

PageRank

Laura Lehmann

Alina Fastowski

LMU München, CIS

11.02.21

Überblick

- 1) Einführung
- 2) PageRank:
 - Was ist die Idee?
 - Berechnung
 - Beispiele
- 3) Initialisierung und Random Surfer Modell
- 4) Abschließendes

Was? Wer? Warum?

- PageRank: Algorithmus zur „Bewertung“ von Websites
 - > welche sind relevant und wichtig?
- Ende 90er Jahre entwickelt von Larry Page und Sergey Brin
- Idee: Bewertung anhand der Verlinkungsstruktur. Denn:
Wenn man in Literatur Zitaten folgt, wird man feststellen, dass die relevantesten Werke am meisten zitiert werden.
Also müssen oft verlinkte Websites auch irgendwie relevant sein, oder?

**The PageRank Citation Ranking:
Bringing Order to the Web**

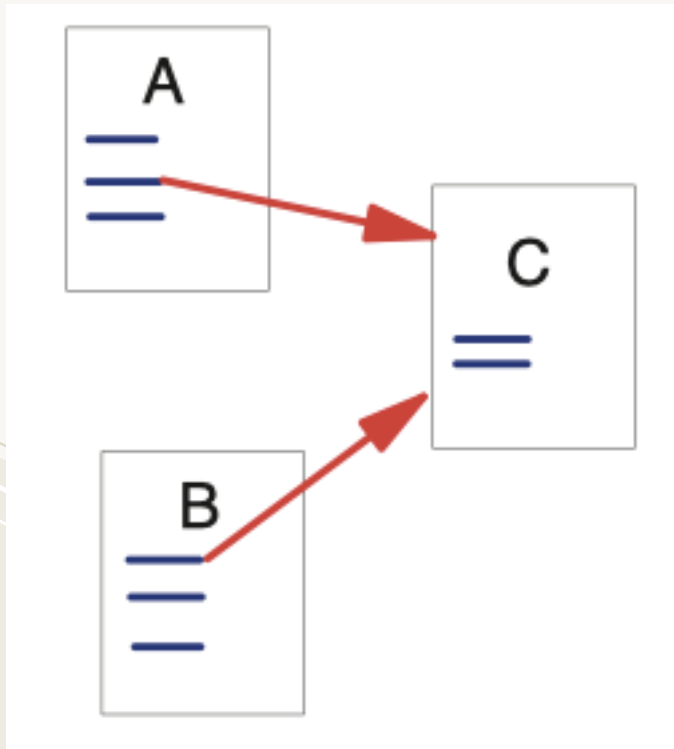
January 29, 1998



Anstoß für Erschaffung von Google

Die Idee

Inhalt von Websites irrelevant – es geht nur um die Links zwischen ihnen.



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

- $PR(A)$ = PageRank einer Seite A
- $PR(T_i)$ = PageRank der Seiten T_i , welche auf A verlinken
- $C(T_i)$ = # Links auf Seite T_i
- d = Dämpfungsfaktor
- N = Gesamtzahl aller Seiten

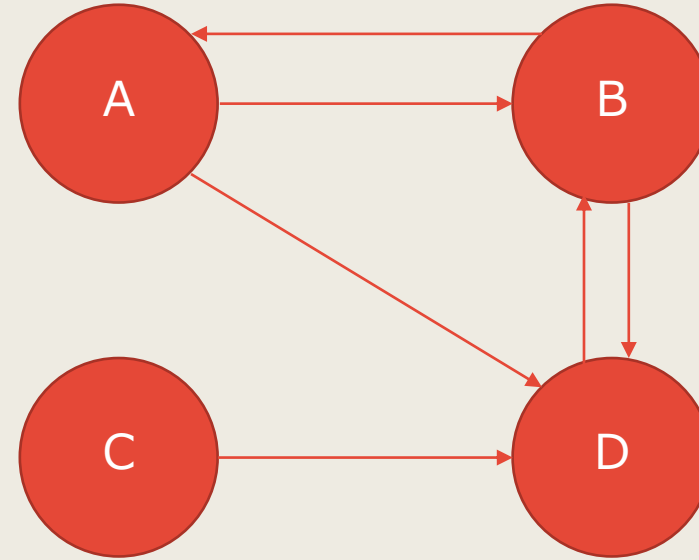
$$(PR(T_1)/C(T_1) + \dots + PR(T_n)/C(T_n))$$

=

„Anteil“ des PR jeder Seite, die auf A verlinkt

PR = Wert zwischen 0 und 1

Beispiel



$$PR(A) = \frac{(1-d)}{4} + d (PR(B)/2)$$

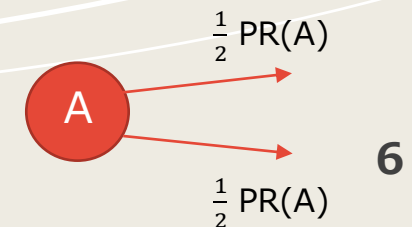
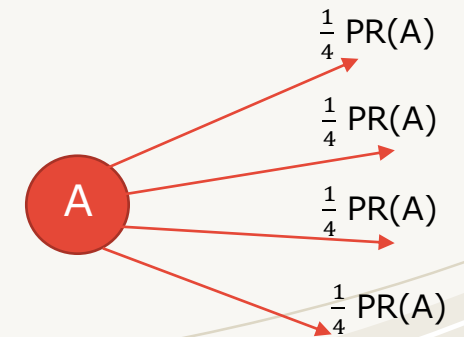
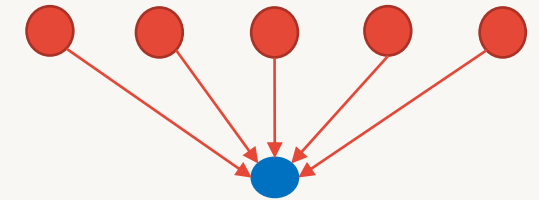
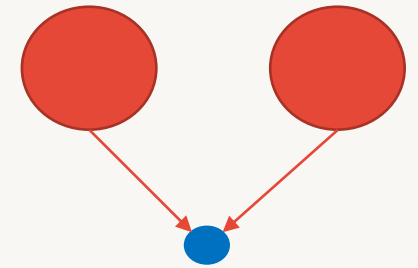
$$PR(B) = \frac{(1-d)}{4} + d (PR(A)/2 + PR(D)/1)$$

$$PR(C) = \frac{(1-d)}{4} + d \times 0$$

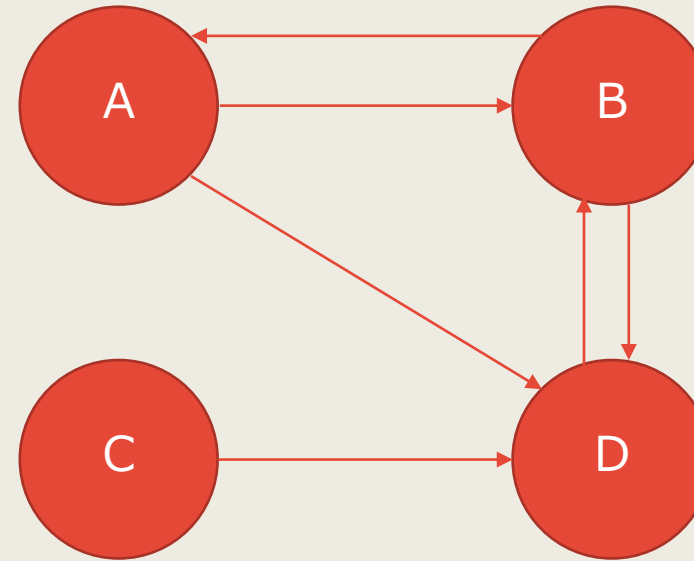
$$PR(D) = \frac{(1-d)}{4} + d (PR(A)/2 + PR(B)/2 + PR(C)/1)$$

Beobachtungen:

- Seite kann hohen PR erreichen, wenn wenige Seiten mit jeweils hohen PR's auf sie verlinken
- Seite kann hohen PR erreichen, wenn viele Seiten mit jeweils niedrigen PR's auf sie verlinken
- Je mehr ausgehende Links auf einer Seite, desto weniger PR kann sie „weitergeben“
- Je weniger ausgehende Links, desto mehr PR kann sie an andere Seiten „weitergeben“



Problem

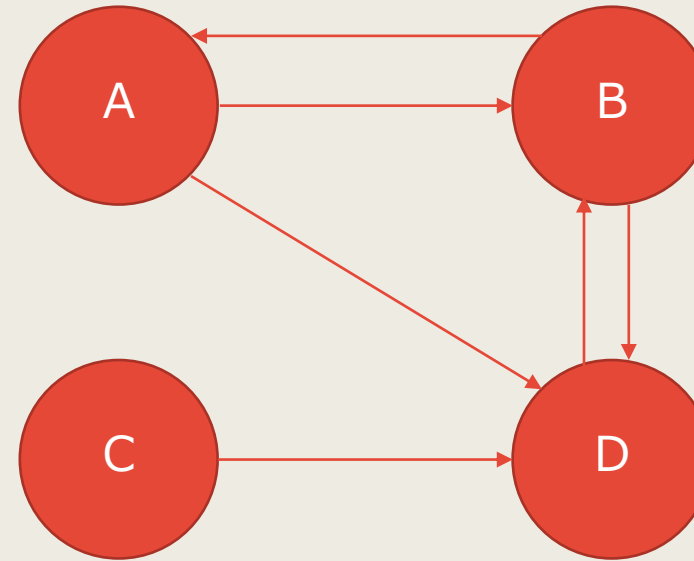


$$PR(A) = \frac{(1-d)}{4} + d (PR(B)/2)$$

PR(A) basiert auf PR(B), welcher wiederum auf anderen PR's basieren usw.

Wie erhält man nun den PageRank einer Seite?

Random Surfer Model



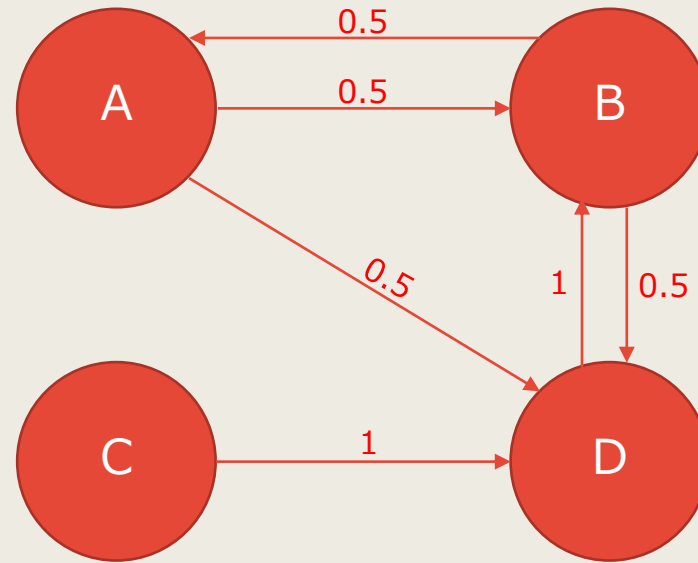
Grundidee:

- Mehrfaches zufälliges Klicken auf Links von Websites
- Auf welcher Seite wird man am wahrscheinlichsten laden?

Intuitiv:

- C am unwichtigsten
- D und B am wichtigsten

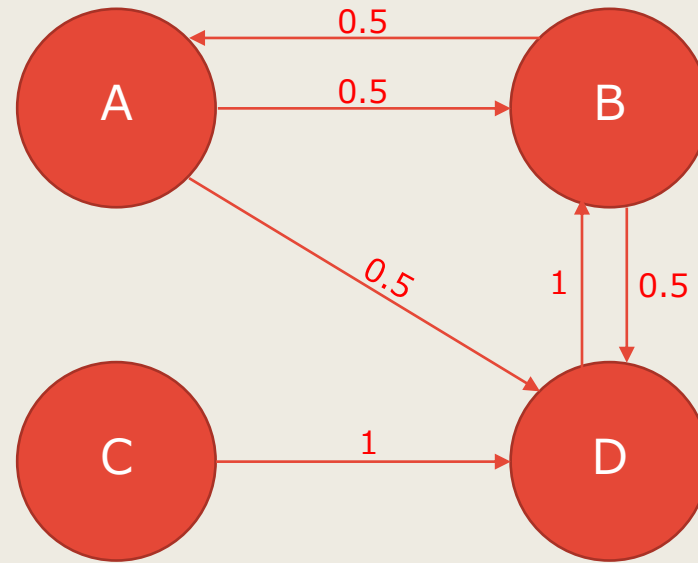
Random Surfer Model



Links zwischen Seiten in Matrixform:

$$M := \begin{array}{c} \begin{array}{cc} & \text{Von} \\ \begin{array}{c} A \\ B \\ C \\ D \end{array} & \begin{array}{cccc} A & B & C & D \end{array} \\ \begin{bmatrix} 0 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0 \end{bmatrix} & \begin{array}{c} A \\ B \\ C \\ D \end{array} \end{array} \begin{array}{c} \text{Nach} \end{array}$$

Random Surfer Model



Links zwischen Seiten in Matrixform:

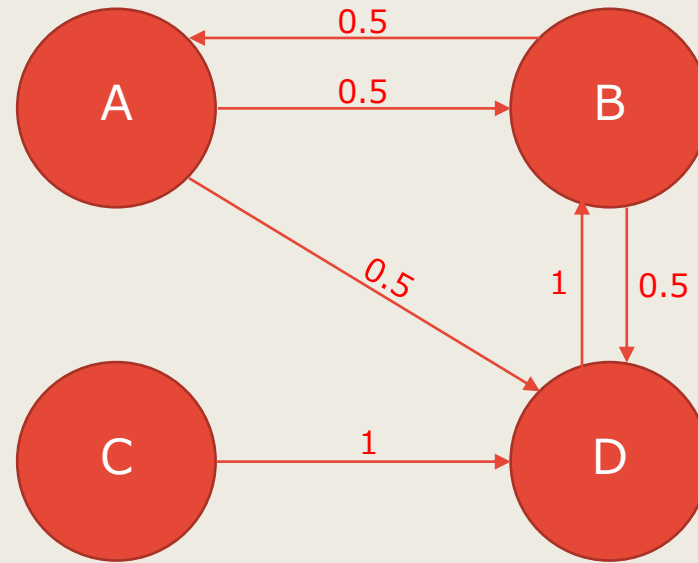
$$M := \begin{matrix} & \begin{matrix} \text{Von} \\ A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} \text{ Nach} & \begin{bmatrix} 0 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0 \end{bmatrix} \end{matrix}$$

$$(m)_{_A} := \begin{bmatrix} 0 \\ 0.5 \\ 0 \\ 0.5 \end{bmatrix} \quad (m)_{A_} := \begin{bmatrix} 0 \\ 0.5 \\ 0 \\ 0 \end{bmatrix}$$

$$(m)_{AA} := (m)_{_A} * (m)_{A_} := 0 * 0 + 0.5 * 0.5 + 0 * 0 + 0.5 * 0 = 0.25$$

$$M^2 := \begin{matrix} & \begin{matrix} \text{Von} \\ A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} \text{ Nach} & \begin{bmatrix} 0.25 & m2_{AB} & m2_{AC} & m2_{AD} \\ m2_{BA} & m2_{BB} & m2_{BC} & m2_{BD} \\ m2_{CA} & m2_{CB} & m2_{CC} & m2_{CD} \\ m2_{DA} & m2_{DB} & m2_{DC} & m2_{DD} \end{bmatrix} \end{matrix}$$

Random Surfer Model



Links zwischen Seiten in Matrixform:

$$M := \begin{array}{c} \text{Von} \\ \begin{matrix} A & B & C & D \end{matrix} \\ \begin{bmatrix} 0 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0 \end{bmatrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} \end{array} \quad \text{Nach}$$

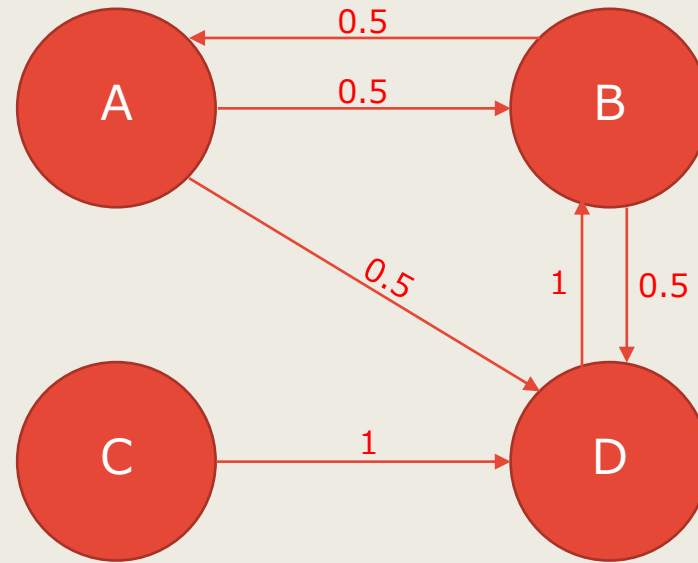
$$(m)_{_A} := \begin{bmatrix} 0 \\ 0.5 \\ 0 \\ 0.5 \end{bmatrix} \quad (m)_{A_} := \begin{bmatrix} 0 \\ 0.5 \\ 0 \\ 0 \end{bmatrix}$$

$$(m)_{AA} := (m)_{_A} * (m)_{A_} := 0 * 0 + 0.5 * 0.5 + 0 * 0 + 0.5 * 0 = 0.25$$

$$M^2 := \begin{array}{c} \text{Von} \\ \begin{matrix} A & B & C & D \end{matrix} \\ \begin{bmatrix} 0.25 & 0 & 0 & 0.5 \\ 0.5 & 0.75 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0 & 0.5 \end{bmatrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} \end{array} \quad \text{Nach}$$

$$M^3 := \begin{array}{c} \text{Von} \\ \begin{matrix} A & B & C & D \end{matrix} \\ \begin{bmatrix} 0.25 & 0.375 & 0.5 & 0 \\ 0.375 & 0.25 & 0 & 0.75 \\ 0 & 0 & 0 & 0 \\ 0.375 & 0.375 & 0.5 & 0.25 \end{bmatrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} \end{array} \quad \text{Nach}$$

Random Surfer Model



Links zwischen Seiten in Matrixform:

$$M := \begin{matrix} & \begin{matrix} \text{Von} \\ A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} \text{ Nach} & \begin{bmatrix} 0 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0 \end{bmatrix} \end{matrix}$$

Vorgehen:

- Potenzieren der Matrix
- Jede weitere Potenz ist ein weiterer zufälliger Klick des Surfers

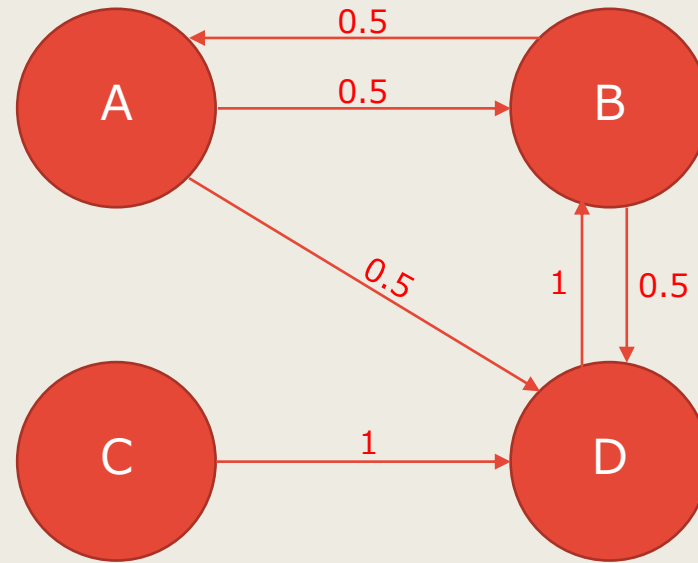
Beispiel:

von B kommend, befindet sich der Surfer nach 3 Klicks mit einer Wahrscheinlichkeit von 0.375 in A, 0.25 in B, 0.0 in C und 0.375 in D

$$M^2 := \begin{matrix} & \begin{matrix} \text{Von} \\ A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} \text{ Nach} & \begin{bmatrix} 0.25 & 0 & 0 & 0.5 \\ 0.5 & 0.75 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0.25 & 0.25 & 0 & 0.5 \end{bmatrix} \end{matrix}$$

$$M^3 := \begin{matrix} & \begin{matrix} \text{Von} \\ A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} \text{ Nach} & \begin{bmatrix} 0.25 & 0.375 & 0.5 & 0 \\ 0.375 & 0.25 & 0 & 0.75 \\ 0 & 0 & 0 & 0 \\ 0.375 & 0.375 & 0.5 & 0.25 \end{bmatrix} \end{matrix}$$

Random Surfer Model



Links zwischen Seiten in Matrixform:

$$M := \begin{matrix} & \begin{matrix} \text{Von} \\ A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} \text{ Nach} & \begin{bmatrix} 0 & 0.5 & 0 & 0 \\ 0.5 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \\ 0.5 & 0.5 & 1 & 0 \end{bmatrix} \end{matrix}$$

Beobachtung:

- Je öfter der Surfer von Seite zu Seite springt, desto mehr pendeln sich die Wahrscheinlichkeiten ein
- Seite B ist die wahrscheinlichste und damit informativste Seite

$$M^{10} := \begin{matrix} & \begin{matrix} \text{Von} \\ A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} \text{ Nach} & \begin{bmatrix} 0.22 & 0.22 & 0.21 & 0.23 \\ 0.45 & 0.45 & 0.46 & 0.44 \\ 0 & 0 & 0 & 0 \\ 0.33 & 0.33 & 0.33 & 0.33 \end{bmatrix} \end{matrix}$$

$$M^{20} := \begin{matrix} & \begin{matrix} \text{Von} \\ A & B & C & D \end{matrix} \\ \begin{matrix} A \\ B \\ C \\ D \end{matrix} \text{ Nach} & \begin{bmatrix} 0.22 & 0.22 & 0.22 & 0.22 \\ 0.44 & 0.44 & 0.44 & 0.44 \\ 0 & 0 & 0 & 0 \\ 0.33 & 0.33 & 0.33 & 0.33 \end{bmatrix} \end{matrix}$$

Vorteile und Nachteile

Vorteile:


- Möglichkeit von Modifikation
- Dient als Grundlage zur Seitenbewertung
- weitere Faktoren bspw. Bewertungen/Gewichtung von Links, maximale Distanz auf weitere Seiten, Nutzerfreundlichkeit, Ladegeschwindigkeit

Nachteile:

- Bietet viele Möglichkeiten zur Manipulation
- Link - "Tauschbörse"
- Künstliches Aufblähen von eigentlich irrelevanten Seiten
- Je mehr Faktoren mit einbezogen, desto höher der Rechenaufwand
- Keine Aussage über inhaltliche Qualität



PageRank anfangs die Basis für die Suchmaschine Google
Heute: PageRank spielt nur noch eine kleine Rolle und ist nur noch einer von vielen Parametern beim Ranking von Seiten



**Danke für Eure
Aufmerksamkeit!**

Quellen

- Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The PageRank citation ranking: Bringing order to the web*. Stanford InfoLab.
- Bianchini, M., Gori, M., & Scarselli, F. (2005). Inside pagerank. *ACM Transactions on Internet Technology (TOIT)*, 5(1), 92-128.
- <https://webworkshop.net/pagerank/>