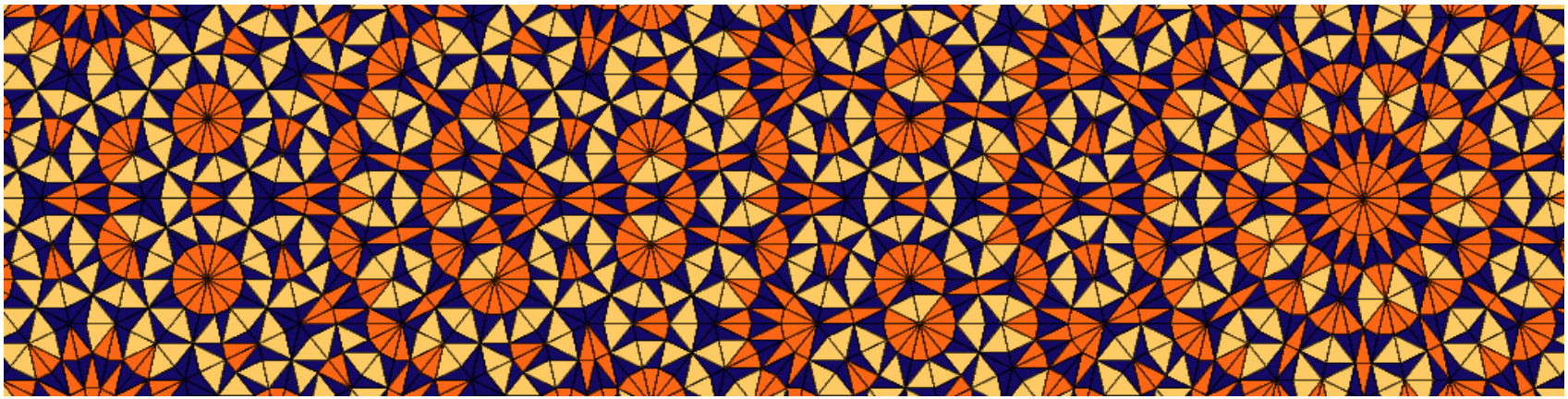
Panorama der Mathematik und Informatik

17: Algorithmen II: Google

Dirk Frettlöh
Technische Fakultät / Richtig Einsteigen

9.6.2015



Gründe für den Erfolg von google:

- Kein Schnickschnack (schlichte Seiten, kluges Bezahlmodell für Werbung)
- Relevante Seiten zuerst: kluge Berechnung (PageRank^{TM})
- Schnell (nur Text, cleveres Hashing, kein Schnickschnack)
- Verdient Geld. Vgl. **Dotcom-Blase**: AOL, pets.com (1998-2000), geocities, flooz.com (1999-2001) waren mal so bekannt wie Amazon. Aber: keine Gewinne. Anderes Bsp:

Die **Kabel New Media** Gruppe ist ein E-Business-Enabler im Full-Service-Bereich interaktiver Kommunikations- und Sales-Lösungen. Die Kernkompetenz liegt in der ganzheitlichen Beratung und Betreuung von etablierten Unternehmen im Bereich E-Business. Das Dienstleistungsportfolio reicht von der Erstellung von Business-Modellen und der Workflow-Organisation über IT-Consulting und Implementierung bis hin zur Entwicklung und Pflege von Inhalten und Marken sowie zu einem effizienten Customer Relationship Management. Das Kabel New Media Netzwerk hat Standorte in Deutschland, Schweden, Großbritannien, Österreich und der Schweiz.

Über das Vermögen der Kabel New Media AG ist am 01.09.2001 das Insolvenzverfahren eröffnet worden.

Circa 1999: *Jetzt macht Kabel New Media eine Kapitalerhöhung. Eins der führenden Unternehmen der Zukunftskommunikation. Und Quality-Leader im E-Business-Bereich. Was sagt Ihre Nase? Zeichnen Sie jetzt!*

NASDAQ:

Dotcom-Blase: "im März 2000 geplatzte Spekulationsblase, die insbesondere die sogenannten Dotcom-Unternehmen der New Economy betraf und vor allem in Industrieländern zu Vermögensverlusten für Kleinanleger führte." (wikipedia)

Google

- Googol $= 10^{100}$. (Googolplex $= 10^{googol}$)
- Begonnen als Doktorarbeit 1996 in Stanford[1]
- Online 1998, Werbung seit 2000, Börsengang 2004
- bis 2004: Fast alle suchen mit google. Auch z.B. Yahoo (!)
- 2004 geht google an die Börse. Insgesamt kommen 19 605 052 Aktien in den Handel, davon 14 142 135 von google ($\sqrt{2} = 1,4142135...$)
- 2005 bringt google weitere 14 159 265 ($\pi = 3,14159265\ldots$) Aktien auf den Markt.
- Google Earth, Books, Gmail, Docs, Maps, Streetview, Glasses, Chrome, Netbooks, Tablets, Android OS, Mobiltelefone...

2011: 96% der Einnahmen aus Werbung.

[1] S. Brin, L. Page, in: The Anatomy of a Large-Scale Hypertextual Web Search Engine, *Computer Networks and ISDN Systems* 35 (1996)

Eine Stellenanzeige von google:

{ Erste 10-stellige Primzahl in aufeinanderfolgenden Ziffern von e }.com

$$e = 2.71828182845904523536028747135266249775724709369995$$

$$957496696762772407663035354759457138217852516642\underline{7427466391}\ldots$$

Unter 7427466391.com war zu lesen:

$f(1) = 7182818284$
$f(2) = 8182845904$
$f(3) = 8747135266$
$f(4) = 7427466391$
$f(5) =$???

Die ersten Zehnergruppen in den Nachkommastellen von e mit Ziffernsumme 49. (Die Lösungen können heute ergoogelt werden)

Nun zum Algorithmus PageRank^{TM}:

Der Satz von Perron-Frobenius

Gegeben eine $n \times n$-Matrix A, mit Einträgen a_{ij} $(1 \leq i, j \leq n)$.

Es seien alle $a_{ij} \geq 0$ (wir schreiben kurz: "$A \geq 0$")

Sowie: In einer Potenz A^k seien alle Einträge positiv ("$A^k > 0$")

Beispiel: Leslie-Matrizen.

Dienen zur Modellierung von Populationen mit n Altersgruppen.
Überlebensrate jeweils u_i, Fruchtbarkeitsrate jeweils f_i $(1 \leq i \leq n)$.

$$\begin{pmatrix} f_1 & f_2 & f_3 & f_4 & \dots & f_n \\ u_1 & 0 & 0 & 0 & \dots & 0 \\ 0 & u_2 & 0 & 0 & \dots & 0 \\ 0 & 0 & u_3 & 0 & \dots & 0 \\ 0 & 0 & 0 & \ddots & \dots & 0 \\ 0 & 0 & 0 & \dots & u_{n-1} & 0 \end{pmatrix}$$

Konkretes Beispiel:

$$A = \begin{pmatrix} 0 & 2 & 1 \\ 0,5 & 0 & 0 \\ 0 & 0,8 & 0 \end{pmatrix} \geq 0 \qquad \left(A^5 = \begin{pmatrix} 0,8 & 2,32 & 1 \\ 0,5 & 0,8 & 0,2 \\ 0,16 & 0,8 & 0,4 \end{pmatrix} > 0 \right)$$

Startpopulation: 10 Jungtiere, als Vektor: $w := \begin{pmatrix} 10 \\ 0 \\ 0 \end{pmatrix}$

Zeitliche Entwicklung:

$$w, \quad Aw, \quad A(Aw) = A^2 w, \quad A^3 w, \quad A^4 w, \quad \ldots\ldots\ldots\ldots\ldots$$

Hier:

$$\begin{pmatrix} 10 \\ 0 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 0 \\ 5 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 10 \\ 0 \\ 4 \end{pmatrix}, \quad \begin{pmatrix} 4 \\ 5 \\ 0 \end{pmatrix}, \quad \begin{pmatrix} 10 \\ 2 \\ 4 \end{pmatrix}, \quad \ldots \quad \begin{pmatrix} 30,4 \\ 13,128 \\ 9,024 \end{pmatrix}, \ldots$$

Theorem (Perron-Frobenius)

Sei $A \in \mathbb{R}^{n \times n}$ wie oben, also $A \geq 0$, $A^k > 0$ für ein $k \in \mathbb{N}$. Dann gilt:

- *A hat einen Eigenwert λ_{PF}, der größer ist als alle anderen: $|\lambda_{PF}| > |\lambda|$ (λ Eigenwert von A, $\lambda \neq \lambda_{PF}$).*
- *λ_{PF} ist einfach, reell und positiv.*
- *Zu λ_{PF} gibt's einen positiven Eigenvektor v: $Av = \lambda_{PF} v$. Der ist eindeutig, wenn man fordert: $v > 0$, $\|v\| = 1$.*
- *Es gibt keinen weiteren positiven Eigenvektor w mit $\|w\| = 1$.*
- *Für alle Vektoren $w \geq 0, w \neq 0$ gilt: $\frac{1}{\lambda_{PF}^n} A^n w \to v$ ($n \to \infty$).*

Zum Beweis:
Lang und technisch.

Zu den ersten zwei Punkten siehe Perron 1907.

Zum Rest siehe etwa

- E. Seneta, “Nonnegative Matrices and Markov Chains”,
- C.D. Meyer, “Matrix Analysis and Applied Linear Algebra”.

Für unser Populationsmodell von oben (Leslie-Matrix):

$$A = \begin{pmatrix} 0 & 2 & 1 \\ 0,5 & 0 & 0 \\ 0 & 0,8 & 0 \end{pmatrix}$$

Eigenwerte:

$$\underline{1,1597...}; \quad 0,57985... + i0.0932...; \quad 0,57985... - i0.0932....$$

Eigenvektor zu $\lambda_{PF} = 1,1597...$ ist $v = \begin{pmatrix} 0{,}5785... \\ 0{,}2494... \\ 0{,}172... \end{pmatrix}$.

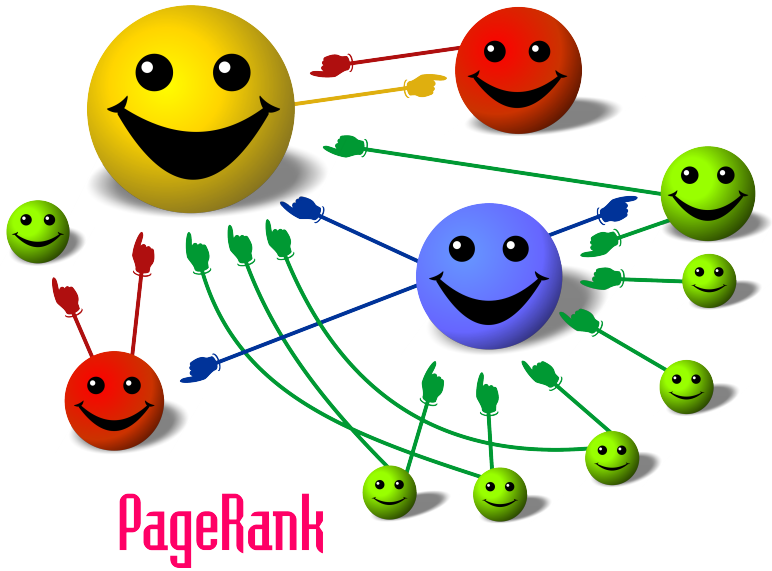
Wegen Perron-Frobenius: $\frac{1}{\lambda_{PF}^n} A^n w \to v$ $(n \to \infty)$.

Populationsverteilung strebt gegen rund 58% : 25% : 17%.

Beispiel: Googles PageRank.

Idee: Webseite i ist wichtig, wenn viele **wichtige** Seiten auf i verlinken.

Sei $p_{ij} = 1$, falls Seite j einen Link auf Seite i enthält; sonst 0. (*Inzidenzmatrix*)



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

Googles PageRank funktioniert nun so:

Berechne den Perron-Frobenius-Eigenvektor λ_{PF} der Inzidenzmatrix des Internets

Wieso ist das sinnvoll? **Idee:** Webseite i ist wichtig, wenn viele wichtige Seiten auf i verlinken.

Sei $p_{ij} = 1$, falls Seite j einen Link auf Seite i enthält; sonst 0.

Sei w_i die Wichtigkeit von Seite i. Dann ist

$$w_i \sim \sum_{j=1}^{N} w_j p_{ij} \quad \text{also} \quad \lambda w_i = \sum_{j=1}^{N} w_j p_{ij}$$

für ein geeignetes λ. (N: Anzahl der Webseiten im Netz.) Also

$$\lambda w = Pw \qquad \text{mit } P \in \mathbb{R}^{N \times N} \text{ gegeben (Inzidenzmatrix)}$$

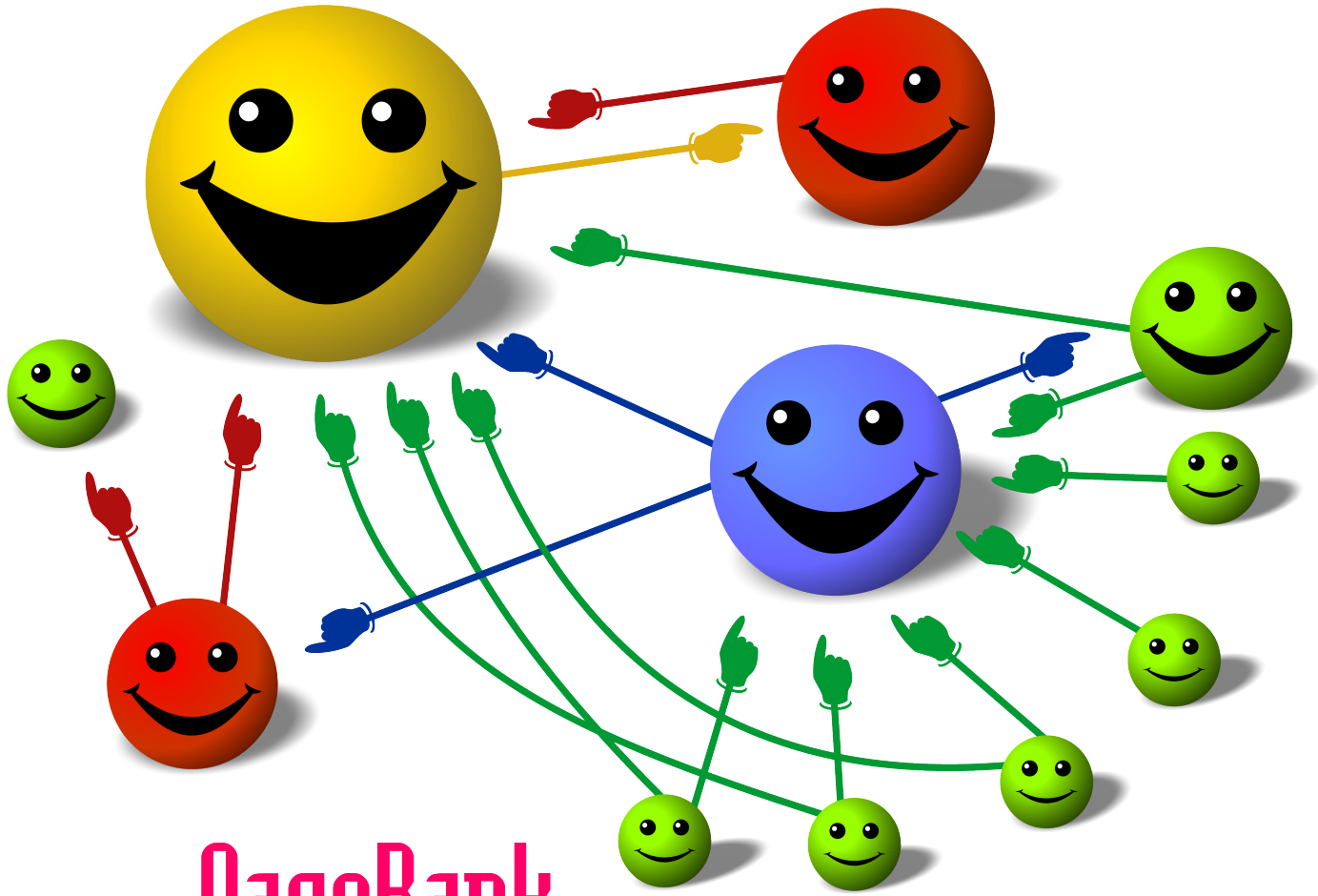
Gesucht ist der (Wichtigkeits-)Vektor w.

Gilt $P^n > 0$ für ein n, so ist λ der Perron-Frobenius-Eigenwert, und w der zugehörige Eigenvektor.

Den berechnet google (im Prinzip) mit Punkt 5 des Satzes von Perron-Frobenius:

- Jeden Monat (Woche? Tag?) aktualisiere Inzidenzmatrix P
- Skaliere die Spalten von P so, dass alle Zeilensummen 1 sind. (Dann ist $\lambda_{PF} = 1$ (Markoffketten!))
- Berechne neues w aus P und dem alten $\tilde{w}$:
  1. $w := P\tilde{w}$
  2. Falls w sehr nah an $\tilde{w}$: Ausgabe w, sonst
  3. $\tilde{w} := w$, weiter bei 1.

Sehr gut parallelisierbar.


PageRank

Problem:

Evtl gibt's kein k mit $P^k > 0$.

D.h. es gibt "Sackgassen" oder "Inseln".

Daher "Dämpfungsfaktor", hier 0.15:

$$w_i = \frac{0.15}{N} + 0.85 \sum_{j=1}^{N} w_j p_{ij}$$

Heute fließen viele weitere Faktoren in den Rang einer Seite ein. (Suchhistorie, Sprache, Standort, Suchwort im Titel einer Seite...)

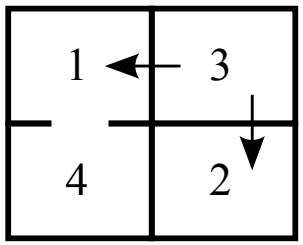
PageRank ist nicht mehr der (einzige? entscheidende?) Faktor bei der Reihenfolge der Suchergebnisse. (Googeln: "Google Panda")

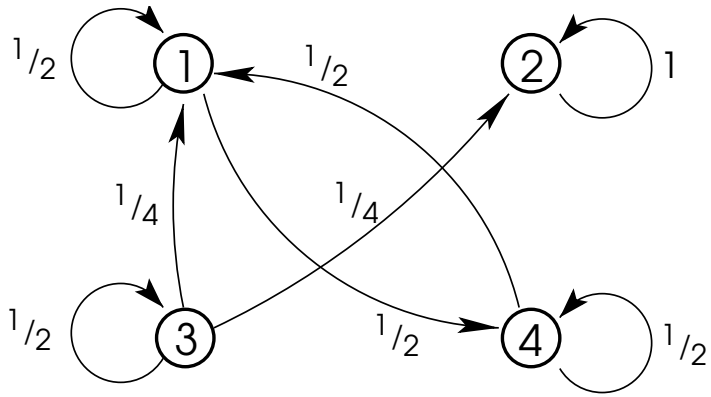
Der Ansatz war nicht ganz neu:

Zitate: "Starke" wissenschaftliche Arbeiten sind solche, die von "starken" Arbeiten zitiert werden.

Sport: "Starke" Teams / Spieler sind die, die starke Gegner schlagen[2].

Interpretation als **Markoffprozess**: Ratte im Labyrinth entspricht zufällig Links anklicken im Netz, berechnet wird Aufenthaltswahrscheinlichkeit auf Seite i.





[auch gezeigt: google hoaxes, siehe wikipedia]

[2] D. Frettlöh: Die Perron-Frobenius-Fußballbundesligatabelle, online