

Wie die Relevanz von Webseiten bestimmt

Alexander Pohl

Gliederung

1. Einleitung
2. Das Web als Graph
3. Das Random-Surfer-Modell
4. Gleichgewicht im Random Surfer-Modell (?)
5. Vervollständigung des Modells: die Google Matrix
6. Berechnung des PageRanks mit der Power-Methode
7. Schlussbemerkungen

Einleitung

Warum brauchen wir **Suchmaschinen**?

- Das Web besitzt keine bekannte Struktur
- Viele Daten (Dokumente) im Web sind kurzlebig
- Es gibt kein zentrales Inhaltsverzeichnis des Web und selbst wenn, wäre es wenig hilfreich, ...
- ...denn das Web ist riesig: es umfasst über **12 Milliarden Dokumente** (Stand 2009)

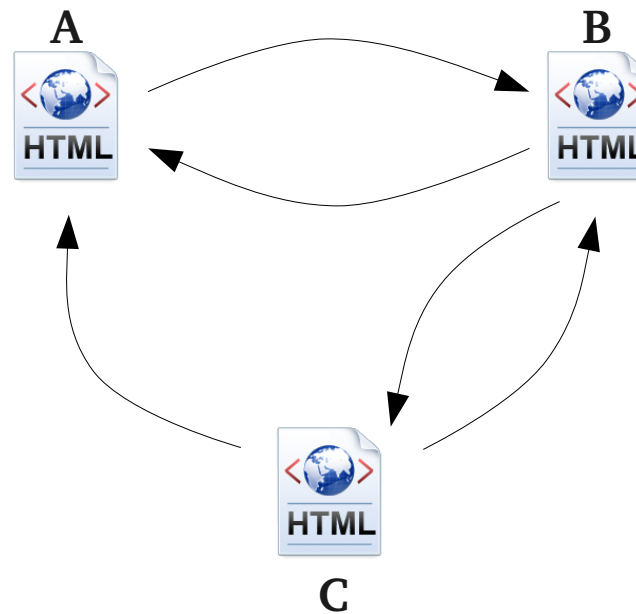
Einleitung

Bekannte Ansätze des Information Retrieval reichen für qualitativ hochwertige Antworten auf Suchanfragen im Web nicht aus.

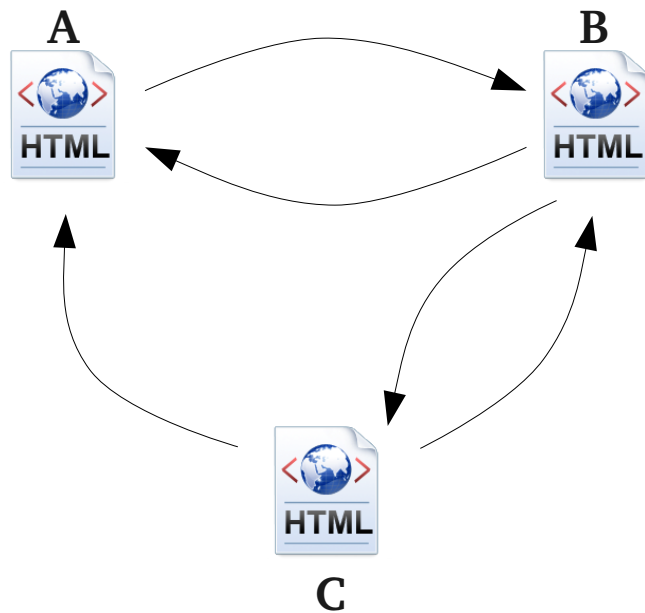
1997 wurde zwei Ansätze vorgeschlagen, wie man die Relevanz (das **Ranking**) von Webseiten anhand der **Link-Struktur** bestimmen kann:

- **HITS** (Hypertext Induced Topic Search) von *Jon Kleinberg*
- **PageRank** von *Sergey Brin* und *Larry Page* (die Gründer von Google)

Das Web als Graph



Das Web als Graph

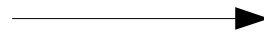


	A	B	C
A			
B			
C			

Das Web als Graphstruktur

Von der „Link-Tabelle“ zur **Adjazenz-Matrix**

	A	B	C
A	0	1	0
B	1	0	1
C	1	1	0



$$\mathbf{M} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Adjazenz-Matrix

Das Random-Surfer-Modell

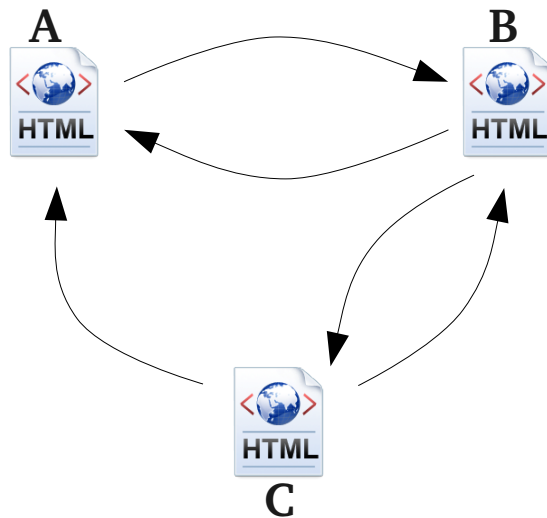
Google simuliert Web-Surfer, um die Relevanz der Webseiten zu ermitteln. Der erste Modellierungs-Ansatz ist der **Random-Surfer**:

Gegeben sei ein Web-Surfer, der in jeder **Klick-Runde** folgendes macht:

- Immer dann, wenn auf der Webseite Links vorhanden sind, folgt er einem Link, indem er auf diesen klickt
- Er tut dies ohne Vorlieben, er wählt jeden Link der Seite also mit gleicher Wahrscheinlichkeit

Das Random-Surfer-Modell

Von der Adjazenz-Matrix zur **Hyperlink-Matrix**



$$\mathbf{M} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

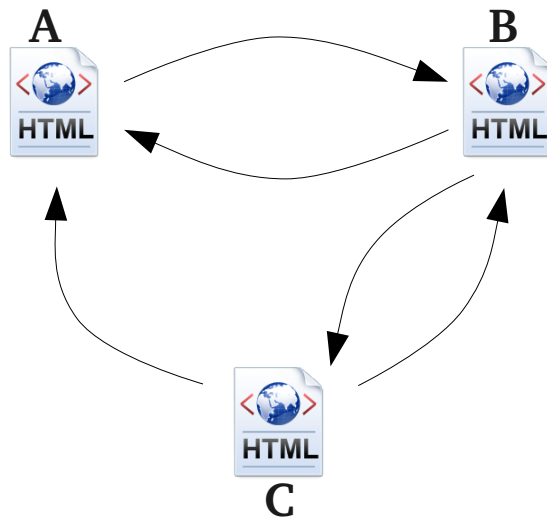
Wahrscheinlichkeit, dass der Random-Surfer von A nach B geht?

Wahrscheinlichkeit, dass der Random-Surfer von B nach C geht?

Wahrscheinlichkeit, dass der Random-Surfer von C nach A geht?

Das Random-Surfer-Modell

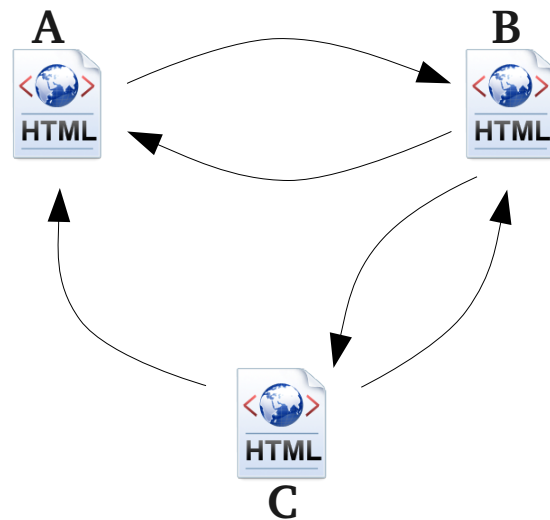
Von der Adjazenz-Matrix zur **Hyperlink-Matrix**



$$\mathbf{M} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

$$\mathbf{H} = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

Das Random-Surfer-Modell



$$\mathbf{H} = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$

Die Einträge der **Hyperlink-Matrix** kann man auf zweierlei Arten interpretieren:

1. Wahrscheinlichkeit eines Random-Surfers, mit der er einen der Links auswählt (haben wir zur Herleitung der Hyperlink-Matrix verwendet)
2. Anzahl der Random-Surfer einer Seite, die einen Link verfolgen

Das Random-Surfer-Modell

$$\mathbf{H} = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 \end{bmatrix}$$



Das Random-Surfer-Modell

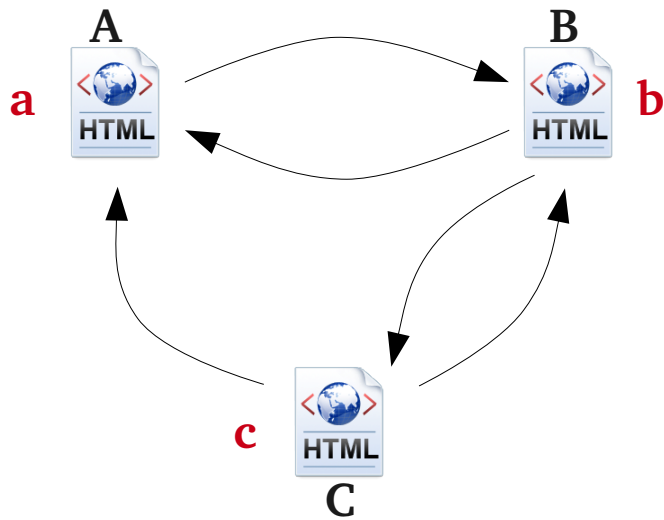


Frage: Wenn man nur ausreichend viele Klick-Runden durchführt, stellt sich dann ein Gleichgewicht ein? Anhand der (dann stabilen) Anzahl der Random-Surfer könnten wir die Relevanz (das Ranking) bestimmen!

Gleichgewicht: In einer weiteren Klick-Runde ändert sich die Anzahl der Random-Surfer auf allen Seiten nicht mehr.

Das Random-Surfer-Modell

Im Gleichgewicht gilt folgendes **Gleichungssystem**: Seien a , b und c die Anzahl der Random-Surfer auf den Seiten A, B und C.



$$a = 0 \cdot a + \frac{1}{2} \cdot b + \frac{1}{2} \cdot c$$

$$b = 1 \cdot a + 0 \cdot b + \frac{1}{2} \cdot c$$

$$c = 0 \cdot a + \frac{1}{2} \cdot b + 0 \cdot c$$

Die rechten Seiten entsprechen den eingehenden Kanten

Das Random-Surfer-Modell

Wir können die Gleichgewichts-Situation auch in Matrix-Schreibweise darstellen.

H' ist die Matrix, in der in den Zeilen die **e**ingehenden Kanten stehen (in H standen in den Zeilen die **a**usgehenden Kanten). Damit lässt sich das Gleichungssystem (mit geeigneter Verknüpfung $\circ$) wie folgt schreiben:

$$\underbrace{\begin{bmatrix} a \\ b \\ c \end{bmatrix}}_{\substack{\text{Anzahl Random-} \\ \text{Surfer in der nächsten} \\ \text{Klick-Runde}}} = \underbrace{\begin{bmatrix} 0 & \frac{1}{2} & \frac{1}{2} \\ 1 & 0 & \frac{1}{2} \\ 0 & \frac{1}{2} & 0 \end{bmatrix}}_{H'} \circ \underbrace{\begin{bmatrix} a \\ b \\ c \end{bmatrix}}_{\substack{\text{Anzahl Random-} \\ \text{Surfer in der aktuellen} \\ \text{Klick-Runde}}}$$

Das Random-Surfer-Modell

Allgemeiner:

$$\begin{bmatrix} a \\ b \\ c \\ \vdots \end{bmatrix} = \begin{bmatrix} & & & \\ & H' & & \\ & & & \\ & & & \end{bmatrix} \circ \begin{bmatrix} a \\ b \\ c \\ \vdots \end{bmatrix}$$

Zentrale Fragen sind nun aber folgende:

1. Gibt es eine Lösung dieser Gleichung mit realen und positiven Komponenten?
2. Falls ja, ist diese Lösung eindeutig?

Das Random-Surfer-Modell

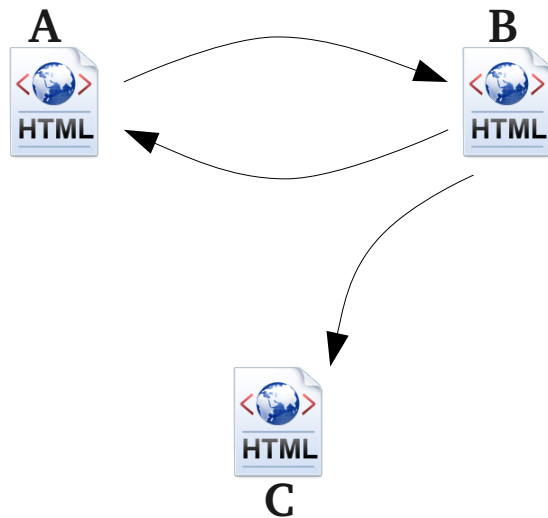
$$\begin{bmatrix} a \\ b \\ c \\ \vdots \end{bmatrix} = \begin{bmatrix} & \\ & H' \\ & \end{bmatrix} \circ \begin{bmatrix} a \\ b \\ c \\ \vdots \end{bmatrix}$$

Die Antwort liefert der Satz von Perron-Frobenius:

Es existiert eine eindeutige und positive Lösung, wenn im zugrundeliegenden Graphen alle Knoten von allen Knoten aus direkt erreichbar sind (d.h. mit einer Kante verbunden sind).

In der Link-Struktur des Web ist die erforderliche **starke Verbundenheit** aber **nicht** der Fall!

Vervollständigung des Modells: die Google Matrix

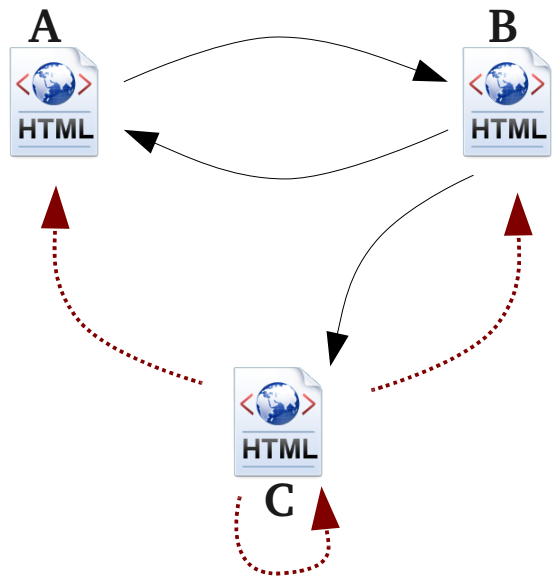


$$\mathbf{H} = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & 0 & 0 \end{bmatrix}$$

C ist eine **Senke**. Man erkennt dies in der (Adjazenz- oder) Hyperlink-Matrix daran, daß in der Zeile für C nur 0-Einträge vorhanden sind.

Was könnte wohl das Verhalten eines Random-Surfers sein, wenn er bei C ankommt?

Vervollständigung des Modells: die Google Matrix



$$\mathbf{H}_G = \begin{bmatrix} 0 & 1 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{bmatrix}$$

In der nächsten Klick-Runde springt er wieder ohne Vorlieben (d.h. gleichwahrscheinlich) auf irgendein anderes Dokument (durch Eingabe der URL, Back-Button des Browser, ...) oder er bleibt auf C.

Vervollständigung des Modells: die Google Matrix

Google verwendet letztendlich folgendes Surfer-Modell:

- Ein Surfer folgt (wie oben beschrieben) der Link-Struktur mit Wahrscheinlichkeit p (zwischen 0 und 1)
- Mit der Komplementär-Wahrscheinlichkeit $(1-p)$ beendet er das Verfolgen der Link-Struktur und springt (ohne Vorlieben) zu irgendeinem Dokument im Web.

Die Google-Matrix beschreibt Wege (entlang der Link-Struktur und zufällige Sprünge) im Web-Graphen. Der der Google-Matrix zugrundeliegende Web-Graph ist stark verbunden.

Bemerkenswert: Dieses Modell ist äußerst intuitiv und erfüllt zugleich alle Voraussetzungen, die im Satz von Frobenius-Perron genannt werden. D.h. mit der Google-Matrix existiert eine positive und eindeutige Lösung!

Vervollständigung des Modells: die Google Matrix

Die Google-Matrix setzt sich wie folgt zusammen:

$$\mathbf{G} = (1 - p)\mathbf{H}'_{\mathbf{G}} + p \underbrace{\begin{bmatrix} \frac{1}{n} & \cdots & \frac{1}{n} \\ \vdots & \ddots & \vdots \\ \frac{1}{n} & \cdots & \frac{1}{n} \end{bmatrix}}_{\text{Personalisierungsmatrix}}$$

p heißt **Random Leap**-Faktor („zufälliges Herumspringen“), $\mathbf{H}_{\mathbf{G}}$ ist die Hyperlink-Matrix, mit „Behebung von Senken“, in $\mathbf{H}'_{\mathbf{G}}$ stehen die eingehenden Kanten in der Zeile und in der **Personalisierungsmatrix** stehen die Wahrscheinlichkeiten, mit welchem ein Surfer von einer Seite zu einer anderen springt.

Berechnung des PageRanks mit der Power-Methode

$$\underbrace{\begin{bmatrix} a \\ b \\ c \\ \vdots \end{bmatrix}}_{\mathbf{v}} = \begin{bmatrix} & \\ & \mathbf{G} & \\ & & \end{bmatrix} \circ \underbrace{\begin{bmatrix} a \\ b \\ c \\ \vdots \end{bmatrix}}_{\mathbf{v}}$$

Wir versuchen immernoch, das obige Gleichungssystem zu lösen. Bei der Power-Methode machen wir folgenden Ansatz.

$$\mathbf{v}_{(t+1)} = \mathbf{G} \circ \mathbf{v}_t$$

Berechnung des PageRanks mit der Power-Methode

$$\mathbf{v}_{(t+1)} = \mathbf{G} \circ \mathbf{v}_t$$

Einmal Auswickeln der rechten Seite

$$\mathbf{v}_{(t+1)} = \mathbf{G} \circ \mathbf{v}_t = \mathbf{G} \circ (\mathbf{G} \circ \mathbf{v}_{(t-1)})$$

Umklammern (Voraussetzung: $\circ$ ist assoziativ)

$$\mathbf{v}_{(t+1)} = \mathbf{G} \circ \mathbf{v}_t = (\mathbf{G} \circ \mathbf{G}) \circ \mathbf{v}_{(t-1)} = \underbrace{\mathbf{G}^2}_{\text{alle Wege der Länge 2 durch den Graphen, der durch G beschrieben ist}} \circ \mathbf{v}_{(t-1)}$$

Alle Wege der Länge 2 durch den Graphen, der durch G beschrieben ist

Berechnung des PageRanks mit der Power-Methode

Kontinuierliches Auswickeln der rechten Seite ergibt dann:

$$\mathbf{v} = \lim_{t \rightarrow \infty} \mathbf{G}^t \circ \mathbf{v}_0$$

Es stellt sich ein Gleichgewicht ein, wenn man nur lange genug durch den Graphen, der durch G beschrieben ist, umherwandert (entlang der Linkstruktur oder beliebige Sprünge)!

Schlussbemerkung

Die Theorie der Matrizen (und Vektoren) studiert ihr in der Vorlesung **Lineare Algebra**

Die Lineare Algebra ist ein sehr wichtiges Gebiet (auch für die Informatik):

- Data Mining, Information Retrieval
- Social Network Analysis
- Optimierung, Risikoabschätzung
- usw.

→ **Vorlesung Webinformationssysteme** von Prof. Bry

Quellen

- **Wie Google Webseiten bewertet**, Vortrags-Folien von *François Bry*
- **Web Information Systems**, Vorlesungsskriptum der Vorlesung Webinformationssysteme von *François Bry*, *Norbert Eisinger* und *Tim Furché*
- **Google's PageRank: The Math Behind the Search Engine**, *Rebecca S. Wills*