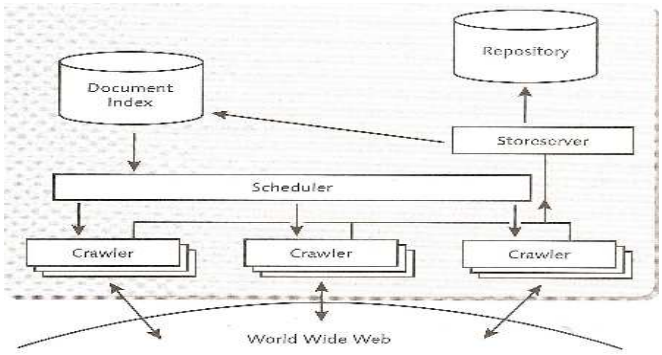
Recommended Search Engine Optimization

by Ralf Kuschel
Student Freie Universität Berlin

I. Einleitung

Jeder kennt sie und für viele gehört sie zum Internet wie die Butter zum Brot. Die Rede ist von der Suchmaschine deren Marktführer seit einigen Jahren Google darstellt. Doch die Einführung der Suchmaschine kam nicht mit der Erfindung von Google, sondern mit der Erfindung des Internets selber. Mit der Erfindung dieses globalen Netzwerkes und den ersten Websites gab es Suchwerkzeuge um sich in diesem Netzwerk zu Recht zu finden. Zum Anfang geschah dies durch Webkataloge wo die wenigen Websites die bis dahin existierten aufgeführt wurden und durchsucht werden konnten. Im Zuge der weiten Verbreitung des Internets und dem immer größer werdenden Volumen sind konnte man Webkataloge nun schwerer verwalten, dies war der eigentliche Anbeginn der Suchmaschinen. Der Grundstein der bekanntesten Suchmaschine der Welt wird 1995 an der Standford University gelegt wo sich Lawrence Page und Sergey Brin treffen und 3 Jahre später, am 7. September 1998 die Google inc. gründen. 10 Jahre später ist Google eine der größten Firmen der Welt und jedes Kind kennt die Adresse www.google.de. Und obwohl Google die am meisten genutzte Suchmaschine des Internets ist haben sich viele bei einer Sucheingabe schon einmal gefragt warum dieser Link so weit oben steht obwohl er nichts mit dem Thema zu tun hat. Im nachfolgenden werde ich erläutern wie es möglich ist das Google Ranking im Rahmen der legalen Möglichkeiten zu beeinflussen denn ein Google Ranking kann über den Erfolg oder Misserfolg einer Website entscheiden.

II. Architektur



Quelle: Suchmaschinenoptimierung von Sebastian Erlhofer

Doch bevor wir darüber sprechen wie man als Webentwickler seine Website so verändert das sie im Google Ranking nach oben klettert müssen wir uns erst einmal mit der Architektur von Suchmaschinen auseinander setzen. Eine Suchmaschine besteht grundsätzlich aus 3 Komponenten nämlich der Datengewinnung der Datenanalyse und der Datenverwaltung.
Die Architektur der Suchmaschine besteht hauptsächlich aus dem Webcrawler-System. Der Webcrawler lädt Dokumente aus dem World Wide Web die der Suchmaschine unbekannt sind. Der Großteil der geladenen URLs sind aus dem Web akquiriert. Da das Volumen des Webs von Jahr zu Jahr wächst müssen Dokumente auch regelmäßig auf Aktualität geprüft werden und gegeben falls muss der Datenbestand modifiziert werden.

Die Datensammlung erfolgt komplett durch die Crawler welche die Schnittstelle der Suchmaschine zum World Wide Web. Im Webcrawler-System gibt es 3 Arten von Modulen. Zum einen gibt es die

Protokoll Module welche als Clients direkten Kontakt zum World Wide Web haben. Zu den Protokoll Modulen gehören die einzelnen Crawler welche die Schnittstelle zum World Wide Web sind.
Das zweite Modul ist das Verarbeitungsmodul. Durch den Storeserver und Scheduler werden die gesammelten Daten nun verarbeitet und abgespeichert damit die beiden Datenmodule, der Document Index und das Repository, welche das dritte Modul bilden die Informationen abrufen kann.

II.I Document Index

Der Document Index übernimmt die Datenverwaltung in der Form das er einzelnen Dokumenten eine DocID zuteilt welches in Form eines eindeutigen Schlüssels geschieht. In diesem Schlüssel sind wichtige Informationen verankert wie den Status des Dokuments und Informationen über den Zustand (zum Beispiel ob das Dokument bereits indexiert wurde). Ein weiterer Pointer im Document Index verweist auf die lokale Kopie welche im Repository abgespeichert ist. Um zwei verschiedene Dokumente zu vergleichen nutzt der Dokumentenindex eine sogenannte Checksumme.
Die Checksumme ist eine Zeichenabfolge von Buchstaben und Zahlen. Außerdem enthält der Dokumentenindex statistische Daten über das Dokument. Dazu gehören die Länge des Dokuments, ein Zeitstempel der das Erstelldatum und des letzten Besuchs beinhaltet, die Änderungshäufigkeit, Seitentitel und der Hostname und IP-Adresse des Hosters. Die Erfassung der IP-Adresse wird aus dem Grund gemacht, da zum Beispiel bei Northernlight, wenn eine URL mit pornografischen Inhalten entdeckt wird, so werden sämtliche URLs mit der gleichen IP auch aus der Datenbank gelöscht. Der Dokumentenindex wird häufig als URL-Datenbank bezeichnet da sie viele Links selbstständig bewertet und verwaltet.

II.II Scheduler

Das Verwaltungsorgan der Suchmaschine ist der Scheduler, welcher die einzelnen Webcrawler organisiert und ihnen die Aufträge erteilt. Die dafür notwendigen Informationen bezieht der Scheduler aus der Datenbank des Dokumentenindex. Die Hauptaufgabe des Schedulers ist die Pflege und Erweiterung des bereits vorhandenen Datenbestandes. Dies geschieht in Form eines alternierenden Wellensystems. Den Begriff alternierend welcher in der Mathematischen Form bei einer Folge für einen dauerhaften Vorzeichenwechsel steht in der Suchmaschine für einen Wechsel zwischen neu zu erfassenden Seiten und zu pflegenden Seiten. Der Scheduler errechnet anhand des Verhältnisses zwischen beiden, welche Aufgaben an die Crawler weitergegeben werden müssen. Der Scheduler weiß immer über den Zustand der Crawler bescheidt.

II.III Webcrawler

Die Crawler ,welche im allgemeinen als Cluster vorliegen, das heißt in einem kompletten Rechensystem auftreten können entweder frei für neue Aufträge sein, mit dem Server in Verbindung treten und einen HTTP – Request senden, auf Antwort warten oder den HTTP – Response verarbeiten. Nach erfolgreicher Verarbeitung schickt der Crawler das Ergebnis an den Storeserver. Eine erfolgreiche Verarbeitung heißt nicht, dass in jedem Fall die Website aktualisiert wurde, sondern kann auch bedeuten dass diese nicht mehr vorliegt. Wenn dies der Fall ist gibt der Storeserver dem Dokumentenindex den Befehl das Dokument aus seiner Datenbank zu löschen, oder wenn sie vorhanden ist den Datenbestand der Website zu aktualisieren. Anhand der Funktionsweise des Schedulers kann man schon einen ersten Optimierungsaspekt erkennen, denn der Scheduler gewichtet die zu besuchenden URLs nach bestimmten Verfahren. Dabei hält er sich zum Beispiel an die Daten die im Zeitstempel stehen, denn wenn ein Dokument sehr oft aktualisiert wird geht der Scheduler davon aus, das die Informationen die das Dokument beinhaltet relevanter und zeitgemäßer sind. Daraus folgt das die Website bei häufiger Aktualisierung auch häufig vom Webcrawler abgefragt wird.
Zum anderen kann auch die Tiefe des Dokumentes über die Gewichtung entscheiden. Wenn

der Pfad eines Dokuments zum Beispiel: http://www.domain.de/index/anderes/info.html ist, so ergibt sich für das Dokument ‘Info‘ eine Tiefe von Zwei. Der Hintergrundgedanke dabei ist, dass der Webcrawler davon ausgeht das tieferliegende Dokumente unwichtiger sind als Dokumente die auf der Root-Ebene liegen. Im Zuge dessen sollte man also prägnante Informationen über das Thema der Website bereits auf die Root-Ebene zu schreiben. Daher werden sie seltener sie seltener auf Änderungen überprüft. Die genaue Gewichtung solcher Aspekte, sind allerdings von Anbieter zu Anbieter unterschiedlich. Die wichtigste Komponente im System der Suchmaschine ist der Crawler welcher auch die einzige Komponente ist die außerhalb des abgeschlossenen Systems arbeitet. Der Kontaktpartner für die Crawler sind die Web- und DNS-Server. Die Crawler bekommen vom Scheduler den Auftrag eine bestimmte URL zu suchen und von dort entweder die Ressource zu downloaden oder zu prüfen ob die Ressource verändert oder gelöscht wurde. Wie bereits erwähnt treten Crawler immer in Cluster auf und alleine Google betreibt 10.000 Server auf denen wiederum ca. 200 Crawler Prozesse laufen. Der Vorteil eines Linux-Rechners mit üblicher Hardwareausstattung ist zum einen, dass er günstiger ist und wenn ein Rechner ausfällt kann er problemlos vom Cluster getrennt werden können.
Der Crawler sendet mittels DNS-Caches einen HTTP-Request an die betreffende IP-Adresse und fordert den Server mit der GET-Methode auf die Ressource und die Header-Informationen zu übertragen, welche er dann an den Storeserver weiterleitet. Außerdem gibt es spezielle Crawler welche nur Flash-Animationen oder PDF-Dateien abfragen.

II.IV Storeserver

Der Storeserver ist für die Sicherung der Daten verantwortlich die er vom Crawler erhält. Er wertet die aus und bringt den Dokumentenindex auf den aktuellen Stand. Gleichzeitig unterzieht der Storeserver für die Ressource eine Aufnahmeprüfung. Wenn der Storeserver die Information bekommt, dass die Ressource gelöscht wurde bekommt er einen Statuscode und wertet diesen dann aus. Durch die Prüfung von Speicherwürdigkeit und Verarbeitbarkeit der Ressource ergeben sich für Webautoren wieder einige Formalitäten die sie beachten müssen. Die Ressource durchläuft hierbei eine Kette von Filtern von denen einige Suchmaschinen abhängig sind, es jedoch drei Filter gibt die sehr einheitlich sind. Der erste Filterprozess ist die Prüfung des Dokumenttyps. Hierbei überprüft der Storeserver den Medientyp, da aus Audio- und Videoressourcen keine Informationen ausgelesen werden können. Um Medientypen auszulesen analysiert der Storeserver den MIME-Typ und den Content-Type. Der zweite Filterprozess ist die Dublettenerkennung. Hierbei geht es darum URLs zu erkennen die auf dieselbe Ressource linken. Da viele Pfade einer URL sehr häufig verwendet werden und nicht auf die gleiche Ressource linken verwendet der Storeserver hierbei die weiter oben erwähnte Checksumme. Existiert im Datenbestand bereits eine Ressource die die gleiche Checksumme besitzt so wird die überprüfte Ressource nicht noch einmal in die Datenbank aufgenommen. Der dritte Filter ist der URL-Filter bei dem überprüft wird ob die Ressource hinter der URL geändert oder komplett ersetzt wurde. Dabei achtet der Filter besonders aus dynamisch generierte Dokumente da bei diesen Dokumenten die Gefahr sehr hoch ist. Zur Erkennung werden aus der URL-Spezifikation definierte Sonderzeichen wie ?, &, =, % benutzt. Außerdem untersucht der Filter die Domain sowie Verzeichnis- und Dateinamen auf Wörter und Phrasen welche auf der allgemeinen Blacklist stehen. Ein weiterer Aspekt der gerade jungen Webautoren zum Verhängnis werden kann ist das auch die Dokumenttiefe analysiert wird. Da viele kostenlosen Anbieter auf ihrem Webspace, Werbebanner installieren welche eine hohe Verzeichnistiefe haben kann es passieren, dass die Website aufgrund der hohen Verzeichnistiefe abgelehnt wird.

II.V. Repository

Hat eine Ressource nun alle Überprüfungen überwunden, so wird sie als lokale Kopie in das Repository gespeichert. Der Datenspeicher vom Repository enthält überwiegend Webseiten mit HTML-Code. Das Repository enthält die Arbeit des gesamten Webcrawler Systems.
Doch warum ist Google heutzutage die größte Suchmaschine der Welt und was unterscheidet Google von anderen Suchmaschinen.

III. Ranking-Verfahren

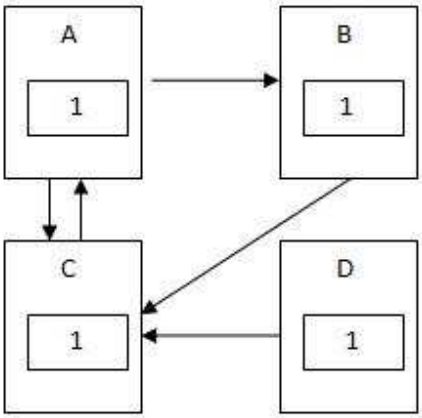
III.I. Pagerank-Verfahren

Google verdankt seinen enormen Erfolg dem Pagerank System. Pagerank verdankt seinen Namen dem Erfinder Lawrence Page dem Mitbegründer von Google. Das Pagerank-Verfahren wird im allgemeinen Kontext auch als Link-Popularity bezeichnet. Der Hintergedanke dabei ist, dass Veröffentlichungen bedeutender sind wenn sie öfter zitiert werden. Das bedeutet im Internet, je mehr eingehende Links eine Webseite hat umso bedeutsamer ist nach der Theorie auch ihr Inhalt. Dafür muss die Anzahl der eingehenden Links sehr hoch sein, da die Suchmaschine dann aus vielen subjektiven Empfehlungen eine quasiobjektive Meinung bildet. Diese Links sieht man besonders oft bei informativen Webseiten, in dem diese sich per Quellenlink auf die Hauptinformationsseite beziehen und ihr durch diesen Link eine höhere Link Popularity verschaffen. Doch wie schnell zu erkennen ist können die Webautoren auch hier das Ranking wieder stark beeinflussen indem sie neue Webseiten entwerfen deren einziger Zweck es darstellt auf die Hauptseite des Webautors zu linken. Das Pagerank-Verfahren hat aus diesem Grund noch einen weiteren Aspekt der betrachtet wird. So besitzen alle Dokumente einen gewissen Rang und wenn die Verlinkung einer Webseite von einem Dokument mit hohem Rang erfolgt, so ist die Bewertung dann wesentlich besser als wenn sie von einer unbedeutenden Webseite erfolgt. Weiterhin ist auch entscheidend wie viele ausgehende Links die Webseite hat, das heißt je weniger verschiedene Links auf der Webseite sind desto besser ist die Bewertung, weil die Chance einen bestimmten Link zu verfolgen wesentlich höher ist.

Beispiel:

Zu Beginn des Verfahrens würde jeder Seite der Pagerank von 1 zugewiesen. In der folgenden Abbildung symbolisieren A, B, C, D verschiedene Webseiten die mithilfe der Pfeile aufeinander verlinken. Zusätzlich rechnen wir mit einer Dämpfung d = 0,85. Der Dämpfungsfaktor bezeichnet in dem Algorithmus von Page eine Individualisierungsvariable und reflektiert den Wert, den eine Webseite einer anderen Webseite von seinem eigenen Pagewert vererben kann. Es soll damit eine Feineinstellung der Berechnungsmethode erreicht werden, so dass ein Webdokument mit einem Verweis nicht seinen vollen PageRank-Wert weiterreichen kann.

$$PR(A) = (1 - 0{,}85) + 0{,}85 * \left(\frac{PR(C)}{1}\right)$$
$$PR(A) = 0{,}15 + 0{,}85 * 1 - 1$$



Die Berechnung des Pageranks von B setzt sich nun aus dem berechneten Wert für A zusammen. In der Formel wird PR(A) durch 2 geteilt da A 2 ausgehende Links besitzt.

$$PR(B) = (1 - 0{,}85) + 0{,}85 * \left(\frac{PR(A)}{2}\right)$$
$$PR(B) = 0{,}15 + 0{,}85 * 0{,}5 = 0{,}58$$

$$PR(C) = (1 - 0{,}85) + 0{,}85 * \left(\frac{PR(A)}{2} + \frac{PR(B)}{1} + \frac{PR(D)}{1}\right)$$
$$PR(C) = 0{,}15 + 0{,}85 * (0{,}5 + 0{,}58 + 1) = 1{,}91$$

$$PR(D) = (1 - 0{,}85) + 0{,}85 * 0 = 0{,}15$$

Mit den neuen Werten beginnt die neue Iteration das heißt das Verfahren wird wiederholt.

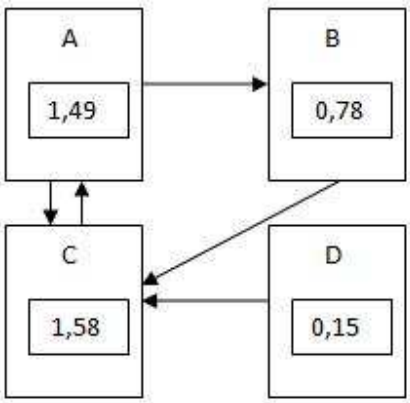
$$PR(A) = (1 - 0{,}85) + 0{,}85 * \left(\frac{PR(C)}{1}\right)$$
$$PR(A) = 0{,}15 + 0{,}85 * 1{,}91 = 1{,}77$$

$$PR(B) = (1 - 0{,}85) + 0{,}85 * \left(\frac{PR(A)}{2}\right)$$
$$PR(B) = 0{,}15 + 0{,}85 * 0{,}89 = 0{,}91$$

$$PR(C) = (1 - 0{,}85) + 0{,}85 * \left(\frac{PR(A)}{2} + \frac{PR(B)}{1} + \frac{PR(D)}{1}\right)$$
$$PR(C) = 0{,}15 + 0{,}85 * (0{,}89 + 0{,}91 + 0{,}15) = 1{,}80$$

$$PR(D) = (1 - 0{,}85) + 0{,}85 * 0) = 0{,}15$$

Man müsste diese Iterationen etwas Zwanzig mal wiederholen bevor man für jede Webseite einen stabilen Wert erhält der sich nichtmehr verändert. Bei D kann man jedoch schon sehen das der Pagerankwert nach der zweiten Iteration schon konvergiert. In dem angegebenen Beispiel welches ich aus dem Buch Suchmaschinenoptimierung verwendet habe bekommt man nach zwanzig Iterationen einen kovergierenden Wert.



Daraus können wir lesen, dass die Seite D die nur einen ausgehenden Link hat die schlechteste Pagerankbewertung hat, während C mit 3 eingehenden und einem ausgehenden Link die höchste Priorität ausweist.
Beim Vergleich zwischen A und C ist auch gut zu erkennen, dass obwohl A mehr ausgehende Links als C hat die Bewertung von C trotzdem höher ist. Daraus ergibt sich das eingehende Links höher bewertet werden als ausgehende Links.

Google bot als erster Suchmaschinen-Betreiber eine Toolbar an, welche den Pagerank einer Website anzeigt. Dieser Pagerank bewegt sich zwischen 0 und 10. Da es später viele Programme gab welche den Pagerank anzeigten nahm Google in ihrer Berechnung immer einen Code der bereits 3-4 Monate veraltet war. Dennoch bot die Pagerank-Anzeige einen groben Anhaltspunkt ob man seine Website besser angepasst hatte oder nicht. So wie das Pagerank-System von Google positive Bewertungen vergibt, gibt es aber auch ein Bad-Rank System. So besitzt die Google Suchmaschine eine Kategorie die sich „Bad Neighbourhood“ nennt. Wenn man also auf seiner Website, Links auf Seiten setzt die im Bad-Rank System erfasst sind riskiert man auch eine schlechte Bewertung für die eigene Website.
Die Seiten die bereits im Bad-Rank System erfasst sind gehören meist zu der bereits früher erwähnten Blacklist welche die Suchmaschinenbetreiber führen. Als Webautor kann man diese Bad-Rank Seiten meist daran erkennen, dass die eine Pagerank-Bewertung von 0 aufweisen, obwohl sie bereits seit längerem indexiert sind. Allgemein kann man sagen, das man auch eingehende Links so hingegen wählen sollte, dass die Webseiten die die Website verlinken, meist eine ähnliche Klientel ansprechen und eine gute Pagerank-Bewertung haben.

III.II Cluster-Verfahren

Ein weiteres Ranking-Verfahren welches von den bisherigen abweicht, ist das Clustering oder

auch Cluster-Verfahren. Man kann sich das Clustering als eine Art Gruppenzuweisung von ähnlichen Dokumenten vorstellen, die aufgrund ihrer Dokumenteninhalte und Eigenschaften als ähnlich deklaritert werden. Der Vorteil des Clusterverfahrens ist, das er durch die Gruppenzuweisung auch Ergebnisse liefert die nicht direkt auf die Suchanfrage passen.
Außerdem erfolgt durch das Clustering eine thematische Gliederung der Suchergebnisse. Wenn der User nun als Beispiel „Auto“ eingibt, so erhält er durch das Clustering auch thematische Seiten für Versicherungen, Autokauf, Zeitschriften und Technisches. Google band als einer der ersten Suchmaschinenbetreiber das Cluster-Verfahren mit in die Ergebnisdarstellung mit ein. Über den Link „Ähnliche Seiten“ kann sich der Benutzer das Cluster zu einem betreffenden Eintrag anzeigen lassen. Die Cluster-Bildung bei Google basiert auf einer Hyperlink-Struktur. Daraufhin zeigt das Cluster-System das Dokument und alle darauf zuweisenden Dokumente an.
Somit wären die durch die Architektur bedingten Anpassungen der Website aufgezeigt, doch was kann der Webautor nun an seiner Webseite verändern das er im Ranking weiter oben steht.
Hierbei unterscheidet man zwischen On-Page Optimierung und Off-Page Optimierung.

IV. On-Page Optimierung

Die Grundlage der On-Page Optimierung ist der korrekte Umgang mit den Webtechnologien HTML und CSS. Da die Parser der Webcrawler-Systeme oftmals nicht so fehlertolerant sind sollte man dringend auf „sauberes“ HTML achten. Die Systeme extrahieren aus dem Quelltext das <title> tag und Überschriften um diese gesondert zu gewichten. Dies ist allerdings nur möglich solange der HTML-Code das Erkennen der Auszeichnungen wie <title> oder <h1> zulässt. Im schlimmsten Fall kann der Parser bei einer vergessenen schließenden Klammer ganze Teile des Dokuments nicht auslesen. Eine weitere Fehlerquelle von Webautoren ist die unglückliche Verwendung von Anführungszeichen. Hierbei sollte man darauf achten das diese immer im Paar auftreten, sodass der Wert des Attributs umschlossen ist. Andernfalls wird der Text von einem bis zum nächsten Anführungszeichen interpretiert, und somit können Beispielsweise Überschriften nicht ausgelesen werden die bekanntlich in eine höhere Gewichtung erhalten. Diese Fehler geschehen meist nicht aus Unwissenheit sondern ausUnachtsamkeit. Weiterhin kommt es auch oft vor das eine falsche URL angegeben wird

<a href=”http://mindblast.de//projekte/index.html“>

Hier wird der Besucher nicht nur den 404-Fehlercode angezeigt bekommen, sondern die Webseite ist auch für die Suchmaschine nicht zu indexieren.

IV.I Webstandards

Da viele Webseiten heutzutage auch nicht mehr von Hand programmiert werden, sondern mit Programmen kommt es häufig vor, dass diese falsche oder veraltete Code-Fragmente benutzen die nicht dem allgemeinen HTML-Standard entsprechen. Dazu kommt das in den letzten Jahren ein Konkurrenzkampf zwischen den Browserherstellern ausgebrochen ist, wodurch es zunehmend proprietäre HTML-Tags eingeführt wurden. Die proprietären HTML-Tags werden von den Suchmaschinen jedoch nicht unterstützt. Um zu prüfen ob der HTML-Code fehlerfrei ist bietet das W3C mit dem W3C-Validator einen Gültigkeitscheck an. Der W3C-Standard ist der allgemeine Standard dem HTML unterliegt. Auf sauberen HTML-Code sollte auch deswegen besonders viel Wert gelegt werden da ein Dokument mit sauberem HTML-Code ein Beweis für Sorgfalt und Professionalität des Webautors darstellt und darüber hinaus auch ein anwendbares Gütekriterium für Suchmaschinen ist.

IV.II. Cascading Style Sheet

Neben HTML ist auch CSS in den letzten Jahren beliebter geworden, da es durch die

Trennung von Inhalt und Design auf vielen Feldern mehr Möglichkeiten bietet. Bei CSS lässt sich das Layout des Dokuments beinahe komplett aus den HTML-Dokumenten auslagern und somit kann man das Aussehen der gesamten Website zentral bestimmen. Ein Vorteil bei der Suchmaschinen-Optimierung ist nun das diese keine Cascading Style Sheets interpretieren.
Dennoch begeht man hier eine Gradwanderung denn solange Suchmaschinen kein CSS berücksichtigen kann der Webautor seine Inhalte so positionieren, dass sie zwar für den Benutzer sichtbar sind, jedoch für die Parser der Suchmaschinen nicht. Da es sich hierbei jedoch um einen Täuschungsversuch der Suchmaschinen handelt, sollte man als Nutzer dieser Methodik immer auf dem Laufenden sein, denn irgendwann werden Suchmaschinen auch CSS interpretieren können und dann läuft man Gefahr das sie eigene Seite aus dem Index geschmissen wird. Daher sollte man wenn man CSS zur Optimierung benutzt die CSS-Datei immer auslagern, da dies die Erfassung der Suchmaschine erschwert da der Crawler jedesmal zwei Dokumente runterladen und interpretieren müsste. Ein weiterer häufig von jungen Webautoren gemachter Fehler ist, dass sie Schriftstücke die sie oder andere in Word geschrieben haben bei Copy and Paste in das Webdokument einfügen. Da Word jedoch meist einer anderen Formatierung unterliegt als das das Programm mit dem das Webdokument erstellt wird, kommt es hierbei zu Formatierungsproblemen.

IV.III. HTML - Tags

Neben dem Einsatz von sauberem HTML, sollte man auch darauf achten das man die HTML-Tags korrekt verwendet. Bei der Optimierung einer Webseite kommt es vor, dass man statt einer <h1>-Überschrift zum Beispiel so etwas vorfindet:

* Über Baumwurzeln *

Diese Überschrift erscheint für den Benutzer zwar genauso wie die <h1>-Überschrift, dennoch sollte im Zuge einer guten Optimierung auf die speziellen Formatierungen der entsprechenden HTML-Tags geachtet werden. Die korrekte Bezeichnung für die Überschrift wäre in diesem Fall:

<h1 class="ueberschrift_gross">Über Baumwurzeln>/h1>

Doch warum sind HTML-Tags so wichtig, denn für den Benutzer ändert sich ja nichts und doch gibt es einen großen Unterschied. Denn während der Benutzer zwar das gleiche sieht, muss der Crawler ja den gegebenen Quelltext auslesen und dieser weiß nicht das es sich im oberen Beispiel um eine Überschrift handelt und außerdem nicht das es sich um eine Überschrift der Ebene 1 handelt. Durch solche Spielereien kann die semantische Struktur des Dokuments nicht korrekt erkannt werden. In der Überschrift wäre dies zum Beispiel der Fall, da Keywords in Überschriften wesentlich höher bewertet werden als Keywords in einem losen Fließtest.

IV.IV. Seitenstruktur

Ein weiterer Aspekt ist die Seitenstruktur die in logisch **hierarchischer Beziehun**g stehen muss und für den User transparent sein muss. Denn was bringt eine gut bewertete Seite wenn der Benutzer sie nicht nutzt beziehungsweise sich nicht auf ihr zu Recht findet. So sollten die wichtigen Dokumente auf einer Webseite direkt erreichbar sein.

Normalerweise nutzt man hierfür die Navigation-Frames die man links, rechts oder auf beiden Seiten anbringt und dort eine gewisse Gliederung der Website einbringt, damit der Besucher über die Index-Seite alle wichtigen Teile der Website über Hyperlinks (welche meist durch Buttons verarbeitet werden) schnellstmöglich zu erreichen, und ihnen so auch keine große Dokumenttiefe zu geben. Zu vermeiden sind auch Sackgassen auf der eigenen Website.

Als Sackgassen kann man das Ende von Verlinkungen ansehen, wo es dem Benutzer nichtmehr möglich ist durch einen Link

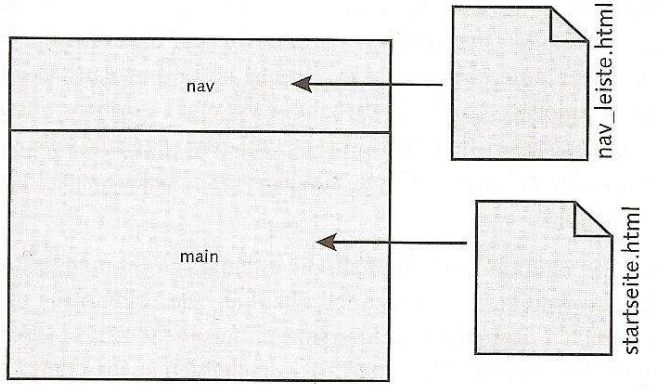
wieder zurück zukommen, sondern er muss den „Zurück“-Button des Browsers verwenden.
Dies ist nicht nur nervig, sondern lässt sich auch einfach die ein dauerhaften Navigation-Frame verhindern da der Benutzer so immer wieder auf die Indexseite zurück kommt. Ein weiterer Aspekt ist die **Dokumententiefe**, denn bei einer Tiefe die höher als 4 ist, verliert so mancher Benutzer bereits die Orientierung auf der Website. Da wir dies nicht erzielen wollen sollte man die Tiefe so wählen das der User praktisch immer weiß wo er sich befindet.
Hierbei kann auch das einblenden des Pfades nicht schaden, um dem Benutzer zu ermöglichen auf die vorherige Ebene zurück zu kommen ohne wieder auf die Root-Ebene zurück zu müssen.

Seit 2007 beeinflussen auch seiteninterne Suchfunktionen die Bewertung denn Google macht sich die Suchmöglichkeit auf Webseiten zum Vorteil und sucht über diese, relevante Suchbegriffe.

IV.V. Frames

Die Verwendung von Frames war früher sehr beliebt, doch heutzutage verwendet man immer weniger Frames da diese sehr viele Tücken aufweisen. Suchmaschinen tun sich mit Frameseiten sehr schwer und indexieren diese meist immer noch fehlerhaft. Doch warum tut sich die Google Suchmaschine so schwer mit Frames. Normale HTML-Dokumente enthalten den darzustellenden Inhalt in Form von Text und Bild in der betreffenden Datei. Wenn man nun Frames verwendet unterteilt man das Browserfenster in verschiedene Bereiche eben diese Frames. Eine solche Seite enthält keinen Inhalt, jedoch enthalten HTML-Dokumente die innerhalb dieses Rahmens liegen die Informationen und so kommt es zu einer Verschachtelung der einzelnen Seiten da diese nebeneinander gestellt werden können.
Innerhalb einer Frameset-Seite gibt es folgenden HTML-Code:

```
<html>
<head>
</head>
   <frameset rows="20  %,80  %">
      <frame src="nav_leiste.html" name="nav">
      <frame src="startseite.html" name="main">
   </frameset>
<body>
</body>
</html>
```



Quelle: Suchoptimierung von Sebastian Erlhofer

Wie in der oberen Grafik zu erkennen ist definiert, definiert das Frameset-Tag eine Zweiteilung der Fläche im Verhältnis 20% (nav) zu 80% (main).
Hierbei kommt es bei der Erfassung durch die Suchmaschinen schon zu Problemen, denn wie im oberen HTML-Code sichtbar ist, ist das <body>, welcher den eigentlichen Inhalt der Seite definiert, leer ist. Neuere Webcrawler sind jedoch in der Lage aus dem Framesets Inhalte zu extrahieren, allerdings ergibt sich hier bereits das nächste Problem. Denn dadurch, dass die Inhalte indexiert werden, wird der nav und der main Bereich getrennt von der Suchmaschine indexiert. Auch durch eingehende Links ergibt sich ein Problem, denn wenn ein Besucher von einem externen Werbebanner direkt auf die Produktseite verwiesen werden möchte, ist das bei Frames nicht ohne weiteres möglich.

IV.VI Dateitypen

Ein weiterer wichtiger Aspekt, ist die Frage welchem Dateityp das erstellte Dokument angehören soll. Da Suchmaschinen wie bereits erwähnt ein Information-Retrieval-System benutzen, und dieses hauptsächlich auf der Auswertung von Textelementen basiert, bilden HTML-Dokumente den dominierenden Standard im

Web. Im Zuge der Weiterentwicklung der Suchmaschinen stellen aber heutzutage auch PDF-Dokumente und MS Office-Programmdateien vergleichbare Alternativen dar, da auch ihre Textformate ausgewertet werden können. Ein Format das den Suchmaschinen lange Zeit Probleme bereitete, ist die Auswertung von Flash-Dateien. Seit der Erfindung der FAST-Technologie ist es auch möglich Flash Seiten zu indexieren, jedoch wurde Google und Yahoo! diese Technik erst 2008 von Adobe bereitgestellt. Nun hat man also die Wahl zwischen dem bewährten HTML-Format und dem neuen Flash-Dateien, wobei letzteres viele Vorzüge bietet, denn so weiß jeder das eine gut programmierte Flash-Seite ein wesentlich besseren Anblick liefert, als ein HTML-Dokument. Doch wer seine Webseite suchmaschinenoptimierend gestalten will greift immer noch auf HTML zurück, denn trotz neuer Technik haben die Webcrawler bei einem HTML-Dokument leichteres Spiel. So steht in diesem Fall einfach Schönheit gegen Effizienz. Wenn man nun seinem HTML-Dokument durch Grafiken ein schöneres Aussehen verleihen will, sollte man wiederrum Vorsichtig sein. Denn viele Webautoren beginnen damit Fließtexte nicht in den HTML-Code zu schreiben, sondern sie durch Grafiken zu ersetzen. Da für einen Webcrawler eine Grafik jedoch nur eine Ansammlung von Pixeln ist, wird er diese auch nicht bewerten. Somit empfiehlt es sich wichtige Texte immer auch als Text zu schreiben und nicht aufgrund grafischer Finessen auf eine Optimierung in der Suchmaschine zu verzichten.

IV.VII Dynamisches/Statisches HTML

Der Unterschied zwischen diesen beiden Formaten ist lediglich darin, dass sie anders vom Webserver verarbeitet werden. Mit der zunehmenden Verbreitung von Shop-Systemen und CMS (Content Management Systemen) wächst die Anzahl der dynamischen Webseiten. Während bei dem Statischen HTML alles als Text gespeichert und hochgeladen wird, ist eine dynamische Webseite eher ein Skelett das mit Variablen aufgefüllt wird. Der Vorteil der dynamischen Webseiten liegt hierbei bei den Shop-Systemen, denn bei einem Skelett muss eine Produktsuche nur einmal programmiert werden. Dennoch besitzen dynamische Webseiten im Zuge der Suchmaschinenoptimierung große Nachteile, die besonders bei der Indexierung auftreten. Da bei einer dynamischen Webseite nur ein Skelett vorgegeben ist, ist es möglich den Inhalt einer Webseite relativ schnell zu verändern, sodass der Inhalt während der Indexierung bereits nichtmehr aktuell sein kann. Dynamisch generierte Webseiten haben viel mit Eingaben bzw. POST- und GET-Formularen zu tun, welche für Suchmaschinen nicht greifbar sind. Je nach Programmierung ist es auch möglich, dass sich die URL verändert und der Suchmaschine beim erneuten Abrufen ein Fehlercode ausgegeben wird. Daran kann man sehen, das ein Dateityp der scheinbar viele Vorteile hat, im Bezug auf die Suchoptimierung viele Nachteile besitzt und das ein statisch generiertes HTML der sichere Weg ist, eine gute Position im Google-Ranking zu erreichen.

V. Off-Page Optimierung

Bei der Off-Page Optimierung geht es darum wie man seine Webseite mit den Gegebenheiten der Umwelt am besten optimiert. Darein fällt bereits das bereits erwähnte anpassen an die Architektur der Suchmaschinen, sodass diese nicht aus dem Index fallen. Eine wichtige Frage im Bezug auf die Off-Page Optimierung ist das Webhosting, da diese eine entscheidende Bedeutung hat. Viele Webautoren achten hierbei meist auf den Geldbeutel, da die Mietung eines Webservers, mit einem großen Webspace sehr teuer ist. Die Auswahl des Webhosters ist dahingehend entscheidend, da beispielsweise aufgrund einer Blacklist-Seite, alle IP-Adressen des Webhosters seitens der Suchmaschine gesperrt wurden. Ein weiterer Stichpunkt in der Off-Page Optimierung sind die Public Relations. Durch effizientes werben, für die eigene Webseite durch Comments, Forenbeiträge und Einträge in Gästebüchern versucht man die eigene Webseite populärer zu machen und neue Benutzer für die eigene Webseite zu begeistern. Weiterhin sollte man als Webautor Social Networking betreiben. Social Networking ist das angeben und einbinden der eigenen Webseiten auf Community-Plattformen wie www.digg.com und

www.slashdot.org, welche Seiten mit hoher Besucherzahl sind. Ein besonderer Fall, ist die Seite www.myspace.com, da man durch ein Anlegen des Profils mit einer minimalen Optimierung, alleine durch den Bekanntheitsgrad von MySpace im Google Ranking sehr weit oben steht.
Junge Webautoren sollten besonders bei der Domain nicht sparen, so sollte man eine .de Domain immer einer kostenlosen Domain (.de.vu etc.) vorziehen. Das der Domainname wichtig ist sieht man beispielsweise an dem Suchwort BMW. Während das Wort als Keyword auf einigen Fahrzeugbörsen und Ersatzteilläden wesentlich öfter auftaucht steht im Goggle-Ranking die Seite www.bmw.de am weitesten oben. Somit zeigt sich das eine Domain im Ranking sehr viel verändern kann. Dazu kommt auch, dass das Alter der Domain eine wichtige Rolle spielt, denn eine ältere Domain hat bessere Chancen im Google Ranking weiteroben zu stehen als eine junge Domain. Nachdem nun die Domain genau benannt ist, sollte man auch darauf achten, dass sämtliche Grafiken und Videos richtig benannt sind, denn mittlerweile gibt es auch Bildersuchen bei Suchmaschinen. Ein weiterer Punkt der mit Public Relations einhergeht ist die Form des Link Exchanges. Beim Link Exchange tauscht man mit Betreibern anderer Webseiten den Link aus, sodass diese gegenseitig aufeinander verlinken. Wie früher schon mal erklärt erhöht sich im Pagerank-System der Wert einer Seite durch eingehende Links und somit empfiehlt es sich immer, Kontakte mit anderen Webautoren zu knüpfen. Man sollte aber vielleicht darauf achten, das man Kontakte mit Webautoren knüpft die Webpräsenzen besitzen die auf das gleiche Metier abzielen wie die eigene Seite.Weitergehend sollte man auch prüfen, wie die linkende Webseite im Goggle Ranking bewertet wird, denn eine Webseite die durch Google eine gute Bewertung hat, liefert automatisch auch für die eigene Webseite eine bessere Bewertung. Link Exchange kann auch dadurch stattfinden, dass man die eigene Seite durch eine Signatur in Foren verlinkt. Eine Webseite die sich bestens dafür eignet auf die eigene Webseite zu verlinken ist Wikipedia, denn „Wiki“ ist jedem bekannt auch Leuten, die nicht viel mit dem Internet zu tun haben. Doch auch hier gilt es wieder einige Sachen zu beachten, damit das scheinbar positive Feature nicht nach hinten los geht. Hierbei entsteht nämlich schnell der Verdacht auf Missbrauch, denn neben einem fundierten Wikipedia-Text muss gegeben sein, dass die verlinkte Webseite den Benutzer nicht in die Irre führt, sondern die Themeninhalte die auf der Wikipedia-Seite weiterführen müssen. Da Wikipedia ein Webkatalog ist, wird jeder Beitrag vorher von einer Person angeschaut, und somit ist die Möglichkeit seine Webseite bei Wikipedia zu verlinken ohne dabei auf das Thema eingehen zu müssen sehr gering.

VI. Keyword-Recherche

Mit meinem letzten Punkt möchte ich nochmal auf ein Thema eingehen das bei der Suchmaschinenoptimierung eine ganz besondere Rolle spielt. So erwähnte ich bereits, dass Keywords in bestimmten Quelltext-Phrasen eine höhere Wertung haben, zum Beispiel in Überschriften doch wie achtet eine Suchmaschine denn nun auf Keywords. Ein wichtiges Kriterium ist es exakte Keywords für die Zielgruppe zu setzen die man erreichen möchte, denn die idealen Keywords gibt es nicht.
Keywords unterliegen gewissen Gütekriterien und im Grunde genommen bieten diese auch lediglich ein besseres Ausgangsmaterial für das Information-Retrieval-System der Suchmaschinen. Sie stellen nämlich den Schlüssel zur Erschließung des Inhalts der Webseite dar. An den richtigen Stellen und an der Wahl der richtigen Schlüsselwörter kann der Einsatz der Keywords enorme Auswirkungen für den Erfolg einer Webseite haben. Wie der deutschen Sprache bereits zu entnehmen ist stellen Substantive im Satz einen wichtigen Baustein dar. Genauso verhält sich dies auch bei Keywords, denn Substantive werden meist eher gesucht als Tätigkeiten oder Adjektve. Man sollte dabei auch darauf achten keine Abkürzungen zu verwenden, denn wenn jemand nach Voice over IP sucht werden die wenigsten VoIP eingeben sondern den kompletten Begriff. Doch wie finde ich nun die passenden Keywords für meine Webseite. Grundsätzlich sollten Keywords das Thema der Seite so genau wie möglich beschreiben, denn selbst wenn ein Besucher auf ihre Seite kommt, jedoch sieht das seine Suche mit ihrer Webseite nichts zu

tun, verlässt er diese umgehend wieder. Desweiteren sollte man sich über das Nutzungspotenzial der Keywords im Klaren sein. Denn was bringt das beste Keyword wenn es nicht in die Google-Suche eingetragen wird. Hierbei gilt es heraus zu filtern, welche Schlüsselwörter auch häufig gesucht werden die zu der jeweiligen Webseite passen. Und auch hier sollte man wieder einen weiteren Aspekt analysieren, denn vielleicht haben auch viele andere das gleiche sensationelle Keyword entdeckt, welches man selber entdeckt hat und schon geht man in der Menge der Mitbewerber um einen Topplatz bei Google unter. Deswegen sollte man im Vorfeld analysieren wie hoch bei einem bestimmten Keyword die Mitbewerbergröße ist denn je niedriger diese ist umso größer ist die Chance im Ranking ganz oben zu stehen. Zum Anfang sollte man sich hierbei eine Keyword-Liste erstellen wo man alle Keywords einträgt die man für geeignet hält und welche man vergleichen möchte. Hat man nun ein Keyword das man unbedingt nehmen möchte, trotz großer Mitbewerberzahl empfiehlt es sich auf Synonyme zurück zu greifen, denn viele deutsche Bedeutungen haben mehrere Wörter die sie bezeichnen. Da man seine Mitbewerberdichte jedoch bei der Größe des Webs selten überblicken kann hat Google für Webautoren das AdWords-Keywords-Tool bereit gestellt. Unter der Seite

https://adwords.google.de/select/KeywordToolExternal

kann der Webautor sein Keyword eingeben und erhält genaue Informationen über die Mitbewerberdichte und die Benutzung des Keywords erhält.
Auch bei Keywords sollte man genauestens auf die Rechtschreibung achten, es sei denn man möchte darauf hoffen, dass sein Kundenkreis das Wort falsch schreibt um auf seine Seite zu gelangen. Somit kann man sagen das bei Keywords sehr viel Recherchearbeit zu erledigen ist, da man zwar einen bekannten Term nehmen sollte, jedoch keinen Term der vor Mitbewerbern nur so überquillt. Darum sollte man sich im Vorfeld einige Gedanken darüber machen, und notfalls über Mitbewerber und Güte des Schlüsselwortes recherchieren.

VI.I Keyword-Dichte

Da Schlüsselbegriffe auf der Webseite eine wichtige Rolle einnehmen meinen es Webautoren teilweise zu gut und verwenden Keywords sehr häufig. Die Keyword-Dichte setzt voraus, dass ein Text maximal eine gewisse Häufigkeit an Keywords besitzen darf. Wenn man nun jedoch zuviele Keywords verwendet kommt man in den Verdacht des Betruges, in dem man Keywords spammt. Die Häufigkeit steht im Verhältnis zu gesamten Worthäufigkeit. Das bedeutet Keywords sollten maximal zwischen 3% und 8% aller Wörter sein. Berechnet wird der genaue Wert mit dem TF-Algorithmus.

VI.I.I TF-Algorithmus

Der TF-Algorithmus weist einem Keyword einen TF-Wert zu, also sollte dieses 40x vorkommen so erhält das Wort einen Wert von 40. Der TF-Algorithmus betrachtet nun das Keyword in mit der Gesamtwortzahl und vergleicht diesen Wert mit anderen Dokumenten. Daraus lässt sich erkennen inwiefern das Wort zwar für das gesamte Dokument aussagekräftig ist, jedoch zeugt ein zu häufig eingesetztes Keyword auch von der Quantität des Textes und das dieser eine niedrigere Qualität besitzt.

VII. Conclusion

Wenn ich im Zuge der Recherchen über dieses Thema eins gelernt habe, dann das, dass die Optimierung einer Webseite für Google eine Gratwanderung ist. Die Möglichkeit am Extrem zu arbeiten ohne dabei ständig Gefahr zu laufen das die eigene Seite aus dem Index fliegt ist sehr schwer. Außerdem muss man im Zuge der Google optimierten Webseite auf einige Features verzichten. Beispielsweise muss man am Design sparen, denn wie wir nun wissen besteht eine optimierte Seite zum größten Teil aus Textelementen und dort ist kein Platz für grafische Kreativität. Der bloße Fakt das die Webcrawler scheinbar lieber mit den alten Standards arbeiten macht einem als Webautor das Leben schwer. So sind Cascading Style Sheets schwer zu benutzen, da sie die Google Optimierung nicht gerade verbessern. Auch grafisch anspruchsvolle Flash-Seiten laufen sehr schnell Gefahr das sie von Google nicht erfasst werden und auch das Arbeiten mit kompletten Framesets behindert die Arbeit der Crawler. Somit kann man sagen das eine optimierte Seite im Quelltext nur aus HTML-Fragmenten besteht und gleichzeitig trotzdem noch nicht automatisch bei Google weit oben steht. Denn trotz alledem muss man im Bereich der Off-Page Optimierung sehr viel formelle Arbeit leisten in dem man analysiert, recherchiert und virtuelle Kontakte knüpft damit die eigene Webseite einen gewissen

Stellenwert erreicht. Doch wann hat man sein Ziel erreicht? Wenn die Seite bei Google ganz oben steht oder auf der ersten Seite? Meiner Meinung nach kommt es darauf gar nicht an denn letztendlich muss man die eigene Webseite nur so bekannt machen, dass sie die

Kundschaft die sie anziehen möchte erhält und die Leute die Seite nicht erst bei Google suchen müssen sondern sie im Internet direkt anwählen. Da hat man seine Website nicht nur für Google optimal programmiert, sondern die Webseite genießt in dem Metier auch einen gewissen Stellenwert. Wenn diese beiden Faktoren zusammentreffen kann man von einer volloptimierten Webseite ausgehen.

Optimierung	Effekt
HTML-Standards beachten	- Webseite kommt in den Index - Crawler können die Webseite besser auslesen
Link-Popularity	- Erhöhung des Pageranks wenn man Kontakte zu kompetenten Webseiten hegt
Keywords	- Keywords an den wichtigen Stellen plazieren - Mit der Dichte nicht übertreiben
„sauberes HTML“	- Zeugt von der Qualität der Webseite - Bescheinigt dem Programmierer eine hohe Kompetenz
Hierarchische Seitenstruktur	- Bessere Benutzbarkeit - Webseite gewinnt zusätzlich an Qualität
Domain	- Eine geeignete Domain die der Thematik der Webseite angepasst ist, erhöht die Chance bei gesuchtem Begriff weit oben zu stehen
Public Relation	- Die Bekanntmachung der Seite kann einen größeren Effekt haben als die Optimierung selber - Wenn eine Seite als besonders wertvoll empfunden wird, so wird diese auch weiterempfohlen

Quellen:

1. Suchmaschinen-Optimierung – Das umfassende Handbuch von Sebastian Erlhofer
2. http://www.wikipedia.de
3. http://www.seo-solutions.de/artikel/geschichte-der-suchmaschine-google.html
4. http://www.lousigerblick.de/archives/141-Grundlagen-der-On-Page-SEO.html
5. http://www-wi.uni-muenster.de/pi/lehre/ss05/seminarSuchen/Ausarbeitungen/ChristophLehrke.pdf