Proseminar: Matrixmethoden in Datenanalyse und Mustererkennung

Page Ranking for a Web Search Engine

REFERENT: SIMON PETER

BETREUER: SARAH SCHÄFFER

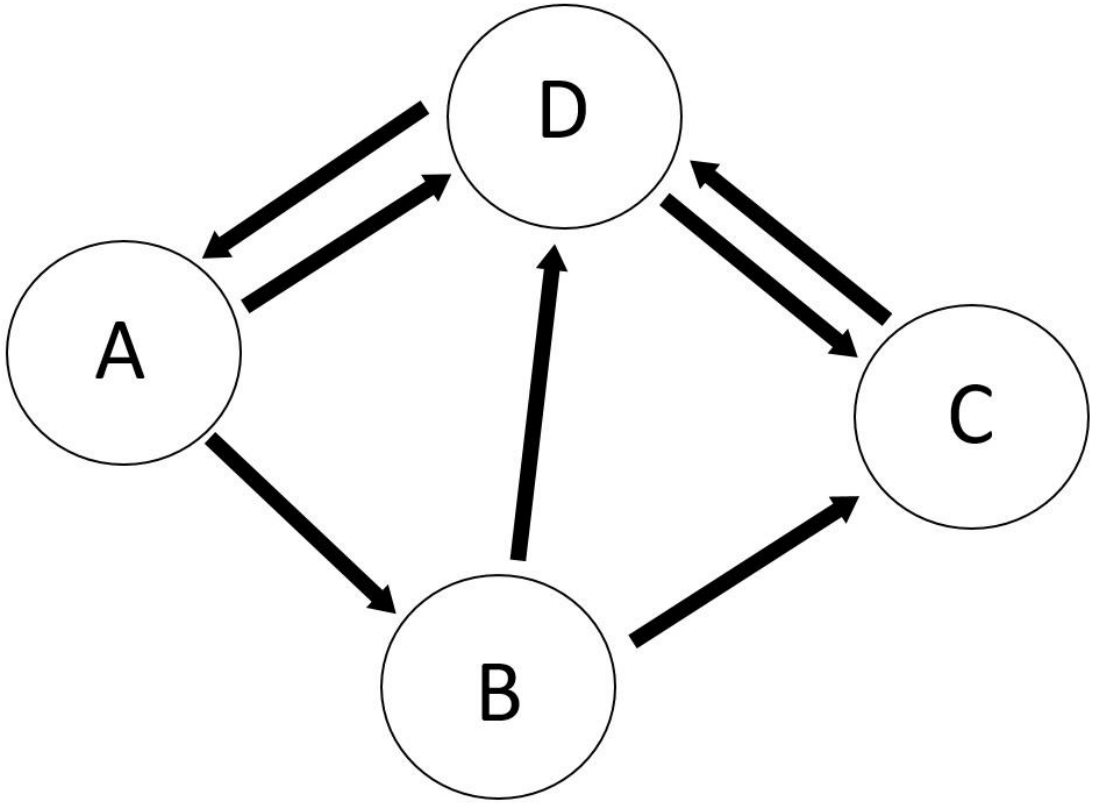
Gliederung

1. Motivationsbeispiel
2. Page Ranking
3. Random Walk
4. Zusammenfassung
5. Literaturverzeichnis

Wer kennt diese Herren?

Quelle: http://static4.businessinsider.com/image/55a93e16371d2211008b767e/google-cofounders-larry-page-and-sergey-brin-just-made-about-8-billion-in-one-day.jpg

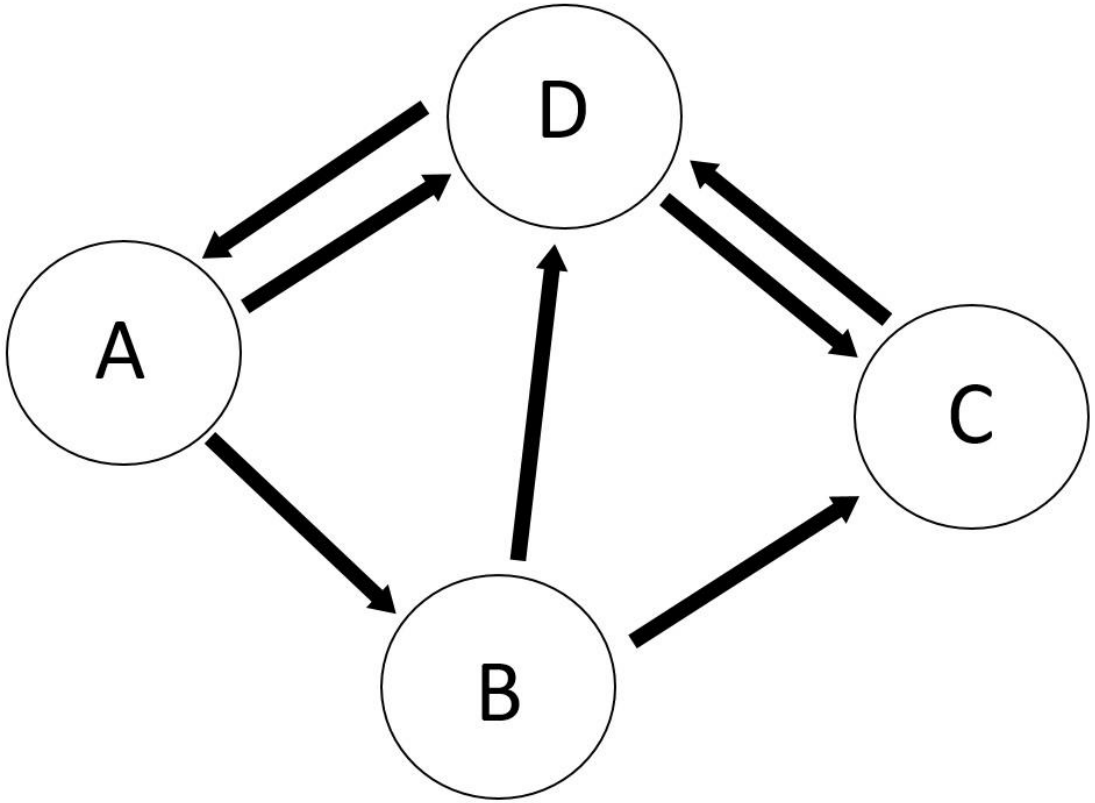
Motivationsbeispiel



- Netzwerk aus vier Webseiten
- jede Webseite hat „outlinks“
- jede Webseite hat „inlinks“
- zufälliger Start bei irgendeiner Seite

Welche Seite wird am häufigsten besucht?

Motivationsbeispiel



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

Motivationsbeispiel

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

$$\left(\begin{array}{cccc|c} 0 & 0 & 0 & \frac{1}{2} & a \\ \frac{1}{2} & 0 & 0 & 0 & b \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & c \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 & d \end{array}\right)$$

Lösungsvektor: $\begin{pmatrix} 4t \\ 2t \\ 5t \\ 8t \end{pmatrix}$ mit $t \in \mathbb{R}$

Motivationsbeispiel

Lösungsvektor $\mathbf{r} = \begin{pmatrix} 4t \\ 2t \\ 5t \\ 8t \end{pmatrix}$ mit $t \in \mathbb{R}$

→ Seite D wird also am ehesten besucht

Page Ranking

Grundprinzip:

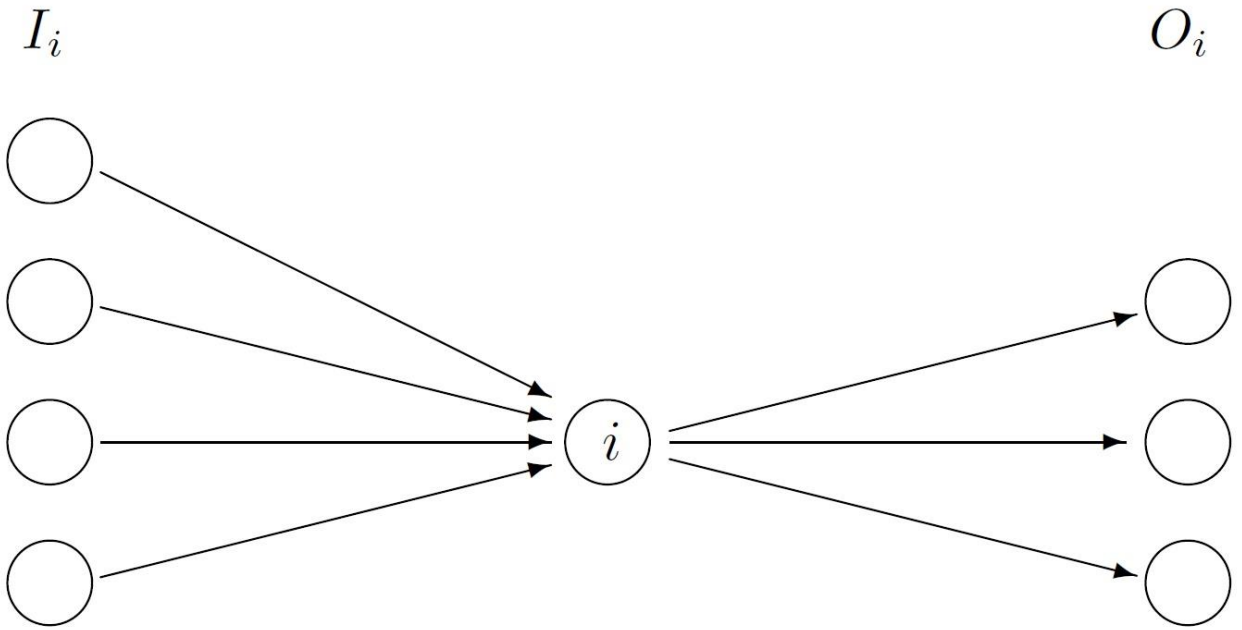
Seite i ist umso „wichtiger“, je mehr inlinks sie hat.

Problem: leicht zu manipulieren!

Quelle: L. Eldén: Matrix methods in data mining and pattern recognition. Volume 4, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

Page Ranking

- alle Webseiten geordnet von 1 bis n
- i sei irgendeine Webseite
- I_i bzw. O_i sind inlinks bzw. outlinks



Quelle: L. Eldén: Matrix methods in data mining and pattern recognition. Volume 4, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

Page Ranking

Abhilfe:

- PageRank r_i einer Seite i als gewichtete Summe der PageRanks, die outlinks zu i haben

$$r_i = \sum_{j \in I_i} \frac{r_j}{N_j}$$

mit der Anzahl der outlinks N_j von Seite j

→ rekursiv, PageRanks können nicht direkt berechnet werden!

Quelle: L. Eldén: Matrix methods in data mining and pattern recognition. Volume 4, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

Page Ranking

Abhilfe:

- wähle Iteration mit Ranking-Vektor $\boldsymbol{r_i}$ und Startvektor $\boldsymbol{r^{(0)}}$

$$\boldsymbol{r_i^{(k+1)}} = \sum_{j \in I_i} \frac{\boldsymbol{r_j^{(k)}}}{N_j}$$

mit $k = 0{,}1 \ldots$

→ Problem: es ist nicht klar, ob diese Iteration auch konvergiert

Quelle: L. Eldén: Matrix methods in data mining and pattern recognition. Volume 4, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

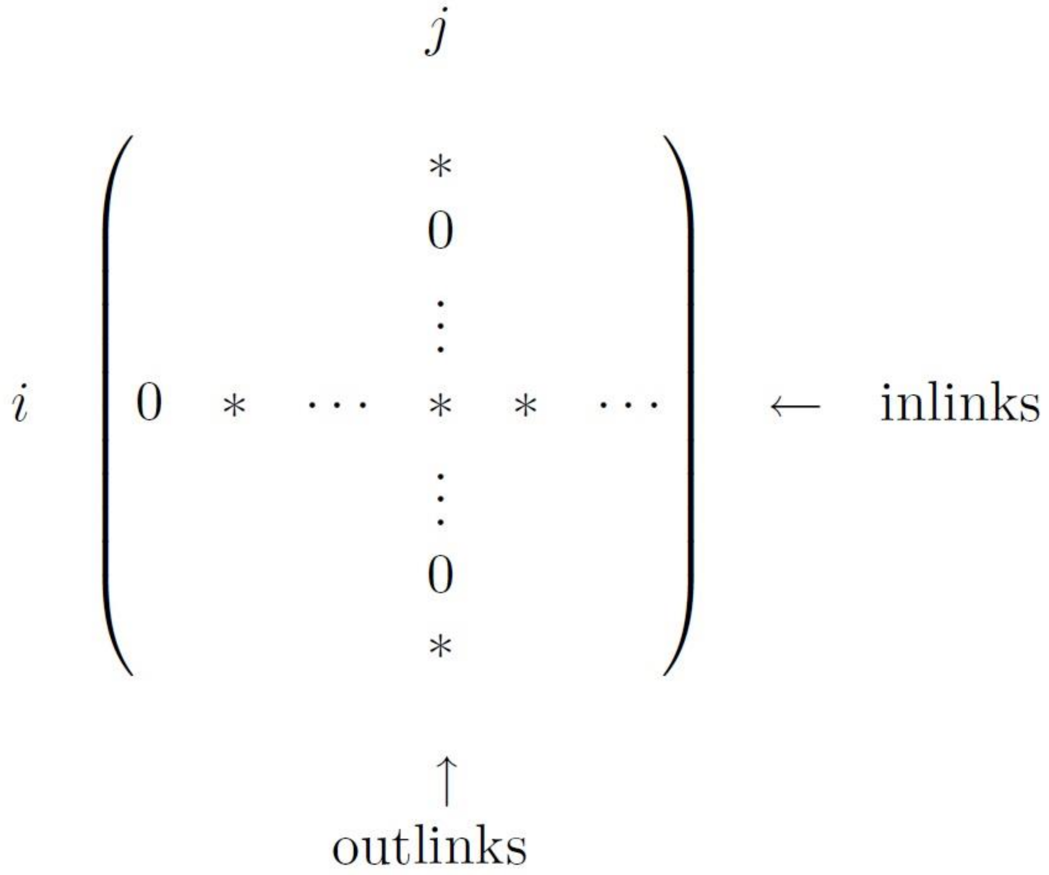
Page Ranking

- mehr Erkenntnis durch Eigenwertproblem
- Q_{ij} quadratische Matrix der Dimension n (Hyperlink-Matrix)
- n Anzahl der Seiten
- $Q_{ij} = \frac{1}{N_j}$ falls es einen Link von j nach i gibt, 0 sonst

Page Ranking

- Eigenwertproblem
- $Q_{ij} = \frac{1}{N_j}$

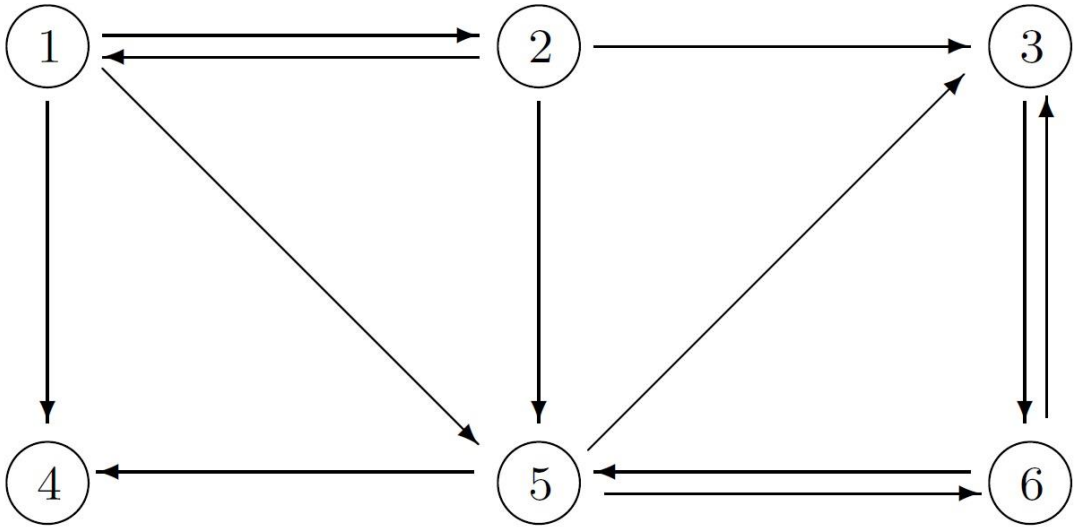
falls es einen Link von j nach i gibt, 0 sonst



Quelle: L. Eldén: Matrix methods in data mining and pattern recognition. Volume 4, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

Page Ranking

- betrachte folgendes Netzwerk



$$Q = \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & \frac{1}{3} & 0 \end{pmatrix}$$

Quelle: L. Eldén: Matrix methods in data mining and pattern recognition. Volume 4, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

Page Ranking

$$Q = \begin{pmatrix} 0 & \frac{1}{3} & 0 & 0 & 0 & 0 \\ \frac{1}{3} & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{3} & 0 & 0 & \frac{1}{3} & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 \\ \frac{1}{3} & \frac{1}{3} & 0 & 0 & 0 & \frac{1}{2} \\ 0 & 0 & 1 & 0 & \frac{1}{3} & 0 \end{pmatrix}$$

Es gilt also für Q:

- $\lambda \cdot \boldsymbol{r} = Q \cdot \boldsymbol{r}$
- mit Eigenwert $\lambda = 1$

Iteration:

$$\boldsymbol{r}^{(k+1)} = Q\boldsymbol{r}^{(k)}$$

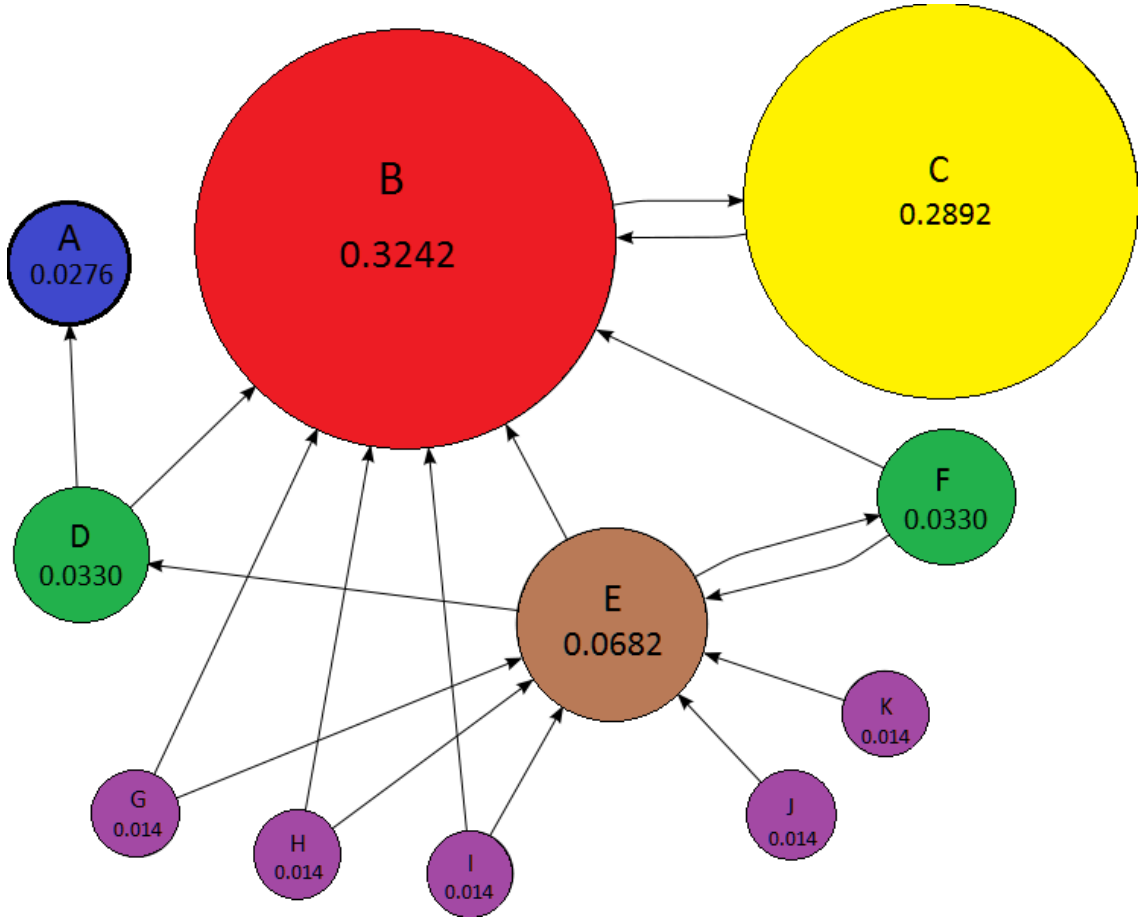
Quelle: L. Eldén: Matrix methods in data mining and pattern recognition. Volume 4, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

Page Ranking

Zwischenergebnis:

Eine Seite ist wichtig, wenn viele wichtige Seiten auf diese verweisen!

Page Ranking



Quelle: https://de.wikipedia.org/wiki/PageRank#/media/File:PageRank-Beispiel.png

Page Ranking

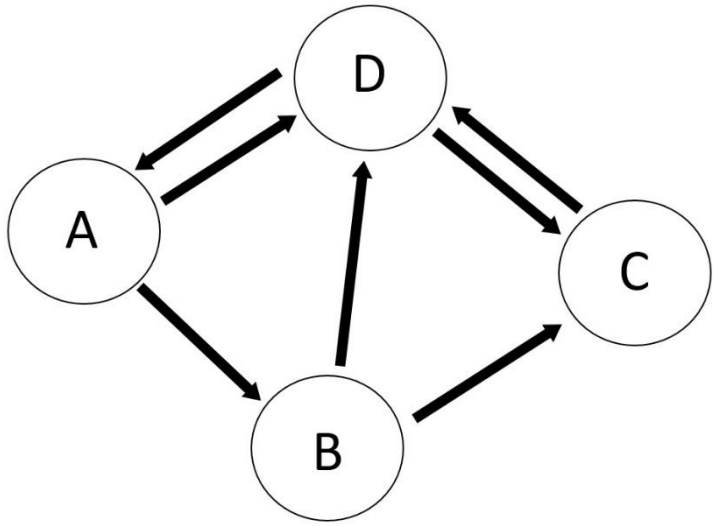
- $\boldsymbol{r}^{(k+1)} = Q\boldsymbol{r}^{(k)}$ (mit Hyperlink-Matrix Q)
- Problem
  - Seiten ohne ausgehende Links verfälschen PageRank
  - Konvergenz?

Random Walk

- wahrscheinlichkeitstheoretische Interpretation
- PageRank auf 1 normieren
- Gewicht einer Seite entspricht Wahrscheinlichkeit
- Surfer klickt zufällig entlang der Links durchs Internet

Random Walk

- Erinnerung



$$\left(\begin{array}{cccc|c} 0 & 0 & 0 & \frac{1}{2} & a \\ \frac{1}{2} & 0 & 0 & 0 & b \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & c \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 & d \end{array}\right)$$

$$P\boldsymbol{r} = \boldsymbol{r}$$

Random Walk

- $P\boldsymbol{r} = \boldsymbol{r}$
- P ist stochastisch
- Eigenwert 1
- ein Lösungsvektor $\boldsymbol{r} = \begin{pmatrix} 4 \\ 2 \\ 5 \\ 8 \end{pmatrix}$

$$\left(\begin{array}{cccc|c} 0 & 0 & 0 & \frac{1}{2} & a \\ \frac{1}{2} & 0 & 0 & 0 & b \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} & c \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 & d \end{array}\right)$$

$$Pr = r$$

Random Walk

- Einführung einer Matrix S
- rein zufälliges Surfverhalten ohne Beachtung der Links
- alle Elemente sind $\frac{1}{n}$ mit n Seiten insgesamt

Quelle: L. Eldén: Matrix methods in data mining and pattern recognition. Volume 4, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.

Random Walk

- Google-Matrix G

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

- quasi eine Überlagerung aus P und S
- $\alpha = 0$ → rein zufälliges Surfverhalten ohne Beachtung von Links
- $\alpha = 1$ → Matrix P
- Page und Brin wählten $\alpha = 0{,}85$
- soll verhindern, dass Anteil des Pageranks vollständig weitergegeben wird

Quelle: http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=6285999

Random Walk

- Google-Matrix G

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

- daher gilt für den Pagerank einer Seite i alternativ auch:

$$r_i = \frac{1-\alpha}{n} + \alpha \cdot \sum_{j \in I_i} \frac{r_j}{N_j}$$

Quelle: https://www.math.uzh.ch/index.php?file&key1=22601

Random Walk

$$r_i = \frac{1-\alpha}{n} + \alpha \cdot \sum_{j \in I_i} \frac{r_j}{N_j}$$

- man kann zeigen (Perron-Frobenius):
  - für $0 < \alpha < 1$ konvergiert jede Iteration für beliebigen Startvektor!
  - und zwar gegen Gleichung

$$\boldsymbol{r} = G\boldsymbol{r}$$

  - mit Eigenvektor **r**

Quelle: https://www.math.tugraz.at/mathc/diskmath/2008/Google.pdf

Random Walk

- Google-Matrix G, $\alpha = 0{,}85$

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

$$0{,}85 \begin{pmatrix} 0 & 0 & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & 0 & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 1 & 0 \end{pmatrix} + 0{,}15 \begin{pmatrix} \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \end{pmatrix} = \begin{pmatrix} 0,0375 & 0,0375 & 0,0375 & 0,4625 \\ 0,4625 & 0,0375 & 0,0375 & 0,0375 \\ 0,0375 & 0,4625 & 0,0375 & 0,4625 \\ 0,4625 & 0,4625 & 0,8875 & 0,0375 \end{pmatrix}$$

Random Walk

- Google-Matrix G

$$\begin{pmatrix} 0,0375 & 0,0375 & 0,0375 & 0,4625 \\ 0,4625 & 0,0375 & 0,0375 & 0,0375 \\ 0,0375 & 0,4625 & 0,0375 & 0,4625 \\ 0,4625 & 0,4625 & 0,8875 & 0,0375 \end{pmatrix}$$

- auch G ist stochastisch
- Eigenwert 1
- Eigenvektor gerundet:

$$\mathbf{r} = \begin{pmatrix} 0{,}39 \\ 0{,}23 \\ 0{,}49 \\ 0{,}75 \end{pmatrix}$$

- also genau das gleiche Page Ranking!

Zusammenfassung

- Seite umso bedeutender, je mehr inlinks sie hat
- Links von bedeutenden Seiten sollen stärker zählen
- Link einer Webseite, die viele outlinks hat, soll weniger beitragen

Problem:

- entscheidend nicht das Interesse der Leser, sondern das anderer Seitenbetreiber!
- Page Ranking liefert keinen Beitrag zur qualitativen Messung des Inhalts!

Vielen Dank für die Aufmerksamkeit!

Literaturverzeichnis

- *L. Eldén: Matrix methods in data mining and pattern recognition. Volume 4, Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA, 2007.*

- URL

- https://www.math.uzh.ch/index.php?file&key1=22601
- http://www-i1.informatik.rwth-aachen.de/~algorithmus/algo10.php
- https://de.wikipedia.org/wiki/PageRank
- http://patft.uspto.gov/netacgi/nph-Parser?patentnumber=6285999
- https://www.math.tugraz.at/mathc/diskmath/2008/Google.pdf