TU Clausthal

Google's PageRank

Eine Anwendung von Matrizen und Markovketten

Vortrag im Rahmen der Lehrerfortbildung an der TU Clausthal

23. September 2009

Dr. Werner Sandmann

Institut für Mathematik
Technische Universität Clausthal

Vortragskontext: Web Information Retrieval

World Wibe Web: Dokumentensammlung, die sich von anderen Dokumentensammlungen (Bibliotheken etc.) wesentlich unterscheidet.

- riesig: mehrere Milliarden Webseiten, d.h. im Web verfügbare Dokumente aller Art (html, xml, php, pdf, doc, xls,...);
- hochdynamisch: 40% aller Seiten ändern sich innerhalb einer Woche, Milliarden neue Seiten jedes Jahr;
- selbstorganisiert: keine Standards, keine Qualitätkontrollen o.ä.
  $\Rightarrow$ fehlerhafte Informationen, Betrügereien, Spam, Datenmüll;
- unübersichtlich: zu fast jedem Thema/Stichwort Millionen potentiell relevanter Seiten;
- teilweise strukturiert: Verweise (Hyperlinks) zwischen verschiedenen Seiten.

Suchmaschinen

- finden potentiell relevante Seiten zu Suchanfragen (queries)
- ordnen die gefundenen Seiten nach deren Relevanz

  ☞ die meisten Nutzer probieren nur wenige (max. 10) angezeigte Seiten

Bewertung von Webseiten durch Suchmaschinen

Suchmaschinen liefern relevante Seiten gemäß bestimmter Bewertungskriterien:

- Content Score: wie gut paßt der Seiteninhalt zur Suchanfrage?
  Problem: ist leicht manipulierbar, als alleiniges Kriterium unzureichend.
- Popularity Score: wie gut ist die Qualität einer Seite?
  - anfrageabhängig (query-dependent):
    Qualität einer Seite wird bezüglich einer Suchanfrage bestimmt,
    ist leicht manipulierbar, erfordert zudem Online-Berechnungen $\Rightarrow$ Wartezeiten.
  - anfrageunabhängig (query-independent):
    Qualität einer Seite wird unabhängig von Suchanfragen fest zugeordnet,
    ist weniger leicht manipulierbar, kann offline berechnet werden.
- Overall Score: Kombination von Content Score und Popularity Score

Der wesentliche Unterschied zwischen verschiedenen Suchmaschinen besteht in der Art der Berechnung dieser Kriterien, insbesondere in der Berechnung des Popularity-Scores.

Zentrale Frage: wie bewertet man die Qualität einer Webseite?

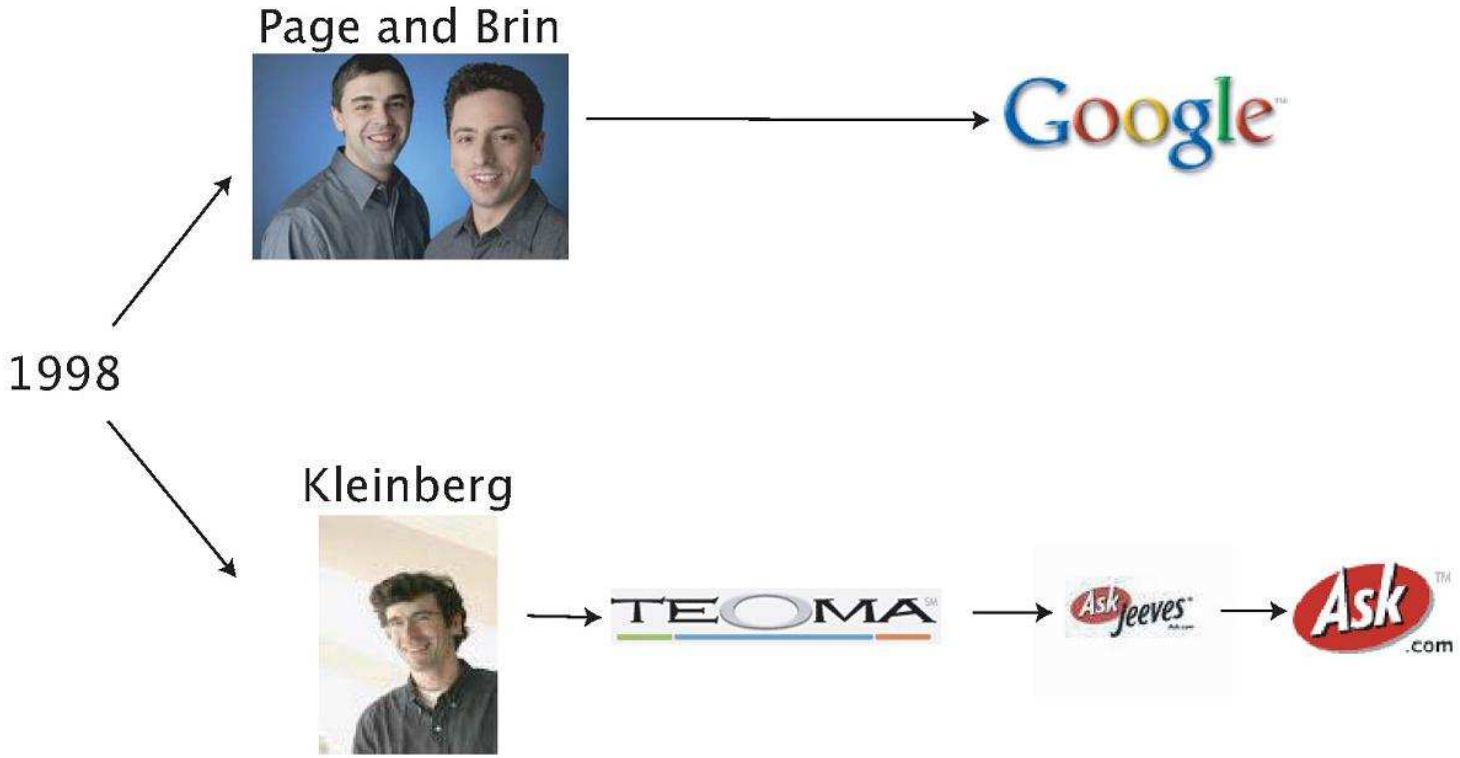
Google-Homepage: „Das Herz unserer Software ist PageRank.“

Hyperlink–Analyse

Idee: Nutze Verbindungsstruktur (Konnektivität) von Webseiten zur Seitenbewertung

Annahmen:

- Verweise (Hyperlinks) verbinden oft verwandte Seiten,
- Ein Verweis auf eine Seite ist eine Empfehlung für diese Seite.



PageRank–Motivation

Google's Bewertungsrichtlinie:

- Eine Webseite ist wichtig und qualitativ hochwertig, wenn viele andere wichtige und qualitativ hochwertige Webseite auf sie verweisen.

Google's Konkretisierung:

- Ein Verweis (Hyperlink) von Seite P_i nach Seite P_j ist eine Empfehlung der Seite P_j durch den Verfasser der Seite P_i.
- Qualität einer Webseite wird bestimmt durch ihren Eingangsgrad, d.h. durch die Anzahl anderer Seiten, die auf sie verweisen.
- Rekursion: Qualität einer Webseite wird bestimmt durch ihren Eingangsgrad und die Qualität der Seiten, die auf sie verweisen

Google's PageRank

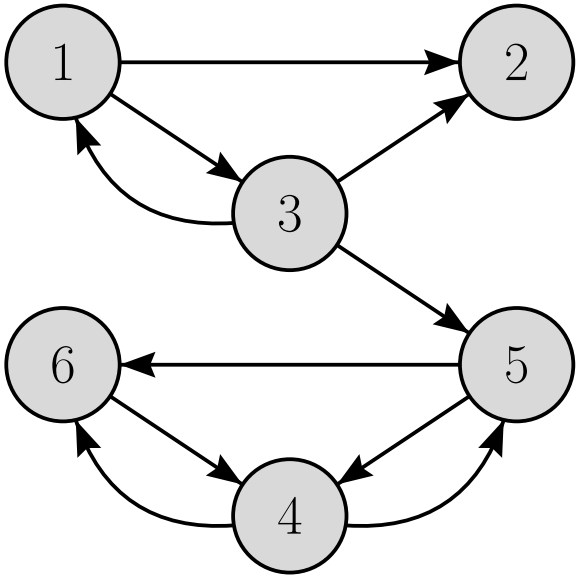
- wird nach diesem Grundprinzip bestimmt,
- ist im Gegensatz zu HITS (Hypertext Induced Topic Search) anfrageunabhängig.

Web–Struktur als gerichteter Graph

Betrachte Menge aller Seiten $\{P_1, \ldots, P_n\}$ und gerichteten Graphen $(\mathcal{V}, \mathcal{E})$ mit

- Knotenmenge $\mathcal{V} = \{1, \ldots, n\}$, Menge der Seitenindizes als Repräsentation der Seiten,
- Kantenmenge $\mathcal{E} \subseteq \mathcal{V}^2$, wobei $(i, j) \in \mathcal{E}$ genau dann, wenn es einen Verweis (Hyperlink) von der Seite P_i zur Seite P_j gibt, $P_i \longrightarrow P_j$.

Hyperlink-Webgraph

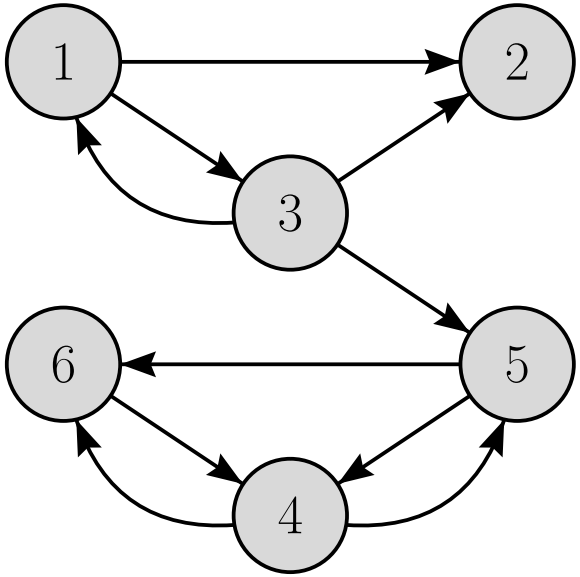


Adjazenzmatrix

$$A = \begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

PageRank–Grundprinzip

- $\mathcal{B}_i :=$ Menge aller Seiten mit Verweis auf die Seite P_i, „Backlinking Pages“
- $\mathcal{O}_i :=$ Menge aller Seiten, auf die die Seite P_i verweist. „Outlinked Pages“



$$\begin{pmatrix} 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 \\ 0 & 0 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

i	$\mathcal{B}_i$	$\mathcal{O}_i$
1	$\{P_3\}$	$\{P_2, P_3\}$
2	$\{P_1, P_3\}$	$\emptyset$
3	$\{P_1\}$	$\{P_1, P_2, P_5\}$
4	$\{P_5, P_6\}$	$\{P_5, P_6\}$
5	$\{P_3, P_4\}$	$\{P_4, P_6\}$
6	$\{P_4, P_5\}$	$\{P_4\}$

- Berechne PageRank $r(P_i)$ gemäß

$$r(P_i) = \sum_{P_j \in \mathcal{B}_i} \frac{r(P_j)}{|\mathcal{O}_j|}$$

für alle Seiten $P_1, \ldots, P_n$, also für $i = 1, \ldots, n$.

- **Problem:** PageRanks $r(P_j)$ der Seiten, die auf P_i verweisen, sind unbekannt!

Iteratives Berechnungsschema

- Starte mit initialem PageRank, z.B. $r^{(0)}(P_i) = \dfrac{1}{n}$ für alle Seiten, also für $i = 1, \ldots, n$.
- Iteriere

$$r^{(1)}(P_i) = \sum_{P_j \in \mathcal{B}_i} \frac{r^{(0)}(P_j)}{|\mathcal{O}_j|} \qquad \text{für } i = 1, \ldots, n,$$

$$r^{(2)}(P_i) = \sum_{P_j \in \mathcal{B}_i} \frac{r^{(1)}(P_j)}{|\mathcal{O}_j|} \qquad \text{für } i = 1, \ldots, n,$$

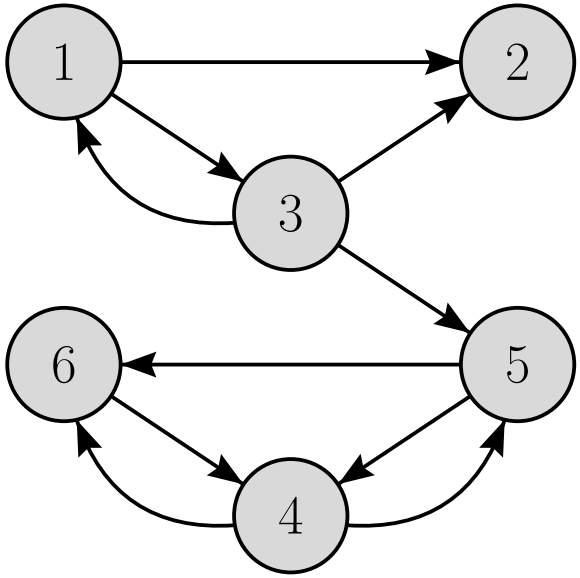
$$\vdots$$

$$r^{(k+1)}(P_i) = \sum_{P_j \in \mathcal{B}_i} \frac{r^{(k)}(P_j)}{|\mathcal{O}_j|} \qquad \text{für } i = 1, \ldots, n,$$

$$\vdots$$

bis die Iteration (hoffentlich!) konvergiert.

Beispiel



i	$\mathcal{B}_i$	$\mathcal{O}_i$
1	$\{P_3\}$	$\{P_2, P_3\}$
2	$\{P_1, P_3\}$	$\emptyset$
3	$\{P_1\}$	$\{P_1, P_2, P_5\}$
4	$\{P_5, P_6\}$	$\{P_5, P_6\}$
5	$\{P_3, P_4\}$	$\{P_4, P_6\}$
6	$\{P_4, P_5\}$	$\{P_4\}$

$$r^{(k+1)}(P_i) = \sum_{P_j \in \mathcal{B}_i} \frac{r^{(k)}(P_j)}{|\mathcal{O}_j|} \qquad \text{für } i = 1, \ldots, n.$$

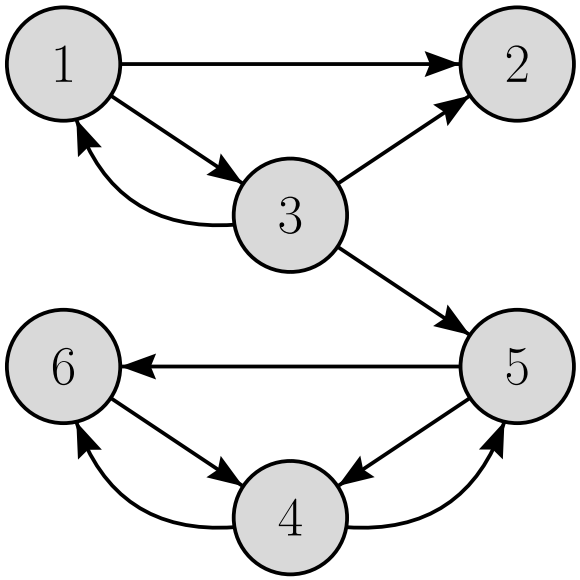
	$k=0$	$k=1$	$k=2$	$\cdots$
$r^{(k)}(P_1)$	$\frac{1}{6}$	$\frac{1}{18}$	$\frac{1}{36}$	$\cdots$
$r^{(k)}(P_2)$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{1}{18}$	$\cdots$
$r^{(k)}(P_3)$	$\frac{1}{6}$	$\frac{1}{12}$	$\frac{1}{36}$	$\cdots$
$r^{(k)}(P_4)$	$\frac{1}{6}$	$\frac{1}{4}$	$\frac{17}{72}$	$\cdots$
$r^{(k)}(P_5)$	$\frac{1}{6}$	$\frac{5}{36}$	$\frac{11}{72}$	$\cdots$
$r^{(k)}(P_6)$	$\frac{1}{6}$	$\frac{1}{6}$	$\frac{14}{72}$	$\cdots$

Hyperlink–Matrix

➳ Matrix $H = (h_{ij})_{i,j=1,\ldots,n}$ mit

$$h_{ij} := \begin{cases} \dfrac{1}{|\mathcal{O}_i|}, & \text{falls } P_i \longrightarrow P_j, \\ 0, & \text{sonst.} \end{cases}$$

Hyperlink-Webgraph



„Hyperlink-Matrix“ = Normalisierte Adjazenzmatrix

$$H = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Iteration in Vektor-Matrixform

- Definiere Zeilenvektor $r^{(k)} := (r^{(k)}(P_1), \ldots, r^{(k)}(P_n))$.
- Iteration in Vektor-Matrix-Form

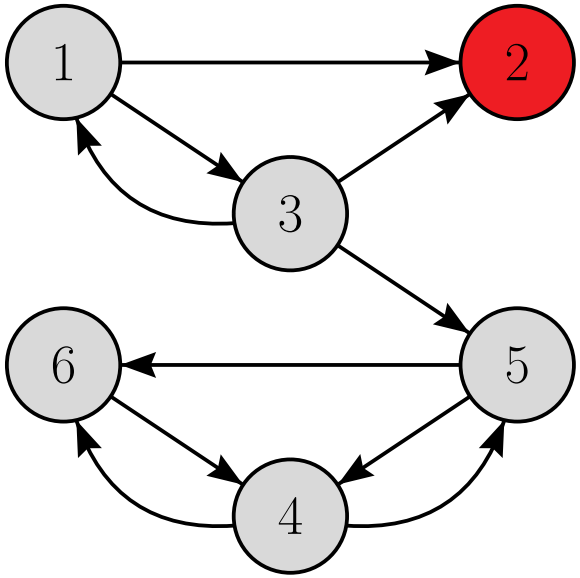
$$r^{(k+1)} = r^{(k)} H \quad \text{bzw.} \quad {r^{(k+1)}}^T = H^T {r^{(k)}}^T.$$

  - Potenzmethode zur Berechnung des betragsmäßig größten Eigenwertes λ_1 von H, bei der der zugehörige Eigenvektor als Beiprodukt abfällt.
- Intendierter PageRank-Vektor $r = \lim_{k \to \infty} r^{(k)}$, falls der Grenzwert existiert,

  soll Eigenvektor der Matrix H zum Eigenwert 1 sein, $r = rH$ bzw. $r^T = H^T r^T$.
- Lösbarkeit des Eigenwertproblems?
  - Existiert ein solcher Eigenvektor?
  - Ist 1 dominanter Eigenwert?
  - Konvergiert die Potenzmethode?
  - Ist der Grenzwert eindeutig, unabhängig vom initialen PageRank?
  - Gesichert für irreduzible aperiodische stochastische Matrizen

Probleme mit der Hyperlink–Matrix

Hyperlink-Webgraph



„Hyperlink-Matrix“ = Normalisierte Adjazenzmatrix

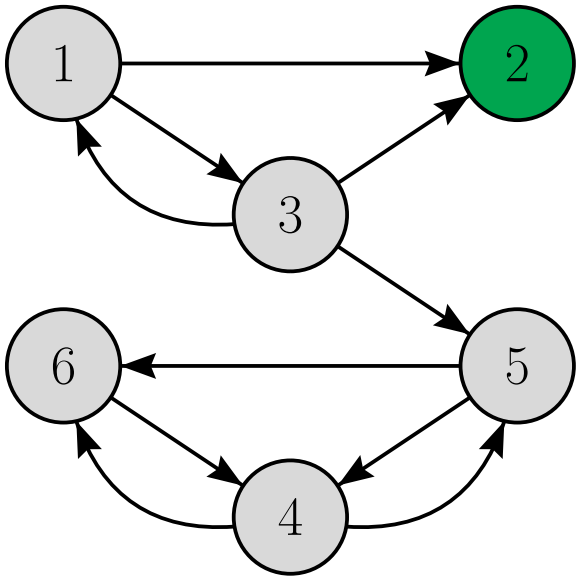
$$H = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

- Es gibt Seiten ohne Hyperlinks, Nutzer landen in einer “Sackgasse“.
- Der Hyperlink-Webgraph enthält im allgemeinen „Senken“ (Dangling Nodes).
- Die Hyperlink-Matrix ist im allgemeinen nur substochastisch, sie enthält Nullzeilen.
- Im allgemeinen keine Konvergenz der Potenzmethode gegen eindeutige Lösung.

Zufällige Sprünge und Surf–Matrix

- **Modifikation:** Wenn auf einer Seite kein Hyperlink auf eine andere Seite vorhanden ist, dann wählt der Nutzer zufällig eine andere Seite (**Random Jumps**).
- Nullzeilen werden durch Gleichverteilung ersetzt.

Hyperlink-Webgraph

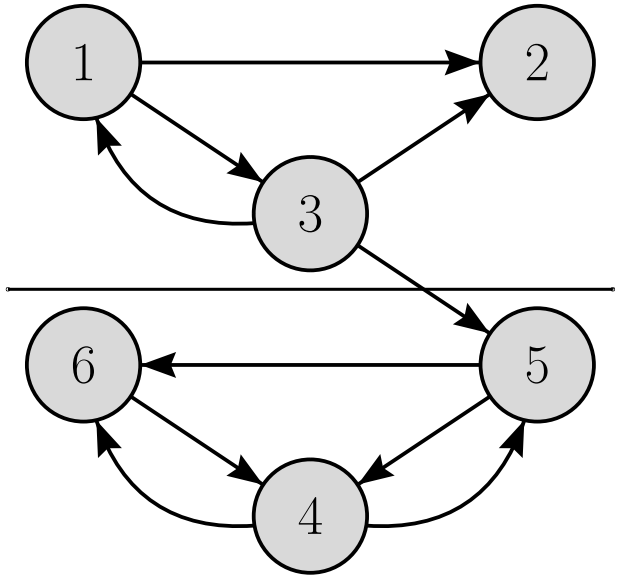


„Surf-Matrix“ = Stochastische Matrix

$$S = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}$$

Probleme mit der Surf–Matrix

Hyperlink-Webgraph



„Surf-Matrix“ = Stochastische Matrix

$$S = \left(\begin{array}{ccc|ccc} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ \hline 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{array}\right)$$

- Der Hyperlink-Webgraph ist nicht stark zusammenhängend, weist Cluster auf.
- Die Surf-Matrix ist nicht irreduzibel, Markovkette enthält transiente Zustände.
- Die Markovkette hat keine von der Anfangsverteilung unabhängige Grenzverteilung.
- Im allgemeinen keine Konvergenz der Potenzmethode gegen eindeutige Lösung.

Erinnerung an Markovketten

- Eine homogene diskrete Markovkette mit Zustandsraum $\mathcal{S}$, Übergangsmatrix $\mathbf{P}$ und beliebiger Anfangsverteilung $\pi^{(0)}$
  - hat eine Grenzverteilung π, falls $\lim_{n\to\infty} \pi^{(n)} = \pi$,
  - hat eine stationäre Verteilung π, falls $\pi = \pi\mathbf{P}$.
- Ist π stationäre Verteilung, dann gilt

$$\pi^{(0)} = \pi \;\Rightarrow\; \pi^{(n)} = \pi \text{ für alle } n \in \mathbb{N}.$$

- Eine Grenzverteilung ist immer stationär, die Umkehrung gilt im allgemeinen nicht.
- Stationäre Verteilungen sind im allgemeinen nicht eindeutig.

Ziel: Garantiere Bedingungen für

- die Existenz von Grenzverteilungen,
- die Eindeutigkeit stationärer Verteilungen.

Diese sind abhängig von bestimmten Eigenschaften der Markovkette $\rightsquigarrow$ Klassifizierungen.

Existenz- und Eindeutigkeitskriterien

Existenz- und Eindeutigkeitskriterien sind insbesondere:

- Für jede aperiodische homogene diskrete Markovkette existiert eine Grenzverteilung π.
- Für jede irreduzible aperiodische homogene diskrete Markovkette ist die Grenzverteilung π unabhängig von der Anfangsverteilung.
- Für jede ergodische (irreduzible, aperiodische, positiv rekurrente) homogene diskrete Markovkette ist die Grenzverteilung π eindeutige stationäre Verteilung.
- Eine eindeutige stationäre Verteilung kann mittels des Gleichungssystems $\pi = \pi\mathbf{P}$ mit Normierungsbedingung berechnet werden.
- Eine eindeutige stationäre Verteilung ist (normierter) Eigenvektor der Übergangsmatrix zum (dominanten) Eigenwert 1.

  $\Rightarrow$ Konvergenz der Potenzmethode für Start mit beliebiger Anfangsverteilung.

Google's Ansatz:

- Modifiziere Surf-Matrix zur Übergangsmatrix einer ergodischen Markovkette!

Google–Matrix

- **Modifikation:** Auch wenn auf einer Seite Hyperlinks auf andere Seiten vorhanden sind, wählt der Nutzer mit gewisser Wahrscheinlichkeit eine andere Seite, auf die die aktuelle Seite nicht verweist.
- Damit kann auf verschiedene Arten Ergodizität erreicht werden.
- Google verwendet eine konvexe Kombination der Surf-Matrix mit einer weiteren stochastischen Matrix, der sogenannten Teleportationsmatrix $E = ee^T/n$, deren Einträge alle gleich $\frac{1}{n}$ sind.

Google-Matrix

$$G = \alpha S + (1-\alpha)E, \qquad E = ee^T/n, \qquad 0 < \alpha < 1.$$

Interpretation:

- Mit Wahrscheinlichkeit α surft ein Nutzer durch Verfolgen von Hyperlinks, mit Wahrscheinlichkeit $1-\alpha$ wählt er zufällig irgendeine Seite.
- $100 \cdot \alpha$ Prozent seiner Zeit im Web folgt ein Nutzer Hyperlinks, $100 \cdot (1-\alpha)$ Prozent seiner Zeit im Web surft er vollkommen zufällig.

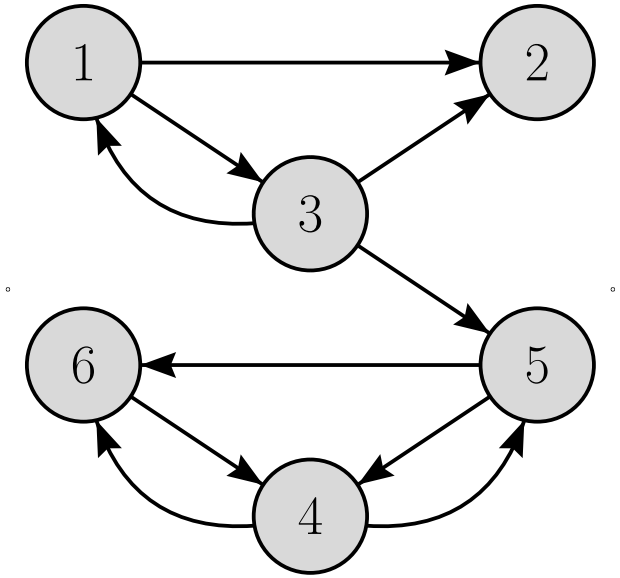
Beispiel zur Google–Matrix

$$S = \begin{pmatrix} 0 & 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}, \qquad E = \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}.$$

Mit $\alpha = 0.9 = \frac{9}{10}$ erhält man die Google-Matrix

$$G = \frac{9}{10} \cdot S + \frac{1}{10} \cdot E = \begin{pmatrix} 1/60 & 7/15 & 7/15 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 19/60 & 19/60 & 1/60 & 1/60 & 19/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 7/15 & 7/15 \\ 1/60 & 1/60 & 1/60 & 7/15 & 1/60 & 7/15 \\ 1/60 & 1/60 & 1/60 & 11/12 & 1/60 & 1/60 \end{pmatrix}.$$

PageRank-Vektor



$$\pi \cdot \begin{pmatrix} 1/60 & 7/15 & 7/15 & 1/60 & 1/60 & 1/60 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 19/60 & 19/60 & 1/60 & 1/60 & 19/60 & 1/60 \\ 1/60 & 1/60 & 1/60 & 1/60 & 7/15 & 7/15 \\ 1/60 & 1/60 & 1/60 & 7/15 & 1/60 & 7/15 \\ 1/60 & 1/60 & 1/60 & 11/12 & 1/60 & 1/60 \end{pmatrix} = \pi.$$

Google's PageRank-Vektor ergibt sich als stationäre Verteilung der Google-Matrix:

$$r := \pi = (0.03721 \quad 0.05396 \quad 0.04151 \quad 0.3751 \quad 0.206 \quad 0.2862).$$

Daraus ergibt sich die Seitenbewertungsreihenfolge $P_4, P_6, P_5, P_2, P_3, P_1$.

Diskussionspunkte

Das Verhalten von Websurfern ist extrem unterschiedlich.

- Der Einfluß des Parameters α bedarf weiterer Untersuchungen.
- Wahl von α ermöglicht prinzipiell Personalisierung,
- Wahl von α beeinflußt Konvergenzgeschwindigkeit der Potenzmethode,
- Google verwendet (angeblich) $\alpha = 0.85$.

Das Web ist riesengroß und enthält Milliarden von Seiten.

- Google berechnet den PageRank-Vektor regelmäßig vollständig neu.
- Google verwendet (angeblich) die Potenzmethode.
- Das ist sicher nicht der Weisheit letzter Schluss!
  - schnellere Berechnung durch Aktualisierung des alten PageRank-Vektors,
  - effizientere Verfahren zur Berechnung stationärer Verteilungen,
  - Kombination beider Ansätze.

Google deckt seine Betriebsgeheimnisse natürlich nicht vollständig auf!

Zusammenfassung

Google's PageRank:

- Bewertung von Webseiten anfrageunabhängig, gestützt auf Hyperlink-Analyse
  - Je mehr Verweise auf eine Seite, desto höher ihr PageRank
  - Normierung bezüglich der Anzahl von Verweisen auf verweisenden Seiten
  - Gewichtung durch PageRank der verweisenden Seiten
- Web-Struktur als Hyperlink-Webgraph aufgefaßt und durch Matrizen dargestellt
  - Gewichtung durch normalisierte Adjazenzmatrix $\longrightarrow$ Hyperlink-Matrix
  - Vermeidung von Sackgassen durch zufällige Sprünge $\longrightarrow$ Surf-Matrix
  - Garantie der Existenz und Eindeutigkeit des PageRank $\longrightarrow$ Google-Matrix
- Google-Matrix ist stochastisch, Übergangsmatrix einer ergodischen Markovkette
- Google's PageRank ist eindeutige stationäre Verteilung dieser Markovkette
- Enormer Berechnungaufwand aufgrund der Größe des Webs

  $\Rightarrow$ Herausforderungen an Forschung: effiziente numerische Analyse von Markovketten

Zum Weiterlesen

Amy N. Langville & Carl D. Meyer:

Google's PageRank and Beyond –
The Science of Search Engine Rankings.

Princeton University Press, 2006.