



***Wie eine Suchmaschine schnell zu guten
Suchergebnissen kommt***

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und ohne fremde Hilfe verfasst und keine anderen als die im Literaturverzeichnis angegebenen Hilfsmittel verwendet habe.

(Ort, Datum)

(Unterschrift)

Inhaltsverzeichnis

<i>Einleitung</i>	2
<i>Die Suchmaschine</i>	3
<i>Google Architektur</i>	3
Crawler	3
Doppelt hält besser	5
Der Weg einer Suchanfrage	7
<i>Der PageRank</i>	9
Random Surfer Modell	10
Aus rein Mathematischer Sicht	11
Veranschaulichung an einem Beispiel	12
Ausgehenden Links und neue Seiten	13
Eingehenden Links	15
Dangling Links	17
<i>Andere Bewertungskriterien</i>	18
<i>Fazit</i>	20
<i>Literaturverzeichnis</i>	21

Einleitung

Jeder benutzt Google täglich ein paar Mal. Schnell mal für eine Antwort auf eine kleine Frage, oder für die Suche nach vielen Informationsquellen zu einem bestimmten Thema. Ohne eine Suchmaschine ist man bei der Größe des heutigen World Wide Webs absolut ratlos und man weiß nicht wo man anfangen soll.

Da ich ein Informatikstudium anstrebe, ist es für mich wichtig, Dinge die ich täglich am Computer benutze, auch zu verstehen. Google ist ein Grundstein des Internets, ohne den man nur noch schwer auskommt. Herauszufinden was Google ist, wie es funktioniert und was es von mir für diesen Dienst verlangt, ist mein persönliches Ziel dieses Seminar-kurses.

In dieser Arbeit werde ich nur den wissenschaftlichen Teil Googles ausarbeiten und wirtschaftliche und ethische Faktoren so weit wie möglich ungeachtet lassen. Ich werde auf die Fragen, wie Google es schafft genau das zu finden, was ich suche, warum Google dafür nur Bruchteile einer Sekunde braucht und warum es bisher keine andere Suchmaschine zu dieser Perfektion gebracht hat, eine Antwort zu finden.

Um einen möglichst echten Eindruck von Google zu gewinnen, habe ich mich insbesondere auf die Literatur der Google Gründer Lawrence Page und Sergey Brin gestützt. Meine zweite Hauptquelle ist die Internet-Agentur *efactory.de*, die zur Optimierung von Websites für die Googlesuche entwickelt wurde und mit ihrem guten, und verständlichen Internetauftritt überzeugt.

Die Suchmaschine

„Eine Suchmaschine ist ein Programm zur Recherche von Dokumenten, die in einem Computer oder einem Computernetzwerk wie z. B. dem World Wide Web gespeichert sind.“¹

Das Ziel einer Suchmaschine ist es, wie ein Bibliothekar zu arbeiten. Sie muss alle verfügbaren Quellen kennen, die Frage des Suchenden verstehen und interpretieren, und auf seine Nachfrage hin die besten Ergebnisse herausfinden und dem Suchenden zurückgeben.

Diese Funktionsweise einer Suchmaschine wird nun Stück für Stück am Beispiel der Suchmaschine Google erörtert.

Google Architektur

Das Google die meist genutzte Suchmaschine im World Wide Web ist, erklärt sich durch die Qualität ihrer Suchergebnisse. Um zu dieser Qualität zu gelangen ist es erforderlich systematisch vor zu gehen. Dies zeigt sich bei Google durchweg von den Programmen, die für die Suche Notwendig sind, bis hin zum Aufbau der großen Rechenzentren.

Crawler

Um wie ein Bibliothekar suchen zu können, muss man erst einmal wissen, was es alles für Dokumente gibt. Dieses Aufspüren von Quellen im Internet übernimmt, wie bei allen großen Suchmaschinen, ein Crawler. Ein Crawler ist ein Programm, welches sich auto-

¹ <http://de.wikipedia.org/wiki/Suchmaschine>

matisch von Seite zu Seite klickt und alles speichert was es findet. Die Seite wird sowohl analysiert, indexiert, als auch gespeichert. Laut Google wird nicht nur der Text sondern „der gesamte Inhalt einer Seite unter Berücksichtigung von Faktoren wie Schriftarten, Unterteilungen und der genauen Position aller Begriffe analysiert.“²

Der indexierte Text wird samt den Informationen über die Häufigkeit einzelner Wörter, deren Position auf der Website und die Schriftgröße auf den „Indexservern“ in alphabetischer Reihenfolge gespeichert, um die Suchbegriffe möglichst schnell zu finden. Der Rest der Informationen eines Dokuments wird auf den „Documents Server“ gespeichert, die den Großteil der gesamten gespeicherten Daten ausmachen.

Wenn eine Crawler eine Seite analysiert, findet er auch meistens Hyperlinks³, die zu weiteren Seiten verweisen. Diese werden dann in einer Art To-Do-Liste auf den „URL Servern“ gespeichert und für andere Crawler zur Verfügung gestellt, auf die ein Crawler zurückgreift, wenn er eine vorgegebene Anzahl von Weiterleitungen gefolgt ist, oder auf einer Seite ohne weiterführende Links stößt (**Dangling Links** Seite 17). Diese Seiten werden dann als Nächstes durchsucht. Mit dieser Methode können rein theoretisch alle Seiten im World Wide Web gefunden werden, sofern sie irgendwo verlinkt sind. Allerdings sind einige Teile des Webs nur mit Zugangsdaten betretbar, die ein Crawler nicht besitzt, und sie somit auch nicht bearbeiten kann.

Da es aber vergleichsweise wenige Crawler für mehrere Milliarden von Dokumenten im World Wide Web gibt, ist die Googleuche nie auf dem neuesten Stand. In der Regel hinkt sie etwa 2 bis 4 Wochen hinterher. Um aber trotzdem Googlefunktionen wie Google News auf dem neuesten Stand zu halten, wurden so genannte „fresh crawls“⁴ eingeführt, die auf Geschwindigkeit und den täglichen Abruf von Nachrichten spezialisiert sind. Die Geschwindigkeit hat zur Folge, dass sie auch dementsprechend schlampig arbeiten.

Um seine eigene neu erstellte Seite in den Suchergebnissen bei Google finden zu können, muss man seine Seite bei den Crawlern anmelden, damit sie durchsucht wird.

² Auszug aus <http://www.google.de/corporate/tech.html>

³ = Link. Verknüpfung aus einem Webdokument zu einem anderen.

⁴ http://www.googleguide.com/google_works.html

Andernfalls kann es Wochen lang dauern bis ein Crawler zufällig auf die Seite trifft und dem Index hinzufügt.

Wenn ein Crawler dann aber mal eine Seite besucht, findet die Bewertung der Seite via PageRank statt (eine Iteration⁵), der aus der Zahl der ein und ausgehenden Links und deren Qualität errechnet wird (siehe **Der PageRank** Seite 9). Dieser PageRank ist später für die Sortierung der Suchergebnisse wichtig.

Doppelt hält besser

Da Google sein Geld nahezu ausschließlich mit seiner Websuche und der zu den Ergebnissen passenden Werbung verdient, muss verhindert werden, dass die Websuche durch einen Fehler im System nicht mehr verfügbar ist. Denn dann bleiben auch die Einnahmen aus. Aus diesem Grund besteht das Googlenetzwerk aus einem Cluster, der sich in verschiedenen Ländern in über 36⁶ Rechenzentren auf der Welt befindet. Der Cluster ist dadurch sogar vor Naturkatastrophen sicher, um eine ununterbrochene Verfügbarkeit zu gewährleisten.

Ein Cluster ist ein Netzwerk aus vielen Einzelcomputern. Der Vorteil eines Clusters gegenüber einzelner Hochleistungscomputer, ist die Ausfallsicherheit. Jede Komponente dieses Systems muss ausfallsicher, das heißt mehrfach vorhanden und ersetzbar sein.

Die vielen Computer eines Clusters haben ein besonders gutes Preis-Leistungs-Verhältnis, da nicht die Qualität einzelner Komponenten entscheidend ist. Die Stärke eines Clusters ist es nicht eine Aufgabe schnell nacheinander zu rechnen, sondern sie in kleine Teile zu teilen, und dann parallel auf vielen Computern zu erledigen, damit eine mit einem Hochleistungscomputer vergleichbare Geschwindigkeit erreicht wird und das bei geringeren Ausgaben. Hochleistungscomputer wären eventuell etwas schneller, und mit weniger Wartungskosten verbunden als tausende einzelner Server, aber die Entwicklung dieser Computer lässt den Preis enorm anschwellen, da nur wenige Exemplare verkauft werden.

⁵ Wiederholung des selben Rechenverfahrens

⁶ <http://www.googlewatchblog.de/2008/04/12/karte-von-googles-rechenzentren-rund-um-die-welt/>

Bei Google kommen heute speziell angefertigte Computer zum Einsatz. Das Mainboard⁷ wurde von Gigabyte extra für Google entwickelt. Es besitzt nicht einmal einen Grafikchip, dafür aber 8 Speicherslots für mehr als 4 Gigabyte Arbeitsspeicher⁸ und Platz für 2 Multicore CPUs⁹. Die Entwicklung von diesen Prozessoren war für Google von wichtiger Bedeutung, da so mehrere Rechenschritte gleichzeitig ausführbar waren und das ohne, dass mehr Computer benötigt wurden und der Preis für diese zusätzliche Rechenleistung drastisch anstieg. Um einen schnellen Datenzugriff von Servern zu ermöglichen besitzt ein Server bei Google 2 Festplatten. Darüber hinaus verfügt jeder Server über einen Akku, der den Server im Falle eines Stromausfalls mit Energie versorgt bis die Diesel-Generatoren für die alternative Stromversorgung angegangen sind.

Untergebracht sind die Server nicht wie bei einem Desktop-PC, sondern in speziell angefertigten 19'' Gehäusen, die in ein Google Rack passen. Ein Google Rack ist ein genormter Schrank der vorne und hinten Platz für 20 Google Server und einen Switch für die Vernetzung der Server hat. Von jedem Switch geht eine Verbindung zu zwei 128 Slot Switches aus, um die Ausfallsicherheit zu gewährleisten.



Google Racks nebeneinander

Ein Rechenzentrum im Jahr 2000 war standardmäßig über eine 2488 Mbit/s Leitung an das Internet angeschlossen und über eine 622 Mbit/s Leitung mit einem anderen Rechenzentrum verbunden, sollte die schnellere Leitung ausfallen.

Gekühlt wird ein Rechenzentrum mit einer Wasserkühlung, die über 2 große Pumpen verfügt und von der jede alleine das System am laufen halten kann.

⁷ Die Hauptplatine eines Computers. Die Verbindung der Verschiedenen Bausteine eines Computers.

⁸ Kurzzeitspeicher eines Computers.

⁹ Hauptprozessoren in einem Gehäuse.

Die Unterhaltskosten einer solchen Serverfarm sind 1,25 mal so groß wie die Kosten für den Stromverbrauch des gesamten IT-Equipment in einem solchen Rechenzentrum.

Diese Daten sind heutzutage mit Sicherheit nicht mehr aktuell, aber veranschaulichen gut das System, wie Google arbeitet. Google ist stets gegen Ausfälle gewappnet. Zahlen wie 2 kaputte Server pro Tag in jedem Rechenzentrum zeigen aber, dass Google nicht ohne Grund auf Nummer sicher geht.

Der Weg einer Suchanfrage

Bei einer Suchanfrage bei Google muss der Browser¹⁰ zuerst über einen DNS-Server die Adresse (z.B. www.google.de) in eine IP Adresse¹¹ auflösen. Da Google mehrere Rechenzentren besitzt, findet hier die erste Sortierung der Suchanfragen statt. Es wird immer auf das geographisch am nächsten gelegene Rechenzentrum verwiesen.

Als ersten Schritt im Rechenzentrum wird die Suchanfrage von einem „Load Balancing Server“ bearbeitet. Dieser prüft die Auslastung des Rechenzentrums, um die Anfrage im Falle einer Überlastung an ein anderes, weniger ausgelastetes Rechenzentrum weiterzuleiten. Wenn die Auslastung eine schnelle Bearbeitung zulässt wird die Anfrage an einen „Google Web Server“ (GWS) des jeweiligen Rechenzentrums übergeben.

Der GWS schickt die Suchanfrage an die Indexserver des Rechenzentrums. Bei diesem Schritt wird die Geschwindigkeit der Googlesuche erreicht, die nur ein Cluster erzeugen kann. Hunderte von Indexservern suchen nach dem Suchbegriff und geben eine Nummer (DocIDs) zu den passenden Dokumenten an den Indexserver zurück. Ein Indexserver muss allerdings nicht den kompletten Google-Index durchsuchen, sondern durchsucht nur den jeweiligen Teil des Indexes aus 5 Milliarden indexierten Seiten, der auf seiner Festplatte gespeichert ist. Jede Seite ist hundertfach im ganzen Googlecluster gespeichert, somit ist ein Ausfall einer Festplatte kein Grund zur Sorge.

¹⁰ Internet Explorer, Mozilla Firefox, Opera, Google Chrome

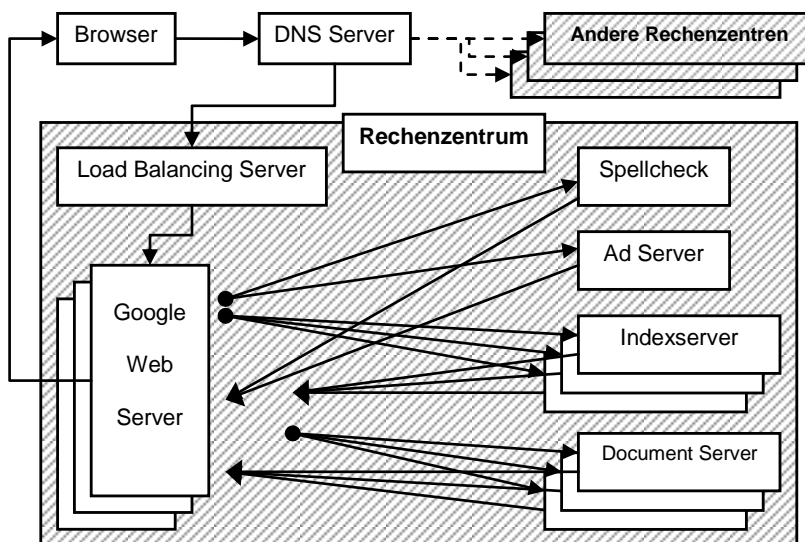
¹¹ Identifizierungsnummer eines Systems im Internet

Google durchsucht also nicht das Internet sondern nur die Seiten, die im Google Cluster gespeichert, also bereits gefunden und analysiert wurden.

Gleichzeitig wird die Suchanfrage auch an einen „Spellchecker“ und einen „Ad Server“ gesendet. Der Spellchecker überprüft die Suchanfrage nach möglichen Schreibfehlern um sie dem Suchenden als Möglichkeit für eine neue Suche vorzuschlagen.

Bei dem Ad Server wird die zu der Suchanfrage passende Werbung aus einer Liste mit Werbungen für die jeweiligen Suchbegriffe gesucht. Sollte eine oder mehrere Werbungen gefunden worden sein, werden diese dem GWS geschickt.

Der GWS gibt die DocIDs von den Indexservern nun an die Documents Server weiter, die parallel die verschiedenen Dokumente raussuchen und Beschreibungen für die Suchergebnisse erstellen, welche dann an den GWS zurückgeben werden. Nun muss der GWS die Dokumente nur noch ordnen, das heißt die Werbung vom Ad Server und die Vorschläge vom Spellchecker einbauen, und an den Browser des Suchenden zurückgeben. Dieser Vorgang ist auf dem Schaubild dargestellt.



Hier besteht ein starker Unterschied zu anderen Clustern. Da die Server im Googlecluster spezielle Aufgaben besitzen, die zwar immer noch dynamisch, aber speziell auf die jeweiligen Aufgabe eines Servers zugewiesen

werden, wird die klare Googlearchitektur nicht gestört, und die verschiedenen Serverarten können, je nach Anforderung, speziell angepasst werden.

Der PageRank

Der PageRank geht auf die Forschungsarbeit von Sergey Brin und Lawrence Page „The Anatom of a Large-Scale Hypertextual Web Search Engine“ zurück. Da seit dieser Veröffentlichung 1998 viel Zeit vergangen ist und sich das Internet durch seine Dynamik stark verändert hat, ist es höchst wahrscheinlich, dass der dort beschriebene PageRank Algorithmus¹² nicht mehr in dieser Form verwendet wird, sondern stark modifiziert ist. Der Erfolg der Google Suche ist allerdings auf den PageRank Algorithmus zurückzuführen, was zur Folge hat, dass die Grundzüge des PageRank Verfahrens immer noch die Gleichen sind.

Das PageRank Verfahren ist die Konsequenz aus vielen verschiedenen Versuchen eine erfolgreiche Suche im World Wide Web zu entwickeln. Vor allem ist es natürlich wichtig, dass der gesuchte Begriff möglichst häufig auf einer Website zu finden ist oder dass der Abstand der Suchbegriffe innerhalb des Dokuments nicht zu weit auseinander liegt und in der richtigen Reihenfolge ist. Er sollte groß und am bestens als Überschrift geschrieben sein. Außerdem sollte die Website auf vielen verschiedenen anderen Dokumenten verlinkt sein, damit die Qualität gewährleistet ist. Diese Kriterien sind bei Google genau wie bei allen anderen großen Suchmaschinen vorhanden und trotzdem gibt es noch einen entscheidenden Faktor.

Wenn man nämlich nach den bisher beschriebenen Kriterien eine Suchmaschine programmieren würde, sollte man auch gute Suchergebnisse bekommen, wenn es nicht auch Menschen geben würde der ein solches Suchverfahren für seine Zwecke ausnutzen würde.

Sollte beispielsweise eine Website erstellt werden, deren Überschrift „Auto Auto [...] Auto“ lautet und zusätzlich hunderte von Websites mit einer Verlinkung zu dieser Website erstellt würden, würde mit Sicherheit bei jeder Suchanfrage mit „Auto“ diese Web-

¹² Eine Rechenvorschrift

site als erstes und „bestes“ Suchergebnis von der Suchmaschine interpretiert und dem Suchenden ausgegeben werden, ungeachtet vom weiteren Inhalt der Seite.

Das größte Problem liegt darin, dass nur auf die absolute Anzahl der Verlinkungen geschaut wird, nicht aber auf die Qualität der verlinkenden Websites. Dort setzt das PageRank Verfahren ein. Es beurteilt jede Website anhand der zu ihr verlinkenden Websites, deren Qualität wiederum aus der Qualität anderer Websites errechnet wird. Dieses rekursive Verfahren beschreibt also die Linkstruktur des gesamten World Wide Webs, und nicht die Qualität der Seite selbst, in dem jedes Dokument, auch wenn über viel hintereinander folgende Links hinweg, Auswirkungen auf die Qualität anderer Dokumente hat.

Random¹³ Surfer Modell

Das PageRank Verfahren wird am besten durch das Modell eines zufälligen Benutzers des World Wide Webs dargestellt.

Ein Zufalls-Surfer befindet sich auf einer zufälligen Seite im World Wide Web, deren Wahrscheinlichkeit durch den PageRank hergeleitet ist. Er klickt nun mit einer bestimmten Wahrscheinlichkeit auf einen weiterführenden Link. Diese Wahrscheinlichkeit besteht einzig und allein darin, wie viele Links auf der Website vorhanden sind. Der PageRank für die weiterführende Seite wird aus den Wahrscheinlichkeiten aller zu ihr linkenden Seiten, deren Qualität und deren PageRank errechnet.

Da der Zufallssurfer sich nicht durch das Klicken auf unendlich viele Links durch das World Wide Web bewegt, sondern irgendwann seine Recherche aufgibt, muss die Übertragung der Qualität einer Website auf eine weiterführende abgedämpft werden.

Dieses Verfahren lässt sich durch einen relativ einfachen Algorithmus veranschaulichen.

$$PR(A) = (1-d) + d [PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n)]$$

¹³ Zufällig

PR(A) ist der PageRank der zu betrachtenden Seite die im Folgenden als Seite A bezeichnet wird.

PR(T_n) ist der PageRank der Seiten T_n, von denen ein Link auf die Seite A zeigt.

C(T_n) ist die Zahl der Links auf der Seite T_n

d ist der Dämpfungsfaktor, der zwischen 0 und 1 liegt und als (1-d) die Wahrscheinlichkeit angibt, mit der der Zufallssurfer die Verfolgung von weiteren Links abbricht. Bei Google ist dies der Wert 0,85 der aus realem Surfverhalten ermittelt wurde.

Aus rein mathematischer Sicht

Mathematisch gesehen, könnte man auch einfach versuchen, die Wahrscheinlichkeit anzugeben, mit der man eine beliebige Seite besucht. Hierzu muss man lediglich die absolute Anzahl der im World Wide Web zur Verfügung stehenden Seiten N in die PageRank Formel einbauen.

$$PR(A) = (1-d) / N + d [PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n)]$$

Die summe aller PageRank Werte nach der 2. Formel ist 1.

Bei Google wird allerdings nur die erste Formel verwendet, da dort die ständig wachsende Zahl der Seiten im World Wide Web nicht berücksichtigt werden muss. Aus der ersten Formel lassen sich zwei einfache Regeln ableiten:

Jede Seite kann einen minimalen Wert von (1-d) annehmen.

Der theoretische maximale PageRank Wert beträgt $d N + (1-d)$, der nur zustande kommt, wenn alle Seiten nur auf ein Seite verlinken würden und diese auch nur auf sich selbst.

Veranschaulichung an einem Beispiel

Anhand eines Beispiels versuche ich nun den PageRank zu veranschaulichen.

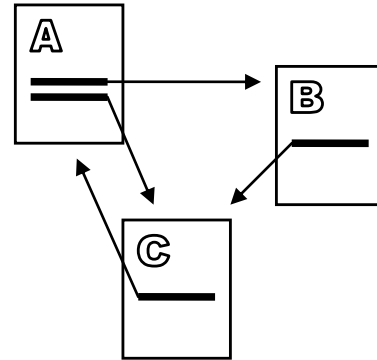
Zu Beginn wird ein Web aus drei Seiten betrachtet. Seite A verlinkt auf Seite B und C, Seite B nur auf C und Seite C nur auf A. Um ein anschauliches Ergebnis zu erhalten wird der Dämpfungsfaktor auf 0,6 und nicht wie sonst bei Google üblich auf 0,85 gesetzt.

Es werden nun 3 Gleichungen gebildet.

$$PR(A) = 1 - 0.6 + 0.6 PR(C)$$

$$PR(B) = 1 - 0.6 + 0.6 [PR(A) / 2]$$

$$PR(C) = 1 - 0.6 + 0.6 [PR(A) / 2 + PR(B)]$$



Aufgelöst auf den Pagerank der 3 Seiten bekommt man folgende Werte:

$$PR(A) = 1.1011271168545 \quad (\sim 1.163732)$$

$$PR(B) = 0.730327906816 \quad (\sim 0,643642)$$

$$PR(C) = 1.16854505824 \quad (\sim 1,192626)$$

Die Werte mit dem Dämpfungsfaktor von 0,85 sind in Klammern angegeben.

In diesem drei Seiten Web ist es nicht schwer das Gleichungssystem zu lösen um an die Lösung zu gelangen. Das World Wide Web besteht aber aus so vielen Seiten, dass es schlicht weg unmöglich ist ein Gleichungssystem aufzustellen und zu lösen. Aus diesem Grund wird der PageRank in Iterationen¹⁴ berechnet. Es ist nicht zwingend notwendig, dass einer Seite von Anfang an ein Startwert

	PR(A)	PR(B)	PR(C)
Iteration 1	1	0.7	1.3
Iteration 2	1.18	0.7	1.12
Iteration 3	1.072	0.754	1.174
Iteration 4	1.104	0.722	1.174
Iteration 5	1.104	0.731	1.164
Iteration 6	1.099	0.731	1.17
Iteration 7	1.102	0.73	1.168
Iteration 8	1.101	0.731	1.168
Iteration 9	1.101	0.73	1.169
Iteration 10	1.101	0.73	1.168
Iteration 11	1.101	0.73	1.169

zugewiesen wird, aber um mit so wenigen Iterationen wie möglich auf einen möglichst exakten Wert zu kommen, wird einer Seite der Startwert 1 zugewiesen, der ungefähr dem

¹⁴ Wiederholung des selben Rechenverfahrens.

Durchschnittlichen Wert aller Seiten entspricht. Die Entwicklung dieser Berechnung ist in der Tabelle auf der vorherigen Seite, mit Hilfe des drei Seiten Web mit einem Dämpfungsfaktor von 0,6 veranschaulicht.

Ausgehende Links und neue Seiten

Wenn man sich den PageRank Algorithmus anschaut könnte man meinen, dass jeder eingehende Link einer Website den PageRank erhöht. Aus dem Algorithmus geht hervor, dass sich der PageRank um genau $d * PR(X) / C(X)$ erhöht. Bei dieser Annahme ist allerdings noch nicht daran gedacht, dass das World Wide Web eine weit reichende Linkstruktur hat.

Man stelle sich ein Web aus vier Seiten vor. Seite A, B und C mit einem PageRank von 1, verlinken sich ausschließlich im Kreis, und Seite X mit einem PageRank von 10 verlinkt nur auf Seite A. Der Dämpfungsfaktor beträgt 0,75.

Nun werden 3 Gleichungen gebildet,

$$PR(A) = 1 - 0.75 + 0.75 [10 + PR(C)]$$

$$PR(B) = 1 - 0.75 + 0.75 \times PR(A)$$

$$PR(C) = 1 - 0.75 + 0.75 \times PR(B)$$

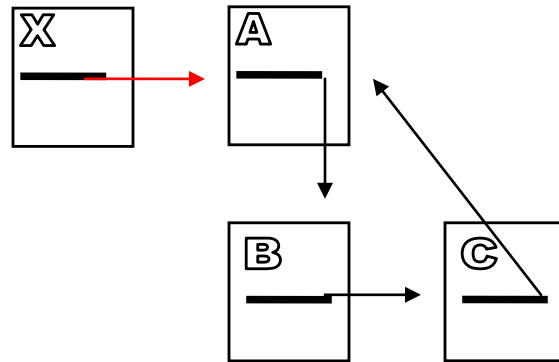
Deren Lösungen wie folgt sind:

$$PR(A) = 517/37 = 13.97$$

$$PR(B) = 397/37 = 10.73$$

$$PR(C) = 307/37 = 8.30$$

Die Summe des aufsummierten PageRanks von A, B und C ist 33 (vor der Verlinkung von X zu A war die Summe 3).



Nach der Formel $d * PR(X) / C(X)$ hätte sich der PageRank der Seite A um 7,5 auf 8,5 erhöhen müssen. Die Differenz von dem zu erwartendem und dem tatsächlichen Wert

von 5,47 ist erstaunlich, erscheint aber anhand des Random Surfer Modells leicht erklärbar. Wenn der Zufalls-Surfer in ein geschlossenes System kommt (hier die Seiten A, B und C) dann besucht er im Schnitt $(d/1-d)$ Seiten innerhalb des geschlossenen Systems. Diese Behauptung stellte Raph Levien und stützt sich darauf, dass sich der PageRank eines geschlossenen Systems um $(d / (1 - d)) * (PR(X) / C(X))$ erhöht, wenn eine Verlinkung der Seite X zu einem geschlossenem System stattfindet. In diesem Fall verändert sich der PageRank des geschlossenen Systems um

$$(0,75 / 0,25) * (10/1) = 30. \text{ (Hier mit } d = 0,75; (d / (1 - d)) = 3)$$

Mit dem original Dämpfungsfaktor von 0,85 ergibt sich eine Erhöhung eines abgeschlossenen Systems von $5,67 * (PR(X) / C(X))$. Hier zeigt sich zum besseren Verständnis, dass je größer der Dämpfungsfaktor ist, desto mehr haben Verlinkungen Einfluss auf Unterseiten einer Website. Allerdings wird der PageRank der Link erhaltenden Seite um so weniger vergrößert je größer der Dämpfungsfaktor ist.

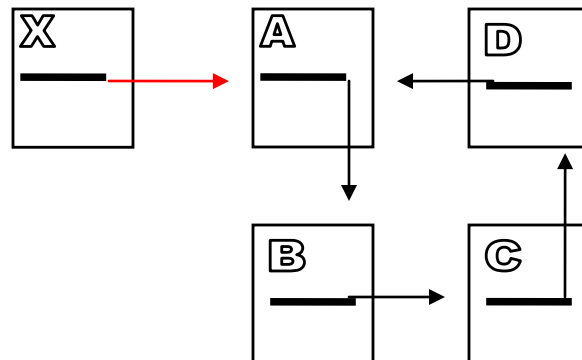
Wenn zu diesem abgeschlossenen System eine Seite D hinzugefügt wird, die sich in die Kreisverlinkung einfügt, ändert sich der PageRank der einzelnen Seiten und des Ganzen abgeschlossenen Systems wie folgt:

$$PR(A) = 0.25 + 0.75 [10 + PR(D)]$$

$$PR(B) = 0.25 + 0.75 \times PR(A)$$

$$PR(C) = 0.25 + 0.75 \times PR(B)$$

$$PR(D) = 0.25 + 0.75 \times PR(C)$$



Die Lösungen des Gleichungssystems:

$$PR(A) = 419/35 = 11.97$$

$$PR(B) = 323/35 = 9.23$$

$$PR(C) = 251/35 = 7.17$$

$$PR(D) = 197/35 = 5.63$$

Die Summe der PageRanks ist 34 (vor der Hinzufügung der Seite D war die Summe des PageRanks im abgeschlossenen Systems 33).

Durch eine hinzugefügte Seite steigt der PageRank des abgeschlossenen Systems um eins. Allerdings verlieren die bereits bestehenden Seiten an PageRank. Aus diesem Grund scheint es so, dass der PageRank vor allem kleine und kompakte Seiten bevorzugt behandelt. Große, und meist auch bekannte Seiten haben, im Gegenzug dazu meistens viele Seiten, auf denen Sie verlinkt sind, und können sich dem entsprechend auch viele Unterseiten leisten.

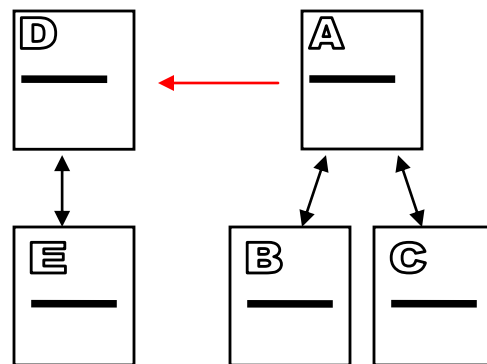
Wenn eine neue Seite zum World Wide Web hinzugefügt wird, vergrößert sich der aufsummierte PageRank um eins, wenn die Seite einen ausgehenden Link besitzt (siehe **Dangling Links** Seite 17). Um die Vergrößerung des PageRanks für seine Seite zu nutzen muss also darauf geachtet werden ihr so wenig PageRank wie möglich zukommen zu lassen, d.h. mehr ausgehende Links auf die eigene Seite, als eingehende von der eigenen Seite zu haben.

Eingehende Links

Bisher wurden nur ausgehende Links und deren Auswirkung auf Dokumente, oder in sich geschlossene Systeme behandelt. Anhand des Random Surfers kann erklärt werden, dass der Zufalls Surfer mit höherer Wahrscheinlichkeit auf einen Link klickt, der von einer Website mit Unterseiten zu einer andere Seite linkt, als auf einen Link der zu einer Seite innerhalb einer des Systems verweist, was zur Folge hat, dass der PageRank absinkt. Dass dies der Fall ist wird nun anhand des nächsten Beispiels erklärt.

Dem geschlossene System A, B, C in dem A auf B und C verlinkt und B und C jeweils auf A verlinken wird ein Link zum System D, E hinzugefügt, in dem D auf E und E auf D verlinkt. Der Dämpfungsfaktor beträgt 0,6.

Vor der Verlinkung beträgt der PageRank der einzelnen Seiten folgende Werte:



$$\text{PR}(A) = 1.374184$$

$$\text{PR}(B) = 0.812908$$

$$\text{PR}(C) = 0.812908$$

$$\text{PR}(D) = 1$$

$$\text{PR}(E) = 1$$

System A, B, C hat einen PageRank von 3; System D, E von 2.

Nach der Verlinkung von A zu D verändern sich die Werte für Beide Systeme drastisch.

$$\text{PR}(A) = 1.157865$$

$$\text{PR}(B) = 0.631649$$

$$\text{PR}(C) = 0.631649$$

$$\text{PR}(D) = 1.362518$$

$$\text{PR}(E) = 1.216319$$

System A, B, C hat einen PageRank von 2,421163; System D, E von 2,578837. Der aufaddierte PageRank des Gesamten Systems bleibt erhalten.

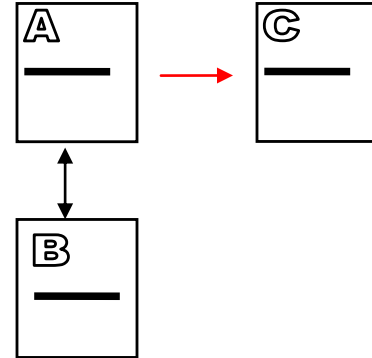
Da im PageRank Algorithmus durch die Anzahl der ausgehenden Links geteilt wird $\text{PR}(T1)/C(T1)$, ist es nicht weiter verwunderlich, dass der PageRank einer Seite mit einem neu auf ihr hinzugefügten Link absinkt. Da der aufaddierte PageRank des gesamten Systems erhalten bleibt, ist der Verlust einer Seite genau so groß wie der Gewinn der anderen Seite. Diesen Verlust kann man also auch mit der Formel $(d / (1-d)) \times (\text{PR}(X) / C(X))$ errechnen.

Um den Verlust, den man durch eine neu hinzugefügte Seite bekommt, so gering wie möglich zu halten, versucht man so wenige ausgehende Links wie möglich in einem Dokument zu haben, oder wenn überhaupt, dann nur zu qualitativ hochwertigen Seiten, die durch andere Kriterien, außer der PageRank Berechnung, einen positiven Einfluss auf die Qualität der Seite hat. Das Problem für Google besteht darin, dass Entwickler ihre Seiten auf die GoogleSuche ausrichten. Sie sind meistens auf einen hohen PageRank ausgelegt und besitzen somit keine unnötigen Links. Die Crawler können so viele kleine Seiten fast gar nicht finden, da sie nirgends verlinkt sind.

Dangling¹⁵ Links

Seit ein paar Jahren durchsucht Google nicht mehr nur Websites, sondern auch PDF und Word Dokumente, oder Powerpoint Präsentationen.

In einer solchen Datei befinden sich aber meistens keine weiterführenden Links. Solche Dokumente sind bei Google als „Dangling Links“ bekannt. Sie haben bei der Berechnung der PageRanks eine besondere Stellung, die sie nicht zu Unrecht haben, wie das folgende Beispiel zeigt.



Bei der Betrachtung eines drei Seiten Webs, in dem sich Seite A und B gegenseitig verlinken, und von Seite A ein Dangling Link auf eine dritte Seite C verweist, die selber keine ausgehenden Links besitzt. Bei einem Dämpfungsfaktor von 0,6 erhält man folgende PageRanks für die Seiten A, B und C:

$$A = 0,435699$$

$$B = 0,336117$$

$$C = 0,336117$$

Der aufaddierte PageRank der drei Seiten bleibt nicht erhalten. Er sinkt von 3 auf ~1,1. Der PageRank der Seite A sinkt um mehr als die Hälfte und das nur, weil sie zum Beispiel auf ein Referat im PDF Format verweist. Um diesen ungewollten Effekt zu verhindern, werden die Dangling Links aus der Datenbank entfernt. Die Seite C würde dann in diesem Beispiel einen PageRank von $(1 - 0,6) + 0,6 * (PR(A) / 2) = 0,7$ haben, da durch das entfernen des Links auf C die Seite A, beide wieder einen PageRank von 1 haben. Somit ist der aufaddierte PageRank des Systems bei 2,7. Es fehlt zwar noch etwas, aber der Dangling Link hat keinen Einfluss mehr auf die Seite A oder B.

¹⁵ Verflixt oder hängend, baumelnd

Andere Bewertungskriterien

Das grundlegende Geheimnis des PageRanks ist gelüftet. Trotzdem schaffen es die anderen Suchmaschinen immer noch nicht, an die Qualität der Suchergebnisse von Google heranzukommen. Der Grund dafür ist, dass es noch andere Faktoren gibt, die die Anordnung der Suchergebnisse beeinflussen. Vorne an steht dabei die Folge der gesuchten Begriffe innerhalb eines Dokuments.

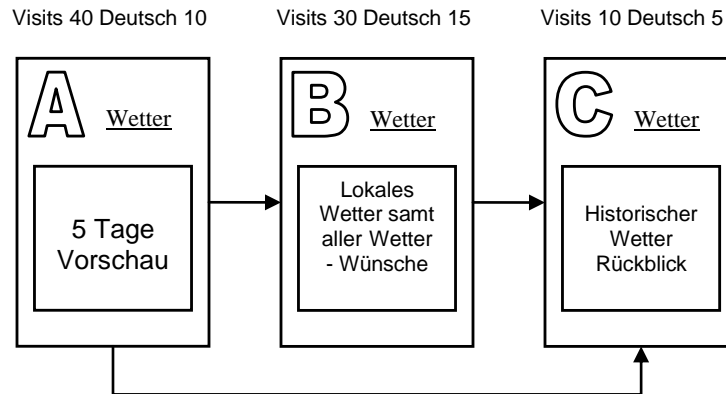
Ein Beispiel von Wolfgang G. Stock von der Heinrich-Heine-Universität Düsseldorf zeigt dies sehr gut. Wort A und Wort B sind hier die beiden Suchbegriffe. Nun wird der Abstand dieser beiden Suchbegriffe in den zu sortierenden Texten errechnet.

- Text 1. Wort A: Position 22 – Wort B: Position 25: Diff.: 3
- Text 2. Wort A: Position 25 – Wort B: Position 22: Diff.: -3
- Text 3. Wort A: Position 1 – Wort B: Position: 2: Diff.: 1

Nun wird aufsteigend nach der Differenz der Wörter bei 1 beginnend sortiert. Sollte die Differenz negativ sein werden sie hinten angestellt, da die Reihenfolge der Suchbegriffe beachtet wird. Das Ranking dieser 3 Texte wäre dann: Text 3 vor Text 1 vor Text 2.

Ein anderes Kriterium könnte natürlich auch die Häufigkeit der Besuche in einem bestimmten Zeitraum einer Website sein. Es könnten aber auch die Häufigkeit der Besuche aus einem bestimmten Land, oder eine Verbindung aus beidem ein Kriterium sein. An dem folgenden Beispiel werden noch einmal die verschiedenen Bewertungskriterien deutlich.

In diesem 3 Seitenweb wird nun nach den verschiedenen Kriterien durchsucht.



Geordnet nach der oben beschriebenen Textstatistik und dem Suchbegriff Wetter, würde die Reihenfolge so aussehen: B vor C vor A.

Nach dem PageRank wären sie in der Reihenfolge C vor A vor B.

Nach der Anzahl der Besuche einer Seite in der Reihenfolge A vor B vor C

und bei der Bewertung der spezifischen Besuche aus einem bestimmten Land (hier Deutschland) in der Reihenfolge B vor A vor C.

Hier zeigt sich sehr anschaulich, dass der Erfolg Googles nicht einzig und allein vom PageRank abhängig ist, sondern dass Google es schafft, verschiedene Bewertungskriterien so miteinander zu verrechnen, dass für die Suchenden meistens das optimale Ergebnis erscheint.

Fazit

Das Problem von Google ist es, dass man wenige gute Informationen von Google selbst bekommt. Die Theorien wie Google nun wirklich funktioniert sind zwar meistens in sich schlüssig aber nicht bewiesen, und werden von Google erst gar nicht kommentiert. Selbst mit Google lassen sich kaum brauchbaren Informationen von Google über die Funktionsweise Googles finden.

Diese Problematik zeigt wie verschlossen Google ist. Doch genau diese Verschlossenheit gewährleistet, dass Google Marktführer bei den Suchmaschinen im World Wide Web bleibt. Das aber wahrscheinlich nur so lange, bis jemand auf die gleiche Idee kommt. Für diesen Fall sorgt Google Heute schon vor, in dem sie in andere Gebiete vordringen und sich nicht mehr nur auf die Onlinesuche beschränken.

Der Aufbau und das Ausmaß des Googlecluster und die daraus resultierende Geschwindigkeit für die Googlesuche sind etwas das mich persönlich beeindruckt hat. Angefangen haben sie mit einem einfachen Algorithmus, den sie bis heute zu einem der größten Geheimnisse gemacht haben. Somit ist Google mit einfachsten Mitteln zu einem der größten und bekanntesten und vielleicht sogar gefährlichsten Unternehmen der Welt geworden.

Literaturverzeichnis

Primärliteratur

Lawrence Page, United States Patent: **Method for node ranking in a linked database**
<http://v3.espacenet.com/publicationDetails/biblio?CC=US&NR=6285999&KC=&FT=E>
(09.05.2008)

Sergey Brin and Lawrence Page, **The Anatomy of a Large-Scale Hypertextual Web Search Engine**, <http://ilpubs.stanford.edu:8090/361/1/1998-8.pdf> (08.05.2008)

Google Architektur

Architektur von Google http://www.beyond-media.net/architektur_google.html
(08.05.2008)

IEEE Computer Society, **WEB SEARCH FOR A PLANET:THE GOOGLE CLUSTER ARCHITECTURE** <http://labs.google.com/papers/googlecluster-ieee.pdf> (08.05.2008)

Lars Geiger, Google Hardware Architektur, <http://www.fmi.uni-stuttgart.de/szs/teaching/ws0506/google/ausarbeitungen/geiger.pdf> (08.05.2008)

Official Google Data Center Video Tours, <http://www.greenm3.com/2009/04/official-google-data-center-video-tours.html> (06.05.2008)

Stephen Shankland, Google uncloaks once-secret server, http://news.cnet.com/8301-1001_3-10209580-92.html (10.05.2008)

Videotour durch ein Google-Rechenzentrum <http://www.golem.de/0904/66376.html>
(08.05.2008)

Crawler

Google, Corporate Information, <http://www.google.com/corporate/tech.html> (12.05.2008)

How Google Works, http://www.googleguide.com/google_works.html (09.05.2008)

Ronny Harbich, Webcrawling – Die Erschließung des Webs, <http://www-e.uni-magdeburg.de/harbich/webcrawling/webcrawling.pdf> (08.05.2008)

PageRank Algorithmus

Andreas Jaster, Der PageRank-Algorithmus von Google,
<http://knol.google.com/k/andreas-jaster/pagerank/35kq0u83x334w/2#> (08.05.09)

Der PageRank-Algorithmus, <http://www.suchmaschinen-doktor.de/algorithmen/pagerank/beispielrechnungen.html> (09.05.09)

eFactory.de, Überblick über das PageRank-Verfahren der Suchmaschine Google,
<http://pr.efactory.de/d-index.shtml> (09.05.09)

Klaus Patzwaldt, Das PageRank Verfahren der Suchmaschine Google, <http://www.at-web.de/suchmaschinenoptimierung/pagerank.htm> (11.05.09)

PageRank Rechner, www.pagerank.dk/PageRank.xls (11.05.09)

Wolfgang G Stock, Information Retrieval: Informationen suchen und finden,
 Veröffentlicht von Oldenbourg Wissenschaftsverlag, 2006,
<http://books.google.de/books?id=mjRxvoh4NHYC> (08.05.09)

Sonstige

<http://de.wikipedia.org/wiki>
 /Suchmaschine (12.08.09)
 /Google_File_System (09.05.09)
 /Webcrawler (10.05.09)
 /Computercluster (10.05.09)

<http://en.wikipedia.org/wiki/>
 /PageRank (08.05.2008)
 /Google_platform (08.05.2008)