

Ranking

Grundlegendes

Ziel: Bewertung der Relevanz der Suchergebnisse

Technik:

- Weise jeder Seite v einen Zahlenwert $r(v) \geq 0$ zu.
- Je höher $r(v)$, desto relevanter ist die Seite.
- **global:** Bezogen auf alle Seiten.
- **lokal:** Bezogen auf eine Auswahl von Seiten.

Frage: Wie kann Relevanz gemessen werden?

Ansätze:

- Bewerte den textuellen Inhalt
- Bewerte den textuellen Inhalt bezogen auf die Anfrage
- Verwende die Links

Wir verfolgen den Link-Ansatz.

Der Eingangsgrad

Annahme: Links werden bewusst gesetzt.

⇒ Jeder Link (u, v) ist eine **Stimme** für das Ziel v .

Damit ergibt sich ein einfaches Ranking: **Der Eingangsgrad**

$$r(v) = \text{in}(v)$$

Experimente: Der Eingangsgrad ist als globales Ranking nur sehr beschränkt geeignet.

Verbesserung: Beschränkung auf einen Bereich um die Suchergebnisse.

Eine Lokalisierung

Kleinberg (1999): Einfache Art der Lokalisierung

Idee: Ergänze die Suchergebnisse um benachbarte Dokumente und bewerte diesen Ausschnitt.

Eingabe:

- Ein Graph $G = (V, E)$
- Eine Menge von Seiten $Q \subseteq V$ (Suchergebnis)
- Eine Zahl $d \geq 1$

Ausgabe: Eine Menge R von Seiten mit $Q \subseteq R$

$R = Q$

für alle $v \in Q$ **tue**

- | Füge d verschiedene Vorgänger von v zu R hinzu
- | Füge d verschiedene Nachfolger von v zu R hinzu

Ende

Eine Lokalisierung

Experimente zeigen, dass dies die Qualität des Rankings bereits verbessert.

Problem: Muss zur Laufzeit berechnet werden

⇒ Erhöht die Antwortzeit

Da Nutzer sehr ungeduldig sind, haben sich globale Rankings durchgesetzt, die im Vorfeld berechnet und gespeichert werden können.

Ranking:

PageRank

Der Naive PageRank

S.Brin und L.Page schlugen 1999 eine Verfeinerung vor.

Idee:

- Jede Seite kann nur die Stimmen verteilen, die sie erhält.
- Jede Seite gibt an jede Seite auf die sie verlinkt den gleichen Anteil weiter.

Damit ergibt sich für jede Seite v die folgende Gleichung:

$$r(v) = \sum_{u|u \rightarrow v} \frac{r(u)}{\text{out}(u)}$$

Frage: Wie kann $r(v)$ berechnet werden?

Antwort: Lineare Algebra

Die Normalisierte Adjazenzmatrix

Definition

Die **Normalisierte Adjazenzmatrix** $M = (m_{u,v})_{u,v \in V}$ ist eine $V \times V$ -Matrix mit

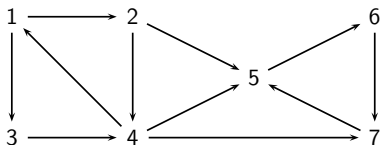
$$m_{u,v} = \begin{cases} 0 & \text{falls } \text{out}(u) = 0 \\ \frac{1}{\text{out}(u)} & \text{falls } \text{out}(u) \neq 0 \text{ und } u \rightarrow v. \end{cases}$$

Normalisiert, da die Zeilensummen entweder 0 oder 1 sind:

$$\sum_{v \in V} m_{u,v} = \begin{cases} 0 & \text{falls } \text{out} = 0 \\ 1 & \text{falls } \text{out} \neq 0 \end{cases}$$

Die Seiten u mit $\text{out}(u) = 0$ nennen wir **Senken**.

Die Normalisierte Adjazenzmatrix



Adjazenzmatrix und normalisierte Adjazenzmatrix:

$$\begin{pmatrix}
 0 & 1 & 1 & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
 1 & 0 & 0 & 0 & 1 & 0 & 1 \\
 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
 0 & 0 & 0 & 0 & 1 & 0 & 0
 \end{pmatrix}
 \quad
 \begin{pmatrix}
 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 & 0 \\
 0 & 0 & 0 & \frac{1}{2} & \frac{1}{2} & 0 & 0 \\
 0 & 0 & 0 & 1 & 0 & 0 & 0 \\
 \frac{1}{3} & 0 & 0 & 0 & \frac{1}{3} & 0 & \frac{1}{3} \\
 0 & 0 & 0 & 0 & 0 & 1 & 0 \\
 0 & 0 & 0 & 0 & 0 & 0 & 1 \\
 0 & 0 & 0 & 0 & 1 & 0 & 0
 \end{pmatrix}$$

Die Normalisierte Adjazenzmatrix

Mittels der normalisierten Adjazenzmatrix M ergibt sich

$$r = M^T r$$

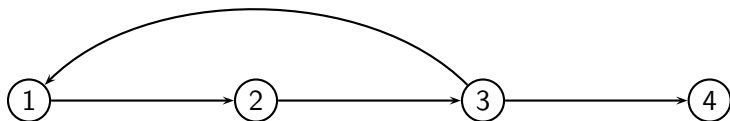
wobei r der Ranking-Vektor ist.

⇒ r ist ein **Eigenvektor** von M^T zum **Eigenwert** 1.

Problem: Häufig gibt es keine Lösung ausser $r = 0$.

Wesentliche Ursache: Die Senken.

Ein Beispiel



Es ergibt sich das folgende Gleichungssystem:

$$r_1 = \frac{r_3}{2}$$

$$r_2 = r_1$$

$$r_3 = r_2$$

$$r_4 = \frac{r_3}{2}$$

Dies führt zu

$$r_3 = r_2 = r_1 = \frac{r_3}{2} \Leftrightarrow r_3 = 0$$

und somit

$$r_1 = r_2 = r_3 = r_4 = 0.$$

PageRank

Um dieses Problem zu umgehen definierten Brin und Page:

Definition

Sei e ein Vektor von Werten für jede Seite und $0 < d < 1$ eine reelle Zahl. Dann ist **PageRank** definiert als:

$$\text{PageRank}(v) = d \sum_{u|u \rightarrow v} \frac{\text{PageRank}(u)}{\text{out}(u)} + (1 - d)e(v).$$

- Jede Seite verteilt ihren Rang gleichmäßig auf ihre Nachfolger
- Der weitergegebene Rang wird mit Faktor d gedämpft
- Jede Seite erhält $(1 - d)e(v)$ als Rang „geschenkt“
- e heißt **Personalisierungsvektor**
- d ist der **Dämpfungsfaktor**

PageRank

Technische Voraussetzungen an e :

- $e(v) \geq 0$ Für alle $v \in V$
- $\|e\|_1 := \sum_{v \in V} |e(v)| = 1$

Übliche Werte:

- $e(v) = \frac{1}{|V|}$
- $d = 0.85$ (ursprünglicher Vorschlag von Brin und Page)

Mittels der normalisierten Adjazenzmatrix ergibt sich

$$\text{PageRank} = dM^T \text{PageRank} + (1 - d)e$$

Interessanter als die eigentliche Definition ist die Interpretation mit Hilfe von **Markov-Ketten**.

Ranking:

Markow-Ketten

Stochastische Prozesse

Grundidee:

- Ein System hat eine (endliche) Menge S von **Zuständen**.
- Das System befindet sich zu jedem Zeitpunkt t in einem eindeutig bestimmten Zustand $X_t \in S$.
- der Zustand X_{t+1} ergibt sich durch ein Zufallsexperiment, das nur vom Zustand X_t und vom Zeitpunkt t abhängt.
- Die **Übergangswahrscheinlichkeiten** sind

$$p_{x,y}^t = P(X_{t+1} = y \mid X_t = x).$$

- Es **muss** ein Zustand angenommen werden, d,h

$$\sum_{y \in S} p_{x,y}^t = 1.$$

Stochastische Prozesse

Definition

Eine **(einfache) Markow-Kette** ist eine Folge $X_0, X_1, X_2, X_3, \dots$ von Zufallsvariablen über einer (endlichen) Menge S , so dass die Wahrscheinlichkeit für $X_{t+1} = x$ nur vom Wert von X_t und t abhängt, d.h.

$$P(X_{t+1} = x \mid X_t = x_t, \dots, X_0 = x_0) \\ = P(X_{t+1} = x \mid X_t = x_t) = p_{x,y}^t.$$

Eine Markow-Kette heisst **homogen**, wenn die Übergangswahrscheinlichkeiten $p_{x,y}^t$ nicht von t abhängen, d.h.

$$p_{x,y}^t = p_{x,y}^{t-1} = p_{x,y}^0 = p_{x,y}.$$

Wir werden uns nur mit homogenen Markow-Ketten beschäftigen.

Stochastische Matrizen

Die Übergangswahrscheinlichkeiten ergeben eine Matrix Π :

$$\Pi = (p_{x,y})_{x,y \in S}.$$

Für die Zeilensummen von Π ergibt sich:

$$\sum_{y \in S} p_{x,y} = 1.$$

Definition

Eine $(n \times n)$ -Matrix $\Pi = (p_{ij})$ heisst **stochastisch**, wenn die Zeilensummen 1 sind, d.h. für $1 \leq i \leq n$ gilt:

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

Konsequenz: Stochastische Matrix \simeq einfache homogene Markow-Kette

Die zeitliche Entwicklung

Die Wahrscheinlichkeit, das das System zum Zeitpunkt t im Zustand x ist, bezeichnen wir mit

$$p_t(x) = P(X_t = x).$$

$\Rightarrow \sum_{x \in S} p_t(x) = 1$ für jeden Zeitpunkt t . Im Allgemeinen gilt:

$$\begin{aligned} p_{t+1}(x) &= P(X_{t+1} = x) \\ &= \sum_{y \in S} P(X_{t+1} = x \mid X_t = y) P(X_t = y) = \sum_{y \in S} p_{y,x} p_t(y) \end{aligned}$$

und somit:

$$p_t = \Pi^T x_{t-1} = (\Pi^T)^t p_0.$$

Die Stationäre Verteilung

Uns interessieren insbesondere die **stationären Verteilungen**.

Definition

Sei Π eine stochastische $n \times n$ -Matrix und p_0 eine Anfangsverteilung (d.h. $p_0(x) \geq 0$ und $\sum_{x \in S} p_0(x) = 1$).

Die **stationäre Verteilung** p_∞ zu p_0 ist der Grenzwert

$$p_\infty = \lim_{t \rightarrow \infty} (\Pi^T)^t p_0,$$

sofern er existiert.

Es bleibt zu klären, ob solche stationären Verteilungen existieren.

Stationäre Verteilungen und Eigenvektoren

Satz

p_∞ ist genau dann eine stationäre Verteilung, wenn es ein **Eigenvektor** von Π^T zum **Eigenwert** 1 ist, so dass

- $p_\infty(x) \geq 0$ und
- $\sum_{x \in S} p_\infty(x) = 1$.

Stationäre Verteilungen und Eigenvektoren

Beweis.

Existiert eine stationäre Verteilung p_∞ , so gilt

$$\Pi^T p_\infty = \Pi^T \lim_{t \rightarrow \infty} (\Pi^T)^t p_0 = \lim_{t \rightarrow \infty} (\Pi^T)^{t+1} p_0 = p_\infty,$$

d.h. p_∞ ist ein **Eigenvektor** von Π^T zum **Eigenwert** 1.

Umgekehrt gilt für jeden Eigenvektor p zum Eigenwert 1 mit positiven Komponenten

$$(\Pi^T)^t p = p \quad \text{und somit} \quad \lim_{t \rightarrow \infty} (\Pi^T)^t p = p.$$

D.h. der Vektor $p_\infty := \frac{p}{\|p\|_1}$ ist eine stationäre Verteilung. □

Die Existenz der Stationären Verteilungen

Satz

Sei $\Pi = (p_{x,y})$ eine stochastische Matrix, so dass

$$p_{x,y} > 0 \text{ für alle } x, y \in S.$$

Dann existiert genau eine stationäre Verteilung p_∞ mit

$$p_\infty = \lim_{t \rightarrow \infty} (\Pi^T)^t p_0$$

für jede beliebige Anfangsverteilung p_0 .

Sind die Bedingung des obigen Satzes erfüllt, konvergiert

$$p_t = \Pi^T p_{t-1}$$

für eine beliebige Anfangsverteilung p_0 gegen die eindeutig bestimmten stationären Verteilung p_∞ .

Ranking:

PageRank als Markow-Kette

Der Teleportierende Surfer

Wir werden PageRank mit Hilfe eines Zufalls-Surfers deuten.

$$\text{PageRank}(v) = d \sum_{u|u \rightarrow v} \frac{\text{PageRank}(u)}{\text{out}(u)} + (1 - d)e(v)$$

- Der Surfer ist zu jedem Zeitpunkt t auf einer Seite.
- Mit Wahrscheinlichkeit d wählt er einen ausgehenden Link.
- Der Link wird gleichverteilt unter allen Links gewählt.
- Mit Wahrscheinlichkeit $(1 - d)$ teleportiert er zu einer beliebigen Seite.
- Mit Wahrscheinlichkeit $e(v)$ wird v als Ziel der Teleportation gewählt.

Die rechte Seite der Gleichung entspräche also der Wahrscheinlichkeit, dass der Surfer auf Seite v ist, sofern er vorher mit Wahrscheinlichkeit $\text{PageRank}(u)$ auf Seite u war.

Der stochastische Prozess

Wir präzisieren die vorangegangene Beschreibung mit Hilfe von Markow-Ketten.

- Die Zustände sind die Seiten: $S = V$.
- Jede Seite wird mit Wahrscheinlichkeit $\frac{1}{|V|}$ als Startseite gewählt.
- Die Übergangswahrscheinlichkeiten $p_{u,v}$ ergeben sich als:

$$p_{u,v} = \begin{cases} \frac{d}{\text{out}(u)} + (1-d)e(v) & \text{falls } u \rightarrow v \\ (1-d)e(v) & \text{falls } u \not\rightarrow v \end{cases}$$

Der stochastische Prozess

X_0, X_1, X_2, \dots seien die besuchten Seiten.

Damit gilt:

$$\begin{aligned}
 P(X_t = v) &= \sum_u P(X_t = v \mid X_{t-1} = u)P(u) \\
 &= \sum_{u|u \rightarrow v} \frac{d}{\text{out}(u)} P(X_{t-1} = u) \\
 &\quad + \sum_u (1 - d)e(v)P(X_{t-1} = u) \\
 &= d \sum_{u|u \rightarrow v} \frac{P(X_{t-1} = u)}{\text{out}(u)} + (1 - d)e(v)
 \end{aligned}$$

Damit entspricht PageRank einer stationären Verteilung der beschriebenen Markow-Kette.

Die Senken

Problem: Es ist kein Stochastischer Prozeß!

Grund: Die Senken haben keine ausgehenden Kanten!

Die resultierende „stochastische“ Matrix hat die Form

$$\Pi = dM + (1 - d)e$$

Konsequenz: Für die Zeilensummen gilt:

$$\sum_v p_{u,v} = \begin{cases} 1 & \text{falls } \text{out}(u) \neq 0 \\ (1 - d) & \text{falls } \text{out}(u) = 0 \end{cases}$$

Damit ist Π **keine** stochastische Matrix!

Die Senken

Möglichkeiten zu Korrektur:

- Entferne alle Senken (ursprünglicher Vorschlag von Brin und Page)
- Füge zu jeder Senke Kanten zu allen anderen Seiten hinzu (erzwinge die Teleportation)

Die zweite (und favorisierte) Möglichkeit ergibt die folgenden Übergangswahrscheinlichkeiten:

$$p'_{u,v} = \begin{cases} \frac{d}{\text{out}(u)} + (1-d)e(v) & \text{falls } u \rightarrow v \\ \frac{d}{|V|} + (1-d)e(v) & \text{falls } u \not\rightarrow v \end{cases}$$

Die resultierende stochastische Matrix nennen wir Π' .

PageRank als Wahrscheinlichkeit

Da $p'_{u,v} > 0$, hat der durch Π' beschriebene stochastische Prozess eine eindeutige Grenzverteilung PageRank'.

Konsequenz

PageRank'(v) ist die Wahrscheinlichkeit, dass der teleportierende Surfer sich nach unendlich vielen Schritten auf Seite v befindet.

In der Literatur wird in der Regel der adaptierte Wert PageRank'(v) verwendet.

Wir werden sehen, dass das resultierende Ranking äquivalent zum ursprünglichen Ansatz ist.

Insbesondere ist die besondere Behandlung der Senken unnötig.