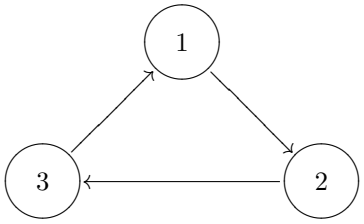
Max-Planck-Institut Saarbrücken
AG5: Datenbanken und Informationssysteme
Dozenten: Prof. Dr.-Ing. G. Weikum, Dr. Ralf Schenkel
Betreuer: Dipl.-Inform. Sergej Sizov, Dipl.-Inform. Christian Zimmer

Informationssysteme (SS 05)
Beispiellösungen zu Übungsblatt 3

10.Mai 2005

Aufgabe 3.1: Markov-Ketten

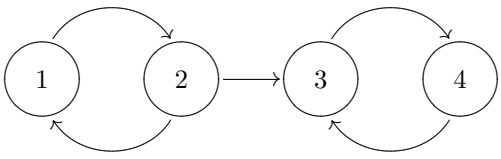
a) Die folgende Markov Kette hat die Periode 3:



(1)

Die dazugehrende Adjazenzmatrix lautet: $\begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{pmatrix}$. Ausgehend von $\Pi^{(0)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, haben wir $\Pi^{(n \bmod 3=1)} = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}$, $\Pi^{(n \bmod 3=2)} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$ und $\Pi^{(n \bmod 3=0)} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$. Somit ist die Markov Kette periodisch. Da die Matrix divergiert, existiert kein Grenzwert $\Pi_j = \lim_{n \mapsto \infty} \Pi_j^{(n)}$, d.h. es gibt keine stationären Zustandswahrscheinlichkeiten und somit ist die Kette nicht ergodisch.

b) Die folgdene Markov Kette ist reduzibel, da nicht jeder Zustand von jedem beliebigen anderen Zustand mit einer positiven Wahrscheinlichkeit erreichbar ist - z.B. kann Zustand 2 so nicht von Zustand 3 erreicht werden.



(2)

Da Zustand 3 von Zustand 2 aus mit positiver Wahrscheinlichkeit erreicht werden kann, ist die rekurrente Wahrscheinlichkeit von Zustand 2 kleiner als 1. Und daher ist die Kette nicht poistiv rekurrent und auch nicht ergodisch.

Aufgabe 3.2:

a)
- Die Kette ist reduzibel da die Zustände 1 und 2 nicht von den Zuständen 3,4 und 5 erreichbar sind.
- Sie ist periodisch, da fr Zustand 1 die einzige Rekurrenz ber Zustand 2 in $2n$ Schritten ist. Daher ist der größte gemeinsame Teiler aller Rekurrenzwerte von Zustand 1 genau 2 ($ggT\{2n\} = 2$ für $n = 1, 2, 3, \ldots$).
- Sie ist nicht ergodisch, da sie weder irreduzibel noch aperiodisch ist.

b)
- Die Kette ist irreduzibel, da alle Zustände von jedem anderen Zustand mit positiver Wahrscheinlichkeit erreichbar sind.
- Sie ist periodisch, da jeder Zustand die Periode 3 hat, d.h. jeder Zustand kann in nur 3 Schritten erreicht werden ($ggT\{3n\} = 3$ für $n = 1, 2, 3, \ldots$).
- Sie ist nicht ergodisch, da sie nicht aperiodisch ist.

c)
- Die Kette ist irreduzibel, da alle Zustände von jedem anderen Zustand mit positiver Wahrscheinlichkeit erreichbar sind.
- Sie ist aperiodisch, da alle Zustände die Periode 1 haben. Zustand 1 und 3 haben Rekurrenzwerte von $2n$ und $3n$ ($ggT\{2, 3\} = 1$); Zustand 2 hat Rekurrenzwert $3n$ und $5n$ ($ggT\{5, 3\} = 1$).
- Die Kette ist auch homogen, da alle Übergangswahrscheinlichkeiten von n unabhängig sind. Somit ist die Kette positiv rekurrent und ergodisch.

d)
- Die Kette ist reduzibel, da Zustand 1 nicht von den Zuständen 2, 3, 4, oder 5 erreicht werden kann.
- Sie ist aperiodisch, da von den Zuständen 2,3 und 4 man Rekurrenzwerte größer oder gleich 2 für alle Pfadlängen finden kann ($ggT\{3, 4, 5, \ldots\} = 1$). Zustand 1 besitzt keinen Rekurrenzwert.
- Sie ist nicht ergodisch, da sie nicht irreduzibel ist.

Anmerkungen:

1) Eine Markov Kette ist ergodisch, wenn sie homogen, irreduzibel, aperiodisch und positiv rekurrent ist.
2) Jede homogene, aperiodische, irreduzibele Markov Kette mit endlichem Zustandsraum ist auch positiv rekurrent und ergodisch.
3) Betrachtet man Schleifen in einer Kette, so dürfen Selbstschleifen von Knoten n zu Knoten n für Aperiodizität oder positive Rekurrenz nicht betrachtet werden.
4) In den meisten Fällen ist es leichter, ein Gegenbeispiel zu finden statt alle möglichen Kombinationen zu testen.

Aufgabe 3.3: Markov-Ketten

a)
- Die Kette ist homogen da alle Übergangswahrscheinlichkeiten von n unabhängig sind.
- Die Kette ist reduzibel, da alle Zustände von jedem anderen Zustand mit positiver Wahrscheinlichkeit erreichbar sind.
- Die Kette ist aperiodisch, da es eine Selbstschleife an jedem Zustand gibt, so dass wir von jedem Zustand zu einem anderen in beliebig vielen Schritten gehen können, bevor wir zum Ausgangszustand zurückkehren; daher können wir alle Rekurrenzwerte größer 1 finden und $ggT\{2, 3, 4, \ldots\} = 1$.
- Die Kette ist positiv rekurrent und ergodisch, da sie homogen, aperiodisch, irreduzibel ist und einen endlichen Zustandsraum besitzt.

b) Da die Kette ergodisch ist, existieren stationäre Zustandswahrscheinlichkeiten π_j und diese konvergieren zu $\pi_j = \lim_{n\to\infty} \pi_j^{(n)}$. Jede Zustandwahrscheinlichkeit ist die Summe alle eingehenden Übergangswahrscheinlichkeiten multipliziert mit den Zustandswahrscheinlichkeiten des Vorgängerzustandes, z.B., $\pi_j = \sum_i \pi_i p_{ij}$. Die Summe aller Zustandswahrscheinlickeiten muss 1 ergeben, d.h., $\sum_j \pi_j = 1$. Für die gegebene Markov Kette erhalten wir folgendes Gleichgewicht für unsere Zustandswahrscheinlichkeiten:

$$\pi_1 = 0.5\pi_1 + 0.1\pi_2 \quad (3)$$

$$\pi_2 = 0.2\pi_2 + 0.5\pi_1 + 0.1\pi_3 \quad (4)$$

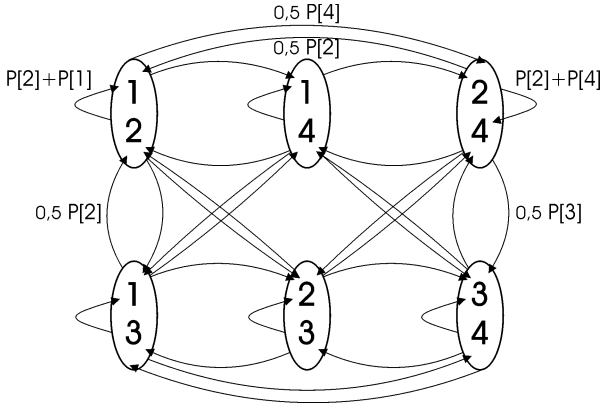
$$\pi_3 = 0.6p3 + 0.7\pi_2 + 0.1\pi_4 \quad (5)$$
$$\pi_4 = 0.9\pi_4 + 0.3\pi_3 \quad (6)$$
$$1 = \pi_1 + \pi_2 + \pi_3 + \pi_4 \quad (7)$$
$$(8)$$

Unter Verwendung der Gauss Eliminierung:

$$\pi_1 = \frac{1}{5}\pi_2 \quad (9)$$
$$\pi_2 = \frac{1}{5}\cdot\pi_2 + \frac{1}{2}\cdot\frac{1}{5}\pi_2 + \frac{1}{10}\pi_3 = \frac{1}{7}\pi_3 \quad (10)$$
$$\pi_3 = \frac{3}{5}p3 + \frac{7}{10}\cdot\frac{1}{7}\pi_3 + \frac{1}{10}\pi_4 = \frac{1}{3}\pi_4 \quad (11)$$
$$1 = \pi_1 + \pi_2 + \pi_3 + \pi_4 = \frac{1}{5}\cdot\frac{1}{7}\cdot\frac{1}{3}\pi_4 + \frac{1}{7}\cdot\frac{1}{3}\pi_4 + \frac{1}{3}\pi_4 + \pi_4 \quad (12)$$
$$\Longrightarrow \quad (13)$$
$$\pi_4 = 105/146 \quad (14)$$
$$\pi_3 = \frac{1}{3}\cdot\frac{105}{146} = 35/146 \quad (15)$$
$$\pi_2 = \frac{1}{7}\cdot\frac{35}{146} = 5/146 \quad (16)$$
$$\pi_1 = \frac{1}{5}\cdot\frac{5}{146} = 1/146 \quad (17)$$
$$(18)$$

Aufgabe 3.4: Anwendung von Markov-Modellen

a) Es sind $n = 6$ verschiedene Zustände vorhanden, wobei jeder eine mögliche Konfiguration des Caches modelliert mit 2 Pages im Speicher (zur besseren Übersicht betrachten wir nur Zustände mit gefülltem Cache; die Startzustände mit einer oder keiner Page werden nicht betrachtet). Die Übergangswahrscheinlichkeiten zwischen zwei Zuständen $c_1 = \{k, j\}$ und $c_2 = \{i, j\}$ sind nur für einige Bespielkanten berücksichtigt. Die Übergangswahrscheinlichkeiten haben die Werte $p_{c_1,c_2} = \frac{1}{2}P[next\ access\ needs\ page\ i]$ für den Fall, dass die Page i in den Cache aufgenommen wird und zur Konfiguration c_2 führt.



b) Die stationären Zustandswahrscheinlichkeiten können durch Lsen des Gleichungssystems der Markov Kette berechnet werden. Aufgrund der *Zipf* Verteilung der Zugriffswahrscheinlichkeiten erhalten wir::

$$P[1] = \frac{(1/1)}{\sum_{i=1]}^{4}\frac{1}{i}} = \frac{(1/1)}{1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4}} = \frac{1}{25/12} = \frac{12}{25} \quad (19)$$
$$P[2] = \frac{(1/2)}{25/12} = \frac{6}{25} \quad (20)$$

$$P[3] = \frac{(1/3)}{25/12} = \frac{4}{25} \tag{21}$$

$$P[4] = \frac{(1/4)}{25/12} = \frac{3}{25} \tag{22}$$

$$\tag{23}$$

Daraus ergibt sich die Matrix P mit allen Übergangswahrscheinlickeiten:

$$P = \begin{pmatrix} 0.72 & 0.08 & 0.06 & 0.08 & 0.06 & 0 \\ 0.12 & 0.64 & 0.06 & 0.12 & 0 & 0.06 \\ 0.12 & 0.08 & 0.60 & 0 & 0.12 & 0.08 \\ 0.24 & 0.24 & 0 & 0.40 & 0.06 & 0.06 \\ 0.24 & 0 & 0.24 & 0.08 & 0.36 & 0.08 \\ 0 & 0.24 & 0.24 & 0.12 & 0.12 & 0.28 \end{pmatrix} \tag{24}$$

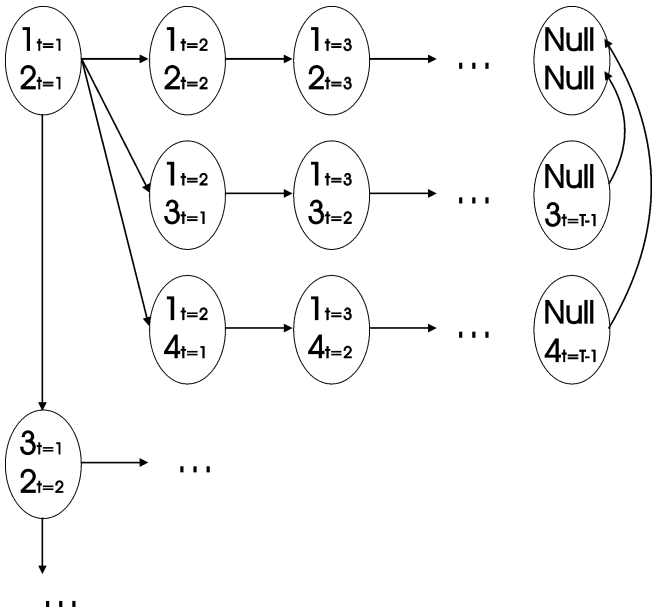
Um die Wahrscheinlichkeiten r_i zu berechnen, müssen wir das folgende Gleichungssystem lösen (z.B. mit Gauss oder Tools wie Maple).

$$\vec{r} = P^T \cdot \vec{r} \quad and \quad 1 = \sum_{i=1}^{|V|} r_i \tag{25}$$

$$\tag{26}$$

(Das Ausrechnen von Hand sollte nicht im Mittelpunkt der Aufgabe hier stehen.)

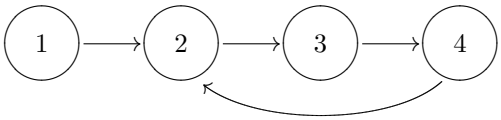
c) Betrachten wir Zeitstempel für gechachte Seiten, so muss der Markov Prozess erweitert werden. Im Vergleich zu den 6 initialen Zuständen müssen wir nun $(T \times T \times (6 + 2 + 2)) = (T \times T \times 10)$ Zustände betrachten



(Hier ist nur ein Auszug der Idee. Alle anderen Zustandsübergänge müssen analog betrachtet werden.)

Aufgabe 3.5: Page-Rank-Verfahren

Wir wollen die Übergangsmatrix P für den Random Walk eines Web surfer auf dem Graphen $G = (V, E)$ betrachten mit $V = \{1, 2, 3, 4\}$ und $E = \{(1, 2), (2, 3), (3, 4), (4, 2)\}$, und $\varepsilon = \frac{1}{10}$.



(27)

Berechnung der Übergangsmatrix:

$$
\begin{aligned}
notequal\,(i,j) &= \begin{cases} 0 & \text{if } i=j \\ 1 & \text{else} \end{cases} && (28)\\
A_{i,j} &= \begin{cases} \frac{1}{outdegree(i)} & \text{if } (i,j)\in E \\ 0 & \text{else} \end{cases} && (29)\\
&= \begin{cases} 1 & \text{if } (i,j)\in E \\ 0 & \text{else} \end{cases} \\
A &= \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix} && (30)\\
P_{i,j} &= \frac{\varepsilon}{n}\cdot notequal\,(i,j) + (1-\varepsilon)\cdot A_{i,j} && (31)\\
&= \frac{0,1}{3}\cdot notequal\,(i,j) + (1-0,1)\cdot A_{i,j} \\
&= 0,0\overline{3}\cdot notequal\,(i,j) + 0,9\cdot A_{i,j} \\
r_i &= \frac{\varepsilon}{n} + (1-\varepsilon)\cdot\sum_{(j,i)\in G}\frac{r(j)}{outdegree\,(j)} && (32)\\
&= 0,0\overline{3} + 0,9\cdot\sum_{(j,i)\in G}\frac{r(j)}{outdegree\,(j)} && (33)\\
P &= \begin{pmatrix} 0 & 0,0\overline{3}+0,9\cdot 1 & 0,0\overline{3}+0,9\cdot 0 & 0,0\overline{3}+0,9\cdot 0 \\ 0,0\overline{3}+0,9\cdot 0 & 0 & 0,0\overline{3}+0,9\cdot 1 & 0,0\overline{3}+0,9\cdot 0 \\ 0,0\overline{3}+0,9\cdot 0 & 0,0\overline{3}+0,9\cdot 0 & 0 & 0,0\overline{3}+0,9\cdot 1 \\ 0,0\overline{3}+0,9\cdot 0 & 0,0\overline{3}+0,9\cdot 1 & 0,0\overline{3}+0,9\cdot 0 & 0 \end{pmatrix} \\
&= \begin{pmatrix} 0 & 0,9\overline{3} & 0,0\overline{3} & 0,0\overline{3} \\ 0,0\overline{3} & 0 & 0,9\overline{3} & 0,0\overline{3} \\ 0,0\overline{3} & 0,0\overline{3} & 0 & 0,9\overline{3} \\ 0,0\overline{3} & 0,9\overline{3} & 0,0\overline{3} & 0 \end{pmatrix} && (34)\\
P^T &= \begin{pmatrix} 0 & 0,0\overline{3} & 0,0\overline{3} & 0,0\overline{3} \\ 0,9\overline{3} & 0 & 0,0\overline{3} & 0,9\overline{3} \\ 0,0\overline{3} & 0,9\overline{3} & 0 & 0,0\overline{3} \\ 0,0\overline{3} & 0,0\overline{3} & 0,9\overline{3} & 0 \end{pmatrix} && (35)
\end{aligned}
$$

wobei $0,9\overline{3} = \frac{28}{30} = \frac{14}{15}$ und $0,0\overline{3} = \frac{1}{30}$.

a) Iterativ, beginnend mit $r_1 = r_2 = r_3 = r_4 = \frac{1}{4}$

$$
\begin{aligned}
\Pi^{(0)} &= \begin{pmatrix} 0,25 \\ 0,25 \\ 0,25 \\ 0,25 \end{pmatrix} && (36)\\
\Pi^{(1)} &= P^T\cdot\Pi^{(0)} = \begin{pmatrix} 0 & 0,0\overline{3} & 0,0\overline{3} & 0,0\overline{3} \\ 0,9\overline{3} & 0 & 0,0\overline{3} & 0,9\overline{3} \\ 0,0\overline{3} & 0,9\overline{3} & 0 & 0,0\overline{3} \\ 0,0\overline{3} & 0,0\overline{3} & 0,9\overline{3} & 0 \end{pmatrix}\cdot\begin{pmatrix} 0,25 \\ 0,25 \\ 0,25 \\ 0,25 \end{pmatrix} = \begin{pmatrix} 0,025 \\ 0,475 \\ 0,25 \\ 0,25 \end{pmatrix} && (37)\\
\Pi^{(2)} &= P^T\cdot\Pi^{(1)} = \begin{pmatrix} 0 & 0,0\overline{3} & 0,0\overline{3} & 0,0\overline{3} \\ 0,9\overline{3} & 0 & 0,0\overline{3} & 0,9\overline{3} \\ 0,0\overline{3} & 0,9\overline{3} & 0 & 0,0\overline{3} \\ 0,0\overline{3} & 0,0\overline{3} & 0,9\overline{3} & 0 \end{pmatrix}\cdot\begin{pmatrix} 0,025 \\ 0,475 \\ 0,25 \\ 0,25 \end{pmatrix} = \begin{pmatrix} 0,0325 \\ 0,265 \\ 0,4525 \\ 0,25 \end{pmatrix} && (38)
\end{aligned}
$$

$$\Pi^{(3)} = P^T \cdot \Pi^{(2)} = \begin{pmatrix} 0 & 0,0\overline{3} & 0,0\overline{3} & 0,0\overline{3} \\ 0,9\overline{3} & 0 & 0,0\overline{3} & 0,9\overline{3} \\ 0,0\overline{3} & 0,9\overline{3} & 0 & 0,0\overline{3} \\ 0,0\overline{3} & 0,0\overline{3} & 0,9\overline{3} & 0 \end{pmatrix} \cdot \begin{pmatrix} 0,0325 \\ 0,265 \\ 0,4525 \\ 0,25 \end{pmatrix} = \begin{pmatrix} 0,03225 \\ 0,27851\overline{6} \\ 0,256741\overline{6} \\ 0,432241\overline{6} \end{pmatrix} \tag{39}$$

$$\Pi^{(4)} = P^T \cdot \Pi^{(3)} = \begin{pmatrix} 0 & 0,0\overline{3} & 0,0\overline{3} & 0,0\overline{3} \\ 0,9\overline{3} & 0 & 0,0\overline{3} & 0,9\overline{3} \\ 0,0\overline{3} & 0,9\overline{3} & 0 & 0,0\overline{3} \\ 0,0\overline{3} & 0,0\overline{3} & 0,9\overline{3} & 0 \end{pmatrix} \cdot \begin{pmatrix} 0,03225 \\ 0,27851\overline{6} \\ 0,256741\overline{6} \\ 0,432241\overline{6} \end{pmatrix} = \begin{pmatrix} 0,03225 \\ 0,44208361\overline{1} \\ 0,27543194\overline{4} \\ 0,24998\overline{4} \end{pmatrix} \tag{40}$$

b) Durch Lösen des Gleichungssystems:

$$\vec{r} = P^T \cdot \vec{r} \tag{41}$$

$$1 = \sum_{i=1}^{|V|} r_i \tag{42}$$

$$\Longleftrightarrow$$

$$\begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{pmatrix} = \begin{pmatrix} 0 & 0,0\overline{3} & 0,0\overline{3} & 0,0\overline{3} \\ 0,9\overline{3} & 0 & 0,0\overline{3} & 0,9\overline{3} \\ 0,0\overline{3} & 0,9\overline{3} & 0 & 0,0\overline{3} \\ 0,0\overline{3} & 0,0\overline{3} & 0,9\overline{3} & 0 \end{pmatrix} \cdot \begin{pmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{pmatrix}$$

$$1 = r_1 + r_2 + r_3 + r_4$$

$$\Longleftrightarrow$$

$$r_1 = \frac{1}{30} \cdot r_2 + \frac{1}{30} \cdot r_3 + \frac{1}{30} \cdot r_4 \tag{43}$$

$$r_2 = \frac{14}{15} \cdot r_1 + \frac{1}{30} \cdot r_3 + \frac{14}{15} \cdot r_4 \tag{44}$$

$$r_3 = \frac{1}{30} \cdot r_1 + \frac{14}{15} \cdot r_2 + \frac{1}{30} \cdot r_4 \tag{45}$$

$$r_4 = \frac{1}{30} \cdot r_1 + \frac{1}{30} \cdot r_2 + \frac{14}{15} \cdot r_3 \tag{46}$$

$$1 = r_1 + r_2 + r_3 + r_4 \tag{47}$$

Als Lösung erhalten wir:

$$r_1 = 0,03225806452 \tag{48}$$

$$r_2 = 0,3328056985 \tag{49}$$

$$r_3 = 0,3221210922 \tag{50}$$

$$r_4 = 0,3128151448 \tag{51}$$

Und daher gilt.

$$\vec{r} = \lim_{i \mapsto \infty} \Pi^{(i)} \approx \begin{pmatrix} 0,03 \\ 0,33 \\ 0,32 \\ 0,31 \end{pmatrix}. \tag{52}$$

Aufgabe 3.6: HITS

Wir haben folgende Matrix A für G:

$$A = \begin{pmatrix} 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix} \tag{53}$$

und weiter,

$$A^T = \begin{pmatrix} 0 & 0 & 0 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 & 0 \end{pmatrix}. \tag{54}$$

Zur Vereinfachung vertzichten wir auf eine Normalisierung:

$$x^{(0)} = y^{(0)} = \frac{1}{8} \cdot \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}^T \tag{55}$$

$$y^{(1)} = A \cdot x^{(0)} = \frac{1}{8} \cdot \begin{pmatrix} 2 & 1 & 2 & 3 & 2 & 1 & 0 & 0 \end{pmatrix}^T \tag{56}$$

$$x^{(1)} = A^T \cdot y^{(0)} = \frac{1}{8} \cdot \begin{pmatrix} 2 & 1 & 3 & 0 & 0 & 2 & 1 & 2 \end{pmatrix}^T \tag{57}$$

$$y^{(2)} = A \cdot x^{(1)} = \frac{1}{8} \cdot \begin{pmatrix} 4 & 3 & 4 & 5 & 5 & 2 & 0 & 0 \end{pmatrix}^T \tag{58}$$

$$x^{(2)} = A^T \cdot y^{(1)} = \frac{1}{8} \cdot \begin{pmatrix} 5 & 3 & 5 & 0 & 0 & 5 & 2 & 3 \end{pmatrix}^T \tag{59}$$

$$y^{(3)} = A \cdot x^{(2)} = \frac{1}{8} \cdot \begin{pmatrix} 7 & 5 & 8 & 13 & 10 & 3 & 0 & 0 \end{pmatrix}^T \tag{60}$$

$$x^{(3)} = A^T \cdot y^{(2)} = \frac{1}{8} \cdot \begin{pmatrix} 10 & 5 & 12 & 0 & 0 & 9 & 4 & 6 \end{pmatrix}^T \tag{61}$$

$$y^{(4)} = A \cdot x^{(3)} = \frac{1}{8} \cdot \begin{pmatrix} 16 & 12 & 15 & 24 & 22 & 6 & 0 & 0 \end{pmatrix}^T \tag{62}$$

$$x^{(4)} = A^T \cdot y^{(3)} = \frac{1}{8} \cdot \begin{pmatrix} 23 & 13 & 22 & 0 & 0 & 21 & 7 & 11 \end{pmatrix}^T \tag{63}$$

$$y^{(5)} = A \cdot x^{(4)} = \frac{1}{8} \cdot \begin{pmatrix} 29 & 22 & 32 & 57 & 45 & 11 & 0 & 0 \end{pmatrix}^T \tag{64}$$

$$x^{(5)} = A^T \cdot y^{(4)} = \frac{1}{8} \cdot \begin{pmatrix} 46 & 24 & 50 & 0 & 0 & 39 & 16 & 21 \end{pmatrix}^T \tag{65}$$

$$y^{(6)} = A \cdot x^{(5)} = \frac{1}{8} \cdot \begin{pmatrix} 86 & 50 & 60 & 109 & 96 & 21 & 0 & 0 \end{pmatrix}^T \tag{66}$$

$$x^{(6)} = A^T \cdot y^{(5)} = \frac{1}{8} \cdot \begin{pmatrix} 102 & 57 & 96 & 0 & 0 & 89 & 29 & 43 \end{pmatrix}^T \tag{67}$$

$$y^{(7)} = A \cdot x^{(6)} = \frac{1}{8} \cdot \begin{pmatrix} 125 & 96 & 132 & 248 & 198 & 43 & 0 & 0 \end{pmatrix}^T \tag{68}$$

$$x^{(7)} = A^T \cdot y^{(6)} = \frac{1}{8} \cdot \begin{pmatrix} 205 & 109 & 212 & 0 & 0 & 169 & 66 & 81 \end{pmatrix}^T \tag{69}$$

$$y^{(8)} = A \cdot x^{(7)} = \frac{1}{8} \cdot \begin{pmatrix} 278 & 212 & 250 & 483 & 417 & 81 & 0 & 0 \end{pmatrix}^T \tag{70}$$

$$x^{(8)} = A^T \cdot y^{(7)} = \frac{1}{8} \cdot \begin{pmatrix} 446 & 248 & 419 & 0 & 0 & 380 & 125 & 175 \end{pmatrix}^T \tag{71}$$

Wir erhalten das folgende Ranking für Authorities und Hubs:

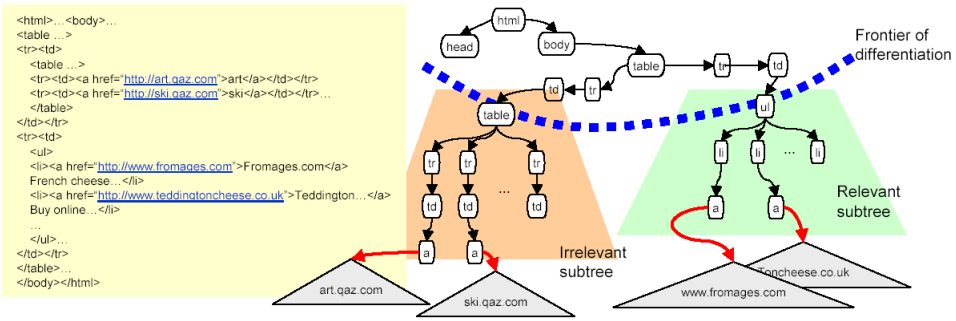
Ranking	1	2	3	4	5	6	7
Authorities	1	3	6	2	8	7	4,5
Hubs	4	5	1	3	2	6	7,8

Aufgabe 3.7: Erweitertes HITS

Both link analysis algorithms, HITS and Page Rank, use a coarse-grained model of the Web, where each page is a node in a graph with two scores (one score for Page Rank, respectively) associated with it. The model takes no notice either of text or the markup structure on each page. HITS leaves the selection of the base set to an external IR algorithm, e.g., an external search engine.

a) A special problem with decoupling a user's query from the link based ranking strategy is that often hubs are *mixed* without any attempt on the part of the hub writer to confound a search engine, i.e. hubs often contain multiple collections of links on totally different topics. For example, a hub u containing links relevant to the query computer scientistmmay also have some links on hobbies of researchers and other topics. If a significant amount of different hubs point to an out-of-topic page v_1, this page gains authority, and the HITS algorithm would diffuse the score through u to some descendant node v_2 of v_1, which could be a different topic but determined to be good hub for the initial query. However, a system may succeed at suppressing the ill effects by filtering all result pages on their keywords at query time.

b) An obvious clue that helps users identify relevant zones on a multitopic page is text similarity. A simple approach for segmenting hubs would be to give preferential treatment to document parts where query terms appear frequently. If different topics can be separated well by the HTML tag structure, e.g., when links on research are preceded by a certain tag <heading2> *My research*, and the ones on hobbies are well separated and preceded by <heading2> *My hobbies*. The tags would mark different DOM (**D**ocument **O**bject **M**odel) subtrees (see Tab. 7), each identified with a characteristic root tag. If we knew these separating tags in the DOM, there would be no problem in identifying an appropriate frontier for segmentation, and we could simply compare the similarity of different parts using standard similarity measure such as scalar product or cosine measure.



Humans are rarely misled by irrelevant links because they interpret HTML page idioms to locate content-bearing regions (i.e., HTML tag-subtrees), assisted by text in those regions. Pages are differentiated into relevant (green) and irrelevant (red) regions by their term distribution as well as links to known relevant or irrelevant sites.

However, this scheme will not work well for some queries. In practice, HTML does not provide this clear tagging structure that would be required for an automated system to simply decide where to cut different DOM subtrees (this would more likely be the case with XML). On the other hand, authors of pages are rarely willing to invest additional effort in annotating their pages in detail; and sometimes it can even be a strategy *not* to give detailed descriptions about contents. For example, for the query *Japanese car maker*, DOM subtrees with links to *www.honda.com* and *www.toyota.com* rarely use any of these query words; they just use names of the companies, such as *Honda* and *Toyota*.

Therefore, depending on direct syntactic matches between query terms and the text in DOM subtrees can be unreliable. A solution for this would be the use of the *centroid* of the root set features for *Honda* and *Toyota* with large weights. To estimate the relevance of a DOM subtree rooted at node u with regard

to a query q, we can simply measure the vector-space similarity between the root set centroid for q and the text in the DOM subtree, associating u with this score.

Aufgabe 3.8: HITS mit SVD

There is a direct mapping between finding the singular value decomposition (SVD) of the adjacency matrix E of a Web graph and the Eigenvectors of EE^T or E^TE. Let the SVD of E be $U\Delta V^T$, where $U^TU = I$ and $V^TV = I$ and Δ is the diagonal matrix $diag(\delta_1, \ldots, \delta_r)$ of singular values, where r is the rank of E, and I is an identity matrix of suitable size. Then $EE^T = U\Delta\underline{V^TV}\Delta U^T = U\Delta^2U^T$, which implies that $EE^TU = U\Delta^2$. Here if E is a $n \times n$ matrix with rank r, then U is $n \times r$; Δ and Δ^2 are $r \times r$. Specifically, $\Delta^2 = diag(\delta_1^2, \ldots, \delta_r^2)$. $U\Delta^2$ is $n \times r$ as well. If $U(j)$ is the jth column of U, we can write $EE^TU(j) = \delta_j^2U(j)$, which means that $U(j)$ is an Eigenvector of EE^T with corresponding Eigenvalue δ_j^2, for $j = 1, \ldots, r$. If Δ^2 is arranged such that $\delta_1^2 \geq \ldots \geq \delta_r^2$, it turns out that finding the hubs scores for E is the same as finding $U(1)$, and more generally, finding multiple hubs/authorities corresponds to finding many singular values of EE^T and E^TE.