Page Rank | **HITS**

Sergey Brin & Larry Page (1955) | Jon Kleinberg (1998)

Literatur: A. Bonato, 2008, AMS, A course on the web graph | Langville & Meyer, 2006, Princeton, Google's PageRank and beyond

Thema: · Reihung von Suchanfragen aufgrund der Linkstruktur des Internets.

· Bewertung von Seiten:

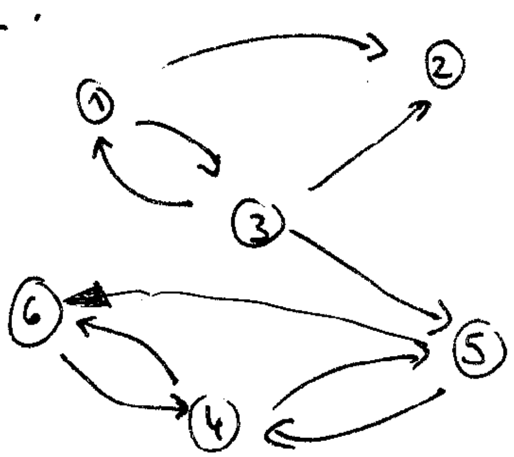
– Gewicht / Bedeutung einer Seite wächst mit der „Anzahl" und dem Gewicht der verweisenden Seiten.

→ Interpretation: Modell des Zufallssurfers

Math. Modell: Website $\hat{=}$ Knoten, Link $\hat{=}$ gerichtete Kante → gerichteter Graph.

Def Ein gerichteter Graph mit Eckenmenge $V = \{1, \dots, n\}$ ist ein Paar (V, E) mit $E \subset V \times V$.

Beispiel.



$V = \{1, \dots, 6\}$

$E = \{(1,2), (1,3), (3,1), (3,2), (3,5), (4,5), (5,4), (6,4), (4,6)\}$.

Mit p_{ij} wird die Wahrscheinlichkeit bezeichnet, dass ein Surfer von (i) nach (j) in einem Schritt geht.

Es sollte dann gelten:

$$\sum_{j=1}^{n} p_{ij} = 1 ,$$

zumindest dann, wenn man „Schleifen" zulässt/einbaut.

Gesucht sind Gewichte / Bewertungen $\pi_1, \ldots, \pi_n \geq 0$ der Knoten $①, \ldots, ⓝ$ mit $\pi_1 + \ldots + \pi_n = 1$ (Normierung), so dass gilt

$$\pi_j = \sum_{i=1}^{n} \pi_i \cdot p_{ij} , \qquad j = 1, \ldots, n .$$

Matrixschreibweise:

$$\pi^T = \pi^T \cdot P , \tag{1}$$

wenn

$$\pi^T = (\pi_1, \ldots, \pi_n) , \quad P = ((p_{ij})) .$$

Def. Eine Matrix $P = ((p_{ij})) \in [0,1]^{n \times n}$ mit

$$\sum_{j=1}^{n} p_{ij} = 1 , \qquad i = 1, \ldots, n$$

heißt stochastische Matrix. Ein Vektor $\pi \in [0,1]^n$ mit

$$\sum_{i=1}^{n} \pi_i = 1$$

heißt stochastischer Vektor bzw. Wahrscheinlichkeitsvektor.

Bedingung (1) ist gleichwertig zu

$$P^T \cdot \pi = \pi , \quad \pi \text{ W-vektor}$$

(ein Eigenwertproblem / Eigenvektorproblem unter Nebenbedingungen)

löse zu einem LGS mit Nebenbedingung:

$$(P^T - E_n) \cdot \pi = 0 \quad , \quad \pi \text{ Wektor.}$$

Im Allgemeinen haben diese Probleme keine eindeutige bzw. keine Lösung, für Google's Page Rank wird jedoch eine spezielle Matrix zugrunde gelegt:

Ansatz 1: $N(i) :=$ Anzahl der von (i) ausgehenden gerichteten Kanten.

$$p_{ij} := \begin{cases} \frac{1}{N(i)} & , \quad (i,j) \in E \\ 0 & , \quad \text{sonst.} \end{cases}$$

Andere Gewichtungen sind denkbar. Hierbei können etwa zusätzlich Schleifen zur Vermeidung von Nullzeilen in P berücksichtigt werden.

Obiges Beispiel:

$$P := \begin{array}{c} \\ 1 \\ 2 \\ 3 \\ 4 \\ 5 \\ 6 \end{array} \overset{\text{nach}}{\begin{array}{c} \begin{array}{cccccc} 1 & 2 & 3 & 4 & 5 & 6 \end{array} \\ \begin{pmatrix} 0 & 0.5 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 1/3 & 0 \\ 0 & 0 & 0 & 0 & 1/2 & 1/2 \\ 0 & 0 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 1 & 0 & 0 \end{pmatrix} \end{array}$$

Problem:
- Sackgassen, d.h. Knoten ① mit $N(i)=0$.
- Unrealistisch in Bezug auf das Surfer-Verhalten

Ansatz 2: Modifikation von P durch Ersetzen von Nullzeilen durch $\frac{1}{n}\cdot e^T$, $e^T := (1,\dots,1)$.

Allgemeiner: Statt $\frac{1}{n}\cdot e^T$ wird ein stochastischer „Personalisierungsvektor" $v^T > 0$, $v^T = (v_1,\dots,v_n)$, $v_i > 0$, verwendet.

Sei $a^T = (a_1,\dots,a_n)$ mit $a_i = \begin{cases} 1, & N(i)=0, \\ 0, & \text{sonst.} \end{cases}$

$$\bar{P} := P + a\cdot v^T.$$

Problem: Nun ist zwar $\bar{P}$ stets eine stochastische Matrix, dennoch kann der gerichtete Graph in Zusammenhangskomponenten zerfallen, die verschieden, gleiss (1) eindeutig lösbar ist.

Ansatz 3: Erneute Modifikation von $\bar{P}$ zu

$$\begin{aligned}\bar{\bar{P}} &:= \alpha\cdot\bar{P} + (1-\alpha)\, e\, v^T \\ &= \alpha P + \left[\alpha\cdot a + (1-\alpha)\cdot e\right] v^T,\end{aligned}$$

wobei $\alpha \in (0,1)$ fest gewählt wird. Dann ist $\bar{\bar{P}}$ erstehin eine stochastische Matrix und zudem sind alle Matrixeinträge positiv.

Vorteile: • α ist flexibel wählbar ($0.85, 0.9, 0.99, ..$)

• Freie Wahl von α, v erschwert mögliche Manipulationen.

• n ist groß, aber P ist „dünn besetzt". Dies gilt nicht für $\bar{\bar{P}}$, dennoch erbt $\bar{\bar{P}}$ diesen Rechenvorteil (s.u.).

Zwei nützliche Fakten zu stochastischen Matrizen:

Satz 1) Ist $A \in \mathbb{R}^{n\times n}$ stochastisch, so ist 1 ein EW von A und A^T. Insbesondere ex. $\pi \in \mathbb{R}^n \setminus \{0\}$ mit $\pi^T = \pi^T \cdot A$.

2) Ist λ ein EW von A bzw A^T, so gilt $|\lambda| \leq 1$.

Beweis 1) Man rechnet sofort nach, dass $(A - E_n)e = 0$ gilt, d.h. $\det(A^T - E_n) = \det(A - E_n) = 0$.

2) A und A^T haben dieselben EWe. Ist λ ein EW von A^T zum EV $v \neq 0$, so ist $A^T v = \lambda v$ und daher

$$\sum_i |\lambda|\,|v_i| = \sum_i |(A^T v)_i| = \sum_i \Big|\sum_j a_{ji} v_j\Big|$$

$$\leq \sum_j \sum_i a_{ji}\,|v_j| = \sum_j |v_j| ,$$

d.h. $|\lambda| \leq 1$. □

Def. Eine Matrix $A \in \mathbb{R}^{n\times n}$ heißt positiv, falls $(A)_{ij} > 0$ $\forall i,j = 1, \ldots, n$.

Da $\bar{\bar{P}}$ eine positive, stochastische Matrix ist, ergibt der Satz von Perron (1905) folgende weitergehende Informationen

- Der Eigenraum zum EW 1 von $\bar{\bar{P}}$ bzw $\bar{\bar{P}}^T$ ist 1-dim.
- Jeder Eigenvektor zum EW 1 ist positiv (d.h. alle Komponenten sind positiv).
- Alle anderen EWe haben Betrag < 1.

Man kann ferner zeigen, dass das folgende Berechnungsverfahren für π mittels Iteration (Potenzmethode) konvergiert, und zwar um so schneller, je kleiner α ist, mindestens aber wie „ $\alpha^k \to 0$ ".

Approximation / Berechnung von π:

- Wähle einen beliebigen W-Vektor x_0^T als Startvektor.
- Definiere rekursiv:

$$x_k^T := \underbrace{x_{k-1}^T \cdot \bar{\bar{P}}}_{= x_0^T \cdot \bar{\bar{P}}^k} = \alpha\, x_{k-1}^T \cdot P + \left[\alpha\, x_{k-1}^T a + (1-\alpha)\right] v^T.$$

(x_{k-1}^T ist absolut.)

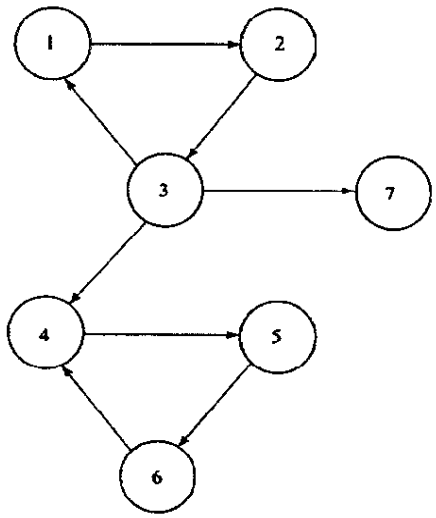
- Dann gilt $x_k \to \pi$ für $k \to \infty$.

Obiges Beispiel führt mit $\alpha = 0.9$ und $v = \frac{1}{6}e$ auf ($k=50$):

$$\pi^T = (0.037\ldots,\ 0.0539\ldots,\ 0.0415\ldots,\ 0.375\ldots,\ 0.206\ldots,\ 0.286\ldots).$$

↑ 6. Platz ↑ 1. Platz

Kleiner Webgraph mit insgesamt 7 Knoten (Webseiten)



		$\alpha = .8$			$\alpha = .9$			$\alpha = .99$		
$\sigma(\mathbf{H})$	$\sigma(\mathbf{S})$	$\sigma(\mathbf{G})$	π^T	Rank	$\sigma(\mathbf{G})$	π^T	Rank	$\sigma(\mathbf{G})$	π^T	Rank
1	1	1	.0641	6	1	.0404	6	1	.0054	6
-.50+.87i	-.50+.87i	-.40+.69i	.0871	5	-.45+.78i	.0558	5	-.50+.86i	.0075	5
-.50-.87i	-.50-.87i	-.40-.69i	.1056	4	-.45-.78i	.0697	4	-.50+.86i	.0096	4
-.35+.60i	**.7991**	.6393	.2372	1	.7192	.2720	1	.7911	.3253	1
-.35-.60i	-.33+.61i	-.26+.49i	.2256	2	-.30+.55i	.2643	2	-.33+.60i	.3240	2
.6934	-.33-.61i	-.26-.49i	.2164	3	-.30-.55i	.2573	3	-.33-.60i	3231	3
0	0	0	.0641	6	0	.0404	6	0	.0054	6

Verändern des Parameters alpha führt in diesem Beispiel zwar zu einer Veränderung der Gewichte, nicht jedoch zu einer Änderung des Ranges. Mit alpha lässt sich zugleich die Sensitivität gegenüber Änderungen der Struktur des Webgraphen beeinflussen, da alpha den zweitgrößten Eigenwert der Google-Matrix G beschränkt. Ist alpha nahe 1, so steigt in der Regel die Empfindlichkeit gegenüber Modifikationen des Graphen.

Einfügen eines weiteren Links von Seite 6 zu Seite 5 ergibt folgendes Bild:

		$\alpha = .8$			$\alpha = .9$			$\alpha = .99$		
$\sigma(\mathbf{H})$	$\sigma(\mathbf{S})$	$\sigma(\mathbf{G})$	π^T	Rank	$\sigma(\mathbf{G})$	π^T	Rank	$\sigma(\mathbf{G})$	π^T	Rank
1	1	1	.0641	6	1	.0404	6	1	.0054	6
-.50+.50i	.7991	.6393	.0871	5	.7192	.0558	5	.7911	0075	5
-.50-.50i	-.50+.50i	-.40+.40i	.1056	4	-.45+.45i	.0697	4	- 50+.50i	0096	4
.6934	-.50-.50i	-.40- 40i	1637	3	-.45-.45i	.1765	3	-.50-.50i	1968	3
-.35+.60i	-.33+.61i	-.26-.49i	.2664	1	-.30+.55i	3145	1	-.33+ 60i	.3885	1
-.35-.60i	-.33-.61i	- 26-.49i	.2491	2	-.30-.55i	.3025	2	-.33-.60i	.3848	2
0	0	0	.0641	6	0	.0404	6	0	.0054	6

Hier ist Seite 4 auf Rang 3 abgefallen, nur die Seiten 4, 5, 6 haben eine neue Bewertung erhalten.

$H = P,\ S = \overline{P},\ G = \overline{\overline{P}}$