

Hauptseminar im Wintersemester 06/07

Inside Google - Algorithmen für Suchmaschinen

Thema: PageRank Grundlagen

Sergej Bors

7. Oktober 2006

Betreuer: Michael Luttenberger

# Inhaltsverzeichnis

<b>1</b>	<b>Einführung</b>	<b>3</b>
1.1	Linkstruktur von Web . . . . .	3
<b>2</b>	<b>Grundlagen</b>	<b>4</b>
2.1	Grundidee von PageRank . . . . .	4
2.2	Rank Sink . . . . .	6
2.3	Zufallssurfer-Modell . . . . .	6
2.4	Alternative Definition von PageRank . . . . .	7
<b>3</b>	<b>PageRank Berechnung</b>	<b>7</b>
3.1	Markov-Ketten Modell von PageRank . . . . .	7
3.2	PageRank Algorithmus . . . . .	9
3.3	Dangling Links . . . . .	11
<b>4</b>	<b>Effiziente Methoden</b>	<b>11</b>
4.1	Speichernutzung . . . . .	12
4.2	Techniken . . . . .	12
4.2.1	BlockRank . . . . .	12
4.2.2	Adaptive Pagerank . . . . .	12
4.2.3	Quadratische Extrapolation . . . . .	13
<b>5</b>	<b>Zusammenfassung</b>	<b>13</b>
<b>6</b>	<b>Quellenverzeichnis</b>	<b>14</b>

# 1 Einführung

Wie bekannt, wächst die Anzahl von verschiedenen Informationsquellen jeder Art im Internet sehr stark, was man aber über die Qualität von diesen nicht sagen kann. Ohne eine gute Suchmethode könnte ein Websurfer auf der Suche nach der passenden Information (z.B. durch zufällige Wahl eines Links) sein ganzes Leben verbringen. Die Möglichkeit alleine den Textinhalt nach den Suchbegriffen durchzusuchen ist auch nicht so hilfreich. Denn es kann Hunderte von Seiten geben, die diese Begriffe enthalten, jedoch keine vernünftige Antworten auf die Fragen des Benutzers bereitstellen bzw. seinen Anforderungen entsprechen. Das Sortieren dieser Menge von Webseiten nach der Relevanz ist keine einfache Aufgabe und stellt an die Suchmaschine, wenn man intuitiv vorgeht, Anforderung die Schätzungsfähigkeit eines Menschen den gegebenen Textinhalt als wichtig zu beurteilen zu simulieren. Das Letzte ist heutzutage klar fast unmöglich. Ausserdem, die Existenz von vielen ähnlichen Dokumenten kann selbst sehr gute Techniken zur Relevanzberechnung, die nur auf der Beurteilung der Textinhalte von Webseiten basiert sind, sinnlos machen, da hier auch die Popularität des Autors, der die Informationen auf seiner Homepage bereitstellt, auch eine sehr große Rolle spielt. So könnte ein Websurfer z.B. einen Computer kaufen wollen und Suchbegriffe, die er der Suchmaschine gibt, wären „Computer“ und „kaufen“. Klar, dass diese beide Wörter eine große Anzahl von Webseiten betreffen können. Jedoch braucht der Benutzer die Homepage von einem Computerverkäufer, dessen Popularität auch für die Qualität und den Preis der Ware sprechen kann.

Im alltäglichen Leben kann eine Person einen guten Ruf dadurch gewinnen, indem andere Personen sie weiterempfehlen. Je mehr Personen eine Person  $A$  weiterempfehlen und je populärer diese Personen sind, umso mehr wird die Popularität von Person  $A$  zunehmen.

Dies führt zum Gedanken, ob man diese Situation nicht auch auf das Leben der Webseiten im Internet zurückführen kann. Wie kann man denn die Link-Popularität berechnen? Verweise (abgegebene Stimmen) von anderen Webseiten auf die zu beurteilende können ja auch verschieden sein. Allein die Anzahl von eingehenden Links (*backlinks* in englischsprachigen Quellen genannt) auf der Seite ist somit kein gutes Kriterium für die Berechnung der Link-Popularität. Denn mit der Verfügbarkeit von kostenlosen Hompages ist es sehr leicht geworden, die Anzahl von eingehenden Links zu manipulieren. Aber die Wichtigkeit von solchen Links ist praktisch gleich Null im Vergleich zu Verweisen von bekannten Webressourcen. **PageRank** ist gerade ein Mittel, die Bedeutsamkeit einer Webseite zu beurteilen.

## 1.1 Linkstruktur von Web

Bevor ich mit der Erklärung des PageRank-Algorithmus anfangen möchte, möchte ich zuerst der Deutlichkeit halber kurz auf die Struktur des Web eingehen, wie sie in [2] dargestellt ist. Dies wird im Weiteren dann verwendet. Jede Webseite hat eine bestimmte Anzahl von eingehenden und ausgehenden Links, wie es in der Abb.1 zu sehen ist.

Im kleinen Teilgraph in Abb.1 sind also die Transitionen  $(1,4)$ ,  $(2,4)$ ,  $(3,4)$  die eingehenden Links und  $(4,5)$ ,  $(4,6)$  die ausgehenden Links für die Webseite 4.

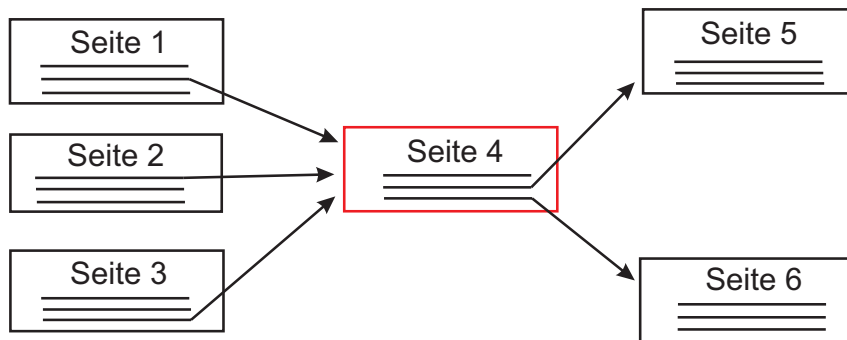


Abbildung 1: Darstellung von ein- / ausgehenden Links

## 2 Grundlagen

### 2.1 Grundidee von PageRank

Das PageRank-Konzept nutzt also nicht nur die absolute Anzahl von eingehenden Links für die Beurteilung der Relevanz einer Webseite. Laut Angaben von Google-Gründer Sergey Brin und Lawrence Page in [1] soll es einem Dokument völlig unabhängig von seinem Inhalt ein hoher Rang zugewiesen werden, wenn dieses Dokument von anderen vielleicht nicht vielen aber wichtigen Webressourcen 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 der auf sie verlinkenden Dokumente. Dies führt zum Gedanken von der rekursiven Berechnung der Bedeutsamkeit eines Dokuments aus der Bedeutsamkeit anderer Dokumente, die auf dieses verweisen. Da die Linkstruktur des gesamten Webs nämlich sehr kompliziert ist und die PageRank-Berechnung ausschließlich auf dieser beruht, sollte die Berechnung relativ schwierig sein. Den Google-Gründer gelang es jedoch den PageRank mittels eines relativ trivialen Algorithmus zu berechnen.

Der ursprüngliche PageRank-Algorithmus, wie er von Brin und Page in [1] und [2] beschrieben wurde, soll sich laut der Angaben in [3] von der gegenwärtig verwendeten Methode in Google um einige wesentliche Faktoren abweichen. Dies ist auch verständlich, da PageRank von den Google-Gründer patentiert wurde und geheim gehalten wird. Jedoch bilden die vorgestellten Grundideen die Basis vom PageRank und man kann aufgrund von dieser weitere Optimierungen vorschlagen.

Wir nehmen eine Variante, die in [1] beschrieben wurde.

Für eine Webseite  $u$  hat der PageRank  $R(u)$  folgende Gestalt:

$$R(u) = (1 - d) + d \sum_{v \in P(u)} \frac{R(v)}{N_v} \quad (1)$$

Dabei gilt folgendes:

- $R(u)$  der PageRank einer Seite  $u$
- $P(u)$  Menge der Seiten, die auf  $u$  verlinken
- $S(u)$  Seiten, auf die  $u$  verlinkt
- $N_u$  Anzahl der Seiten, auf die  $u$  verlinkt
- $N = |V|$  - Anzahl aller Seiten im Web
- $d$  ist ein Dämpfungsfaktor (mit  $0 \leq d \leq 1$ ), zu dem noch einige Worte im Folgenden gesagt werden

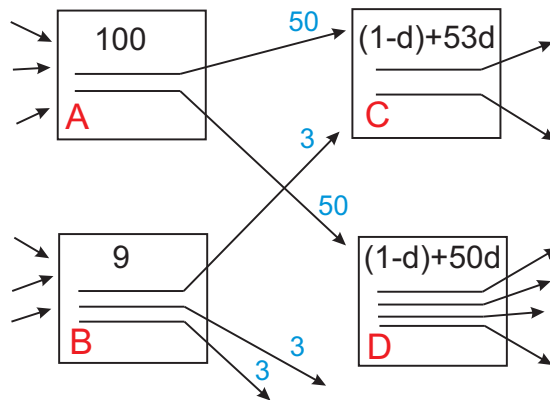


Abbildung 2: PageRank Fortpflanzung

Die Abbildung 2 illustriert die Fortpflanzung des PageRank entsprechend Formel 1, wobei Zahlen in den Kästchen den PageRank von der jeweiligen Seite bezeichnen. Zum Beispiel:

$$R(C) = (1 - d) + d\left(\frac{R(A)}{2} + \frac{R(B)}{3}\right) = (1 - d) + 53d$$

Das Verfahren basiert auf der Beziehung einzelner Seiten zueinander. Der PageRank einer Seite  $u$  bestimmt sich rekursiv aus dem PageRank von derjenigen Seiten  $v$ , die auf  $u$  verlinken. Damit nicht der gesamte Rang einer Seite zu einer anderen, auf die diese verlinkt, weitergegeben wird, wird ihr Rang durch die Anzahl ihrer ausgehenden Links geteilt. Dies stellt sicher, dass bei der Fortpflanzung von PageRank das Ergebnis nicht viel zu groß am Ende der Berechnung ist, was man auch als Wahrscheinlichkeitsverteilung betrachten kann. Ohne dies könnte eine Webseite, die einen hohen PageRank besitzt, eine andere in der Wirklichkeit nicht so wichtige Seite sehr wichtig machen, indem sie auf die Letztere verlinkt. Mittels  $N_v$  wird also der PageRank der verlinkenden Seite gewichtet. Also, je mehr ausgehende Links eine Seite  $v$  hat, desto weniger gibt sie PageRank an die Seite  $u$  weiter. Durch die Multiplikation der Summe mit dem Dämpfungsfaktor  $d$  wird der Ausmaß der Weitergabe des PageRanks verringert.

## 2.2 Rank Sink

Der Zweck der Größe „ $d$ “ besteht auch darin, eine Pathologie zu vermeiden. Es tritt nämlich ein Problem auf, wenn zwei oder mehrere Seiten, die aufeinander verlinkt sind, zu keinen anderen Seiten aus dem Webgraphen verweisen (Abbildung 3).

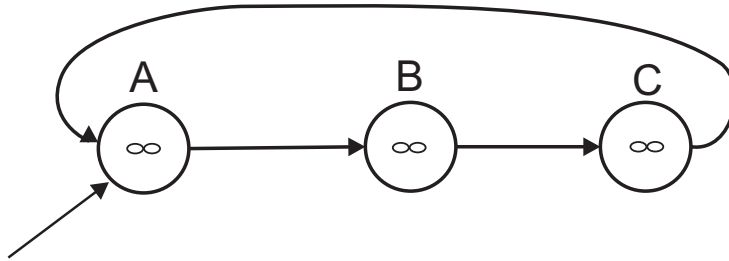


Abbildung 3: Rank Sink

Angenommen es gibt nun eine weitere Webseite, die auf eine der Seiten aus solcher Schleife zeigt, dann wird ihr PageRank bei der rekursiven Berechnung von PageRanks laut Formel 1 von der einen zur anderen Webseite in einem ständigen Kreislauf weitergegeben. Dies beeinträchtigt eine normale Verteilung von PageRanks, sodass sich der PageRank der Seiten innerhalb der Schleife permanent erhöht und nicht korrekt an andere Seiten weiterverteilt wird, da keine ausgehenden Links aus dieser Schleife existieren. Dies führt dazu, dass die PageRanks innerhalb der Schleife zu hoch und außerhalb der Schleife zu niedrig sind. Das Problem beseitigt man, indem man der ständigen Erhöhung der PageRanks innerhalb der Schleife mit dem Dämpfungsfaktor  $d$  entgegenwirkt. Desweiteren wird die Nichtweitergabe der PageRanks nach Außen, durch die Einführung sogenannter „Rank Sources“ kompensiert. Der zusätzliche Summand  $(1 - d)$  in der Formel 1 ist gerade eine Rank Source. Laut [2] wird dieses Problem auch als „Rank Sink“ bezeichnet.

L.Page und S.Brin geben in [1] für  $d$  den Wert 0.85 als erprobt optimal an.

## 2.3 Zufallssurfer-Modell

Die Google-Gründer geben in [1] und [2] eine andere Erklärung für PageRank. Sie betrachten das PageRank-Konzept als ein Verfahren zur Abbildung des Verhaltens eines Websurfers, der von einer Webseite zu einer anderen anhand beliebiger Links übergeht, ohne dabei auf die Inhalte zu achten. Dies vergleicht man mit der Wahrscheinlichkeitsverteilung eines zufälligen Wanderns über den Webgraphen. Der Zufallssurfer befindet sich mit einer bestimmten Wahrscheinlichkeit auf einer Seite, die sich aus deren PageRank bestimmen lässt. Die Wahrscheinlichkeit, dass der Zufallssurfer einen bestimmten Link anklickt, berechnet sich also aus der Anzahl der Links, die er zur Auswahl hat. Aus diesem Grund macht man auch die Gewichtung bei der rekursiven Berechnung. Unendlich lange kann ja ein realer Webserfer die Links nicht verfolgen. Irgendwan wird er gelangweilt, wenn bisher nichts gefunden wurde oder wenn er in eine Schleife geraten ist, und startet die Suche mit einer anderen Quelle. Dies begründet, warum der Dämpfungsfaktor  $d$  ins Spiel kommt. Gedämpft wird erstens, weil der Surfer nicht unendlich lange einen Pfad verfolgt, und zweitens gibt  $d$  die Wahrscheinlichkeit an, mit der er in einem Pfad

weiterklickt. Die Konstante  $(1 - d)$  gibt die Wahrscheinlichkeit des Abbruchs gleichverteilt an alle Webseiten an. Also, je höher  $d$  ist, umso wahrscheinlicher ist es, dass der Surfer die Links weiterverfolgt.

## 2.4 Alternative Definition von PageRank

Das Zufallssurfermodell liefert uns eine andere Version des PageRank-Problems, auf die Page und Brin in ihren Veröffentlichungen eingehen. In dieser Definition wird der PageRank folgendermaßen bestimmt:

$$\tilde{R}(u) = \frac{(1 - d)}{N} + d \cdot \sum_{v \in P(u)} \frac{\tilde{R}(v)}{N_v} \quad (2)$$

Zur Erinnerung:  $N$  ist hier die Anzahl aller Webseiten in Web. Im Gegensatz zur Formel (1) beschreibt Formel (2) die tatsächliche Wahrscheinlichkeit, mit der eine Webseite erreicht wird.

Dabei gilt  $R(u) = N \cdot \tilde{R}(u)$ , wie man durch Einsetzen leicht überprüft. Wir werden im folgenden Kapitel sehen, dass diese Formulierung des Problems uns erlaubt, die Existenz und Eindeutigkeit von  $\tilde{R}$  und damit auch von  $R$  zu zeigen.

In der vorigen Formulierung von PageRank, in der nicht durch die Anzahl aller Webseiten dividiert wird, beschreibt der PageRank dagegen den Wert, welchen eine Seite in der Rangliste hat, wenn man den Seiten ihre Ränge entsprechend der Hyperlinkstruktur vom Web zuweist. Dabei konvergiert die Summe der PageRanks aller Seiten gegen die Anzahl aller Seiten, wobei der durchschnittliche PageRank aller Seiten gegen 1 geht. Jede Seite hat einen minimalen PageRank von  $(1 - d)$ . Der theoretisch maximale PageRank einer Seite beträgt  $d \cdot N + (1 - d)$ , wobei  $N$  die Anzahl aller Webseiten ist. Diesen Wert bekommt man in dem Fall, wenn alle Webseiten ausschließlich auf die Seite, die so einen PageRank haben könnte, verweisen, und auch diese Seite wiederum nur auf sich selbst verweist.

Wir verwenden diese zweite Interpretation von PageRank für das sogenannte „Zufallssurfer-Modell“

Diese Interpretation von PageRank, wo es schon von den tatsächlichen Wahrscheinlichkeiten gesprochen wird, kann man auf das Markov-Ketten Modell zurückführen und damit auch den Ansatz für die Berechnung des PageRank erhalten.

## 3 PageRank Berechnung

### 3.1 Markov-Ketten Modell von PageRank

Dieser Abschnitt stützt sich im Wesentlichen auf Auseinandersetzungen in [4].

Zurückgreifend auf die Formel (2) betrachten wir das Websurfen als speziellen stochastischen Prozess auf dem Web-Graphen  $V$  und bestimmen die Wahrscheinlichkeit, dass ein Websurfer auf einer bestimmten Webseite landet. Ein Websurfer wählt ausgehend von einer Webseite  $u$  als nächste Webseite entweder

- mit Wahrscheinlichkeit  $d$  eine Seite  $v$ , auf die  $u$  verlinkt oder
- mit Wahrscheinlichkeit  $(1 - d)$  eine beliebige andere Seite im Web (z.B. per Eingabe einer URL). Hierbei liege eine Gleichverteilung zu Grunde. Somit hat jede Seite aus  $V$  die Wahrscheinlichkeit  $\frac{(1-d)}{N}$ , mit  $N = |V|$ .

Die Anzahl eingehender Links einer Webseite beeinflusst die Wahrscheinlichkeit, dass der Surfer auf dieser Webseite landet. Dies soll ebenfalls modelliert werden. Um zu verhindern, dass bestimmte Seiten durch sehr viele ausgehende Links den Rang anderer Seiten erhöhen, erhält ein ausgehender Link einer Seite  $u$  das Gewicht  $\frac{1}{N_u}$  ( $N_u$  ist ja Anzahl der Ausgehenden Links von  $u$ ).

Aus dieser Motivation macht man für den PageRank folgenden Ansatz. Gesucht ist eine Rang-Funktion

$$\tilde{R} : V \rightarrow \mathbb{R}_0,$$

so dass die rekursive Beziehung  $\tilde{R}(u) = \frac{(1-d)}{N} + d \sum_{v \in P(u)} \frac{\tilde{R}(v)}{N_v}$  erfüllt ist.

Der erste Term stellt hier die Wahrscheinlichkeit dar, dass der Surfer zufällig auf die Seite springt, während der zweite Term die Wahrscheinlichkeit beschreibt, dass der Surfer einen Pfad weiterverfolgt. Diese Wahrscheinlichkeit hängt wiederum rekursiv vom Rang  $\tilde{R}(v)$  der Vorgängerwebseite und der Anzahl von  $v$  ausgehender Links ab.

Aufgrund der Auswahl der Wahrscheinlichkeiten modelliert man das Zufällige Surfen als Markov-Prozess.

*Bemerkung:*  $\tilde{R}(u)$  bestimmt, wie oben erwähnt, die Wahrscheinlichkeit vom Landen des Websurfers auf ein einer Seite  $u$ . Deshalb muss gelten:  $\tilde{R}(V) = \sum_{v \in V} \tilde{R}(v) = 1$ .

Wir definieren eine gewichtete Adjazenzmatrix  $A \in \mathbb{R}^{n \times n}$  durch

$$A_{uv} := \begin{cases} \frac{1}{N_v}, & \text{falls } v \in P(u) \text{ (} v \text{ ist Vorgänger von } u \text{)} \\ 1, & \text{falls } u = v \text{ und } N_u = 0 \\ 0, & \text{sonst} \end{cases}$$

Mit dem Vektor  $\vec{R}$  der Dimension  $N$  bezeichnen wir dann die Bildmenge der Abbildung  $\tilde{R}$ , so dass gilt  $\vec{R}_u = \tilde{R}(u)$

Unter der Annahme, dass eine beliebige, aber feste Reihenfolge von Knoten des Webgraphen zu Grunde liegt, kann die Gleichung (2) in Matrixnotation geschrieben werden als

$$\vec{R} = d \cdot A \cdot \vec{R} + \frac{(1-d)}{N} \cdot \mathbf{1}_N, \quad (3)$$

wobei  $\mathbf{1}_N := (1, \dots, 1)^T$ . Wegen der Bemerkung gilt:  $\mathbf{1}_N^T \cdot \vec{R} = 1$

Die letzte Beziehung in (3) eingesetzt, erhalten wir

$$\begin{aligned}
\vec{R} &= d \cdot A \cdot \vec{R} + \frac{(1-d)}{N} \cdot \mathbf{1}_N \cdot \mathbf{1}_N^T \cdot \vec{R} \\
&= (d \cdot A + \frac{(1-d)}{N} \cdot J_N) \cdot \vec{R}.
\end{aligned} \tag{4}$$

Die Einträge von  $J_N$  sind dabei alle gleich 1. Laut Konstruktion gilt  $\forall v \in V : \sum_{u \in V} A_{uv} = 1$  sowie

$$\forall 1 \leq j \leq N : \frac{1}{N} \sum_{i=1}^N (J_N)_{ij} = 1.$$

Daher wegen  $0 \leq d \leq 1$  auch die Summen der Einträge in den Spaltenvektoren der Matrix

$$M := d \cdot A + \frac{(1-d)}{N} \cdot J_N$$

gleich 1. Das heißt Matrix  $M$  ist **stochastisch**. Damit erfüllt  $M$  die Eigenschaften einer homogenen Markovkette.

Der folgende Hauptsatz der **Perron-Frobenius Theorie** liefert die Existenz des PageRank Vektors, sowie ein Kriterium für dessen Eindeutigkeit.

**THEOREM.** Für jede stochastische Matrix  $\Pi$  existiert ein Wahrscheinlichkeitsvektor  $\vec{x}$  mit  $\Pi \cdot \vec{x} = \vec{x}$ ,  $\vec{x} \geq 0$  und  $\sum_i x_i = 1$ . Gibt es darüberhinaus eine natürliche Zahl  $k$ , so dass  $\Pi^k$  nur von Null verschiedene Einträge besitzt, dann ist  $\vec{x}$  eindeutig bestimmt. Insbesondere existiert  $\lim_{k \rightarrow \infty} \Pi^k$  und in jeder Zeile resultierender Matrix  $\Pi^\infty$  steht der Wahrscheinlichkeitsvektor  $\vec{x}$ .

In unserem Fall ist also  $\vec{R}$  (PageRank) die stationäre Verteilung der Markovkette mit Matrix  $M$ , so dass gilt

$$M \cdot \vec{R} = \vec{R} \tag{5}$$

Das Perron-Frobenius-Theorem garantiert dabei, dass genau eine Lösung existiert.

Um  $\vec{R}$  nun zu berechnen reicht es erstmal das Lineare Gleichungssystem (5) zu lösen. Da aber die so definierte Matrix von Webgraphen sehr groß ist, kann man die Methoden zum Lösen von LGS nicht immer anwenden.

## 3.2 PageRank Algorithmus

Die Grenzverteilung  $\vec{R}$  von oben kann man mit genügend vielen Iterationen anhand der von Brin und Page in [2] vorgestellten Methode **Power Method** annähern.

Sei  $\vec{R}_0 : V \rightarrow [0, 1]$  eine Startverteilung mit  $\sum_{v \in V} |\vec{R}_0(v)| = 1$  und  $\vec{R}_0$  der zugehörige Wahrscheinlichkeitsvektor, dann gilt

$$\vec{R}_k = M \cdot \vec{R}_{k-1},$$

also nach der  $k$ -facher Iteration

$$\vec{R}_k = M^k \cdot \vec{R}_0.$$

Die Konvergenz gegen die Grenzverteilung  $\vec{R}$  ist in [3] gezeigt. Der folgende Algorithmus berechnet den Pagerank.  $\vec{S}$  sei dabei die Startverteilung mit  $\vec{S} = (S_1, \dots, S_N)^T$  und  $S_i = \frac{1}{N} \forall i \in 1, \dots, N$ .

```

 $\vec{R}_0 \leftarrow \vec{S}$ 
 $i = 1;$ 
repeat
   $\vec{R}_i \leftarrow M * \vec{R}_{i-1}$ 
   $\delta = \|\vec{R}_i - \vec{R}_{i-1}\|_1$ 
   $i = i + 1$ 
until  $\delta < \epsilon$ 

```

Hierbei wird abgebrochen, falls die Verbesserung im  $(i + 1)$ -ten Schritt <sup>1</sup>  $\|\vec{R}_{i+1} - \vec{R}_i\|_1$  die Schranke erfüllt.

Das in Abbildung 4 ausgeführte Beispiel illustriert, wie PageRanks für einen kleinen Webgraph aus drei Seiten bestehend anhand der Power Methode berechnet werden.

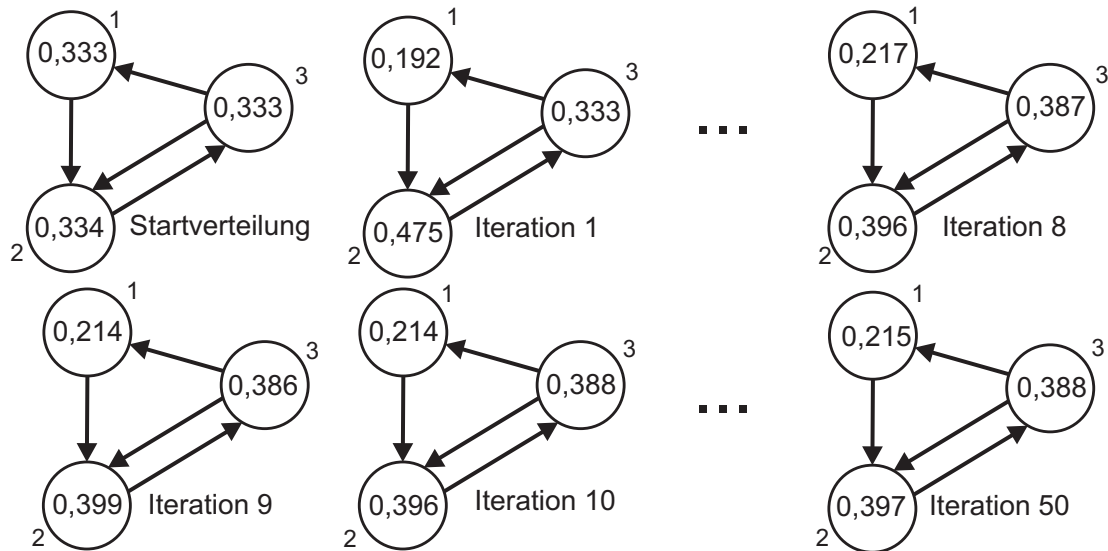


Abbildung 4: Berechnung von PageRank mit Power Methode

<sup>1</sup>Vektornorm:  $\|a\|_1 = \sum_{i=1}^n |a_i|$ , für  $a \in \mathbb{R}^n$

Als Startverteilung wurde also ein Vektor  $\vec{S} = (0.333, 0.333, 0.334)^T$  gewählt. Die Markov-Matrix  $M$  wurde dann entsprechend den Überlegungen aus dem vorigen Abschnitt aus der Graphstruktur in Abb.4 interpretiert ( $d = 0,85$  gewählt).  $M$  sieht also so aus:

$$M = \begin{pmatrix} 0,05 & 0,05 & 0,475 \\ 0,9 & 0,05 & 0,475 \\ 0,05 & 0,9 & 0,05 \end{pmatrix}$$

In der Abbildung 4 sind also Kreise Webseiten. In Kreisen sind PageRank-Werte für einen aktuellen Iterationsschritt eingetragen. Wie es in der Abbildung 4 zu sehen ist, konvergieren die Werte ziemlich schnell gegen die stationäre Verteilung, die ungefähr folgenden Wert hat, wenn man LGS (5) löst:  $\vec{R} = (0.2148, 0.3974, 0.3878)^T$ . Laut Angaben von Page und Brin sind für das gesamte Web ca. 50 Iterationen hinreichend.

### 3.3 Dangling Links

Links, die auf Webseiten ohne ausgehende Links verweisen, die man auch als sogenannte „Dangling Links“ bezeichnet, stellen noch ein gewisses Problem dar. Sie beeinflussen die korrekte Berechnung des PageRanks, weil nicht bekannt ist, wo ihr Gewicht weiterverteilt werden soll. „Dangling Links“ werden aus dem Modell vor der Berechnung einfach entfernt, weil sie den PageRank einer Seite nicht direkt beeinflussen, bis alle PageRanks berechnet sind. Das Entfernen von diesen Links kann neue „Dangling Links“ erzeugen, deshalb kann es notwendig sein, diesen Vorgang in mehreren Iterationen durchzuführen. Wenn nun alle PageRanks berechnet sind, werden auf Basis der PageRanks von Seiten, die auf Dangling Links verlinken, den „Dangling-Links“-Seiten entsprechende Werte zugewiesen. Wobei, wie beim Entfernen, können wieder mehrere Iterationen entstehen. Was man aber nicht vergessen darf, dass sich beim Entfernen von Dangling Links die Normalisierung der anderen Links auf der gleichen Seite im gewissen Maße verändert. Jedoch beeinflusst diese Veränderung das Ergebnis im Endeffekt sehr wenig, sodass diese Tatsache bei der Berechnung vernachlässigt werden kann.

## 4 Effiziente Methoden

Der Grundbaustein des PageRank-Algorithmus ist die Berechnung der stationären Verteilung für die Markov-Kette. Da aber die Matrix des Webgraphen sehr groß ist und besteht nämlich aus mehreren Milliarden von Knoten (Webseiten) und einigen Billionen von Links, kann die Berechnung von PageRank unheimlich lange dauern, denn mindestens 50 Iterationen der Power Methode sind notwendig, um mehr oder weniger gute PageRanks zu erhalten. Außerdem kommt in Frage eine effiziente Nutzung und Verwaltung des Speichers bei der Bearbeitung. Denn bei der Arbeit mit so einer riesigen Struktur, wie Webgraph, treten wiederum Probleme auf. Da die Berechnung der PageRanks für alle Webseiten sehr viel Zeit in Anspruch nimmt, so dass man nicht gleich, wenn eine Anfrage kommt, diese durchführen kann und rechtzeitig antworten, wird PageRank im Allgemeinen „offline“ berechnet, d.h. nachdem der Web-Crawler das Web durchsucht

hat und bevor irgendwelche Anfragen gestellt werden. Um also die Konvergenz der PageRank-Berechnung zu beschleunigen, sind Forscher der Universität Stanford damit beschäftigt, Verbesserungen zu finden bzw. neue Methoden zu erfinden.

## 4.1 Speichernutzung

Die Matrix der Markov-Kette kann also in den Hauptspeicher vielleicht nicht passen. Wenn die Berechnung auf einer genügend kleinen Teilmenge des Webs durchgeführt werden sollte, so kann man die üblichen Methoden anwenden. Andernfalls muss man kreativ sein um jedoch die Berechnung zu ermöglichen. Im Falle, wenn man die Matrix die Speicherkapazität übersteigt, kann man die Daten komprimieren und dann auf komprimierte Daten effiziente Methoden anwenden.

## 4.2 Techniken

Es gibt drei wichtige Verbesserungen von PageRank, auf die man im Rahmen dieses Seminars eingehen kann. Zu diesen sind ausführliche Informationen in den Quellen der Universität Stanford in [5],[6] und [7] zu finden.

### 4.2.1 BlockRank

Die Forscher der Universität Stanford haben die Linkstruktur des Web beobachtet und gefunden, dass 80% der Links jeder beliebigen Webseite auf interne Seiten derselben Webseite verlinken. Die Behandlung von solchen relativ geschlossenen Systemen ist bei der PageRank-Berechnung wesentlich einfacher. Diese Blockstruktur des Webgraphen lässt sich ausnutzen und macht eine Beschleunigung der Berechnung des PageRank möglich. Der Algorithmus arbeitet zunächst die internen Links einer Webseite ab, bevor das gesamte Web durchsucht wird. Damit wird der PageRank für Seiten einer Domäne unabhängig berechnet. Die so lokal berechnete PageRanks werden anschließend am PageRank der zugehöriger „Hauptseite“ gewichtet. Das ganze liefert eine Geschwindigkeitsverbesserung um den Faktor zwei.

### 4.2.2 Adaptive Pagerank

Auch „Modified Adaptive PageRank“ genannt. Diese Technik ist vor Allem an der iterativen Berechnung vom PageRank orientiert. Der Verlauf der Konvergenzrate von PageRanks einzelner Webseiten ist ungleichmäßig. Nämlich konvergiert eine große Anzahl der Seiten sehr schnell, wobei die Berechnungen für einige Seiten sehr viel Zeit in Anspruch nehmen, bis ein vernünftiger Wert erzielt ist. Die Seiten deren PageRank-Berechnung langsam ist, sind in den meisten Fällen Seiten mit einem hohen PageRank-Wert. Im Gegensatz zu diesen ist die Konvergenzrate von weniger relevanten Webseiten, weil diese weniger verlinkt sind, viel schneller. Aufgrund dieser Beobachtung konnten die Forscher einen einfachen Algorithmus entwickeln. Die redundanten Berechnungen werden

vermieden, indem die Berechnung der PageRanks für viele wenig relevante Seiten nicht bei jeder folgenden Iteration neu durchgeführt wird. Durch diesen einfachen Trick wird der PageRank-Algorithmus um ungefähr 30% schneller.

### 4.2.3 Quadratische Extrapolation

Diese Verbesserung kann man besonders auf mathematisches Problem der Markov-Ketten zurückführen. Es wird angenommen, dass die stationäre Verteilung der Markov-Kette für das gesamte Web als Linearkombination der in den ersten drei Iterationen der Power Methode berechneten Vektoren bestimmt werden kann. Die Vektoren werden einfach geschätzt, auf der Annahme basierend, dass die Linkstruktur des Webs wesentlich einfacher ist, als es uns Realität bietet. Bei der Schätzung wird die Tatsache genutzt, dass der erste Eigenwert einer stochastischen irreduziblen aperiodischen Markov-Kette immer gleich eins ist. Im Anschluß wird die PowerMothode noch einige Male iteriert, um Folgen der näherungsweise Berechnung zu bereinigen. Das Verfahren liefert nicht immer eine völlig korrekte Lösung, jedoch ist nah an dem gewünschten Ergebnis und stellt damit eine Verbesserung dar.

## 5 Zusammenfassung

Der PageRank ist nicht der einzige Erfolgsfaktor von Google, obwohl man diesen als Kern dieser Suchmaschine bezeichnen kann. Im Spiel sind noch andere mehrere Einzelheiten für die Relevanzvestimmung von Webseiten, die Google zum Durchbruch auf dem Markt geführt haben. Diese werden von der Firma selbstverständlich geheim gehalten.

Die Anzahl von Webseiten im Web wächst exponentiell und diese Tatsache erfordert immer neue Optimierungen an Suchverfahren. Das heißt PageRank ist immer ein breites Feld für Forschungen. Die Kombination von hier erwähnten Verfahren, die die Effizienz von PageRank steigern, kann die PageRank-Berechnung wesentlich beschleunigen. Zu diesem Thema wird viel diskutiert und man kann erwarten, dass zukünftig andere innovative Suchverfahren erscheinen.

## 6 Quellenverzeichnis

- [1] Sergey Brin and Lawrence Page  
The Anatomy of a Large-Scale Hypertextual Web Search Engine  
<http://www-db.stanford.edu/pub/papers/google.pdf>
- [2] Page, Lawrence; Brin, Sergey; Motwani, Rajeev; Winograd, Terry.  
The PageRank Citation Ranking: Bringing Order to the Web  
<http://dbpubs.stanford.edu:8090/pub/1999-66>
- [3] Amy N. Langville, Carl D. Meyer  
Deeper Inside PageRank  
[http://meyer.math.ncsu.edu/Meyer/PS\\_Files/DeeperInsidePR.pdf](http://meyer.math.ncsu.edu/Meyer/PS_Files/DeeperInsidePR.pdf)
- [4] Prof Dr. Michael Clausen  
Grundlagen des Multimediaretrivals  
[http://www-mmdb.iai.uni-bonn.de/lehre/materialMMRws0506/Skript\\_MMR\\_WS0506.pdf](http://www-mmdb.iai.uni-bonn.de/lehre/materialMMRws0506/Skript_MMR_WS0506.pdf)
- [5] Kamvar, Sepandar; Haveliwala, Taher; Golub, Gene.  
Adaptive Methods for the Computation of PageRank  
<http://dbpubs.stanford.edu:8090/pub/2003-26>
- [6] Kamvar, Sepandar; Haveliwala, Taher; Manning, Christopher; Golub, Gene.  
Exploiting the Block Structure of the Web for Computing PageRank  
<http://dbpubs.stanford.edu:8090/pub/2003-17>
- [7] Kamvar, Sepandar; Haveliwala, Taher; Manning, Chris; Golub, Gene.  
Extrapolation Methods for Accelerating PageRank Computations  
<http://dbpubs.stanford.edu:8090/pub/2003-16>