Kapitel 9: Markov-Ketten als Grundlage der Funktionsweise von Suchmaschinen im Internet

Situation:

Eingabe: eine Anfrage, bestehend aus einem oder mehreren Stichworten

Ziel: Bestimme diejenigen Webseiten, deren Inhalt hinsichtlich der Anfrage-Stichworte am "informativsten" ist.

Probleme:

- Es gibt sehr viele Webseiten (in 2005: mehr als 8 Milliarden Stück)
- Ständig kommen neue hinzu
- Viele Webseiten werden (fast) täglich aktualisiert; viele werden nach einiger Zeit auch wieder gelöscht.
- Suchanfragen müssen in Echtzeit beantwortet werden
- Welches sind die "informativsten" Webseiten?

9.1 Die Architektur von Suchmaschinen

Herausforderung:

Für einen sich rasant ändernden Suchraum gigantischer Größe müssen Anfragen ohne merkliche Reaktionszeit beantwortet werden.

Um dies zu gewährleisten, nutzen Suchmaschinen folgende Komponenten:

- Der Crawler durchforstet das Internet, um neue oder veränderte Webseiten zu bestimmen.
- Abspeichern der wichtigsten Informationen: Die vom Crawler gefundene Information muss aufbereitet und gespeichert werden
- Indizierung: Die Datenstruktur der gespeicherten Informationen muss in Echtzeit alle (relevanten) Webseiten bestimmen, die die Anfrage-Stichworte enthalten.

- <u>Bewertung der Webseiten:</u> Die ausgewählten Webseiten müssen im Hinblick auf ihren Informationsgehalt bewertet werden.

<u>Details zum Thema "Indizierung":</u>

- Es werden alle Worte einer jeden Webseite erfasst.
- Fundamentale Datenstruktur: der <u>invertierte Index</u>, der zu jedem Stichwort s alle Webseiten bestimmt, die s enthalten. Oft liefert der invertierte Index noch Zusatzinformationen, die die Wichtigkeit des Stichworts innerhalb der Webseite beschreiben (etwa: Häufigkeit des Stichworts, seine Schriftgröße, das Vorkommen des Stichworts in Beschriftungen von Links <u>auf</u> die Webseite).

  Insgesamt: Der invertierte Index ist Text-basiert.
- Eine weitere wichtige Datenstruktur: der <u>Link-Index</u>, mit dem die Graph-Struktur des Internets modelliert wird.

Ansatz: Jede Webseite wird durch einen Knoten repräsentiert.
Eine Kante von Knoten i zu Knoten j entspricht einem Link von Webseite i auf Webseite j.

Der Link-Index (oder: Webgraph) wird üblicherweise als Adjazenzliste gespeichert.

Details zum Thema "Bewertung der Webseiten":

- Ziel: Bestimme die für die gegebenen Anfrage-Stichworte informativsten Webseiten, d.h. sortiere die Webseiten, die die Anfrage-Stichworte enthalten, nach ihrer "Relevanz"

Dabei werden u.a. folgende Kriterien berücksichtigt:

1.) Häufigkeit und Positionierung der Suchbegriffe auf der jeweiligen Webseite sowie in der Beschriftung von Links, die auf diese Webseite verweisen.

2.) Die grundlegende Bedeutung einer Webseite.
Bei Google wird dies durch den so genannten Page-Rank der Webseite modelliert.

Ein etwas anderer Ansatz zur Bewertung der grundlegenden Bedeutung einer Webseite ist der "Hubs and Authorities"-Ansatz von Kleinberg.

Beide Ansätze betrachten allein den Webgraphen (d.h. die Link-Struktur des Internets, nicht den textuellen Inhalt der Webseiten) und gehen von der folgenden Annahme aus:

Wenn eine Webseite i einen Link auf Webseite j enthält, dann

- gibt es eine inhaltliche Beziehung zwischen beiden Webseiten und
- der Autor der Webseite i hält die Informationen auf Webseite j für wertvoll.

Beide Ansätze versuchen, die "relative Wertschätzung" zwischen Webseiten in eine "absolute Relevanz" der Webseiten umzurechnen.

Im Rest von Kapitel 9 werden wir uns etwas genauer ansehen, wie die "absolute Relevanz" einer Webseite modelliert und berechnet werden kann.

9.2 Page-Rank, Zufalls-Surfer und Markov-Ketten

Sei $G = (V, E)$ mit $V = \{1, 2, \dots, n\}$ und $E \subseteq V \times V$ der Webgraph.

D.h.: Knoten von G repräsentieren Webseiten, und jede Kante $(i, j) \in E$ modelliert einen Link von Webseite i auf Webseite j.

Für jeden Knoten $i \in V$ sei

$$a_i := \text{Aus-Grad}_G(i)$$

der Ausgangsgrad von i in G.
D.h. a_i ist die Anzahl der Hyperlinks, die von Webseite i auf andere Webseiten verweisen.

Die "grundlegende Bedeutung" einer Webseite i wird im Folgenden durch eine Zahl PR_i (dem sogenannten Page-Rank von i) modelliert.

Der Wert PR_i soll die Qualität von Webseite i widerspiegeln (je höher, desto besser) und wird durch einen "Peer-Review" bestimmt:

- Der Page-Rank PR_j von Webseite j ist hoch, wenn viele Webseiten i mit hohem Page-Rank PR_i auf die Seite j verweisen.
- Eine Webseite i mit a_i ausgehenden Links "vererbt" ihren Page-Rank dabei anteilig an alle Webseiten j mit $(i,j) \in E$ um den Anteil $\frac{PR_i}{a_i}$ weiter.

In dieser Sichtweise müsste also f.a. $j \in V$ gelten:

$$PR_j = \sum_{\substack{i \in V: \\ (i,j) \in E}} \frac{PR_i}{a_i} . \qquad (*)$$

Aus praktischen Gründen (mehr dazu: später) wird die Vererbung von PR_i auf die "Nachfolgeseiten" j mit $(i,j) \in E$ meistens um einen "Dämpfungsfaktor" d (mit $0 \leq d \leq 1$) abgeschwächt und die Werte werden so normiert, dass gilt: $PR_1, \dots, PR_n \geq 0$ und $\sum_{i \in V} PR_i = 1$:

Definition 9.1 (Page-Rank)

Sei d eine reelle Zahl mit $0 \le d \le 1$;
d wird im Folgenden auch Dämpfungsfaktor genannt.

Sei $G = (V, E)$ mit $V = \{1, \dots, n\}$ und $E \subseteq V \times V$
der Webgraph und sei f.a. $i \in V$ $a_i := \text{Aus-Grad}_G(i)$.

Ein Tupel $PR = (PR_1, \dots, PR_n)$ von reellen Zahlen
≥ 0 hat die Page-Rank-Eigenschaft (bzgl. Dämpfungsfaktor d),
wenn gilt:

(a) $\sum_{i \in V} PR_i = 1$ und

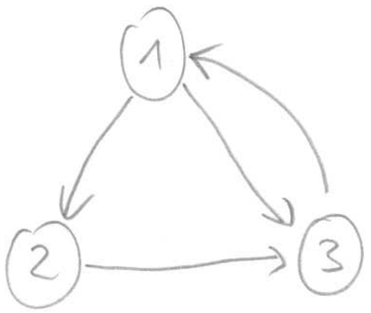
(b) f.a. $j \in V$ gilt:

$$PR_j = \frac{1-d}{n} + d \cdot \sum_{\substack{i \in V: \\ (i,j) \in E}} \frac{PR_i}{a_i} \qquad (**)$$

(Beachte: Für $d = 1$ erhält man gerade die Gleichung (*);
für $d = 0$ ist $PR_1 = PR_2 = \dots = PR_n = \frac{1}{n}$.)

Beispiel 9.2

Zur Veranschaulichung der Page-Rank-Eigenschaft betrachten wir den Dämpfungsfaktor $d := 1/2$ und den folgenden (sehr kleinen) Webgraphen, der aus nur drei Dokumenten besteht:



Wir suchen ein Tupel $PR = (PR_1, PR_2, PR_3)$ von reellen Zahlen ≥ 0, das die Page-Rank-Eigenschaft bzgl $d = 1/2$ hat, d.h es gilt:

0) $PR_1 + PR_2 + PR_3 = 1$

1) $PR_1 = \frac{1}{2 \cdot 3} + \frac{1}{2} \cdot PR_3$

2) $PR_2 = \frac{1}{6} + \frac{1}{2 \cdot 2} \cdot PR_1$

3) $PR_3 = \frac{1}{6} + \frac{1}{2 \cdot 2} \cdot PR_1 + \frac{1}{2} \cdot PR_2$

Eine Lösung dieses linearen Gleichungssystems (z.B. mittels Gauß-Elimination) liefert:

$PR_1 = \frac{14}{39}$, $PR_2 = \frac{10}{39}$, $PR_3 = \frac{15}{39}$.

Eine etwas andere Schreibweise der Bedingung (b) in Definition 9.1 zeigt, dass die Tupel $PR = (PR_1, \dots, PR_n)$, die die Page-Rank-Eigenschaft (bzgl d) besitzen, gerade die Eigenvektoren zum Eigenwert 1 der folgenden Matrix

$$P = \left(p_{i,j} \right)_{i,j=1,\dots,n} = \begin{pmatrix} p_{1,1} & \cdots & p_{1,n} \\ \vdots & & \vdots \\ p_{n,1} & \cdots & p_{n,n} \end{pmatrix}$$

sind: F.a. $i,j \in \{1,\dots,n\}$ sei

$$p_{i,j} := \begin{cases} \frac{1-d}{n} + \frac{d}{a_i} & \text{, falls } (i,j) \in E \\ \frac{1-d}{n} & \text{, falls } (i,j) \notin E \end{cases}$$

Beobachtung 9.3: Für jedes Tupel $X = (X_1, \dots, X_n)$ von reelen Zahlen mit $\sum_{i=1}^{n} X_i = 1$ gilt:

$$X \cdot P = \left(\frac{1-d}{n} + d \cdot \sum_{\substack{i \in V: \\ (i,j) \in E}} \frac{X_i}{a_i} \right)_{j=1,\dots,n}$$

Beweis:

$$X \cdot P = (X_1, \dots, X_n) \cdot \begin{pmatrix} p_{1,1} & \cdots & p_{1,n} \\ \vdots & & \vdots \\ p_{n,1} & \cdots & p_{n,n} \end{pmatrix}.$$

Die j-te Komponente von $X \cdot P$ ist also:

$$\sum_{i=1}^{n} X_i \cdot p_{i,j} \underset{\text{Def } p_{i,j}}{=} \sum_{i=1}^{n} X_i \cdot \frac{1-d}{n} + \sum_{\substack{i \in V: \\ (i,j) \in E}} X_i \cdot \frac{d}{a_i}$$

$$= \frac{1-d}{n} \cdot \sum_{i=1}^{n} X_i + d \cdot \sum_{\substack{i \in V: \\ (i,j) \in E}} \frac{X_i}{a_i}$$

$$\underset{\sum_{i=1}^{n} X_i = 1}{=} \frac{1-d}{n} + d \cdot \sum_{\substack{i \in V \\ (i,j) \in E}} \frac{X_i}{a_i}. \qquad \square$$

Folgerung 9.4:

Ein Tupel $PR = (PR_1, \dots, PR_n)$ von reellen Zahlen ≥ 0 hat genau dann die Page-Rank-Eigenschaft (bzgl d), wenn gilt:

(a) $\sum_{i=1}^{n} PR_i = 1$ und

(b) $PR \cdot P = PR$,

d.h. PR ist ein Eigenvektor (zum Eigenwert 1) der Matrix P.

(bzw. eine Lösung des linearen Gleichungssystems

$PR \cdot \left(\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} - P \right) = 0$

Um den Page-Rank der einzelnen Webseiten zu berechnen, müssen wir also "nur" einen Eigenvektor zum Eigenwert 1 der Matrix P bestimmen

(z.B. durch Lösen des linearen Gleichungssystems

$$PR \cdot \left(\begin{pmatrix} 1 & & 0 \\ & \ddots & \\ 0 & & 1 \end{pmatrix} - P \right) = 0 \quad).$$

Probleme:

- Das ist sehr aufwändig, da der Webgraph sehr groß ist.
- Ist "der" Eigenvektor zum Eigenwert 1 überhaupt eindeutig bestimmt? Muss es einen solchen geben? Kann es evtl. verschiedene solche Eigenvektoren von P geben?

Um diese Probleme zu bewältigen, hilft uns die Theorie der Markov-Ketten sowie die folgende Sichtweise auf die Matrix P:

Der Zufalls-Surfer (Random-Surfer):

Ein Zufalls-Surfer beginnt auf einer beliebigen Webseite und verfolgt beliebige Links, ohne dabei auf Inhalte zu achten.

Der Eintrag $p_{i,j}$ in der Matrix P gibt dabei die Wahrscheinlichkeit an, mit der der Zufalls-Surfer in einem Schritt von Seite i zu Seite j wechselt. Die Wahl

$$p_{i,j} = \begin{cases} \frac{1-d}{n} + \frac{d}{a_i}, & \text{falls } (i,j) \in E \\ \frac{1-d}{n}, & \text{falls } (i,j) \notin E \end{cases}$$

bedeutet dabei: Wenn der Zufalls-Surfer auf Webseite i ist, so wählt er

- mit Wahrscheinlichkeit d einen Link, der von Seite i ausgeht — und jeder der a_i ausgehenden Links wird mit derselben Wahrscheinlichkeit, d/a_i, ausgewählt
- mit Wahrscheinlichkeit $(1-d)$ eine <u>beliebige</u> Webseite im Webgraphen — und jede der n Webseiten wird mit derselben Wahrscheinlichkeit, $\frac{1-d}{n}$, ausgewählt.

<u>Beachte:</u> Die Werte $p_{i,j}$ können wir tatsächlich als Wahrscheinlichkeiten interpretieren, da gilt: $p_{i,j} \geq 0$ (f.a. $i,j \in \{1,\dots,n\}$) und

f.a. $i \in \{1,\dots,n\}$ gilt: $\sum_{j=1}^{n} p_{i,j} = 1$

(denn: $\sum_{j=1}^{n} p_{i,j} = \sum_{j=1}^{n} \frac{1-d}{n} + \sum_{\substack{j \in V: \\ (i,j) \in E}} \frac{d}{a_i} = (1-d) + a_i \cdot \frac{d}{a_i} = 1$).

Mit dieser Sichtweise auf die Matrix P gilt für den "modifizierten Webgraphen" $G' = (V, E')$ mit $V = \{1, \dots, n\}$ und $E' = V \times V$, dass (G', P) eine Markov-Kette im folgenden Sinn ist:

(Schreibweisen: Markov, Markow, Markoff.)

<u>Definition 9.5</u> (Markov-Kette)

Eine <u>Markov-Kette</u> $(\hat{G}, \hat{P})$ besteht aus einem gerichteten Graphen $\hat{G} = (\hat{V}, \hat{E})$ mit $\hat{V} = \{1, \dots, n\}$ (für ein geeignetes $n \in \mathbb{N}$) und einer Matrix

$$\hat{P} = (\hat{p}_{i,j})_{i,j=1\dots n} = \begin{pmatrix} \hat{p}_{1,1} & \cdots & \hat{p}_{1,n} \\ \vdots & & \\ \hat{p}_{n,1} & \cdots & \hat{p}_{n,n} \end{pmatrix}$$

so dass gilt:

<u>(1)</u> f.a. $i,j \in \{1 \dots n\}$ mit $(i,j) \notin \hat{E}$ ist $\hat{p}_{i,j} = 0$

und

<u>(2)</u> für jede Zeile $i \in \{1, \dots, n\}$ gilt:

$$\sum_{j=1}^{n} p_{i,j} = 1$$

Die Matrix $\hat{P}$ heißt Übergangsmatrix.

Falls zusätzlich gilt:

<u>(3)</u> f.a. $i,j \in \{1 \dots n\}$ ist $\hat{p}_{i,j} \geq 0$,

so heißt $\hat{P}$ auch <u>stochastische Matrix</u>.

Der Eintrag $\hat{p}_{i,j}$ kann dann als Wahrscheinlichkeit aufgefasst werden, mit der ein "Zufalls-Surfer im Graphen $\hat{G}$" in einem Schritt von Knoten i zu Knoten j springt.

<u>Definition 9.6</u> (Stationäre Verteilung)

Sei $(\hat{G}, \hat{P})$ eine Markov-Kette mit $\hat{G} = (\hat{V}, \hat{E})$ und $\hat{V} = \{1, \ldots, n\}$.

(a) Eine <u>Verteilung auf $\hat{V}$</u> ist ein Vektor $x = (x_1, \ldots, x_n)$ von reellen Zahlen ≥ 0 mit $\sum_{i=1}^{n} x_i = 1$.

(x_i kann dabei als Wahrscheinlichkeit dafür aufgefasst werden, dass der "Zufalls-Surfer auf $\hat{G}$" sich zu Beginn auf Knoten i befindet).

(b) Eine Verteilung x auf $\hat{V}$ heißt <u>stationäre Verteilung</u>, falls gilt: $x \cdot P = x$

(d.h. x ist ein Eigenvektor zum Eigenwert 1 von P)

Beachte: Gemäß Folgerung 9.4 gilt für den "modifizierten Webgraphen G'" und die Matrix P folgendes:

Ein Tupel $PR = (PR_1, \dots, PR_n)$ hat genau dann die Page-Rank-Eigenschaft (bzgl d), wenn PR eine stationäre Verteilung (für die Markov-Kette (G', P)) ist.

Die Theorie der Markov-Ketten wurde in der Literatur gut untersucht. Insbesondere ist ein nützliches Kriterium bekannt, das garantiert, dass eine stationäre Verteilung existiert, eindeutig bestimmt ist und sehr effizient näherungsweise berechnet werden kann: Dies ist bei so genannten ergodischen Markov-Ketten (siehe Definition 9.7) der Fall.

Notation:

Ist $\hat{P}$ eine $n \times n$-Matrix und $k \geq 1$ eine natürliche Zahl, so ist

$$\hat{P}^k := \underbrace{\hat{P} \cdot \ldots \cdot \hat{P}}_{k\text{-mal}}$$ die $n \times n$-Matrix, die aus

dem Matrix-Produkt von k Kopien von $\hat{P}$ entsteht.

Wenn $\hat{P}$ eine stochastische Matrix ist, so können wir den Eintrag $(\hat{P}^k)_{i,j}$ in Zeile i und Spalte j der Matrix $\hat{P}^k$ auffassen als die Wahrscheinlichkeit, dass der "Zufalls-Surfer auf $\hat{G}$" innerhalb von genau k Schritten von Knoten i zu Knoten j gelangt.

Definition 9.7 (Ergodische Markov-Ketten)

Eine Markov-Kette $(\hat{G}, \hat{P})$ mit $\hat{G} = (\hat{V}, \hat{E})$ und $\hat{V} = \{1, \ldots, n\}$ heißt ergodisch, wenn f.a. $i_1, i_2 \in \{1, \ldots, n\}$ und alle $j \in \{1, \ldots, n\}$ gilt: Die Grenzwerte $\lim\limits_{k \to \infty} (\hat{P}^k)_{i_1,j}$ und $\lim\limits_{k \to \infty} (\hat{P}^k)_{i_2,j}$ existieren und es gilt: $\lim\limits_{k \to \infty} (\hat{P}^k)_{i_1,j} = \lim\limits_{k \to \infty} (\hat{P}^k)_{i_2,j} > 0$.

Beachte:

Ist $(\hat{G}, \hat{P})$ ergodisch, so gilt:

1) Die Matrix

$$\hat{P}^{\infty} := \lim_{k\to\infty} (\hat{P}^k) := \left(\lim_{k\to\infty} (\hat{P}^k)_{ij} \right)_{i,j=1 \dots n}$$

ist wohldefiniert (da die Grenzwerte existieren),

und

2) alle Zeilen in $\hat{P}^{\infty}$ sind identisch.

Notation: $\pi^{\infty} := (\pi^{\infty}_1, \dots, \pi^{\infty}_n)$ sei die 1. Zeile von $\hat{P}^{\infty}$.

Somit ist $\hat{P}^{\infty} = \begin{pmatrix} \pi^{\infty} \\ \vdots \\ \pi^{\infty} \end{pmatrix} = \begin{pmatrix} \pi^{\infty}_1 & \cdots & \pi^{\infty}_n \\ \pi^{\infty}_1 & \cdots & \pi^{\infty}_n \\ \vdots & & \vdots \\ \pi^{\infty}_1 & \cdots & \pi^{\infty}_n \end{pmatrix}$, und

für jede Verteilung $X = (X_1, \dots, X_n)$ gilt: $X \cdot \hat{P}^{\infty} = \pi^{\infty}$.

Daher gilt: Wenn der „Zufalls-Surfer auf $\hat{G}$“ seinen Startknoten gemäß einer beliebigen Anfangsverteilung $X = (X_1, \dots, X_n)$ wählt und hinreichend viele zufällige Schritte macht, so ist die Wahrscheinlichkeit, bei Knoten $j \in \hat{V}$ zu landen, beliebig nah bei π^{∞}_j.

Die Wahl des Anfangsknotens ist für einen hinreichend langen "Random Walk" also ohne Belang, und die "Grenzverteilung" π^∞ einer ergodischen Markov-Kette "vergisst" den Startpunkt eines "Random Walk".

Daher gilt folgender Satz:

<u>Satz 9.8</u>

Jede ergodische Markov-Kette $(\hat{G}, \hat{P})$ besitzt eine <u>eindeutig bestimmte stationäre Verteilung</u> γ, und es gilt: $\underline{\gamma = \pi^\infty}$.

<u>Beobachtung 9.9</u>:

Ist G' der "modifizierte Webgraph" und P die Übergangsmatrix für Dämpfungsfaktor d $(0 \le d \le 1)$, so gilt:

(a) Falls $d < 1$, so ist die Markov-Kette (G', P) ergodisch (hier ohne Beweis), und gemäß Satz 9.8 ist π^∞ das eindeutig bestimmte Tupel, das die Page-Rank-Eigenschaft bzgl d hat.

(b) Falls $d=1$ ist, so ist (G', P) nicht unbedingt ergodisch, denn der Zufalls-Surfer wird z.B. auf Webseiten, aus denen keine Hyperlinks heraus führen gefangen.

(Das ist der auf Seite 9.7 erwähnte "praktische Grund" wegen dem der Dämpfungsfaktor d eingeführt wurde).

9.3 Die effiziente Berechnung des Page-Rank

Wir wählen einen Dämpfungsfaktor $d<1$ und nutzen Beobachtung 9.9 (a).

Als Anfangsverteilung für den Zufalls-Surfer wählen wir die Gleichverteilung $X^{(0)} := (\underbrace{\tfrac{1}{n}, \dots, \tfrac{1}{n}}_{n\text{-mal}})$

Für $k=1,2,3,\dots$ berechnen wir nach und nach die Vektoren $X^{(k)} := X^{(k-1)} \cdot P$.

Somit gilt f.a. $k \in \mathbb{N}$: $X^{(k)} = X^{(0)} \cdot P^k$.

Da (G', P) ergodisch ist, wissen wir, dass

$PR := \pi^{\infty}$ das eindeutig bestimmte Tupel ist, das die Page-Rank-Eigenschaft bzgl. d hat.

Außerdem gilt: $\pi^{\infty} = \lim_{k\to\infty} \left(X^{(0)} \cdot P^k\right) = \lim_{k\to\infty} \left(X^{(k)}\right)$

Zur näherungsweisen Berechnung von π^{∞} beginnen wir mit $X^{(0)} = \left(\frac{1}{n}, \dots, \frac{1}{n}\right)$ und berechnen nacheinander für $k = 1, 2, 3, \dots$ den Vektor $X^{(k)} := X^{(k-1)} \cdot P$.

Aufgrund des hohen Zusammenhangs des Webgraphen konvergiert die Folge $\left(X^{(0)}, X^{(1)}, X^{(2)}, \dots\right)$ sehr schnell gegen π^{∞}.

Für eine schnelle Berechnung des Vektor-Matrix-Produkts $X^{(k)} := X^{(k-1)} \cdot P$ wird ausgenützt, dass P viele "identische" Einträge der Form $\frac{1-d}{n}$ hat. Außerdem ist die Berechnung des Vektor-Matrix-Produkts sehr gut parallelisierbar.

Derzeit werden mehrere Tausend PCs eingesetzt, die mehrere Stunden zur Berechnung des Page-Ranks benötigen.

(... was in Anbetracht der Tatsache, dass es mehrere Milliarden Webseiten gibt, erstaunlich gering ist).

Mehr zum Thema "Suchmaschinen" und "Markov-Ketten" können Sie in der Veranstaltung "Internet Algorithmen" bei Prof. Georg Schnitger lernen!