

Softwaresimulation der Google-Suchmaschinentechnologie

Google - Pagerank

Von www.getstyle.net

Dresden, 01. Oktober 2005

I. INHALTSVERZEICHNIS

	Seite
I. INHALTSVERZEICHNIS	1
II. ABBILDUNGSVERZEICHNIS	3
III. TABELLENVERZEICHNIS	3
1 VORWORT	4
2 EINLEITUNG	5
2.1 SUCHMASCHINEN	6
2.2 BOOLESCHE SUCHMASCHINEN	6
2.3 SUCHMASCHINEN MIT RANGORDNUNGSVERFAHREN	7
2.4 METASUCHMASCHINEN	8
2.5 LERNENDE SUCHMASCHINEN	8
3 GOOGLE	9
3.1 KONZEPT DES PAGERANK	10
3.2 PAGERANK-ALGORITHMUS	11
3.3 ALTERNATIVE DEFINITION DES PAGERANK-ALGORITHMUS	12
3.4 RANDOM SURFER MODEL	13
3.5 FUNKTIONSWEISE DES PAGERANK	14
3.6 DIE ITERATIVE BERECHNUNG DES PAGERANK	15
3.7 PAGERANK ALS MATRIZE	17
3.8 HAUPTINFLÜSSE AUF DEN PAGERANK	19
3.8.1 <i>Ausgehende Kanten</i>	19
3.8.1.1 Wirkung ausgehender Kanten	19
3.8.1.2 Rank Sinks und Rank Source	21
3.8.1.3 Dangling Links	22
3.8.1.4 Konzentration ausgehender Kanten.....	24
3.8.1.5 Wirkung auf den Pagerank durch zusätzliche Dokumente.	27
3.8.2 <i>Eingehende Kanten</i>	30
3.8.2.1 Wirkung eingehender Kanten	30
3.8.2.2 Warum der Dämpfungsfaktor?	33

3.9	ERWEITERUNG DES PAGERANK-VERFAHRENS	36
3.9.1	<i>Differenzierte Beurteilung von Kanten</i>	36
3.9.1.1	Position der Kanten.....	36
3.9.1.2	Aktualität der Dokumente.....	36
3.9.1.3	Geografischer Abstand zwischen Dokumenten	36
3.9.2	<i>Themenbasierter Pagerank</i>	38
3.9.2.1	Intelligenter Surfer von Richardson und Domingos.....	38
3.9.2.2	Topic Sensitiv Pagerank nach Taher Haveliwala	39
4	ZUSAMMENFASSUNG	41
5	GLOSSAR	42
IV.	INTERNETLINK-VERZEICHNIS	43
V.	QUELLENVERZEICHNIS	44

II. ABBILDUNGSVERZEICHNIS

	Seite
ABBILDUNG 3.5-1 FUNKTIONSWEISE DES PAGERANK (BEISPIEL 1).....	14
ABBILDUNG 3.8.1-1 WIRKUNG AUSGEHENDER KANTEN (BEISPIEL 2)	19
ABBILDUNG 3.8.1-2 RANK SINKS (BEISPIEL 3)	21
ABBILDUNG 3.8.1-3 DANGLING LINKS	23
ABBILDUNG 3.8.1-4 KONZENTRATION AUSGEHENDER KANTEN (BEISPIEL 4)	24
ABBILDUNG 3.8.1-5 KONZENTRATION AUSGEHENDER KANTEN (BEISPIEL 5)	25
ABBILDUNG 3.8.1-6 EINFÜGEN ZUSÄTZLICHER DOKUMENTE (BEISPIEL 6).....	27
ABBILDUNG 3.8.1-7 EINFÜGEN ZUSÄTZLICHER DOKUMENTE (BEISPIEL 7).....	28
ABBILDUNG 3.8.2-1 WIRKUNG EINGEHENDER KANTEN (BEISPIEL 8).....	30
ABBILDUNG 3.8.2-2 WIRKUNG EINGEHENDER KANTEN (BEISPIEL 9).....	31
ABBILDUNG 3.8.2-3 DÄMPFUNGSFAKTOR (BEISPIEL 9)	33

III. TABELLENVERZEICHNIS

TABELLE 3.6-1 ITERATIVE BERECHNUNG	16
TABELLE 3.8.2-4 DÄMPFUNGSFAKTOR	35

1 Vorwort

Als angehender Diplomingenieur für Medieninformatik und als jemand, der das Internet fast täglich nutzt, hat es mich bisher schon immer interessiert, wie Suchmaschinen arbeiten. Aus diesem Grund war ich positiv überrascht, als das Seminar „Information-Retrieval“ bzw. Informationswiedergewinnung oder Informationsbeschaffung als Wahlpflichtfach an der Fachhochschule Lausitz angeboten wurde. In diesem Seminar weihte mich Herr Professor Dr.-Ing. habil. Horst Zuse in die Techniken der Suchmaschinen ein. Am Ende des Seminars bot der Herr Professor eine Wissenschaftliche Arbeit zum Thema „Softwaresimulation der Google-Suchmaschinentechologie“ den Studenten an. Um mein im Seminar erlangtes Wissen weiter auszubauen, nahm ich die Gelegenheit wahr, und bot mich für die Wissenschaftliche Arbeit an. Mit dieser äußerst interessanten Arbeit kann ich die Technik von Google auf spielerische Weise anderen Studierenden nahe bringen.

2 Einleitung

Ziel dieser Arbeit ist es, ein Programm zu erstellen, das die Pagerank-Technologie von Google simuliert. Anhand dieser Simulation soll die Problematik der Sortierung der gefundenen Dokumente (Webseiten) in Google veranschaulicht werden. Diese Sortierung basiert im Wesentlichen auf der Verlinkung der einzelnen Dokumente des World Wide Web.

Das Herzstück der wohl beliebtesten Suchmaschine Google ist Pagerank™, ein Verfahren zur Beurteilung von Dokumenten. Dieser Algorithmus wurde von Larry Page und Sergey Brin an der University of Stanford entwickelt. Er teilt allen Dokumenten des World Wide Web eine „Wichtigkeit“ zu, die unter anderem dazu dient, die Suchergebnisse zu sortieren. Um diese Wichtigkeit zu berechnen, wird momentan einmal im Monat das zurzeit wohl größte Eigenwertproblem der Welt gelöst. Hierbei wird der Eigenvektor π_* einer Matrix $P \in \mathbb{R}^{n \times n}$ zum größten Eigenwert dieser Matrix berechnet.

Um die verschiedenen Beispiele nachvollziehen zu können, besteht die Möglichkeit, die mit „Beispiel“ gekennzeichneten Abbildungen in meinem Pagerank-Calculator zu testen. Dieser Pagerank-Calculator ist unter der Adresse <http://www.getstyle.net/pagerank-calculator/> zu finden. Auf dieser Homepage ist auch eine ausführliche Hilfe zu dem Pagerank-Calculator vorhanden.

2.1 Suchmaschinen

Unter dem Begriff Suchmaschinen fallen verschiedene Arten von Suchmaschinen. Die vier Grundarten sind Web-, Meta-, Desktop- und Echtzeitsuchmaschinen. In dieser Arbeit wird nur auf die Websuchmaschinen eingegangen, die anderen Typen werden daher nicht näher betrachtet. Die Arbeit einer Websuchmaschine beginnt mit dem so genannten Spidern des World Wide Web. Hierbei durchkämmen die Spider jedes einzelne Dokument nach seinen Inhalten und analysieren zum Beispiel Überschriften, Texte und Tabellen. Die gefundenen Inhalte werden anschließend in einer Datenbank gespeichert. Danach hangeln sich die Spider über die ausgehenden Kanten (Links) der Dokumente weiter zu anderen Dokumenten. Somit wird der gesamte Inhalt des Web in einer Datenbank abgelegt. Dieser wird ständig erneuert, da sich die Inhalte der Dokumente in vielen Fällen auch ständig ändern. Die Datenbank wird dazu benutzt, die Suchanfragen der Suchmaschinenbenutzer mit dem Datenbankinhalt abzugleichen. Somit löst jede Suchanfrage eine Suche im Bestand der Datenbank der Suchmaschine aus. Diese Suche wird dann in Form einer Trefferliste auf dem Bildschirm ausgegeben.

2.2 Boolesche Suchmaschinen

Die relativ einfache Implementierung und Umsetzung macht die Boolesche Suchmaschine zu einer der einfachsten in dieser Hinsicht. Eine sehr bekannte Boolesche Suchmaschine ist Lycos. Die meist sehr komplizierten Suchanfragen bestehen aus booleschen Operatoren (UND, ODER), die in vielen Möglichkeiten miteinander kombiniert werden können. Umso mehr Begriffe mit UND bzw. ODER verbunden werden, desto genauer wird das Ergebnis. Aber es kann durchaus passieren, dass man überhaupt kein Ergebnis mehr bekommt, wenn man zu viele Begriffe verbindet. Ein weiterer großer Nachteil ist, dass es bei der Ergebnisdarstellung keine Rangliste gibt. Das bedeutet, dass die gefundenen Dokumente einfach so wie sie gefunden wurden, also ungeordnet, aufgelistet werden. Die Qualität, das heißt die

absteigende Sortierung nach Relevanz, der Suchergebnisse lässt ebenfalls oft zu wünschen übrig. In experimentellen Untersuchungen hat das boolesche Verfahren gegenüber anderen Retrievalverfahren mit Abstand die schlechtesten Ergebnisse erbracht. Durch die fehlende Benutzerfreundlichkeit und die komplizierte Anfragesprache ist eine Boolesche Suchmaschine denkbar ungeeignet für den Endanwender. Mit einer Kombination aus Boolescher Suchmaschine und einer Ähnlichkeits-Suchmaschine kann man aus der ungeordneten Liste eine Liste erstellen, die durch Ähnlichkeiten absteigend mit dem Suchbegriff geordnet wird.

2.3 Suchmaschinen mit Rangordnungsverfahren

Suchmaschinen mit Rangordnungsverfahren sind in der Benutzerfreundlichkeit und der Anfragesprache den Booleschen Suchmaschinen weit überlegen. Bei diesen Suchmaschinen gibt der Benutzer einfach seine Anfrage in Form einer Menge von Suchwörtern ein. In der einfachsten Form dieser Suchmaschine werden die in der Suchanfrage vorkommenden Wörter im Dokument gesucht. Anschließend wird die Anzahl des Auftretens der Suchworte gezählt, und nachfolgend die gefundenen Dokumente nach der abnehmenden Anzahl der Treffer geordnet ausgegeben. Dieses absteigende Sortieren nennt man „Ranking“. Durch dieses Ranking stehen die für die Suchanfrage relevanten Dokumente vermehrt am Anfang der Trefferliste. Das Problem der Booleschen Suchmaschine, dass die Anzahl der Treffer sehr begrenzt sein kann, gibt es bei Rangordnungsverfahren so nicht. Da die Ergebnisse nach Relevanz geordnet sind, kann der Benutzer die Rangordnung so weit durchgehen, wie er es für nötig hält. Beispiele für Suchmaschinen, die unter anderem auch das Rangordnungsverfahren nutzen, sind MSN und Google.

2.4 Metasuchmaschinen

Metasuchmaschinen sind solche, deren wesentliches Merkmal darin besteht, dass eine Suchanfrage an andere Suchmaschinen gleichzeitig weitergeleitet, die Ergebnisse gesammelt und aufbereitet werden. Das bedeutet, dass Metasuchmaschinen nicht selbst im World Wide Web suchen, sondern andere für sich suchen lassen. Die Aufbereitung der Treffer geschieht meist durch die Entfernung von doppelten Einträgen bei Überschneidung der Suchmaschinenergebnisse, durch die Bewertung der Ergebnisse und durch Aufstellung eines eigenen internen Rankings. Die Ergebnisse werden dann einheitlich, wie in anderen Suchmaschinen, dargestellt. Klassische Metasuchmaschinen sind zum Beispiel MetaCrawler und Metaspinner.

2.5 Lernende Suchmaschinen

Lernen bedeutet an dieser Stelle, dass die Suchmaschine die Suchstrategie individuell, basierend auf den Wünschen der Nutzer, dynamisch ändern kann und so individuell angepasste Suchergebnisse liefert. Die lernende Suchmaschine verarbeitet das Nutzerurteil (favorisierte oder abgelehnte Suchergebnisse) und generiert daraus einen neuen Vorschlag an Suchergebnissen. Eine solche Technologie wird erstmals in der Bildersuchmaschine Foto-Scout-Zuse verwendet. [HZ05]

3 Google

Im Verlauf der Internetgeschichte stellt eine Suchmaschine alle anderen in den Schatten: Google. Dieser Suchmaschine gelang es als erste, den Benutzern eine Liste mit den für sie relevanten Treffern herabsteigend sortiert auszugeben. Das bedeutet, dass die für den Benutzer anscheinend wichtigsten Seiten an den ersten zehn Positionen in der Suchmaschine erschienen. Die teilweise weit überlegende Qualität der Suchergebnisse, der Performance sowie der Benutzerfreundlichkeit der Suchmaschine sorgten dafür, dass Google zu einer der beliebtesten Suchmaschinen bis heute wurde. Der Erfolg dieser Suchmaschine beruht ganz essentiell auf dem Pagerank-Verfahren. Das Pagerank-Verfahren hat seinen Namen nicht etwa, weil es den Rank von Webseiten ermittelt, sondern weil es erstmalig in einer Arbeit von Lawrence Page und Sergey Brin vorgestellt wurde [PB98].

Das Bewertungsverfahren Pagerank wurde im Jahr 1998 von den zwei Studenten Sergey Brin und Lawrence Page an der Stanford University entwickelt. Obwohl seit der Veröffentlichung des Papiers „The Anatomy of a Large-Scale Hypertextual Web Search Engine“ [PB98] von Brin und Page einige Zeit vergangen ist, ist trotz zahlreicher Änderungen, Anpassungen und Modifikationen der Ansatz des ursprünglichen Pagerank-Algorithmus erhalten geblieben. Der jetzige Pagerank-Algorithmus ist ein von Google gut gehütetes Geheimnis, weshalb zu dem heutigen Pagerank-Algorithmus nur Vermutungen durch Beobachtung geäußert werden können. Da das Konzept des Pagerank-Algorithmus noch immer ein grundlegender Baustein sein dürfte, wird sich in dieser Arbeit darauf bezogen.

3.1 Konzept des Pagerank

Angenommen, ein Dokument besitzt zwei eingehende Kanten, auch Backlinks genannt, einmal von „www.pauls-private-homepage.de“ und einmal von „www.microsoft.de“. Bei diesen zwei Kanten sollte klar werden, dass es eine unterschiedliche Bewertung von Kanten geben muss. Die Kante von Microsoft wird als wichtiger angesehen, da diese viele eingehende Kanten besitzt. Im Gegensatz zum einfachen Konzept der Link-Popularität, welches nur die Anzahl der eingehenden Kanten in Betracht zieht, wird beim Pagerank-Konzept die Bedeutsamkeit eines jeden Dokumentes beachtet.

Page und Brin argumentieren, dass ein Dokument zwar bedeutsam ist, wenn es von vielen Dokumenten verlinkt wird, jedoch ist nicht jedes Dokument gleichwertig zu jedem anderen. Jedes Dokument sollte eine hohe Bedeutsamkeit erhalten, wenn es von anderen bedeutsamen Dokumenten verlinkt wird, zunächst unabhängig davon, welche Inhalte es hat.

Für jede Seite in der Datenbank von Google, welche eine möglichst große Teilmenge des gesamten World Wide Web repräsentieren sollte, wird aus der Struktur ihrer Referenzen (Backlinks) iterativ ein Bedeutsamkeitswert berechnet. [PA98]

Der Pagerank-Algorithmus basiert auf folgenden Annahmen:

- Jeder Autor eines Dokumentes macht implizit eine Aussage über seine subjektive hohe Meinung von Dokumenten, auf die er durch eine Kante verweist.
- Die Gesamtheit der subjektiven Meinungen aller Autoren kann als objektive Einschätzung eines Dokumentes gewertet werden.
- Je mehr Backlinks ein Dokument bekommt, desto bedeutender scheint dieses zu sein.
- Je weniger Kanten ein Dokument erhält, desto bedeutender ist jeder einzelne Backlink.

- Je bedeutender ein Dokument ist, desto bedeutsamer sind auch dessen ausgehende Kanten.
- Je bedeutender die Backlinks sind, die ein Dokument bekommt, desto bedeutender scheint das Dokument selbst zu sein.

3.2 Pagerank-Algorithmus

Auf den oben genannten Annahmen basierend, entwickelten Page und Brin den folgenden Algorithmus zur Berechnung eines Bewertungsindex (Pagerank).

$$\mathbf{PR(A) = (1-d) + d (PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n))}$$

Hierbei ist:

- $PR(A)$ der Pagerank (Bewertungsindex) eines Dokumentes A ,
- $PR(T_i)$ der Pagerank der Dokumente T_i , von denen eine Kante auf das Dokument A zeigt,
- $C(T_i)$ die Gesamtanzahl der Kanten auf Seite T_i und
- d ein Dämpfungsfaktor (Damping Factor), wobei $0 \leq d \leq 1$ ist.

Verbal lässt sich der Algorithmus wie folgt beschreiben:

1. Jedes Dokument des World Wide Web wird mit einem Startwert initialisiert. Theoretisch kann der Startwert beliebig gewählt werden, da der Algorithmus immer konvergiert. Jedoch hat die Wahl des Startwertes einen wesentlichen Einfluss darauf, wie schnell eine akzeptable Konvergenz erreicht wird.
2. Der Pagerank eines Dokumentes wird berechnet, indem man den Pagerank der Dokumente der ausgehenden Kanten nimmt und diesen durch die jeweilige Anzahl der ausgehenden Kanten teilt.
3. Aus dem Pagerank der eingehenden Kanten (Backlinks) wird der Pagerank neu berechnet.
4. Dieses Verfahren wird ab Punkt 2 so oft wiederholt, bis der Pagerank aller Dokumente konvergiert bzw. bis eine hinreichende Annäherung erreicht ist. In ihrer Veröffentlichung "The Anatomy of a Large-Scale

"Hypertextual Web Search Engine" zeigen Page und Brin, dass 100 Iterationen für eine hinreichende Annäherung ausreichend sind.

Wie man hier sieht, bewertet das Pagerank-Verfahren ein Dokument nicht anhand seines Inhaltes, seiner Größe oder seines Auftretens, sondern lediglich nach der Linkstruktur des World Wide Web, mit der die einzelnen Dokumente miteinander verbunden sind.

3.3 Alternative Definition des Pagerank-Algorithmus

In ihren Veröffentlichungen gehen Brin und Page auch auf eine von ihnen etwas abgewandelte Form des Pagerank-Algorithmus ein. In dieser Definition wird der Pagerank folgendermaßen bestimmt:

$$\mathbf{PR(A) = (1-d)/N + d (PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n))}$$

Um die gesamte Anzahl der Dokumente im World Wide Web mit in die Berechnung des Pagerank einzubeziehen, wurde hier eine weitere Variable N hinzugefügt, die die Anzahl der Dokumente repräsentiert. Diese zweite Variante des Pagerank-Algorithmus unterscheidet sich aber nicht grundlegend von der ersten. Durch das Hinzufügen der Anzahl der Dokumente kann die zweite Version vielmehr als eine Definition des Pagerank gesehen werden, die die tatsächliche Wahrscheinlichkeit beschreibt, mit der ein anderes Dokument erreicht wird. Somit kann diese Definition als Wahrscheinlichkeitsverteilung über alle Dokumente des World Wide Web angesehen werden, wobei die Summe aller Pagerank 1 ergibt. In der ersten Version dagegen stellt der Pagerank einen Erwartungswert für das Aufrufen eines Dokumentes durch den Zufall-Surfer dar.

3.4 Random Surfer Model

Sergey Brin und Lawrence Page bieten in ihren Veröffentlichungen eine sehr einfache, instinktive Rechtfertigung des Pagerank-Algorithmus an. Sie sehen den Pagerank als ein Model, das das Verhalten eines Websurfers simuliert. Das Random Surfer Model verfolgt das Prinzip, dass ein Benutzer auf einem zufällig gewählten Dokument zu surfen beginnt, und ohne Rücksicht auf irgendwelche Inhalte weitere zufällig ausgewählte Kanten verfolgt, um eine anderes Dokument zu erreichen. Nach einer beliebigen Zeit verlässt der „Zufall-Surfer“ das Dokument, und sucht ein völlig neues Dokument auf. Die Wahrscheinlichkeit, mit der der „Zufall-Surfer“ sich auf einem bestimmten Dokument aufhält, entspricht genau dessen Pagerank. Das bedeutet, dass der Pagerank als Maß für die Wahrscheinlichkeit, wie oft das Random Surfer Model ein Dokument besucht, bezogen auf die komplette Anzahl der in Google registrierten Dokumente, verstanden werden kann. Angenommen, man setzt jetzt die Anzahl der Dokumente und die Anzahl der Durchläufe auf 100, und eine bestimmte Webseite hat einen Pagerank von 4, das bedeutet, dass der „Zufall-Surfer“ bei 100 Durchläufen diese Dokumente im Mittel 4 Mal besucht.

Da nach dem Random Surfer Model der „Zufall-Surfer“ nicht unendlich vielen Kanten eines Dokumentes folgt und nach einer gewissen Zeit „gelangweilt“ wird, und sich aus diesem Grund ein beliebig anderes Dokument sucht, wird die Wahrscheinlichkeit, mit der der „Zufall-Surfer“ ein neues Dokument aufsucht, um den Faktor d gedämpft. Dieser Dämpfungsfaktor d ist auch der Grund, warum der Pagerank nicht komplett an ein Dokument weitergegeben wird, sondern sich auf die ausgehenden Kanten verteilt. Dieser Faktor kann zwischen 0 und 1 liegen. Je größer der Faktor ist, desto wahrscheinlicher ist es, dass der „Zufall-Surfer“ die Kanten weiter verfolgt und das Surfen nicht abbricht. In der Praxis wird oft ein Wert von 0,85 als Dämpfungsfaktor d als ein guter Kompromiss zwischen Linkverfolgung und Abbruch des Surfvorganges angesehen.

3.5 Funktionsweise des Pagerank

Die Funktionsweise des Pagerank soll nun anhand des in Abbildung 3.5.1-1 gezeigten Beispiels veranschaulicht werden. Das kleine Miniweb besteht aus drei Dokumenten, wobei Dokument A auf Dokument B und Dokument C verlinkt, Dokument B auf Dokument C verlinkt und Dokument C wiederum auf Dokument A verlinkt.

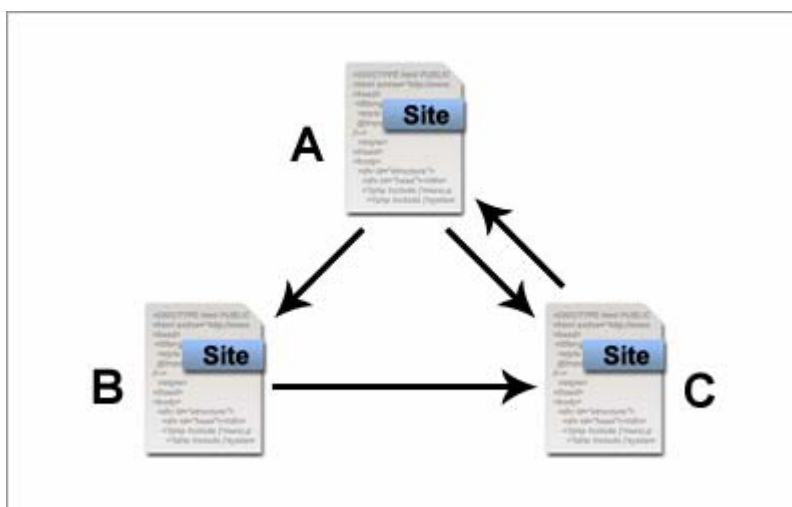


ABBILDUNG 3.5-1 FUNKTIONSWEISE DES PAGERANK (BEISPIEL 1)

Als Dämpfungsfaktor d wird hier der von Lawrence Page und Sergey Brin in den Veröffentlichungen genannte Wert von 0,85 genommen, um ein möglichst realitätsnahes Ergebnis zu bekommen. Nach dem Einsetzen in den Pagerank-Algorithmus ergeben sich folgende drei Gleichungen für den Pagerank der einzelnen Webseiten:

$$PR(A) = 0,15 + 0,85 * PR(C)$$

$$PR(B) = 0,15 + 0,85 * (PR(A) / 2)$$

$$PR(C) = 0,15 + 0,85 * (PR(B) + PR(A) / 2)$$

Dieses Gleichungssystem lässt sich sehr einfach lösen und es ergeben sich folgende Werte für die einzelnen Dokumente:

$$PR(A) \approx 1,16$$

$$PR(B) \approx 0,64$$

$$PR(C) \approx 1,19$$

In diesem Beispiel zeigt sich, dass das Dokument C, da es die meisten eingehenden Kanten hat, das scheinbar bedeutendste Dokument der drei ist. Aus Sicht der Wahrscheinlichkeit kommt die Suchmaschine mit der Wahrscheinlichkeit von 1,19 bei drei Anläufen in dem gezeigten Miniweb auf das Dokument C.

Für dieses kleine Miniweb lässt sich das Gleichungssystem noch relativ schnell und unproblematisch lösen. Da das eigentliche World Wide Web jedoch aus mehreren Milliarden Dokumenten besteht, lässt sich ein solches Gleichungssystem nicht mehr in akzeptabler Zeit lösen.

3.6 Die iterative Berechnung des Pagerank

Google sieht sich auf Grund der Größe des World Wide Web in der Praxis gezwungen, ein iteratives Verfahren zu verwenden. Hierbei kommt das Iterationsverfahren von Von-Mises zum Einsatz [VM03]. Das bedeutet, dass das Ergebnis näherungsweise bestimmt wird. Zunächst wird allen von Google erfassten Dokumenten ein Anfangswert für den Pagerank von 1 zugewiesen. Die Höhe des Anfangswertes hat aber keinen Einfluss auf das Ergebnis, da dieses irgendwann konvergiert. Aber wie schnell, nach wie vielen Iterationen, es konvergiert, kann durchaus durch eine gute Wahl eines Startwertes beeinflusst werden. Nach der Zuweisung der Startwerte für die Pagerank wird der Pagerank aller Dokumente in mehreren Berechnungsrunden ermittelt. In jedem Iterationsschritt wird der jeweils zuvor berechnete Näherungswert verwendet, bis das Ergebnis konvergiert. In dem nun folgenden Beispiel soll ein solches Näherungsverfahren demonstriert werden. Dabei wird wieder auf das Miniweb aus der Abbildung 3.5-1 zurückgegriffen.

Bei Startwert für Pagerank von 1			
Iteration	PR(A)	PR(B)	PR(C)
0	1	1	1
1	1	0.575	1.425
2	1.36125	0.575	1.06375
3	1.0541875	0.72853125	1.21728125
4	1.1846890625	0.5980296875	1.21728125
5	1.1846890625	0.6534928516	1.1618180859
6	1.137545373	0.6534928516	1.2089617754
7	1.1776175091	0.6334567835	1.1889257074
8	1.1605868513	0.6504874414	1.1889257074
9	1.1605868513	0.6432494118	1.1961637369
10	1.1667391764	0.6432494118	1.1900114118
11	1.1615097	0.64586415	1.19262615
12	1.1637322275	0.6436416225	1.19262615
13	1.1637322275	0.6445861967	1.1916815758
14	1.1629293394	0.6445861967	1.1924844639
15	1.1636117943	0.6442449693	1.1921432364
16	1.163321751	0.6445350126	1.1921432364
17	1.163321751	0.6444117442	1.1922665049
18	1.1634265291	0.6444117442	1.1921617267
19	1.1633374677	0.6444562749	1.1922062574
20	1.1633753188	0.6444184238	1.1922062574

TABELLE 3.6-1 ITERATIVE BERECHNUNG

Aus der Tabelle 3.6-1 ist ersichtlich, dass eine sehr gute Näherung an die tatsächlichen Werte schon nach wenigen Iterationen erreicht wird. Aber schon bei einem Startwert von 10 braucht der Algorithmus ca. 80 Durchläufe für das Miniweb, bis er konvergiert. Daher wird als Startwert oftmals 1 genommen. Für die Berechnung des Pagerank für das komplette World Wide Web werden in den Veröffentlichungen von Lawrence Page und Sergey Brin ca. 100 Iterationen als hinreichend genannt.

3.7 Pagerank als Matrize

Das Gleichungssystem des Pagerank-Algorithmus aus Abschnitt 3.3 kann auch in Matrixform dargestellt werden. Zuerst werden alle Dokumente in beliebiger Reihenfolge durchnummeriert mit x_1, x_2, \dots, x_n . Anschließend wird eine $n \times n$ -Matrix A definiert, die erst einmal ganz allgemein angibt, welches Dokument mit welchem verlinkt ist. Diese Matrix wird auch Linkmatrix genannt. Nun wird jedes einzelne Dokument durchgegangen und geprüft, ob es verlinkt wurde oder nicht. Wenn es verlinkt wurde, wird der Wert mit $1/C(x_j)$ berechnet, ansonsten wird der Wert dieser Position in der Matrix auf 0 gesetzt.

$$A_{jk} = \begin{cases} 1/C(x_j), & \text{wenn } x_j \text{ auf } x_k \text{ verweist} \\ 0 & \text{sonst.} \end{cases}$$

Alle Diagonalelemente dieser Linkmatrix müssen 0 sein, da ein Dokument nicht auf sich selbst zeigen darf. Außerdem sind alle Werte, die von 0 verschieden sind, in einer Zeile gleich, da die Summe der Werte in jeder von 0 verschiedenen Zeile 1 sein muss. Für das in Abbildung 3.5.1-1 gezeigte Beispiel sieht die Linkmatrix folgendermaßen aus:

$$A = \begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Anschließend werden die Werte zu einem n -dimensionalen Zeilenvektor D , $R = (R(x_1), R(x_2), \dots, R(x_n))$ zusammengeführt. Alle Komponenten des Zeilenvektors sind gleich $1-d$, wobei d den Dämpfungsfaktor darstellt. Das Gleichungssystem des Pagerank-Algorithmus aus Punkt 3.3 kann dann in der Form

$$R = D + dRA$$

dargestellt werden. Für die 3×3-Matrix aus dem oben genannten Beispiel würde das Gleichungssystem wie folgt aussehen:

$$\begin{pmatrix} R(x1) \\ R(x2) \\ R(x3) \end{pmatrix} = \begin{pmatrix} R(1-d) \\ R(1-d) \\ R(1-d) \end{pmatrix} + d * \begin{pmatrix} R(x1) \\ R(x2) \\ R(x3) \end{pmatrix} \begin{pmatrix} 0 & 1/2 & 1/2 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$$

Nach dem Umstellen von $\mathbf{R} = \mathbf{D} + d\mathbf{R}\mathbf{A}$ erhält man folgende Gleichung, wobei \mathbf{I} die $n \times n$ -Einheitsmatrix ist.

$$\mathbf{R} (\mathbf{I} - d\mathbf{A}) = \mathbf{D}$$

Ein solches System besitzt dann eine eindeutige Lösung \mathbf{R} , wenn die Matrix $\mathbf{I} - d\mathbf{A}$ invertierbar ist. Das Ergebnis wird dann durch $\mathbf{R} = \mathbf{D}(\mathbf{I} - d\mathbf{A})^{-1}$ gegeben. Eine Matrix $\mathbf{I} - d\mathbf{A}$ ist immer dann invertierbar, wenn $0 < d < 1$ ist. Wenn man jetzt für den Dämpfungsfaktor d den üblichen Wert von 0,85 einsetzt, bekommt man als Lösung folgende Matrix:

$$\mathbf{R} = \begin{pmatrix} 1,16 \\ 0,64 \\ 1019 \end{pmatrix}$$

Diese Matrix gibt den Bewertungsindex (Pagerank) aller Seiten an. [FE03]

3.8 Haupteinflüsse auf den Pagerank

3.8.1 Ausgehende Kanten

3.8.1.1 Wirkung ausgehender Kanten

Wie oben beschrieben, bildet das Pagerank-Verfahren die Linkstruktur des gesamten World Wide Web ab. Aus diesem Grund sollte klar sein, dass sowohl eingehende als auch ausgehende Kanten wesentlichen Einfluss auf den Pagerank eines Dokumentes nehmen. Im Folgenden wird anhand eines Beispiels letzteres näher erläutert.

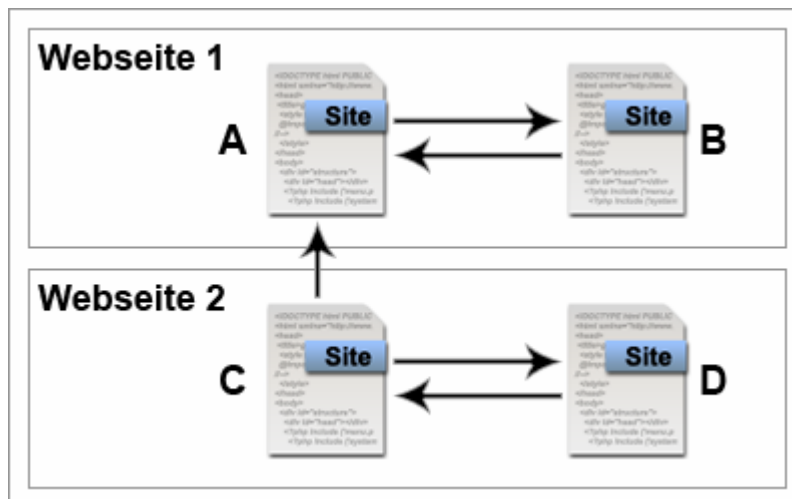


ABBILDUNG 3.8.1-1 WIRKUNG AUSGEHENDER KANTEN (BEISPIEL 2)

In diesem Beispiel wird, wie in Abbildung 3.8.1-1 gezeigt, das gesamte World Wide Web auf zwei Webseiten reduziert. Beide Webseiten bestehen jeweils aus zwei Dokumenten, die sich gegenseitig verlinken. Webseite 1 besteht aus Dokument A und B. Webseite 2 besteht aus Dokument C und D. Jedes Dokument startet mit einem Pagerank von 1. Jetzt wird dem Dokument C eine ausgehende Kante hinzugefügt. Bei einem angenommenen Dämpfungsfaktor d von 0,85 ergeben sich folgende Gleichungen für den Pagerank der einzelnen Dokumente:

$$PR(A) = 0,15 + 0,85 * (PR(C) / 2 + PR(B))$$

$$PR(B) = 0,15 + 0,85 * PR(A)$$

$$PR(C) = 0,15 + 0,85 * PR(D)$$

$$PR(D) = 0,15 + 0,85 * (PR(C) / 2)$$

Durch das Lösen des Gleichungssystems ergeben sich folgende Werte für die einzelnen Dokumente:

$$PR(A) \approx 1,66$$

$$PR(B) \approx 1,56$$

$$PR(C) \approx 0,43$$

$$PR(D) \approx 0,33$$

Es ergeben sich folgende Pagerank für beide Webseiten durch Aufsummieren des Pagerank der Webseiten:

$$PR(\text{Webseite 1}) \approx 3,22$$

$$PR(\text{Webseite 2}) \approx 0,76$$

Wie man sieht, entspricht die Summe aller Dokumente dem aufsummierten Pagerank aller Dokumente, in diesem Fall 4. Somit sollte auch klar werden, dass das Hinzufügen von Kanten keinen Einfluss auf den aufsummierten Pagerank des Web hat. Noch klarer sollte werden, dass der Gewinn an Pagerank des verlinkten Dokumentes genauso groß wie der Verlust des verlinkenden Dokumentes sein muss.

Wie man in dem oben gezeigten Beispiel gesehen hat, verliert das verlinkende Dokument deutlich an Pagerank. Der Verlust des Pagerank des verlinkenden Dokumentes ist auf das Verhalten des „Zufall-Surfer“ aus dem Random Surfer Model zurückzuführen. Durch das Hinzufügen einer externen Kante sinkt die Wahrscheinlichkeit, dass eine interne Kante verfolgt wird. Somit wird deutlich, dass jede weitere ausgehende Kante eines Dokumentes dessen Pagerank mindert. Wenn jetzt aber jeder auf die Idee kommen sollte, auf seiner Webseite keine ausgehenden Kanten zu

setzen, würde er die Grundlage des World Wide Web zerstören. Das World Wide Web lebt durch die Verlinkungen von Webmastern zu anderen Webseiten. Somit geben sie mit dem Setzen einer externen Kante ein repräsentatives Urteil zu dem anderen Dokument ab. Außerdem beziehen sie in gewisser Art und Weise den Inhalt des verlinkten Dokumentes in ihr Dokument ein.

3.8.1.2 Rank Sinks und Rank Source

Die vereinfachte Version des Pagerank-Verfahrens wirft ein Problem auf. Es können so genannte Rank Sinks auftreten. Diese entstehen, wenn zwei oder mehrere Webseiten so miteinander verlinkt sind, dass sie zu keinen anderen Dokumenten verweisen. Wie in Abbildung 3.8.1-2 zu sehen ist, handelt es sich hierbei um zyklisch verlinkte Dokumente mit einem Eingang aber ohne Ausgang. Bei der Berechnung des Pagerank dieser Dokumente wird dieser bei jeder Iteration weitergegeben. Dies ist ein ständiger Kreislauf, der dazu führt, dass der Pagerank der Webseiten stetig erhöht wird und nicht laut Definition an andere Seiten weiterverteilt wird, da keine ausgehenden Kanten bei dieser Schleife existieren.

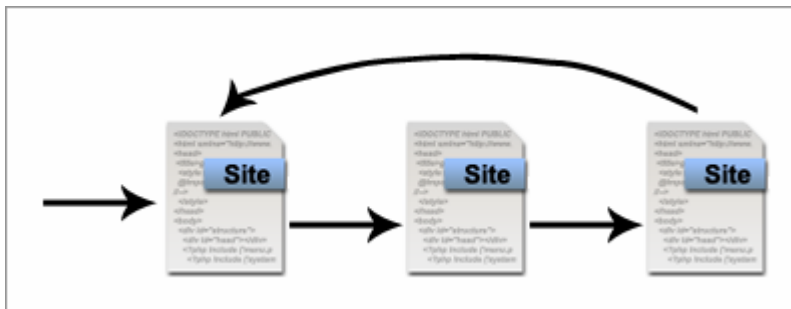


ABBILDUNG 3.8.1-2 RANK SINKS (BEISPIEL 3)

Daraus folgt, dass der Pagerank einerseits außerhalb der Schleife zu niedrig und innerhalb der Schleife viel zu hoch ist. Um den Verlust von Rank in Rank Sinks auszugleichen, führten Lawrence Page und Sergey Brin so genannte Rank Source ein. Zusätzlich zu dem Dämpfungsfaktor d , der der zu hohen Einstufung von Dokumenten entgegenwirkt, kommt noch ein Vektor E , den man als Summand sehen kann, hinzu. Eine Rank Source ist sozusagen ein „Bonus“, welcher jedem Dokument bei jeder Iteration zugeschrieben wird. Die Rank Source und der Dämpfungsfaktor d sind

immer so abgestimmt, dass die Summe des im System vorhandenen Pagerank konstant bleibt. In der Praxis hat sich ein Dämpfungsfaktor von 0,85 etabliert.

3.8.1.3 Dangling Links

Ein Problem, welches immer wieder auftritt, ob gewollt oder nicht gewollt, sind die so genannten „Dangling Links“. Diese Kanten sind Verweise, die auf Dokumente verweisen, die selbst keine ausgehenden Kanten besitzen. Das Model des Pagerank wird dadurch beeinflusst, da nicht festgelegt ist, wo der Pagerank weiterverteilt werden soll. In den meisten Fällen verweisen diese Dangling Links auf Dokumente, die noch nicht von der Suchmaschine erfasst worden sind. Es gibt mehrere Möglichkeiten, warum diese noch nicht erfasst worden sein könnten. Zum einen könnten dies schlecht für Suchmaschinen lesbare Dokumente sein, des Weiteren ist es so gut wie unmöglich, das gesamte World Wide Web zu erfassen, da es auf Grund der explosionsartig wachsenden Anzahl der Dokumente ungeahnte Maße annimmt. Zum anderen könnte es sein, dass ein Dokument-Ersteller mit Absicht die Suchmaschinen ausschließt, indem er mithilfe einer robots.txt Datei der Suchmaschine zu verstehen gibt, dass seine Dokumente nicht durchsucht werden sollen. Nach Übereinkunft des Robots Exclusion Standard-Protokolls wird in einer robots.txt Datei festgelegt, ob eine Suchmaschine eine bestimmte Webseite durchsuchen darf oder nicht [RO94]. Da Google aber auch andere Dokumententypen, z.B. PDF oder Word-Dokumente, durchsucht, welche meist keine ausgehenden Kanten beinhalten, ist es verständlich, dass es nicht negativ bewertet wird, keine ausgehenden Kanten zu haben.

Somit wird klar, dass Dangling Links den Pagerank nicht direkt beeinflussen, sondern einfach zur Berechnung aus dem Model entfernt werden, bis der Pagerank der Dokumente berechnet ist. Wie in Abbildung 3.8.1-3 gezeigt, ist die Entfernung von Dangling Links ein iterativer Vorgang, da das Entfernen erneut zu Dangling Kanten führen kann. Im Anschluss an die Berechnung des Pagerank wird den Dokumenten, die keine ausgehenden Kanten besitzen, der Pagerank der auf sie zeigenden Dokumente zugeschrieben.

Die Anzahl der Berechnungsschritte für den Pagerank entspricht hierbei der genauen Anzahl der Iterationen, die für die Entfernung der Dangling Links nötig waren. Durch die Entfernung solcher Dangling Links kann es jedoch passieren, dass sich der Pagerank auf die anderen ausgehenden Kanten anders verteilt. Diese Veränderung hat allerdings keinen großen Effekt, so dass sie außer Acht gelassen werden kann.

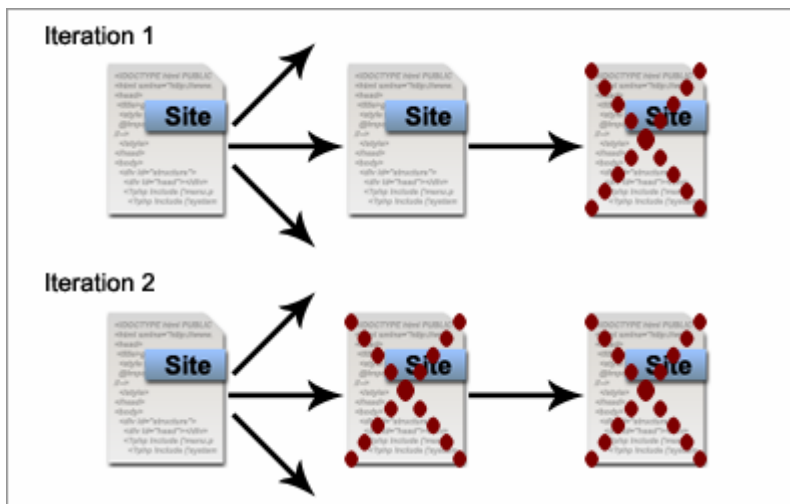


ABBILDUNG 3.8.1-3 DANGLING LINKS

3.8.1.4 Konzentration ausgehender Kanten

Wie in Punkt 3.8.1-1 gezeigt wurde, haben ausgehende Kanten eher negativen Einfluss auf den Pagerank eines Dokumentes und dessen Unterseiten. Dieser negative Einfluss kann aber minimiert werden, indem alle ausgehenden Kanten in einem Dokument zusammengefasst werden.

In Abbildung 3.8.1-4 wird ein Dokument mit drei Unterseiten simuliert, die folgendermaßen miteinander verlinkt sind. Dokument A zeigt auf alle Unterseiten und alle drei Unterseiten zeigen auf A zurück. Außerdem haben die Unterseiten B, C, D jeweils noch eine ausgehende Kante.

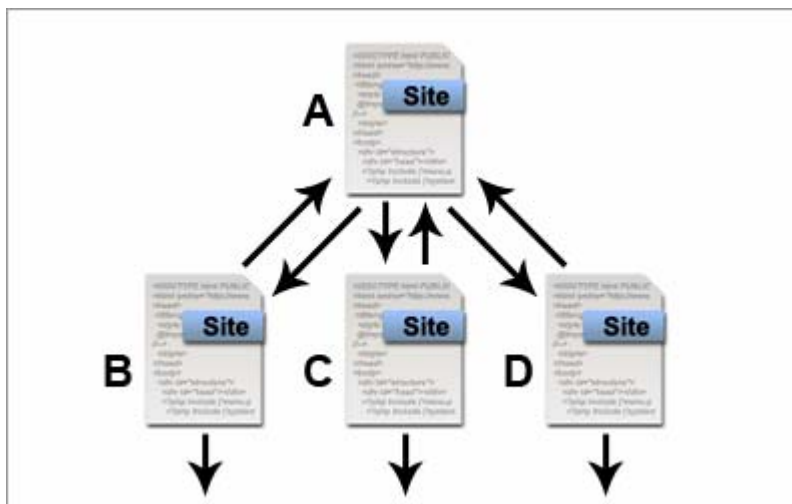


ABBILDUNG 3.8.1-4 KONZENTRATION AUSGEHENDER KANTEN (BEISPIEL 4)

Bei einem angenommenen Dämpfungsfaktor d in Höhe von 0,85 ergeben sich die folgenden Gleichungen für die Pagerank-Berechnung:

$$PR(A) = 0,15 + 0,85 * (PR(B) / 2 + PR(C) / 2 + PR(D) / 2)$$

$$PR(B) = 0,15 + 0,85 * (PR(A) / 3)$$

$$PR(C) = 0,15 + 0,85 * (PR(A) / 3)$$

$$PR(D) = 0,15 + 0,85 * (PR(A) / 3)$$

Durch das Lösen des Gleichungssystems ergeben sich folgende Werte für die einzelnen Dokumente:

$$PR(A) \approx 0,53$$

$$PR(B) \approx 0,31$$

$$PR(C) \approx 0,31$$

$$PR(D) \approx 0,31$$

In dem nun folgenden Beispiel wird das kleine Miniweb so modifiziert, dass die ausgehenden Kanten nun alle auf der Unterseite D konzentriert werden, und somit B und C keine ausgehenden Kanten mehr besitzen.

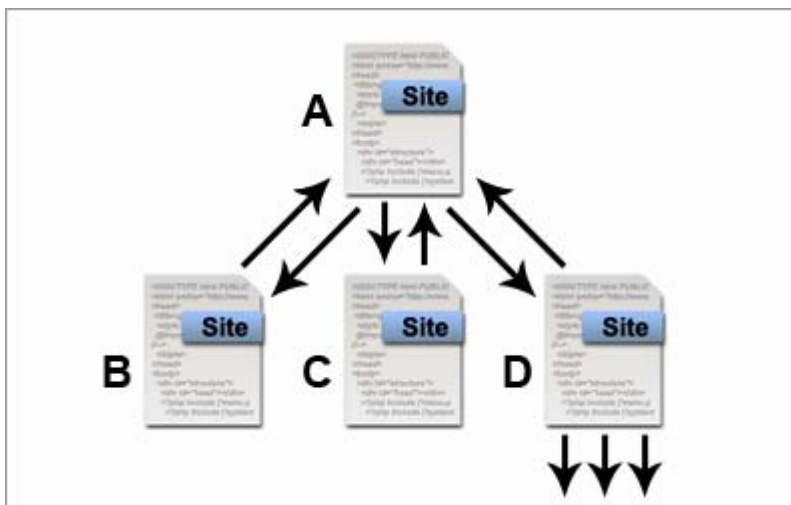


ABBILDUNG 3.8.1-5 KONZENTRATION AUSGEHENDER KANTEN (BEISPIEL 5)

Bei einem angenommenen Dämpfungsfaktor d in Höhe von 0,85 ergeben sich die folgenden Gleichungen für die Pagerank-Berechnung:

$$PR(A) = 0,15 + 0,85 * (PR(B) + PR(C) + PR(D) / 4)$$

$$PR(B) = 0,15 + 0,85 * (PR(A) / 3)$$

$$PR(C) = 0,15 + 0,85 * (PR(A) / 3)$$

$$PR(D) = 0,15 + 0,85 * (PR(A) / 3)$$

Durch das Lösen des Gleichungssystems ergeben sich folgende Werte für die einzelnen Dokumente:

$$\text{PR}(A) \approx 0,95$$

$$\text{PR}(B) \approx 0,42$$

$$\text{PR}(C) \approx 0,42$$

$$\text{PR}(D) \approx 0,42$$

Wie man in diesem Beispiel gesehen hat, bekommen bei der Konzentration der ausgehenden Kanten alle Dokumente einen höheren Pagerank. Des Weiteren macht es auch im Hinblick auf Suchmaschinenoptimierung Sinn, alle ausgehenden Kanten zu konzentrieren. Aber es sollte bei der Optimierung auch darauf geachtet werden, dass die Benutzerfreundlichkeit nicht zu kurz kommt.

3.8.1.5 Wirkung auf den Pagerank durch zusätzliche Dokumente

Da die Anzahl aller Dokumente des Web gleich deren aufaddierter Pagerank ist, folgt geradewegs, dass der aufaddierte Pagerank des Web bei jedem zusätzlichen Dokument um genau eins erhöht wird. Aber wie wirkt sich auf den Pagerank eines bestimmten Dokumentes und dessen Unterseiten aus, wenn eine weitere Unterseite hinzukommt? In dem nun folgenden Beispiel soll dieser Frage auf den Grund gegangen werden.

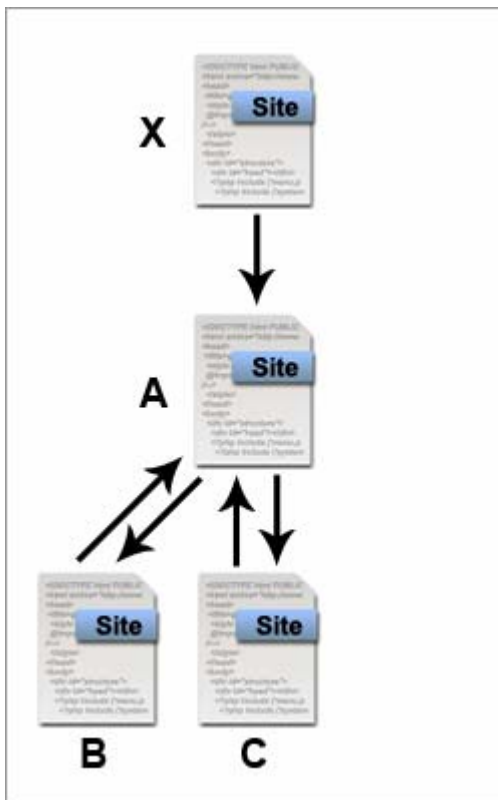


ABBILDUNG 3.8.1-6 EINFÜGEN ZUSÄTZLICHER DOKUMENTE (BEISPIEL 6)

Die in Abbildung 3.8.1-6 gezeigte kleine Webstruktur dient zur Veranschaulichung. Die Dokumente A, B und C gehören zu einer Webseite, wobei B und C hier die Unterseiten darstellen. Dokument X symbolisiert in diesem Beispiel eine externe Webseite, die mit einem Pagerank von 10 auf A verweist. Bei einem angenommenen Dämpfungsfaktor d in Höhe von 0,85 ergeben sich die folgenden Gleichungen für die Pagerank-Berechnung:

$$PR(A) = 0,15 + 0,85 * (10 + PR(B) + PR(C))$$

$$PR(B) = 0,15 + 0,85 * (PR(A) / 2)$$

$$PR(C) = 0,15 + 0,85 * (PR(A) / 2)$$

Durch das Lösen des Gleichungssystems ergeben sich folgende Werte für die einzelnen Dokumente:

$$PR(A) \approx 32,23$$

$$PR(B) \approx 13,85$$

$$PR(C) \approx 13,85$$

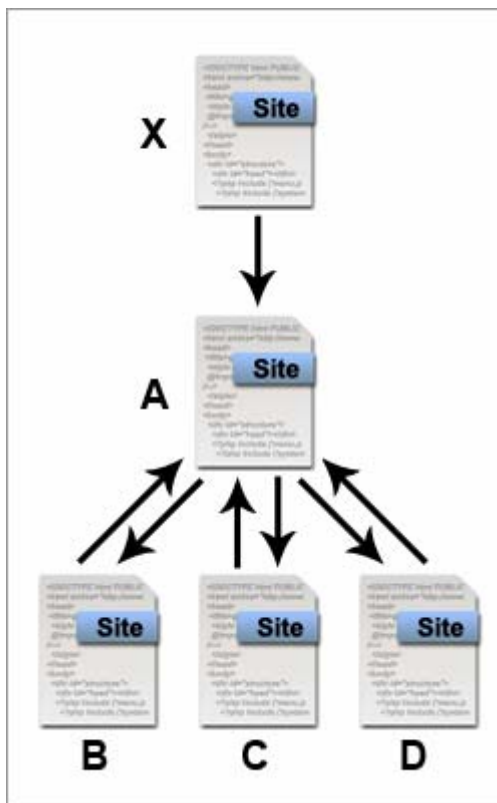


ABBILDUNG 3.8.1-7 EINFÜGEN ZUSÄTZLICHER DOKUMENTE (BEISPIEL 7)

Jetzt wird, wie in Abbildung 3.8.1-7 gezeigt, ein weiteres Dokument D hinzugefügt. Nach dem Hinzufügen von Dokument D lauten die Gleichungen für die Pagerank-Berechnung folgendermaßen:

$$PR(A) = 0,15 + 0,85 * (10 + PR(B) + PR(C) + PR(D))$$

$$PR(B) = 0,15 + 0,85 * (PR(A) / 3)$$

$$PR(C) = 0,15 + 0,85 * (PR(A) / 3)$$

$$PR(D) = 0,15 + 0,85 * (PR(A) / 3)$$

Durch das Lösen des Gleichungssystems ergeben sich folgende Werte für die einzelnen Dokumente:

$$PR(A) \approx 32,43$$

$$PR(B) \approx 9,35$$

$$PR(C) \approx 9,35$$

$$PR(D) \approx 9,35$$

Da unsere Beispielwebseite keine ausgehenden Kanten aufweist, steigt der aufaddierte Pagerank aller Dokumente nach dem Hinzufügen von Seite D erwartungsgemäß um genau 1 von 59 auf 60. Ferner steigt der Pagerank von Dokument A marginal an. Der Pagerank der Dokumente B und C jedoch sinkt um ein beträchtliches Maß, da sich der Bewertungsindex von A jetzt auf drei Dokumente verteilen muss.

3.8.2 Eingehende Kanten

3.8.2.1 Wirkung eingehender Kanten

Eine noch wichtigere Rolle als die ausgehenden Kanten spielen die eingehenden Kanten eines Dokumentes. Sie repräsentieren in einer gewissen Art und Weise die Meinung anderer Webseitenbetreiber zu der eigenen Webseite. Daher erhöht jede eingehende Kante eines Dokumentes deren Pagerank. Dieser nun erhöhte Pagerank wird teilweise an die externen Kanten dieses Dokumentes weitergegeben. Dadurch profitiert nicht nur das Dokument, das die eingehende Kante bekommen hat, sondern auch die Dokumente, die von ihr verlinkt worden sind. In dem nun folgenden Beispiel soll die Wirkung eingehender Kanten näher erläutert werden.

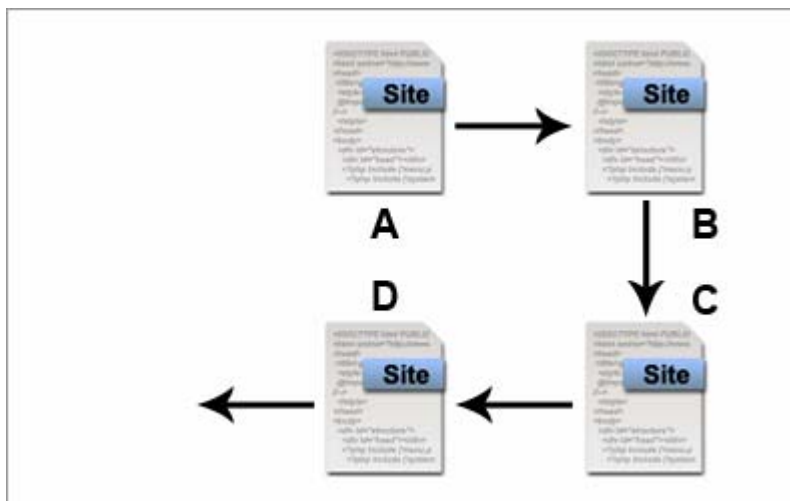


ABBILDUNG 3.8.2-1 WIRKUNG EINGEHENDER KANTEN (BEISPIEL 8)

In dem in Abbildung 3.8.2-1 gezeigten Miniweb handelt es sich um eine Aneinanderreihung von verschiedenen Dokumenten, die jeweils eine ausgehende Kante und, bis auf A, eine eingehende Kante besitzen. Bei einem angenommenen Dämpfungsfaktor d in Höhe von 0,85 und einem Startwert von 1 ergeben sich die folgenden Gleichungen für die Pagerank-Berechnung:

$$PR(A) = 0,15$$

$$PR(B) = 0,15 + 0,85 * PR(A)$$

$$PR(C) = 0,15 + 0,85 * PR(B)$$

$$PR(D) = 0,15 + 0,85 * PR(C)$$

Durch das Lösen des Gleichungssystems ergeben sich folgende Werte für die einzelnen Dokumente:

$$PR(A) \approx 0,15$$

$$PR(B) \approx 0,28$$

$$PR(C) \approx 0,39$$

$$PR(D) \approx 0,48$$

In dem nun folgenden Beispiel herrschen dieselben Bedingungen, wie in Abbildung 3.8.2-1, außer dass jetzt noch ein Dokument X mit dem festen Pagerank-Wert von 10 hinzukommt.

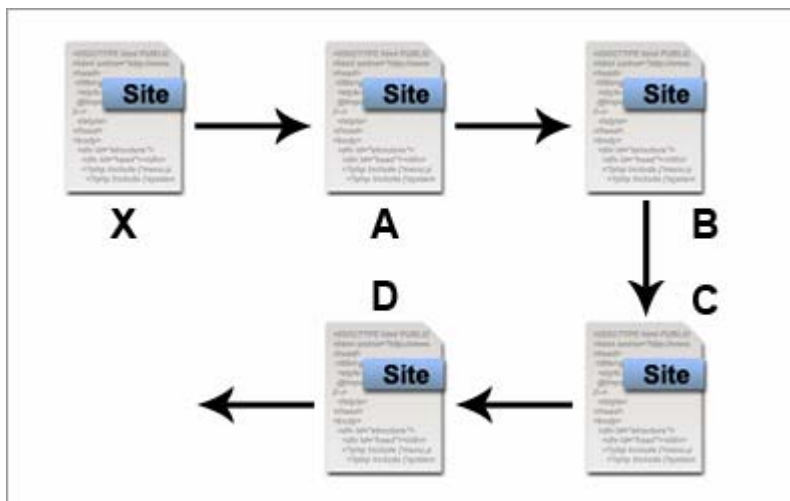


ABBILDUNG 3.8.2-2 WIRKUNG EINGEHENDER KANTEN (BEISPIEL 9)

Bei einem angenommenen Dämpfungsfaktor d in Höhe von 0,85 ergeben sich die folgenden Gleichungen für die Pagerank-Berechnung:

$$PR(A) = 0,15 + 0,85 * PR(X)$$

$$PR(B) = 0,15 + 0,85 * PR(A)$$

$$PR(C) = 0,15 + 0,85 * PR(B)$$

$$PR(D) = 0,15 + 0,85 * PR(C)$$

Durch das Lösen des Gleichungssystems ergeben sich folgende Werte für die einzelnen Dokumente:

$$PR(A) \approx 8,65$$

$$PR(B) \approx 7,51$$

$$PR(C) \approx 6,53$$

$$PR(D) \approx 5,69$$

Die unmittelbare Wirkung der zusätzlichen Kante auf Dokument A setzt sich also über die Verlinkung der einzelnen Dokumente untereinander fort. Die Höhe des Pagerank, der an das verlinkte Dokument weitergegeben wird, hängt ganz entscheidend mit der Wahl des Dämpfungsfaktors d zusammen.

3.8.2.2 Warum der Dämpfungsfaktor?

In dem folgenden Beispiel wird aufgezeigt, was mit dem Pagerank passiert, wenn der Dämpfungsfaktor verändert wird. Der Dämpfungsfaktor bewegt sich im Bereich zwischen 0 und 1. Das in Abbildung 3.8.2-3 gezeigte kleine Miniweb soll an dieser Stelle als Beispiel dienen.

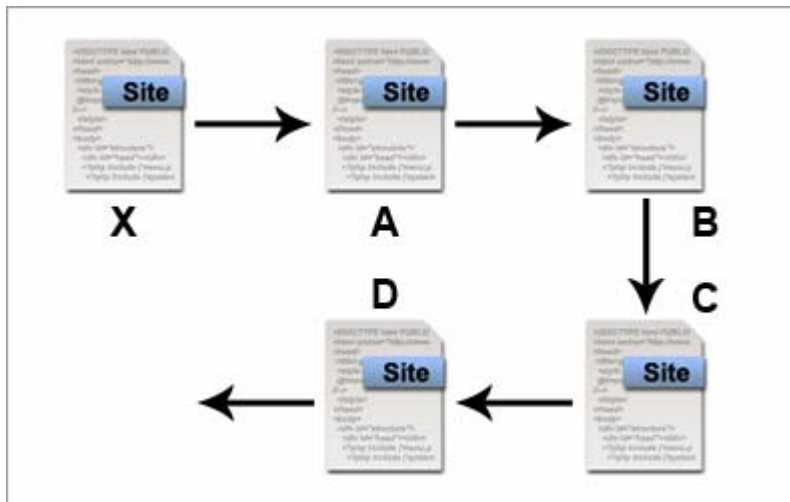


ABBILDUNG 3.8.2-3 DÄMPFUNGSFAKTOR (BEISPIEL 9)

In diesem Beispiel handelt es sich erneut um eine Aneinanderreihung von Dokumenten, die jeweils nur eine ausgehende und entsprechend nur eine eingehende Kante besitzen. Das Dokument X hat in diesem Beispiel einen statischen Pagerank von 10. Alle anderen Dokumente werden mit einem Startwert von 1 initialisiert.

Die Wahl des Dämpfungsfaktors nimmt einen wesentlichen Einfluss auf den Grad der Weitergabe des Pagerank von X. Nimmt man zum Beispiel einen extrem großen Dämpfungsfaktor $d = 1$, so ergibt sich nach der Formel:

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

für A, B, C, D ein Pagerank von 10. Das bedeutet, dass der komplette Pagerank von X an A, dann von A an B, von B an C und von C an D weitergegeben wurde. Wie aus der Tabelle 3.8.2-4 ersichtlich wird, bekommen A, B, C und D mit Annäherung des Dämpfungsfaktors d an den Wert 1 einen viel zu hohen Pagerank.

Wenn man sich jetzt das andere Extrem, Dämpfungsfaktor $d = 0$, ansieht, stellt man fest, dass das Dokument X, welches das bedeutendste Dokument in unserem Miniweb darstellt, keinerlei Einfluss auf die anderen Dokumente, auf die es verlinkt, hat.

Benutzt man nun einen Dämpfungsfaktor d von 0,85 für das in Abbildung 3.8.2-3 gezeigte Miniweb, so ergeben sich, nach dem Einsetzen in die Definition des Pagerank, folgende Gleichungen:

$$PR(A) = 0,15 + 0,85 * PR(X)$$

$$PR(B) = 0,15 + 0,85 * PR(A)$$

$$PR(C) = 0,15 + 0,85 * PR(B)$$

$$PR(D) = 0,15 + 0,85 * PR(C)$$

Mit den folgenden Werten für die einzelnen Dokumente:

$$PR(A) \approx 8,65$$

$$PR(B) \approx 7,5$$

$$PR(C) \approx 6,53$$

$$PR(D) \approx 5,69$$

Die folgende Tabelle 3.8.2-4 zeigt noch einmal zusammenfassend, welchen Einfluss der Dämpfungsfaktor d auf das Miniweb in Abbildung 3.8.2-3 hat.

Dämpfungsfaktor d	PR(A)	PR(B)	PR(C)	PR(D)
1	10	10	10	10
0,9	9,1	8,29	7,56	5,69
0,85	8,65	7,5	6,53	5,69
0,8	8,2	6,76	5,6	4,68
0,7	7,3	5,41	4,09	3,1
0,6	6,4	4,24	2,95	2,12
0,5	5,5	3,24	2,13	1,56
0,4	4,6	2,44	1,58	1,2
0,3	3,7	1,81	1,2	1,06
0,2	2,8	1,36	1,07	1,01
0,1	1,9	1,09	1	1
0	1	1	1	1

TABELLE 3.8.2-4 DÄMPFUNGSFAKTOR

Wie man gesehen hat, ist der Grad der Weitergabe der Pagerank im Großen und Ganzen von der Höhe des Dämpfungsfaktors d abhängig. Die Tabelle 3.8.2-4 zeigt, dass mit steigendem Dämpfungsfaktor die ausgehenden Kanten bevorzugt werden und sich deren Pagerank erhöht. Für den Pagerank des verlinkenden Dokumentes, in diesem Fall C, fungieren hohe Dämpfungsfaktoren jedoch als „Senke“.

3.9 Erweiterung des Pagerank-Verfahrens

3.9.1 Differenzierte Beurteilung von Kanten

3.9.1.1 Position der Kanten

Es macht durchaus Sinn, die Wichtigkeit von verschiedenen Kanten an dessen Position im Dokument oder an der Position in dessen Struktur festzumachen. Zum Beispiel werden Kanten, die direkt von der Hauptseite (Startseite) eines wichtigen Dokumentes kommen, wesentlich höher bewertet, als Kanten, die in der Struktur weiter hinten sind. Kanten, die in ihrer Farbe und Formatierung stark hervorgehoben sind, werden ebenfalls höher bewertet. Auch Kanten mit großer Schrift oder Betonung auf anderem Weg werden höher bewertet. Aber auch Kanten, die sich im oberen Teil eines Dokumentes befinden, werden höher gewichtet als Kanten im unteren Teil. Auf diesem Wege kann das Surfverhalten von Benutzern annähernd nachgeahmt werden, da ein Benutzer bewiesenermaßen eine große farbige Kante einer kleinen unscheinbaren Kante vorzieht.

3.9.1.2 Aktualität der Dokumente

Ein weiteres wichtiges Merkmal für die Bedeutsamkeit eines Dokumentes ist die Aktualität der eingehenden Kanten. In vielen Fällen ist es auch ratsam, den Pagerank zu erhöhen, wenn die Backlinks von Dokumenten kommen, die immer aktuellen und wechselnden Inhalt haben. Daraus kann man schließen, dass die Kanten, die von solchen aktuellen Dokumenten kommen, nicht veraltet sind.

3.9.1.3 Geografischer Abstand zwischen Dokumenten

Der Grundgedanke des Pagerank war, dass ein Dokument umso bedeutsamer ist, desto bedeutsamer die auf dieses Dokument verlinkenden Dokumente sind. Aber auch die Anzahl der Backlinks spielt eine große Rolle. Einige Webseitenersteller wollten diesen Fakt zur Erhöhung ihres Pagerank ausnutzen und erstellten Dokumente, deren Dasein nur darin bestand, auf die eigentliche Hauptwebseite zu verlinken, so genannte

„Linkfarmen“. Durch diese Manipulation sollte erreicht werden, dass sie bei Google an Platz eins stehen. Das wäre nicht nur ein Problem für die Qualität der Suche von Google, denn der so erschwindelte erste Platz stellt den Benutzer der Google-Suchmaschine auf keinen Fall zufrieden, sondern auch für das gesamte World Wide Web, denn auf diese Weise würde das Internet nur unnötig vergrößert werden. Eine einfache Möglichkeit, den „Linkfarmen“ entgegenzuwirken, besteht darin, den geografischen Abstand zwischen den Dokumenten mit in die Bewertung von Kanten einzubeziehen. Das bedeutet, dass zum Beispiel Kanten, die sich auf ein und demselben Web-Server befinden, weniger in die Bewertung mit einfließen, als Kanten, die von anderen möglichst weit entfernten Web-Servern kommen. Warum sollten wohl 100 Links von ein und derselben Domain auf das gleiche Dokument zeigen? Daraus folgt, dass der Pagerank für Dokumente steigt, deren Backlinks von Organisationen und Autoren aus verschiedenen geografischen Orten kommen.

3.9.2 Themenbasierter Pagerank

3.9.2.1 Intelligenter Surfer von Richardson und Domingos

Zur Erläuterung des Ansatzes des „intelligenten Surfer“ [RD02] nehmen sich Matthew Richardson und Pedro Domingos das Random Surfer Model zur Hand. Da das Random Surfer Model darauf basiert, dass es wahllos und zufällig Kanten verfolgt und sich nicht nach einer Suchanfrage richtet, ist es nicht sehr effektiv. Der „intelligente Surfer“ soll Kanten nur entsprechend seiner Suchanfrage verfolgen und nach dem Abbruch des Surfvorgangs nur Webseiten aufrufen, deren Themengebiet mit dem der Suchanfrage übereinstimmt.

Somit sind für den „intelligenter Surfer“ nach Richardson und Domingos nur Dokumente relevant, die den vom Surfer gesuchten Begriff auch tatsächlich enthalten. Das bedeutet, dass es eine eigens für den Suchbegriff neue Pagerank-Berechnung geben muss. Diese Berechnung stützt sich dabei ausschließlich auf Kanten zwischen Dokumenten, die den Suchbegriff enthalten.

Das Model wirft aber einige Probleme auf. Zum Beispiel bei Suchbegriffen, die nicht sehr häufig im World Wide Web vorkommen. Um einen guten Pagerank bekommen zu können, müssten die Dokumente alle aufeinander verlinken. Und es könnte auch passieren, dass einige sehr relevante Dokumente wenig verlinkt sind und deswegen wenig Berücksichtigung finden. Des Weiteren ist ein kleiner Subbereich des Web wesentlich anfälliger für Spam (Linkfarmen).

Ein weiteres großes Problem stellt die Berechnung eines solchen themenbasierenden Pagerank dar. In ihrer Veröffentlichung „The Intelligent Surfer“ [RD02] schätzen Richardson und Domingos sowohl den Speicher- als auch den Rechenbedarf für mehrere 100.000 Begriffe auf das 100- bis 200-fache des ursprünglichen Pagerank-Verfahrens. Noch problematischer ist der Zeitbedarf für die Berechnung. Wenn man mit fünf Stunden für eine herkömmliche Berechnung des Pagerank auskam, steigt die Berechnung

bei themenbasierter Ansicht auf etwa drei Wochen. Dadurch scheint der Einsatz eines solchen Berechnungssystems außer Frage. Aber in einigen Foren wird diskutiert, dass man sehr wohl darauf achten sollte, dass die eingehenden und besonders die ausgehenden Kanten zu einem Themengebiet gehören, in dem man seine Dokumente selbst auch wieder finden möchte.

3.9.2.2 Topic Sensitiv Pagerank nach Taher Haveliwala

Die herkömmliche Berechnung des Pagerank sieht für jedes Dokument nur einen Pagerank vor. Nach Taher Haveliwalas Idee des „Topic Sensitiv Pagerank“ [HA02] bekommt jede Webseite mehrere Pagerank zugewiesen, die jeweils auf bestimmte Themen abgestimmt sind. Haveliwala sieht es vor, das World Wide Web in 16 Themengebiete zu unterteilen. Als repräsentative Themengebiete werden in diesem Zusammenhang die 16 Hauptkategorien des Open Directory Projekt (ODP) vorgeschlagen. Es handelt sich hierbei um das größte und umfangreichste von Menschen erstellte Linkverzeichnis, bei dem Webseiten Themengebieten zugeordnet werden.

Der Ansatz von Haveliwala basiert auf der Grundlage, dass man bestimmten Dokumenten eines Themengebietes einen gewissen Bonus, den so genannten „Yahoo-Bonus“, zuspricht. Das bedeutet, dass für die Berechnung jedes einzelnen Themengebietes alle Dokumente des Themengebietes in dem Moment höher bewertet werden, als Dokumente anderer Themengebiete. Soll zum Beispiel ein Dokument mit dem Thema „Jaguar“ indiziert werden, wird der Pagerank der Themengebiete „Gesellschaft: Tiere“ oder „Wirtschaft: Kraftfahrzeuge“ mit hoher Wahrscheinlichkeit viel höher bewertet als der Pagerank derselben Dokumente für jedes andere Themengebiet. Jetzt wird während einer Suchanfrage eines Surfers das eingegebene Suchwort sofort klassifiziert, das heißt, einem bestimmten Themengebiet zugeordnet, und er bekommt nur Ergebnisse des Themengebietes.

Aber woher weiß die Suchmaschine, zu welchem Themengebiet zum Beispiel das Suchwort „Jaguar“ gehört? Für die Einschätzung, zu welchem

Thema ein Suchwort passen könnte, gibt es zwei Ansätze. Einmal wird anhand der Wahrscheinlichkeit entschieden, zu welchem Themengebiet das Suchwort gehören könnte. Diese Möglichkeit ist aber augenscheinlich nicht sehr genau. Der zweite Ansatz analysiert die Surfer-History. Dabei wird untersucht, welche Begriffe vorher vom Surfer gesucht worden sind. Waren es zum Beispiel Begriffe wie „Leopard“ oder „Puma“, kann davon ausgegangen werden, dass mit dem jetzigen Suchwort „Jaguar“ kein Auto, sondern das Tier gemeint ist.

Durch eine Einteilung in verschiedene Themen für die Berechnung des Pagerank wird das Problem verhindert, dass stark verlinkte Dokumente bei einer Suchanfrage weit oben in der Ergebnisliste aufgeführt werden, obwohl sie mit dem thematischen Hintergrund dieser speziellen Anfrage nichts zu tun haben.

4 Zusammenfassung

Die Google-Suchmaschine ist die Suchmaschine weltweit, die am effektivsten arbeitet. Einer der Hauptgründe dafür ist der Pagerank, der die Wichtigkeit eines jeden Dokumentes widerspiegelt. Der Pagerank-Algorithmus basiert im Wesentlichen auf dem Random Surfer Model, das das Surfverhalten eines Benutzers nachahmt. Dieses Model verfolgt die ausgehenden Links eines jeden Dokumentes und hangelt sich auf diese Weise durch das gesamte World Wide Web. Probleme können auftreten, wenn ein Dokument keine ausgehenden Links besitzt. Diese werden dann zur Berechnung entfernt und später wieder hinzugefügt.

Bei der Berechnung des Pagerank der einzelnen Dokumente werden zum Beispiel auch die Art der Hervorhebung der Links oder der geografische Abstand zwischen den Dokumenten berücksichtigt. Das Pagerank-Verfahren beruht also auf der Linkstruktur des gesamten World Wide Web, was bedeutet, dass das Hinzufügen oder Entfernen einer kompletten Webseite oder von Unterseiten bereits bestehender Webseiten einen wesentlichen Einfluss auf den Pagerank der Dokumente hat.

Im Rahmen der Suchmaschinenoptimierung kann zum Beispiel das Konzentrieren der ausgehenden Links auf nur ein Dokument zu einem höheren Pagerank führen. Außerdem macht jeder Autor eines Dokumentes sozusagen implizit eine Aussage über seine subjektive hohe Meinung von Dokumenten, auf die er durch einen Link verweist. Somit trägt jeder Autor dazu bei, die Wichtigkeit der Dokumente zu bestimmen.

5 Glossar

Backlinks	Eingehende Links einer Webseite
Dangling Links	Kanten zu Dokumenten die keine ausgehenden Kanten besitzen
Dämpfungsfaktor	Dieser Faktor wird benutzt, um den Grad der Weitergabe des Pagerank zu regulieren
Dokumente	Andere Bezeichnung für Webseiten
Kanten	Andere Bezeichnung für Links
PHP	Hypertext Preprocessor; Skriptsprache mit einer an C bzw. Perl angelehnten Syntax, die hauptsächlich zur Erstellung dynamischer Webseiten verwendet wird.
Ranking	Oder auch Rangordnung, ist das Ergebnis einer Sortierung
Rank Sinks	Sind zyklisch verlinkte Seiten (Schleifen)
Zufall-Surfer	Simuliert das Surfverhalten eines Internetnutzers

IV. INTERNETLINK-VERZEICHNIS

<http://www.abakus-internet-marketing.de/foren/>

<http://www.avaris-webdesign.de/suchmaschinen/suchmaschinen.html>

<http://www.cgl.uwaterloo.ca/Projects/Vanish/webquery-1.html>

<http://www.cs.cornell.edu/home/kleinber/auth.pdf>

<http://www.efactory.de/>

<http://hilbert.math.uni-mannheim.de/hm1/teil07.pdf>

<http://information-retrieval.de/>

<http://www.wi.uni->

[muenster.de/pi/lehre/ss05/seminarSuchen/Ausarbeitungen/JanKorves.pdf](http://www.wi.uni-muenster.de/pi/lehre/ss05/seminarSuchen/Ausarbeitungen/JanKorves.pdf)

<http://www.markhorrell.com/seo/pagerank.shtml>

<http://searchenginewatch.com/reports/article.php/2156451>

[muenchen.de/m3/teaching/numerik05/vorlesung/2005_04_29/42478.pdf](http://www.wi.uni-muenchen.de/m3/teaching/numerik05/vorlesung/2005_04_29/42478.pdf)

<http://searchengineforums.com/>

<http://www.uni-leipzig.de/~debatin/webeval.htm>

<http://www.wikipedia.de>

<http://www.webworkshop.net/>

V. QUELLENVERZEICHNIS

Alle hier aufgeführten Quellen befinden sich zusätzlich auf der beigelegten CD.

[FE03] Franz Embacher: Von Graphen Genen und dem WWW (2003),
[<http://homepage.univie.ac.at/Franz.Embacher/Lehre/aussermathAnw/Google.html>], 10.05.2005.

[HA02] Taher H. Haveliwala: Topic-Sensitive PageRank – Slides (2002),
[<http://www.stanford.edu/~taherh/papers/slides/tspr-slides.pdf>],
03.06.2005.

[HZ05] Professor Dr.-Ing. habil. Horst Zuse: Foto-Scout-Zuse (2005),
[<http://www.horst-zuse.homepage.t-online.de/wb/foto-scout-zuse.html>], 01.09.2005.

[PA98] Lawrence Page: Pagerank Patent 6,285,999 (1998),
[http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=/netahtml/src_hnum.htm&r=1&f=G&l=50&s1=6,285,999.WKU.&OS=PN/6,285,999&RS=PN/6,285,999], 23.07.2005.

[PB98] Sergey Brin and Lawrence Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine (1998), [<http://www-db.stanford.edu/pub/papers/google.pdf>], 20.08.2005.

[PO94] Martijn Koster: A Standard for Robot Exclusion (1994),
[<http://www.robotstxt.org/wc/norobots.html>], 06.04.2005.

[RD02] Matthew Richardson Pedro Domingos: The Intelligent Surfer: Probabilistic Combination of Link and Content Information in PageRank (2002),

[<http://www.cs.washington.edu/homes/pedrod/papers/nips01b.pdf>],
08.06.2005.

[VM03] Robert Barsch: Verbesserte Suchstrategien im WWW (2003),

[http://www.mathe.tu-freiberg.de/~ernst/Lehre/ALA/WebPages02/Ausarbeitungen/Thema8/google_vs_clever.pdf], 20.10.2005.