Googles PageRank-Verfahren

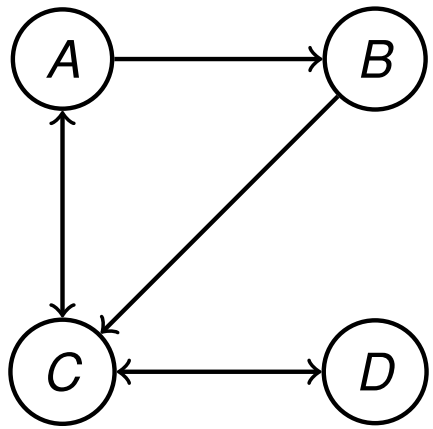
TECHNISCHE UNIVERSITÄT DARMSTADT

- Hauptkriterium, nach dem die Suchmaschine Google die Ergebnisse einer Suchanfrage ordnet
  - Maß für die „Wichtigkeit“ einer Webseite
- Methode beruht auf dem Konzept der *Markovketten* aus der Stochastik
- Sergey Brin & Lawrence Page: The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 33:107–117, 1998.

WWW mit vier Seiten

TECHNISCHE UNIVERSITÄT DARMSTADT

Betrachte eine Webstruktur aus vier Seiten *A*, *B*, *C*, *D*, die wie folgt verlinkt sind:



- modelliere Verhalten eines zufälligen Websurfers ...
- und bestimme wie oft sich der Surfer auf welcher Seite im Erwartungswert aufhält

Wahrscheinlichkeit $\mathrm{P}_t(X)$

TECHNISCHE UNIVERSITÄT DARMSTADT

Wir bezeichnen die Wahrscheinlichkeit, dass sich der Surfer zu einem diskreten Zeitpunkt $t \in \mathbb{N}$ auf der Webseite X aufhält mit $\mathrm{P}_t(X)$. Die Wahrscheinlichkeit auf einer der vier Seiten zu beginnen, beträgt jeweils $1/4$:

$$\mathrm{P}_0(A) = \mathrm{P}_0(B) = \mathrm{P}_0(C) = \mathrm{P}_0(D) = 1/4\,.$$

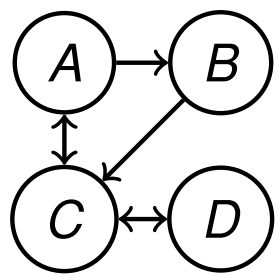
Im nächsten Schritt kann etwa die Seite C von A oder von B aus erreicht werden, wobei A zwei Links hat und B nur eins. Also gilt

$$\mathrm{P}_{t+1}(C) = \frac{\mathrm{P}_t(A)}{2} + \frac{\mathrm{P}_t(B)}{1} + \frac{\mathrm{P}_t(D)}{1}\,,$$

d.h.

$$\mathrm{P}_1(C) = 1/8 + 1/4 + 1/4 = 5/8$$

und entsprechend für die übrigen Seiten.



Wahrscheinlichkeit $\mathrm{P}_t(X)$

TECHNISCHE UNIVERSITÄT DARMSTADT

Wir bezeichnen die Wahrscheinlichkeit, dass sich der Surfer zu einem diskreten Zeitpunkt $t \in \mathbb{N}$ auf der Webseite X aufhält mit $\mathrm{P}_t(X)$. Die Wahrscheinlichkeit auf einer der vier Seiten zu beginnen, beträgt jeweils $1/4$:

$$\mathrm{P}_0(A) \;=\; \mathrm{P}_0(B) \;=\; \mathrm{P}_0(C) \;=\; \mathrm{P}_0(D) \;=\; 1/4\,.$$

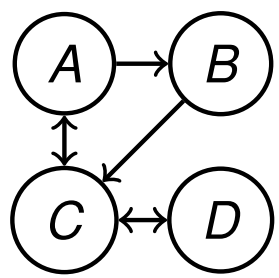
Im nächsten Schritt kann etwa die Seite C von A oder von B aus erreicht werden, wobei A zwei Links hat und B nur eins. Also gilt

$$\mathrm{P}_{t+1}(C) \;=\; \frac{\mathrm{P}_t(A)}{2} + \frac{\mathrm{P}_t(B)}{1} + \frac{\mathrm{P}_t(D)}{1}\,,$$

d.h.

$$\mathrm{P}_1(C) \;=\; 1/8 + 1/4 + 1/4 \;=\; 5/8$$

und entsprechend für die übrigen Seiten.



Wahrscheinlichkeit $\mathrm{P}_t(X)$

TECHNISCHE UNIVERSITÄT DARMSTADT

Wir bezeichnen die Wahrscheinlichkeit, dass sich der Surfer zu einem diskreten Zeitpunkt $t \in \mathbb{N}$ auf der Webseite X aufhält mit $\mathrm{P}_t(X)$. Die Wahrscheinlichkeit auf einer der vier Seiten zu beginnen, beträgt jeweils $1/4$:

$$\mathrm{P}_0(A) \;=\; \mathrm{P}_0(B) \;=\; \mathrm{P}_0(C) \;=\; \mathrm{P}_0(D) \;=\; 1/4\,.$$

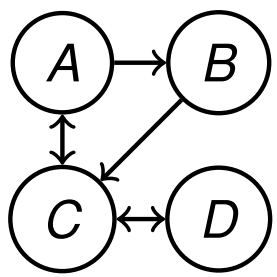
Im nächsten Schritt kann etwa die Seite C von A oder von B aus erreicht werden, wobei A zwei Links hat und B nur eins. Also gilt

$$\mathrm{P}_{t+1}(C) \;=\; \frac{\mathrm{P}_t(A)}{2} + \frac{\mathrm{P}_t(B)}{1} + \frac{\mathrm{P}_t(D)}{1}\,,$$

d.h.

$$\mathrm{P}_1(C) \;=\; 1/8 + 1/4 + 1/4 \;=\; 5/8$$

und entsprechend für die übrigen Seiten.



Wahrscheinlichkeit $\mathrm{P}_t(X)$

TECHNISCHE UNIVERSITÄT DARMSTADT

Wir bezeichnen die Wahrscheinlichkeit, dass sich der Surfer zu einem diskreten Zeitpunkt $t \in \mathbb{N}$ auf der Webseite X aufhält mit $\mathrm{P}_t(X)$. Die Wahrscheinlichkeit auf einer der vier Seiten zu beginnen, beträgt jeweils $1/4$:

$$\mathrm{P}_0(A) \;=\; \mathrm{P}_0(B) \;=\; \mathrm{P}_0(C) \;=\; \mathrm{P}_0(D) \;=\; 1/4\,.$$

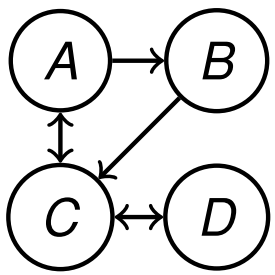
Im nächsten Schritt kann etwa die Seite C von A oder von B aus erreicht werden, wobei A zwei Links hat und B nur eins. Also gilt

$$\mathrm{P}_{t+1}(C) \;=\; \frac{\mathrm{P}_t(A)}{2} + \frac{\mathrm{P}_t(B)}{1} + \frac{\mathrm{P}_t(D)}{1}\,,$$

d.h.

$$\mathrm{P}_1(C) \;=\; 1/8 + 1/4 + 1/4 \;=\; 5/8$$

und entsprechend für die übrigen Seiten.



Wahrscheinlichkeit $\mathrm{P}_t(X)$

TECHNISCHE UNIVERSITÄT DARMSTADT

Wir bezeichnen die Wahrscheinlichkeit, dass sich der Surfer zu einem diskreten Zeitpunkt $t \in \mathbb{N}$ auf der Webseite X aufhält mit $\mathrm{P}_t(X)$. Die Wahrscheinlichkeit auf einer der vier Seiten zu beginnen, beträgt jeweils $1/4$:

$$\mathrm{P}_0(A) \;=\; \mathrm{P}_0(B) \;=\; \mathrm{P}_0(C) \;=\; \mathrm{P}_0(D) \;=\; 1/4\,.$$

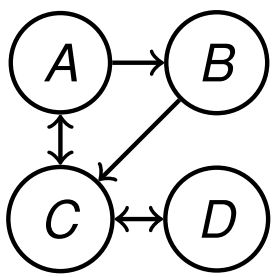
Im nächsten Schritt kann etwa die Seite C von A oder von B aus erreicht werden, wobei A zwei Links hat und B nur eins. Also gilt

$$\mathrm{P}_{t+1}(C) \;=\; \frac{\mathrm{P}_t(A)}{2} + \frac{\mathrm{P}_t(B)}{1} + \frac{\mathrm{P}_t(D)}{1}\,,$$

d.h.

$$\mathrm{P}_1(C) \;=\; 1/8 + 1/4 + 1/4 \;=\; 5/8$$

und entsprechend für die übrigen Seiten.



Schema

TECHNISCHE UNIVERSITÄT DARMSTADT

Für Webseiten $p_1, p_2, \ldots, p_n$ entsteht dann folgendes Rechenschema:

$$\mathrm{P}_{t+1}(p_k) \;=\; \frac{1-d}{n} + d \sum_{p_i \in M(p_k)} \frac{\mathrm{P}_t(p_i)}{\ell(p_i)}$$

wobei $M(p_k)$ die Menge der Seiten von $p_1, \ldots, p_n$ ist, die einen Link auf p_k gesetzt haben, und

$$\ell(p_i) \;=\; \#\text{Links von } p_i \text{ auf andere Seiten}\,.$$

Empirisch: $d = 0.85$

Zurück zum Beispiel

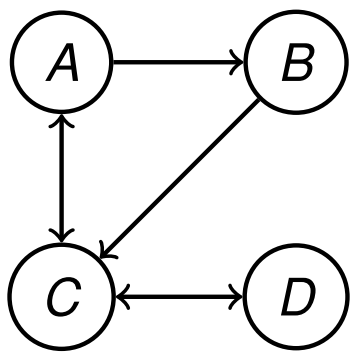
TECHNISCHE UNIVERSITÄT DARMSTADT

In unserem Beispielweb ist

$$M(A) = \{C\}\,, \quad M(B) = \{A\}\,, \quad M(C) = \{A, B, D\}\,, \quad M(D) = \{C\}$$

und

$$\ell(A) = 2\,, \quad \ell(B) = 1\,, \quad \ell(C) = 2\,, \quad \ell(D) = 1\,.$$



Das Verhalten des Surfers im „Unendlichen“

TECHNISCHE UNIVERSITÄT DARMSTADT

- vorige Formel berücksichtigt nicht direkt, dass eine Seite ohne ausgehende Links wiederum mit einer Zufallsseite startet
- dies lässt sich (leicht) beheben

Damit kann man kann dann zeigen, dass der *Grenzwert*

$$\lim_{t \to \infty} \mathrm{P}_t(p_k) \;=:\; \mathrm{P}(p_k)$$

existiert.

Interpretiere diesen Grenzwert als „Wichtigkeit“ der Webseite.

Was hat das mit linearer Algebra zu tun?

TECHNISCHE UNIVERSITÄT DARMSTADT

Setze

$$R := \begin{pmatrix} \mathrm{P}(p_1) \\ \vdots \\ \mathrm{P}(p_n) \end{pmatrix} \in \mathbb{R}^n .$$

Dann gilt

$$R = \begin{pmatrix} (1-d)/n \\ \vdots \\ (1-d)/n \end{pmatrix} + d \begin{pmatrix} s_{11} & \dots & s_{1n} \\ \vdots & \ddots & \vdots \\ s_{n1} & \dots & s_{nn} \end{pmatrix} \cdot R , \tag{1}$$

wobei

$$s_{ki} = \begin{cases} 1/\ell(p_i) & \text{falls } p_i \in M(p_k) \\ 0 & \text{sonst} . \end{cases}$$

Eine stochastische Matrix

TECHNISCHE UNIVERSITÄT DARMSTADT

Für alle $i \in \{1, 2, \ldots, n\}$ gilt, dass

$$\sum_{k=1}^{n} s_{ki} = 1 .$$

Anders ausgedrückt, jede Spaltensumme der Matrix $S = (s_{ki})$ beträgt 1 (oder 0 für Seiten ohne Links). In unserem Beispiel ist

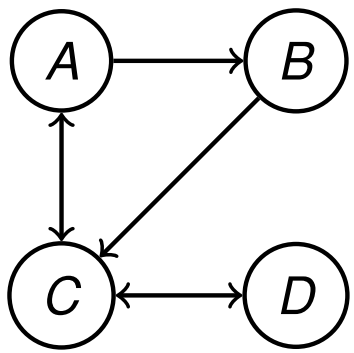
$$S = \begin{pmatrix} 0 & 0 & 1/2 & 0 \\ 1/2 & 0 & 0 & 0 \\ 1/2 & 1 & 0 & 1 \\ 0 & 0 & 1/2 & 0 \end{pmatrix} .$$

Eine stochastische Matrix

und ein Eigenvektor zum Eigenwert 1

TECHNISCHE UNIVERSITÄT DARMSTADT

- ohne Dämpfung (d.h. für $d = 1$) besagt Formel (1), dass R Eigenvektor der Matrix S zum Eigenwert 1 ist
- Im Beispiel hat S die Eigenwerte 1, 0 und $-1/2$. Der Eigenraum zum Eigenwert 1 ist eindimensional, und ein Eigenvektor ist $(2, 1, 4, 2)^T$.
- skaliere mit $1/9$, um dies als Vektor von Wahrscheinlichkeiten zu interpretieren



1. C
2. A, D
3. B

Zum Schluss noch mit Dämpfung

TECHNISCHE UNIVERSITÄT DARMSTADT

Will man die Dämpfung einbeziehen, so stellt man fest, dass (1) äquivalent ist zu

$$\begin{pmatrix} 1 \\ R \end{pmatrix} = \begin{pmatrix} 1 & 0 & \dots & 0 \\ (1-d)/n & ds_{11} & \dots & ds_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ (1-d)/n & ds_{n1} & \dots & ds_{nn} \end{pmatrix} \cdot \begin{pmatrix} 1 \\ R \end{pmatrix}$$

d.h. $\left(\begin{smallmatrix} 1 \\ R \end{smallmatrix}\right)$ ist Eigenvektor der $(n+1) \times (n+1)$-Matrix

$$\begin{pmatrix} 1 & 0 & \dots & 0 \\ (1-d)/n & ds_{11} & \dots & ds_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ (1-d)/n & ds_{n1} & \dots & ds_{nn} \end{pmatrix}$$

zum Eigenwert 1.