Webmaster aufgepaßt: Was Sie schon immer über den PageRank von Google wissen wollten ...

... und nie zu fragen wagten

Ein Aufklärungsbericht, den ich im Internet für meine Leser herausgegraben habe. Sehr interessant!

Überblick über das PageRank-Verfahren der Suchmaschine Google

Im Verlauf der letzten Jahre hat sich Google weltweit zur bedeutendsten Suchmaschine entwickelt. Maßgebend verantwortlich hierfür war neben einer hohen Performance und einer großen Benutzerfreundlichkeit vor allem die anderen Suchmaschinen teilweise weit überlegene Qualität der Suchergebnisse. Diese Qualität der Suchergebnisse beruht ganz wesentlich auf dem PageRank-Verfahren.

An dieser Stelle soll ein möglichst breiter Überblick über alle Aspekte des PageRank-Verfahrens wiedergegeben werden. Unser Überblick stützt sich dabei im Kern auf Veröffentlichungen der Google-Gründer Lawrence Page und Sergey Brin aus ihrer Zeit als Graduiertenstudenten an der Stanford University.

Vielerorts wird angeführt, dass seit den Forschungsarbeiten am PageRank-Verfahren vor allem angesichts der Dynamik des Internets zu viel Zeit vergangen ist, als dass die veröffentlichten Dokumente immer noch für die Bewertungsmethodik der Suchmaschine Google maßgebend sind. Es soll auch nicht bezweifelt werden, dass im Verlauf der letzten Jahre mit großer Wahrscheinlichkeit zahlreiche Änderungen, Anpassungen und Modifikationen am ursprünglichen PageRank-Algorithmus stattgefunden haben. Allerdings war gerade das PageRank-Verfahren ein wichtiger Faktor für den Erfolg der Suchmaschine Google, womit zumindest das Konzept des PageRank-Verfahrens immer noch grundlegend sein sollte

Das PageRank-Konzept

Im Zuge der Entwicklung des World Wide Webs wurden verschiedene Verfahren zur Bewertung von Webseiten mit dem Ziel der Relevanzbeurteilung durch Suchmaschinen entwickelt. Ein aus unmittelbar einleuchtenden Gründen auch heute immer noch von praktisch allen Suchmaschinen genutzter Maßstab ist das Vorkommen eines Suchbegriffs in den Inhalten einer Webseite. Dieses Vorkommen wird nach den verschiedensten Kriterien wie etwa der relativen Häufigkeit des Vorkommens (der sog. Keyword-Dichte), den Stellen des Vorkommens des Suchbegriffs oder auch der Exponiertheit des Suchbegriffs im Dokument gewichtet.

Aus der Absicht, Suchmaschinen resistent gegen Webseiten zu machen, die auf der Basis von Analysen der inhaltsspezifischen Bewertungskriterien generiert wurden (Doorway Pages), entstand das Konzept der Link-Popularität. Dabei fließt die Anzahl der eingehenden Links für ein Dokument als ein grundsätzliches Kriterium für die Bedeutung einer Webseite in die Relevanzbeurteilung ein. Diesem Ansatz liegt zu Grunde, dass ein Dokument um so wichtiger

ist, je häufiger es von anderen verlinkt wird. Hierdurch wird weitestgehend verhindert, dass automatisch generierte "suchmaschinenoptimierte" Webseiten ohne jeglich Einbindung in das WWW oben in den Suchmaschinenergebnissen erscheinen. Es zeigte sich allerdings, dass auch das Konzept der Link-Popularität schnell von Webmastern antizipiert werden konnte, indem sie von ebenso unbedeutenden, automatisch generierten Seiten eingehende Links für Doorway Pages schufen.

Im Gegensatz zum Konzept der Link-Popularität nutzt das PageRank-Konzept nicht einfach die absolute Anzahl eingehender Links für die Beurteilung der Bedeutung einer Webseite. Die Argumentation der Google-Gründer gegen das Konzept der einfachen Link-Popularität war, dass ein Dokument zwar bedeutsam ist, wenn es von vielen anderen verlinkt wird, nicht jedes verlinkende Dokument ist jedoch gleichwertig. Vielmehr sollte einem Dokument - völlig unabhängig von seinen Inhalten - ein hoher Rang zugewiesen werden, wenn es von anderen bedeutenden Dokumenten verlinkt wird.

Die Bedeutsamkeit eines Dokuments bestimmt sich im Rahmen des PageRank-Konzepts also aus der Bedeutsamkeit der darauf verlinkenden Dokumente. Deren Rang wiederum bestimmt sich ebenfalls aus dem Rang verlinkender Dokumente. Die Bedeutsamkeit eines Dokuments definiert sich stets rekursiv aus der Bedeutsamkeit anderer Dokumente. Da - wenn auch über viele hintereinanderfolgende Links hinweg - der Rang eines jeden Dokuments eine Auswirkung auf den Rang eines jeden anderen hat, beruht das PageRank-Konzept letztlich auf der Linkstruktur des gesamten Webs. Obwohl diese ganzheitliche Betrachtung des WWW es nicht vermuten lässt, gelang es Page und Brin das PageRank-Konzept mittels eines relativ trivialen Algorithmus umzusetzen.

Der PageRank-Algorithmus

Der ursprüngliche PageRank-Algorithmus wurde von Lawrence Page und Sergey Brin mehrfach beschrieben. Er hat die folgende Form:

$$PR(A) = (1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

Hierbei ist:

- PR(A) der PageRank einer Seite A,
- PR(Ti) der PageRank der Seiten Ti, von denen ein Link auf die Seite A zeigt,
- C(Ti) die Gesamtanzahl der Links auf Seite Ti und
- d ein Dämpfungsfaktor (Damping Factor),
  wobei $0 <= d <= 1$ ist.

Das PageRank-Verfahren bewertet damit grundsätzlich nicht Websites in ihrer Gesamtheit, sondern basiert ausschließlich auf der Beziehung einzelner Webseiten zueinander. Der PageRank einer Seite A bestimmt sich dabei rekursiv aus dem PageRank derjenigen Seiten, von denen ein Link auf die Seite A zeigt.

Der PageRank der Seiten Ti, die auf eine Seite A verlinken, fließt nicht gleichmäßig in den PageRank von Seite A ein. Der PageRank einer Seiten T wird stets anhand der Anzahl C(T) der von Seite T ausgehenden Links gewichtet. Das bedeutet, dass je mehr ausgehende Links eine Seite T hat, umso weniger PageRank gibt sie an Seite A weiter.

Der anhand der Anzahl an ausgehenden Links gewichtete PageRank der Seiten Ti wird nun

addiert. Dies hat zur Folge, dass jeder zusätzliche eingehende Link für eine Seite A stets den PageRank dieser Seite A erhöht.

Schließlich wird die Summe der gewichteten PageRanks der Seiten Ti mit dem Dämpfungsfaktor d, der stets zwischen 0 und 1 liegt multipliziert. Hierdurch wird das Ausmaß der Weitergabe des PageRanks von einer Seite auf einer andere verringert.

Das Random Surfer Modell

Lawrence Page und Sergey Brin bieten in ihren Veröffentlichungen eine sehr einfache, intuitive Rechtfertigung des PageRank-Algorithmus an. Sie betrachten PageRank-Verfahren als ein Modell zur Abbildung von Benutzer-Verhalten. Hierzu führen sie einen Zufalls-Surfer an, der von einer Webseite zur nächsten jeweils beliebige Links verfolgt, ohne dabei auf Inhalte zu achten.

Der Zufalls-Surfer befindet sich mit einer bestimmten Wahrscheinlichkeit auf einer Website, die sich aus deren PageRank herleiten lässt. Die Wahrscheinlichkeit, dass der Zufalls-Surfer nun einen bestimmten Link verfolgt, ergibt sich dann einzig und allein daraus, aus wievielen Links er die Auswahl hat. Aus diesem Grunde fließt der PageRank einer verlinkenden Seite stets nach der Anzahl Ihrer ausgehenden Links gewichtet in die PageRank Berechnung einer verlinkten Seite ein.

Die Wahrscheinlichkeit, dass der Zufalls-Surfer auf eine Seite gelangt, ist also die Summe der Wahrscheinlichkeiten, mit der er von einer verlinkenden Seite den entsprechenden Link verfolgt. Nun wird allerdings die Wahrscheinlichkeit, mit der der Zufalls-Surfer auf eine Seite gelangt, um den Faktor d gedämpft. Dies hat im Rahmen des Random Surfer Modells den Hintergrund, dass der Zufalls-Surfer nicht unendlich viele Links verfolgt. Nach einer bestimmten Zeit wird er gelangweilt und ruft eine beliebige andere Webseite auf.

Die Wahrscheinlichkeit, mit der der Zufalls-Surfer die Verfolgung von Links nicht abbricht und somit weiterklickt, wird durch den Dämpfungsfaktor d angegeben, der abhängig von der Höhe der Wahrscheinlichkeit einen Wert von 0 bis 1 annimmt. Je höher d ist, um so wahrscheinlicher ist es, dass der Zufalls-Surfer Links verfolgt. Da der Zufalls-Surfer nach dem Abbruch der Link-Verfolgung eine beliebige Seite aufruft, geht die Wahrscheinlichkeit mit er er dies tut, mit dem Wert (1-d) als Konstante in die Berechnung des PageRanks einer jeden Seite ein.

Abweichende Formulierung des PageRank-Algorithmus

Lawrence Page und Sergey Brin bieten in ihren Veröffentlichungen zwei unterschiedliche Versionen des PageRank-Algorithmus an. In dieser zweiten Version bestimmt sich der PageRank einer Seite A wie folgt:

$$PR(A) = (1-d) / N + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

Hierbei ist N die Anzahl aller Seiten des Webs. Diese zweite Version des PageRank-Algorithmus unterscheidet sich allerdings nicht grundlegend von der ersten. In der zweiten Version beschreibt der PageRank einer Seite im Sinne des Random Surfer Modells lediglich die tatsächliche Wahrscheinlichkeit, mit der der Zufalls-Surfer nach dem Verfolgen vieler Links eine Seite erreichen wird. Dieser Algorithmus bildet damit eine Wahrscheinlichkeitsverteilung über alle Seiten des Webs ab. Die Summe aller PageRank-

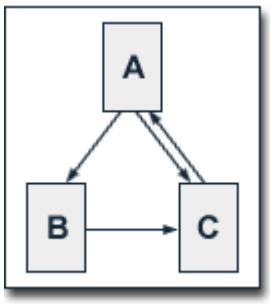
Werte des Webs ist damit bei dieser Version des Algorithmus gleich 1.

In der oben genannten, ersten Version erfolgt eine Gewichtung der Wahrscheinlichkeit des Besuchs einer Seite nach der Anzahl der Seiten des Webs. Demnach ist der PageRank in dieser Version im Grunde der Erwartungswert für den Besuch des Zufalls-Surfers auf einer Seite, wenn er hierfür Anläufe in genau der Höhe der Anzahl der Seiten des Webs nimmt. Bestünde das Web also aus 100 Seiten, und eine Seite hat einen PageRank von 2, so würde der Zufalls-Surfer sie bei 100 "Surfgängen" im Mittel zweimal erreichen.

Wie bereits erwähnt, unterscheiden sich die beiden Versionen des Algorithmus sich nicht grundlegend. Letztlich muss der PageRank einer Seite aus der Algorithmus-Version 2 lediglich mit der Anzahl der Webseiten multipliziert werden, um zum PageRank der Algorithmus-Version 1 zu gelangen. Selbst Page und Brin ist in Ihrer wohl bekanntesten Veröffentlichung "The Anatomy of a Large-Scale Hypertextual Web Search Engine" der Fehler unterlaufen, die erste Version des PageRank-Algorithmus als Wahrscheinlichkeitsverteilung zu charakterisieren, bei der die Summe der PageRank-Werte aller Seiten gleich eins sei.

Im Folgenden wird für die weiteren Betrachtungen der oben zuerst genannte Algorithmus verwandt. Dies hat den einfachen Hintergrund, dass Berechnungen hiermit wesentlich einfacher sind, da die Größe des Webs vollkommen außer Acht gelassen werden kann.

Die Eigenschaften des PageRank



Die Eigenschaften des PageRank sollen jetzt anhand eines Beispieles veranschaulicht werden.

Hierzu wird ein kleines 3-Seiten-Web aus den Seiten A, B und C betrachtet, wobei Seite A sowohl auf Seite B als auch auf Seite C verlinkt. Seite B verlinkt lediglich auf Seite C und Seite C wiederum verlinkt auf Seite A. Der Dämfungsfaktor d wird Angaben von Lawrence Page und Sergey Brin zufolge für tatsächliche Berechnungen üblicherweise auf 0.85 gesetzt. Der Einfachheit halber wird d an dieser Stelle ein Wert von 0.5 zugewiesen, wobei die Höhe von d zwar Auswirkungen auf den PageRank hat, das hier zu erläuternde Prinzip jedoch nicht beeinflusst. Es ergeben sich die folgenden Gleichungen für den PageRank der einzelnen Seiten:

PR(A) = 0.5 + 0.5 PR(C)
PR(B) = 0.5 + 0.5 (PR(A) / 2)
PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B))

Dieses Gleichungssystem lässt sich sehr einfach für den PageRank der einzelnen Seiten lösen. Es ergeben sich die folgenden Werte:

PR(A) = 14/13 = 1.07692308
PR(B) = 10/13 = 0.76923077
PR(C) = 15/13 = 1.15384615

Es zeigt sich, dass die Summe der PageRanks aller Seiten gleich drei und somit gleich der Anzahl der Seiten ist. Dies ist keine spezifisches Ergebnis für unser Beispiel, da der PageRank Algorithmus einen Erwartungswert für den Besuch von Seiten bei Anläufen in

Höhe der Anzahl der Seiten darstellt.

Für ein kleines 3-Seiten-Beispiel lässt sich ein Gleichungssystem unproblematisch lösen. Das tatsächliche WWW besteht jedoch mittlerweile aus mehreren Milliarden Webseiten, so dass die Lösung eines entsprechenden Gleichungssystems nicht mehr möglich ist.

Die iterative Berechnung des PageRank

Aufgrund der Größe des Webs erfolgt in der Praxis der Suchmaschine Google eine näherungsweise, iterative Berechnung des PageRank. Dies bedeutet, dass zunächst jeder Seite ein PageRank zugewiesen wird, und anschließend der PageRank aller Seiten in mehreren Berechnungsrunden ermittelt wird. Diese näherungsweise Berechung soll wiederum anhand unseres kleinen Beispiels demonstriert werden, wobei als Ausganswert für den PageRank einer jeden Seite zunächst 1 angenommen wird.

Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	0.75	1.125
2	1.0625	0.765625	1.1484375
3	1.07421875	0.76855469	1.15283203
4	1.07641602	0.76910400	1.15365601
5	1.07682800	0.76920700	1.15381050
6	1.07690525	0.76922631	1.15383947
7	1.07691973	0.76922993	1.15384490
8	1.07692245	0.76923061	1.15384592
9	1.07692296	0.76923074	1.15384611
10	1.07692305	0.76923076	1.15384615
11	1.07692307	0.76923077	1.15384615
12	1.07692308	0.76923077	1.15384615

Es zeigt sich, dass sich in unserem Beispiel bereits nach sehr wenigen Iterationen eine sehr gute Näherung an die tatsächlichen Werte ergibt. Für die Berechnung des PageRanks für das komplette WWW werden von Lawrence Page und Sergey Brin ca. 100 Iterationen als hinreichend genannt.

Entscheidend ist, dass die Summe der PageRanks aller Seiten nach der Durchführung der iterativen Berechnung gegen die Anzahl aller Seiten konvergiert. Der durchschnittliche PageRank aller Seiten geht mithin gegen 1. Jede Seite hat einen minimalen PageRank von (1-d). Der theoretisch maximale PageRank einer Seite beträgt dN+(1-d), wobei N die Anzahl aller Webseiten ist. Dieser theoretische Wert käme zustande, wenn sämtliche Webseiten ausschließlich auf eine Seite verlinken, und auch diese wiederum ausschließlich auf sich selbst verlinkt.

Die Implementierung des PageRank in die Suchmaschine Google

Für die Implementierung des PageRank ist von zentraler Bedeutung, auf welche Art und Weise der PageRank in die generelle Bewertung von Webseiten durch die Suchmaschine Google einfließt. Das Verfahren wurde von Lawrence Page und Sergey Brin mehrfach in

ihren Veröffentlichungen beschrieben. Ursprünglich basierte die Seitenbewertung durch Google auf drei Faktoren:

- Seitenspezifische Faktoren
- Ankertext eingehender Links
- PageRank

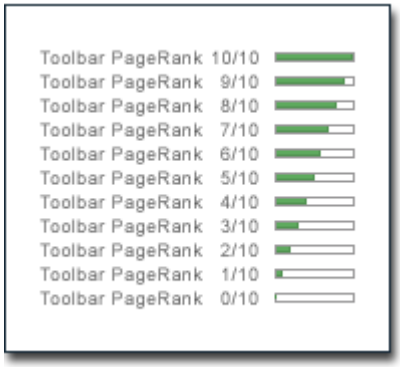
Zu den seitenspezifischen Faktoren zählen neben den konkreten Textinhalten etwa auch der Inhalt des Title-Tags und die URL einer Seite. Es ist mehr als wahrscheinlich, dass seit der Veröffentlichung dieser Punkte weitere Faktoren hinzugekommen sind. Dies soll an dieser Stelle jedoch nicht interessieren.

Bei Suchanfragen wird aus den seitenspezifischen Faktoren und den Ankertexten eingehender Links für den Suchbegriff eine nach Position und Grad der Hervorhebung gewichteter IR-Wert berechnet. Die Bewertung für die Relevanz einer Webseite für eine konkrete Suchanfrage wird nun mit dem PageRank als Indikator für die ganz allgemeine Bedeutsamkeit der Webseite kombiniert. Dieses Kombinieren erfolgt in multiplikativer Form. Dass hier kein additives Verfahren eingesetzt wird ist unmittelbar einleuchtend, da ansonsten Seiten mit einem sehr hohen PageRank auch auf Suchanfragen hin gefunden werden könnten, obwohl sie in keinerlei Zusammenhang zum gesuchten Begriff stehen.

Insbesondere bei aus mehreren Begriffen bestehenden Suchanfragen zeigt sich ein deutlich größerer Einfluss der inhaltsspezifischen Bewertungskomponenten. Der Einfluss des PageRank hingegen wird eher bei unspezifischen, aus lediglich einem Suchbegriff bestehenden Anfragen deutlich. Gerade für Mehr-Begriffs-Anfragen ist es möglich, mit den klassischen Mitteln der Suchmaschinen-Optimierung Listungen vor Seiten zu erlangen, die einen weitaus höheren PageRank-Wert inne haben.

Bei der Optimierung für Suchbegriffe, für die in den Suchmaschinen ein großer Wettbewerb herrscht, ist ein hoher PageRank-Wert unerlässlich für eine hohe Suchmaschinen-Position, selbst wenn die Seite selbst den klassischen Kriterien der Suchmaschinen-Optimierung folgt. Dies liegt darin begründet, dass die Wertung des zusätzlichen Vorkommens eines Suchbegriffs innerhalb eines Dokuments sowie in den Ankertexten von eingehenden Links mit der Häufigkeit des Vorkommens abnimmt, um Spam durch oftmalige Keyword-Wiederholungen zu vermeiden. Damit sind die Möglichkeiten zur Seitenoptimierung im klassischen Sinne beschränkt, und bei hohem Wettbewerb in Suchmaschinen für einen Suchbegriff wird der PageRank zum ausschlaggenden Faktor

Die PageRank Anzeige der Google Toolbar



Einen großen Bekanntheitsgrad erlangte der PageRank durch seine Anzeige in der Google Toolbar. Die Google Toolbar ist ein Browser-Plug-In für den Microsoft Internet Explorer, das von der Google Website herunter geladen werden kann und zahlreiche Erleichterungen für die Google-Suche bereithält.

Die Google Toolbar zeigt den PageRank einer Seite auf einer Skala von 0 bis 10 an. Zunächst ist der PageRank an

der Breite des grünen Balkens in der Anzeige ersichtlich. Fährt der Benutzer mit der Maus über die Anzeige, gibt die Toolbar darüberhinaus den Wert des Toolbar-PageRank an.

Vorsicht: Die PageRank-Anzeige zählt zu den "Advanced Features" der Google Toolbar. Sobald diese "Advanced Features" aktiviert sind, sammelt Google über die Toolbar Daten über das Benutzerverhalten. Außerdem führt die Toolbar selbstständig Updates durch, ohne dass der Benutzer über das Herunterladen der neuen Version informiert wird. Dies bedeutet letztlich, dass Google Zugriff auf die Festplatte des Benutzers hat.

Der tatsächliche PageRank, der für eine Seite theoretisch maximal einen Wert von dN+(1-d) annehmen kann, wobei N die Anzahl aller Seiten des Webs ist und d üblicherweise auf 0.85 gesetzt wird, muss für die Anzeige in der Google Toolbar skaliert werden. Es wird im Allgemeinen davon ausgegangen, dass die Skalierung nicht linear sondern logarithmisch erfolgt. Bei einem Dämpfungsfaktor von 0.85 und einem damit verbundenen minimalen PageRank von 0.15 sowie einer angenommenen logarithmischen Basis von 6 ergäbe sich das folgende Bild für die Skalierung:

Toolbar-PageRank	Tatsächlicher PageRank
0/10	0.15 - 0.9
1/10	0.9 - 5.4
2/10	5.4 - 32.4
3/10	32.4 - 194.4
4/10	194.4 - 1,166.4
5/10	1,166.4 - 6,998.4
6/10	6,998.4 - 41,990.4
7/10	41,990.4 - 251,942.4
8/10	251,942.4 - 1,511,654.4
9/10	1,511,654.4 - 9,069,926.4
10/10	9,069,926.4 - 0.85 × N + 0.15

Ob tatsächlich eine mathematisch strikte logarithmische Skalierung erfolgt ist natürlich ungewiss. Wahrscheinlich erfolgt eine manuelle Skalierung, die einem logarithmischen Schema folgt, damit Google die volle Kontrolle darüber behält, wie viele Seiten einen bestimmten Toolbar-PageRank inne haben. Diesem Schema dürfte allerdings eine logarithmische Basis von 6 bis 7 zu Grunde liegen, was sich etwa ansatzweise aus der Anzahl der von Google angezeigten eingehenden Links mit einem Toolbar-PageRank größer 4 für Seiten mit einem sehr hohen Toolbar-Pagerank herleiten lässt.

Die Datenkommunikation der Toolbar

Auch Webmaster, die aufgrund von Sicherheitsbedenken die Google Toolbar oder auch den Internet Explorer nicht dauerhaft nutzen möchten, haben eine Möglichkeit zum Einblick in die PageRank-Werte ihrer Seiten. Google übermittelt den PageRank in einfachen Textdateien an die Toolbar. Früher geschah dies per XML. Der Wechsel zu Textdateien fand im August 2002 statt.

Die PageRank-Textdateien können direkt von der Domain www.google.com abgerufen werden. In ihrer Grundform sehen die URLs der Dateien folgendermaßen aus (ohne

Zeilenumbrüche):

http://www.google.com/search?
client=navclient-auto&
ch=0123456789&
features=Rank&
q=info:http://www.domain.com/

Die PageRank-Dateien bestehen aus einer Zeile. Der PageRank-Wert ist die letzte Ziffer in dieser Zeile.

Die oben in der URL dargestellten Parameter sind unerlässlich für die Anzeige der PageRank-Dateien im Browser. So identifiziert sich mit dem Wert "navclient-auto" für den Parameter "client" die Toolbar; mit dem Parameter "q" wird die abgefragte URL übermittelt. Der Wert "Rank" für den Parameter "features" legt fest, dass die PageRank-Dateien abgerufen werden. Wird dieser Parameter weggelassen, werden auch weiterhin XML-Dateien übermittelt. Der Parameter "ch" wiederum übergibt eine Prüfsumme für die URL, wobei sich diese Prüfsumme im Zeitablauf für einzelne URLs lediglich bei Updates der Toolbar ändern kann.

Um die Prüfsummen einzelner URLs herauszufinden ist es damit erforderlich, die Toolbar zumindest einmal zu installieren. Hierbei wird dann vielerorts der Einsatz von Packet Sniffern, lokalen Proxies und ähnlichem empfohlen, um die Kommunikation zwischen Toolbar und Google aufzuzeichnen. Dies ist allerdings nicht zwingend erforderlich, da die PageRank-Dateien vom Internet Explorer gecached werden und somit die Prüfsummen im Ordner Temporary Internet Files eingesehen werden können. Die PageRank-Dateien können hiermit dann auch z.B. in anderen Browsern als dem Internet Explorer angezeigt werden, ohne dass Googles 36-Jahres-Cookies akzeptiert werden müssen.

Da die PageRank-Dateien im Browser-Cache gespeichert werden und somit offen einsehbar sind, und sofern eine Abfrage nicht automatisiert erfolgt, sollte dies keine Verletzung von Googles Dienstleistungsbedingungen darstellen. Es ist allerdings Vorsicht geboten. Die Toolbar übermittelt einen eigenen User-Agent an Google. Es ist:

Mozilla/4.0 (compatible; GoogleToolbar 1.1.60-deleon; *OS* SE 4.10)

Hierbei ist 1.1.60-deleon eine Toolbar-Version, die sich natürlich ändern kann, und OS das Betriebssystem des jeweils eingesetzten Rechners. Google kann also nachprüfen, ob eine direkte Anfrage über den Browser erfolgt, sofern kein Proxy zwischengeschaltet und der User-Agent entsprechend modifiziert wird.

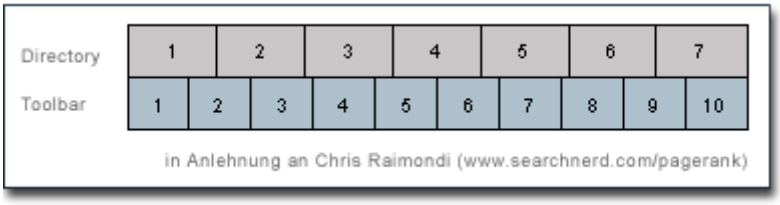
Beim Blick in den Cache des IE wird man in der Regel feststellen, dass die PageRank-Dateien nicht von der Domain www.google.com, sondern von IPs wie z.B. 216.239.33.102 abgerufen werden. Ebenso enthalten die URLs häufig einen weiteren Parameter "failedip" mit Werten wie z.B. "216.239.35.102;1111". Die IPs sind jeweils einem der derzeit sieben sich im Einsatz befindlichen Rechenzentren Googles zugeordnet. Wozu der Parameter "failedip" tatsächlich genutzt wird, ist unklar. Hintergrund der unmittelbaren Abfrage der PageRank-Dateien bei einzelnen IPs ist wohl der Versuch, die PageRank-Anzeige insbesondere in den Zeiten des "Google Dance" besser zu steuern.

Die PageRank Anzeige der Google Directory

Denjenigen, denen der Abruf der PageRank-Dateien zu kompliziert ist, bleibt schließlich mit der Google Directory (directory.google.com) noch eine eingeschränkte Möglichkeit, etwas über den PageRank ihrer Site zu erfahren.

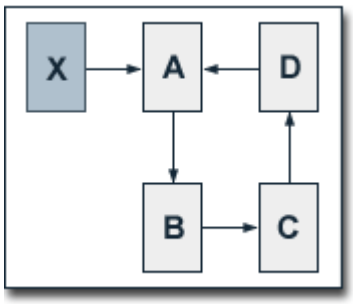
Bei der Google Directory handelt es sich um einen Dump des Open Directory Projects (dmoz.org), der neben den Seiteneinträgen ähnlich der Google Toolbar den skalierten PageRank für die in das ODP eingetragene Seite in Balkenform anzeigt. Allerdings erfolgt die PageRank-Anzeige in der Google-Directory auf einer Skala von 1 bis 7. Der exakte Wert wird nicht angezeigt, kann aber über die zweigeteilte Balkengrafik bzw. die Breite von deren Einzelgrafiken bestimmt werden, falls der Betrachter sich beim einfachen Augenschein unsicher ist.

Durch den Vergleich des Toolbar-PageRanks mit dem Directory-PageRank kann vom tatsächlichem PageRank von Seiten, die in das ODP eingetragen sind, ein etwas genauerer Eindruck gewonnen werden. Dieser Zusammenhang wurde zuerst von Chris Raimondi (www.searchnerd.com/pagerank) aufgezeigt.



Insbesondere für Seiten mit einem Toolbar-PageRank von 5 oder 6 ergibt sich hier die Möglichkeit der Einschätzung, ob sich die Seite eher am unteren oder am oberen Ende eines Bereichs der Toolbar-Skalierung befindet. Es sei an dieser Stelle angemerkt, dass für die Darstellung des Vergleichs der beiden PageRank-Anzeigen der Toolbar-PageRank von 0 nicht berücksichtigt wurde. Dass dies sinnvoll ist, kann anhand von Seiten mit einem Directory-PageRank von 3 nachvollzogen werden. Hier ist allerdings zu berücksichtigen, dass zur Überprüfung eine Seite der Google Directory mit einem Toolbar-PageRank von maximal 4 ausgewählt werden sollte, da sich sonst in der Regel keine von dort verlinkten Seiten mit einem Toolbar-PageRank von 3 finden lassen.

Der Effekt eingehender Links



Es wurde bereits gezeigt, dass ein jeder eingehender Link auf ein Webseite deren Pagerank stets erhöht. Bei oberflächlicher Betrachtung des ursprünglichen PageRank-Algorithmus

PR(A) = (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))

möchte man meinen, ein zusätzlicher eingehender Link erhöht den PageRank der verlinkten Seite um

d × PR(X) / C(X)

wobei PR(X) der PageRank der verlinkenden Seite X und C(X) deren Anzahl ausgehender Links ist. Bei genauerer Betrachtung zeigt sich allerdings, dass eine Webseite, die einen zusätzlichen eingehenden Link erhält, selbst auch auf andere Webseiten verlinken kann. Diese erhalten nunmehr ebenfalls einen höheren PageRank, den sie gegebenfalls wiederum über Links an unsere Seite mit dem zusätzlichen eingehenden Link zurückgeben.

Die einzelnen Effekte zusätzlicher eingehender Links sollen nun anhand eines Beispiels beschrieben werden.

Wir betrachten eine kleine Website aus den Seiten A, B, C und D, die im Kreis verlinkt sind und nicht selbst auf andere Webseiten verlinken. Ohne eingehende Links von externen Seiten ergibt sich ganz offensichtlich ein PageRank von 1 für jede der betrachteten Seiten. Wir wählen nunmehr eine Seite X, für die ein PageRank PR(X) von 10 angenommen wird. Seite X verlinkt auf Seite A und sonst auf keine andere Seite. Bei einem willkürlich gewählten Dämpfungsfaktor von 0.5 ergibt sich das folgende Gleichungssystem für den PageRank der einzelnen Seiten unserer Site:

PR(A) = 0.5 + 0.5 (PR(X) + PR(D)) = 5.5 + 0.5 PR(D)
PR(B) = 0.5 + 0.5 PR(A)
PR(C) = 0.5 + 0.5 PR(B)
PR(D) = 0.5 + 0.5 PR(C)

Da die Anzahl der ausgehenden Links jeder Seite gleich 1 ist, müssen diese hier nicht berücksichtigt werden. Die Lösung des Gleichungssystems ergibt folgende Werte für den PageRank der einzelnen Seiten:

PR(A) = 19/3 = 6.33
PR(B) = 11/3 = 3.67
PR(C) = 7/3 = 2.33
PR(D) = 5/3 = 1.67

Der unmittelbare Effekt des zusätzlichen Links auf Seite A in Höhe von

d × PR(X) / C(X) = 0,5 × 10 / 1 = 5

setzt sich also über die Verlinkung der einzelnen Seiten untereinander fort.

Der Einfluss des Dämpfungsfaktors

Der Grad der Weitergabe von PageRank ist vor allem auch abhängig von der Höhe des Dämpfungsfaktors d. Wird für diesen beispielsweise ein Wert von 0.75 angenommen, ergibt sich für das obige Beispiel das folgende Gleichungssystem:

PR(A) = 0.25 + 0.75 (PR(X) + PR(D)) = 7.75 + 0.75 PR(D)
PR(B) = 0.25 + 0.75 PR(A)
PR(C) = 0.25 + 0.75 PR(B)
PR(D) = 0.25 + 0.75 PR(C)

Die Lösung dieses Gleichungssystems ergibt folgende Werte für den PageRank der einzelnen Seiten:

PR(A) = 419/35 = 11.97
PR(B) = 323/35 = 9.23
PR(C) = 251/35 = 7.17
PR(D) = 197/35 = 5.63

Es zeigt sich zunächst ein wesentlich höherer unmittelbarer Effekt des zusätzlichen eingehenden Links auf den PageRank von Seite A in Höhe von

d × PR(X) / C(X) = 0.75 × 10 / 1 = 7.5

Dieser Effekt setzt sich nun aber noch verstärkt durch die interne Verlinkung der Seiten fort, so dass der PageRank von Seite A bei einem Dämpfungsfaktor von 0.75 beinahe doppelt so hoch ist wie bei einem Dämpfungsfaktor von 0.5. Ist der PageRank von Seite A bei einem Dämpfungsfaktor von 0.5 noch beinahe viermal größer als der Pagerank von Seite D, so ist er bei einem Dämpfungsfaktor von 0.75 nur noch etwas mehr als doppelt so groß. Je höher der Dämpfungsfaktor ist, um so stärker ist einerseits der Effekt auf den PageRank der den Link erhaltenden Seite und um so gleichmäßiger verteilt sich andererseits der PageRank auf die anderen Seiten der Site.

Der tatsächliche Effekt eingehender Links

Die Summe der PageRank-Werte aller Seiten bei einem Dämpfungsfaktor von 0.5 beträgt in unserem Beispiel

PR(A) + PR(B) + PR(C) + PR(D) = 14

Dadurch, dass eine Seite mit einem PageRank von 10 mit ihrem einzigen Link auf eine Seite der Beispiel-Site verlinkt, erhöht sich also deren aufaddierter PageRank um 10. (Vor Erhalt des Links hatte jede Seite einen PageRank von 1.) Bei einem PageRank von 0.75 beträgt die Summe der PageRank-Werte

PR(A) + PR(B) + PR(C) + PR(D) = 34

Der aufaddierte PageRank erhöht sich also um 30. Es zeigt sich, dass sich die Summe des PageRanks stets um

$(d / (1-d)) \times (PR(X) / C(X))$

erhöht, wenn X die verlinkende Seite, PR(X) deren PageRank und C(X) die Anzahl der ausgehenden Links von Seite X ist. Dieser Wert ist allerdings daran gebunden, dass die Verlinkung in ein geschlossenes System von Webseiten, also etwa eine Website ohne ausgehenden Link erfolgt. Sofern von der Website Links auf andere, externe Webseiten gesetzt sind, verringert sich der Faktor entsprechend.

Die Begründung für den oben angegebenen Wert liefert uns Raph Levien und sie bezieht sich auf das Random Surfer Modell. Die Länge eines Surf-Vorgangs des Zufalls-Surfers ist eine Exponentialverteilung mit einem Mittel von (d/(1-d)). Wenn also der Zufalls-Surfer einen Link auf ein geschlossenes System von Webseiten verfolgt, besucht er im Schnitt genau (d/(1-d)) Seiten innerhalb dieses geschlossenen Systems. Und genau so viel mehr PageRank der ursprünglich verlinkenden Seite - gewichtet nach der Anzahl der ausgehenden Links - muss damit an das geschlossene System übertragen werden.

Lawrence Page und Sergey Brin geben regelmäßig einen Dämpfungsfaktor von 0.85 für die tatsächliche PageRank-Berechnung an. Damit ergibt sich ein Faktor für die Erhöhung des aufaddierten PageRanks einer geschlossenen Site durch einen zusätzlichen eingehenden Link von Seite X in Höhe von

$(0.85 / 0.15) \times (PR(X) / C(X)) = 5.67 \times (PR(X) / C(X))$

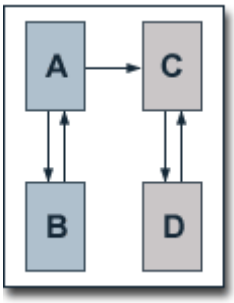
Eingehende Links haben also einen weitaus größeren Effekt auf den PageRank als man bei oberflächlicher Betrachtung annehmen mag.

Die PageRank-1 Regel

Viele Nutzer der Google Toolbar stellen fest, dass oftmals Seiten mit einem bestimmten Toolbar-PageRank eine darauf verlinkende Seite mit einem Toolbar-PageRank aufweisen, der um den Wert 1 höher ist als der der verlinkten Seite. Diese Beobachtung dient vielfach dazu, den hier präsentierten PageRank-Algorithmus in Frage zu stellen. Dagegen soll an dieser Stelle gezeigt werden, dass die Beobachtung vollkommen im Einklang mit dem hier präsentierten PageRank-Algorithmus steht.

Zuallererst stützt die PageRank-1 Regel das grundlegende Konzept des PageRank-Verfahrens. Webseiten sind genau dann bedeutsam, wenn andere bedeutsame Webseiten auf sie verweisen. Es ist nicht erforderlich, dass eine Website viele eingehenden Links erhält, um einen hohen PageRank zu bekommen. Ein einzelner Link von einer Website mit einem hohen PageRank reicht hierzu aus.

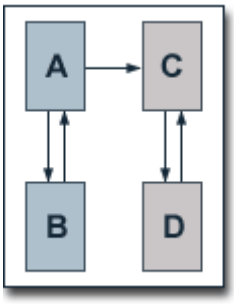
Dafür, dass die PageRank-1 Regel auch mit dem hier präsentierten PageRank-Algorithmus in Einklang steht, sind mehrere Faktoren verantwortlich. Zunächst ist Toolbar-PageRank eine logarithmisch skalierte Version des tatsächlichen PageRank. Wenn der PageRank einer verlinkenden Seite im Sinne der Toolbar um eins höher ist als derjenige der verlinkten Seite, so kann ihr tatsächlicher PageRank immer mindestens um einen Faktor höher sein, der der Basis des für die Skalierung eingesetzten Logarithmus entspricht. Ist also die Basis des Logarithmus gleich 6, und der Toolbar-PageRank der verlinkenden Seite gleich 5, so kann der tatsächliche PageRank der verlinkten Seite immer mindestens 6 Mal kleiner sein, damit diese in jedem Fall noch einen Toolbar-PageRank von 4 erreicht.



Der Basis des Logarithmus wirkt nun die Anzahl der ausgehenden Links auf der verlinkenden Seite entgegen, da deren PageRank praktisch unter allen verlinkten Seiten aufgeteilt wird. Es wurde allerdings oben auf dieser Seite bereits gezeigt, dass der über einen Link an eine Seite weitergegebene PageRank weitaus größer sein kann, als der im Algorithmus erscheinenden Term d(PR(Ti)/C(Ti)) vermuten lässt. Dies hängt damit zusammen, dass intern in der Regel weitere Seiten auf die von außen verlinkte Seite verlinken und somit weiteren PageRank auf diese Seite verteilen. Gehen wir nun etwa davon aus, dass die logarithmische Basis für die Skalierung 6 beträgt und weiterhin bei einem hohen Dämpfungsfaktor das doppelte des nach ausgehenden Links gewichteten PageRank der verlinkenden Seite auf die verlinkte Seite übertragen wird, so könnte die verlinkende Seite mindestens 12 ausgehende Links haben, damit die verlinkte Seite immer noch einen Toolbar-PageRank aufweist, der maximal um 1 kleiner ist als derjenige der verlinkenden Seite.

Eine Zahl von 12 ausgehenden Links erscheint an dieser Stelle zugegebenermaßen sehr gering. Allerdings ist es in aller Regel so, dass wenn eine Webseite von außen verlinkt wird, dies nicht nur von einer einzelnen Seite geschieht, und der betrachteten Seite somit noch weiterer PageRank übertragen wird. Falls sich Beispiele finden, bei der eine Seite durch einen einzigen externen Link ein PageRank übertragen wird, der der PageRank-1 Regel entspricht, und die verlinkende Seite eine hohe Zahl ausgehender Links hat, so ist dies vor allem ein Indiz dafür, dass der PageRank der verlinkenden Seite sich im oberen Bereich ihres Toolbar-PageRank-Wertes befindet. Schließlich könnte die verlinkende Seite aus unserem Beispiel eine "hohe" 5 und die verlinkte Seite eine "tiefe" 4 sein. In diesem Falle könnte die verlinkende Seite bis zu 72 ausgehende Links aufweisen. Diese Zahl würde sich weiter erhöhen, wenn wir von einer größeren logarithmischen Basis für die Skalierung des Toolbar-PageRanks ausgehen.

Der Effekt ausgehender Links



Da das PageRank-Verfahren die Link-Struktur des gesamten Webs abbildet, ist es unausweichlich, dass wenn eingehende Links einen Einfluss auf den PageRank haben, das gleiche auch für ausgehende Links gilt. Zur Darstellung der Effekte ausgehender Links soll wieder ein kleines Beispiel dienen.

Betrachtet wird ein Web aus zwei Websites, die jeweils zwei Seiten beinhalten. Die eine Site besteht aus den Seiten A und B, die andere aus den Seiten C und D. Die beiden Seiten einer jeden Site verlinken sich jeweils gegeneinander. Es wird unmittelbar deutlich, dass jede der Seiten einen ursprünglichen PageRank von 1 inne hat. Nun wird Seite A ein externer Link auf Seite C hinzugefügt. Für den PageRank der einzelnen Seiten ergeben sich bei einem angenommenen Dämpfungsfaktor d von 0.75 die folgenden Gleichungen:

PR(A) = 0.25 + 0.75 PR(B)
PR(B) = 0.25 + 0.375 PR(A)
PR(C) = 0.25 + 0.75 PR(D) + 0.375 PR(A)
PR(D) = 0.25 + 0.75 PR(C)

Die Lösung dieses Gleichungssystems ergibt die folgenden Werte:

PR(A) = 14/23
PR(B) = 11/23

und somit einen aufsummierten PageRank von 25/23 für die erste Site,

PR(C) = 35/23
PR(D) = 32/23

und damit einen aufsummierten PageRank von 67/23 für die zweite Site. Der aufsummierte PageRank beider Sites in Höhe von 92/23 = 4 bleibt also erhalten. Das Hinzufügen von Links hat also keinen Einfluss auf den aufsummierten PageRank des Webs. Ferner ist damit der Gewinn der verlinkten Site genauso groß wie der Verlust der anderen.

Der tatsächliche Effekt ausgehender Links

Wie bereits gezeigt, ist der Gewinn eines geschlossenen Systems auf das ein zusätzlicher Link gesetzt wird, gegeben durch

$(d / (1-d)) \times (PR(X) / C(X))$,

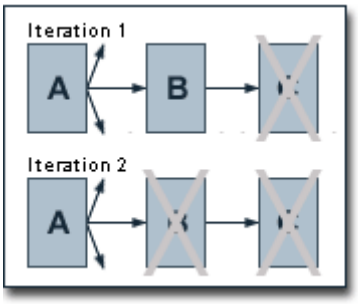
wobei X die verlinkende Seite, PR(X) deren PageRank und C(X) die Anzahl der ausgehenden Links von Seite X ist. Dieser Wert beschreibt damit auch den PageRank-Verlust, der einem vormals geschlossenen System daraus entsteht, dass einer Seite X innerhalb dieses Systems ein ausgehender Link hinzugefügt wird.

Bedingung für die angegebene Formel ist, dass die verlinkte Site nicht etwa direkt wieder auf die verlinkende Site zurückverlinkt, da die verlinkende Site wieder einen Teil des verlorenen PageRanks zurückgewinnen würde. Solange eine Rückverlinkung sich über eine gar nicht so große Anzahl von Webseiten erstreckt, ist dieser Effekt jedoch durch die Wirkungsweise des Dämpfungsfaktors zu vernachlässigen. Ferner Bedingung für die Gültigkeit der Formel ist, dass die verlinkende Site nicht bereits vorher ausgehende Links besitzt. Sollte dies jedoch der Fall sein, vermindert sich die Höhe des Verlustes der betrachteten Site, und gleichzeitig haben die bereits verlinkten Webseiten einen entsprechenden Verminderung des PageRank hinzunehmen.

Selbst wenn für eine tatsächlich existierende Website die PageRank-Werte der einzelnen Webseiten bekannt wären, könnte allerdings dennoch nicht ohne weiteres im Vorhinein ermittelt werden, wie sehr das Hinzufügen eines externen Links den PageRank der einzelnen Seiten vermindert, da die oben genannten Formel den Status nach der Verlinkung betrachtet.

Intuitive Begründung für den Effekt ausgehender Links

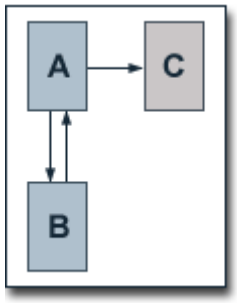
Intuitiv lässt sich der Verlust von PageRank für die eigenen Seiten damit erklären, dass der Zufalls-Surfer aus dem Random Surfer Modell durch das Hinzufügen eines externen Links mit einer geringeren Wahrscheinlichkeit einen Link auf eine der internen Seiten verfolgt. Damit sinkt in der Folge auch die Wahrscheinlichkeit, mit der sich der Surfer auf einer derjenigen Seiten aufhält, die wiederum auf diejenige Seite verlinken, der der externe Link



hinzugefügt wurde, womit auch deren PageRank sinken muss.

Es bleibt letztlich festzuhalten, dass ausgehende externe Links den aufsummierten PageRank aller Webseiten einer Website und gegebenenfalls auch den PageRank jeder einzelnen Seite einer Site vermindern. Da jedoch die Verlinkung zwischen Websites gerade die Grundlage des PageRank-Verfahrens darstellt und für sein Funktionieren unabdingbar ist, besteht durchaus die Möglichkeit, dass ausgehende Links an einer anderen Stelle innerhalb der Bewertung von Webseiten durch die Suchmaschine Google positiven Einfluss nehmen. Schließlich machen gerade auch relevante ausgehende Links die Qualität einer Website aus, und Webmaster, die Links auf andere Websites setzen, beziehen gewissermaßen deren Content in das eigene Web-Angebot mit ein.

Dangling Links



Ein nicht ganz unwichtiger Aspekt ausgehender Links ist das Fehlen ausgehender Links. Sobald einzelne Webseiten keine ausgehenden Links aufweisen, versickert der PageRank gewissermaßen an diesen Stellen. Lawrence Page und Sergey Brin bezeichnen Verweise auf derartige Seiten als "Dangling Links".

Die Auswirkungen von Dangling Links sollen anhand eines kleinen Beispiels veranschaulicht werden. Wir betrachten eine Website die aus aus den drei Seiten A, B und C besteht. Die Seiten A und B verlinken sich gegenseitig. Seite A verlinkt zudem auf Seite C, die ihrerseits jedoch keine ausgehenden Links aufweist. Für den PageRank der einzelnen Seiten ergeben sich bei einem angenommenen Dämpfungsfaktor d von 0.75 die folgenden Gleichungen:

PR(A) = 0.25 + 0.75 PR(B)
PR(B) = 0.25 + 0.375 PR(A)
PR(C) = 0.25 + 0.375 PR(A)

Die Lösung dieses Gleichungssystems ergibt die folgenden PageRank-Werte:

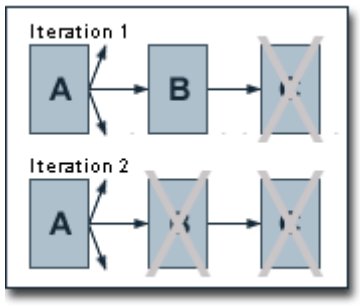
PR(A) = 14/23
PR(B) = 11/23
PR(C) = 11/23

Damit beträgt der aufaddierte PageRank aller Seiten 36/23, also nur etwas mehr als die Hälfte dessen, was zu erwarten gewesen wäre, wenn Seite C auf eine der beiden Seiten A oder B verlinkt hätte. Die Anzahl von Dangling Links ist nach Angaben von Page und Brin nicht unbeträchtlich - und sei es, weil zahlreiche verlinkte Seiten von Google nicht indexiert sind, etwa weil die Indexierung per robots.txt verhindert wird. Hier ist zusätzlich zu berücksichtigen, dass Google mittlerweile auch andere Dokumenten-Typen als HTML wie zum Beispiel PDF oder Word Dateien indexiert, die keine wirklichen ausgehenden Links haben. Dangling Links könnten also nicht unbeträchtliche Folgen für das PageRank-Verfahren haben.

Um die negativen Effekte von Dangling Links auszuschalten, werden diese Angaben von

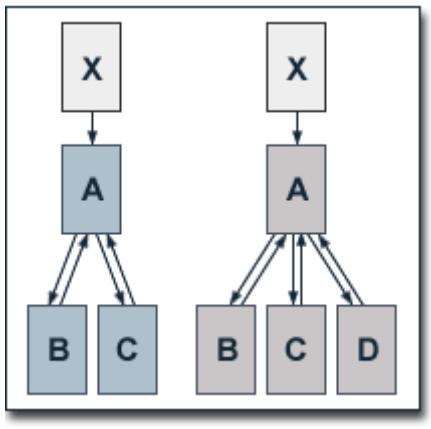
Page und Brin zufolge vor der PageRank-Berechnung aus der Datenbank unter Anpassung der jeweiligen Anzahl von ausgehenden Links entfernt bis alle PageRank-Werte berechnet sind. Bei der Entfernung von Dangling Links handelt es sich um einen iterativen Vorgang, da das Entfernen wiederum neue Dangling Links erzeugen kann, wie aus unserer einfachen Abbildung ersichtlich. Nachdem die eigentliche PageRank-Berechnung abgeschlossen ist, wird auch den Dangling Links PageRank - auf der Basis der PageRank-Werte der auf sie verweisenden Seiten und unter Rückgriff auf den PageRank-Algorithmus - zugewiesen. Dies erfordert ebenso viele Iterationen wie bei der Entfernung der Dangling Links. Um bei unserer Abbildung zu bleiben, könnte schließlich Seite C vor Seite B bearbeitet werden. Seite B weist dann im ersten Bearbeitungsdurchlauf bei der Bearbeitung von Seite C noch keinen PageRank auf, womit Seite C wiederum ein PageRank von 0 zugewiesen würde. Erst anschließend erhält Seite B ihren PageRank und im zweiten Bearbeitungsschritt würde Seite C einen tatsächlichen PageRank zugewiesen bekommen.

Für unser ursprüngliches Beispiel hat das Entfernen von Seite C aus der Datenbank zur Folge, dass die Seiten A und B nach Abschluss der Berechnungen jeweils einen PageRank von 1 erhalten. Seite C wird dann im Anschluss ein PageRank in Höhe von 0.25 + 0.375 PR(A) = 0.625 zugewiesen. Damit enspricht der aufaddierte PageRank zwar nicht der Anzahl der Seiten, doch zumindest diejenigen Seiten mit ausgehenden Links nehmen keinen Schaden durch Dangling Links.



Durch die Eliminierung von Dangling Links haben diese also keinen negativen Einfluss auf den PageRank der übrigen Seiten. Und wie bereits erwähnt, sind Verweise auf Dokumententypen, die keine ausgehenden Links aufweisen können, grundsätzlich Dangling Links. Damit wird auch unmittelbar deutlich, dass etwa Links auf PDF-Dokumente den PageRank einer darauf verlinkenden Seite bzw. Site nicht reduzieren können. PDF-Dokumente können also ein sehr gutes Instrument der Suchmaschinenoptimierung für Google sein.

Der Einfluss der Anzahl der Seiten auf den PageRank



Da der aufaddierte PageRank aller Seiten des Webs gleich der Anzahl der Seiten ist, folgt unmittelbar, dass eine zusätzliche Seite den aufaddierten PageRank des Webs um eins erhöht. Wesentlich interessanter als die Auswirkungen zusätzlicher Seiten auf den aufaddierten PageRank des gesamten Webs sind die Auswirkungen auf den PageRank der Seiten einer konkreten Site.

Um die konkreten Auswirkungen zusätzlicher Seiten zu veranschaulichen, betrachten wir zunächst eine hierarchisch strukturierte Beispielsite bestehend aus den drei Seiten A, B und C, der auf der unteren Ebene eine zusätzliche Seite D hinzugefügt wird. Die Site hat keine ausgehenden Links. Auf Seite A verlinkt eine externe Seite X mit einem PageRank von 10 durch ihren einzigen ausgehenden Link. Bei einem Dämpfungsfaktor d in Höhe von 0.75 ergeben sich vor dem Hinzufügen von Seite D die folgenden Gleichungen für den PageRank der einzelnen Seiten:

PR(A) = 0.25 + 0.75 (10 + PR(B) + PR(C))
PR(B) = PR(C) = 0.25 + 0.75 (PR(A) / 2)

Die Lösung des Gleichungssystems ergibt die folgenden PageRank-Werte:

PR(A) = 260/14
PR(B) = 101/14
PR(C) = 101/14

Nach dem Hinzufügen von Seite D lauten die Gleichungen für die PageRank-Berechnung folgendermaßen:

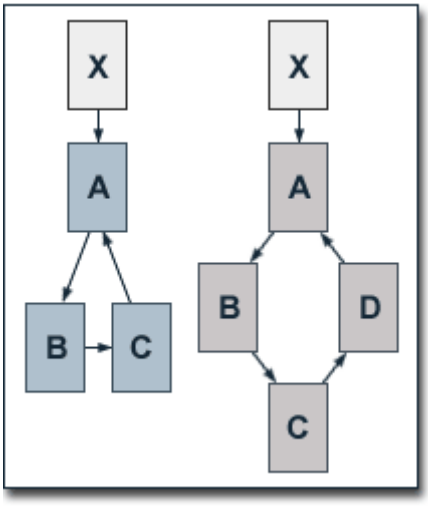
PR(A) = 0.25 + 0.75 (10 + PR(B) + PR(C) + PR(D))
PR(B) = PR(C) = PR(D) = 0.25 + 0.75 (PR(A) / 3)

Die Lösung dieses Gleichungssystems ergibt die folgenden PageRank-Werte:

PR(A) = 266/14
PR(B) = 70/14
PR(C) = 70/14
PR(D) = 70/14

Da unsere Beispielsite keine ausgehenden Links aufweist, steigt der aufaddierte PageRank aller Seiten nach dem Hinzfügen von Seite D erwartungsgemäß um genau 1 von 33 auf 34. Ferner steigt der PageRank von Seite A marginal an. Der PageRank der Seiten B und C jedoch sinkt um ein beträchtliches Maß.

Die Reduzierung des PageRanks durch zustätzliche Seiten



Bei dem Hinzufügen zusätzlicher Seiten zu einer Website mit strikt hierarchischer Struktur sind die Auswirkungen auf den PageRank der bereits bestehenden Seiten uneinheitlich. Welche Auswirkungen das hinzufügen von Seiten zu Sites mit anderer Struktur hat, soll wiederum anhand eines Beispiels erläutert werden.

Betrachtet wird jetzt eine Website, deren Seiten A, B und C untereinander im Kreis verlinken und der eine zusätzliche Seite D hinzugefügt wird, die sich in die bestehende Linkstruktur einpasst. Die Site hat ebenfalls keine ausgehenden Links. Auf Seite A verlinkt wiederum eine externe Seite X mit einem PaeRank von 10 durch ihren einzigen ausgehenden Link. Bei einem Dämpfungsfaktor d in Höhe von 0.75 ergeben sich vor dem Hinzufügen von Seite D die folgenden Gleichungen für den PageRank der einzelnen Seiten:

PR(A) = 0.25 + 0.75 (10 + PR(C))
PR(B) = 0.25 + 0.75 × PR(A)
PR(C) = 0.25 + 0.75 × PR(B)

Die Lösung des Gleichungssystems ergibt die folgenden PageRank-Werte:

PR(A) = 517/37 = 13.97
PR(B) = 397/37 = 10.73
PR(C) = 307/37 = 8.30

Nach dem Hinzufügen von Seite D lauten die Gleichungen für die PageRank-Berechnung folgendermaßen:

PR(A) = 0.25 + 0.75 (10 + PR(D))
PR(B) = 0.25 + 0.75 × PR(A)
PR(C) = 0.25 + 0.75 × PR(B)
PR(D) = 0.25 + 0.75 × PR(C)

Die Lösung dieses Gleichungssystems ergibt die folgenden PageRank-Werte:

PR(A) = 419/35 = 11.97
PR(B) = 323/35 = 9.23
PR(C) = 251/35 = 7.17
PR(D) = 197/35 = 5.63

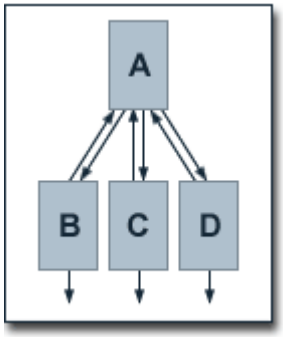
Wiederum steigt der aufaddierte PageRank aller Seiten nach dem Hinzfügen von Seite D um genau 1 von 33 auf 34. Jetzt allerdings verlieren alle bereits vorher existierenden Seiten an PageRank. Dieser Effekt zeigt sich stets um so eher, je gleichmäßiger der PageRank auf die einzelnen Seiten einer Site verteilt werden.

Damit wird auch deutlich, dass der PageRank-Algorithmus grundsätzlich kleinere Websites bevorzugt. Dies ist allerdings dadurch zu relativieren, dass Sites mit mehr Content dies ausgleichen können, indem andere Seitenbetreiber um so eher auf sie verlinken.

Es ist allerdings auch möglich, durch zusätzliche Seiten den PageRank bereits existierender Seiten zu steigern. Hierbei ist jedoch darauf zu achten, dass auf die zusätzlichen Seiten möglichst wenig PageRank verteilt wird.

Die Distribution von PageRank im Rahmen der Suchmaschinenoptmierung

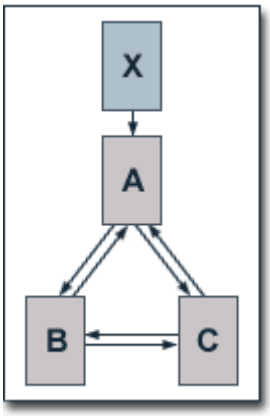
Bislang wurde erörtert, wie durch die Anzahl von ein- und ausgehenden Links sowie durch die Anzahl der Webseiten einer Site der PageRank beinflusst werden kann. An dieser Stelle hingegen soll hauptsächlich besprochen werden, wie mittels der internen Link-Struktur einer Site zum Zwecke der Suchmaschinenoptimierung Einfluss auf den PageRank genommen werden kann.



In den meisten Fällen sind Websites zumindest bedingt hierarchisch struktieriert. Dabei ist in der Regel die Startseite für den wichtigsten Suchbegriff bzw. die wichtigste Suchphrase optimiert. In unserem Beispiel erhält die optimierte Startseite A einen eingehenden Link von einer Seite X mit einem PageRank von 10 und einem einzigen ausgehenden Link. Die Seiten B und C erhalten einen Link von Seite A und verlinken auch

wieder auf diese zurück. Hieraus ergeben sich bei einem angenommenen Dämpfungsfaktor d in Höhe von 0.5 die folgenden Gleichungen für die PageRank-Berechnung:

PR(A) = 0.5 + 0.5 (10 + PR(B) + PR (C))
PR(B) = 0.5 + 0.5 (PR(A) / 2)
PR(C) = 0.5 + 0.5 (PR(A) / 2)

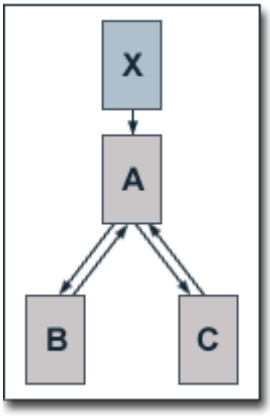


Die Lösung dieses Gleichungssystems ergibt die folgenden PageRank-Werte:

PR(A) = 8
PR(B) = 2.5
PR(C) = 2.5

Nun ist es in der Regel nicht ausreichend, im Rahmen der Suchmaschinenoptimierung lediglich die Startseite für einen Suchbegriff zu optimieren. Es ist vielmehr ratsam, alle Seiten auf die Optimierung für jeweils unterschiedliche Suchbegriffe auszurichten.

Sobald die Startseite für den optimierten Suchbegriff hinreichend gute Suchmaschinenergebnisse erzielt, die anderen Seiten hingegen noch nicht, empfiehlt gegebenenfalls die Linkstruktur entsprechend der folgenden Vorgehensweise bei unserer Beispielsite zu modifizieren. Die hierachisch nachgeordneten Seiten B und C verlinken sich nunmehr gegenseitig, und bei ansonsten gleichen Bedingungen ergibt sich das folgende Gleichungssystem für die PageRank-Berechnung:



PR(A) = 0.5 + 0.5 (10 + PR(B) / 2 + PR(C) / 2)
PR(B) = 0.5 + 0.5 (PR(A) / 2 + PR(C) / 2)
PR(C) = 0.5 + 0.5 (PR(A) / 2 + PR(B) / 2)

Hieraus ergeben sich nun die folgenden PageRank-Werte für die einzelnen Seiten:

PR(A) = 7
PR(B) = 3
PR(C) = 3

Es zeigt sich, dass die PageRank-Werte für die Seiten B und C steigen, so dass diese wahrscheinlich für die optimierten Suchbegriffe höher in den Suchmaschinenergebnissen erscheinen werden. Andererseits besteht die Möglichkeit, dass die Startseite in den Suchmaschinenergebnissen absinkt.

Grundsätzlich zeigt sich, dass sich im Rahmen der Suchmaschinenoptimierung der PageRank innerhalb einer Seite um so gleichmäßiger verteilt, je stärker die hierarchisch nachrangigen Seiten untereinander verlinkt sind.

Gezielte Distribution von PageRank durch Konzentration der ausgehenden Links

Dass ausgehende Links sich grundsätzlich eher negativ auf den PageRank der Seiten einer Website auswirken, wurde bereits gezeigt. An dieser Stelle soll erörtert werden, wie dieser Effekt durch die gezielte Platzierung der ausgehenden Links minimiert werden kann.

Betrachtet wird nun eine Beispielsite aus den Seiten A, B, C und D, wobei Seite A auf die anderen Seiten verlinkt, und diese neben einem Link auf Seite A jeweils auch noch einen ausgehenden Link haben. Bei einem angenommen Dämpfungsfaktor d in Höhe von 0.5 ergeben sich die folgenden Gleichungen für die PageRank-Berechnung:

PR(A) = 0.5 + 0.5 (PR(B) / 2 + PR(C) / 2 + PR(D) / 2)
PR(B) = PR(C) = PR(D) = 0.5 + 0.5 (PR(A) / 3)

Die Lösung dieses Gleichungssystems ergibt die folgenden PageRank-Werte:

PR(A) = 1
PR(B) = 2/3
PR(C) = 2/3
PR(D) = 2/3

Nunmehr wird die Beispiel-Website so modifiziert, dass bei ansonsten gleichen Voraussetzungen nurmehr Seite D alle ausgehenden Links auf sich vereint und die Seiten B und C keinerlei ausgehenden mehr besitzen. Bei einem Dämpfungsfaktor d in Höhe von 0.5 ergeben sich die folgenden Gleichungen für die PageRank-Berechnung:

PR(A) = 0.5 + 0.5 (PR(B) + PR(C) + PR(D) / 4)
PR(B) = PR(C) = PR(D) = 0.5 + 0.5 (PR(A) / 3)

Die Lösung dieses Gleichungssystems ergibt die folgenden PageRank-Werte:

PR(A) = 17/13
PR(B) = 28/39
PR(C) = 28/39
PR(D) = 28/39

Es zeigt sich unmittelbar, dass für unsere Beispiel-Website die PageRank-Werte aller vier Seiten erhöhen. Vor dem Hintergrund der Suchmaschinenoptimierung kann es also durchaus ratsam sein, die ausgehenden Links einer Website auf einer einzelnen Seite zu konzentrieren, wobei allerdings durchaus nicht vergessen werden darf, dass dies der Benutzerfreundlichkeit abträglich sein kann.

Linktausch zum Zwecke der Suchmaschinenoptimierung

Viele Webmaster streben zum Zwecke der Suchmaschinenoptimierung den Linkaustausch mit möglichst vielen anderen Websites an, um auf diese Weise ihre Link-Popularität zu erhöhen.

Da das Hinzufügen von Links allerdings keinerlei Effekte auf den aufaddierten PageRank innerhalb geschlossener Systeme von Webseiten hat, stellt sich die Frage, in wie fern ein Linkaustausch zwischen Websites überhaupt Auswirkungen auf den PageRank hat.

Wir betrachten zwei hierarchisch strukturierte Websites aus den Seiten A, B und C bzw. D, E und F. Seite A verlinkt auf die Seiten B und C und diese wiederum verlinken zurück auf Seite A. Da die zweite Site exakt gleich strukturiert ist, ergeben sich für sie die gleichen PageRank-Werte, die deshalb an dieser Stelle nicht berücksichtigt werden müssen. Bei einem Dämpfungsfaktor d in Höhe von 0.5 ergeben sich die folgenden Gleichungen für die PageRank-Berechnung:

PR(A) = 0.5 + 0.5 (PR(B) + PR(C))
PR(B) = PR(C) = 0.5 + 0.5 (PR(A) / 2)

Die Lösung des Gleichungssystems ergibt die folgenden PageRank-Werte für die einzelnen Seiten:

PR(A) = 4/3
PR(B) = 5/6
PR(C) = 5/6

und analog

PR(D) = 4/3
PR(E) = 5/6
PR(F) = 5/6

Nunmehr findet für unsere Beispiel-Websites ein Linktausch statt. Seite A verlinkt auf Seite D und umgekehrt. Bei ansonsten gleichen Voraussetzungen mit einem Dämpfungsfaktor d in Höhe von 0.5 ergibt sich jetzt das folgende Gleichungssystem für die Berechnung der PageRank-Werte:

PR(A) = 0.5 + 0.5 (PR(B) + PR(C) + PR(D) / 3)
PR(B) = PR(C) = 0.5 + 0.5 (PR(A) / 3)
PR(D) = 0.5 + 0.5 (PR(E) + PR(F) + PR(A) / 3)
PR(E) = PR(F) = 0.5 + 0.5 (PR(D) / 3)

Die Lösung des Gleichungssystems ergibt die folgenden PageRank-Werte:

PR(A) = 3/2
PR(B) = 3/4
PR(C) = 3/4
PR(D) = 3/2
PR(E) = 3/4
PR(F) = 3/4

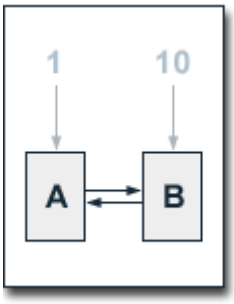
Es zeigt sich also, dass durch den Linktausch die beiden Seiten A und D profitieren und die hierarchisch nachgeordneten Seiten PageRank einbüßen. Für die Suchmaschinenoptimierung bedeutet dies in erster Linie, dass hier ein genau entgegengesetzter Effekt wie bei der stärkeren internen Verlinkung von hierarchisch nachgeordneten Seiten stattfindet. Der Linktausch bietet sich also insbesondere an, wenn nur mit einer Seite auf einen bestimmten

Suchbegriff abgezielt werden soll.

Voraussetzung für die genannten positiven Wirkungen durch einen Linktausch ist in jedem Falle, dass die jeweils verlinkenden Seiten einen ähnlich großen PageRank an die jeweils andere Site weitergeben. Sollte etwa eine Seite einen wesentlich größeren PageRank oder aber wesentlich weniger ausgehende Links haben, so besteht die Möglichkeit, dass alle Seiten ihrer Site an PageRank einbüßen. Ein nicht zu unterschätzender Einflussfaktor ist hier auch die Größe der beiden Sites. Je mehr Seiten eine Website besitzt, um so mehr des PageRanks aus eingehenden Links wird auf andere Seiten der Site verteilt, unabhängig davon, weie viele ausgehende Links die am Linktausch beteiligte Seite hat. Damit profitiert die am Linktausch beteiligte Seite selbst relativ wenig vom Linktausch, und kann an die andere am Linktausch beteiligte Seite nur relativ wenig PageRank zurückgeben. Letzlich sollten die genannten Faktoren stets gegeneinander abgewogen werden, bevor ein Linktausch eingegangen wird.

Abschließend bleibt anzumerken, dass ein Linktausch auch positive Effekte für alle Seiten einer Site haben kann, ohne dass die andere am Linktausch beteiligte Site geschädigt wird. Dies kann der Fall sein, wenn die am Linktausch beteiligte Seite bereits eine bestimmte Anzahl ausgehender Links auf Seiten aufweist, die nicht in direkter oder indirekter Form auf die betrachtete Site zurückverlinken. Mit dem Linktausch geht der betrachteten Site dann weniger PageRank durch die bereits vorher existierenden ausgehenden Links verloren.

Der Yahoo-Bonus und seine Auswirkungen auf die Suchmaschinen-Optimierung



Vielfach wird angenommen, das einige Websites von Google eine spezielle PageRank-Bewertung erhalten, die einen manuellen Eingriff erfordert und sich nicht direkt aus dem ursprünglichen PageRank-Algorithmus ergibt. Zu diesen Websites zählen z.B. die Verzeichnisse Yahoo und Open Directory Project (dmoz.org). Im Rahmen der Suchmaschinen-Optimierung hätte diese Annahme zur Folge, dass ein Eintrag in die genannten Verzeichnisse für den PageRank von besonderer Bedeutung ist.

Ein häufig genannter Ansatz für die besondere Bewertung spezieller Websites ist, dass diesen für die iterative Berechnung des PageRank ein höherer Startwert zugewiesen wird. Diese mögliche Vorgehensweise soll anhand eines sehr einfachen Beispiels überprüft werden. Wir betrachten ein 2-Seiten-Web, bei dem jede der beiden Seiten jeweils ausschließlich auf die andere verlinkt. Der einen Seite wird ein Startwert von 10 zugewiesen, der anderen ein Startwert von 1. Der Dämpfungsfaktor d wird in diesem Beispiel auf 0.1 gesetzt, da bei einem geringen Dämpfungsfaktor der PageRank im Zuge der Iterationen schneller konvergiert. Damit ergeben sich folgende Formeln für die PageRank-Berechnung:

PR(A) = 0.9 + 0.1 PR(B)
PR(B) = 0.9 + 0.1 PR(A)

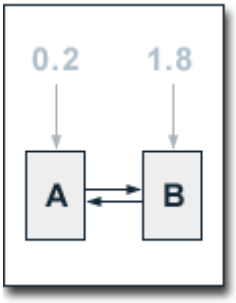
Die PageRank-Werte ergeben sich im Laufe der Iterationen wie folgt:

Iteration	PR (A)	PR (B)
0	1	10
1	1.9	1.09
2	1.009	1.0009
3	1.00009	1.000009

Es wird unmittelbar ersichtlich, dass die PageRank-Werte trotz der Vergabe besonderer Startwerte für die Berechnung jeweils gegen 1 konvergieren, so wie es auch ohne die Vergabe spezieller Startwerte zu erwarten gewesen wäre. Bei ausreichend vielen Iterationen hat somit der Startwert keinerlei Auswirkung auf den PageRank. Auswirkungen würden sich lediglich ergeben, wenn nur wenige Iterationen durchgeführt werden. Hier ist allerdings zu bedenken, dass sich etwa in unserem Beispiel die PageRank-Relation zwischen den beiden Seiten direkt nach der ersten Iteration umkehrt. Hierzu sei angemerkt, dass für die rekursive Berechnung jeweils die PageRank-Werte der aktuellen Iteration und nicht etwa der vorherigen genutzt wurden. Wären die Werte der vorherigen Iteration genutzt worden, würde die PageRank-Relation alterieren.

Modifikation des PageRank-Algorithmus

Dass eine Zuweisung spezieller Startwerte ohne Auswirkungen bleibt, bedeutet jedoch nicht, dass Websites nicht durch einen Eingriff in den PageRank-Algorithmus bevorzugt werden können. So beschreibt Lawrence Page bereits in seiner Patentschrift zum PageRank-Verfahren (United States Patent 6,285,999) die Möglichkeit für die besondere Bewertung spezieller Webseiten. Der Ausgangspunkt für seine Überlegungen ist, dass der Zufalls-Surfer aus dem Random Surfer Modell zwar mit einer starr festgelegten Wahrscheinlichkeit aufhört, Links zu verfolgen, dann aber im Gegensatz zum ursprünglichen PageRank-Algorithmus nicht mehr mit der gleichen Wahrscheinlichkeit eine Webseite für einen erneuten Start seines Surf-Vorgangs auswählt. Es entspricht schließlich dem normalen Verhalten eines Internet-Nutzers, dass er als Ausgangspunkt mit einer höheren Wahrscheinlichkeit etwa eines der genannten Verzeichnisse Yahoo oder ODP wählt.



Damit die besondere Bewertung einzelner Webseiten in dieser Form in den ursprünglichen PageRank Algorithmus einfließen kann, muss er um einen weiteren Erwartungswert erweitert werden. Die entsprechende Formel hat dann folgendes Aussehen:

$$PR(A) = E(A)\,(1-d) + d\,(PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

Hierbei ist (1-d) jetzt die Wahrscheinlichkeit, mit der der Zufalls-Surfer das Weiterverfolgen von Links abbricht und
E(A) die nach der Anzahl der Webseiten gewichtete Wahrscheinlichkeit, mit der der Zufalls-Surfer die Seite A danach aufruft. Bei E handelt es sich dabei wiederum um einen Erwartungswert, dessen Durchschnitt über alle Seiten gleich 1 ist, damit der Durchschnitt der PageRank-Werte weiterhin gegen 1 konvergiert und nicht etwa durch die besondere Bewertung spezieller Seiten schwankt und somit der PageRank einen unregelmäßigen Einfluss auf die Gesamtbewertung von Seiten einnimmt.

In unserem Beispiel liege nach dem Abbruch des Surfvorgangs durch den Zufalls-Surfer die Wahrscheinlichkeit für den Aufruf von Seite A bei 10% und die Wahrscheinlichkeit für den Aufruf von Seite B bei 90%. Damit ist bei einem 2-Seiten-Web E(A)=0.2 und E(B)=1.8. Für die Ermittlung der PageRank Werte der beiden Seiten ergeben sich bei einem Dämpfungsfaktor d=0.5 hierdurch die folgenden Gleichungen:

$PR(A) = 0.2 \times 0.5 + 0.5 \times PR(B)$
$PR(B) = 1.8 \times 0.5 + 0.5 \times PR(A)$

Die Lösung dieses Gleichungssystems ergibt die folgenden PageRank-Werte:

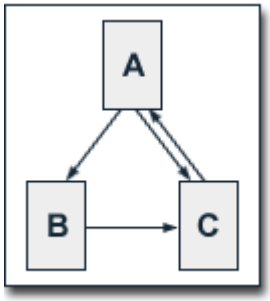
$PR(A) = 11/15$
$PR(B) = 19/15$

Die Summe der beiden PageRank-Werte liegt weiterhin bei 2. Die höhere Wahrscheinlichkeit für das Aufrufen von Seite B nach dem Abbruch spiegelt sich in ihrem höheren PageRank-Wert wider. Die gleichmäßige Verlinkung der beiden Seiten untereinander vermindert jedoch ganz deutlich die Auswirkung der höheren Aufrufwahrscheinlichkeit auf den PageRank.

Es ist also möglich, eine besondere Gewichtung einzelner Seiten in den PageRank-Algorithmus einfließen zu lassen, ohne dass dessen Charakter grundsätzlich verändert werden müsste. Fraglich bleibt jedoch, nach welchen Kriterien die Gewichtung erfolgen kann. In der Patentschrift zum PageRank-Verfahren nennt Lawrence Page hierzu explizit die Nutzung tatsächlichen Benutzerverhaltens. Daten zum tatsächlichen Nutzerverhalten werden von Google über die Google Toolbar gesammelt. Das besondere hierbei ist, dass nicht einmal allzu große Datenmengen verarbeitet werden müssten, wie dies der Fall wäre, wenn eine Bewertung ausschließlich auf Nutzerverhalten basieren würde. Eine begrenzte Stichprobe wäre durchaus ausreichend, um zumindest die 1.000 oder 10.000 wichtigsten Anlaufstellen im Web zu ermitteln. Der PageRank-Algorithmus wäre dann in der Lage, über die Link-Struktur des Webs die Lücken zu füllen.

Die Ausführungen zum Einfließen tatsächlichen Benutzerverhaltens in das PageRank-Verfahren sind natürlich pure Spekulation. Ob überhaupt eine besondere Gewichtung spezieller Seiten stattfindet, wird letztlich ein Geheimnis der Google-Verantwortlichen bleiben.

Dennoch Zuweisung bestimmter Startwerte?



Obwohl die Zuweisung bestimmter Startwerte für die PageRank-Berechnung bei hinreichend vielen Iterationen wirkungslos für das Ergebnis der Berechnung bleibt, kann eine entsprechende Vorgehensweise durchaus sinnvoll sein.

Wir betrachten hierzu unser 3-Seiten-Beispiel aus den Seiten A, B und C, wobei Seite A sowohl auf Seite B als auch auf Seite C verlinkt. Seite B verlinkt lediglich auf Seite C und Seite C wiederum verlinkt auf Seite A. Den Dämfungsfaktor d setzen wir in diesem Falle für die Berechnungen auf 0.75. Hierdurch ergeben sich die folgenden Gleichungen für die iterative

Berechnung des PageRanks der einzelnen Seiten:

PR(A) = 0.25 + 0.75 PR(C)
PR(B) = 0.25 + 0.75 (PR(A) / 2)
PR(C) = 0.25 + 0.75 (PR(A) / 2 + PR(B))

Grundsätzlich muss den einzelnen Seiten kein Startwert vor Beginn der Iterationen zugewiesen werden. Sie haben in diesem Falle einen Wert von 0 und es ergibt sich das folgende Bild:

Iteration	PR(A)	PR(B)	PR(C)
0	0	0	0
1	0.25	0.34375	0.60156
2	0.70117	0.51294	0.89764
3	0.92323	0.59621	1.04337
4	1.03253	0.63720	1.11510
5	1.08632	0.65737	1.15040
6	1.11280	0.66730	1.16777
7	1.12583	0.67219	1.17633
8	1.13224	0.67459	1.18054
9	1.13540	0.67578	1.18261
10	1.13696	0.67636	1.18363
11	1.13772	0.67665	1.18413
12	1.13810	0.67679	1.18438
13	1.13828	0.67686	1.18450
14	1.13837	0.67689	1.18456
15	1.13842	0.67691	1.18459
16	1.13844	0.67692	1.18460
17	1.13845	0.67692	1.18461
18	1.13846	0.67692	1.18461
19	1.13846	0.67692	1.18461
20	1.13846	0.67692	1.18461
21	1.13846	0.67692	1.18461
22	1.13846	0.67692	1.18462

Bei einer Zuweisung eines Startwertes von 1 ergibt sich das folgende Bild für die Durchführung der Iterationen:

Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	0.625	1.09375
2	1.07031	0.65137	1.13989
3	1.10492	0.66434	1.16260
4	1.12195	0.67073	1.17378
5	1.13034	0.67388	1.17928
6	1.13446	0.67542	1.18199
7	1.13649	0.67618	1.18332
8	1.13749	0.67656	1.18398

9	1.13798	0.67674	1.18430
10	1.13823	0.67684	1.18446
11	1.13835	0.67688	1.18454
12	1.13840	0.67690	1.18458
13	1.13843	0.67691	1.18460
14	1.13845	0.67692	1.18461
15	1.13845	0.67692	1.18461
16	1.13846	0.67692	1.18461
17	1.13846	0.67692	1.18461
18	1.13846	0.67692	1.18461
19	1.13846	0.67692	1.18462

Wird nunmehr den Seiten ein initialer PageRank zugewiesen, der der tatsächlichen PageRank-Verteilung etwas mehr entspricht (1.1 für Seite A, 0.7 für Seite B und 1.2 für Seite C), ergibt sich das folgende Bild:

Iteration	PR(A)	PR(B)	PR(C)
0	1.1	0.7	1.2
1	1.15	0.68125	1.19219
2	1.14414	0.67905	1.18834
3	1.14126	0.67797	1.18645
4	1.13984	0.67744	1.18552
5	1.13914	0.67718	1.18506
6	1.13879	0.67705	1.18483
7	1.13863	0.67698	1.18472
8	1.13854	0.67695	1.18467
9	1.13850	0.67694	1.18464
10	1.13848	0.67693	1.18463
11	1.13847	0.67693	1.18462
12	1.13847	0.67692	1.18462
13	1.13846	0.67692	1.18462

Es zeigt sich, dass je näher die zugewiesenen Startwerte der tatsächlichen Verteilung kommen, die PageRank-Werte offenbar um so schneller konvergieren. Damit wären weniger Iterationen für die PageRank-Berechnung erforderlich, was insbesondere angesichts eines stets wachsenden Webs die Lieferung von auf einer aktuelleren Datanbasis gestützten Suchmaschinenergebnissen ermöglichen kann. Ausgangspunkt für eine hinreichend exakte Annahme könnten dabei für Seiten, die bereits den jeweils vorhergegangenen Berechnungszyklus durchlaufen haben, die PageRank-Werte aus diesem vorhergegangenen Berechnungszyklus sein. Neu in den Index aufgenommenen Seiten könnte dann ein initialer PageRank von 1 zugewiesen werden, der sich dann bereits nach der ersten Iteration sehr schnell dem tatsächlichen Zustand angleicht.

Weitere Einflussfaktoren im Rahmen des PageRank-Verfahrens

Es wurde bereits vielerorts diskutiert, ob für die PageRank-Berechnung seit der Veröffentlichung der wissenschaftlichen Arbeiten durch Lawrence Page und Sergey Brin

weitere Kriterien als nur die einfache Link-Struktur des Webs für die Berechnung des PageRanks hinzugezogen wurden. Lawrence Page selbst skizziert in der Patentschrift zum PageRank-Verfahren die folgenden potentiellen Einflussfaktoren:

- Die Stärke der Hervorhebung eines Links
- Die Position eines Links innerhalb des Dokuments
- Die Distanz zwischen Webseiten
- Die Bedeutung einer verweisenden Seite
- Die Aktualität einer verweisenden Seite

Die Implementierung dieser weiteren Einflussfaktoren würde zunächst auf bessere Annäherung des Random Surfer Modells an tatsächliches Nutzerverhalten abzielen. Mit der Einbeziehung von Hervorhebung und Position eines Links geht man davon aus, das ein Benutzer nicht völlig wahllos klickt, sondern unabhängig vom Ankertext eher die deutlich erkennbaren und unmittelbar sichtbaren Links verfolgt. Mit der Berücksichtigung der anderen Faktoren könnte Google darüber hinaus eine weit größere Flexibilität in der Bestimmung der Bedeutung eines eingehenden Links für eine Webseite erreichen, als durch die bereits erwähnten Methoden.

Ob einzelne dieser Faktoren tatsächlich in das PageRank-Verfahren implementiert sind, ist empirisch kaum zu belegen, und soll deshalb an dieser Stelle auch nicht ausführlich diskutiert werden. Es soll vielmehr erörtert werden, auf welche Art und Weise weitere Einflussfaktoren in den PageRank-Algorithmus implementiert werden könnten und welche Möglichkeiten zur Einflussnahme auf den PageRank seitens Google sich hierdurch ergeben.

Modifizierung des PageRank-Algorithmus

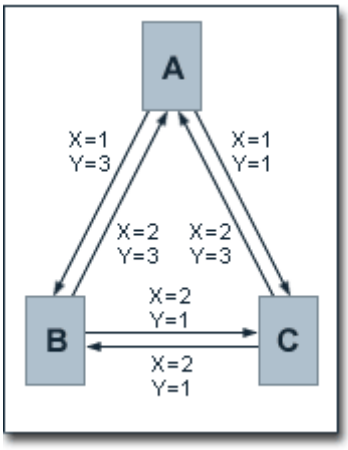
Um weitere Faktoren in das PageRank-Verfahren einfliessen zu lassen, ist der ursprüngliche PageRank-Algorithmus wiederum zu modifizieren. Da wir davon ausgehen müssen, dass für die PageRank-Berechnung weiterhin zahlreiche Iterationen durchgeführt werden, ist hierbei allerdings zu berücksichtigen, dass im Sinne einer möglichst schnellen PageRank-Berechnung für die Einbeziehung weiterer Faktoren zusätzliche Datenbank-Zugriffe im Laufe der Iterationen weitestgehend vermieden werden sollten. Aus diesem Grunde bietet sich der folgende, modifizierte PageRank-Algorithmus an:

$$PR(A) = (1-d) + d\ (PR(T1)\times L(T1,A) + ... + PR(Tn)\times L(Tn,A))$$

Hier bei stellt L(Ti,A) eine Bewertung des Links, der von der Seite Ti auf die Seite A zeigt, dar. L(Ti,A) tritt damit an die Stelle der Gewichtung des PageRanks von Seite Ti nach der Anzahl deren ausgehender Links durch den Faktor 1/C(Ti). Der Wert L(Ti,A) würde sich aus mehreren einzelnen Faktoren zusammensetzen, die jedoch nur einmal ermittelt werden müssten und dann vor der eigentlichen PageRank-Berechnung in einen einzigen Wert einfließen. Hierdurch vergrößert sich die Anzahl der Datenbankzugriffe nicht, wobei allerdings angemerkt werden muss, dass durch die hier vorgeschlagene Modifikation des PageRank-Algorithmus im Laufe jeder Iteration bei der Bestimmung jedes einzelnen PageRanks ein Zugriff auf eine wesentlich größere Datenbank zu erfolgen hat, als im Falle des ursprünglichen PageRank-Algorithmus, da nun nicht mehr nur die Bewertung von Seiten (anhand der Anzahl ihrer ausgehenden Links) sondern die Bewertung jedes einzelnen Links in die Berechnung einfließt.

Unterschiedliche Bewertung von Links innerhalb einzelner Seiten

Zwei wesentliche von Lawrence Page in der Patentschrift zum PageRank-Verfahren genannte Bewertungskriterien für Links sind deren Grad der Hervorhebung und Position innerhalb eines Dokuments. Es handelt es sich hierbei also um Kriterien, die im Rahmen des Random Surfer Modells einzig die Wahrscheinlichkeit widerspiegeln, mit der der Zufalls-Surfer einen bestimmten Link auf einer Website in Relation zu einem anderen Link auf dieser Website verfolgt. Im ursprünglichen PageRank-Algorithmus entspricht diese Wahrscheinlichkeit dem Term (1/C(Ti)), wobei die Wahrscheinlichkeiten für das Verfolgen von Links von einer Seite dabei jeweils gleich waren.



Eine Zuweisung unterschiedlicher Wahrscheinlichkeiten für das Verfolgen von Links könnte beispielhaft etwa folgendermaßen erfolgen:

Wir betrachten ein Web aus den drei Seiten A, B und C. Jede der Seiten verlinkt jeweils auf jede andere. Links werden nach zwei Bewertungskriterien X und Y gewichtet. X stellt die Hervorhebung eines Links dar. X ist gleich 1, sofern der Links nicht hervorgehoben und gleich 2, sofern der Link etwa fett oder kursiv hervorgehoben ist. Y stellt die Position eines Links im Dokument dar. Y ist gleich 1, sofern der Link in der unteren Hälfte des Dokuments und gleich 3, sofern der Link in der oberen Hälfte des Dokuments erscheint. Nehmen wir einen multiplikativen Zusammenhang zwischen X und Y an, werden die Links aus unserem Beispielweb wie folgt bewertet:

$X(A,B) \times Y(A,B) = 1 \times 3 = 3$
$X(A,C) \times Y(A,C) = 1 \times 1 = 1$
$X(B,A) \times Y(B,A) = 2 \times 3 = 6$
$X(B,C) \times Y(B,C) = 2 \times 1 = 2$
$X(C,A) \times Y(C,A) = 2 \times 3 = 6$
$X(C,B) \times Y(C,B) = 2 \times 1 = 2$

Zur Ermittlung der einzelnen Faktoren L sind schließlich die Bewertungen der Links nicht mehr allein nach der Anzahl der ausgehenden Links zu gewichten. Vielmehr erfolgt eine Gewichtung nach der wiederum bewerteten Anzahl der ausgehenden Links. Hierdurch ergeben sich die folgenden Gewichtungsquotienten Z(Ti) für die einzelnen Seiten Ti:

$Z(A) = X(A,B) \times Y(A,B) + X(A,C) \times Y(A,C) = 4$
$Z(B) = X(B,A) \times Y(B,A) + X(B,C) \times Y(B,C) = 8$
$Z(C) = X(C,A) \times Y(C,A) + X(C,B) \times Y(C,B) = 8$

Die einzelnen Bewertungsfaktoren L(T1,T2) für einen Link von Seite T1 nach Seite T2 ergeben sich damit als

$L(T1,T2) = X(T1,T2) \times Y(T1,T2) / Z(T1)$

und haben in unserem Beispiel die folgenden Werte:

$L(A,B) = 0.75$

L(A,C) = 0.25
L(B,A) = 0.75
L(B,C) = 0.25
L(C,A) = 0.75
L(C,B) = 0.25

Bei einem Dämpfungsfaktor d in Höhe von 0.5 ergeben sich damit die folgenden Gleichungen für die PageRank-Berechnung:

PR(A) = 0.5 + 0.5 (0.75 PR(B) + O.75 PR(C))
PR(B) = 0.5 + 0.5 (0.75 PR(A) + 0.25 PR(C))
PR(C) = 0.5 + 0.5 (0.25 PR(A) + 0.25 PR(B))

Aus der Lösung dieses Gleichungssystems folgen als PageRank-Werte für die einzelnen Seiten:

PR(A) = 819/693
PR(B) = 721/693
PR(C) = 539/693

Zuallererst sehen wir, dass Seite A den höchsten PageRank erhält. Dies ist darin begründet, dass Seite A sowohl von Seite B als auch von Seite C den jeweils stärker bewerteten Link erhält.

Es zeigt sich ferner, dass auch hier bei der Bewertung der einzelnen Links die Summe der PageRank-Werte aller Seiten mit 2079/693 gleich 3 und damit gleich der Anzahl der Seiten ist. Somit können die mittels des derart modifizierten PageRank-Algorithmus ermittelten Werte ohne weitere Normalisierung in die allgemeine Bewertung von Webseiten durch Google einfließen.

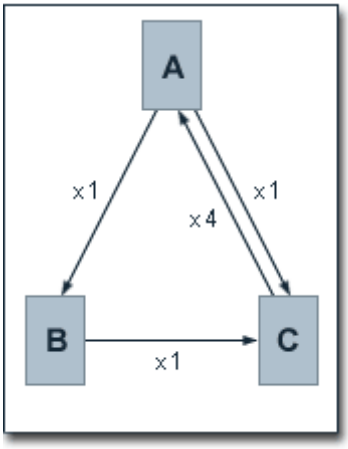
Unterschiedliche Bewertung von Links nach Eigenschaften der verweisenden Seite

Neben der unterschiedlichen Bewertung von Links innerhalb einer Seite führt Lawrence Page in der Patentschrift zum PageRank-Verfahren an, dass Links auch nach Kriterien gewichtet werden können, denen eine Bewertung der verweisenden Seite zu Grunde liegt. Dies scheint auf den ersten Blick überflüssig, da es bereits der Grundgedanke des PageRank-Konzepts ist, dass Links einen um so größeren Einfluss haben, je bedeutender die verlinkende Seite ist. Page und Brin erkannten allerdings sehr früh, dass ihr Algorithmus anfällig gegen das "künstliche Aufblähen" des PageRank einzelner Seiten ist.

Eine Beinflussung des PageRank kann in erster Linie dadurch erfolgen, dass Webmaster eine Vielzahl von Webseiten generieren, deren Links den PageRank so distribuieren, dass einzelne Seiten im System eine besondere Bedeutung erlangen. Diese Seiten können dann einen hohen PageRank inne haben, ohne dass von anderen Websites mit hoher Relevanz auf sie verlinkt wird. Hierdurch wird nicht nur das Konzept des PageRank unterwandert, sondern insbesondere auch der Suchmaschinenindex mit einer Vielzahl von Webseiten überflutet, die lediglich zum Zwecke der Beeinflussung des PageRank geschaffen wurden.

Als ein Mittel der Verhinderung dieser Form der Beinflussung zeigt Lawrence Page in seiner Patentschrift die Bewertung von Links anhand der Distanz zwischen verlinkender und

verlinkter Seite auf. Hintergrund ist, dass je größer die Distanz zwischen zwei Seiten ist, um so geringer ist die Wahrscheinlichkeit, dass ein und die selbe Person beide Seiten kontrolliert. Kriterium der Distanz zwischen Seiten kann etwa sein, ob Sie auf der selben Domain liegen oder nicht. Damit würden interne Links weniger gewichtet als externe. Aber auch jedes andere Kriterium der Distanz käme laut Page in Frage, also etwa ob Seiten sich auf dem selben Webserver befinden. Letztlich sei auch gerade die Verlinkung durch Websites aus unterschiedlichen geographischen Regionen ein deutliches Indiz für die Relevanz einer Seite.



Als weiteres Indiz für die Bedeutung einer Seite nennt Page die Aktualität der verlinkenden Seite. Die Informationen einer Seite sind mit viel geringerer Wahrscheinlichkeit veraltet, je mehr kürzlich modifizierte Seiten auf sie verlinken. Dagegen bevorzugt das eigentliche PageRank-Verfahren wie auch jedes Verfahren der Messung der Link-Popularität eher ältere Dokumente, die erst im Laufe ihrer Existenz eine Vielzahl eingehender Links erhalten haben und mit einer geringeren Wahrscheinlichkeit als neue Dokumente kürzlich verändert wurden. Grundsätzlich könnten aktuelle Dokumente mittels der bereits erwähnten Gewichtung des Faktors (1-d) besser bewertet werden. Hierdurch erhielten sowohl die aktuellen Dokumente selbst als auch diejenigen Dokumente auf die sie verlinken einen höheren PageRank. Die Aktualität einer Seite ist allerdings nicht zwingend ein Indiz für die Qualität der auf Ihr präsentierten Informationen. Daher ist es unbedingt ratsam, wie von Page vorgeschlagen, nicht die Aktualität einer Seite selbst zu bewerten, sondern vielmehr die Aktualität ihrer eingehenden Links.

Schließlich nennt Page als Kriterium für die Bedeutung eines Links noch die grundsätzliche Bedeutung der verlinkenden Seite. Als Beispiel wird in der Patentschrift zum PageRank Verfahren der Link von der Root-Seite einer Domain genannt. Hier könnte allerdings letztlich seitens Google ganz willkürlich auf das PageRank-Verfahren Einfluss genommen werden.

Um die Bewertung verlinkender Seiten in den PageRank-Algorithnmus aufzunehmen, muss der Bewertungsfaktor aus unserem modifizierten PageRank-Algorithmus nunmehr aus mehreren Einzelfaktoren bestehen. Für einen Link von Seite Ti nach Seite A könnte er wie folgt notiert werden:

$$L(Ti,A) = K(Ti,A) \times K1(Ti) \times ... \times Km(Ti)$$

Hierbei stellt K(Ti,A) die weiter oben vorgestellte Bewertung der einzelnen Links innerhalb einer Seite dar. Dazu erfolgt eine Bewertung der Seite Ti nach m Kriterien, die durch die Faktoren Kj(Ti) repräsentiert werden. Für eine Implementierung dieser Modifikationen muss im Falle der Bewertung von Seiten nun nicht mehr nur der PageRank-Algorithmus abgeändert werden, sondern auch das PageRank-Berechnungsverfahren. Dies soll wieder anhand eines Beispiels demonstriert werden.

Wir betrachten das 3-Seiten-Web aus den Seiten A, B und C, wobei Seite A sowohl auf Seite B als auch auf Seite C verlinkt. Seite B verlinkt auf Seite C und Seite C wiederum verlinkt auf Seite A. Alle ausgehenden Links einer Seite werden jeweils als gleichwertig betrachtet. Es erfolgt eine Bewertung der Links nach genau einem seitenspezifischen Kriterium. Ein Link von Seite C sei viermal bedeutender als ein Link von anderen Seiten. Nach entsprechender Gewichtung nach der Anzahl der Seiten ergibt sich das folgende Bild für unsere

Bewertungsfaktoren:

K(A) = 0.5
K(B) = 0.5
K(C) = 2

Bei einem Dämpfungsfaktor d in Höhe von 0.5 ergeben sich die folgenden Gleichungen für die Berechnung der PageRank-Werte der einzelnen Seiten:

PR(A) = 0.5 + 0.5 × 2 PR(C)
PR(B) = 0.5 + 0.5 × 0.5 × 0.5 PR(A)
PR(C) = 0.5 + 0.5 (0.5 PR(B) + 0.5 × 0.5 PR(A))

Die Lösung dieses Gleichungssystems ergibt die folgenden PageRank-Werte für die einzelnen Seiten:

PR(A) = 4/3
PR(B) = 2/3
PR(C) = 5/6

Es zeigt sich also, dass die Summe der PageRank-Werte nicht mehr gleich der Anzahl der Seiten ist. Dies liegt darin begründet, dass die erfolgte Gewichtung der Seitenbewertung nach der Anzahl der Seiten nicht korrekt war. Zur Ermittlung der korrekten Gewichtung müsste allerdings vorab die Linkstruktur des Webs antizipiert werden, was im Falle des WWW jedoch nicht möglich ist. Aus diesem Grunde ist bei der Bewertung von Links nach seitenspezifischen Faktoren der ermittelte PageRank zu normalisieren, damit kein unbegründet hoher oder geringer Einfluss des PageRank innerhalb der Gesamtbewertung von Seiten entsteht. Bei der iterativen PageRank-Berechnung hätte die Normalisierung nach jeder Iteration zu erfolgen, um unerwünschte Effekte zu minimieren.

Im Falle eines sehr kleinen Webs zeigen sich Verzerrungen des PageRank durch die Bewertungen von Links nach seitenspezifischen Kriterien sehr deutlich. Im Falle des tatsächlichen WWW dürften sich diese Verzerrungen weitestgehend ausgleichen. Es wäre allerdings zu befürchten, dass etwa die Bewertung der Distanz zwischen verlinkenden Webseiten durchaus zu Verzerrungen führen kann, da stark verlinkte Seiten sicherlich überdurchschnittlich dazu tendieren, aus unterschiedlichen geographischen Regionen verlinkt zu werden. Hier besteht allerdings die Möglichkeit zur Antizipation durch Erfahrungswerte aus vorangegangenen Berechnungszyklen, so dass die Normalisierung immer nur minimal sein müsste. Eine Einbeziehung zusätzlicher Bewertungskriterien in das PageRank-Verfahren ist in jedem Falle möglich, dabei allerdings mit einem erhöhten Rechenaufwand verbunden.

Themen-basierter PageRank

Die themen- bzw. themengebietsbezogene Homogenität von Webseiten wird schon seit geraumer Zeit als mögliches Ranking-Kriterium von Suchmaschinen diskutiert. Für die Integration von Themen in Suchmaschinenalgorithmen gibt es die verschiedensten Denkansätze. Ihnen gemein ist, dass Webseiten nicht mehr allein aufgrund Ihrer eigenen Inhalte bewertet werden, sondern dass auch die Inhalte anderer Webseiten hierzu berücksichtigt werden. So könnten also beispielsweise die Inhalte aller Seiten einer Website Einfluss auf die Bewertung einer einzelnen Seite dieser Website nehmen. Andererseits ist es

auch denkbar, dass eine Seite an den Inhalten derjenigen Seiten gemessen wird, auf die sie verlinkt oder aber von denen sie selbst verlinkt wird.

Sehr kontrovers diskutiert wird der mögliche Einsatz eines themenbasierten Rankings für die Suchmaschine Google. Immer wieder finden sich in einschlägigen Foren und auf Websites zum Thema Suchmaschinenoptimierung Ratschläge, dass eingehende Links von Seiten mit thematischer Ähnlichkeit einen größeren Einfluss auf den PageRank haben als Links von anderen Seiten. Diese Annahme soll hier kritisch beleuchtet werden. Zunächst werden hierzu zwei relativ neue Ansätze zur Integration von Themen in das PageRank-Verfahren diskutiert: auf der einen Seite das Modell des "intelligenten Surfers" von Matthew Richardson und Pedro Domingos und auf der anderen Seite der Topic-Sensitive PageRank von Taher Haveliwala. Anschließend sollen Möglichkeiten aufgezeigt werden, inwieweit Inhaltsanalysen und -vergleiche dazu eingesetzt werden können, thematische Ähnlichkeiten zwischen Seiten zu berechnen, um auf dieser Basis dann eine Gewichtung von Links im Rahmen des PageRank-Verfahrens vorzunehmen.

Der "intelligente Surfer" von Richardson und Domingos

Matthew Richardson und Pedro Domingos ziehen zur Erläuterung ihres Ansatzes zur Implementierung von Themengebieten in das PageRank-Verfahren zunächst das Random Surfer Modell heran. Sie schlagen anstelle eines Surfers, der wahllos Links verfolgt, einen intelligenteren Surfer vor, der einerseits Links nur entsprechend seiner Suchanfrage verfolgt und andererseits auch nach dem Abbruch des Surf-Vorgangs nur Seiten aufruft, die seiner Suchanfrage entsprechen.

Im Rahmen des Ansatzes von Richardson und Domingos sind für den "intelligenten Surfer" also nur Seiten relevant, die den von ihm gesuchten Begriff auch tatsächlichen enthalten. Das Random Surfer Modell ist jedoch nichts als ein Abbild des PageRank-Verfahrens. Zur Umsetzung muss also für jeden im Web existierenden Begriff eine eigene PageRank-Berechnung stattfinden. Diese Berechnung stützt sich dabei ausschließlich auf Links zwischen Seiten, die den jeweiligen Begriff enthalten.

Das Modell von Richardson und Domingos wirft einige Probleme auf. Vor allem entstehen diese im Bereich von Suchbegriffen, die nicht sehr häufig im Web vorkommen. Da diese wenigen Seiten sich auch noch verlinken müssen, um in die PageRank-Berechnung eingehen zu können, basieren die Resultate auf nur sehr kleinen Subbereichen des Webs und lassen gegebenenfalls sehr relevante Seiten außen vor. Ferner ist natürlich ein kleiner Subbereich des Webs wesentlich anfälliger für Spam im Sinne der Generierung zahlreicher Webseiten.

Zudem ergeben sich gravierende Probleme bezüglich der Skalierbarkeit. Richardson und Domingos schätzen sowohl den Speicher- als auch den Rechenbedarf für mehrere 100.000 Begriffe und entsprechende PageRank-Berechnungen auf das 100-200-fache des ursprünglichen PageRank-Verfahrens. Diese Zahlen klingen angesichts der großen Zahl relativ kleiner Subbereiche des Webs realistisch.

Der erhöhte Speicherbedarf sollte kein grundsätzliches Problem darstellen, da Richardson und Domingos hierzu richtig anführen, dass die begriffsspezifischen PageRank-Werte nur einen Bruchteil des Datenvolumens des inversen Index Google's ausmachen dürften. Wirklich problematisch ist der Zeitbedarf für die Berechnung. Kalkulieren wir nur mit fünf Stunden für eine herkömmliche PageRank-Berechnung, so würde diese im Falle des Modells von

Richardson und Domingos etwa drei Wochen in Anspruch nehmen. Dies stünde für den tatsächlichen Einsatz nicht zur Diskussion.

Taher Haveliwala's Topic-Sensitive PageRank

Der Ansatz von Taher Havilewala scheint für den tatsächlichen Einsatz vielversprechender. Auch Havilewala regt die Berechnung unterschiedlicher PageRanks für unterschiedliche Themenbereiche an. Hierbei sollen jedoch nicht hunderttausende PageRanks für verschiedene Subbereiche des Webs, sondern vielmehr wenige PageRanks auf der Basis des gesamten Webs berechnet werden. Bei dieser Berechnung wird zwar das gesamte Web berücksichtigt, es erfolgt jedoch jeweils eine dem Themengebiet entsprechende, unterschiedliche Gewichtung.

Die Grundlagen für den Ansatz von Havilewala wurden hier schon im Abschnitt zum "Yahoo-Bonus" beschrieben. Dabei wurde die Möglichkeit aufgezeigt, spezifischen Webseiten eine besondere Bedeutung im Rahmen des PageRank-Verfahrens zukommen zu lassen. Auf das Random Surfer Modell übertragen geschah dies dadurch, dass die Wahrscheinlichkeit erhöht wird, dass der Zufalls-Surfer nach dem Abbruch eines Surf-Vorgangs eine bestimmte Seite aufsucht. Diese Einflussnahme auf das PageRank-Verfahren wirkt sich dann über Links auf den PageRank aller Seiten des Webs aus. Konkret erreicht wurde diese Einflussnahme durch die Implementierung eines weiteren Wertes E in den PageRank Algorithmus:

$$PR(A) = E(A) (1-d) + d (PR(T1)/C(T1) + ... + PR(Tn)/C(Tn))$$

Havilewala geht in seinem Ansatz zum Topic-Sensitive PageRank einen Schritt weiter. Er weist keiner Site oder Seite eine grundlegende und allgemeingültige höhere Wertung zu, sondern differenziert diese auf der Basis bestimmter Themengebiete. Für jedes dieser Themengebiete identifiziert er jeweils andere Seiten mit besonderer Kompetenz. Auf der Grundlage dieser Bewertungen werden dann unterschiedliche PageRanks jeweils für das gesamte Web berechnet.

In seiner Arbeit zum Topic-Sensitive PageRank wählte Haveliwala die 16 Hauptkategorien des Open Directory Projekt sowohl zur Identifizierung von Themengebieten als auch für die besondere Bewertung innerhalb der PageRank-Berechung aus. Konkret weist Haveliwala für die einzelnen PageRank-Berechnungen den jeweiligen Seiten unter den Hauptkategorien des ODP einen hohen Wert E innerhalb des PageRank Algorithmus zu. Wird etwa der PageRank für das Thema Gesundheit berechnet, erhalten die ODP-Seiten der Kategorie Gesundheit jeweils einen relativ höheren Wert E, der sich dann auf die von dort verlinkten Seiten auswirkt. Dies setzt sich natürlich fort, und unter der Annahme, dass Websites zum Thema Gesundheit sich tendenziell verstärkt gegenseitig verlinken, haben all diese Seiten im Rahmen des Themas Gesundheit einen relativ höheren PageRank.

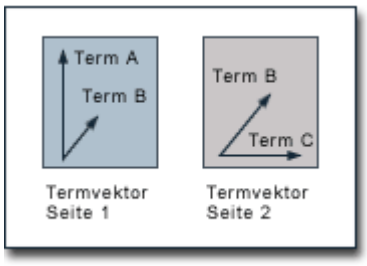
Haveliwala konstatiert die Unvollkommenheit der Wahl des Open Directory Project, die sich etwa in einer großen Abhängigkeit von den Editoren des ODP und in einer nur sehr groben Untergliederung in Themengebiete äußert, sie liefert allerdings offensichtlich bereits gute Ergebnisse und kann sicherlich ohne großen Aufwand verbessert werden.

Ein Schwerpunkt der Arbeit zum Topic-Sensitive PageRank ist die Identifizierung der Präferenzen des Benutzers. Über themenspezifische Bewertungsmöglichkeiten zu verfügen ist

nutzlos, so lange man nicht darüber informiert ist, welche Themengebiete für den Benutzer interessant sind. Schließlich soll für jeweilige Suchanfragen immer nur derjenige PageRank in die Seitenbewertung einfließen, der für die Suchanfrage des Benutzers auch tatsächlich relevant ist. Erst hierdurch kann der Topic-Sensitive PageRank tatsächlich genutzt werden.

Auch zur Identifikation der Benutzerpräferenzen liefert Haveliwala allerdings praktikable Ansätze. So beschreibt er beispielsweise die Suche im Kontext durch Markieren eines Begriffes innerhalb eines Dokuments - und damit den Inhalt dieses Dokuments als Anhaltspunkt für die Identifizierung von Benutzerpräferenzen. An dieser Stelle soll dazu wiederum an die Möglichkeiten der Google Toolbar erinnert werden. Die Toolbar überträgt Daten zu Suchbegriffen und besuchten Seiten an Google und könnte damit leicht zur Erstellung von Benutzerprofilen dienen. Doch auch ohne Installation der Toolbar wäre letztlich eine aktive Auswahl eines Themengebiets durch den User jeweils vor seiner Suche denkbar.

Die Bewertung von Links auf der Basis von Inhaltsanalysen



Dass grundsätzlich eine Gewichtung einzelner Links im Rahmen des PageRank-Verfahrens möglich ist, wurde auf der vorigen Seite bereits gezeigt. Der Hintergrund einer Gewichtung von Links auf der Basis von Inhaltsanalysen würde in erster Linie in der Verhinderung der Korrumpierung des Pagerank-Verfahrens liegen. So könnte theoretisch mittels Inhaltsanalysen erreicht werden, dass Links, die ausschließlich zum Zwecke der Steigerung des PageRanks bestimmter Sites gesetzt werden, in vielen Fällen in weitaus geringerem Maße auf den PageRank Einfluss nehmen. Fraglich ist allerdings, ob eine derartige Bewertung auf der Basis von Inhaltsanalysen auch tatsächlich umgesetzt werden kann.

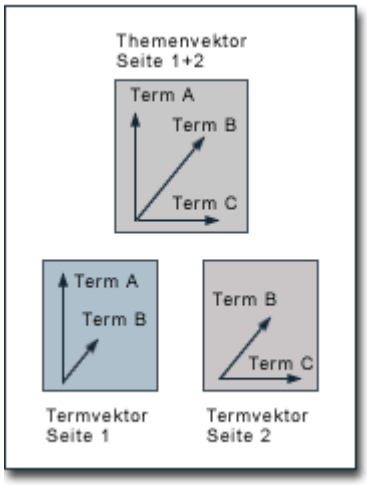
Die Grundlagen zum Vergleich von Inhalten wurden bereits in den 60er und 70er Jahren des 20. Jahrhunderts von Gerard Salton erarbeitet. Sein Vektorraummodell bildet Dokumente als Vektoren aus Termen (Begriffe innerhalb eines Dokuments) und deren Gewichtung ab. Diese Termvektoren können dann miteinander verglichen werden, indem z.B. das Kosinus-Maß (Skalarprodukt) berechnet wird, um inhaltliche Ähnlichkeiten zwischen den Dokumenten zu messen. In seiner einfachen Form weist das Vektorraummodell einige Schwächen auf. So wird etwa die grundsätzliche Annahme kritisiert, dass die Ähnlichkeit zwischen Dokumenten daran bemessen wird, ob und in welchem Ausmaß einzelne Terme tatsächlich in zwei zu vergleichenden Dokumenten vorkommen. Mittlerweile gibt es jedoch zahlreiche Erweiterungen und Verfeinerungen des Vektorraummodells die viele der Probleme beheben.

Mit Arbeiten die auf Saltons Vektorraummodell aufbauen hat sich vor allem auch Krishna Bharat hervorgetan. Dies ist in erster Linie von Interesse, weil Bharat mittlerweile zu Googles Mitarbeiterstab zählt und insbesondere als Entwickler der "Google News" (news.google.com) gilt. Google News ist ein Service, der Nachrichten-Websites spidert, die einzelnen Nachrichten auswertet und anschließend in unterschiedlichen Kategorien zu unterschiedlichen Themen zusammenfasst. Nach Angaben Googles erfolgen all diese Vorgänge vollkommen automatisiert. Hierzu werden weitere Kriterien wie etwa der Zeitpunkt des Erscheinens eines jeweiligen Artikels herangezogen, sofern jedoch keinerlei manuelle Eingriffe stattfinden, ist eine Zusammenfassung unter inhaltlichen Gesichtspunkten nur möglich, wenn die Inhalte der einzelnen Nachrichten zunächst einmal tatsächlich miteinander verglichen werden. Es stellt

sich nur die Frage, wir dies realisiert werden kann.

In Ihrer Veröffentlichung zum Aufbau einer Termvektor-Datenbank beschreiben Raymie Stata, Krishna Bharat und Farzin Maghoul sehr anschaulich, wie Vergleiche zwischen Inhalten auf der Basis von Termvektoren realisiert und vor allem auch, wie verschiedene Hürden bei Umsetzung überwunden werden können. Zunächst besteht die Problematik, dass zahlreiche Begriffe innerhalb eines Dokuments nicht für einen Inhaltsvergleich geeignet sind. So wird aus der Gesamtheit aller Begriffe zuerst das am häufigsten vorkommende Drittel gefiltert, da diese Begriffe nur zu einem sehr geringen Grad dazu beitragen können, die Inhalte von Dokumenten voneinander zu unterscheiden. Da relativ selten vorkommende Begriffe, die z.B. auch aus Tippfehlern resultieren können, gegebenenfalls thematisch sehr unterschiedliche Dokumente sehr ähnlich erscheinen lassen, weil die entsprechenden Begriffe insgesamt sehr selten vorkommen, wird auch das am wenigsten auftretende Drittel gefiltert, womit für die Durchführung von Vergleichen nurmehr ein Drittel aller Begriffe genutzt wird.

Auch wenn bereits zwei Drittel aller Begriffe nicht in die Termvektoren gelangen können, ist diese Auswahl für einen Vergleich noch wenig effizient. Stata, Bharat und Maghoul führen deshalb vor dem Aufbau der Termvektoren eine weitere Filterung durch, so dass ein Termvektor jeweils auf maximal 50 Begriffen basiert. Diese 50 Begriffe sind jedoch nicht etwa die 50 am häufigsten innerhalb eines Dokuments auftretenden Begriffe. Vielmehr werden die 50 Begriffe genutzt, für die die Relation aus dem Vorkommen innerhalb eines Dokuments zum Vorkommen innerhalb der Gesamtheit aller Dokumente am größten ist. Gerade hierdurch wird es möglich, die Inhalte von Dokumenten tatsächlich voneinander abzugrenzen.



Die beschriebenen Maßnahmen sind Standards im Rahmen der Nutzung von Termvektoren. Wenn z.B. das Skalarprodukt aus zwei derart ermittelten Termvektoren relativ hoch ist, sind die beiden entsprechenden Seiten einander unter thematischen Gesichtspunkten tendenziell ähnlich. Diese Vorgehensweisen ermöglichen Inhaltsvergleiche in vielen Bereichen, ob sie allein jedoch für unser Ziel der Gewichtung von Links im Rahmen des PageRank-Verfahrens ausreichend sind, ist zu bezweifeln. Schließlich können vor allem Synonyme, aber auch andere Begriffe, die Ähnliches umschreiben, mittels der beschriebenen Vorgehensweisen nicht identifiziert werden. Für das Problem der Zusammenfassung von Singular und Plural etwa, existieren für die englische Sprache relativ einfache Algorithmen. In anderen Sprachen ist dies jedoch ungleich schwerer zu bewältigen. Unterschiedliche Sprachen sind dabei ein grundsätzliches Problem. Bis auf die Ausnahme von z.B. Lehnwörtern oder Markennamen werden verschiedensprachige Texte in der Regel keine gemeinsamen Begriffe enthalten, oder aber gemeinsame Begriffe haben eine völlig unterschiedliche Bedeutung, so dass ein Vergleich zwischen Texten in unterschiedlichen Sprachen nicht möglich ist. Doch auch hierfür bieten Stata, Bharat und Maghoul einen Lösungsansatz.

Stata, Bharat und Maghoul zeigen eine sehr konkrete Anwendungsmöglichkeit ihrer Termvektor-Datenbank auf, indem Sie für einzelne Dokumente ein entsprechendes, vordefiniertes Themengebiet identifizieren. Über diese Identifizierung von Themengebieten hat Bharat auch gemeinsam mit Monika Henzinger - derzeit Research Director bei Google - veröffentlicht, und sie funktioniert folgendermaßen: Zunächst werden sogenannte

Themenvektoren berechnet. Themenvektoren sind selbst auch Termvektoren, nur dass Sie nicht auf den Inhalten einer einzelnen Webseite basieren, sondern auf den Inhalten vieler Webseiten, denen eine inhaltliche Ähnlichkeit gemein ist. Um einen Themenvektor aufbauen zu können, muss es für jedes vordefinierte Themengebiet eine bestimmte Anzahl an Webseiten geben, für die bekannt ist, welchem Themengebiet sie zugeordnet werden können. Zu diesem Zwecke greifen Stata, Bharat und Maghoul auf Web-Verzeichnisse zurück.

In einer konkreten Anwendung von Themenvektoren haben sie auf der Basis von jeweils ca. 30.000 Links innerhalb der damals 12 Hauptkategorien des Yahoo-Verzeichnisses Themenvektoren mit einem Umfang von jeweils ca. 10.000 Begriffen gebildet. Um das Thema einer Webseite identifizieren zu können, haben sie anschließend die Ähnlichkeiten zwischen dem entsprechenden Termvektor und den einzelnen Themenvektoren berechnet. Derjenige Themenvektor, für den der höchste Wert ermittelt wird, bestimmt das Thema der Webseite. Dass die Einordnung von Themen in der Praxis gut funktioniert kann wiederum anhand von Google News beobachtet werden. Einzelne Artikel werden nicht nur zu einem konkreten Thama zusammengefasst, sondern auch noch in eine der Kategorien World, U.S., Business, Sci/Tech, Sports, Entertainment und Health eingeordnet. Solange eine derartige Kategorisierung nicht über die Website-Strukturen der Quellen für die Artikel erfolgt (was unwahrscheinlich ist), muss tatsächlich das Thema eines Artikels bzw. einer Gruppe von Artikeln berechnet werden.

Krishna Bharat beschäftigte sich zum Zeitpunkt seiner Veröffentlichungen nicht mit PageRank, sondern vielmehr mit dem Kleinberg-Algorithmus, so dass er weniger die Gewichtung von Links als vielmehr das Filtern von inhaltlich unzusammenhängenden Links verfolgt hat. Der Schritt zu einem Vergleich von Inhalten für die Gewichtung von Links im Rahmen des PageRank ist jedoch nur kurz. Anstatt der Inhalte von zwei sich verlinkenden Seiten werden nurmehr die für sie identifizierten Themengebiete verglichen. So könnten beispielsweise die Grade der Zugehörigkeit eines jeden Dokuments zu jeweils allen Themengebieten in einem Themenzugehörigkeitsvektor erfasst werden. Diese Vektoren können dann als Grundlage für den Vergleich zweier sich verlinkender Webseiten gewählt werden und somit der Gewichtung der Links dienen.

Die Nutzung von Themenvektoren bietet gegenüber dem direkten Vergleich von Termvektoren einen wesentlichen Vorteil: Ein Themenvektor kann auf Begriffen aus unterschiedlichen Sprachen basieren. Hierzu müssen lediglich z.B. Seiten aus den nationalen Yahoo-Versionen berücksichtigt werden. Mögliche Abweichungen in den Verzeichnis-Strukturen können sicherlich manuell angepasst werden. Besser wäre gegebenenfalls ein Rückgriff auf das ODP, dessen Strukturen sich innerhalb der Kategorie "World" an die Struktur der Hauptkategorien anlehnen. Hierdurch wäre die Feststellung thematischer Ähnlichkeiten zwischen verlinkenden Seiten auch multilingual zu realisieren, so dass eine sinnvoll geartete Gewichtung von Links auf der Basis von Inhaltsanalysen durchaus möglich erscheint.

Gibt es eine tatsächliche Implementierung von Themen in das PageRank-Verfahren?

Dass die Ansätze von Haveliwala sowie Richardson und Domingos zwar durchaus interessant sind, aber derzeit nicht eingesetzt werden, ist offensichtlich. Man könnte es unmittelbar bei der Nutzung Googles erkennen. Eine Gewichtung von Links auf der Basis von Inhaltsanalysen hingegen wäre nicht unmittelbar zu bemerken. Dass sie theoretisch möglich ist, wurde gezeigt. Ob sie aber auch praktisch umgesetzt wird, ist durchaus zweifelhaft.

Es soll an dieser Stelle nicht der Anspruch erhoben werden, den einzig möglichen Weg zur Gewichtung von Links aus der Basis von Inhaltsanalysen aufgezeigt zu haben. Es gibt in der Tat sicherlich dutzende andere. Der hier vorgestellte orientiert sich jedoch an Veröffentlichungen wichtiger Google-Mitarbeiter, was ihn dazu qualifiziert, auf ihn eine kritische Beurteilung zu stützen.

Wie immer im Rahmen des PageRank-Verfahrens, so stellt sich auch hier die Frage, ob ein Einsatz der vorgestellten Lösung hinreichend skalierbar ist. Einerseits erfordert sie zusätzliche Speicherkapazitäten. Die zitierte Arbeit von Stata, Bharat und Maghoul beschreibt schließlich gerade die Architektur einer Termvektoren-Datenbank, die sich in Ihrer Struktur grundlegend von Google's inversem Index unterscheidet, da sie aus Effizienzgründen von Seiten-IDs auf Terme referenziert und damit kaum in bestehende Architekturen integriert werden kann. Der zusätzliche Speicherbedarf dürfte für die aktuelle Indexgröße zwischen mehreren hundert GB und wenigen TB liegen. Dies sollte angesichts eines um ein Vielfaches größeren Index allerdings nicht sehr ins Gewicht fallen. Problematischer ist der Zeitbedarf für den Aufbau der Datenbank und die Berechnung der Gewichtungen.

Der Aufbau einer Termvektor-Datenbank sollte sich unter zeitlichen Aspekten etwa in der Größenordnung des Aufbaus des inversen Index bewegen. Natürlich können wir davon ausgehen, dass etliche Prozesse gleichsam für den Aufbau beider Datenbanken genutzt werden können. Sobald jedoch zum Beispiel die Gewichtung der Terme innerhalb einzelner Termvektoren von ihrer Gewichtung innerhalb des Index abweichen muss, bleibt der Zeitbedarf erheblich. Sofern wir davon ausgehen, dass wie in unserem Lösungsansatz hier, das Skalarprodukt der aus Term- und Themenvektoren errechneten Themenzugehörigkeitsvektoren bestimmt werden soll, so können wir davon ausgehen, dass dieser Prozess einen Zeitaufwand darstellt, der sich im Rahmen der eigentlichen PageRank-Berechnung bewegt. Natürlich muss auch hier bedacht werden, dass die PageRank-Berechnung selbst durch die Gewichtung von Links zusätzlich an Komplexität gewinnt.

Der zusätzliche Aufwand wäre also gewiss nicht unerheblich. Vor allem auch deshalb stellt sich die Frage, ob eine Gewichtung von Links überhaupt sinnvoll ist. Links, die zwischen thematisch unzusammenhängigen Seiten allein zum Zwecke der PageRank-Erhöhung einer der beiden Seiten gesetzt werden, mögen zwar ärgerlich sein, sie dürften jedoch nur einen minimalen Anteil an der Gesamtheit aller Links ausmachen. Andererseits ist das Web an sich vollkommen inhomogen. Google, Yahoo oder das ODP verdanken ihren hohen PageRank sicherlich nicht nur eingehenden Links von anderen Suchdiensten. Ein großer Teil der Links innerhalb des Webs werden einfach nicht mit dem Ziel gesetzt, Besuchern einen Weg zu weiteren, thematisch verwandten Informationen zu weisen. Die Motivation für das Setzen von Links ist vielmehr vielfältig. Weiterhin sind die wohl beliebtesten Websites in sich vollkommen inhomogen. Man denke nur an Portale wie Yahoo oder aber an Nachrichten-Websites, deren Artikel allen Bereichen menschlichen Lebens entstammen. Eine starke Gewichtung von Links in der hier beschriebenen Form würde sich drastisch auf ihren PageRank auswirken.

Eine Gewichtung von Links dürfte also nur sehr eingeschränkt stattfinden, wenn das PageRank-Verfahren nicht ad absurdum geführt werden soll. Dies wirft dann natürlich die Frage auf, ob dann der erforderliche Aufwand gerechtfertigt ist. Schließlich gibt es durchaus andere Möglichkeiten, den Spam, der beispielsweise durch erkaufte, thematisch unzusammenhängende Links in den Suchergebnissen nach vorn kommen kann, an das das Ende der Suchergebnisse zu verbannen.

PR0 - Die PageRank 0 Bestrafung

Seit Ende des Jahres 2001 greift die Bestrafung von Websites mit einem PageRank von 0 um sich. In einschlägigen Suchmaschinenoptimierungs-Foren hat sich hierfür die Kurzform PR0 eingebürgert und diese soll auch hier benutzt werden. PR0 ist dadurch gekennzeichnet, dass alle - oder zumindest viele - Seiten einer Website in der Google Toolbar einen PageRank von 0 aufweisen, obwohl diese mitunter qualitativ hochwertige eingehende Links aufweisen können. Sie sind nicht vollkommen aus dem Index entfernt, erscheinen aber in Suchergebnissen stets ganz unten und sind somit praktisch nicht aufzufinden.

Einem PageRank von 0 muss natürlich nicht immer eine Bestrafung zu Grunde liegen. Vielen vermeintlich bestraften Seiten mangelt es schlicht an eingehenden Links mit entsprechend hohem PageRank. Wenn aber die Seiten einer Site, die vormals gut in den Suchergebnissen platziert waren, plötzlich die gefürchtete weiße PageRank-Anzeige aufweisen, und sich hinsichtlich der eingehenden Links der Site nichts wesentliches verändert hat, liegt nach herrschender Meinung eine Bestrafung durch Google vor.

Über die tatsächlichen Ursachen des PR0 kann natürlich nur spekuliert werden. Da seitens Google mittlerweile nicht mehr über technische Details und grundlegende Algorithmen publiziert wird, sind schließlich erforderliche Hintergrundinformationen kaum oder gar nicht verfügbar. Nichtsdestotrotz soll wegen der tiefgreifenden Auswirkungen von PR0 ein theoretischer Ansatz hierfür geliefert werden.

Suchmaschinen-Spam ist eines der großen Probleme mit denen Suchmaschinen-Betreiber seit jeher zu kämpfen haben. Die übliche Vorgehensweise gegen Spam war immer, dass - sobald Spam identifiziert wird - die entsprechenden Domains oder auch gleich IP-Adressen in der Regel für unbestimmte Zeit aus dem Index verbannt werden.

Ein derartiges manuelles Entfernen von Websites aus dem Index ist immer mit einem hohen Personalaufwand verbunden. Dies läuft der stets von Google angestrebten hohen Skalierbarkeit der Suchmaschine zuwider. Es ist hiermit also erforderlich, Spam automatisiert zu filtern. Hierdurch entsteht jedoch die Gefahr, auch viele unschuldige Webmaster zu bestrafen. Die eingesetzten Filter dürfen also nur sehr sensibel auf potentiellen Spam reagieren. Um dabei dennoch effektiv zu sein, kann es - wie auch im Rahmen des PageRank-Verfahrens - sinnvoll sein, Linkstrukturen zu analysieren.

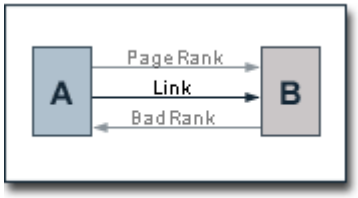
Eine derartige Vorgehensweise wurde von einem Mitarbeiter Google's, der unter dem Pseudonym GoogleGuy auftritt, mehrmals mehr oder weniger eindeutig im Google-Forum von WebmasterWorld bestätigt. Bekannt wurde sie als Bestrafung für das "linking to bad neighborhoods". Im folgenden soll erörtert werden, wie eine derartige Identifikation von Spam über die Analyse von Linkstrukturen realisiert werden kann. Insbesondere soll dabei gezeigt werden, wie mittels solcher Verfahren ganze Netzwerke von Spam-Seiten, die gegebenenfalls auch auf viele verschiedene Domains verteilt sind, ans EnSuchmaschinen-Spam ist eines der großen Probleme mit denen Suchmaschinen-Betreiber seit jeher zu kämpfen haben. Die übliche Vorgehensweise gegen Spam war immer, dass - sobald Spam identifiziert wird - die entsprechenden Domains oder auch gleich IP-Adressen in der Regel für unbestimmte Zeit aus dem Index verbannt werden.

Ein derartiges manuelles Entfernen von Websites aus dem Index ist immer mit einem hohen Personalaufwand verbunden. Dies läuft der stets von Google angestrebten hohen Skalierbarkeit der Suchmaschine zuwider. Es ist hiermit also erforderlich, Spam automatisiert zu filtern. Hierdurch entsteht jedoch die Gefahr, auch viele unschuldige Webmaster zu bestrafen. Die eingesetzten Filter dürfen also nur sehr sensibel auf potentiellen Spam reagieren. Um dabei dennoch effektiv zu sein, kann es - wie auch im Rahmen des PageRank-Verfahrens - sinnvoll sein, Linkstrukturen zu analysieren.

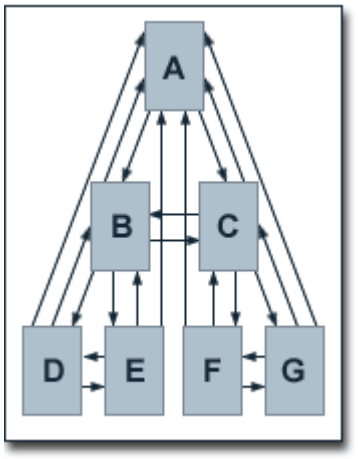
Eine derartige Vorgehensweise wurde von einem Mitarbeiter Google's, der unter dem Pseudonym GoogleGuy auftritt, mehrmals mehr oder weniger eindeutig im Google-Forum von WebmasterWorld bestätigt. Bekannt wurde sie als Bestrafung für das "linking to bad neighborhoods". Im folgenden soll erörtert werden, wie eine derartige Identifikation von Spam über die Analyse von Linkstrukturen realisiert werden kann. Insbesondere soll dabei gezeigt werden, wie mittels solcher Verfahren ganze Netzwerke von Spam-Seiten, die gegebenenfalls auch auf viele verschiedene Domains verteilt sind, ans Ende der Ergebnisseiten verbannt werden können.

BadRank als Umkehrung von PageRank

Der hier präsentierte theoretische Ansatz zum PR0 wurde grundlegend zuerst von Raph Levien (www.advogato.org/person/raph) formuliert. Basis dieses Ansatzes ist es, dass neben PageRank ein weiteres Verfahren eingeführt wird, das nicht wie PageRank die grundsätzliche Bedeutung einer Webseite im positiven Sinne bestimmt, sondern vielmehr die negativen Eigenschaften von Webseiten misst. Der Einfachheit halber soll dieses Verfahren hier BadRank genannt werden.



Das "linking to bad neighborhoods" bildet die Grundlage für den BadRank. Eine Seite, die auf eine andere Seite mit einem hohen BadRank verlinkt, erhält hierdurch tendenziell selbst einen hohen BadRank. Hiermit werden die Parallelen zu PageRank bereits offentsichtlich, nur dass BadRank nicht auf der Bewertung der eingehenden Links einer jeweiligen Webseite beruht, sondern vielmehr auf deren eigenen ausgehenden Links. In diesem Sinne ist BadRank gewissermaßen eine Umkehrung von PageRank. In einer direkten Adaption des PageRank Algorithmus würde sich die folgende Formel für den BadRank ergeben:



$$BR(A) = E(A) (1-d) + d (BR(T1)/C(T1) + ... + BR(Tn)/C(Tn))$$

Hierbei ist

- BR(A) der BadRank von Seite A,
- BR(Ti) der BadRank derjenigen Seiten T, auf die Seite A verlinkt,
- C(Ti) die Anzahl der eingehenden Links der jeweiligen Seite Ti und
- d der auch hier erforderliche Dämpfungsfaktor.

Der Wert E(A) entsprach im Rahmen einer der hier diskutierten Modifikationen des PageRank Algorithmus der manuellen Höherbewertung spezifischer Webseiten. Im Rahmen des BadRank Algorithmus reflektiert

dieser Wert, ob eine Seite beim spidern des Webs von einem Spam-Filter erfasst wurde. Ohne diesen Wert E(A) wäre der BadRank Algorithmus vollkommen nutzlos, da es sich wiederum lediglich um eine Analyse von Linkstrukturen handeln würde, der aber keine weiteren Kriterien zu Grunde lägen.

Mit Hilfe des hier präsentierten BadRank-Algorithmus können also zunächst Spam-Seiten erfasst werden. Ihnen wird dann über E(A) ein numerischer Wert zugewiesen, der beispielsweise der Schwere des Spammings entspricht oder sich vielleicht besser am PageRank einer jeweiligen Seite orientiert, wobei die Summe aller E(A) gleich der Anzahl der Webseiten sein muss. Im Laufe einer iterativen Berechnung überträgt sich dieser zugewiesene Wert dann nicht nur als BadRank auf diejenigen Seiten, die auf Spam-Seiten verlinken. Vielmehr wäre BadRank in der Lage, Regionen des Webs zu identifizieren, in denen Spam besonders häufig auftritt, ganz ähnlich wie PageRank Regionen des Webs identifiziert, denen eine grundlegende Bedeutsamkeit zukommt.

BadRank und PageRank weisen dabei natürlich gravierende Unterschiede auf, die vor allem darin begründet sind, dass die Verteilung von eingehenden und ausgehenden Links ganz entscheidend voneinander abweicht. Unser Beispiel stellt eine einfache, hierarchisch strukturierte Website dar, die natürliche Linkstrukturen wohl recht gut abbildet. Dabei verlinken alle Seiten jeweils auf all diejenigen Seiten, die hierarchisch über ihnen angeordnet sind. Zudem verlinken sie auf die ihnen direkt untergeordneten Seiten und diejenigen Seiten innerhalb einer solchen Kategorie verlinken wiederum einander.

Die Verteilung der eingehenden und ausgehenden Links innerhalb einer derartigen Site gibt die folgende Tabelle wieder.

Ebene	eingehende Links	ausgehende Links
0	6	2
1	4	4
2	2	3

Wie zu erwarten, erfolgt hinsichtlich der eingehenden Links eine hierarchische Abstufung von oben nach unten. Die Anzahl der ausgehenden Links ist hingegen in der mittleren Hierarchiestufe am höchsten. Ein ganz ähnliches Bild zeigt sich, wenn wir eine weitere Ebene unten in unsere Beispiel-Site einfügen, die den oben beschriebenen Richtlinien folgt:

Ebene	eingehende Links	ausgehende Links
0	14	2
1	8	4
2	4	5
3	2	4

Wiederum konzentriert sich die Zahl der ausgehenden Links in den mittleren Hierarchiestufen. Vor allem aber, ist die Verteilung der ausgehden Links wesentlich gleichmäßiger als die der eingehenden Links.

Wenn wir in unserem ursprünglichen Beispiel der Index-Seite A einen Wert E(A) gleich 100 zuweisen, wobei alle anderen Werte E gleich 1 sind, ergeben sich bei einem Dämpfungsfaktor

d von 0.85 die folgenden BadRank-Werte:

Seite	BadRank
A	22.39
B/C	17.39
D/E/F/G	12.21

Es zeigt sich, dass der BadRank sich von der Index-Seite aus weiter auf alle anderen Seiten der Site verteilt. Auf die Kombination von BadRank und PageRank wird weiter unten noch detaillierter eingegangen, ganz gleich jedoch, wie diese erfolgt, ist es unmittelbar ersichtlich, dass beide sich sehr gut neutralisieren könnten. Schließlich können wir davon ausgehen, dass auch der PageRank abnimmt, je weiter wir uns in der Seitenhierarchie nach unten bewegen. Mit einer derartigen Neutralisierung kann in jedem Falle ein PR0 für alle Seiten erreicht werden.

Nehmen wir nun andererseits an, dass die hierachisch nachgeordnete Seite G auf eine Seite X mit einem fixen BadRank BR(X)=10 verlinkt, wobei der Link von Seite G der einzige eingehende Link von Seite X ist und alle Werte E für unser Beispiel-Site gleich 1 sind, ergeben sich bei einem Dämpfungsfaktor d von 0.85 die folgenden Werte:

Seite	BadRank
A	4.82
B	7.50
C	14.50
D	4.22
E	4.22
F	11.22
G	17.18

Hier ist die Verteilung des BadRank weit weniger homogen als im vorangegangenen Szenario. Nichtsdestotrotz erfolgt eine Distribution des BadRank über die gesamte Site. Bemerkenswert ist, dass der BadRank der Index-Seite A relativ gering ist. Es wäre damit problematisch, einen im Vergleich zu den anderen Seiten höheren PageRank gleichermaßen zu neutralisieren. Dieser Effekt mag wenig wünschenswert sein, er spiegelt jedoch die Beobachtungen zahlreicher Webmaster wider: Relativ häufig tritt das Phänomen auf, dass alle Seiten bis auf die jeweilige Index-Seite einer Site einen PR0 aufweisen, wobei die Index-Seite oft einen Toolbar-PageRank von 2 bis 4 hat. Es drängt sich damit die Vermutung auf, dass diese spezielle Form des PR0 nicht darauf beruht, dass die entsprechende Website von einem der Spam-Filter identifiziert wurde, sondern dass sie ihre Bestrafung aufgrund eines "linking to bad neighborhoods" erhalten hat. Ferner wäre es natürlich auch möglich, dass diese Form des PR0 darin begründet ist, dass lediglich hierarchisch nachgeordnete Bereiche einer Website von einem Spam-Filter erfasst wurden.

Die Kombination von PageRank und BadRank zum PR0

Wenn wir davon ausgehen, dass ein BadRank in der hier präsentierten Form existiert, stellt sich nun die Frage, in welcher Form BadRank und PageRank kombiniert werden können, um

einerseits möglichst viele Spammer aus den Suchergebnissen zu eliminieren und andererseits möglichst wenige unschuldige Seitenbetreiber ungerechtfertigterweise zu bestrafen.

Rein intuitiv bietet sich eine Verwendung der BadRank-Werte im Rahmen der eigentlichen PageRank-Berechnung an. So könnte beispielsweise im Zuge der iterativen Berechnung der PageRank einer Seite direkt durch ihren BadRank dividiert werden. Dies hätte den Vorteil, dass eine Seite mit hohem BadRank auch keinen bzw. nur einen minimalen PageRank weitergeben kann. Schließlich kann man argumentieren, dass wenn eine Seite auf eine suspekte Seite verlinkt, die anderen Seiten, auf die sie verlinkt, ebenfalls suspekt sind. Eine direkte Verbindung birgt allerdings große Gefahren. So sind beispielsweise die Auswirkungen auf den PageRank in keinster Weise vorab einzuschätzen. Insbesondere ist zu beachten, dass eine Seite, der die Möglichkeit genommen wird PageRank weiterzugeben, zu einem Dangling Link wird. Wie jedoch im Abschnitt zu den ausgehenden Links erörtert wurde, ist es unbedingt erforderlich, Dangling Links im Rahmen der PageRank-Berechnung zu vermeiden.

Es ist also sinnvoll, PageRank und BadRank getrennt voneinander zu berechnen. Die anschließende Kombination der beiden kann dabei auf einfachen arithmetischen Berechnungen beruhen. Eine Subtraktion hätte die grundsätzlich wünschenswerte Folge, dass relativ geringe BadRank-Werte bei relativ hohen PageRank-Werten kaum Einfluss hätten. Allerdings wäre es mit der Subtraktion problematisch, tatsächlich einen PR0 für viele Seiten zu erreichen. Es würde vielmehr eine breite Abstufung in niedrige PageRank-Regionen stattfinden. Mit der Division von PageRank durch BadRank wäre ein PR0 leicht zu erreichen. Dies würde jedoch implizieren, dass der BadRank eine extrem große Bedeutung erhält. Vor allem jedoch ist ein sehr großer Teil der BadRank kleiner als 1, da auch der Durchschnitt des BadRanks 1 ist, womit eine Normalisierung erforderlich wäre. Eine Normalisierung und Skalierung des BadRanks auf Werte zwischen 0 und 1, so dass "gute" Seiten Werte nahe 1 und "schlechte" Seiten Werte nahe 0 aufweisen, und eine anschließende Multiplikation dieser Werte mit dem PageRank einer Seite dürfte hier die besten Ergebnisse liefern.

Womöglich am effektivsten und am einfachsten zu realisieren wäre jedoch eine schlichte, abgestufte Beurteilung von PageRank und BadRank. Denkbar ist, dass sofern der BadRank einen bestimmten Wert überschreitet, es stets zum PR0 kommt. Gleiches gilt, wenn die Relation aus PageRank zu BadRank einen bestimmten Wert unterschreitet. Daneben ist es sinnvoll, dass wenn der BadRank und/oder die Relation aus BadRank zu PageRank unter einem bestimmten Wert liegen, der BadRank keinen Einfluss nimmt. Nur wenn keiner dieser Fälle eintritt, wäre eine tatsächliche Kombination von PageRank und BadRank, etwa durch Division von PageRank durch BadRank, erforderlich. Auf diese Weise sollten alle unerwünschten Effekte vermieden werden können.

Kritische Beurteilung von BadRank und PR0

Wie die Kombination von PageRank und BadRank tatsächlich erfolgt, ist eher von nachrangiger Bedeutung. Eine getrennte Berechnung und anschließende Kombination von beiden hat allerdings zur Folge, dass man gegebenenfalls nicht am Toolbar PageRank messen kann, wie hoch tatsächlich der BadRank einer Seite ist. Denn falls eine Seite einen hohen PageRank im ursprünglichen Sinne hat, muss der Einfluss des BadRank nicht unbedingt ersichtlich sein. Verlinkt eine andere Seite darauf, kann dies jedoch durchaus gravierende Folgen haben.

Die weitaus größere Problematik liegt in der hier präsentierten, direkten Umkehrung des PageRank-Algorithmus: Genauso, wie ein zusätzlicher eingehender Link einer Seite deren PageRank immer nur erhöhen kann, kann ein zusätzlicher ausgehender Link einer Seite auch deren BadRank immer nur erhöhen. Dies liegt darin begründet, dass im Rahmen der BadRank-Berechnung sich die übertragenen Werte einfach aufaddieren. Somit ist es vollkommen gleich, auf wie viele untadelige Sites eine Seite verlinkt - ein einziger Link auf eine Spam-Site kann gegebenenfalls ausreichen, um zu einem PR0 zu führen.

Diese Problematik stellt sich allerdings wohl nur in Ausnahmefällen. Da sich schließlich bei einer direkten Umkehrung des PageRank-Algorithmus der BadRank einer Seite unter deren eingehenden Links aufteilt, wird bei einzelnen Links auf Seiten mit hohem BadRank immer nur jeweils ein Bruchteil des BadRank übertragen. Google's Matt Cutts sagt hierzu: "If someone accidentally does a link to a bad site, that may not hurt them, but if they do twenty, that's a problem." (searchenginewatch.com/sereport/02/11-searchking.html)

Solange jedoch alle Links im Rahmen des BadRank gleichermaßen gewertet werden, besteht dennoch auch bei einzelnen Links ein Problem. Haben schließlich zwei Seiten einen sehr unterschiedlich hohen PageRank und verlinken auf die gleiche Seite mit hohem BadRank, kann es nach Art und Weise der Kombination von PageRank und BadRank dazu kommen, dass die Seite mit dem höheren PageRank weit weniger unter dem auf sie übertragenen BadRank leidet als diejenige Seite mit dem niedrigeren PageRank. Wir können allerdings zuversichtlich sein, dass Google mit derartigen Problemen umzugehen weiß. Nichtsdestotrotz soll nochmals angemerkt werden, dass ausgehende Links im Rahmen der hier beschriebenen Verfahren immer nur schaden können.

Dass die hier vorgestellten Verfahren tatsächlich auch dieser Form eingesetzt werden, ist natürlich reine Spekulation. Grundsätzlich sollte jedoch die Bewertung von Linkstrukturen in Analogie zum PageRank-Verfahren genau die Art und Weise sein, wie nur Google mit Spam umzugehen versteht.