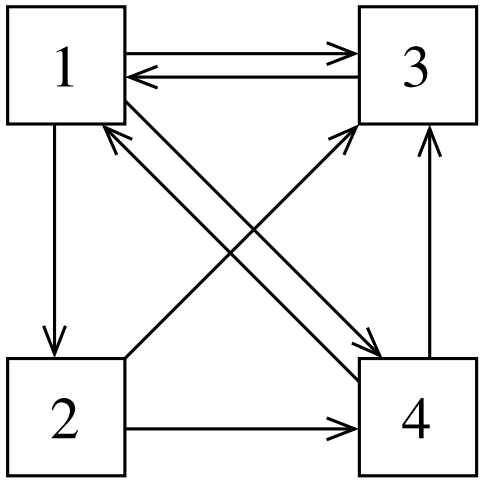
Das Prinzip der Suchmaschine Google™

Numerische Mathematik 1
WS 2011/12

Basieren auf dem Paper “The $25,000,000,000 Eigenvector: The Linear Algebra behind Google” von Kurt Bryan und Tanya Leise (SIAM Review, Vol. 48)

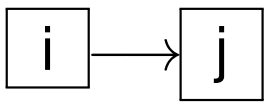
Internet als gerichteter Graph



1. Ein Kasten

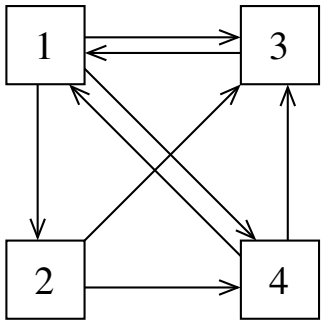
   

   stellt die *Website* Nummer *i* dar.
2. Ein Pfeil

   bedeutet, dass sich auf der Website *i* einen *Link* auf die Website *j* befindet.
   (Links von *i* nach *i* (sogenannte Selbstlinks) werden erst mal nicht berücksichtigt)

Notation



$(n = 4)$

Es bezeichne

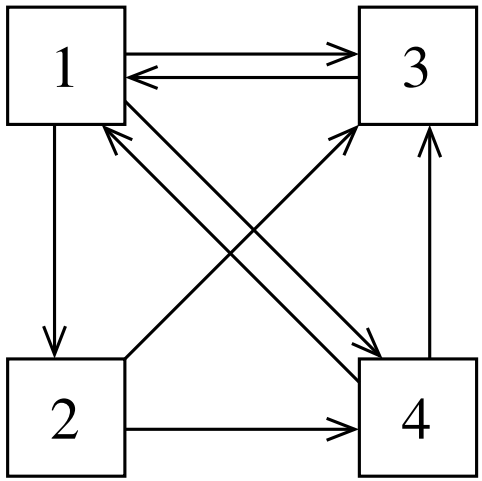
$$W := \{i \mid i = 1, \ldots, n\}$$

die Menge der Websites, welche als Knoten betrachtet werden, und es bezeichne

$$L := \{(i, j) \mid \text{Website } i \text{ verlinkt auf Website } j\}$$

die Menge der gerichteten Kanten in dem Graphen (ohne Selbstlinks, d.h. $(i, j) \in L$ impliziert $i \neq j$).

Problem

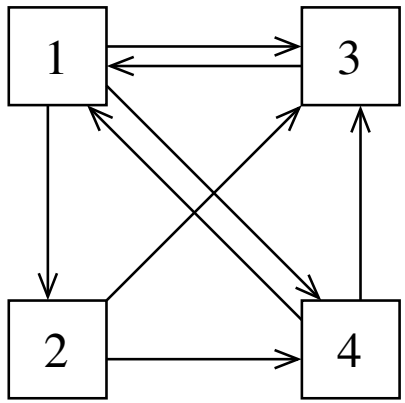


Wie kann man aufgrund eines solchen Graphen bestimmen welche Websites wichtig sind und welche nicht?

Ziel: Bestimme einen sog. *PageRank*^{TM}, d.h. eine endliche Folge $r : W \to \mathbb{R}_{\geq 0}$ welche jeder Website ihre Wichtigkeit zuordnet:

$$r_i > r_j \qquad \Rightarrow \qquad \text{“}i \text{ ist wichtiger als } j\text{”}$$

Was heißt “Wichtigkeit”?

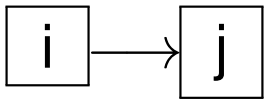


Es scheint sinnvoll, dass

1. Websites mit vielen *Backlinks* (d.h. Websites, auf die viele Links von anderen Seiten verweisen) sind wichtiger als solche mit wenigen.
2. Backlinks von wichtigen Seiten haben stärkeres Gewicht als Backlinks von unwichtigen Seiten.

Ansatz

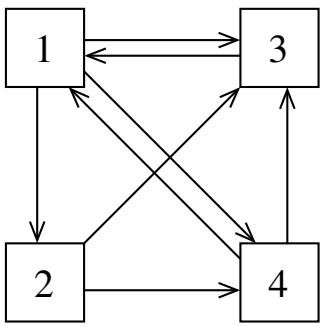
Man interpretiert



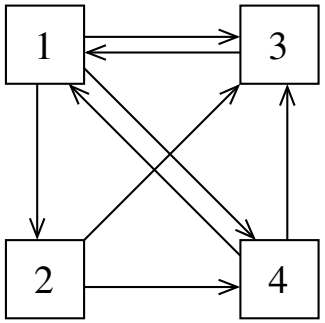
als eine Wählerstimme (engl. vote) von *i* für *j*.

Dabei kann man den Ansatz machen, dass ...

1) ... jede Website ihr Stimmgewicht **gleichmäßig** auf alle verlinkten Seiten verteilt und
2) ... das Stimmgewicht einer Website *i* durch ihre **Wichtigkeit** r_i bestimmt ist.



1) Gleichmäßigkeit



Es bezeichne

$$k_i := \left| \left\{ (i, w) \in L \middle| w \in W \right\} \right|$$

die Anzahl der Links, die von der Website i ausgehen.

Annahme: $k_i \neq 0$ für alle $i \in W$ (d.h. keine hängenden Knoten)

Wenn die Website i auf die Website j verlinkt (d.h. wenn $(i, j) \in L$) dann bekommt die Website j den Anteil

$$\frac{1}{k_i}$$

vom Stimmgewicht, welches der Website i zur Verfügung steht.

2) Stimmgewicht = Wichtigkeit

Wenn die Website i auf die Website j verlinkt (d.h. wenn $(i,j) \in L$) dann bekommt die Website j eine Gesamtstimme von

$$\frac{r_i}{k_i}.$$

Summation über alle Links (die auf die Website j verlinken) liefert

$$r_j = \sum_{\substack{i \in W \text{ mit} \\ (i,j) \in L}} \frac{r_i}{k_i},$$

für $j \in W$.

Grundlegende Gleichung

Schließlich kann man die Gleichungen

$$r_j = \sum_{\substack{i \in W \text{ mit} \\ (i,j) \in L}} \frac{r_i}{k_i}, \qquad \text{für } j \in W,$$

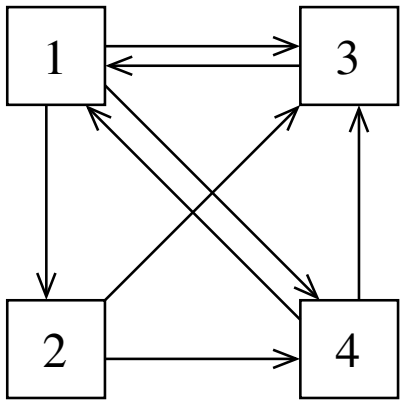
umschreiben in

$$\underbrace{\begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix}}_{=:r} = \underbrace{\begin{bmatrix} \frac{\delta_{1,1}}{k_1} & \cdots & \frac{\delta_{n,1}}{k_n} \\ \vdots & & \vdots \\ \frac{\delta_{1,n}}{k_1} & \cdots & \frac{\delta_{n,n}}{k_n} \end{bmatrix}}_{=:A} \underbrace{\begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix}}_{=r}$$

wobei A *Google-Matrix* heißt und

$$\delta_{i,j} := \begin{cases} 1, & (i,j) \in L \\ 0, & (i,j) \notin L \end{cases}. \qquad \text{(Kronecker-Symbol)}$$

Beispiel

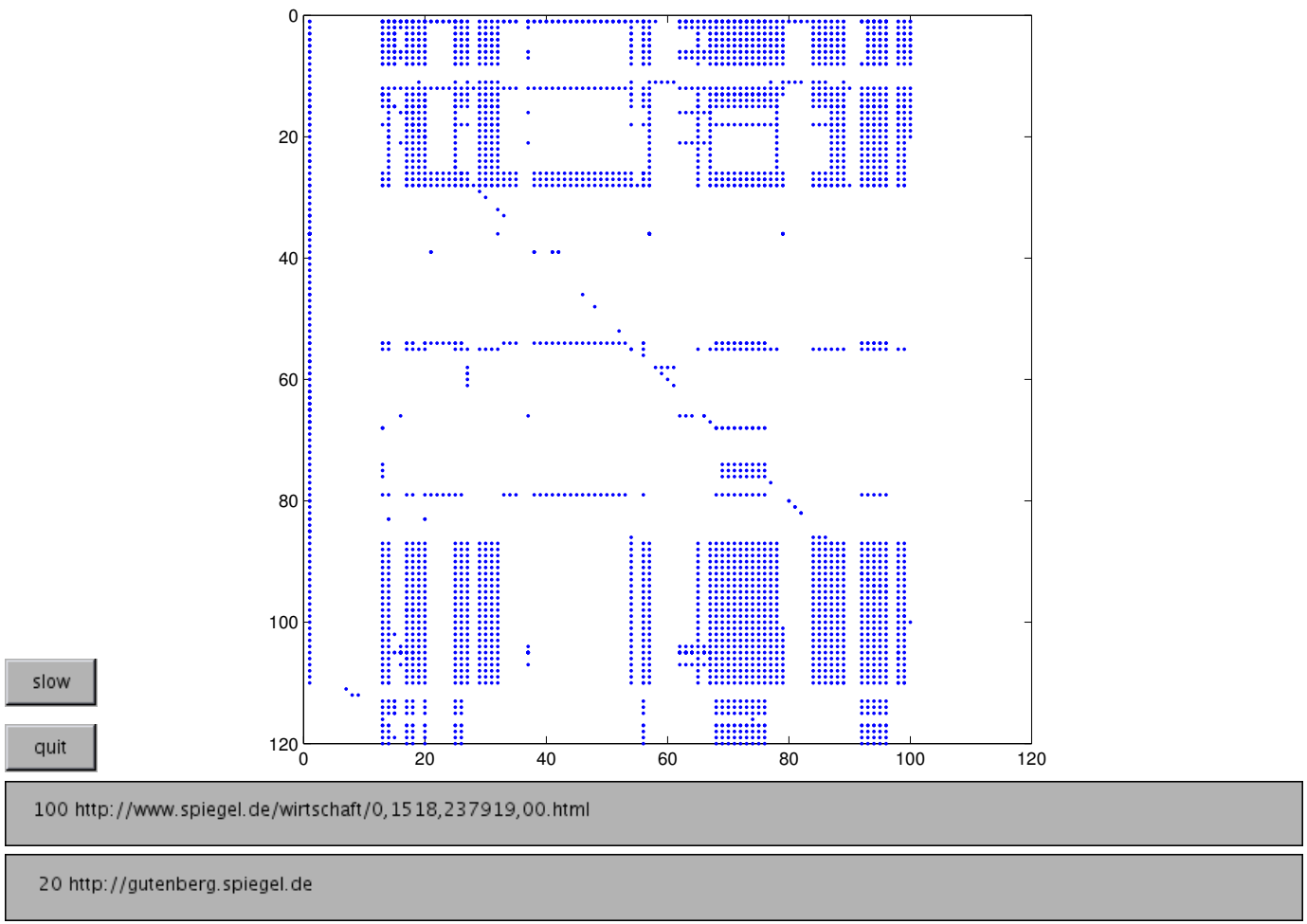


$(n = 4)$

$$\underbrace{\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix}}_{=:r} = \underbrace{\begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}}_{=:A} \underbrace{\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix}}_{=r}$$

Die Summe der Einträge in jeder Spalte von A ergeben stets 1. Das gilt für jede Google-Matrix, falls jede Seite mindestens auf eine andere Seite verlinkt (d.h. keine hängenden Knoten).

Einfacher Webcrawler



`http://www.mathworks.de/moler/ncmfilelist.html`

→ Datei: `surfer.m`

Eigenwertproblem

Es ist also ein Eigenvektor mit nicht-negativen Einträgen zum Eigenwert 1 gesucht, d.h. ein

$$r = \begin{bmatrix} r_1 \\ \vdots \\ r_n \end{bmatrix} \in \mathbb{R}^n$$

mit $r_i \geq 0$ und

$$r = Ar.$$

$\rightarrow$ Existenz? Eindeutigkeit (bis auf Konstante)?

Existenz

Definition

Sei $A \in \mathbb{R}^{n,n}$ eine Matrix deren Einträge alle nicht-negativ sind. Außerdem gelte für jede Spalte, dass die Summe der Einträge gleich 1 ist. Dann heißt A *spalten-stochastisch*.

Lemma

Jede spalten-stochastische Matrix $A \in \mathbb{R}^{n,n}$ hat 1 als Eigenwert.

Beweis: Die Summe der Einträge in jeder Spalte ergibt 1, d.h. definiert man $e := \begin{bmatrix} 1 & \cdots & 1 \end{bmatrix}^T$ so ist $e^T A = e^T$. Daraus folgt

$$A^T e = e,$$

d.h. die Matrix A^T hat den Eigenwert 1 und somit auch die Matrix A.

Es ist nicht klar, dass ein Eigenvektor zum Eigenwert 1 nur nicht-negative Einträge enthält.

Totale Konnektivität

Sei $A \in \mathbb{R}^{n,n}$ spalten-stochastisch und sei

$$S := \frac{1}{n} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & & \vdots \\ 1 & \cdots & 1 \end{bmatrix} \in \mathbb{R}^{n,n}$$

die Google-Matrix, die ein Internet beschreibt, in dem jede Seite auf jede andere Seite verlinkt (inklusive sich selbst). Dann ist

$$M = (1 - m)A + mS$$

für jedes $m \in (0, 1]$ eine spalten-stochastische Matrix mit nur positiven Einträgen.

Perron-Frobenius Theorie

Theorem

Für jede spalten-stochastische Matrix M mit nur positiven Einträgen gilt:

1. *Der Eigenwert* 1 *hat geometrische Vielfachheit* 1*.*
2. *Jeder Eigenvektor zum Eigenwert* 1 *hat nur positive oder nur negative Einträge. Damit gibt es genau einen Eigenvektor r mit nur positiven Elementen, der die Bedingung*

$$1 = \left(e^T r =\right) \sum_{i=1}^{n} r_i = \|r\|_1,$$

erfüllt.

Beweis: "The $25,000,000,000 Eigenvector: The Linear Algebra behind Google" von Kurt Bryan und Tanya Leise (SIAM Review, Vol. 48).

Eigenwerte

Lemma

Für jede spalten-stochastische Matrix M mit nur positiven Einträgen gilt:

Alle Eigenwerte sind betragsmäßig kleiner gleich 1*.*

Beweis: Für jedes Eigenpaar (λ, v) von M gilt

$$|\lambda| = |\lambda| \frac{\|v\|_1}{\|v\|_1} = \frac{\|\lambda v\|_1}{\|v\|_1} = \frac{\|Mv\|_1}{\|v\|_1} \leq \max_{w \neq 0} \frac{\|Mw\|_1}{\|w\|_1} = \|M\|_1 = 1.$$

Fixpunktiteration

Sei M eine Google-Matrix. Setzt man $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ durch

$$\Phi(r) := Mr$$

so sucht man mit

$$r = Mr = \Phi(r)$$

einen Fixpunkt von Φ.

Idee: Fixpunktiteration, d.h. wähle z.B.

$$r_0 := \frac{1}{n} \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \in \mathbb{R}^n$$

und führe die Fixpunktiteration

$$r_{\ell+1} = \Phi(r_\ell) = Mr_\ell$$

durch. $\to$ Potenzmethode

Konvergenz

Theorem

Sei M eine spalten-stochastische Matrix mit nur positiven Einträgen und sei $r_0 \in \mathbb{R}^n$ eine Vektor mit nur positiven Einträgen und $\|r_0\|_1 = 1$. Man definiere die Folge $\{r_\ell\}_{\ell \in \mathbb{N}} \subset \mathbb{R}^n$ rekursiv durch

$$r_{\ell+1} = M r_\ell.$$

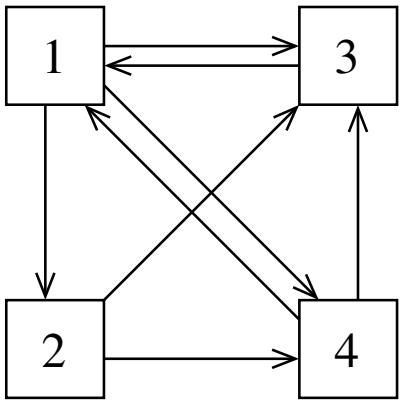
Dann gilt $\|r_\ell\|_1 = 1$ und die Folge $\{r_\ell\}_{\ell \in \mathbb{N}}$ konvergiert gegen

$$\lim_{\ell \to \infty} r_\ell =: r,$$

den eindeutigen Eigenvektor zum Eigenwert 1 mit nur positiven Einträgen und $\|r\|_1 = 1$.

Beweis: "The $25,000,000,000 Eigenvector: The Linear Algebra behind Google" von Kurt Bryan und Tanya Leise (SIAM Review, Vol. 48).

Beispiel



$$r_{\ell+1} = \begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix} r_\ell$$

$$r_0 = \begin{bmatrix} \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \\ \frac{1}{4} \end{bmatrix}, \quad r_1 = \begin{bmatrix} \frac{9}{24} \\ \frac{2}{24} \\ \frac{8}{24} \\ \frac{5}{24} \end{bmatrix}, \quad r_2 = \begin{bmatrix} \frac{63}{144} \\ \frac{18}{144} \\ \frac{39}{144} \\ \frac{24}{144} \end{bmatrix}, \quad \ldots, \quad r_{100} = \begin{bmatrix} 0.3871 \\ 0.1290 \\ 0.2903 \\ 0.1935 \end{bmatrix}$$