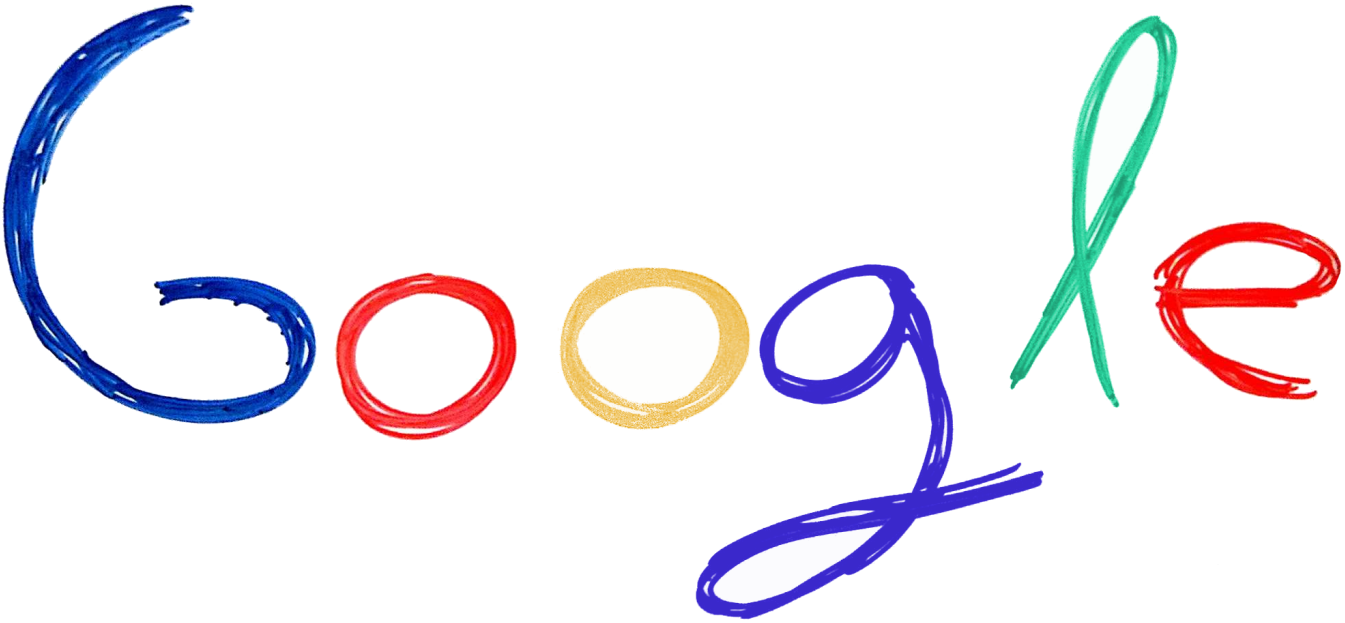


Page Rank

Wie Google tickt - ein Blick hinter die Kulissen

Stefan Trauth

Alle Inhalte dieses Artikels stehen unter einer Creative-Commons-Lizenz (Namensnennung, nicht-kommerzielle Weitergabe unter gleichen Bedingungen, Deutschland-Lizenz 3.0). Attribution: Stefan Trauth

7. August 2011

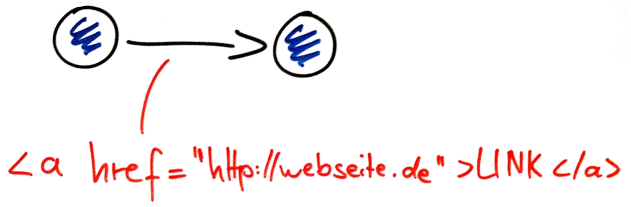
Ich gebe zu: Ich nutze die Google-Suchte täglich. Nein! Stündlich! Mehrmals! Ihr nicht auch? Aber haben wir uns wirklich einmal richtig Gedanken darüber gemacht wie dieses wundervolle Suchwerkzeug es immer wieder schafft uns mit gespenstisch guten Suchergebnissen zu überraschen? Meistens sind die Suchergebnisse so gut, dass etwa die ersten zehn Ergebnisse ausreichen, um uns Websucher glücklich zu machen. Und das bei millionen von Treffern.

Das Finden passender Stellen im Web, die auf unsere Suchanfrage passen ist dabei gar nicht das zentrale Problem. Die Frage ist viel mehr, welche Ergebnisse Google seinem Nutzer präsentiert, nachdem es über eine Million Treffer gefunden hat. Welche Ergebnisse sind nun die Besten? Was interessiert den Sucher wahrscheinlich am meisten? Das sind die Fragen, die die hohe Komplexität aktueller Suchmaschinen ausmachen - aber auch deren Qualität.

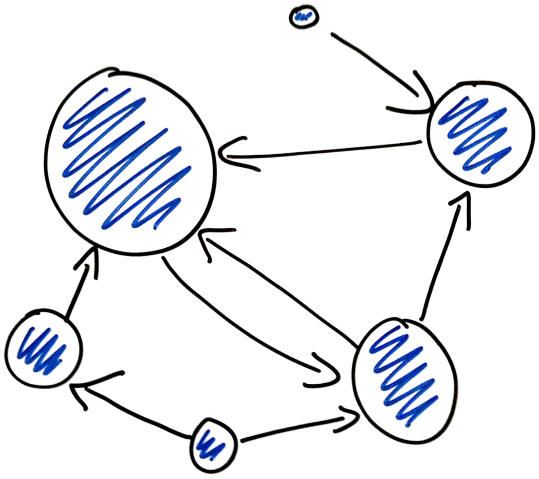
Google verwendet dafür den PageRank Algorithmus, benannt nach seinem Entdecker Larry Page. Die Idee, die Larry Page hatte war simpel: Er wollte die Linkstruktur des Netzes nutzen, um herauszufinden welche Seiten wichtiger und welche weniger wichtig sind. So könne dann eine Sortierung der Ergebnisliste entstehen.

Verlinkung ist eines der Dinge, die das Netz so besonders machen. Nie zuvor war Wissen so vernetzt wie heute.

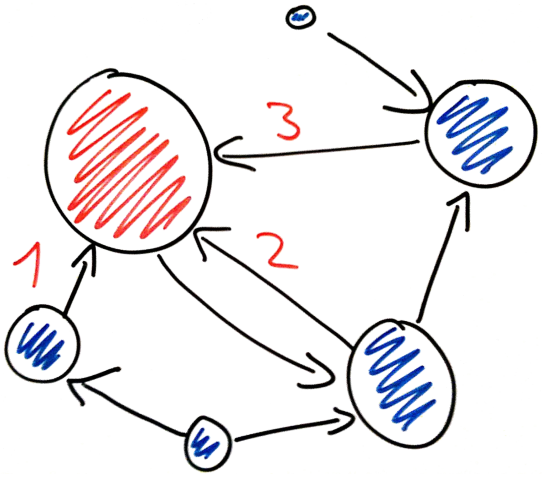
Eine Webseite verweist hier direkt auf eine weitere und der Surfer kann so auf die nächste Seite gelangen:



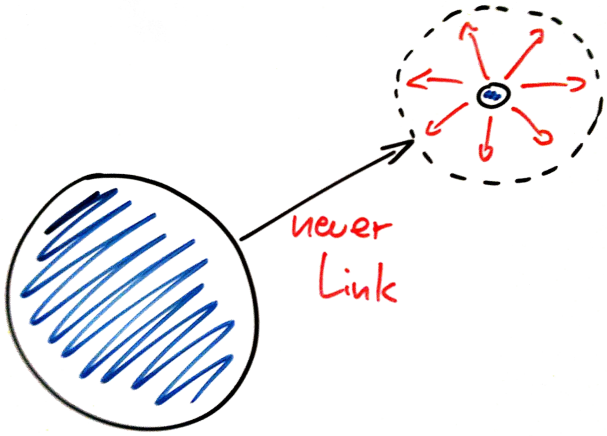
Das Internet besteht nun natürlich aus viel mehr Webseiten. Fügen wir also weitere Seiten zu unserem Modell hinzu:



Die größer gezeichneten Kreise stellen Webseiten dar, die einen höheren PageRank erhalten. Warum? Es verweisen sehr viele andere Seiten via Links dorthin. Es scheint dort also interessantes Material zu liegen, auf das Andere verweisen:

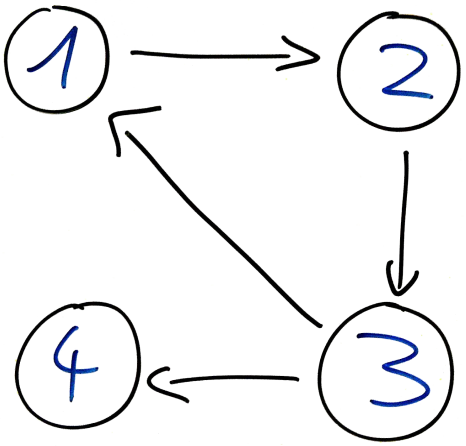


Verweist nun also eine Webseite mit einem bestimmten PageRank auf eine neue Seite, so vergrößert sich der PageRank der neuen Seite:

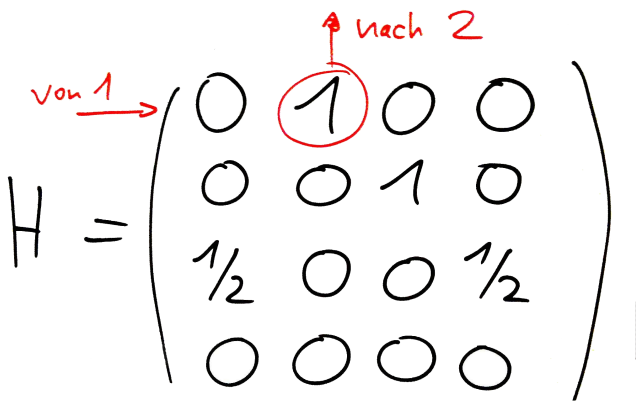


Doch wie hat Larry Page diese Idee umgesetzt? Bis hier klingt das alles noch sehr abstrakt formuliert.

Gehen wir etwas tiefer und bauen wir uns ein einfaches Beispiel-Mini-Kuschel-Internet aus lediglich vier Webseiten:



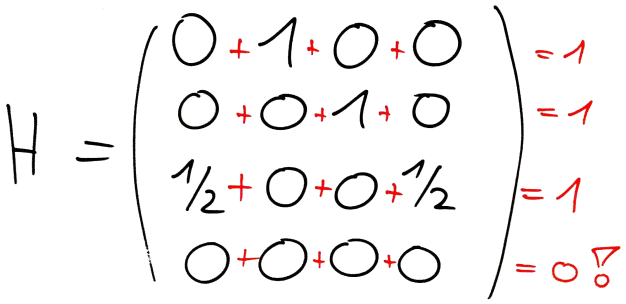
Nun wird dieser Graph zunächst einmal in eine besser handhabbare Form überführt (zumindest aus mathematischer Sicht):



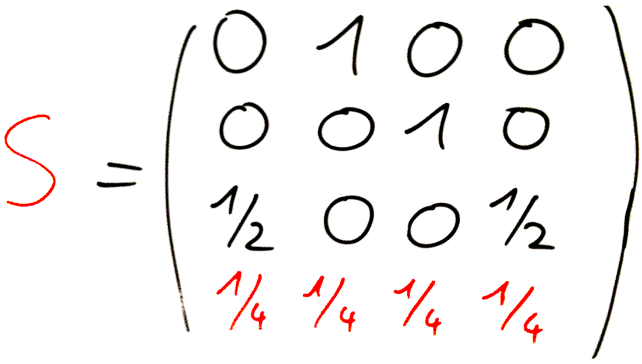
Die Matrix liest sich von links nach rechts, wie hier im Beispiel: Es existiert ein Link von Webseite 1 zu Webseite 2 und die Wahrscheinlichkeit, dass der Surfer auf Seite 2 landet ist 1 (also 100%), da Webseite 1 ja nur auf Webseite 2 verweist und sonst nirgendwohin. Der Surfer hat also keine Wahl.

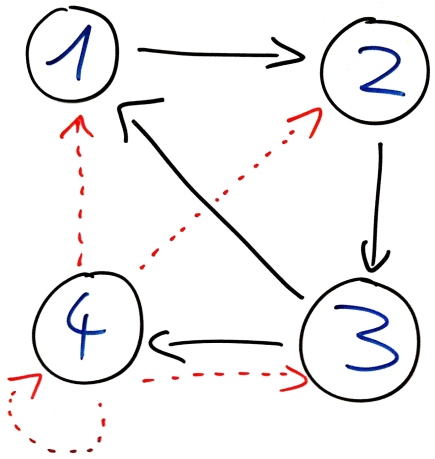
Bei Webseite 3 kann er jedoch wählen zwischen dem Link zu Webseite 1 und dem Link zu Webseite 4. Daher ist in der Matrix nur 1/2 eingetragen: Die Wahrscheinlichkeit beträgt jeweils 50%.

Wir haben nun also eine Matrix, die ausdrückt, wie wahrscheinlich es ist, dass ein Surfer von einer Webseite zu einer anderen springt, indem er die dort platzierten Links nutzt. Daher muss die Summe der Zeilen auch immer 1 ergeben:

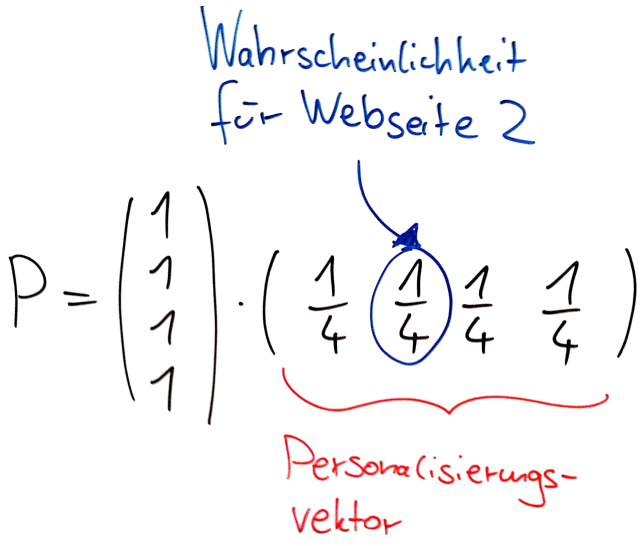


Bei Webseite 4 klappt das leider nicht, da von dort aus gar kein abgehender Link existiert. Man geht daher davon aus, dass der Surfer von dort zu einer beliebigen Webseite springen kann. Bei vier Webseiten insgesamt, macht das für jede Seite eine Wahrscheinlichkeit von 1/4.





Ein Aspekt, der nun noch betrachtet werden muss, ist das generell rein zufällige Springen des Surfers auf eine beliebige Webseite im Netz. Schließlich ist niemand gezwungen nur Links zu nutzen, sondern man kann auch direkt eine URL anspringen. Dazu verwendet man den Personalisierungsvektor v. Er enthält für jede Webseite im Netz die Wahrscheinlichkeit, dass der Surfer dorthin springt. Hier habe ich mich dazu entschieden die Wahrscheinlichkeit für alle Seiten als gleich groß zu betrachten. Was Google hier tatsächlich verwendet ist nicht bekannt.



Damit man besser damit rechnen kann, wird daraus eine Matrix P:

$$P = \begin{pmatrix} 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/4 & 1/4 & 1/4 & 1/4 \end{pmatrix}$$

Diese Personalisierungsmatrix P und die Matrix S, die wir zuvor entwickelt haben, werden nun addiert und dabei mit einem damping Faktor gewichtet.

$$G = \alpha \cdot S + (1-\alpha)P$$

damping Factor

So erhalten wir die finale Googlematrix G. Wie die Gewichtung tatsächlich gewählt ist, lässt Google nicht durchblicken. Bestehende Implementierungen, die frei verfügbar sind, wählen für α Werte zwischen 0,85 und 0,95. So lassen sich die besten Ergebnisse erzielen.

Wir erhalten also für die Matrix G:

mit $\alpha = 0{,}95$

$$G = \begin{pmatrix} \frac{1}{80} & \frac{77}{80} & \frac{1}{80} & \frac{1}{80} \\ \frac{1}{80} & \frac{1}{80} & \frac{77}{80} & \frac{1}{80} \\ \frac{39}{80} & \frac{1}{80} & \frac{1}{80} & \frac{39}{80} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix}$$

Bevor wir nun zum sehr mathematischen, aber auch spannenden Teil kommen, möchte ich noch einmal kurz resümieren, was wir mit der Matrix G nun erhalten haben. Wir haben eine Matrix, die uns darstellt, wie wahrscheinlich es ist, dass ein Surfer eine Webseite aufruft. Immer abhängig davon, wo er sich gerade befindet. Dabei haben wir berücksichtigt, dass der Surfer auf Links klickt (was wahrscheinlicher ist, daher die Gewichtung) oder dass er die Seite direkt über die URL aufruft. Wenn wir das heutige Internet betrachten, dann ist das ein sehr eingeschränkter Blickwinkel, aber wir wollen das Beispiel hier nicht unnötig kompliziert machen.

Nun also zum komplizierten Teil: Zähne zusammenbeisen!

Schauen wir uns die Matrix G noch einmal genauer an. Die Summe der Zeileneinträge ergibt immer 1 und auch alle Werte, die in der Matrix enthalten sind sind kleiner oder gleich 1. Das sind die Voraussetzungen dafür, dass wir von einer **stochastischen Matrix** sprechen.

$$G = \begin{pmatrix} \frac{1}{80} + \frac{77}{80} + \frac{1}{80} + \frac{1}{80} = 1 \\ \frac{1}{80} + \frac{1}{80} + \frac{77}{80} + \frac{1}{80} = 1 \\ \frac{39}{80} + \frac{1}{80} + \frac{1}{80} + \frac{39}{80} = 1 \\ \frac{1}{4} + \frac{1}{4} + \frac{1}{4} + \frac{1}{4} = 1 \end{pmatrix}$$

Für stochastische Matrizen gilt:

- Einer der Eigenwerte der Matrix ist 1.
- Der Eigenraum zum Eigenwert 1 hat die Dimension 1. Wir erhalten also einen Eigenvektor π.

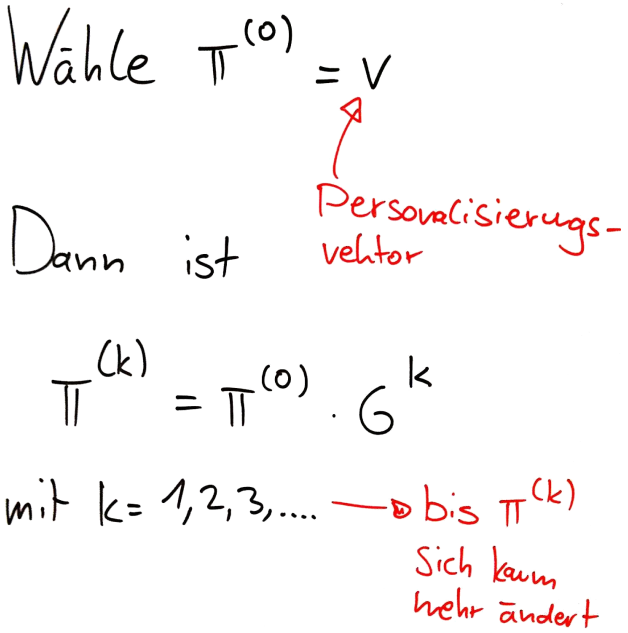
Das Eigensystem mit Matrix G und π hat also eine eindeutige Lösung.

$$G \cdot \pi = 1 \cdot \pi$$

Eigenwert, Eigenvektor

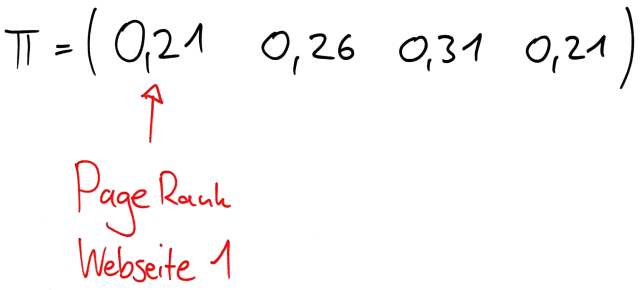
Nun gilt es nur noch den Eigenvektor π zu bestimmen. Das ist in unserem kleinen Beispiel kein Problem, denn unsere Matrix ist sehr klein. Google steht da vor einem etwas größerem Problem: Deren Matrix hat mittlerweile mehr als 25 Milliarden Zeilen.

Man behilft sich also mathematischen Tricks, um den Rechenaufwand zu reduzieren. Wir werden es hier einmal mit der **Power Methode** versuchen. Klingt auf jeden Fall schon mal sehr mächtig.



Wir erhalten unseren gesuchten Eigenvektor π also durch Konvergenz. D.h. wir multiplizieren die Matrix G solange mit sich selbst, bis sich nur noch kleine Differenzen feststellen lassen. Wenn wir uns daran erinnern wie wir die Matrix G aus S und P zusammengesetzt haben, dann könnten wir weitere Optimierungen vornehmen. Da das zum eigentlichen Verständnis des Algorithmus allerdings wenig beiträgt verzichte ich hier darauf.

Was wir jedenfalls auf vielen Wegen erreichen ist eine eindeutige Lösung für den Vektor π:



Das sind nun der Reihe nach die PageRank-Werte für unsere vier Webseiten.

Mit Hilfe der Google Toolbar ist es möglich sich den PageRank einer Webseite anzuschauen. Allerdings sind die Werte nicht direkt vergleichbar mit den Werten die uns der Algorithmus hier liefert, da Google die Ergebnisse noch einmal auf eine (wahrscheinlich - nicht offiziell bestätigt) logarithmische Skala zwischen 0 (am niedrigsten) und 10 (am höchsten) abbildet.

Es sei auch noch erwähnt, dass PageRank nicht alles ist. Ein hoher PageRank ist zwar eine gute Voraussetzung, um ganz oben in der Ergebnisliste von Google zu erscheinen, allerdings spielt es natürlich immer noch eine große Rolle, welche Stichworte der Suchende tatsächlich eingegeben hat.

Heute nutzt Google aber noch weit mehr Technologie, um bessere Suchergebnisse zu liefern. Wie genau diese Daten in die Bereitstellung der Suchergebnisse einfließen, ist natürlich Betriebsgeheimnis. Es lässt sich nur spekulieren...

Wer sich mehr für dieses Thema interessiert und ein bisschen mehr in die Details schauen möchte, dem seien folgende Quellen empfohlen:

- `http://www.cems.uvm.edu/\%7Elakobati/AppliedUGMath/other_Google/Wills.pdf` (PDF)
- `http://tinyurl.com/3dkqp7n` (PDF)
- `http://www.fmi.uni-stuttgart.de/szs/teaching/ws0607/google/material/pagerank.pdf` (PDF)
- `http://www.ams.org/samplings/feature-column/fcarc-pagerank`
- `https://secure.wikimedia.org/wikipedia/en/wiki/PageRank`