] Vgl. Jones, 2008, S. 62.

[113] URL: http://www.google.at/search?hl=de&tbo=1&biw=1280&bih=845&tbs=ww%3A1&q=familienhund&btnG=Suche [04.01.2011]

[114] Vgl. Morville, 1998, S. 23.

Bei Websites nimmt man die Klassifizierung oft nach subjektiven Kriterien vor, die eine Gruppierung etwa nach folgenden Gesichtspunkten festlegen:[115]

- Gruppierung nach Thema
- Gruppierung nach Zielgruppe
- Gruppierung nach Aufgabe

Für eine suchmaschinenfreundliche Struktur empfiehlt sich eine Liste der wesentlichen Inhaltsseiten einer Website anzulegen. Zuerst wird die Kategorisierung für die oberste Ebene auf der Startseite vorgenommen. Danach erfolgt die Zuordnung von Detailseiten im Sinne eines Bottom-Up Prozesses und die Organisation der mittleren Ebenen.[116]

5.3 Website-Navigation

5.3.1 Bedeutung

Eine logisch gestaltete Website-Navigation ermöglicht die Fortbewegung in der im vorigen Kapitel definierten Hierarchie mit einer klaren Kategorie- und Subkategorie-Struktur und das Entdecken der wichtigsten Webseiten. In Kapitel 4 wurde darauf hingewiesen, dass manche Webseiten aufgrund technologischer Aspekte nur bedingt für Suchmaschinen auffindbar sind. Alternative Navigationswege zu Inhalten einer Website können dabei helfen diese Hürden zu überwinden.

Die Navigation repräsentiert die interne Linkstruktur einer Website, über welche die Informationshierarchie abgebildet wird. Unter internen Links versteht man Hyperlinks, die auf die gleiche Domain verweisen, auf der sie gesetzt sind. Das Ziel ist eine effektiv verfolgbare Struktur mit verschiedenen Pfaden für Suchmaschinen-Spider, damit so viele Webseiten wie möglich gefunden und in den Index aufgenommen werden.[117] Es lassen sich gezielt Prioritäten für Suchmaschinen festlegen, indem interne Links besonders auf spezifische Webseiten ausgerichtet sind. Gleichzeitig wird somit der Einfluss anderer Webseitenbereiche abgeschwächt.[118]

[115] Vgl. Kalbach, 2007, S. 218ff.
[116] Vgl. Enge et al., 2010, S. 192f.
[117] Vgl. URL: http://www.seomoz.org/learn-seo/internal-link [08.01.2011]
[118] Vgl. URL: http://www.seomoz.org/ugc/an-intelligent-way-to-plan-your-internal-linking-structure [08.01.2011]

Nielsen und Loranger weisen auf die Wichtigkeit von Konsistenz bei der Gestaltung der Website-Navigation hin:

> *„Consistency is a fundamental concept in navigation. Keeping a consistent navigation structure helps people visualize their current location and options, and minimizes guesswork. Navigational elements act as stepping-stones to help people traverse from one area to the next."*[119]

Dieser Grundsatz ist für menschliche Benutzer wie auch für Suchmaschinen Spider in der Exploration von Websites gleichermaßen von Bedeutung.

5.3.2 Formen der Website-Navigation

In ihrer rudimentärsten Form bildet eine Website-Navigation ab, wie man sich in der hierarchischen Struktur innerhalb der Ebenen einer Kategorie bewegt. Dabei werden die weitreichenden Möglichkeiten der Hyperlinkstruktur nur beschränkt genutzt. Zusätzlich ist es sinnvoll zwischen Knoten unterschiedlicher Kategorien wechseln zu können, die eine thematische Beziehung zueinander aufweisen.[120] Bei der Website-Navigation unterscheidet man drei grundlegende Typen:

- Strukturelle Navigation
- Assoziative Navigation
- Utility Navigation

Die strukturelle Navigation stellt die Verbindung von Webseiten innerhalb der hierarchischen Struktur dar. Sie ermöglicht vom aktuellen Standpunkt aus die über bzw. untergeordnete Ebene aufzurufen. Die assoziative Navigation ermöglicht die Exploration einer Website durch Verweis auf ähnliche Webseiten, unabhängig davon, in welchem Ast der Hierarchie sie sich befinden. Die Utility Navigation kann als supplementäre Navigation betrachtet werden, welche Webseiten wie Login oder AGBs zur Unterstützung des Benutzers beinhaltet.[121] Da diese meist nicht direkt in die Website-Hierarchie eingebunden ist, findet sie in diesem Kapitel keine gesonderte Beachtung.

---

[119] Nielsen/Loranger, 2006, S. 178.

[120] Vgl. Rossi et al., 1999, S. 1670.

[121] Vgl. Kalbach, 2007, S. 86.

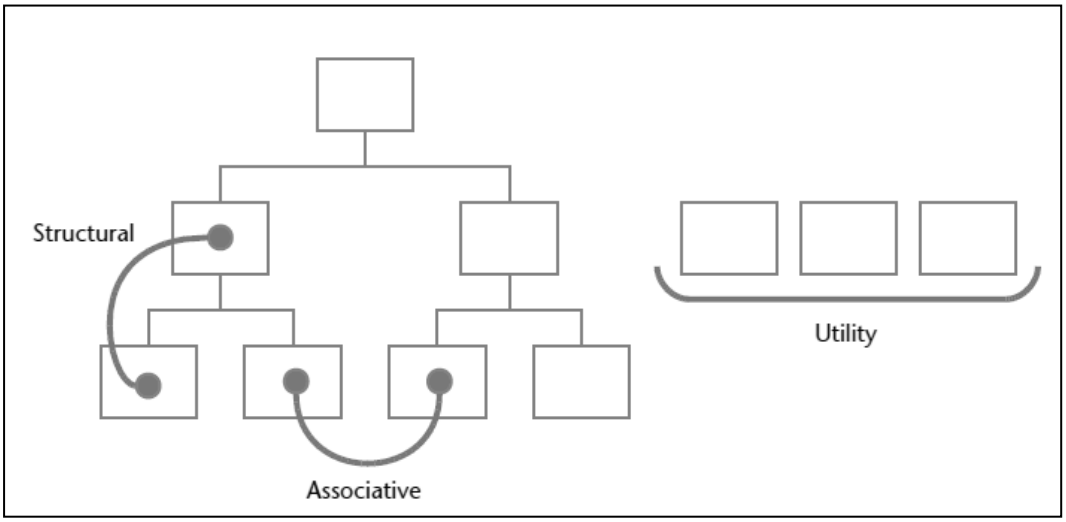

Abbildung 8: Primäre Kategorien der Navigation nach Fiorito und Dalton[122]

Neben den genannten Navigationsformen wird aufgrund ihrer Bedeutung für Suchmaschinen auf die häufig verwendete Breadcrumb-Navigation und Probleme mit dem Einsatz von Nummerierung für die Navigation zwischen einzelnen Webseiten eingegangen.

5.3.3 Strukturelle Navigation

Die strukturelle Navigation setzt sich aus der Hauptnavigation und der Subnavigation zusammen. Über diese können Suchmaschinen sich innerhalb der jeweiligen Kategorien in der hierarchischen Struktur bewegen.

5.3.3.1 Hauptnavigation

Die Hauptnavigation, auch globale Navigation genannt, hat den Zweck Links zu den Top-Level Kategorien einer Website zusammenzufassen und das übergreifende Thema zu definieren.[123] Da diese Links auch auf spezifischen Webseiten niedrigerer Hierarchie-Ebenen enthalten sind, werden diese von Suchmaschinen als besonders wichtig angesehen.[124] Das verdeutlicht die Relevanz einer sinnvollen Kategorisierung der Website-Inhalte.

5.3.3.2 Subnavigation

Die Subnavigation, auch lokale Navigation genannt, ist eine Erweiterung der Hauptnavigation, die ausschließlich Links zu den jeweiligen Unterkategorien enthält. Über sie gelangt man tiefer in die hierarchische Struktur einer Website.[125] Suchmaschinen erschließt sich

[122] Wodtke/Govella, 2009, S. 191.
[123] Vgl. Kalbach, 2007, S. 88.
[124] Vgl. Enge et al., 2010, S. 51.
[125] Vgl. Wodtke/Govella, 2009, S. 195f.

durch diese Pfade die Granularität eines Themas, vorausgesetzt die Kategorisierung wurde nach sinnvollen Merkmalen durchgeführt.

5.3.4 Assoziative Navigation

Die assoziative Navigation ermöglicht Suchmaschinen die alternative Exploration abseits der herkömmlichen Hierarchie einer Website, auch kategorieübergreifend.

5.3.4.1 Kontextuelle Navigation

Als kontextuelle Navigation werden Links verstanden, welche in der Nähe des Content-Bereichs einer Webseite platziert werden. In manchen Fällen sind diese auch direkt im Fließtext zu finden. Die verlinkten Webseiten enthalten meist einen thematischen Bezug zur verweisenden Webseite.[126] Für Suchmaschinen ergibt sich daraus ein inhaltlicher Kontext, der sich auch positiv auf die Relevanzbewertung auswirken kann.

5.3.4.2 Quick Links

Quick Links werden im Gegensatz zu kontextueller Navigation auf Website-Ebene gesetzt. Sie können dazu verwendet werden auf wichtige Inhalte tief in der hierarchischen Struktur zu verweisen.[127] Wie in diesem Kapitel erläutert indexieren Suchmaschinen aufgrund beschränkter Ressourcen Webseiten oft nur bis zu einer bestimmten Ebene, je nach Autorität der Website. Über Quick Links kann man so Webseiten für Suchmaschinen sichtbar machen, die sonst ignoriert worden wären.

5.3.4.3 Navigationslinks im Footer

Zusätzlich zur Hauptnavigation können HTML Textlinks zusätzlich im Footer einer Website eingefügt werden. Dies ist insofern beim Einsatz problematischer Web-Technologien wie JavaScript für Menüs von Bedeutung. Somit erhalten Suchmaschinen Spider einen alternativen Weg zu wichtigen Webseiten, welche über die herkömmliche Website-Architektur nicht von ihnen besucht werden können.[128] Navigationslinks im Footer sind auf jeder Unterseite enthalten. Suchmaschinen-Spider messen ihnen deshalb tendenziell weniger Bedeutung bei als Links der kontextuellen Navigation.[129]

---

[126] Vgl. Kalbach, 2007, S. 92.

[127] Vgl. ebenda S. 96.

[128] Vgl. Ledford, 2008, S. 41.

[129] Vgl. URL: http://www.youtube.com/watch?v=D0fgh5RIHdE [08.01.2011]

5.3.5 Breadcrumb-Navigation

Eine Breadcrumb-Navigation zeigt die aktuelle Position eines Benutzers innerhalb einer Website an. Sie besteht aus einer Kette von Elementen, die auf in der Hierarchie darüber liegende Webseiten verweisen.[130]

Abbildung 9: Breadcrumb-Navigation Yahoo! Directory[131]

Der Einsatz einer Breadcrumb-Navigation ist für Suchmaschinen hilfreich, da sie Verweise mit beschreibendem Ankertext auf interne Webseiten erzeugt. Webcrawler erhalten dadurch einfachen Zugang zu weiteren Informationen einer Website.[132]

5.3.6 Probleme der Navigation mittels Nummerierung (Pagination)

Bei hohem Informationsumfang werden Inhalte auf einzelne Webseiten aufgeteilt und über eine Nummern-Navigation miteinander verbunden. Diese Form findet u.a. im E-Commerce für umfangreiche Produktlisten Anwendung.[133]

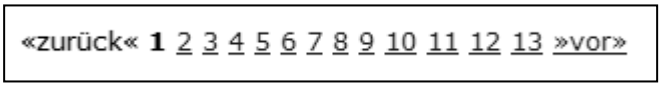

Abbildung 10: Navigation mittels Nummerierung[134]

Für die Sichtbarkeit in Suchmaschinen ist eine Navigation über Nummern nicht zu empfehlen, da diese keine Themenrelevanz vermitteln. Zusätzlich besteht die Gefahr, dass durch die Verschiebung von Inhalten in die unterschiedlichen Nummern-Seiten, Spider diese aufgrund von Duplicate Content ignorieren. In der Folge werden unzählige Webseiten nicht indiziert.[135]

---

[130] Vgl. Kalbach, 2007, S. 60.

[131] URL: http://dir.yahoo.com/Business_and_Economy/Shopping_and_Services/Barter_and_Swap/ [02.01.2011]

[132] Vgl. Erlhofer, 2008, S. 227.

[133] Vgl. Wodtke/Govella, 2009, S. 215.

[134] URL: http://geizhals.at/?cat=tvlcd&xf=33_15%2F38 [02.01.2011]

[135] Vgl. Enge et al., 2010, S. 195.

5.4 Zusammenfassung

In diesem Kapitel wurden strukturelle Kriterien betrachtet, die für eine erfolgreiche Erfassung von Websites durch Suchmaschinen berücksichtigt werden müssen.

Zuerst wurde die Basis der Informationsarchitektur durch die Festlegung einer suchmaschinenfreundlichen Website-Struktur geschaffen. Es wurde gezeigt, welchen Einfluss die Festlegung der Hierarchie-Ebenen auf die Sichtbarkeit von einzelnen Webseiten hat. Einhergehend legte die Kategorisierung von Inhalten nach verschiedenen Gesichtspunkten den Grundstein für die Relevanzbewertung.

Der nächste Abschnitt Website-Navigation veranschaulichte, in welcher Weise einzelne Webseiten innerhalb der Website-Struktur verbunden werden können, um thematische Beziehungen für Suchmaschinen zu schaffen und ihnen alternative Wege zur Exploration einer Webpräsenz anzubieten.

6 Sitemaps zur Optimierung der Indexierung

6.1 Einleitung

Aufgrund technischer Rahmenbedingungen kann es für Suchmaschinen Spider noch immer Probleme bei der Indexierung geben. Aus diversen Gründen wie Kostenaufwand oder Komplexität der Änderung von Websites werden diese Hürden von Betreibern in manchen Fällen nicht beseitigt. Über sogenannte Sitemaps kann man sicherstellen, dass Webcrawler einzelne Dokumente einer Website entdecken und die Wahrscheinlichkeit erhöhen, dass diese in den Index aufgenommen werden. Ohne diese würden bestimmte Webseiten für Suchmaschinen unsichtbar bleiben. Man unterscheidet zwischen HTML Sitemaps und XML Sitemaps, die direkt an Suchmaschinenbetreiber gesendet werden können.

6.2 HTML Sitemaps

6.2.1 Beschreibung

HTML Sitemaps stellen eine Übersicht aller verfügbaren Inhalte einer Website dar, welche aus Links zu Webseiten und optionalen Textbeschreibungen bestehen. Sitemaps vermitteln, welche Inhalte auf einer Website vorhanden sind. Auch thematische Zusammenhänge lassen sich aus der Struktur einfach erschließen.[136]

HTML Sitemaps kommt laut Cutts im Vergleich zu XML Sitemaps besondere Bedeutung zu, da sie neben Suchmaschinen auch Benutzern die Navigation erleichtern. Deshalb sollte der Einsatz einer HTML Sitemap oberste Priorität genießen.[137]

6.2.2 Erstellung und Einsatz

Die Erstellung einer HTML Sitemap erfolgt mittels eines Text-Editors. Es gibt auch die Möglichkeit spezielle Tools einzusetzen, welche eine Sitemap automatisch generieren. Links sollten stets mit einer textuellen Beschreibung versehen werden, damit Suchmaschinen die Bedeutung der Dokumenteninhalte besser verstehen.[138]

Bei bis zu 100 Webseiten können alle Links direkt in eine Sitemap, nach Kategorien geordnet, eingefügt werden. Für eine größere Anzahl an Dokumenten sollte man eine eigene Übersichtsseite je Kategorie anlegen. Enthaltene Links werden nach Wichtigkeit ge-

[136] Vgl. Moran/Hunt, 2009, S. 256.
[137] Vgl. URL: http://www.youtube.com/watch?v=hi5DGOu1uA0 [08.01.2011]
[138] Vgl. Jerkovic, 2010, S. 195.

ordnet, da Suchmaschinen Spider in manchen Fällen nur eine beschränkte Anzahl an Links auf einer Webseite indizieren.[139]

Neue oder geänderte Webseiten sollen zeitnah in die Sitemap hinzugefügt werden, da Suchmaschinen aktuelle Inhalte bevorzugen. Ein Link auf die Sitemap wird in der Regel auf der Startseite platziert. Somit können Suchmaschinen-Spider einfach darauf zugreifen und Links verfolgen.[140]

6.3 XML Sitemaps

6.3.1 Beschreibung

Google hat 2005 das Sitemap Protocol eingeführt, um die Indexierung von Webseiten durch Suchmaschinen zu beschleunigen. 2006 übernahmen andere Suchanbieter wie Microsoft oder Yahoo! diesen Standard, worauf es aus Gründen einer weitläufigeren Verwendung in XML Sitemap Protocol umbenannt wurde. Dieses erlaubt es an Suchmaschinen eine Liste mit URLs zu übermitteln, die anschließend gecrawlt und indiziert werden sollen.[141] Bevor Webcrawler eine Liste mit relevanten Links aus Sitemaps erhalten, entfernt ein Filter URLs zu spambehafteten Webseiten. Erst danach werden Inhalte über Webserver-Anfragen für die Indexierung abgerufen.[142]

Eine XML Sitemap stellt keinesfalls eine Garantie dar, dass sämtliche in Sitemaps enthaltene URLs tatsächlich in den Index aufgenommen werden. Diese ermöglichen in erster Linie, Wissen über die Struktur einer Website zu erhalten und interne Crawl-Vorgänge zu optimieren, wie Google berichtet:

> *„Google doesn't guarantee that we'll crawl or index all of your URLs. However, we use the data in your Sitemap to learn about your site's structure, which will allow us to improve our crawler schedule and do a better job crawling your site in the future."*[143]

6.3.2 Einsatzzweck

XML Sitemaps erleichtern Webcrawlern die Erfassung von Websites mit unoptimierter Linkstruktur und problematischen Web-Technologien. Über die Angabe der Crawling-Frequenz wird definiert, wie oft bestimmte Webseiten besucht werden sollen. So kann meist eine schnellere Indexierung erreicht werden, was für Websites mit einer hohen Än-

[139] Vgl. Moran/Hunt, 2009, S. 256.
[140] Vgl. Jones, 2008, S. 64.
[141] Vgl. Enge et al., 2010, S. 184.
[142] Vgl. Schonfeld/Shivakumar, 2009, S. 993.
[143] Vgl. URL: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=156184 [08.01.2011]

derungsfrequenz einen kritischen Erfolgsfaktor darstellt. Durch das zeitnahe Crawling lässt sich überdies hinaus der tatsächliche Ersteller von Inhalten zuverlässiger ermitteln. Kopien werden somit meist als solche erkannt und in Folge ignoriert. Die Festlegung von Prioritäten für einzelne URLs ist ebenfalls in XML Sitemaps möglich.[144]

Ergänzend dazu sind XML Sitemaps für neue Websites wirksam, auf die erst wenige Links verweisen und über den normalen Crawl-Vorgang schwieriger zu entdecken sind. Auch Archivseiten, die schlecht oder gar nicht intern verlinkt sind, könnten einfacher erreicht werden.[145]

6.3.3 Erstellung

Sitemaps sind Dateien für Suchmaschinen Spider, welche auf XML-Format basieren und mit UTF-8 Encoding abgespeichert werden. Für die Erstellung können unterschiedliche Tools eingesetzt werden. Eine automatische Erstellung erlaubt ein XML Sitemap Generator, der mittels Skript die Datei anlegt und in manchen Fällen auch an Suchmaschinen sendet. Eine andere Option ist das manuelle Anlegen einer simplen Textdatei mit den entsprechenden URLs und Tags. Google nimmt jedoch auch RSS und Atom 1.0 Feeds an. Diese enthalten im Gegensatz zu den ersten beiden Varianten nur Informationen zu neuen URLs.[146]

Bei der Erstellung von XML Sitemaps muss beachtet werden, dass bestimmte Sonderzeichen durch einen Escape Code ersetzt werden. Dies betrifft meist dynamische URLs, bei denen das Et-Zeichen ausgewechselt wird.[147] Die nachfolgende Tabelle enthält eine Auflistung von Zeichen mit den dazugehörigen Escape Codes.

Zeichen	Escape Code
&	&amp;
'	&apos;
"	&quot;
>	&gt;
<	&lt;

Tabelle 4: Entity Escape Characters[148]

Für jede URL kann in der XML Sitemap ein Maximum von vier Attributen festgelegt werden. Das verpflichtende Attribut <loc> enthält die tatsächliche URL. Zusätzlich kann das

[144] Vgl. Jerkovic, 2010, S. 194f.
[145] Vgl. URL: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=156184 [08.01.2011]
[146] Vgl. Enge et al., 2010, S. 185.
[147] Vgl. Ledford, 2008, S. 234.
[148] Vgl. Perkins, 2009, S. 234.

letzte Änderungsdatum, die Änderungsfrequenz sowie die Priorität Webcrawlern mitteilen, ob eine bestimmte Webseite bald besucht werden sollte.[149]

Tag	Verwendung	Beschreibung
<urlset>	Immer benötigt	Tag, der die gesamte Datei umgibt; enthält Protokollstandard
<url>	Immer benötigt	Tag für URL-Eintrag
<loc>	Immer benötigt	URL der Webseite
<lastmod>	Optional	Datum der letzten Änderung YYYY-MM-DD
<changefreq>	Optional	Änderungsfrequenz: Always, Hourly, Daily, Weekly, Monthly, Yearly, Never
<priority>	Optional	Priorität einer URL relativ zu anderen URLs der Website; Wert zwischen 0.0 und 1.0

Tabelle 5: XML Sitemap Tags[150]

Sobald die XML Sitemap erstellt ist, wird diese in der höchsten für die Indizierung relevante Verzeichnisebene gelegt. Befinden sich URLs in einer höheren Ebene, werden diese von Suchmaschinen nicht beachtet. Bei der Löschung von URLs muss die Sitemap in der Regel nicht angepasst werden. Bei einer großen Menge nicht mehr vorhandener URLs ist ihre Entfernung aber ratsam.[151] Webcrawler würden sonst unnötig die Bandbreite des Webservers beanspruchen.

6.3.4 Übermittlung an Suchmaschinen

Die XML Sitemap kann auf verschiedene Wege an Suchmaschinen übermittelt werden:

- Web Account bei Suchmaschinenanbieter
- Ping Methode
- Eintrag in robots.txt

Suchmaschinenanbieter wie Google bieten einen eigenen Account für Webmaster, über den Sitemaps eingereicht werden. Dazu muss die Inhaberschaft der Website per Einfügen eines HTML Files in das Stammverzeichnis des Webservers oder per speziellen Meta Tag in der Index-Datei nachgewiesen werden. Mittlerweile verfügen Google, Yahoo! und

[149] Vgl. Jerkovic, 2010, S. 198ff.
[150] Vgl. Ledford, 2009, S. 235.
[151] Vgl. Enge et al., 2010, S. 186.

Microsoft über ein Ping-Interface. Website-Betreiber können so die Sitemap über den Browser oder automatisch per PHP Skript übermitteln. Eine weitere Möglichkeit besteht darin, den Speicherort der XML Sitemap über die robots.txt den Webcrawlern mitzuteilen.[152]

6.3.5 Limitationen

XML Sitemaps weisen eine Beschränkung von 50.000 URLs bzw. 10 MB pro Datei auf. Um größere Datenmengen zu ermöglichen, kann man für eine Website mehrere Sitemaps anlegen, welche über einen Index mit Verweisen verbunden werden. Auch dieser Methode sind Grenzen gesetzt, da ein Sitemap-Index maximal 1000 Sitemaps aufnimmt.[153]

Laut Jerkovic ist das Limit von 10 MB manches Mal schneller erreicht, als die maximal mögliche Anzahl an URLs ausgenutzt wird. In solchen Fällen könnte man über eine sogenannte Gzip-Komprimierung die Dateigröße reduzieren und mehr URLs in eine XML Sitemap aufnehmen.[154]

6.4 Zusammenfassung

Dieses Kapitel hat verdeutlicht, wie man die Auffindbarkeit von Webseiten durch den Einsatz von Sitemaps neben der Einhaltung technologischer und struktureller Kriterien verbessern kann.

Mit HTML Sitemaps erhalten Benutzer wie auch Suchmaschinen eine einfache Übersicht der Websitestruktur und einen direkten Zugang über Links. Als Ergänzung wurden XML Sitemaps vorgestellt, die ausschließlich für die Kommunikation mit Suchmaschinen gedacht sind. Durch die nähere Spezifikation von Informationen zu einzelnen Dokumenten kann langfristig ein besseres Crawling-Verhalten für die eigene Website erzielt und eine bessere Repräsentation in den Suchergebnisseiten bewirkt werden.

[152] Vgl. Jerkovic, 2010, S. 205ff.
[153] Vgl. Ledford, 2009, S. 237f.
[154] Vgl. Jerkovic, 2010, S. 204.

7 Fazit und Ausblick

Die globale Fragestellung der Arbeit bestand darin, wie die Suchmaschinenindexierung von Websites durch On-Site Maßnahmen effektiv gestaltet werden kann. Dabei wurde zuerst aufgezeigt, welche Rahmenbedingungen im Bereich Web Hosting gegeben sein müssen, um ein optimales Fundament für die Aufnahme von Webseiten in die Datensammlung der Suchmaschinen zu schaffen. Technologische Kriterien und strukturelle Entscheidungen der Informationsarchitektur hinsichtlich ihres Einflusses auf die Indexierung fanden in weiterer Folge Beachtung. Abseits dieser Faktoren wurden weitere Maßnahmen zur Optimierung in Form von Sitemaps identifiziert und näher erläutert.

Grundsätzlich von Bedeutung für die Beantwortung der zentralen Fragestellung war es die Herausforderungen sowie den Ablauf der **Suchmaschinenindexierung** aufzuzeigen. Es wurde deutlich, dass das Sammeln von Dokumenten ein ressourcenintensiver Prozess ist. Aufgrund der enormen Menge an Daten sind Suchmaschinen gezwungen, ihre verfügbaren Kapazitäten aufzuteilen, um eine optimale Repräsentation des World Wide Webs im Index zu erreichen. Des Weiteren stellen komplexe Strukturen und der Einsatz verschiedener Formate auf Websites Hürden für die korrekte Erfassung von Webseiten dar. Ein wichtiger Aspekt ist die Auffindung von Dokumenten über eine Hyperlinkstruktur. Diese unterstreicht die Wichtigkeit, vielfältige Pfade zur Erreichung von Inhalten für Webcrawler anzulegen. In der Analyse wurde klar, dass Suchmaschinen aus Webseiten textuelle Informationen in Form von Schlüsselwörtern extrahieren und diese für den Aufbau des Suchmaschinen-Index heranziehen. Dies stellt im Hinblick des Einsatzes verschiedener Web-Technologien ein kritisches Erfolgskriterium dar.

Zunächst untersuchte eine Fragestellung, welche **Rahmenbedingungen im Web Hosting** für die Suchmaschinenindexierung erfüllt werden müssen. Es wurden relevante Erfolgsfaktoren und verschiedene Hosting-Typen hinsichtlich ihrer Eignung erläutert. Das Ergebnis dieser Betrachtung war, dass eine hohe Verfügbarkeit des Webservers und kurze Auslieferungszeiten von Webseiten an Webcrawler eine wichtige Rolle spielen, damit Suchmaschinen neue Ressourcen einer Website erfassen bzw. bereits im Index enthaltene aktualisieren. Der Betrieb eines eigenen Webservers bietet größtmögliche Chancen der Einflussnahme und Kontrolle, was insbesondere für größere Projekte essentiell ist. Shared Hosting eignet sich ebenfalls für eine hohe Sichtbarkeit in Suchmaschinen, vorausgesetzt man wendet sich an einen kompetenten Dienstleister, der einen entsprechenden Service Level sicherstellt.

In weiterer Folge war es Ziel der Arbeit den Einfluss **technologischer Kriterien** auf die Sichtbarkeit einer Website in Suchmaschinen darzustellen.

Bei dynamischen URLs gilt, auf besucherbezogene Parameter wie Session IDs vollständig zu verzichten, da sonst gleiche Inhalte unter verschiedenen URLs in den Index aufgenommen werden. Wesentlich ist die Erzeugung kurzer URLs und einhergehend die Beschränkung der Parameteranzahl. Im Falle einer Umwandlung in statische URLs ist zu beachten, keine strukturrelevanten Informationen vor den Suchmaschinen zu verstecken. Im Zusammenhang mit verschobenen oder gelöschten Webseiten wurden suchmaschinengerechte Weiterleitungen diskutiert. Eine Umsetzung erfolgt ausschließlich über server-seitige HTTP Status Codes, die von einem Webcrawler korrekt interpretiert werden können. In weiterer Folge wurde mit der robots.txt ein Protokoll angeführt, das Ressourcen einer Website von der Indexierung ausschließt. Bei der Umsetzung müssen Anweisungen mit Bedacht gewählt werden, damit der Zugriff von Suchmaschinen auf wichtige Inhalte erhalten bleibt. Im letzten Abschnitt wurden Empfehlungen für den Einsatz von Web-Technologien abgegeben. Bei der Konzipierung einer neuen Website sollte vor allem von Frames Abstand genommen werden. Moderne Web-Technologien wie Flash oder AJAX bedeuten meist einen großen Optimierungsaufwand für die Indexierung und werden nicht von allen Suchmaschinenanbietern erfolgreich interpretiert. Als sichere Lösung zeigte sich, dass für Webcrawler meist alternative Inhalte in Textform angeboten werden sollen und Pfade über eine herkömmliche Hyperlinkstruktur die Exploration einer Website erleichtern.

Ebenfalls Teil der Zielerreichung waren **strukturelle Entscheidungen der Informationsarchitektur** für eine optimale Indexierung. Im Mittelpunkt der Betrachtung standen beschränkte Ressourcen der Suchmaschinen, welche einer vollständigen Repräsentation einer Website im Index entgegenstehen. Inhalte werden auf Ebenen einer hierarchischen Baumstruktur, nach Kategorien gruppiert, angeordnet – mit den wichtigsten Inhalten in den oberen Ebenen. Die Website-Navigation realisiert Pfade für Suchmaschinen-Spider in der Form einer internen Linkstruktur. Dadurch wird festgelegt, welchen Inhalten in der Indexierung besondere Priorität zukommt.

In der letzten Unterfrage war von Interesse, welche **weiteren Maßnahmen zur Optimierung der Indexierung** beitragen. Diese wurde mit der Diskussion von HTML und XML Sitemaps beantwortet. HTML Sitemaps erfüllen das Bedürfnis von Suchmaschinen nach einer simplen Hyperlinkstruktur. Eine Übersicht aller relevanten Webseiteninhalte auf einer HTML Seite stellen die Auffindbarkeit durch Webcrawler sicher. Über XML Sitemaps werden Suchmaschinen zusätzlich für den Aufbau sowie Pflege des Index Informationen zur Priorität und Aktualisierungsfrequenz von Webseiten übergeben. Diese führen meist zu einer besseren Repräsentation der eigenen Website in den Trefferlisten von Suchmaschinen.

Im Rahmen der Arbeit wurde gezeigt, dass Suchmaschinen die Herausforderungen der Indexierung deutlich besser meistern als dies früher der Fall war. Insbesondere Google

als weltweit führender Anbieter kann mittlerweile Informationen aus den meisten Web-Technologien und Website-Strukturen gewinnen, wenn auch oft nur in beschränktem Maße. Bei der Konzipierung einer Website ist weiterhin zu beachten, dass nicht alle Suchmaschinen ähnlich fortgeschrittene Indexierungsfähigkeiten bei modernen Technologien wie AJAX oder Flash aufweisen und somit die Gefahr besteht, dass Webseiten in den Ergebnisseiten unberücksichtigt bleiben. Websitebetreiber benötigen bei der Auswahl Feingespür, welche Technologie bei der Umsetzung einer Website zweckdienlich ist - hinsichtlich der Indexierung von Suchmaschinen sowie den Bedürfnissen von Benutzern. Waren es in den Anfängen simple textbasierte HTML-Seiten mit Hyperlinkstrukturen, welche effektives Webcrawling ermöglichten, so stellen diese weiterhin eine sichere Lösung für die Sichtbarkeit in Suchmaschinen dar.

Suchmaschinen verfügen über enorme Ressourcen, um dem Anspruch einer möglichst hohen Abdeckung der Dokumente im World Wide Web und fortlaufender Aktualisierung des Datenbestands gerecht zu werden. Aufgrund der steigenden Anzahl an Websites ist eine vollständige Erfassung in absehbarer Zeit weiterhin nicht realistisch. Websitebetreiber sind gefordert, wichtige Webseiten für die Indexierung über eine effektive interne Linkstruktur und Eingliederung in den oberen Ebenen einer Website-Hierarchie sichtbar zu machen.

Literaturverzeichnis

Monographien, Bücher und Sammelbände

Eckert, Claudia: IT-Sicherheit. Konzepte – Verfahren – Protokolle, München, 2009

Enge, Eric / Spencer, Stephan / Fishkin, Rand / Stricchiola, Jessie C.: The Art of SEO. Mastering Search Engine Optimization, Sebastopol, 2010

Erlhofer, Sebastian: Suchmaschinen-Optimierung für Webentwickler. Das umfassende Handbuch, 4. Auflage, 2008

Holdener, Anthony T. III, Ajax. The Definitive Guide, 1. Auflage, Sebastopol, 2008

Jerkovic, John I.: SEO Warrior, Sebastopol, 2010

Jones, Kristopher B.: Search Engine Optimization. Your visual blueprint for effective Internet marketing, Indianapolis, 2008

Kalbach, James: Designing Web Navigation, 1. Auflage, Sebastopol, 2007

King, Andrew B.: Website Optimization, 1. Auflage, Sebastopol, 2008

Koch, Daniel: Suchmaschinen-Optimierung. Website-Marketing für Entwickler, München, 2007

Ledford, Jerry L.: Search Engine Optimization Bible, Indianapolis, 2008

Lewandowski, Dirk: Web Information Retrieval. Technologien zur Informationssuche im Internet, 2005

Moran, Mike / Hunt, Bill: Search Engine Marketing, Inc. Driving Search Traffic to Your Company's Web Site, 2. Auflage, Boston, 2009

Morville, Peter: Information Architecture for the World Wide Web, 1. Auflage, 1998

Nielsen, Jakob / Loranger, Hoa: Prioritizing Web Usability, Berkeley, 2006

Perkins, Todd: Search Engine Optimization for Flash, 1. Auflage, Sebastopol, 2009

Schwarz, Torsten: Leitfaden Online Marketing, 2. Auflage, 2008

Sherman, Chris / Price, Gary: The Invisbile Web. Uncovering Information Sources Search Engines Can't See, 2. Auflage, New Jersey, 2001

Wodtke, Christina / Govella, Austin: Information Architecture. Blueprints for the Web, 2. Auflage, Berkeley, 2009

Fachartikel und Journale

Bozon, Alessandro / Comai, Sara / Fraternali, Piero / Carughi, Giovanni T.: Capturing RIA Concepts in a Web Modeling Language, in: WWW ’06 Proceedings of the 15th international conference on the World Wide Web, 2006, 907-908

Chellapilla, Kumar / Maykov, Alexey: A Taxonomy of JavaScript Redirection Spam, in: Proceedings AIRWeb ’07, 2007, 81-88

Kobayashi, Mei / Takeda, Koichi: Information Retrieval on the Web, in: ACM Computing Surveys, Vol. 32, No. 2, 06/2000, 144-173

Rossi, Gustavo / Schwabe, Daniel / Lyardet, Fernando: Improving Web Information Systems with Navigational Patterns, in: Computer Networks, Vol. 31, 05/1999, 1667-1678

Schonfeld, Uri / Shivakumar, Narayanan: Sitemaps. Above and Beyond the Crawl of Duty, in: WWW ’09 Proceedings of the 18th international conference on the World Wide Web, 2009, 991-1000

Artikel aus dem Web

http://googlewebmastercentral.blogspot.com/2008/09/dynamic-urls-vs-static-urls.html, Google Webmaster Central, Abfrage: 08.01.2011

http://searchengineland.com/an-update-on-javascript-menus-and-seo-16060, Search Engine Land, Abfrage: 08.01.2011

http://www.comscore.com/Press_Events/Press_Releases/2010/1/Global_Search_Market_Grows_46_Percent_in_2009, comScore, Abfrage 13.11.2010

http://www.google.com/support/webmasters/bin/answer.py?answer=40132&cbid=-1pe1wokkq7axt&src=cb&lev=%20answer, Google Webmaster Help, Abfrage: 08.01.2011

http://www.google.com/support/webmasters/bin/answer.py?answer=72746, Google Webmaster Help, Abfrage: 08.01.2011

http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=156184, Google Webmaster Help, Abfrage: 08.01.2011

http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=174992, Google Webmaster Help, Abfrage. 08.01.2011

http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=34445, Google Webmaster Help, Abfrage: 08.01.2011

http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=66355, Google Webmaster Help, Abfrage: 08.01.2011

http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=81766, Google Webmaster Help, Abfrage: 08.01.2011

http://www.mattcutts.com/blog/how-many-links-per-page/, Matt Cutts. Gadgets, Google and SEO, Abfrage: 08.01.2011

http://www.mattcutts.com/blog/myth-busting-virtual-hosts-vs-dedicated-ip-addresses/, Matt Cutts. Gadgets, Google and SEO, Abfrage: 08.01.2011

http://www.seomoz.org/blog/301-redirect-or-relcanonical-which-one-should-you-use, SEOmoz, Abfrage: 08.01.2011

http://www.seomoz.org/learn-seo/internal-link, SEOmoz, Abfrage: 08.01.2011

http://www.seomoz.org/ugc/an-intelligent-way-to-plan-your-internal-linking-structure, SEOmoz, Abfrage: 08.01.2011

http://www.sistrix.com/blog/721-hosting-with-own-ip-address-necessary-for-seo.html, SISTRIX, Abfrage: 08.01.2011

http://www.sistrix.com/blog/902-what-is-the-impact-of-the-webserver-speed.html, SISTRIX, Abfrage: 08.01.2011

http://www.w3.org/TR/REC-html40/present/frames.html, World Wide Web Consortium, Abfrage: 08.01.2011