Inside PageRank
Markov-Chain-Theory

Studienstiftung Salem 2008

29. August 2008

Random Walk im Internet

Modell
Ein Internetbenutzer bewegt sich zufällig durch das Internet entlang dessen Linkstruktur. Die Menge der Seiten wird mit $S = \{S_1, \ldots, S_n\}$ bezeichnet. Die Wahrscheinlichkeiten h_{ij}, mit denen der Surfer dem Link von S_i nach S_j folgt, bilden die *Hyperlinkmatrix* H. In Sackgassen springt der Surfer willkürlich auf eine der Seiten aus S.

Eigenschaften der Hyperlinkmatrix

Stochastische Matrix
H ist stochastische Matrix, d.h. alle Zeilensummen sind gleich 1.

Spektralradius
Der Spektralradius einer stochastischen Matrix $\rho(H) = 1$.

Beweis:
a) $||H||_\infty = 1$; $||H||_\infty \geq \rho(H)$
b) $Pe = e; 1 \leq \rho(H)$
$\Rightarrow 1 \leq \rho(H) \leq 1 \Rightarrow \rho(H) = 1$

Pagerank als Markov Chain

Definition

Unter einer Markov Chain versteht man einen stochastischen Prozess mit Zustandsraum $S = \{S_1, \dots, S_n\}$, der der Markov-Bedingung

$P(X_{t+1} = S_j | X_t = S_{i_t}, X_{t-1} = S_{i_{t-1}}, \dots, X_0 = S_{i_0}) = P(X_{t+1} = S_j | X_t = S_{i_t})$

genügt, wobei X_t den Zustand zum Zeitpunkt t repräsentiert (diskrete Zeit).

Das Modell des Random Surfers stellt eine stationäre Markov Chain dar, wobei H die Rolle der Übergangsmatrix zukommt.

Wahrscheinlichkeitsvektoren

Definition

Spaltenvektoren $p^T = (p_1, \ldots, p_n)$, deren Spaltensumme gleich 1 ist, nennt man Wahrscheinlichkeitsvektoren.

Wahrscheinlichkeitsvektoren in Markov Chains

Zur Initialisierung der Markov Chain wählt man einen Wahrscheinlichkeitsvektor $p^T(0) = (p_1(0), \ldots, p_n(0)$, der angibt, mit welcher Wahrscheinlichkeit der Random Surfer auf welcher Seite aus S beginnt. Der Wahrscheinlichkeitsverteilungsvektor $p^T(t) = (p_1(t), \ldots, p_n(t)$ listet die Wahrscheinlichkeiten auf, dass der Random Surfer zum Zeitpunkt t auf einer bestimmten Seite ist.

Die Markov Chain in Aktion

Sei $^T(0) = (p_1(0), \ldots, p_n(0))$ die Initialisierung einer Markov Chain mit Übergangsmatrix H. Wir berechnen die Wahrscheinlichkeit, dass der Surfer zum Zeitpunkt $t = 1$ auf der Seite S_j ist:

$$
\begin{aligned}
p_j(1) =& P(X_1 = S_j) \\
=& \sum_{i=1}^{n} P\left(X_1 = S_j \wedge X_0 = S_i\right) \\
=& \sum_{i=1}^{n} P\left(X_0 = S_i\right) P\left(X_1 = S_j | X_0 = S_i\right) \\
=& \sum_{i=1}^{n} p_i(0) h_{ij} \\
\Rightarrow& p^T(1) = p^T(0) H \\
\Rightarrow& Allg. : p^T(t) = p^T(0) H^t
\end{aligned}
$$

Ziel

Was wollen wir?

Um ein Ranking zu erstellen, muss unsere Markov Chain gegen einen stationären Zustand konvergieren, der durch den Wahrscheinlichkeitsverteilungsvektor π beschrieben wird. Wir suchen deshalb nach Kriterien, die die Existenz dieses stationären Zustands garantieren und passen unsere Matrix entsprechend an.

Schön wäre: H primitiv, denn...

Satz über Markov Chains

Satz:
Sei $A \epsilon \mathbb{R}^{n \times n}$ stochastisch und primitiv. Dann exisitiert

$$lim_{t \to \infty} p^T(0) A^t = \pi^T$$

und es gilt $\pi^T A = \pi^T$.
Die stationäre Wahrscheinlichkeitsverteilung π^T ist der eindeutige Eigenvektor zum Eigenwert 1.

Primitive Matrizen

Definition

Eine nichtnegative Matrix $A \epsilon \mathbb{R}^{n \times n}$ heißt primitiv, wenn sie
a) irreduzibel ist und
b) nur einen einzigen Eigenwert auf dem Spektralradius besitzt.

Definition

Eine nichtnegative Matrix $A \epsilon \mathbb{R}^{n \times n}$ heißt reduzibel, wenn es eine Permutationsmatrix P gibt, sodass
$P^T A P = \begin{pmatrix} X & Y \\ 0 & Z \end{pmatrix}$, wobei X und Z quadratisch sind.
Sonst heißt sie irreduzibel.

Perron-Frobenius-Theorie

Satz (Frobenius)
Eine nichtnegative Matrix $A\epsilon\mathbb{R}^{n\times n}$ ist genau dann irreduzibel, wenn ihr gerichteter Graph stark verknüpft ist.

Satz (Perron)
Sei $A\epsilon\mathbb{R}^{n\times n}$ positiv und $r = \rho(A)$. Dann gilt:
r ist Eigenwert von A und r ist der einzige Eigenwert auf dem Spektralkreis von A.

Anpassung der Hyperlinkmatrix

Erster Gedanke
Überprüfen, ob H irreduzibel ist und nur einen Eigenvektor mit Betrag 1 besitzt. Aufwändig!

Besser: Störung von H
Modifikation von H zur Google-Matrix G durch

$$G = \alpha H + (1 - \alpha) e v^T,$$

wobei $\alpha \epsilon (0, 1]$ und v Wahrscheinlichkeitsvektor.

Google-Matrix

Eigenschaften von G
a) G ist stochastisch
b) Der Graph ist trivialerweise stark verknüpft und G damit irreduzibel (Frobenius)
b) G ist positiv und besitzt damit nur einen Eigenwert auf dem Spektralkreis (Perron)
$\Rightarrow$ Es existiert ein stationärer Zustand π^T mit $\pi^T G = \pi^T$.
Die Einträge von π^T entsprechen den Zeitanteilen, die der Random Surfer auf der jeweiligen Webseite verbracht hat.