Studienarbeit I

Einflussfaktoren der heutigen Suchmaschinenoptimierung am Beispiel von Google

von

PASCAL LANDAU

Student der Angewandten Informatik an der
Dualen Hochschule Baden-Württemberg

1. Januar 2011

Matrikelnummer:	104375
Kurs:	TAI08B1
Unternehmen:	SAP AG
Betreuer:	Daniel Lindner

Zusammenfassung

Die Informationssuche im Internet erfuhr in den letzten 10 Jahren einen enormen Aufschwung gegenüber herkömmlichen Medien. Dafür ist vor allem die gesteigerte Qualität der Ergebnisse der Suchmaschinen verantwortlich. Diese Qualität ist maßgeblich in den Rankingalgorithmen der Suchmaschinen begründet, die die relevantesten Treffer am weitesten vorn anzeigen. Für den deutschen Raum ist die Suchmaschine Google dabei mit einem Marktanteil von knapp 90% quasi Monopolist und deren genaue Algorithmen und Filter sind ein wohlgehütetes Geheimnis. Die Erforschung dieser Algorithmen und der effektive Einsatz dieser Erkenntnisse zum Verbessern des Rankings einer Webseite wird als Suchmaschinenoptimierung bezeichnet. Das Thema dieser Arbeit ist die Analyse, Beschreibung und Gewichtung verschiedener Faktoren unter Berücksichtigung offizieller Aussagen von Google sowie empirischer Experimente im praktischen Umfeld.

Eidesstattliche Erklärung

Ich erkläre hiermit eidesstattlich, dass ich die vorliegende Arbeit selbstständig und ohne Benutzung anderer als der angegebenen Hilfsmittel angefertigt habe. Aus den benutzten Quellen direkt oder indirekt übernommene Gedanken habe ich als solche kenntlich gemacht. Diese Arbeit wurde bisher in gleicher oder ähnlicher Form oder auszugsweise noch keiner anderen Prüfungsbehörde vorgelegt und auch nicht veröffentlicht.

Karlsruhe, 1. Januar 2011	
Ort, Datum	Unterschrift

Inhaltsverzeichnis

Abbildungsverzeichnis

Quellcodeverzeichnis

Abkürzungsverzeichnis

AJAX	Asynchronous JavaScript and XML
CSS	Cascading Style Sheets
HTML	Hyper Text Markup Language
HTTP	Hyper Text Transfer Protocol
ODP	Open Directory Project
SEO	Search Engine Optimization
SERP	Search Engine Result Pages
URL	Universal Resource Locator
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

1 Einleitung

Es gibt eine Vielzahl von Suchmaschinen, von denen jedoch nur wenige wirklich relevant sind und benutzt werden. Im deutschsprachigen Raum ist die Suchmaschine Google trotz leicht rückläufiger Zahlen der Branchenprimus auf dem Suchmaschinenmarkt mit einer Suchanfragenabdeckung von knapp 90% [Fis09, S. 167–168]. Aus diesem Grund wird sich diese Arbeit speziell mit der Suchmaschine Google beschäftigten.

1.1 Gliederung

Die Arbeit beginnt mit einem kurzen Einstieg zur technischen Arbeitsweise von Google sowie einer Einführung der in [BP98] beschriebenen Grundprinzipien zur Berechnung des Rankings. Diese Grundprinzipien werden im Anschluss gegenüber der heutigen Arbeitsweise evaluiert.

Der Hauptteil der Studienarbeit befasst sich mit den konkreten Einflussfaktoren für das Ranking, die sich auf die beiden großen Bereiche OnPage Optimierung und OffPage Optimierung verteilen. In diesem Teil werden sowohl bekannte und bestätigte Aussagen zusammengefasst als auch empirisch ermittelte Daten beziehungsweise „Best Practices“ beschrieben und ausgewertet. Zum Schluss werden die ermittelten Ergebnisse zusammengefasst und in einem Fazit aufgearbeitet.

Nicht Bestandteil dieser Arbeit ist das Finden relevanter Suchbegriffe beziehungsweise Keywords.

1.2 Ziel der Arbeit

Ziel dieser Arbeit ist die Identifikation der Einflussfaktoren, die das Ranking einer Webseite bei Suchmaschinen bestimmen. Diese Faktoren sind nur zum Teil öffentlich dargelegt, da eine Suchmaschine möglichst resistent gegenüber Manipulationen bleiben muss, weil nur so eine objektive Beurteilung von Webseiten möglich ist. Neben der Identifikation werden die verschiedenen Faktoren erläutert, gewichtet und bezüglich ihrer Praxistauglichkeit bewertet.

1.3 Begriffe und Personen

In dieser Arbeit wird zum Teil auf Fachvokabular beziehungsweise spezielle Begriffe aus dem Bereich der Suchmaschinenoptimierung zurückgegriffen. Diese werden im Folgenden einge-

führt.

PageRank

Der ursprüngliche PageRankalgorithmus wird in Kapitel 2.4.1 genauer vorgestellt. Der dort vorgestellte Algorithmus wird in der heutigen Zeit nicht mehr in dieser Art und Weise verwendet. Wenn in diese Arbeit der Begriff PageRank benutzt wird, dann steht dieser in der Regel stellvertretend für eine ganze Reihe von Faktoren, die Google heutzutage zur Berechnung der Reputation einer Webseite heranzieht.

Linkpower

Mit Bezug auf den PageRank und die Eigenschaft, dass Webseiten PageRank über ihre Verlinkung von anderen Webseiten erben, wird der Begriff Linkpower für das Gewicht einer solchen Verlinkung benutzt. So besitzen zum Beispiel die Links auf Webseiten mit einem hohen PageRank eine größere Linkpower als solche auf Webseiten mit niedrigem PageRank.

SERPs

SERPs[1] bezeichnen die Ergebnisseiten, die aus einer Suche bei Google resultieren. Dieses Akronym ist in etwa mit dem Begriff Google Ranking beziehungsweise schlichtweg Ranking gleichzusetzen.

Keyword

Bei der Suchmaschinenoptimierung werden Webseiten für bestimmte Begriffe optimiert. Diese Begriffe entsprechen den Wörtern, die bei Google zum Suchen eingegeben werden. Der allgemeine Begriff für die Suchbegriffe lautet Keywords.

Link

Kurzform für Hyperlink

[1] Search Engine Result Pages

Backlink

Wenn sich auf Webseite A ein Hyperlink mit der URL[2] von Webseite B als Wert des `href` Attributs befindet, dann besitzt B einen Backlink von A. A verlinkt also zu B beziehungsweise B wird von A angelinkt.

Snippet

Bei der Darstellung der SERPs erzeugt Google in der Regel einen kurzen Textausschnitt, der den Inhalt der Webseite beschreibt. Dieser wird als Snippet bezeichnet und ist beispielhaft in Abbildung 1 (innerhalb der roten Umrandung) dargestellt.

Abbildung 1: Ein angezeigtes Ergebnis zum Suchbegriff Suchmaschinenoptimierung

Google Webmaster Tools

Durch das Anlegen eines Benutzerkontos bei Google hat man Zugang zu den Google Webmaster Tools. Dort kann man die eigene Domain eintragen und nach einer Verifizierung einige statistische Daten einsehen sowie einige Parameter einstellen. Die Google Webmaster Tools sind unter https://www.google.com/webmasters/tools/ zu erreichen.

Google Webmaster Guidelines

Google hat Richtlinien für Webseiten definiert, die bei Google gelistet werden sollen. Diese sind unter http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=35769 zu erreichen und werden auch als Google Webmaster Guidelines bezeichnet.

[2] Uniform Resource Locator

Matt Cutts

Matt Cutts ist der Leiter des Google Webspam Teams, das für die Qualität der Suchergebnisse zuständig ist. Er ist eine der wenigen Personen, die öffentliche Aussagen zu Googles eingesetzten Bewertungskriterien bekanntgeben.

1.4 Suchmaschinen

Das Internet zeichnet sich durch eine enorme Informationsvielfalt aus. Diese Informationen befinden sich an vielen unterschiedlichen Stellen und müssen auf Grund ihrer Masse sinnvoll geordnet werden. Dieses Problem bestand bereits zu Beginn des Internets und wurde mit dessen Entwicklung zum Web 2.0 und der Zunahme von usergenerierten Inhalten zunehmend größer. Es wurden verschiedene Lösungsansätze entwickelt, die im Folgenden erläutert werden.

1.4.1 Webkataloge

Webkataloge sind manuell gepflegte Verzeichnisse, die in verschiedene Kategorien unterteilt sind. Diese Kategorien können wiederum Unterkategorien besitzen, so dass sich eine hierarchische Struktur ergibt. In den einzelnen Kategorien werden die URLs zu passenden Ressourcen inklusive einer kurzen Beschreibung gespeichert. Die berühmtesten Vertreter in diesem Gebiet sind das Open Directory Project[3] und das Yahoo! Directory[4]. Allerdings haben diese Verzeichnisse viele Nachteile, da sie manuell gepflegt und erweitert werden müssen und diese Aufgaben mit dem zunehmenden Wachstum des Internets nicht mehr zu vereinbaren ist. Auch für einen User ist die Informationssuche mühsam, weil er über verschiedene Hierarchieebenen navigieren muss, bevor er zu einem Ergebnis kommt.

1.4.2 Volltextsuchmaschinen

Einen automatisierten Ansatz zur Organisation des Internets stellten die ersten Volltextsuchmaschinen zur Verfügung. Diese basierten auf dem automatisierten Einlesen von Webdokumenten und deren Speicherung in einem Index. Es konnten nun generische Suchanfragen gestellt werden, die mit dem Index abgeglichen wurden und als Resultat alle Dokumente lieferten, die zur Anfrage passten. Bei diesem Konzept entfällt die manuelle Pflege, aber in den meisten Fällen ist die Anzahl der passenden Dokumente schlicht zu groß um einen echten Nutzen für einen

[3] http://www.dmoz.org/
[4] http://dir.yahoo.com/

Suchenden darzustellen. Dieses Problem kommt zum Teil durch ein fehlendes Verständnis der Suchmaschinen für den Inhalt eines Webdokumentes zu Stande.

1.4.3 Einbeziehung von Meta Informationen

Meta Informationen sind „Informationen über Informationen“. Mit Hilfe dieser Informationen ist es möglich, Daten genauer zu beschreiben und somit maschinell verwertbare Informationen daraus herzustellen. Suchmaschinen können damit eine bessere Abschätzung des Inhaltes einer Webseite machen und dadurch die Ergebnismenge einer Suchanfrage einschränken. Die typischen Meta Informationen bei HTML Dokumenten werden innerhalb des `<head>` Tags notiert und sind unter dem Namen Meta Tags bekannt.

1.4.4 Einführung eines Rankings

Selbst Suchmaschinen, die Metadaten verarbeiten, können lediglich eine genauere Ergebnisliste liefern. Wie bereits erwähnt ist diese Liste aber meist immer noch viel zu groß um daraus manuell die Informationen zu extrahieren, die tatsächlich gesucht werden. Aus diesem Grund wurden die rankingbasierten Suchmaschinen entwickelt, die zusätzlich noch eine Bewertung der einzelnen Ergebnisse zu einer Abfrage vornehmen und damit eine Struktur erschaffen, die ein sinnvolles Arbeiten ermöglicht. Die größten Vertreter dieser Gattung im westlichen Raum sind Google, Yahoo! und Bing.

2 Grundlagen

Larry Page, einer der Gründer Googles, beschrieb die perfekte Suchmaschine als etwas, das genau versteht, was man sucht und auch genau das als Ergebnis liefert. Um diesen Anspruch zu erfüllen hat Google eine Technologie entwickelt, die sich auf die drei folgenden Bestandteile stützt:

1. Crawling
2. Indexing
3. Query Processing

2.1 Crawling

Google setzt sogenannte Webcrawler (oft auch Crawler oder Spider gennant) ein um Webseiten zu finden. Der Crawler von Google nennt sich Googlebot. Generell werden dabei nicht zufällig beliebige Webseiten abgerufen, sondern der Crawler arbeitet sich systematisch durch die Verlinkung von Webseiten. Von einer abgerufenen Webseite werden die Hyperlinks extrahiert und in einer Queue gespeichert. Diese Queue wird dann nach und nach abgearbeitet. Um Ressourcen zu schonen wird allerdings zuvor verglichen, welche Webseiten der Crawler bereits abgerufen hat.

In der heutigen Zeit werden zwei verschiedene Crawling Verfahren unterschieden, das Deep-Crawling und das Fresh-Crawling. Dabei entspricht das Deep-Crawling dem oben erklärten Verfahren, während das Fresh-Crawling für die Aktualität der abgerufenen Seiten verantwortlich ist. In diesem Fall werden also bereits bekannte Webseiten erneut gecrawlt um die neusten Änderungen darauf zu erkennen.

Die Ergebnisse des Crawling werden an den sogenannten Indexer übergeben, der im Folgenden erklärt wird.

2.2 Indexing

Das reine Sammeln von Webseiten bietet zunächst nichts anderes als die Archivierung von Informationen. Der Hauptzweck von Suchmaschinen ist jedoch das Suchen (und Finden) von Dokumenten. Da die Dauer dieses Prozesses mit einer steigenden Anzahl von Dokumenten ebenfalls ansteigt, muss eine Technik gefunden werden um diesen Prozess so effizient wie möglich zu gestalten. Aus diesem Grund legt Google für jede gecrawlte Webseite einen Index an, der

aus den einzelnen Wörtern des Dokumentes besteht. Der Index verknüpft ein Wort mit einem Dokument und kann von mehreren Servern parallel durchsucht werden.

Der Index selbst ist für suchende Zugriffe optimiert (Wörter werden zum Beispiel nur in Kleinschreibung gespeichert und alphabetisch sortiert). Die effiziente Anwendung dieses Verfahrens ermöglicht es Google, Suchanfragen in den Bruchteilen einer Sekunde zu beantworten, obwohl theoretisch mehrere Milliarden erfasste Webseiten durchsucht werden müssten.

2.3 Query Processing

Das Query Processing stellt die Schnittstelle von Google zu den Nutzern der Suchmaschine dar. Eine von einem Suchenden eingegebene Begriffsmenge wird von Google aufbereitet und an die Datenbank gesendet. Die Aufbereitung beinhaltet zum Beispiel die Entfernung von Stoppwörtern (zum Beispiel „und“, „in“, „die“, etc.).

Die Anfrage an die Index-Datenbank liefert nun alle Dokumente, die die gesuchten Begriffe enthalten. Diese Dokumentmenge bezeichnet man auch als „posting list“. Die wirkliche Leistung liegt darin, diese posting list so zu sortierten, dass sie die relevantesten Ergebnisse zu Beginn anzeigt. Dazu setzt Google laut [Goo] mehr als 200 Bewertungsfaktoren ein, die zum einen die Relevanz und zum anderen die Reputation einer Seite bewerten. Die Ergebnisse sind das, was man generell unter dem Begriff SERP zusammenfasst und was sich in der aufbereiteten Anzeige für den suchenden User manifestiert.

2.4 Anfängliche Ranking Grundsätze

Einige von Googles Ranking Faktoren werden in [BP98] beschrieben. Diese werden im Folgenden erläutert und es wird eine Evaluation im Bezug auf die heutige Relevanz dieser Faktoren vorgenommen. Die Faktoren sind:

1. PageRank
2. Anchor Text
3. Other Features

Dabei fallen PageRank und Anchor Text in den Bereich der OffPage Optimierung, die in Kapitel 4 genauer beschrieben, während sich die „Other Features“ auf den in Kapitel 3 beschriebenen Bereich der OnPage Optimierung beziehen.

2.4.1 PageRank

Der PageRank Algorithmus wird in [Pag+99] eingeführt und ist das wohl bekannteste Rankingkriterium von Google. Der Algorithmus ist nach dem Google Mitgründer Larry Page benannt und liefert ein Maß für die Relevanz von Webdokumenten basierend auf ihrer Reputation im Internet. Diese Reputation wird auf Basis der eingehenden Hyperlinks berechnet. Der Grundgedanke hinter diesem Prinzip der Bewertung ist die Betrachtung von Links als Empfehlungen und ist in etwa vergleichbar mit der Benutzung von Zitaten in der Literatur beziehungsweise in wissenschaftlichen Ausarbeitungen.

Der PageRank ist ein konkreter Wert, der nach der folgenden Formel berechnet wird:

$$PR(p_i) = \frac{1-d}{N} + d \sum_{p_j \in M(P_i)} \frac{PR(p_j)}{L(p_j)}$$

$PR(x)$ = PageRank von Webseite x,
d = Dämpfungsfaktor,
N = Anzahl aller Seiten im Index,
$L(x)$ = Anzahl der Hyperlinks auf Webseite x,
$M(P_x)$ = Menge aller Seiten, die auf Webseite x verlinken.

Der PageRank ist demnach ein Wert zwischen 0 und 1. Da der Algorithmus rekursiver Natur ist, wird der tatsächliche PageRank in mehreren Iterationen ermittelt. Der Dämpfungsfaktor d stellt sicher, dass Webseiten, die in einem Kreislauf aufeinander verlinken, keinen unendlichen PageRank bekommen. Für d wird ein Wert von ca. 0,85 empfohlen.

Eine weitere Betrachtung, die bei dem PageRank zum Tragen kommt, ist der sogenannte „Random Surfer“. Dieser Begriff bezeichnet einen User, der sich zufällig durch das Internet bewegt (zwischen verschiedenen Webseiten navigiert) und dabei von einer Seite zur nächsten kommt, indem er den Links auf einer Webseite folgt. Ab einer gewissen Stelle bricht der User den Vorgang ab und beginnt ihn auf einer zufällig gewählten anderen Webseite erneut. Bei diesem Modell entspricht die Wahrscheinlichkeit, dass ein User eine Webseite aufruft in etwa dem PageRank dieser Seite. In [Pag+99] wird das PageRank Prinzip simplifizierend wie in Abbildung 2 visualisiert. Die angezeigten Zahlen entsprechen dabei dem aktuellen PageRank einer Seite (Zahl steht innerhalb der Seite) beziehungsweise dem vererbten PageRank (Zahl steht am Pfeil).

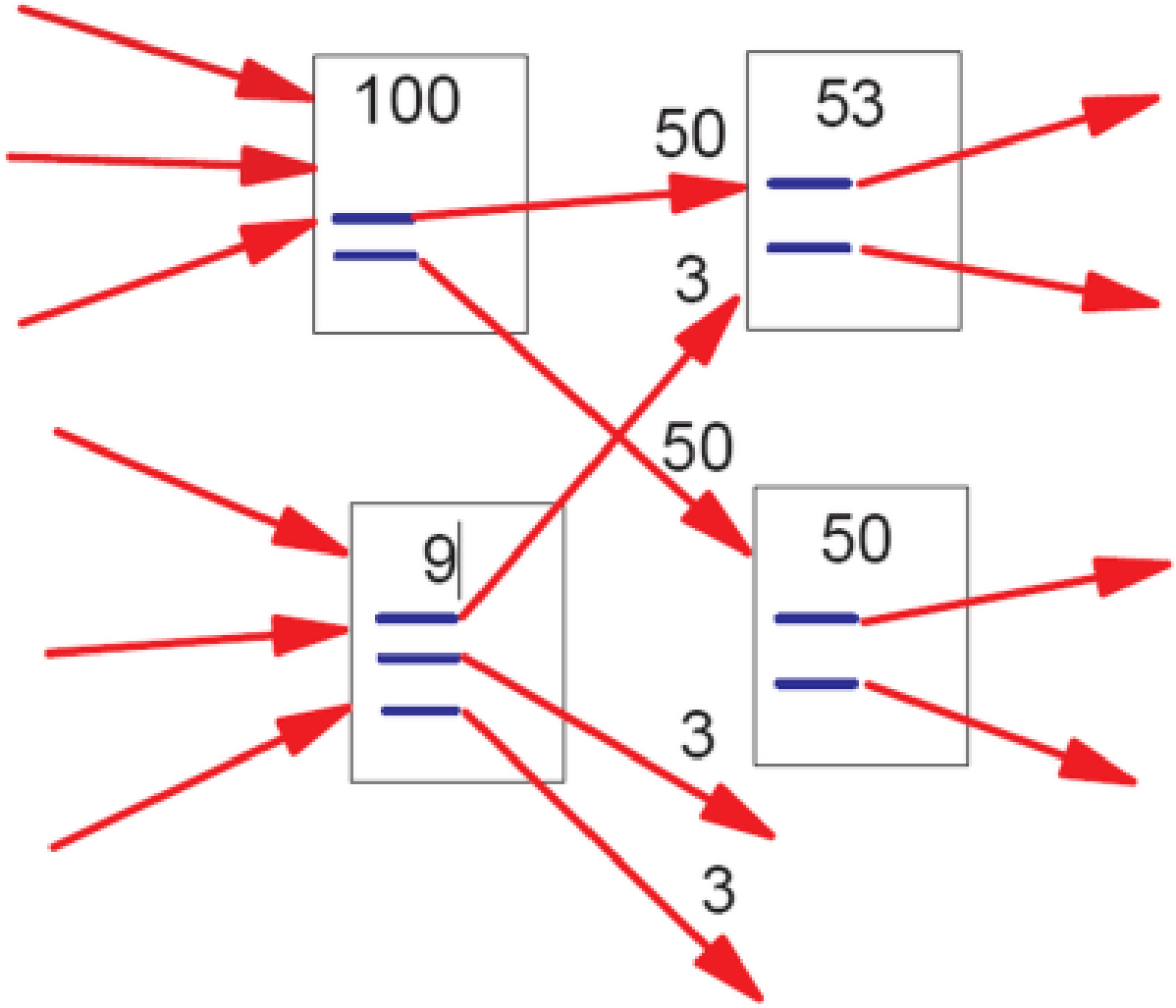

Abbildung 2: Gewichtung von Links bei dem Random Surfer Modell

Evaluation

Das PageRank Prinzip wird mit sehr hoher Wahrscheinlichkeit auch heut zu Tage noch eingesetzt. Allerdings nicht mehr exakt in der oben beschriebenen, ursprünglichen Form. Man geht davon aus, dass heute eher das sogenannte „Reasonable Surfer“ Modell, das in [DAB10] beschrieben ist, bei der Berechnung des PageRanks eingesetzt wird. Der große Unterschied in diesem Modell liegt darin, dass der navigierende User nun nicht mehr durch Zufall auf einen Link klickt, sondern dass das Verhalten des Users von bestimmten Faktoren abhängt. So ist es zum Beispiel wahrscheinlicher, dass ein User einem Link folgt, der einen thematischen Bezug zu der Webseite hat, auf der er sich gerade befindet. Weiterhin spielt die Platzierung des Links eine Rolle. Ein Link im Hauptinhalt (dem sogenannten Content) einer Seite wird mit großer Wahrscheinlichkeit häufiger angeklickt als ein Link im Footer. Abbildung 3 aus [DAB10] verdeutlicht die unterschiedliche Gewichtung. Maximal kann in diesem Beispiel ein Link den Wert

1 besitzen.

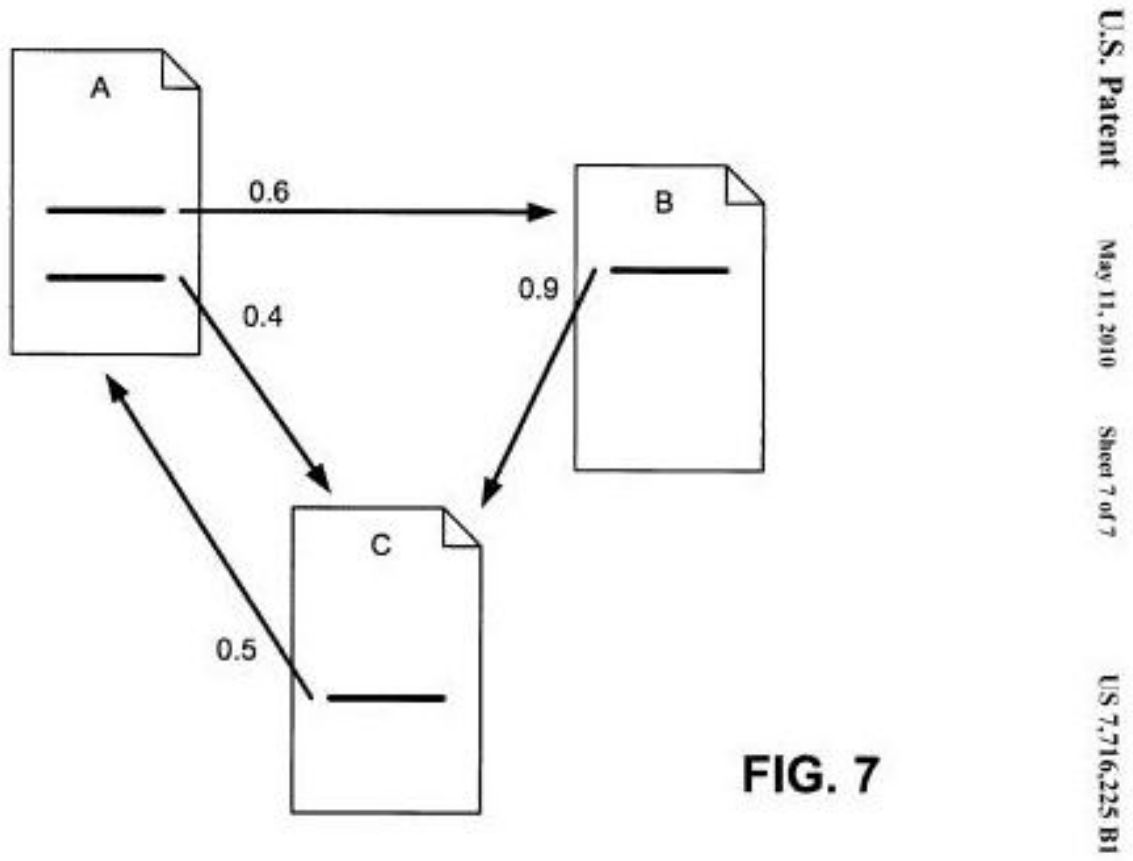

Abbildung 3: Unterschiedliche Gewichtung von Links bei dem Reasonable Surfer Modell

Eine weitere Modifikation bezüglich des PageRanks ist die Einführung des `nofollow` Attributs. Dieses Attribut wurde von Google 2005 in [Blo05] eingeführt um der zunehmenden Menge an Spam-Kommentaren in Blogs sowie dem aufkommenden Verkauf von Backlinks Einhalt zu gebieten. Dadurch sollte es beispielsweise für Webmaster möglich sein, einen Werbelink auf ihrer Seite zu platzieren ohne dabei den Google Algorithmus zu manipulieren. Links, die das nofollow tragen, vererben weder Ankertext (siehe nächstes Kapitel) noch PageRank.

Ein wesentliches Indiz dafür, ob und wie der ursprüngliche PageRank Algorithmus noch bei Google eingesetzt wird, könnte außerdem im Mai 2011 auftauchen. Zu diesem Zeitpunkt laufen die Nutzungsrechte Googles am PageRank Patent aus, dessen Inhaber die Universität St-

anford ist. Sollte der PageRank in seiner beschriebenen Form nicht mehr eingesetzt werden, könnte Google von dem Erwerb weiterer Nutzungsrechte absehen. Bereits Ende 2009 wurde die Anzeige des PageRanks aus den Google Webmaster Tools entfernt und von der Google Mitarbeiterin Susan Moskwa unter [Mos09] wie folgt kommentiert:

> We've been telling people for a long time that they shouldn't focus on PageRank so much; many site owners seem to think it's the most important metric for them to track, which is simply not true. We removed it because we felt it was silly to tell people not to think about it, but then to show them the data, implying that they should look at it. :-)

2.4.2 Anchor Text

Der Begriff „Anchor Text" lässt sich mit „Ankertext" beziehungsweise „Linktext" ins Deutsche übersetzen. Damit ist der im gerenderten HTML sichtbare, anklickbare Text eines Hyperlinks gemeint, der sich im Quelltext zwischen dem öffnenden und schließenden `<a>` Tag befindet. In dem Link in Quellcodeausschnitt 1 ist somit „Beispiel" der Ankertext.

```
<a href="http://www.example.com/">Beispiel</a>
```

Listing 1: Syntax eines Hyperlinks

Der Ankertext wird von Google als Ranking Kriterium mit einbezogen, weil er laut [BP98] zum einen oft eine genauere Beschreibung einer Webseite als der Text auf der Seite selbst liefert und zum anderen die Möglichkeit bietet, nicht textbasierte Inhalte wie Bilder oder Videos für Suchmaschinen mit einer erkennbaren Beschreibung auszustatten. Durch diesen Faktor wurde das sogenannte Google Bombing möglich, da Webdokumente für einen Begriff auf eine gute Position befördert werden konnten, obwohl sie die gesuchten Begriffe gar nicht beinhalten. Ein berühmtes Beispiel dafür waren die Suchergebnisse für die Suche nach „miserable failure", die am 2. Juni 2005 den Lebenslauf von George W. Bush als erstes Ergebnis lieferten. Zwar gab Google in einem offiziellen Blogpost [Blo07] bekannt, dass das Problem algorithmisch gelöst worden sei, aber der Einfluss des Ankertextes ist nach wie vor gegeben.

Evaluation

Der Ankertext wird auch heute noch als eines der wichtigsten Kriterien für das Ranking betrachtet. Einen einfachen Beweis dafür kann man nachvollziehen, wenn man nach dem Begriff „hier" sucht (Stand 28. November 2010). Bei 326.000.000 gefundenen Dokumenten wird die

Download Seite des Adobe Acrobat Readers als erstes Ergebnis zurückgeliefert, obwohl sie das Wort „hier“ kein einziges Mal enthält. Betrachtet man die Version der Webseite im Google Cache, so wird dort der Hinweis

> Diese Begriffe erscheinen nur in Links, die auf diese Seite verweisen: **hier**

angezeigt. Siehe dazu auch Abbildung 4.

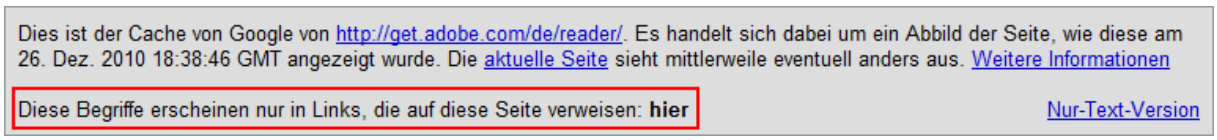

Abbildung 4: Hinweis zu Suchergebnissen, die den Suchbegriff nicht enthalten

2.4.3 Other Features

Unter „Other Features“ werden in der Ursprungsversion von Google 1998 die Faktoren „Keyword Proximity“ und „HTML Markup“ genannt. Unter Keyword Proxmity versteht man dabei die Nähe von Suchbegriffen innerhalb eines Dokumentes zueinander. Dabei wird die Indexposition im Quelltext des ersten Suchbegriffes mit der der weiteren Suchbegriffe verglichen. HTML Markup bezeichnet die syntaktische Textauszeichnung wie zum Beispiel die Schriftgröße und -Farbe.

Evaluation

Die oben genannten Faktoren stellen nur einen sehr geringen Bruchteil der Faktoren dar, die Google heut zu Tage auf einer Webseite selbst zur Berechnung des Rankings zu Rate zieht. Eine ausführliche Erläuterung der heute bekannten Faktoren wird in Kapitel 3 vorgenommen.

3 OnPage Optimierung

Unter dem Begriff OnPage Optimierung fasst man alle Maßnahmen zusammen, die man auf einer Webseite selbst durchführen kann um bei Suchmaschinen besser zu einem Begriff gelistet zu werden. Neben der reinen Nennung des Suchbegriffes verbergen sich dahinter noch eine Reihe weiterer Faktoren, die im Anschluss beschrieben werden.

3.1 Struktur und Aufbau einer Homepage

In diesem Abschnitt wird der grundsätzliche Aufbau einer Homepage behandelt. Die beschriebenen Maßnahmen sollten möglichst früh durchgeführt werden, noch bevor der eigentliche Inhalt der Domain erstellt wird. Da hierbei das Fundament der Homepage gelegt wird, können etwaige Fehler im Nachhinein nur schwer korrigiert werden.

3.1.1 Semantischer Aufbau von HTML Seiten

HTML Seiten haben syntaktisch bedingt eine zweigeteilte Gliederung in Head und Body. Im Head Bereich werden Meta Informationen notiert, auf die unter Punkt 3.2 näher eingegangen wird. Der Body definiert den für einen User sichtbaren Teil der Webseite. Dabei wird häufig (mindestens) eine Gliederung in die Bereiche

- Header
- Navigation beziehungsweise Menu
- Content
- Footer

vorgenommen. Abbildung 5 visualisiert diesen Aufbau exemplarisch.

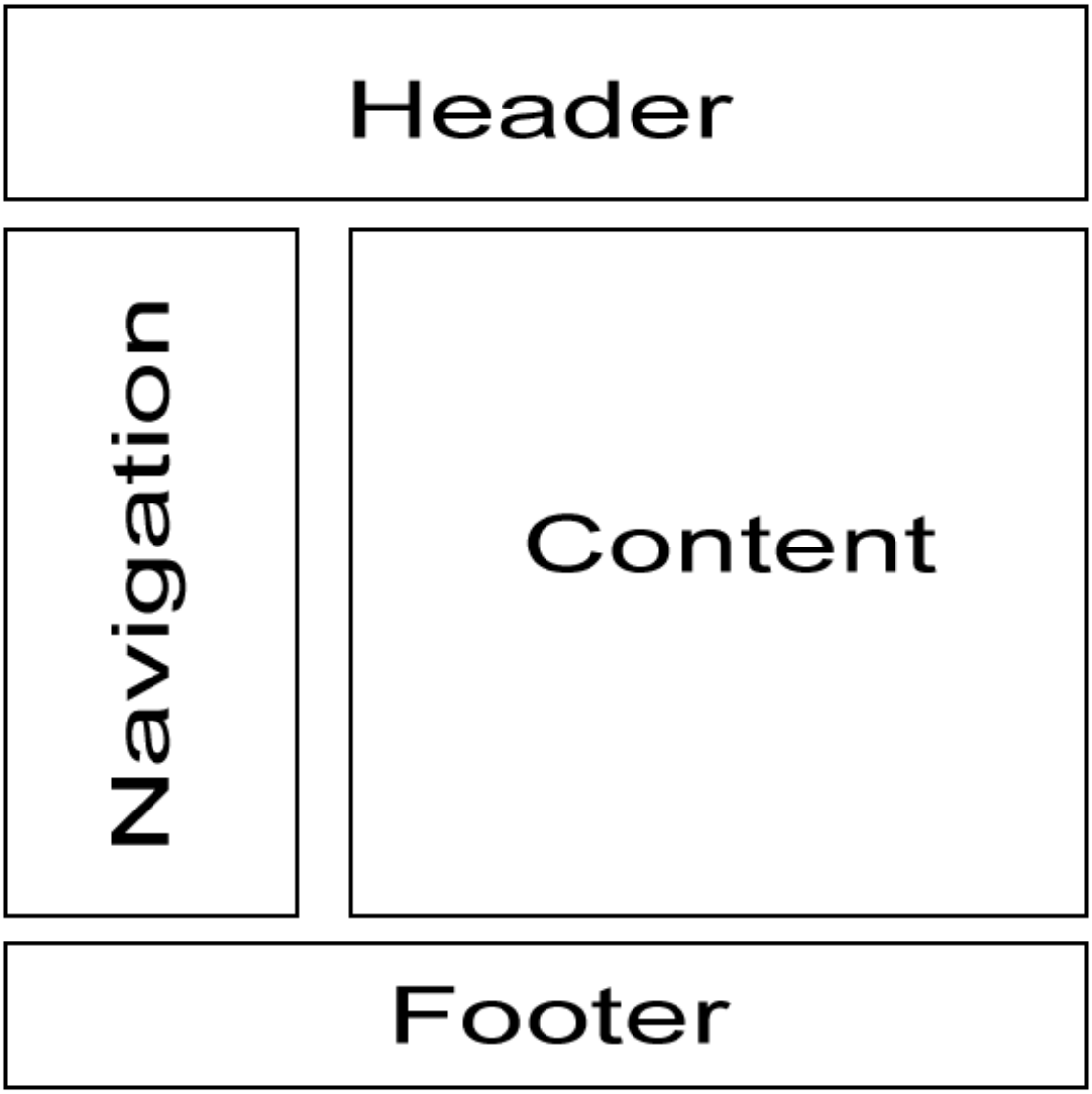

Abbildung 5: Exemplarischer Seitenaufbau von Webseiten

Header, Footer und Navigation enthalten dabei häufig sogenannten Boilerplate Code, also Code, der sich auf jeder Unterseite einer Homepage wiederholt. Die Navigation ist dabei das zentrale Navigationselement für den Benutzer, das die Hauptbereiche einer Homepage verlinkt. Der Content stellt den tatsächlichen Informationsträger dar.

3.1.2 Hierarchie und interne Verlinkungsstrukturen

Eine Homepage hat eine inhärente Gliederung in mindestens zwei Ebenen: Startseite und Unterseiten. Die Startseite kennzeichnet sich dadurch aus, dass sie aufgerufen wird, wenn allein der Domainname angegeben wird. Von der Startseite führen Links auf die Unterseiten der Domain, die dann wiederum untereinander verlinkt sein können. Bei einer größeren Anzahl von Unterseiten bietet sich ein mehrstufigere Gliederung an, so dass verschiedene Unterseiten in Kategorien zusammengefasst werden und lediglich die Kategorien von der Startseite aus verlinkt werden.

Die Hierarchie hat einen Einfluss auf die generelle(, keywordunabhängige) Gewichtung der einzelnen Webseiten einer Homepage. Das lässt sich mit dem PageRank Prinzip erklären, denn generell ist die Startseite einer Domain die am häufigsten verlinkte Seite. Die Power der Startseite wird an die verlinkten Seiten vererbt. Daraus lässt sich schließen, dass eine Unterseite mehr Power erhält, je kürzer sie von der Startseite über die interne Verlinkung zu erreichen ist. Wenn also einzelne Seiten besonders hervorgehoben werden sollen, sollte eventuell sogar in Betracht gezogen werden, diese direkt von der Startseite aus zu verlinken.

Bezüglich der internen Verlinkung nimmt die Navigation eine Sonderrolle ein, denn sie ist in der Regel in jeder Unterseite präsent. Die in der Navigation verlinkten Seiten werden demnach von jeder Unterseite aus verlinkt und tragen dementsprechend mehr Gewicht.

3.1.3 robots.txt

Bei der Datei `robots.txt` handelt es sich um eine Datei, die sich speziell an Webcrawler richtet. Sie trägt den primären Nutzen das Verhalten dieser Crawler zu steuern, was, genauer gesagt, aus dem Erlauben und Verbieten des Crawlens bestimmter Seiten (oder auch ganzer Verzeichnisse) besteht. Dazu werden die Direktiven allow beziehungsweise disallow verwendet.

Die definierten Regeln können entweder für alle Crawler gelten oder sich nur auf einen Bestimmten beziehen. Als Webseitenbetreiber kann man dadurch zum Beispiel Traffic sparen, indem man die Crawler unwichtiger Suchmaschinen aussperrt. Außerdem lassen sich damit zum Beispiel geschützte Verzeichnisse von der Aufnahme in den Google Index ausschließen.

Ein weiteres Einsatzgebiet dieser Datei ist die Lokalisierung der Sitemap einer Domain, die im folgenden Kapitel erläutert wird. In dem in Quellcodeausschnitt 2 gezeigten Beispiel wird das Crawlen aller Unterseiten und Verzeichnisse bis auf die Pfade `/admin/` und `/statistik/` erlaubt. Weiterhin wird die Sitemap unter `http://www.example.com/sitemap.xml`

lokalisiert. Diese Regeln gelten für alle Crawler (spezifiziert durch „*“). Sollen besondere Regeln für bestimmte Crawler gelten, müssen diese über die User Agent Direktive gekennzeichnet werden. Eine Liste der möglichen Robots ist unter http://www.robotstxt.org/db.html zusammengestellt.

```
User Agent: *
Disallow:   /admin/
            /statistik/
Sitemap:    http://www.example.com/sitemap.xml
```

Listing 2: Beispiel einer robots.txt Datei

3.1.4 Sitemap

Bei der internen Verlinkung der Unterseiten einer Homepage sollte vor allem darauf geachtet werden, dass alle Unterseiten erreichbar (also von mindestens einer anderen Stelle aus verlinkt) sind. Dadurch wird die Grundvoraussetzung für das Crawlen und Indexieren dieser Seiten geschaffen. Das allein ist jedoch noch kein Garant dafür, dass ein Crawler diese Seite auch besucht, da auch Google keine unbegrenzten Ressourcen zur Verfügung hat. Wenn eine Webseite nun sehr tief in der Hierarchie angesiedelt ist, dann kann es sehr lange dauern, bis es zu einer Indexierung kommt. Es gibt außerdem Situationen in denen eine interne Verlinkung nicht möglich ist, wie etwa bei einigen dynamischen Technologien wie AJAX[5] oder Flash. Aus diesem Grund gibt es die Möglichkeit eine Sitemap einzureichen. Dabei handelt es sich um eine Datei im XML[6] Format, die nach der in http://www.sitemaps.org/protocol.php definierten Syntax aufgebaut werden muss. In dieser Datei werden alle Seiten der Domain durch ihre URL aufgelistet und haben so eine größere Chance, schnell gecrawlt zu werden. Dabei ist die Anzahl der URLs auf 50.000 beschränkt, so dass eventuell mehrere Sitemaps angelegt werden müssen.

Um Google die Sitemap bekannt zu machen, kann diese entweder in der Datei `robots.txt` eingetragen oder direkt in den Google Webmaster Tools übermittelt werden.

3.1.5 PageRank Sculpting

Die Angaben zur Hierarchie und internen Verlinkung unter Punkt 3.1.2 haben bereits die Bedeutung einer wohlgeplanten Seitenstruktur verdeutlicht. Durch das sogenannte PageRank Sculp-

[5] Asynchronous JavaScript and XML
[6] eXtensible Markup Language

ting lässt sich die interne Verlinkung bezüglich der vererbten Linkpower noch weiter optimieren.

Die Tatsache, dass sich der PageRank einer Seite auf alle verlinkten Seiten aufteilt, legt nahe, dass es für die einzelnen angelinkten Seiten umso besser ist, je weniger Links sich auf der linkgebenden Seite befinden, da dann für die verbleidenden Links eine prozentual größere Power verbleibt. Da es aber unter Umständen nicht möglich ist die Links komplett zu entfernen, muss man auf diverse andere Techniken ausweichen.

Ein gutes praktisches Beispiel für verschwendete Linkpower ist die Verlinkung des Impressums, die in der Regel auf jeder Unterseite vorgenommen wird und zum Beispiel bei kommerziellen Seiten rechtlich vorgeschrieben ist. Allerdings soll ein Impressum nur in den seltensten Fällen ein gutes Ranking in Suchmaschinen erzielen, so dass man durchaus von einer Verschwendung der Linkpower sprechen kann.

Das Ziel des PageRank Sculptings ist also die Erhaltung der Funktionalität (beziehungsweise des Umfangs der Funktionalität) bei gleichzeitiger Optimierung der internen Verlinkungsstrukturen. Dies kann effektiv bisher allerdings nur auf zwei Wegen erreicht werden:

1. dynamische Erzeugung des Links mittels JavaScript
2. ausliefern von verschiedenen Quelltexten für Webcrawler und echte Besucher

Früher gab es noch die Möglichkeit, diesen Effekt mit der Kennzeichnung unwichtiger interner Links mit dem nofollow Attribut zu erreichen. Allerdings wurde im Juni 2009 von Matt Cutts offiziell in [Cut09b] verkündet, dass Links mit diesem Attribut zwar keine Linkpower vererben, aber dennoch dazu führen, dass die anderen Links auf der gleichen Webseite prozentual weniger Power vererbt bekommen.

Dynamische Erzeugung des Links mittels JavaScript

Bei dieser Möglichkeit wird durch den `Document.Write()` Befehl ein Hyperlink dynamisch erzeugt. Google ist jedoch laut eigenem Statement (siehe [Blo08]) in der Lage, JavaScript zu interpretieren:

> [...] We already do some pretty smart things like scanning JavaScript and Flash to discover links to new web pages, [...]

Deshalb muss der entsprechende Code in einer Funktion untergebracht werden, die dann mittels `robots.txt` vor den Zugriffen durch Webcrawler geschützt wird.

Der Nachteil dieser Methode besteht darin, dass ein Link dann für alle User mit deaktiviertem JavaScript nicht mehr erreichbar ist. Bezogen auf das oben genannte Impressumsbeispiel kann dies zu rechtlichen Problemen für kommerzielle Webseiten führen.

Ausliefern verschiedener Quelltexte
Google kennzeichnet seine Crawler durch eine besondere User Agent Kennung, anhand derer man ihn von normalen Usern und anderen Crawlern unterscheiden kann. Durch diese Unterscheidung ist es möglich, dem Crawler eine modifizierte Version einer Webseite zu präsentieren.

Generell fasst man dieses Vorgehen allerdings unter dem Begriff „Cloaking“ zusammen. Cloaking widerspricht den Google Webmaster Richtlinien und kann bei Entdeckung zum Ausschluss einer Webseite aus dem Index führen. Von dieser Praktik ist deshalb strikt abzuraten.

Fazit
Zusammenfassend kann man sagen, dass sich das PageRank Sculpting durch keine der genannten Methoden zufriedenstellend verwirklichen lässt. Die bisher beste Möglichkeit besteht darin, unwichtige Seiten soweit wie möglich zusammenzufassen (zum Beispiel das Impressum und die Kontaktmöglichkeiten). Diese Technik ist auch unter dem Begriff Link Konsolidierung bekannt. Weiterhin muss beachtet werden, dass das PageRank Prinzip inzwischen wahrscheinlich nicht mehr nach dem Random Surfer sondern nach dem Reasonable Surfer Modell arbeitet und man davon ausgehen kann, dass unprominent platzierte Links auch generell weniger Linkpower vererben. Für Unterseiten wie Impressum, Kontakt, etc. bietet sich also eine Platzierung im Footer einer Webseite an.

3.1.6 User Experience

Der Begriff „User Experience“ bezeichnet die Wahrnehmung eines Users bezüglich einer Webseite. Sie ist deshalb zum Beispiel durch Faktoren wie

- gut lesbare Texte
- durchdachte Menuführung
- Geschwindigkeit des Seitenaufbaus

geprägt. Die „User Experience“ ist ein Kriterium für das Ranking einer Webseite bei Google. Sie besitzt weniger Gewicht im Vergleich zu anderen Faktoren, sollte aber nicht außer Acht gelassen werden.

Google kann nicht exakt ermitteln, was eine gute und was eine schlechte User Experience ist, weil dazu zum einen sehr viele Faktoren zu berücksichtigen sind und zum anderen einige davon auf den persönlichen Präferenzen eines Users beruhen. So kann zum Beispiel ein User die Navigation auf der linken Seite bevorzugen während ein anderer sie lieber auf der rechten Seite hätte. Nichtsdestotrotz gibt es objektive Faktoren, die generell positiv oder negativ behaftet sind. Im Folgenden werden zwei Beispiele erläutert.

Antwortgeschwindigkeit
Im April 2010 hat Google in [Blo10] offiziell angekündigt, dass die Antwortgeschwindigkeit einen Einfluss auf das Ranking nimmt. Gleichzeitig wird dort aber bestätigt, dass dieser Einfluss bisher nur sehr wenige Suchanfragen (weniger als 1%) beeinflusst und bisher nur auf dem amerikanischen Markt getestet wird. Dennoch ist es nicht unwahrscheinlich, dass sich dieses Kriterium etablieren wird.

Navigierbarkeit
Dieser Punkt spiegelt sich in gewisser Weise in Punkt 3.1.2 wider. Allerdings liegt der Fokus in diesem Fall darauf, dass sich ein User gut auf einer Seite zurechtfindet. In den meisten Fällen korrelieren beide Navigationsstrukturen, da man zum Beispiel wichtige Inhalte sowohl für Suchmaschinen als auch für Besucher möglichst prominent (zum Beispiel auf der Startseite) platziert.

Ein konkretes Beispiel für eine gute Menuführung ist die sogenannte Breadcrumb Navigation, bei der auf jeder Seite eine Navigationshierarchie angezeigt wird, die den Navigationspfad von der Startseite bis zur aktuell angezeigten Webseite widerspiegelt. Der Einsatz einer Breadcrumb Navigation wird auch in [Inc] empfohlen.

3.2 Meta Informationen

Suchmaschinen bewerten nicht nur den eigentlichen Content einer Webseite sondern auch Meta Informationen über eine Webseite.

3.2.1 Titel

Der Titel einer Webseite wird im `<head>` Bereich einer HTML Seite im `<title>` Tag notiert. Dieser Tag nimmt unter den OnPage Faktoren einen relativ hohen Stellenwert ein. Der Titel sollte deshalb den Kerninhalt einer Seite in wenigen Worten beschreiben und dabei auf

jeden Fall das Hauptkeyword beinhalten, für das die Seite ranken soll. Dabei gilt generell der Grundsatz: So viele Wörter wie nötig, aber so wenig wie möglich. Zudem sollten wichtige Wörter am Anfang stehen. Idealerweise existiert jeder Titel nur ein einziges Mal innerhalb der Seiten einer Domain, was bei Nichteinhaltung als Warnung in den Google Webmaster Tools angezeigt wird.

Ein praktisches Beispiel soll an dieser Stelle die Unterscheidung zwischen einem gutem und einem schlechten Titel verdeutlichen. Angenommen, es soll ein Titel für eine Webseite gefunden werden, die von der Gliederung einer Studienarbeit handelt. Eine für diesen Inhalt passende Keyword Kombination wäre „Gliederung Studienarbeit“. Ein Beispiel eines schlechten Titels wäre „Ein Artikel, der beschreibt, wie man eine Studienarbeit gliedert“, da er viel zu lang ist und die wichtigen Keywords erst am Ende nennt. Ein guter Titel wäre dahingegen „Gliederung einer Studienarbeit“, da er kurz und prägnant ist und die wichtigsten Keywords enthält.

Bei der Wahl des Titels gilt es zu beachten, dass dieser als Link in den SERPs erscheint und damit direkten Einfluss darauf hat, ob ein Suchender darauf klickt oder nicht. Aus diesem Grund wäre (bezogen auf das obige Beispiel) ein Titel, der lediglich aus „Gliederung Studienarbeit“ besteht zwar optimal bezüglich Länge und Keywordpositionierung, wird aber eventuell seltener von einem Suchenden angeklickt, weil er sich intuitiv unvollständig anhört.

Fazit
Der Titel ist ein sehr wichtiges Instrument, das zum einen einen verhältnismäßig großen Einfluss auf das Ranking einer Webseite hat und das zum anderen ein Schnittstellenkriterium zum ersten Kontakt mit dem User darstellt.

3.2.2 Domainname

Der Name einer Domain ist einer der Rankingfaktoren von Google, wie man aus der Zusammenfassung eines Interviews mit Matt Cutts auf [And08] entnehmen kann:

> [...]Domain names are the primary way of mapping where domains are on the web and Matt expects that to continue. Domain names are important and inseparable going forward.
>
> Generic domains that users are likely to remember, will indeed carry more weight than others.[...]

Das macht vor allem dann einen Unterschied, wenn der Domainname eines der Hauptkeywords enthält. Solche Domains werden als Keyword Domains bezeichnet. Wenn die Domain einzig

und allein aus dem Keyword besteht, dann spricht man auch von einer Exact Match Domain. Bei Keywords, die aus mehr als einem Wort bestehen macht es (im Gegensatz zu den Pfadangaben in URLs) für Google keinen Unterschied, ob die verschiedenen Wörter durch ein Trennzeichen getrennt sind oder nicht.

Da der Rankingfaktor „Keyword Domain“ jedoch schon lange bekannt ist, gibt es kaum noch freie generische Domains. Abhilfe schafft hier entweder das Anhängen eines generischen Suffixes (wie zum Beispiel „24“, „info“, „club“, etc.) oder das Ausweichen auf eine generische TLD[7] wie .com, .net oder .org.

Laut eigener Aussage ist die Top Level Domain kein Rankingfaktor für Google. Problematisch kann es aber immer dann werden, wenn man eine länderspezifische Domain besitzt und auf dieser fremdsprachige Inhalte anbietet. Für den deutschen Sprachraum sollte demnach nach Möglichkeit auch eine .de Domain registriert werden.

3.2.3 Domainalter

In [Ach+] wird unter anderem das Registrierungsdatum einer Domain sowie das Datum des erstmaligen Auffindens einer Domain durch einen Crawler als mögliche Rankingfaktoren vorgestellt. In einem Screencast bestätigte Matt Cutts den Einfluss des Domainalters[8], machte aber zugleich deutlich, dass dies vorrangig auf neue (wenige Monate alte) Webseiten Einfluss hat. Ein möglicher Hintergrund ist hier die Bekämpfung von Spam, denn Spam Seiten werden bei Erkennung aus dem Google Index ausgeschlossen und verlieren damit ihre Daseinsberechtigung. Daher ist die Fluktuationsrate bei Domains, die für Spam missbraucht werden wesentlich höher als bei Nicht-Spam-Domains.

3.2.4 URL

URLs sind ebenfalls ein wichtiges Kriterium der OnPage Optimierung. Sie lokalisieren eine Webseite eindeutig und können wertvolle Informationen über die Seite, die sie repräsentieren, enthalten. Suchmaschinen werten den Text aus, der sich in einer URL befindet. Diese Auswertung kann zum Beispiel daran erkannt werden, dass Suchbegriffe bei der Darstellung der SERPs innerhalb der URL fett markiert sind. Deshalb sollten sich die Keywords, für die die Seite ranken soll, auch in der URL befinden. Es hat sich in der Praxis eingebürgert, eine ähnliche Wortwahl wie bei dem Titel zu benutzen, wobei jedoch Stoppwörter wie zum Beispiel

[7] Top Level Domain
[8] http://www.youtube.com/watch?v=-pnpg00FWJY, besucht am 31.12.2010

„und“, „der“, „die“, etc. vermieden werden sollten. Generell gilt auch hier die Regel: So kurz wie möglich und so lang wie nötig.

In URLs können allerdings im Gegensatz zum Titel nicht alle Zeichen verwenden werden. Stattdessen dürfen URLs nur die in [BLFM05] definierten Zeichen enthalten. Wenn in einer URL verschiedene Wörter getrennt werden sollen, dann empfiehlt Google gemäß [Cenc] die Verwendung des Minusszeichen, wobei jedoch auch unter anderem die folgenden Zeichen als Trennungszeichen erkannt werden: !, #, %, ', (,), *, +, [Komma], /, :, =, @.

URL Parameter

Bei einigen Content Management Systemen werden die Inhalte von Webseiten in einer Datenbank gespeichert und die URLs zu diesen Seiten werden dynamisch erzeugt. Dabei wird dann häufig die ID des jeweiligen Datenbankeintrages in der URL als Paramater übergeben, was zum Beispiel zu folgenden URLs führt:

- `http://www.example.com/artikel.php?id=123`
- `http://www.example.com/forum/thread.php?tid=5`
- `http://www.example.com/index.php?category=5&subcategory=3`

Gerade am letzten Beispiel sieht man sehr deutlich, dass eine Suchmaschine aus solchen URLs keinerlei Informationen extrahieren kann. Um dieses Problem zu lösen ohne dabei auf den Komfort dynamisch erzeugter Inhalte zu verzichten, gibt es grundsätzlich zwei Ansätze.

Zum einen kann man einen sogenannten Slug definieren. Damit ist eine Zeichenkette gemeint, die einen Datensatz eindeutig identifiziert. Dadurch kann nun statt der numerischen ID eine Zeichenkette übergeben werden. Diese Methode hat jedoch zwei Nachteile, denn zum einen sind auf Strings basierende Suchoperationen in Datenbanken langsamer als solche mit numerischer Basis und zum anderen führt eine Veränderung des Slugs im Nachhinein zur Nicht-Erreichbarkeit der Seite über eine ehemals bekannte URL.

Der andere Ansatz basiert auf dem Apache Modul mod_rewrite[9]. Dieses Modul ermöglicht die Auswertung von URLs anhand regulärer Ausdrücke, mit deren Hilfe eine URL auf bestimmte Parameter geparst wird, so dass diese dann an eine Webseite übergeben werden können. Dadurch ist es zum Beispiel möglich redundante Zeichen in einer URL unterzubringen.

Unter der Annahme, dass `http://www.example.com/artikel.php?id=123` einen Artikel zum Thema Suchmaschinenoptimierung enthält, wäre es sinnvoller, wenn die URL `http://www.example.com/artikel/suchmaschinenoptimierung.html` (oder

[9] http://httpd.apache.org/docs/1.3/mod/mod_rewrite.html

ähnlich) lauten würde. Diese Form lässt sich nicht ganz erreichen, da zumindest die ID des Datenbankeintrages enthalten sein muss. Eine resultierende URL könnte aber beispielsweise `http://www.example.com/artikel/suchmaschinenoptimierung,123.html` lauten. Es bietet sich hier an die ID hinten zu nennen, da die Gewichtung der Keywords wie auch beim Titel von vorne nach hinten abnimmt.

Um dieses Beispiel lauffähig zu machen muss eine .htaccess Datei angelegt werden, die das mod_rewrite Modul aktiviert und eine entsprechende Regel definiert. Quellcodeausschnitt 3 zeigt dies für das oben genannte Beispiel. Das einzige Problem dieser Lösung besteht darin, dass

```
#Modul aktivieren
RewriteEngine On
#Regel definieren
RewriteRule ^artikel/.*,(.*)\.html$    artikel\.php?id=$1 [QSA]
```

Listing 3: mod_rewrite in einer .htaccess Datei einsetzen

eine Webseite nun über mehrere URLs zugreifbar ist, denn jedes Zeichen nach dem Vorwärts-Slash und vor dem Komma ist hier beliebig. Für Suchmaschinen ist eine andere URL allerdings gleichbedeutend mit einer anderen Webseite, so dass hier eine Duplicate Content Problematik entsteht.

3.2.5 Duplicate Content

Als Duplicate Content bezeichnet man den Inhalt von Dokumenten im Web, die dem Inhalt anderer Dokumente stark ähneln beziehungsweise sogar gleichen, wobei ein Dokument dabei genau von einer URL identifiziert wird. Für Benutzer von Google macht es keinen Sinn, die gleiche Information mehrfach angezeigt zu bekommen. Deshalb versucht man Duplikate zu erkennen und bei der Berechnung der SERPs auf ein einziges, repräsentatives Ergebnis zu beschränken. Auf der letzten Seite der Suchergebnisse finder sich deshalb häufig der Hinweis

> Um Ihnen nur die treffendsten Ergebnisse anzuzeigen, wurden einige Einträge ausgelassen, die den {n} bereits angezeigten Treffern sehr ähnlich sind.

Probleme

Für eine Homepage können sich aus diesem Feature Googles einige Probleme ergeben, wenn diese Homepage Seiten beinhaltet, die den gleichen Inhalt über verschiedene URLs ausliefern. Das prominenteste Beispiel ist die Erreichbarkeit einer Seite mit der beziehungsweise ohne die Eingabe der www. Standardsubdomain:

- `http://www.example.com`
- `http://example.com`

In vielen Fällen ist zusätzlich noch eine Standarddatei definiert, die bei der Eingabe des Domainnamens angezeigt wird. Unter der Annahme, dass es sich dabei um die Datei index.html handelt, ist zusätzlich noch die URL `http://www.example.com/index.html` erreichbar und liefert den gleichen Content.

Das eigentliche Problem dabei ist die Tatsache, dass alle Versionen eigene, eingehende Links besitzen können. Es macht aber keinen Sinn, diese Links auf mehrere URLs aufzuteilen, sondern die Power der Links in einer einzigen URL zu konsolidieren. Zwar unternimmt Google Versuche, diese Konsolidierung automatisch vorzunehmen, aber es gibt keine Garantie dafür, dass dies auch in allen Fällen funktioniert.

Ein weiteres Problem dabei ist die von Google bevorzugte Variante der angezeigten URL. So könnte es zum Beispiel sein, dass Google die Domain in den Suchergebnissen ohne www. darstellt obwohl diese zum Beispiel auf Visitenkarten etc. stets mit www. gedruckt wird. Bevor diverse Lösungen zu der Duplicate Content Problematik vorgestellt werden, folgt zunächst ein kurzer Abschnitt zu Situationen, in denen Duplicate Cotent häufig auftritt.

Druckversionen und Versionen für mobile Geräte

Eine HTML Seite eignet sich nur bedingt für den Druck, da sie in der Regel Elemente enthält, die keine für einen Ausdruck nützlichen Informationen zur Verfügung stellen (Navigation, Footer, etc.). Deshalb macht es aus Sicht der Usability Sicht Sinn, eine gesonderte Druckversion bereitzustellen. Diese muss irgendwie aufrufbar sein und besitzt demnach eine eigene URL. Gleichzeitig hat sie aber den gleichen Inhalt wie die Nicht-Druckversion, so dass eine Duplicate Content Situation entsteht.

Das gleiche Problem ergibt sich bei Homepages, die spezielle, für mobile Endgeräte optimierte Versionen einer Webseite zur Verfügung stellen. Wie auch bei der Druckversion werden hier einige Elemente der „originalen“ Webseite nicht auftauchen, aber der Contentbereich bleibt der Gleiche.

Navigationsstrukturen

Bei Webseiten mit mehrdimensionalen Navigationsstrukturen tritt häufig das Problem auf, dass eine Zielseite über verschiedene Pfade zu erreichen ist. Gerade bei Webshops gibt es häufig eine Einteilung in hierarchische Kategorien, wobei untergeordnete Kategorien zu verschiedenen Oberkategorie zugeordnet sind. Im englischen wird diese Art der Navigation als „faceted navigation“ bezeichnet. Ein Beispiel dazu:

- `http://www.example.com/shop/moebel/stuehle/holzhocker.html`
- `http://www.example.com/shop/material/holz/holzhocker.html`

Auch hier steht der Gedanke der Usability im Vordergrund, da das Produkt „Holzhocker“ verschiedenen Kategorien zugeordnet ist und auch auf den verschiedenen Kategorieseiten gelistet werden sollte. Nichtsdestotrotz entstehen dadurch unterschiedliche URLs mit dem gleichen Inhalt.

Parameterübergaben

Einer URL können nach der Angabe des Pfades, eingeleitet durch ein Fragezeichen, Parameter übergeben werden. URLs mit gleichen Pfaden aber unterschiedlichen Parametern werden wiederum von Suchmaschinen als eigenständige URLs gewertet. Dieses Verhalten macht Sinn, wenn man die unter Punkt 3.2.4 genannten Beispiele bedenkt, bei denen der angezeigte Inhalt dynamisch aus einer Datenbank geladen und dabei durch einen ID-Parameter identifiziert wurde.

Parameter werden aber auch zu anderen Zwecken benutzt, wie zum Beispiel zur Identifizierung einer Session. Bei Benutzern, die Cookies deaktivieren wird dabei die Session ID an jeden internen Link auf einer Webseite angehangen. Dadurch entstehen URLs wie beispielhaft in der folgenden Liste dargestellt:

- `http://www.example.com/?sid=0011e714c078160254e7374a476ab188`
- `http://www.example.com/?sid=be30908222d1a60fd8cf7800cfcaa7c7`
- `http://www.example.com/?sid=e6948c99d871291d0abd4bdbf4d5c7eb`

Jede dieser unterschiedlichen URLs hat den gleichen Inhalt.

Redundante Informationen in URLs

Die vorgestellte Lösung zur Optimierung der URL Strukturen mittels redundanten Informationen hatte das Problem, dass eine Seite über mehrere URLs zugreifbar war. Auch das führt dazu, dass der gleiche Seiteninhalt über verschiedene URLs aufrufbar ist.

Lösungen

Es gibt verschiedene Lösungsstrategien für Duplicate Content Probleme auf der eigenen Webseite, von denen sich drei als praktikabel erwiesen haben. Diese werden im Folgenden evaluiert.

Verwendung des Noindex Meta Tags

Auf Meta Tags allgemein wird unter Punkt 3.2.6 noch genauer eingegangen. Der Noindex Meta Tag signalisiert einer Suchmaschine, dass die damit ausgezeichnete Seite nicht in den Google

Index aufgenommen werden soll. Dadurch wird dem Problem entgangen, dass Google selbständig eine Seite auswählt, die in den SERPs angezeigt wird. Allerdings hat diese Anwendung einen Nachteil, denn sie missachtet die Link Konsolidierung. Gerade bei dem durch verschiedene Navigationsstrukturen entstehenden Duplicate Content kann es gut möglich sein, dass auf verschiedene URLs von außen verlinkt wird. Diese geteilte Linkpower wäre damit verschwendet. Der Einsatz dieses Tags ist also auf diejenigen Bereiche beschränkt, die niemals von außen verlinkt werden. Da das aber niemals zu 100% ausschließbar ist, ist im Normalfall die Verwendung des Canonical Tags vorzuziehen.

Verwendung des Canonical Tags

Der Canonical Tag wurde im Februar 2009 in [Blo09b] vorgestellt und adressiert Duplicate Content Probleme. Der Tag wird im `<head>` Bereich einer HTML Seite notiert und hat die in Quellcodeausschnitt 4 vorgestellte Syntax.

```
<link rel="canonical" href="http://www.example.com/canonical-target" />
```

Listing 4: Syntax des Canonical Tags

Der Tag wird auf derjenigen Seite notiert, die *nicht* in den Ergebnissen der Suchmaschinen auftauchen soll und enthält als `href` Attribut die URL, die beim Ranking angezeigt werden soll. In der Wirkung ergibt sich daraus der gleiche Effekt wie bei der Verwendung des Noindex Meta Tags wobei jedoch Ankertext und PageRank mit einem geringen Malus an das kanonikalisierte Ziel übergeben werden. Der Einsatz des Canonical Tags bietet sich für alle Duplicate Content Probleme an, bei denen die verschiedenen Inhalte zwingend über eigene URLs erreichbar sein müssen, weil sie einen eigenen Zweck erfüllen. Das ist zum Beispiel bei Druckversionen oder Faceted Navigations der Fall.

301 Redirects

Die sauberste und eindeutigste Art Duplicate Content zu bereinigen, ist das Antworten mit einem HTTP Status Code 301 (Moved Permanently) auf den Seiten, die nicht in den SERPs erscheinen sollen. Dabei wird in der HTTP Header Anweisung `location:` eine absolute URL angegeben, die die Ressource lokalisiert, an der sich der angeforderte Inhalt befindet. Wie auch bei dem Einsatz des Canonical Tags wird hierbei Ankertext und PageRank mit einem gewissen Malus an das Ziel der Weiterleitung weitergegeben. Diesen Malus bestätigte Matt Cutts in einem Interview mit Eric Enge auf [Eng10] wie folgt:

> [...]Matt Cutts: That's a good question, and I am not 100 percent sure about the answer. I can certainly see how there could be some loss of PageRank. I am not 100 percent sure whether the crawling and indexing team has implemented that sort of

> natural PageRank decay, so I will have to go and check on that specific case. (Note: in a follow on email, Matt confirmed that this is in fact the case. There is some loss of PR through a 301).[...]

Diese sogenannten 301 Redirects werden vorrangig zur Bewältigung der mit-oder-ohne-www-Problematik eingesetzt und eignen sich vor allem auch dann, wenn Inhalte einer alten Domain unter einer neuen verfügbar gemacht werden sollen.

Einstellungen in den Google Webmaster Tools

Die Google Webmaster Tools bieten Einstellungsmöglichkeiten für zwei Duplicate Content Probleme:

1. Erreichbarkeit mit und ohne www
2. Parameter

Diese Einstellungsmöglichkeiten sind aber zum einen Google spezifisch (und zählen somit nicht für andere Suchmaschinen) und haben zum anderen keine Vorteile gegenüber den oben genannten Möglichkeiten. Sie sind hier nur aus Gründen der Vollständigkeit aufgeführt.

Abbildung 6: Einstellungen zur Parameterbehandlung in den GWT

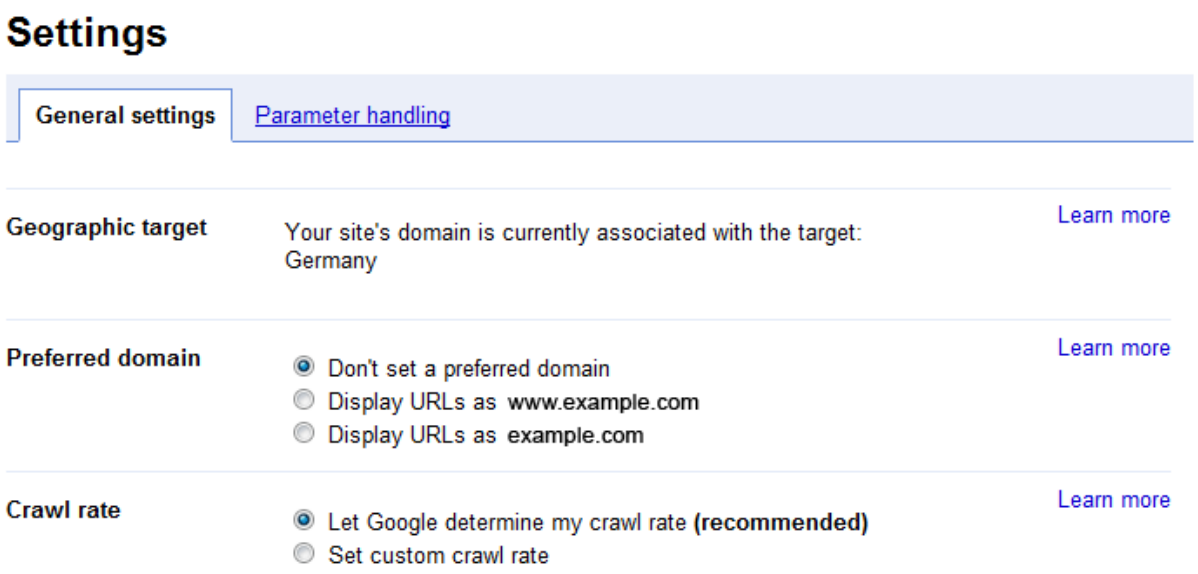

Abbildung 7: Einstellungen zur bevorzugten Domain in den GWT

3.2.6 Meta Tags

Meta Tags werden im `<head>` Bereich einer HTML Seite definiert und geben zusätzliche Informationen zu der Seite an. Für Google haben nur wenige der möglichen Meta Tags eine Bedeutung. Diese werden in [Cenb] genannt und im Folgenden erörtert.

Description

Der Description Meta Tag ist als kurze Zusammenfassung für den Inhalt einer Seite gedacht und sollte für jede Seite einer Domain einzigartig sein. Dieser Tag hat laut [Blo09a] keinen Einfluss auf das Ranking einer Webseite. Er wird jedoch unter Umständen für die Generierung des Snippets verwendet. Allerdings gibt es dafür keine Garantie, denn Google passt das Snippet der Suchanfrage an.

Keywords

Der Keywords Meta Tag hat keinen Einfluss auf das Ranking einer Webseite in den SERPs und hat auch sonst keinerlei Bedeutung für Google. Er wird an dieser Stelle nur deshalb erwähnt, weil es ein weit verbreitetes Gerücht ist, dass dieser Tag das Ranking einer Seite beeinflussen kann. Es gibt diverse Screencasts von Matt Cutts[10][11], die das widerlegen.

[10] http://www.youtube.com/watch?v=jK7IPbnmvVU, besucht am 31.12.2010
[11] http://www.youtube.com/watch?v=_euoDRk1qN0, besucht am 31.12.2010

Robots

Der Robots Meta Tag steuert das Verhalten von Suchmaschinen auf der Seite, auf der er platziert wird. Er kann die folgenden Werte annehmen:

noindex	Verhindert das Indexieren einer Seite
nofollow	Verhindert, dass Hyperlinks für weitere Crawlingvorgänge verwendet werden
nosnippet	Verhindert die Anzeige eines Snippets
noodp	Verhindert, dass der Beschreibungstext des ODP[12] (sofern vorhanden) als Snippet verwendet wird
noarchive	Verhindert, dass Google eine Version dieser Seite im Cache behält
unavailable_after:[date]	Verhindert das Crawlen ab dem durch [date] spezifizierten Datum
noimageindex	Verhindert, dass Bilder dieser Seite bei der Google Bildersuche auftauchen

Keiner dieser Werte hat auf das Ranking Einfluss. Falls der Tag nicht gesetzt ist, wird per Default von dem Wert „index, follow“ ausgegangen. Dieser besagt, dass die Seite indexiert werden soll und dass die verlinkten Webseiten gecrawlt werden dürfen.

Refresh

Der Refresh Meta Tag bewirkt eine Weiterleitung, die von Google wie eine 301 Weiterleitung behandelt wird. Der Tag ist jedoch vom W3C[13] als veraltet eingestuft und sollte nicht mehr verwendet werden.

3.3 Content - Der Inhalt einer Webseite

Der Inhalt ist der für Suchmaschinen interessanteste Teil einer Webseite, da er potenziell die meisten Informationen enthält. Eine der häufigsten Empfehlungen bei der Suchmaschinenoptimierung (und einer der Leitsätze von Matt Cutts) lautet sinngemäß wie folgt: „Produziere guten, qualitativ hochwertigen Content, der für Menschen gemacht ist, dann kommt der Erfolg bei Suchmaschinen von allein“.

[13] World Wide Web Consortium

Diese Aussage mag zu den Anfangszeiten des Internets in dieser Form gestimmt haben, ist aber mit der Entwicklung des Web 2.0 so nicht mehr haltbar. Das liegt zum einen daran, dass es kaum noch ein Thema gibt, zu dem es keine Informationen im Internet gibt und zum anderen daran, dass die Qualität des Contents allein keinen Einfluss auf das Ranking hat. Erst die Empfehlung in Form von Hyperlinks macht den Inhalt einer Webseite auch für Google qualitativ hochwertig. Zusätzlich gibt es noch weitere Faktoren, die bei der Erstellung von Content beachtet werden müssen, weil sie Einfluss auf das Ranking haben. Ein Zitat von Matt Cutts auf [Cena] liefert einen interessanten Ansatz, der als genereller Anhaltspunkt für die Aufbereitung von suchmaschinenoptimiertem Content dient:

> Pretend that you're a search engine. Pick a query like civil war or recycling or whatever you want. Search for the phrase on Google, pick three or four pages from the results, and print them out. On each printout, find the individual words from your query (such as „civil" and „war") and use a highlighter to mark each word with color. Do that for each of the 3-5 documents that you print out. Now tape those documents on a wall, step back a few feet, and squint your eyes. If you didn't know what the rest of a page said, and could only judge by the colored words, which document do you think would be most relevant? Is there anything that would make a document look more relevant to you? Is it better to have the words be in a large heading or to occur several times in a smaller font? Do you prefer it if the words are at the top or the bottom of the page? How often do the words need to appear? See if you can come up with 2-3 things you would look for to see if a document matched a query well. This can help students learn to evaluate website relevance the way a search engine would evaluate it so that they can better understand why a search engine returns certain results over others.

3.3.1 Keyword Density

Im optimalen Fall behandelt eine Webseite genau ein Keyword und ist auch genau auf dieses Keyword optimiert. Die Häufigkeit, mit der dieses Keyword im Content einer Webseite vorkommt, ist einer der Ranking Faktoren für Google. Als konkreten Wert zieht man hierbei die sogenannte Keyword Density zu Rate, die das Verhältnis zwischen der Anzahl der Vorkommen des Keywords zur Anzahl aller Wörter auf der Webseite angibt. Es gibt keinen von Google offiziell bestätigten Wert für eine gute Keyword Density. Empfohlen werden aber zum Beispiel in [Fis09, S. 311] Werte zwischen drei und vier Prozent. Allerdings ist bereits die Berechnung der Keyword Densitiy nicht ganz unproblematisch, denn es ist nicht bekannt, ob der komplette Text (inklusive Boileplate Code) auf einer Webseite herangezogen wird oder wie zum Beispiel

Stoppwörter gehandhabt werden.

Als Ziel sollte man sich aber auf jeden Fall setzen, dass das Keyword am häufigsten von allen Wörtern (ausgenommen Stoppwörtern) vorkommt um dadurch einer falschen Einschätzung des Themas der Seite durch Suchmaschinen vorzubeugen. Dabei gilt es jedoch zu beachten, dass die Keyword Density nicht zu groß wird, da dies einen Manipulationsversuch signalisieren könnte. Dieser wird auch als Keyword Stuffing bezeichnet und widerspricht gemäß [Cen10] den Google Webmaster Guidelines.

3.3.2 Keyword Proximity und Keyword Positioning

Bei Suchbegriffen, die aus mehr als einem Keyword bestehen, spielt die Nähe (engl.: Proximity) und die Reihenfolge dieser Begriffe eine Rolle bei der Bewertung. Je näher die Begriffe beieinander liegen, desto höher die Relevanz. Die Positionierung des Keywords bezieht sich nicht etwa darauf, ob das Keyword weit oben oder unten im Quelltext einer Seite steht, sondern darauf, ob es sich an einer prominenten (zum Beispiel im Content) oder weniger prominenten Stelle (zum Beispiel im Footer) befindet.

3.3.3 Semantische Relevanz

Ein großes Problem bei der Bewertung von Dokumenten ist das Erkennen der Thematik, die dieses Dokument beschreibt. Diese Problematik wird offensichtlich, wenn man zum Beispiel Wörter mit mehreren Bedeutungen („Teekesselchen“) als Beispiel nimmt. Ohne zusätzliche Informationen kann eine Suchmaschine nicht unterscheiden, welche Bedeutung ein Wort auf einer Seite einnimmt.

Um dieses Problem zu umgehen verwendet Google vermutlich Algorithmen, die die semantische Nähe von Wörtern berechnen. Einen Anhaltspunkt dafür findet man zum Beispiel in [Pat08]. Dort wird ein Verfahren vorgestellt bei dem Wortgruppen im Zusammenhang mit einem Keyword oder einer Kombination von Keywords untersucht werden. Wendet man dieses Verfahren auf entsprechend viele Dokumente an, so lassen sich diese Dokumente anhand der Wortgruppen in Clustern zusammenfügen und ermöglichen dadurch eine Kategorisierung des Inhaltes. Für einen suchmaschinenoptimierten Text bedeutet das, dass nicht nur die Nennung des Keywords für das Ranking von Bedeutung ist, sondern auch der Kontext in dem dieses Wort steht.

3.3.4 Interne und externe Verlinkung

Mit dem Aufbau einer Webseite zu einem bestimmten Thema kann es gut sein, dass sich überschneidende Unterthemen auf verschiedenen Unterseiten behandelt werden. In diesem Zusammenhang sollte man unbedingt Gebrauch von der internen Verlinkung der eigenen Webseite machen. Das hilft zum einem dem menschlichen Benutzer, weil er sich über ein angrenzendes Thema weiter informieren kann, und hat zum anderen auch einen positiven Effekt für das Ranking in Suchmaschinen (für beide Unterseiten). Für die Unterseite, die angelinkt wird, ergibt sich ein Vorteil, weil ihr auf diese Weise Linkpower zufließt. Der positive Effekt ist also direkt mit dem bekannten PageRank Prinzip verknüpft. Die maximale Effizienz ergibt sich dabei, wenn als Linktext das Keyword gewählt wird, für das die angelinkte Seite ranken soll. Aber auch die Unterseite, auf der der Link platziert ist, wird positiv im Ranking beeinflusst, denn Google bewertet auch die ausgehenden Links einer Webseite. Man spricht dabei generell von „Good Neighborhood“ und „Bad Neighborhood“.

Good Neighborhood
Unter der „guten Nachbarschaft“ versteht man verlinkte Webseiten, die sich durch positive Eigenschaften auszeichnen. Man kann die Beurteilung von „positiv“ dabei auf zwei Arten vornehmen. Zum einen kann man als Mensch entscheiden, ob eine Webseite qualitativ hochwertige Informationen enthält und diese benutzerfreundlich aufbereitet sind. Zum anderen kann man sich auf eine Beurteilung der Suchmaschinen stützen. Diese spiegelt sich im Ranking wider. Sucht man also zu dem Thema, über das man gerade berichtet, nach weiteren Informationen bei Google, so bieten die SERPs einen guten Anhaltspunkt dessen, was Google als „gute Nachbarschaft“ betrachtet.

Bad Neighborhood
Eine Webseite kann negativ beeinflusst werden, wenn auf ihr gegen die Google Webmaster Guidelines verstoßen wird. Das kann zum Beispiel durch das mutwillige Verstecken von Texten oder durch den Einsatz von Spamtechniken der Fall sein. Gleiches gilt für infizierte Webseiten. Diese sogenannten Bad Sites werden zum einen selbst im Ranking herabgestuft und haben gleichzeitig einen negativen Effekt auf das Ranking der Seiten, die zu ihnen verlinken.

Fazit
Google setzt nach eigenen Angaben immer den Benutzer in den Mittelpunkt. Es führt generell zu einer guten User Experience, wenn eine Webseite auf weitere, relevante Informationen

verlinkt, wohingegen deplatzierte Links zu themenfremden Inhalten entweder einfach ignoriert werden oder im schlimmsten Fall einen negativen Effekt haben. Es sollte daher Gebrauch von relevanter interner und externer Verlinkung gemacht werden, aber bei externen Links mit Vorsicht bezüglich schlechter Nachbarschaft vorgegangen werden.

3.4 Syntaktische Auszeichnung

In diesem Kapitel geht es um die Verwendung von HTML Syntax zur Optimierung des Textes auf einer Webseite. Bereits in [BP98] werden gesondert ausgezeichneten Wörtern eine höhere Bedeutung zugemessen und Google hat sich seit dieser Zeit beständig weiterentwickelt.

3.4.1 Validität

Entgegen einiger Gerüchte ist die Validität von Webseiten gemäß dem W3C kein Kriterium, das beim Ranking von Google genutzt wird. Matt Cutts bestätigte dies in einem entsprechenden Screencast[14]. Ein Großteil der Seiten im Netz enthält keinen validen Quellcode, sei es um alte Browser zu unterstützen oder weil der Validität im Alltag keine Bedeutung zugemessen wird. Generell lässt sich jedoch von der Validität des Quellcodes nicht auf die Qualität des Inhaltes schließen und die meisten aktuellen Browser stellen selbst nicht-validen Code korrekt dar. Die Berücksichtigung würde also potenziell zu inhaltlich schlechteren Ergebnissen führen, weil invalide Webseiten benachteiligt würden. Googles Startseite selbst zeigt übrigens über 30 Validierungsfehler[15].

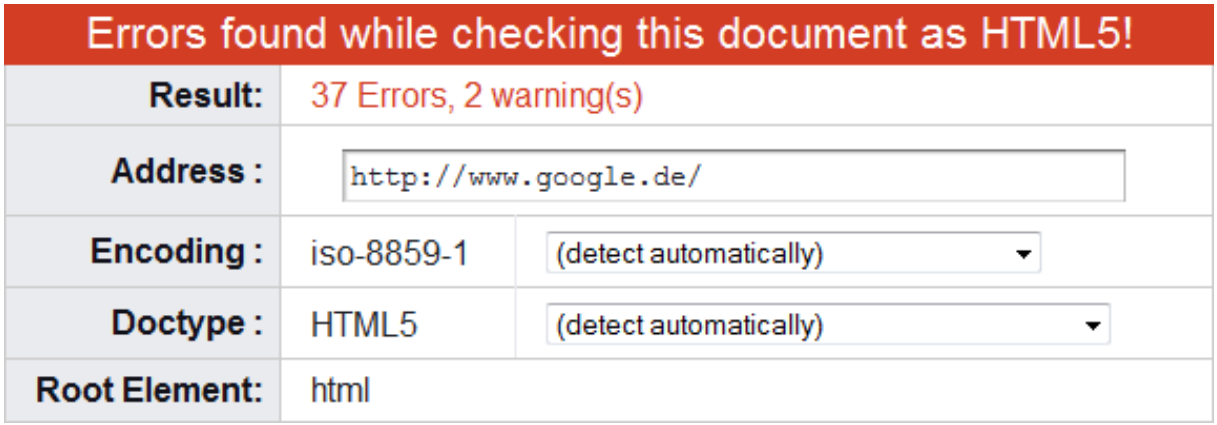

Abbildung 8: Validierungsfehler von http://www.google.de/ am 31.12.2010

[14] http://www.youtube.com/watch?v=FPBACTS-tyg, besucht am 31.12.2010
[15] http://validator.w3.org/check?uri=http%3A%2F%2Fwww.google.de%2F Stand 31.12.2010

3.4.2 Markup

HTML ist eine Auszeichnungssprache für textuelle Inhalte. Die Auszeichnung wird dabei über Tags vorgenommen, die eine gewisse semantische Bedeutung ausdrücken und in entsprechender Weise von Browser gerendert werden, so dass sich diese Bedeutung auch visuell für einen Besucher manifestiert. Über diese Auszeichnung ist es also möglich, dass bestimmte Inhalte mit Semantik belegt werden.

Es gibt keine offizielle Aussage, welchem Tag wie viel Gewicht zugemessen wird und man kann davon ausgehen, dass es keine einfachen mathematischen Zusammenhänge gibt, sondern dass sich ein für das Ranking positiver Wert aus verschiedenen Faktoren zusammensetzt. Es reicht also nicht aus, sämtliche Keywords im Text zum Beispiel in Fettschrift oder gar als Überschrift zu markieren. Derartige Techniken werden eher als Spam gewertet. Es folgen einige Beispiele für Tags, die einen Einfluss auf das Ranking haben.

Überschriften

Überschriften werden in HTML über die `<hx>` Tags gekennzeichnet, wobei x für den Grad der Überschrift steht. `<h1>` Tags kennzeichnen die wichtigsten Überschriften. Das Keyword oder die Keywordkombination sollte auf jeden Fall innerhalb eines `<hx>` Tags auftauchen. Man kann allerdings keine eindeutige Aussage darüber treffen, ob Google die unterschiedlichen `h` Tag Abstufungen in ihrer natürlichen Reihenfolge als Rankingfaktoren einsetzt. Gegen diese Annahme spricht zum Beispiel, dass das weltweit am meisten verbreitete Blogsystem Wordpress die Titel neuer Posts standardmäßig in `<h2>` statt in `<h1>` Tags darstellt.

Es ist allerdings relativ sicher zu sagen, dass Worten in Überschriften eine höhere Bedeutung zufällt, weil sie klar aus dem Text hervorstechen. Eye-Tracking Analysen haben außerdem ergeben, dass Besucher hauptsächlich die Überschriften lesen und den Rest eines Textes nur überfliegen. Es macht also durchaus Sinn, diesen Strukturierungselementen eine größere Bedeutung zuzumessen.

Listen

Listen sind ebenfalls Strukturierungselemente und fallen beim Betrachten einer Webseite ins Auge. In [HH10] werden verschiedene Algorithmen und Ansätze vorgestellt, die die semantische Nähe von Wörtern in einem Text identifizieren sollen. Einer dieser Ansatzpunkte lautet dabei, dass Listen gesondert behandelt werden müssen, da sie zum Beispiel eine wertungsfreie Aufzählung vornehmen könnten. So sind zum Beispiel die Daten in Abbildung 9 als gleichwertig bezüglich der semantischen Distanz zum Titel zu betrachten, obwohl der letzte Punkt rein

örtlich betrachtet weiter entfernt liegt als der Erste. In [Fis09, S. 317] wird sogar vermutet, dass Wörter in Listen auf Grund ihrer optischen Hervorhebung als wichtiger eingestuft werden.

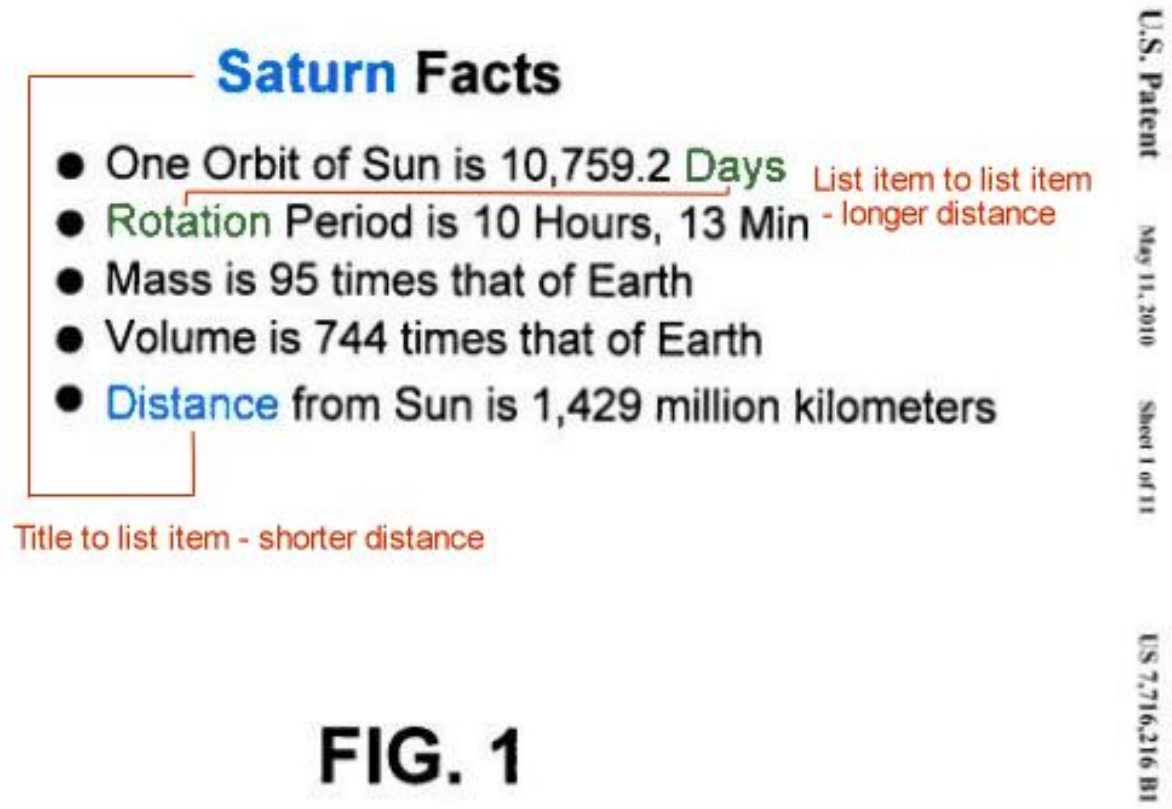

Abbildung 9: Beispiel zur semantischen Nähe von Listenpositionen

Fett, kursiv, etc.

Worte, die innerhalb eines Fließtextes auf eine Weise markiert sind, die sie vom restlichen Text abheben, stechen einem Besucher beim Lesen ins Auge. Diese Worte tragen also scheinbar mehr Bedeutung als die nicht-markierten Worte. An dieser Stelle ist der Begriff „Markierung“ allerdings zu weitläufig benutzt, weil es durch die Einführung von CSS[16] möglich ist, diese visuellen Effekte zu erzeugen, ohne dabei die vorgesehenen HTML Tags zu verwenden. Eine Suchmaschine kann diese Hervorhebung dann nicht nachvollziehen.

Nachweisbaren Einfluss hat laut [Fis09, S. 320] zumindest die Auszeichnung in Fettschrift

[16]Cascading Style Sheets

durch <b> oder Tags. Es ist anzunehmen, dass das Gleiche auch für kursive Schrift (<i> beziehungsweise Tags) und unterstrichenen Text (<u> Tag) gilt.

Aus Sicht der Usability macht es Sinn, bei einem Text die wichtigsten Aussagen besonders zu markieren, so dass ein Leser den Text zur Not nur überfliegen muss um die Kernaussage zu verstehen.

Multimediale Inhalte

Suchmaschinen haben generell Probleme bei dem Verständnis von multimedialen Inhalten wie Videos, Bildern oder Musikdateien. Sie sind hier zum einen auf den umschließenden Text angewiesen und zum anderen auf die Benutzung des `alt` Attributes. Dieses Attribut beschreibt den darzustellenden Inhalt und wird zum Beispiel von Screenreadern benötigt. Weiterhin trägt der Dateiname des Bildes zur Informationsgewinnung bei. Die Optimierung dieser Faktoren beeinflusst auf jeden Fall die Positionierung einer Grafik in der Google Bildersuche und trägt laut [Fis09, S. 325] außerdem auch zu einem besserem Ranking derjenigen Seite bei, die dieses Bild anzeigt (sofern es thematisch passt).

4 OffPage Optimierung

Die Offpage Optimierung ist das Äquivalent zur OnPage Optimierung. Sie befasst sich mit der Reputation einer Seite, maßgeblich ausgedrückt durch deren Verlinkung durch andere Webseiten im Internet. Wie auch bei der OnPage Optimierung gibt es hier unterschiedliche Einflussfaktoren, die sich zunächst in zwei Gruppen einteilen lassen:

1. Quantitative Faktoren
2. Qualitative Faktoren

Seit dem 1. Dezember 2010 gibt es außerdem die offizielle Bestätigung einer weiteren Gruppe von Einflussfaktoren. Dies wurde bei einem Interview von Danny Sullivan mit Vertretern der Suchmaschinen Google und Bing (siehe [Sul10]) bekanntgegeben. Dabei handelt es sich um sogenannte soziale Medien. Da deren Einfluss aber erst seit kurzem bekannt und (noch) sehr gering ist, gibt es keine ausreichenden Daten oder Informationen dazu. Deshalb werden in dieser Arbeit nur die quantitativen und qualitativen Faktoren herkömmlicher Backlinks behandelt.

4.1 Quantitative Faktoren

Quantitative Einflüsse basieren auf der ungewichteten Anzahl von Backlinks zu einer Seite. Der ursprüngliche PageRank Algorithmus ist das beste Beispiel für diese Art von Einflussfaktoren. Es gibt drei wichtige Kennzahlen in diesem Bereich:

1. Linkpopularität
2. Domainpopularität
3. IP-Popularität

Diese Kennzahlen benutzt man zum einen für eine gesamte Domain, zum anderen aber auch für einzelne Unterseiten einer Domain. Die Zahlen für gesamte Domain ergeben sich dabei aus der Summe der einzelnen Ausprägungen der Kennzahlen aller Unterseiten inklusive der Startseite. In den folgenden Erklärungen wird zur Vereinfachung jeweils der Term „Webseite“ benutzt.

4.1.1 Linkpopularität

Unter dem Begriff Linkpopularität versteht man die Anzahl sämtlicher Backlinks, die auf eine Webseite verweisen. Frei nach dem Motto: „Viele Links bedeuten eine hohe Reputation“ gilt hier: „Je größer diese Zahl ist, desto besser ist das für das Ranking einer Webseite“.

Allerdings gilt diese Kennzahl als nicht besonders aussagekräftig, da eine sogenannte seitenweite Verlinkung heutzutage nichts Ungewöhnliches mehr darstellt. Ein konkretes Beispiel dafür sind die sogenannten Blogrolles. Dabei verlinkt ein Blogger zum Beispiel einen befreundeten oder themenrelevanten anderen Blog von der Navigation seines eigenen Blogs aus. Dadurch wird dieser Link auf jeder Unterseite seines Blogs angezeigt. Jeder dieser Links fließt in die Linkpopularität mit ein. Der Ursprungsgedanke, ein Link sei eine spezielle Empfehlung für weitergehende Informationen verliert somit an Bedeutung, denn offensichtlich ist der in der Blogroll gesetzte Link nicht spezifisch auf den Inhalt einer Unterseite zugeschnitten.

4.1.2 Domainpopularität

Die Domainpopularität ist eine Kennzahl, die die Anzahl der auf eine Webseite verlinkenden Domains angibt. Dabei spielt es keine Rolle, wie oft diese Webseite von den Seiten einer Domain verlinkt wurde. Damit entschärft man die Problematik der Verlinkung auf jeder Unterseite maßgeblich. Auch für die Domainpopularität gilt: Je größer diese Kennzahl desto positiver ist der Einfluss auf das Ranking.

4.1.3 IP-Popularität

Eine verschärftere Form der Domainpopularität stellt die IP-Popularität dar. Bei dieser Kennzahl werden sämtliche IP-Adressen gezählt, von denen aus Links auf eine Webseite gesetzt werden. Diese Zahl entstand aufgrund der Tatsache, dass die meisten Webhoster einen gewissen, begrenzten IP-Bereich zur Verfügung haben und diesen für verschiedene Domains zur Verfügung stellen. Dabei ist es möglich, dass unterschiedliche Domains auf der gleichen IP-Adresse gehostet sind.

Um die Domainpopularität einer Webseite zu erhöhen könnte man nun auf den Gedanken kommen, bei einem Hoster schlichtweg eine ganze Reihe von unterschiedlichen Domains zu registrieren und von diesen auf die Zielwebseite zu verlinken. Dadurch würde offensichtlich eine Manipulation der Suchmaschinenalgorithmen stattfinden, denn der gesetzte Link hätte primär das Ziel, das Ranking der Zielwebseite zu verbessern und besäße damit eigentlich keinen Wert im Sinne einer Reputationserhöhung.

Die IP-Popularität gibt es in diversen Ausprägungen. Die oben vorgestellte Variante ist dabei die am wenigsten restriktive, da sie lediglich voraussetzt, dass sich die IP Adressen in irgendeinem Bit unterscheiden. In anderen Varianten wird zum Beispiel ein komplettes Klasse C Netzwerk

(also die ersten drei Oktette einer IPv4[17]-Adresse) als Berechnungsgrundlage gewählt. In jeder Variante gilt jedoch auch hier der Grundsatz, dass eine hohe IP-Popularität einen positiven Einfluss auf das Ranking hat.

4.1.4 Fazit

Die quantitativen Faktoren der OffPage Optimierung geben einen groben Überblick über den Verlinkungsgrad beziehungsweise die Popularität einer Webseite im Internet. Das Problem bei diesen Kennzahlen ist die Gefahr der Manipulation. Zwar werden durch restriktivere Betrachtungsweisen die Manipulationsversuche erschwert, aber noch längst nicht unwirksam gemacht. Es gibt zum Beispiel bereits spezielle Hosting Angebote, bei denen hunderte von Domains auf unterschiedlichen IPs gehostet werden[18].

Ein weiteres Problem dieser Kennzahlen besteht darin, dass sie zu ungerechtfertigten Benachteiligungen führen können. Das einfachste Beispiel dafür sind zwei unterschiedliche Domains, die über ein gemeinsames Thema berichten und sich häufig gegenseitig verlinken - zum Beispiel weil die jeweils andere Domain weiterführende Informationen zur Verfügung stellt. In diesem Falle wäre die IP- oder Domainpopularität zu restriktiv, da es angebracht wäre, jeden einzelnen Link zu werten, auch wenn er von der gleichen Domain stammt.

Aufgrund dieser Probleme kann man davon ausgehen, dass diese Kennzahlen eine erste Einschätzung bezüglich der OffPage Optimierung einer Webseite liefern, aber noch lange nicht das Maß aller Dinge sind.

4.2 Qualitative Faktoren

Nicht jeder Backlink besitzt die gleiche Qualität. Selbst bei dem ursprünglichen PageRank Algorithmus wurden Backlinks von Webseiten mit einem hohen PageRank höher bewertet als von solchen mit niedrigem PageRank. Da aber selbst der PageRank noch relativ leicht zu manipulieren ist, werden noch einige weitere Faktoren zu Rate gezogen.

4.2.1 PageRank

Der PageRank wurde bereits in Kapitel 2.4.1 vorgestellt und wird an dieser Stelle nur der Vollständigkeit halber erwähnt.

[17] Internet Protocol version 4
[18] http://www.multipleiphosting.com/

4.2.2 TrustRank

Ein Konzept, dem heutzutage eine sehr große Bedeutung zugemessen wird, ist der TrustRank. Trust bedeutet in diesem Zusammenhang die Vertrauenswürdigkeit einer Seite bezüglich deren bereitgestellten Informationen und deren Resistenz gegen Webspam.

Suchmaschinen haben ein begründetes Interesse daran, den Benutzern lediglich relevante, auf deren Suche zugeschnittene Ergebnisse zu liefern. Dabei sehen sie sich ständig der Problematik aggressiver Online-Marketing-Methoden ausgesetzt, die sich zum Beispiel durch automatisierten Webspam äußern.

Begriffsklärung

Bevor auf die Umsetzung des TrustRanks eingegangen wird, muss zunächst der Begriff eindeutig identifiziert werden. Im Allgemeinen wird der in [GGMP04] vorgestellten Algorithmus gemeint, wenn von TrustRank die Rede ist. Einer der Co-Autoren dieses Papers war Jan Pedersen, ein Yahoo! Mitarbeiter der ein Jahr später den Patentantrag [BGP] einreichte. Es handelt sich hierbei also nicht um ein von Google eingereichtes Patent. Es ist jedoch davon auszugehen, dass Google ein recht ähnliches Prinzip verwendet. Für einige Verwirrung sorgte in diesem Zusammenhang auch die Tatsache, dass Google fast zur selben Zeit den Trademark auf den Begriff „TrustRank“ hielt, damit jedoch einen Anti-Phishing Filter bezeichnete. Der hier beschriebene TrustRank bezieht sich jedoch auf das in [GGMP04] vorgestellte Konzept.

Algorithmus

Die Grundidee des TrustRanks besteht in der Einteilung in gute und schlechte Webseiten. Unter guten Webseiten versteht man solche, die regelmäßig gepflegt und deren Inhalte überwacht werden. Gute Webseiten zeichnen sich außerdem dadurch aus, dass sie mit sehr geringer Wahrscheinlichkeit auf schlechte Seiten verlinken, dafür aber mit einer hohen Wahrscheinlichkeit auf qualitativ hochwertige Seiten. Schlechte Webseiten sind Spam-Seiten, die zum Beispiel illegale oder betrügerische Absichten verfolgen oder allein zum Zwecke der Suchmaschinenmanipulation existieren.

Das Problem ist an dieser Stelle die Unfähigkeit, die Unterscheidung in gute und schlechte Seiten komplett automatisiert vorzunehmen. Deshalb basiert der Algorithmus auf einer sogenannten Orakelfunktion, bei der ein menschlicher Autor diese Unterscheidung vornimmt. Da es bei einer ständig wachsenden Anzahl an Webseiten unmöglich ist, jede Webseite einzeln zu bewerten, wird ein dem PageRank Algorithmus ähnliches Vererbungsprinzip eingesetzt. Dazu wird zunächst eine automatisierte Vorauswahl an Webseiten getroffen, die möglichst viele

gute Webseiten besitzen sollte. Für diese Vorauswahl kann zum Beispiel der PageRank als Auswahlkriterium dienen. Die ausgewählten Webseiten werden als Seed bezeichnet und bekommen von einem Menschen einen sogenannten Trustscore zugewiesen. Dieser Trustscore wird dann ebenso wie der PageRank an verlinkte Webseiten vererbt. Durch die oben erwähnte Eigenschaft guter Webseiten, nur mit einer sehr geringen Wahrscheinlichkeit auf schlechte, aber mit hoher Wahrscheinlichkeit auf gute Webseiten zu verlinken, bietet der TrustRank jedoch eine größere Sicherheit vor Manipulationen als der PageRank. Da jedoch nicht davon auszugehen ist, dass sämtliche verlinkte Seiten ebenfalls der gleichen inhaltlichen Überwachung und Pflege wie die ursprüngliche Seed unterliegen, wird ein Dämpfungsfaktor bei der Vererbung eingesetzt.

Fazit

Der TrustRank ist ein wirkungsvolles Konzept um die Verbreitung von Spam in den Suchmaschinenergebnissen zu minimieren. Weiterhin kann er außerdem als Rankingfaktor eingesetzt werden, da er ebenso wie der PageRank iterativ ermittelt werden und allen Seiten des Internets einen Wert zuweisen kann, wodurch wiederum eine Metrik entsteht, die einen Vergleich verschiedener Webseiten ermöglicht. Ein hoher TrustRank kann also zu einem besseren Ranking führen. Dieser kann erreicht werden, indem man von einer Seite mit hohem Trust verlinkt wird.

Das Problem beim Einsatz dieses Algorithmus ist zum einen die Wahl der richtigen Seed-Webseiten und zum anderen die Kalibrierung der verschiedenen Parameter (wie zum Beispiel dem Dämpfungsfaktor).

4.2.3 Backlinkeigenschaften

Die Qualität eines Backlinks wird durch diverse Faktoren bestimmt. Dazu zählen zum einen die bereits vorgestellten Kennzahlen, denn eine Webseite die zum Beispiel eine hohe IP-Popularität besitzt, wird von Suchmaschinen höher bewertet. Dementsprechend zählen auch Links von diesen Seiten mehr. Gleiches gilt auch für PageRank und TrustRank. Es gibt aber noch weitere Faktoren, die sich direkt auf diejenige Seite beziehen, auf der sich ein Backlink befindet. Diese werden im Folgenden vorgestellt.

Google kennt die linkgebende Webseite

Ein Backlink kann nur dann eine positive Wirkung haben, wenn Google auch von der Existenz dieses Links Kenntnis besitzt. Das lässt sich zum Beispiel prüfen, wenn man die URL, über die die linkgebende Seite zu erreichen ist, bei Google eingibt um danach zu suchen. Die URL müsste nun in den Suchmaschinenergebnissen auftauchen. Dieses Vorgehen kann jedoch nicht

immer angewandt werden. Ein Beispiel dafür ist der unter Punkt 3.2.5 vorgestellte Noindex Meta Tag. Dieser verhindert zwar die Aufnahme in den Google Index, führt aber nicht dazu, dass ein Backlink von dieser Seite nicht gewertet wird.

Linkmethode

Die herkömmliche Methode einen Hyperlink zu erzeugen besteht in der Verwendung des `<a>` Tags, der die Zielurl als Wert seines `href` Attributs trägt. Die Verwendung dieser Methode garantiert die volle Wertigkeit dieses Backlinks. Damit dieser Link zählt, darf er allerdings nicht mit dem nofollow Attribut ausgestattet sein (siehe Punkt 2.4.1). Es gibt aber noch weitere Methoden, einen Hyperlink zu realisieren. Dazu zählt zum Beispiel eine Weiterleitung mit Hilfe des JavaScript Befehls *`window.location`* oder die Verlinkung in einer HTML Image Map. Google verfolgt auch diese Arten von „Hyperlinks“, es ist jedoch nicht bekannt ob diese auch als gleichwertig zu `<a>` Tags gewertet werden. Einige Webseiten verwenden ein spezielles Skript zur Maskierung ausgehender Links. Dabei wird jeweils auf eine seiteninterne Adresse verwiesen, an die ein Parameter (zum Beispiel die URL der Zielseite) übergeben wird. Das Skript verwendet nun einen Redirect (HTTP Statuscode 3xx) um letztendlich auf die gewünschte Seite weiterzuleiten. Wie bereits unter Punkt 3.2.5 erläutert wurde, unterliegen diese Redirects immer einem gewissen Malus, der den Wert eines solchen Backlinks schmälert.

Ankertext

Der Ankertext wurde ebenfalls bereits unter Punkt 2.4.2 eingeführt und erläutert. Für Suchmaschinen stellt er eine zusätzliche Information zur angelinkten Seite dar und sollte deshalb aus dem gewünschten Zielwort bestehen.

Generell geht man jedoch davon aus, dass es in diesem Bereich einen Filter von Google gibt, der die Verteilung verschiedener Ankertexte überwacht. Der Grundgedanke hinter diesem Filter ist die Erkennung einer unnatürlichen Verlinkung, die wiederum auf eine Manipulation hindeuten könnte. Für den praktischen Einsatz bedeutet das, dass die Ankertexte zu einem gewissen Teil variieren sollten. So wird zum Beispiel häufig der komplette Domainname als Linktext benutzt. Ein anderes Beispiel ist die Verwendung des vollen Namens bei dem Hinterlassen von Kommentaren bei Blogbeiträgen.

Linktitel

Das `title` Attribut gehört zu den sogenannten Universalattributen von HTML und kann als solches in fast allen HTML Tags verwendet werden. Auf selfHTML wird das `title` Attribut unter [SEL] wie folgt beschrieben:

> [Das title Attribut]erlaubt es, HTML-Elemente mit kommentierendem Text beziehungsweise Meta-Information auszustatten. Der kommentierende Text wird gängigerweise vom Browser in einem kleinen Fenster („Tooltip“) oder in der Statusleiste angezeigt, wenn der Anwender mit der Maus über den Anzeigebereich des HTML-Elements fährt.

Mit Hilfe dieses Attributes können einem Backlink also zusätzlich zum Ankertext weitere Informationen über die angelinkte Seite mitgegeben werden. Zwar gibt es seitens Google keine offizielle Aussage, dass der in diesem Attribut enthaltene Text einen Einfluss auf das Ranking besitzt, aber es wäre nur konsequent, wenn dies der Fall wäre.

Themenrelevanz

Der Ursprungsgedanke des Random Surfer Modells bestand in der Annahme eines Users, der über Hyperlinks von einer Webseite zur nächsten navigiert und dabei den Zufall entscheiden lässt, welchem Link konkret gefolgt wird. Der Reasonable Surfer erweitert dieses Modell um eine Wahrscheinlichkeitskomponente, die gewisse Links stärker gewichtet als andere. Die konsequente Fortführung dieses Gedankens ist eine Aufwertung von Backlinks, die sich in einem themenrelevanten Kontext befinden und einem User zum Beispiel weiterführende Informationen zur Verfügung stellen. Dass Google in der Lage ist, Zusammenhänge zwischen verschiedenen Begriffen herzustellen, zeigen verschiedene Services wie zum Beispiel die Anzeige verwandter Suchbegriffe, die beispielhaft in Abbildung 10 dargestellt ist. Auch wenn es sich dabei wahrscheinlich um ein statistisches Modell handelt, so belegt es zumindest die Möglichkeiten, die Google in diesem Zusammenhang besitzt.

Abbildung 10: Anzeige verwandter Suchbegriffe zum Suchbegriff Mallorca

Backlinks aus einem relevanten Kontext können außerdem zur Erkennung von manipulativen (zum Beispiel eingekauften) Backlinks beitragen. Bei einem zu großen Anteil themenirrelevanter Backlinks kann zum Beispiel automatisch ein Flag gesetzt werden, das entweder eine menschliche Kontrolle auslöst oder für eine Rankingabstrafung der entsprechenden Seite oder gar der kompletten Domain sorgt.

Linkplatzierung

In Kapitel 2.4.1 wurde bereits erläutert, dass die Platzierung eines Backlinks auf einer Webseite (Header, Content, Footer, etc.) einen Einfluss auf dessen Gewichtung hat.

Anzahl externer und interner Links

Die Anzahl der externen und internen Links, die sich auf einer Seite befinden, schmälern den Wert jedes einzelnen Links. Dies geht direkt aus dem PageRank Algorithmus hervor. Google selbst nannte als Richtwert max. 100 Links pro Seite. Diese Zahl ist jedoch historisch bedingt und geht vor allem darauf zurück, dass Google zu Beginn lediglich 100kb einer Webseite indexiert hat. Dennoch ist sie auch heute noch gültig - wenn auch aus anderen Gründen, die Matt Cutts in einem Blogpost [Cut09a] wie folgt erläutert:

> [...]These days, Google will index more than 100K of a page, but there's still a good reason to recommend keeping to under a hundred links or so: the user experience. If you're showing well over 100 links per page, you could be overwhelming your users and giving them a bad experience.[...]

Generell geht man davon aus, dass interne Links in diesem Zusammenhang nicht so stark ins Gewicht fallen wie Externe, da es bei vielen Webseiten üblich ist, zum Beispiel eine Navigation auf jeder Seite anzuzeigen, so dass sich fast immer einige interne Links auf jeder Seite befinden. Für eine linkgebende Seite ist es jedoch weiterhin ein positives Kriterium, wenn dieses möglichst wenige ausgehende Links besitzt.

5 Zusammenfassung und Ausblick

5.1 Intention

Das Internet ist eine extrem schnell wachsende Informationsplattform. Die Fülle an Informationen allein macht das Internet jedoch nicht wertvoll, sondern die Ordnung, die durch Suchmaschinen erschaffen wird. Immer mehr Menschen nutzen das Internet tagtäglich und es stellt für Unternehmen einen wesentlichen Wettbewerbsvorteil dar, wenn sie sich in ihrem Umfeld im Internet (repräsentiert durch eine gute Position in den Suchmaschinen) etablieren können. Die Suchmaschinenoptimierung ist deshalb ein Forschungsfeld, dass zum einen in der Zukunft relevant sein wird und das zum anderen auch heute bereits einen Bezug zur Realität hat.

5.2 Probleme

Da die exakten Algorithmen, die eine Suchmaschine einsetzt, nicht veröffentlicht werden können, bleiben zur Erforschung derselben zunächst nur offizielle Aussagen und eingereichte Patente als Anhaltspunkte sowie empirisch durchgeführte Experimente als Belege. Die Durchführung dieser Experimente stellt jedoch ein weiteres Problem dar, da Google über 200 Faktoren bei der Berechnung des Rankings mit einbezieht. Es sind weder alle diese Faktoren bekannt, noch kann deren einzelne Gewichtung verlässlich ermittelt werden und so ist es fast unmöglich eine sterile Testumgebung aufzubauen. Zuverlässige Aussagen können eigentlich nur bezüglich Ja/Nein-Fragen gemacht werden. So kann man zum Beispiel empirisch ermitteln, ob Google Hyperlinks folgt, die mittels der JavaScript Funktion `window.location` realisiert sind, in dem man eine ansonsten unverlinkte Seite über eben jene Funktion verlinkt und in den Logfiles das Erscheinen des Googlebots überwacht. Es ist jedoch kaum möglich eine Aussage darüber zu treffen, ob dieser Link gleichwertig zu einem normalen, mittels `<a>` Tag realisierten Link gewertet wird.

Die größte Herausforderung bei dieser Arbeit bestand in der Zusammenstellung belegbarer Fakten zum Beispiel in Form offizieller Aussagen oder von Google eingereichten Patenten. Man findet zwar viele Informationen zur Suchmaschinenoptimierung im Internet, aber darunter befindet sich sehr viel Halbwissen, Vermutungen und falsche Schlussfolgerungen, so dass es hier galt, die Informationen zu hinterfragen und mit belegbaren Aussagen zu unterstützen oder zu widerlegen.

5.3 Fazit

In dieser Arbeit wurde ein Einblick in verschiedene Faktoren der Suchmaschinenoptimierung bezogen auf die Suchmaschine Google gegeben. Neben der generellen Funktionsweise von Google wurden außerdem die beiden großen Bereiche OnPage und OffPage Optimierung behandelt. Für eine erfolgreiche Suchmaschinenoptimierung sind beide Bereiche wichtig und sollten mit gleicher Sorgfalt bearbeitet werden. In der Realität hat sich jedoch gezeigt, dass die OffPage Optimierung die OnPage Optimierung überwiegt. Das lässt sich zum Beispiel dadurch erklären, dass die OnPage Optimierung in ihren Möglichkeiten begrenzt ist und bei der Vielzahl von Webseiten kein ausreichendes genaues Unterscheidungskriterium mehr darstellt. Die OffPage Optimierung hingegen repräsentiert die Reputation einer Webseite, die wiederum mit jedem eingehenden Backlink erhöht wird und der nach oben hin im Prinzip keine Grenzen gesetzt sind.

Literatur

[Ach+] Anurag Acharya u. a. “Information retrieval based on historical data”. Patent Application 20050071741.

[And08] John Andrews. *What Matt Cutts Said at Domain RoundTable 2008*. 19. Apr. 2008. URL: http://www.johnon.com/543/mattcutts-domainroundtable.html (besucht am 31. 12. 2010).

[BGP] Pavel Barkhin, Zoltan Istvan Gyongyi und Jan Pedersen. “Link-based spam detection”. Patent Application 20060095416.

[BLFM05] T. Berners-Lee, R. Fielding und L. Masinter. *Uniform Resource Identifier (URI): Generic Syntax*. RFC 3986. Internet Engineering Task Force, Jan. 2005. URL: http://www.rfc-editor.org/rfc/rfc3986.txt.

[Blo05] The Official Google Blog. *Preventing comment spam*. 18. Jan. 2005. URL: http://googleblog.blogspot.com/2005/01/preventing-comment-spam.html (besucht am 31. 12. 2010).

[Blo07] Official Google Webmaster Central Blog. *A quick word about Googlebombs*. 25. Jan. 2007. URL: http://googlewebmastercentral.blogspot.com/2007/01/quick-word-about-googlebombs.html (besucht am 31. 12. 2010).

[Blo08] Official Google Webmaster Central Blog. *Crawling through HTML forms*. 11. Apr. 2008. URL: http://googlewebmastercentral.blogspot.com/2008/04/crawling-through-html-forms.html (besucht am 31. 12. 2010).

[Blo09a] Official Google Webmaster Central Blog. *Google does not use the keywords meta tag in web ranking*. 21. Sep. 2009. URL: http://googlewebmastercentral.blogspot.com/2009/09/google-does-not-use-keywords-meta-tag.html (besucht am 31. 12. 2010).

[Blo09b] Official Google Webmaster Central Blog. *Specify your canonical*. 12. Feb. 2009. URL: http://googlewebmastercentral.blogspot.com/2009/02/specify-your-canonical.html (besucht am 31. 12. 2010).

[Blo10] Official Google Webmaster Central Blog. *Using site speed in web search ranking*. 9. Apr. 2010. URL: http://googlewebmastercentral.blogspot.com/2010/04/using-site-speed-in-web-search-ranking.html (besucht am 31. 12. 2010).

[BP98] Sergey Brin und Lawrence Page. “The Anatomy of a Large-Scale Hypertextual Web Search Engine”. In: *COMPUTER NETWORKS AND ISDN SYSTEMS*. Elsevier Science Publishers B. V., 1998, S. 107–117.

[Cena] Google Librarian Central. *How does Google collect and rank results?* URL: http://www.google.com/librariancenter/articles/0512_01.html (besucht am 31. 12. 2010).

[Cenb] Google Webmaster Central. *Meta tags*. URL: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=79812 (besucht am 31. 12. 2010).

[Cenc] Google Webmaster Central. *URL structure*. URL: http://www.google.com/support/webmasters/bin/answer.py?hl=en&answer=76329 (besucht am 31. 12. 2010).

[Cen10] Google Webmaster Central. *Keyword stuffing*. 9. Okt. 2010. URL: http://www.google.com/support/webmasters/bin/answer.py?answer=66358 (besucht am 31. 12. 2010).

[Cut09a] Matt Cutts. *How many links per page?* 9. März 2009. URL: http://www.mattcutts.com/blog/how-many-links-per-page/ (besucht am 31. 12. 2010).

[Cut09b] Matt Cutts. *PageRank sculpting*. 15. Juni 2009. URL: http://www.mattcutts.com/blog/pagerank-sculpting/ (besucht am 31. 12. 2010).

[DAB10] Jeffrey A. Dean, Corin Anderson und Alexis Battle. “Ranking documents based on user behavior and/or feature data”. Patent 7716225. 2010.

[Eng10] Eric Enge. *Matt Cutts Interviewed by Eric Enge*. 14. März 2010. URL: http://www.stonetemple.com/articles/interview-matt-cutts-012510.shtml (besucht am 31. 12. 2010).

[Fis09] Mario Fischer. *Website Boosting 2.0: Suchmaschinen-Optimierung, Usability, Online-Marketing*. Bd. 2. Heidelberg: mitp, 2009. ISBN: 978-3-8266-1703-4.

[GGMP04] Zoltán Gyöngyi, Hector Garcia-Molina und Jan Pedersen. “Combating web spam with trustrank”. In: *In VLDB*. Morgan Kaufmann, 2004, S. 576–587.

[Goo] Google. *Technology overview - Google Corporate Information*. URL: http://www.google.com/corporate/tech.html (besucht am 31. 12. 2010).

[HH10] Georges R. Harik und Monika H. Henzinger. “Document ranking based on semantic distance between terms in a document”. Patent 7716216. 2010.

[Inc] Google Inc. *Search Engine Optimization Starter Guide*. URL: http://www.google.com/webmasters/docs/search-engine-optimization-starter-guide.pdf (besucht am 31. 12. 2010).

[Mos09] Susan Moskwa. *PageRank Distribution Removed From WMT*. 14. Okt. 2009. URL: http://www.google.com/support/forum/p/Webmasters/thread?tid=6a1d6250e26e9e48&hl=en (besucht am 31. 12. 2010).

[Pag+99] Lawrence Page u. a. *The PageRank Citation Ranking: Bringing Order to the Web.* Techn. Ber. 1999-66. Previous number = SIDL-WP-1999-0120. Stanford InfoLab, 1999.

[Pat08] Anna Lynn Patterson. "Automatic taxonomy generation in search results using phrases". Patent 7426507. 2008.

[SEL] Redaktion SELFHTML. *Allgemeine Universalattribute. Übersicht der Universalattribute*. URL: http://de.selfhtml.org/html/attribute/allgemeine.htm#uebersicht (besucht am 31. 12. 2010).

[Sul10] Danny Sullivan. *What Social Signals Do Google & Bing Really Count?* 1. Dez. 2010. URL: http://searchengineland.com/what-social-signals-do-google-bing-really-count-55389 (besucht am 31. 12. 2010).