# Das Google-PageRank-System

## Mit Markoff-Ketten und linearen Gleichungssystemen Ranglisten erstellen

Google ist mittlerweile *die* Suchmaschine im WWW schlechthin geworden. Jede(r) hat viel Erfahrung damit, und es erhebt sich die Frage: Wie schafft es Google, dass *wichtige* Seiten zum gesuchten Thema (bzw. Begriff) am *Anfang* der Liste stehen? Mathematisch betrachtet, stecken „Mehrstufige Prozesse" (bzw. „Markoff-Ketten") hinter dem so genannten PageRank-System, mit dem die einzelnen Seiten, die Google zu einem Stichwort findet, in eine Reihenfolge gebracht werden.

Das Thema „Mehrstufige Prozesse" (in einer elementaren Form!) ist in manchen deutschen Bundesländern möglicher Lehrstoff in der Oberstufe. Es gehört einerseits zur Linearen Algebra („Übergangsmatrizen"), andererseits zur Stochastik (Wahrscheinlichkeiten, relative Häufigkeiten) bzw. zur Analysis (Grenzwerte). In der Tat ist es ein Gebiet, in dem der Vernetzungsgedanke sehr gut verwirklicht werden kann, und *Vernetzung* gehört sicher zu jenen Begriffen, die in den Präambeln von so gut wie allen Lehrplänen als positiv und wünschenswert hervorgehoben werden.

Kamen „Mehrstufige Prozesse" (bzw. „Markoff-Ketten") schon im Unterricht vor, dann bietet das PageRank-System eine weitere Anwendung. Andererseits kann man es als Einstieg in das Thema nutzen. In der didaktischen Literatur werden mehrstufige Prozesse relativ häufig aufgegriffen (vgl. die unten angeführte Literatur), auch in deutschen Schulbüchern (z. B. Lambacher-Schweizer 2001).

### Einstieg in Markoff-Ketten

Als Einstiegsaufgabe ist die Formulierung in **Kasten 1** denkbar. Auch wenn die Schülerinnen und Schüler noch nichts von Markoff-Ketten und Übergangsmatrizen gehört haben, können sie das Problem leicht lösen, am besten vielleicht mit einer Tabellenkalkulation (z. B. EXCEL). Die zugehörigen Rekursionen können aus dem *Übergangsgraphen* direkt ablesen und dort als Formel eingegeben werden:

$$0{,}8A_n + 0{,}3B_n + 0{,}2C_n = A_{n+1}$$
$$0{,}1A_n + 0{,}6B_n + 0{,}1C_n = B_{n+1}$$
$$0{,}1A_n + 0{,}1B_n + 0{,}7C_n = C_{n+1}$$

Durch das „Herunterziehen" der Formel in der Tabellenkalkulation können die entsprechenden Werte für große $n$ schnell abgelesen werden (allein mit einem Taschenrechner wäre das ungleich mühsamer). Die Werte pendeln sich schnell bei $(A_n, B_n, C_n)$ = (55 %, 20 %, 25 %) ein, und zwar unabhängig von der Anfangsverteilung (stabile „Grenzverteilung").

Genau solche Grenzverteilungen werden im Folgenden im (mathematischen) Zentrum stehen. Für das experimentelle Ermitteln solcher Grenzverteilungen zwischen einigen wenigen möglichen „Stationen" (im Einstiegsbeispiel nur drei: A-tel, B-tel, C-tel) ist *Tabellenkalkulation* sehr gut geeignet. Man braucht dabei weder Matrizen noch weitere Theorie dazu. Die Grenzverteilung wird *iterativ* bestimmt. Dieses iterative Vorgehen mit der Tabellenkalkulation ist einerseits ein einfacher Einstieg, andererseits aber gar nicht so weit weg von der prinzipiellen Vorgehensweise, wie solche „Grenzverteilungen" bei riesigen Dimensionen bestimmt werden: Beim PageRank-Algorithmus geschieht das Lösen des zugehörigen linearen Gleichungssystems auch iterativ (näherungsweise) und nicht durch eine geschlossene Formel oder Ähnliches.

Für eine etwas weiter und tiefer gehende Auseinandersetzung mit „Grenzverteilungen" (insbesondere theoretische Aspekte) kommt man mit der Tabellenkalkulation nicht mehr aus, da braucht man dann *Übergangsmatrizen.*

### Google und die Anfänge

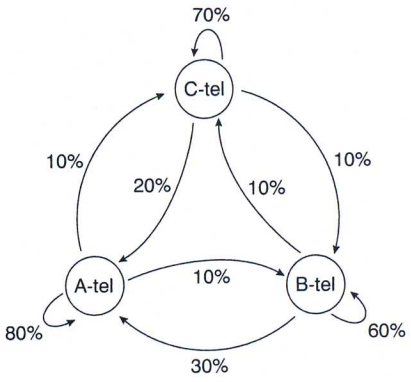
Es gibt sehr viele „Suchmaschinen", die das WWW in Bruchteilen von Sekunden durchforsten können, wobei „wichtige" Seiten jeweils zuerst aufgelistet werden sollen.

Eine besonders häufig verwendete Suchmaschine ist Google, wobei der Name Google „etwas Riesengroßes" bezeichnen sollte – nach der unglaublichen Fülle des WWW. Er kommt von

❶ ZUM AUSPROBIEREN

### Einstiegsaufgabe

Der Telefonmarkt in einem Land sei durch drei Firmen bestimmt (A-tel, B-tel und C-tel). Diese Telefongesellschaften schließen mit ihren Kunden jeweils Jahresverträge ab. Der Einfachheit halber nehmen wir an, dass diese Verträge jeweils genau ein Kalenderjahr gelten mit Wechselmöglichkeiten jeweils zu Jahresende/Jahresbeginn.
Die Kunden bleiben am Ende des Jahres zu einem bestimmten Prozentsatz bzw. wechseln zu anderen Betreibern. Dies kann man am einfachsten mit einem so genannten *gerichteten Graphen* (s. rechts, auch *Übergangsgraph* genannt) beschreiben: Dies bedeutet für die Firma C-tel zum Beispiel: 70 % der C-tel-Kunden bleiben nach Ablauf des Jahresvertrages bei C-tel, 20 % wechseln zu A-tel und 10 % zu B-tel. Analog sind die anderen relativen Übergangshäufigkeiten zu interpretieren.




→ Angenommen, diese Übergangsraten bleiben über 5 (10; 20) Jahre konstant. Wie sieht die Verteilung der Kunden auf die einzelnen Firmen aus, wenn zu Beginn jede Firma 1/3 der Kunden hat? Wie wären die entsprechenden Werte, wenn die „Anfangsverteilung" nicht $(A_0, B_0, C_0) = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, sondern $(A_0, B_0, C_0)$ = (30 %, 50 %, 20 %) wäre?

einer Abwandlung des Begriffes „Googol", ein Wort, das in der ersten Hälfte des 20. Jahrhunderts in Amerika für die riesengroße Zahl $10^{100}$ etabliert wurde. Ein Googol ist zwar deutlich größer als die Anzahl der Atome im sichtbaren Universum (ca. $10^{80}$), andererseits ist ein Googol aber „nur" etwa $70 \cdot 69 \cdot 68 \cdot \ldots \cdot 2 \cdot 1 =$ 70! und entspricht somit der Anzahl der Möglichkeiten, 70 verschiedene Gegenstände in einer Reihe anzuordnen.

Da merkt man, wie mächtig so manche Schreibweisen in der Mathematik sind und wie schnell Fakultäten wachsen: 70! ist ca. $10^{20}$-mal so groß wie die Anzahl der Atome im sichtbaren Universum! Wer würde so aus dem Bauch heraus nicht glauben, dass es doch viel, viel mehr Atome im Universum gibt als Möglichkeiten, 70 Gegenstände in einer Reihe anzuordnen?

Google ging am 7. September 1998 als Testversion online und ist unter den Suchmaschinen die Nummer 1. Der Grund dafür liegt im „PageRank-Algorithmus", der Google damals schneller machte als die Konkurrenz. Gerade beim Aufteilen des potentiellen Marktes ist es wichtig, besser zu sein als die Konkurrenz. Auch wenn die Qualität der anderen Suchmaschinen seitdem gestiegen ist, gibt es selbst bei gleich guter Qualität für die vielen Anwender von Google (ich selbst gehöre auch dazu) keinen Grund zum Umsteigen. Dies hat einfach auch mit Gewohnheit zu tun: Warum soll man wechseln, wenn anderswo die Qualität nicht besser ist?

## Wie arbeitet die Suchmaschine?

Wird ein Stichwort (oder mehrere) eingegeben, beginnen die Suchmaschinen, mit einem so genannten *spider* oder *webcrawler* (spezielles Computerprogramm) das WWW zu „durchforsten": Auf welchen Seiten kommt der gesuchte Begriff vor, wo ist etwas zu ihm zu erfahren? Ziel dieser umfangreichen Suchtätigkeit ist es, eine möglichst gute „Momentaufnahme" der In-

❷ ZUM AUSPROBIEREN

### Das WWW als gerichteter Graph

*A, B, C, D* seien vier verschiedene Internetseiten, die auf eine einfache Weise miteinander verlinkt sind. So gibt es von Seite *A* einen Link zu den Seiten *B* und *C*; von der Seite *B* gibt es Links zu den Seiten *C* und *D*, von der Seite *D* gibt es Links zu *B* und *C* und von Seite *C* kommt man nur zur Seite *A*.


Foto: © Guy Erwood – Fotolia.com

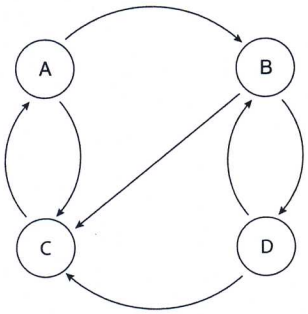
**Surfen ist, von Seite zu Seite „klicken". Wie verteilen sich langfristig die User bei nur vier Seiten (A, B, C, D)?**

**→ AUFGABEN**

1. Zeichnen Sie einen entsprechenden gerichteten Graphen (Übergangsgraph").
2. Angenommen, alle Links auf einer Seite werden mit derselben Wahrscheinlichkeit benutzt und zu Beginn verteilen sich die „User" mit den relativen Anteilen $(A_0, B_0, C_0, D_0) = \left(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}\right)$ auf die Seiten *A, B, C, D*. Die User wechseln in gewissen Zeiteinheiten die Internetseite (indem sie einem Link folgen oder auch nicht).
   Nach *n* Zeitschritten ergibt sich eine Verteilung von $(A_n, B_n, C_n, D_n)$.
   Stellen Sie ein entsprechendes Gleichungssystem mit Hilfe von Matrizen und Vektoren auf, mit dem Sie die Verteilung nach 1, 2, 20, ..., *n*-Schritten iterativ berechnen können.
3. Stellen wir uns vor, sehr viele User nutzen dieses Netz: Welcher Anteil davon wird sich – langfristig – im Zuge der Recherchen bei *A, B, C, D* aufhalten?

### Lösung

**1.** Der Übergangsgraph:




Die zugehörigen Rekursionen lauten:

$$\begin{aligned} & & C_n & & = A_{n+1} \\ 0{,}5A_n & & & + 0{,}5D_n & = B_{n+1} \\ 0{,}5A_n & + 0{,}5B_n & & + 0{,}5D_n & = C_{n+1} \\ & 0{,}5B_n & & & = D_{n+1} \end{aligned}$$

**2.** Beschreibung des Linearen Gleichungssystems durch Matrizen und Vektoren:

$$U \cdot \vec{v} = \vec{v}_{n+1}$$

$$\begin{pmatrix} 0 & 0 & 1 & 0 \\ 0{,}5 & 0 & 0 & 0{,}5 \\ 0{,}5 & 0{,}5 & 0 & 0{,}5 \\ 0 & 0{,}5 & 0 & 0 \end{pmatrix} \cdot \begin{pmatrix} A_n \\ B_n \\ C_n \\ D_n \end{pmatrix} = \begin{pmatrix} A_{n+1} \\ B_{n+1} \\ C_{n+1} \\ D_{n+1} \end{pmatrix}.$$

Die Startverteilung lautet: $\vec{v}_0 = \begin{pmatrix} 0{,}25 \\ 0{,}25 \\ 0{,}25 \\ 0{,}25 \end{pmatrix}$

In den nächsten Schritten beträgt die Verteilung:

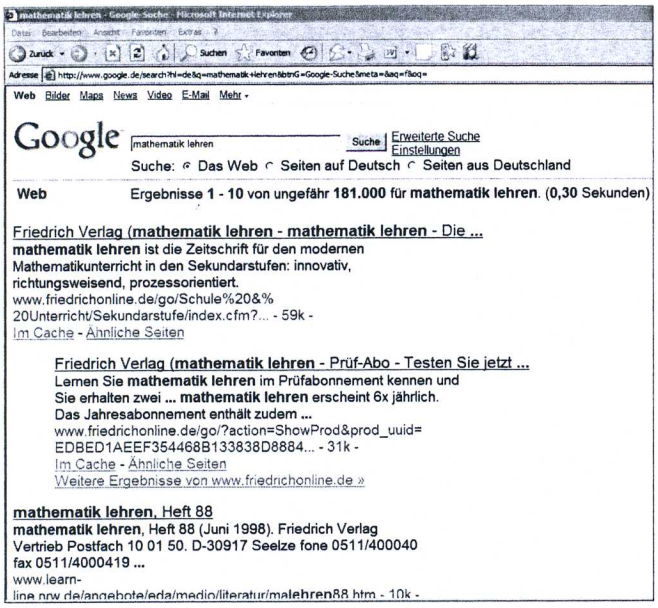
$$\vec{v}_1 = U \cdot \vec{v}_0 = \begin{pmatrix} 0{,}25 \\ 0{,}25 \\ 0{,}375 \\ 0{,}125 \end{pmatrix} \quad \text{und}$$

$$\vec{v}_2 = U \cdot \vec{v}_1 = U^2 \cdot \vec{v}_0 = \begin{pmatrix} 0{,}375 \\ 0{,}1875 \\ 0{,}3125 \\ 0{,}125 \end{pmatrix}.$$

**3.** Die weiteren Verteilungen $\vec{v}_n = U^n \cdot \vec{v}_0$; streben zu einer „Grenzverteilung"

$$\vec{v}_n \to \vec{v}, \text{ mit } \vec{v} = \begin{pmatrix} \frac{3}{9} \\ \frac{2}{9} \\ \frac{3}{9} \\ \frac{1}{9} \end{pmatrix}.$$

Langfristig werden sich also auf den Seiten A und C jeweils $\frac{3}{9}$ der User befinden, bei B nur $\frac{2}{9}$ und bei D nur $\frac{1}{9}$. In diesem Sinne wären A und C gleichberechtigt die „wichtigsten Seiten".




**Mit dem PageRank-Algorithmus kommt die wichtigste Seite bei Google an erster Stelle**

halte und der Vernetzungsstruktur (welche Seite ist mit welcher verlinkt?) des WWW in Bezug auf den Suchbegriff zu erhalten.

Jede Anfrage löst eine Suche nach Textstellen in den erfassten Internetdokumenten aus und wird in Form einer Liste von Treffern beantwortet. Um für wichtige Informationen nicht seitenweise blättern zu müssen, sollen zum jeweiligen Thema wichtige Seiten auch zuerst aufgelistet werden. Es stellt sich also bei jeder Suchmaschine die Frage: In welcher Reihenfolge werden die Treffer dem Anfragesteller präsentiert? Wie kann bewerkstelligt werden, dass die Liste der Treffer mit den „relevanten" und „wichtigen" Webseiten beginnt?

Bei Wikipedia heißt es dazu: „Der *PageRank*-Algorithmus ist ein Verfahren, eine Menge verlinkter Dokumente, wie beispielsweise das WWW, anhand ihrer Struktur zu bewerten bzw. zu gewichten. Dabei wird jedem Element ein Gewicht, der PageRank, aufgrund seiner Verlinkungsstruktur zugeordnet. Der Algorithmus wurde von Larry Page (daher stammt der Name PageRank) und Sergej Brin an der Stanford University entwickelt und von dieser zum Patent angemeldet. Er diente dem von Brin und Page gegründeten Unternehmen Google als Grundlage für die Bewertung von Seiten."

## Das WWW als gerichteter Graph

Wenn jemand einen Suchbegriff in Google eingegeben hat, werden *spider* ausgeschickt, die die in Frage kommenden Dokumente samt ihrer Vernetzung (hier: Verlinkung) als gerichteten Graphen darstellen. Wir wählen zunächst ein ganz einfaches Beispiel. A, B, C, D seien vier verschiedene Internetseiten, die auf eine einfache Weise miteinander verlinkt sind, so gibt es z. B. von Seite A einen Link zu den Seiten B und C, von der Seite B gibt es Links zu den Seiten C und D usw. (vgl. **Kasten 2**, S. 59).

### 1. Modellannahme

Alle Links auf einer Seite mit jeweils derselben relativen Häufigkeit bzw. Wahrscheinlichkeit benutzt werden.

Damit würde im Übergangsgraphen im Falle zweier ausgehender Pfeile bei beiden $\frac{1}{2}$ stehen, bei drei ausgehenden Pfeilen jeweils $\frac{1}{3}$ etc. Wegen dieser bewusst vereinfachenden Annahme sind die Wahrscheinlichkeiten bei den einzelnen Pfeilen gar nicht dazugeschrieben. (Natürlich entspricht diese Vereinfachung nicht ganz der Realität; ein Link am Ende einer langen html-Seite wird wohl nicht mit derselben Wahrscheinlichkeit benutzt wie einer, der prominent ganz oben in der Seite „thront". Aber dieses Vereinfachen und Idealisieren, so dass der „Sache" mit mathematischen Mitteln beizukommen ist, ist ein typischer Schritt bei so gut wie allen Modellierungen.)

Nun kann man sich wie bei den Telefongesellschaften vorstellen, dass eine Menge „User" sich auf die Seiten *A, B, C, D* verteilen, zu Beginn mit den relativen Anteilen $(A_0, B_0, C_0, D_0)$. Der Wechsel der Telefongesellschaft entspricht hier einem Wechsel der Internetseite. Wir stellen uns dieses System wieder in diskreten Schritten vor: Die User wechseln in gewissen Zeiteinheiten die Internetseite (indem sie einem Link folgen oder auch nicht), so dass sich nach $n$ Zeitschritten eine Verteilung von $(A_n, B_n, C_n, D_n)$ ergibt (*relative Anteile*). Wir können die zugehörigen Rekursionen wieder leicht aus dem Übergangsgraphen ablesen und aufschreiben.

Um zu überprüfen, bei welcher Verteilung $(\overline{A}, \overline{B}, \overline{C}, \overline{D})$ sich die User dieser vier Seiten langfristig einpendeln werden, könnten wir wieder mit EXCEL arbeiten. Aber man kann lineare Gleichungssysteme auch gut mit Matrizen und Vektoren beschreiben.

Alle Übergänge $U \cdot \vec{v}_0 = \vec{v}_1$, werden durch dieselbe Matrix $U$ („Übergangsmatrix") vermittelt. Dabei stehen in der Spalte $i$ die Wahrscheinlichkeiten, dass ein User, der sich gerade auf Seite $i$ befindet, sich im nächsten Schritt auf Seite $j$ befinden wird d. h. einen Link zur Seite $j$ benutzt (durch den Eintrag in Zeile $j$ gegeben; $i, j = 1, 2, 3, 4$). So muss z. B. jemand, der sich auf Seite $C$ befindet, im nächsten Schritt zwangsweise (Wahrscheinlichkeit 1) zu Seite $A$ kommen, was im Übergangsgraph und in der Übergangsmatrix anhand der 1 in Spalte 3 und Zeile 1 abzulesen ist.

Für die Übergänge erhält man der Reihe nach:

$$U \cdot \vec{v}_0 = \vec{v}_1,$$

$$\underbrace{U \cdot (\overbrace{U \cdot \vec{v}_0}^{\vec{v}_2})}_{U^2 \cdot \vec{v}_0} = \vec{v}_2,$$

$$\underbrace{U \cdot \left[ \overbrace{U \cdot (\underbrace{U \cdot \vec{v}_0}_{\vec{v}_1})}^{\vec{v}_2} \right]}_{U^3 \cdot \vec{v}_0} = \vec{v}_3, \ldots,$$

$$U^n \cdot \vec{v}_0 = \vec{v}_n.$$

Wir haben mit Hilfe von Matrizen somit eine Möglichkeit, eine *geschlossene* Formel für die Verteilung $\vec{v}_n$ zu erhalten (*nicht* nur *iterativ* wie mit EXCEL). Vektoren, die relative Anteile oder Wahrscheinlichkeiten enthalten, heißen in der Mathematik auch „stochastische Vektoren". (Ein Vektor $\vec{v}$ heißt stochastisch, wenn seine Komponenten Zahlen aus dem Intervall [0; 1] sind mit Summe 1.) Eine quadratische Matrix heißt „stochastisch", wenn ihre Spaltenvektoren stochastisch sind. In manchen Büchern werden bei einer stochastischen Matrix stochastische *Zeilen*vektoren statt *Spalten*vektoren gefordert. Dann muss die zugehörige Multiplikation umgekehrt erfolgen: (Zeilen-)Vektor mal Matrix statt Matrix mal (Spalten-)Vektor. Die zweite Form der Multiplikation ist in der Schule aber üblicher. Beide Begriffe (*stochastischer Vektor* bzw. *stochastische Matrix*) sind zwar für einen möglichen Unterricht in der Schule nicht unbedingt nötig, bieten sich aber hier an.

Übergangsmatrizen sind natürlich stochastische Matrizen, sie sind quadratisch, und in der ersten Spalte z. B. stehen die relativen Häufigkeiten, mit der User von $A$ im nächsten Schritt zu den Seiten *A, B, C, D* wechseln. Diese Zahlen sind naturgemäß aus dem Intervall [0; 1] und ergeben in Summe 1 (analog bei den anderen Spalten).

An dieser Stelle hat man auch die Möglichkeit, leichte allgemeine Begründungen zu thema-

tisieren, zum Beispiel für die Aussagen:

(1) das Produkt einer stochastischen Matrix mit einem stochastischen Vektor ist wieder ein stochastischer Vektor.

(2) das Produkt zweier stochastischer Matrizen ist wieder eine stochastische Matrix.

Eine wichtige Anwendung erfahren hier auch das Potenzieren von Matrizen und das Assoziativgesetz der Matrizenmultiplikation, wie bei

$$\vec{v}_3 = U\cdot\left(U\cdot(U\cdot\vec{v}_0)\right) = (U\cdot U\cdot U)\cdot\vec{v}_0 = U^3\cdot\vec{v}_0.$$

## Wie misst man die Wichtigkeit einer Seite?

Natürlich ist eine Seite umso wichtiger, je mehr Seiten auf diese Seite verweisen, insbesondere dann, wenn es sich bei den verweisenden Seiten selbst um „wichtige" handelt; denn dann nimmt man ja berechtigt an, dass auf dieser Seite gewisse tragende „Standards" bezüglich des Suchbegriffes gesetzt werden und dort also viel Wissenswertes zu finden ist. Welche ist nun die wichtigste Seite im obigen Graphen? Welche die zweitwichtigste, usw.? Wie soll man allgemein die Wichtigkeit einer Seite in einem gerichteten Graphen feststellen?

Stellen wir uns vor, sehr viele User nutzen dieses Netz (gerichteter Graph): Welcher Anteil davon wird sich – langfristig– im Zuge der Recherchen bei *A, B, C, D* aufhalten? Wenn sich herausstellen sollte, dass eine bestimmte Seite 90 % der Suchenden auf sich zieht, so ist wohl klar, dass diese Seite am wichtigsten ist und in der Liste zuerst genannte werden sollte.

Diese „langfristigen relativen Anteile" sind also eine Möglichkeit, die Wichtigkeit einer Seite zu beschreiben, und dafür brauchen wir „Grenzverteilungen".

Angenommen die User beginnen zufällig auf den vier Seiten zu surfen und die „Startverteilung" der User beträgt für alle vier Seiten jeweils $\frac{1}{4}$. Wenn diese User dann zufällig weitersurfen, beträgt die Verteilung im nächsten Schritt (0,25; 0,25; 0,375; 0,125) und im darauf folgenden (0,375; 0,1875; 0,3125; 0,125), wie in Kasten 2 berechnet wird. Die Seiten *A* und *C* scheinen hier also den Löwenanteil abzubekommen. Dies ist auch plausibel: Alle Seiten haben einen Link zur Seite *C* und von dort aus muss man zur Seite *A* . . . . Indem man immer wieder von links mit der Übergangsmatrix *U* multipliziert, erhält man die weiteren Verteilungen $\vec{v}_n = U^n\cdot\vec{v}_0$; diese streben zu einer „Grenzverteilung" $\vec{v}_n \to \vec{v}$ (siehe Kasten 2). Demnach müssten die Seiten *A* und *C* mit einem Anteil von jeweils $\frac{3}{9}$ gleichberechtigt auf Platz 1 gereiht werden, vor Seite $B\left(\frac{2}{9}\right)$ und $D\left(\frac{1}{9}\right)$.

Solche Grenzverteilungen kann man im Unterricht auf mehrere Arten bestimmen:

1. Mit EXCEL die Iteration so lange durchführen, bis sich die Werte nicht mehr ändern.
2. Mit einem ComputerAlgebraSystem (CAS) eine hohe Matrixpotenz von *U* bestimmen, um mit $\vec{v}_n = U^n\cdot\vec{v}_0$ für große *n* wohl nahe der „Grenzverteilung" zu sein.
3. Gesucht ist dabei ein Vektor $\vec{v}$ mit Komponentensumme 1, der sich bei Multiplikation mit *U* nicht mehr ändert: $U\cdot\vec{v} = \vec{v}$. Man muss also ein lineares Gleichungssystem lösen, natürlich auch mit CAS.

### Probleme bei der Berechnung der Grenzverteilung

Kann es mehrere solche Vektoren $\vec{v}$ (Grenzverteilungen, Lösungen) geben? Wenn es mehrere solche Vektoren gibt, streben dann die Verteilungsvektoren $\vec{v}_i$ manchmal (je nach Startverteilung $\vec{v}_0$) gegen den einen und manchmal gegen den anderen? Dies alles wäre natürlich sehr unangenehm, denn immerhin soll diese Grenzverteilung die Basis und Argumentationsgrundlage für das Ranking nach der Wichtigkeit sein.

Eine nicht eindeutige und von der Startverteilung abhängige Grundlage wäre allerdings sehr zweifelhaft. Am besten wäre es, wenn dieser Vektor der Grenzverteilung eindeutig wäre und auch unabhängig von der Startverteilung $\vec{v}_0$.

Alle drei oben angegebenen Möglichkeiten zur Berechnung der Grenzverteilung $\vec{v}$ funktionieren in einfacher Weise natürlich nur für relativ kleine Dimensionen, wie oben bei einer $4\times4$-Matrix, evtl. auch noch bei einer $20\times20$-Matrix, aber klarer Weise nur mehr schwerlich bei einer $100000\times100000$-Matrix. Bei größeren Matrizen werden hier andere *iterative* Algorithmen zur *näherungsweisen* Lösung verwendet. Bei Google-Anwendungen können dies einige hunderttausend oder gar Millionen Seiten sein; außerdem kommen auf Google pro Sekunde sehr viele Anfragen zu, die alle prompt erledigt werden sollten.

## Der Satz von Markoff

Unabhängig davon, ob „Markoff-Ketten" als Begriff thematisiert werden, besonders wichtig in diesem Zusammenhang ist der *Satz von Markoff*, der eine einfache hinreichende Bedingung an die Übergangsmatrix *U* angibt, die garantiert, dass die Grenzverteilung $\vec{v}$ existiert, eindeutig und unabhängig von der Startverteilung $\vec{v}_0$ ist (ohne Beweis).

### Satz von Markoff

Wenn *U* stochastisch ist und $U^n$ für irgendein $n \geq 1$ (mindestens) eine Zeile mit nur positiven Elementen hat, dann streben die Matrixpotenzen $U^n$ für $n \to \infty$ zu einer stochastischen Grenzmatrix *G* mit identischen Spalten (d. h. die Zeileneinträge sind in jeder Zeile konstant).

Es ist klar, dass die Spalten der Grenzmatrix *G* dann den eindeutigen und vom Startvektor $\vec{v}_0$ unabhängigen Grenzvektor $\vec{v}$ angeben. Denn wegen $A_0 + B_0 + C_0 + D_0 = 1$ erhält man z. B. im Fall einer $4\times4$-Matrix für den Grenzvektor $\vec{v}$ mit dieser Grenzmatrix *G* (unabhängig von den konkreten Werten von $A_0, B_0, C_0, D_0$)

$$\vec{v} = G\cdot\vec{v}_0 = \begin{bmatrix} u_1 & u_1 & u_1 & u_1 \\ u_2 & u_2 & u_2 & u_2 \\ u_3 & u_3 & u_3 & u_3 \\ u_4 & u_4 & u_4 & u_4 \end{bmatrix}\cdot\begin{bmatrix} A_0 \\ B_0 \\ C_0 \\ D_0 \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ u_3 \\ u_4 \end{bmatrix}.$$

Da *G* stochastisch ist, gilt $\Sigma u_i = 1$.

Bei unserem Beispiel hat zwar nicht *U* selbst, aber schon $U^2$ eine solche Zeile mit nur positiven Elementen, so dass die Konvergenz und die Unabhängigkeit von der Startverteilung nach dem Satz von Markoff garantiert ist. Wir berechnen in unserem Fall z. B. $U^{20}$ und erhalten mit CAS (vier Nachkommastellen):

$$U^{20} = \begin{bmatrix} 0{,}3333 & 0{,}3333 & 0{,}3333 & 0{,}3333 \\ 0{,}2222 & 0{,}2222 & 0{,}2222 & 0{,}2222 \\ 0{,}3333 & 0{,}3333 & 0{,}3333 & 0{,}3333 \\ 0{,}1111 & 0{,}1111 & 0{,}1111 & 0{,}1111 \end{bmatrix}$$

Die Grenzverteilung

$$\vec{v} = \begin{bmatrix} 3/9 \\ 2/9 \\ 3/9 \\ 1/9 \end{bmatrix}$$ ist hier als Spalte gut abzulesen.

Der Satz von Markoff braucht im Unterricht nicht unbedingt bewiesen zu werden, man kann ihn einfach benutzen, um den PageRank-Algorithmus in seinen Grundzügen nachvollziehen zu können. (Bei Bedarf findet sich ein elementarer Beweis für den Spezialfall von $2\times2$-Matrizen z. B. in Humenberger 2002a.) Auch sonst braucht die zugehörige Theorie nicht breitgetreten zu werden.

## Komplexeres Beispiel: Suche mit Sackgasse

Ein immer noch sehr kleines Netzwerk aus sechs Internetseiten findet sich in **Kasten 3**. Hier gibt es eine Situation, die oben noch nicht gegeben war: Von der Seite 2 gibt es offenbar keine weiterführenden Links, im Surfvorgang könnte man so eine Seite als „Senke" bzw. „*Sackgasse*" bezeichnen. Dies spiegelt sich in der zweiten Spalte der Matrix *U* wider, die nur Nullen enthält. Dies ist natürlich schlecht für unsere Zwecke (stochastische Matrix, deren Spaltensumme sollte 1 sein).

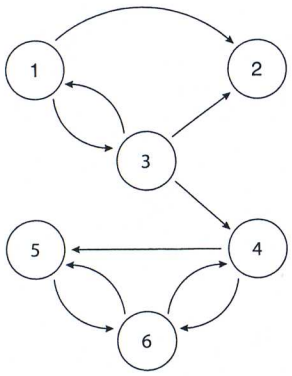
Was wird man in so einer Situation beim Recherchieren mit Google praktisch machen? Es gibt mehrere Möglichkeiten:

1. Den Suchvorgang beenden und bei der Seite ② bleiben; dies würde in der Matrix bedeuten die zweite Null in der zweiten Spalte durch eine 1 zu ersetzen, im gerichteten Gra-

❸ BEISPIEL

## Komplexeres Netzwerk aus sechs Internetseiten

Die Übergangsmatrix lautet:




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

### Sackgasse

Die Seite 2 ist eine „Sackgasse", kein Link führt von ihr weg.

**Annahme**

Trifft man auf die Seite 2, so geht man zurück zur Liste aller Seiten und klickt irgendeine (mit der Wahrscheinlichkeit $\frac{1}{6}$) an. In der Übergangsmatrix fügt man also statt einer reinen „Nullenspalte" den 6-dimensionalen Spaltenvektor (alle Einträge = $\frac{1}{6}$) ein und erhält:

$$U_1 = \begin{pmatrix} 0 & \frac{1}{6} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{6} & \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{2} & \frac{1}{6} & 0 & 0 & 0 & 0 \\ 0 & \frac{1}{6} & \frac{1}{3} & 0 & 0 & \frac{1}{2} \\ 0 & \frac{1}{6} & 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ 0 & \frac{1}{6} & 0 & \frac{1}{2} & 1 & 0 \end{pmatrix}$$

### Neueinstiege

Wenn stets über die Liste gewechselt würde (Neueinsteig), würde in der ganzen Übergangsmatrix die Wahrscheinlichkeit $\frac{1}{6}$ stehen.

Allgemein: Beim Neueinsteigen gilt für die Übergangsmatrix

$$U_2 = \begin{pmatrix} \frac{1}{m} & \cdots & \frac{1}{m} \\ & \vdots & \\ \frac{1}{m} & \cdots & \frac{1}{m} \end{pmatrix}, \text{ denn: } \begin{pmatrix} \frac{1}{m} & \cdots & \frac{1}{m} \\ & \vdots & \\ \frac{1}{m} & \cdots & \frac{1}{m} \end{pmatrix} \cdot \underbrace{\begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix}}_{\sum v_i = 1} = \begin{pmatrix} \frac{1}{m} \\ \vdots \\ \frac{1}{m} \end{pmatrix}$$

**Annahme**

Mit der Wahrscheinlichkeit $\alpha$ folgen die User den Links, mit der Wahrscheinlichkeit $1 - \alpha$ steigen sie über die Liste neu ein. Dann ergibt sich die Übergangsmatrix

$T = \alpha \cdot U_1 + (1 - \alpha) \cdot U_2$

Für unser lineares Gleichungssystem $T \cdot \vec{v} = \vec{v}$ ergibt sich mit

$$\alpha = 0{,}85;\ \vec{v} = \begin{pmatrix} v_1 \\ \vdots \\ v_6 \end{pmatrix};\ \underbrace{v_1 + \ldots + v_6 = 1}_{v_i \geq 0}$$

bei Berechnungen mit einem CAS; 4 Nachkommastellen:

$$\begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{pmatrix} = \begin{pmatrix} 0{,}0517 \\ 0{,}0737 \\ 0{,}0574 \\ 0{,}1999 \\ 0{,}2686 \\ 0{,}3487 \end{pmatrix}.$$

phen käme dann bei ② ein Pfeil zu sich selbst dazu. Diese Möglichkeit wollen wir nicht wählen.

2. Man könnte mit dem Browser eine Seite zurück gehen und von dort andere Links benützen als den zu ② (die dann hoffentlich keine Sackgassen sind). Hier müsste man unterscheiden, von welcher Seite aus man zu ② gekommen ist, was die Sache relativ kompliziert machte.
3. Man verlässt diese Seite, kehrt zur Liste zurück (gleichgültig, ob diese Liste von Google schon nach der Wichtigkeit gereiht wurde oder nicht) und klickt zufällig eine der vielen (anderen) Seiten an.

Wir entscheiden uns für die dritte Variante und formulieren dies noch mal explizit:

### 2. Modellannahme

Wenn man beim Surfvorgang in einer Sackgasse (ohne weiterführenden Link) landet, so kehrt man zur Liste zurück und klickt nun eine der möglichen $m$ Internetseiten an, und zwar alle mit derselben Wahrscheinlichkeit $\frac{1}{m}$.

Man sieht dabei auch davon ab, dass die in Rede stehende Seite selbst normalerweise nun wohl nicht mehr angeklickt wird; aber wenn dies sehr viele Seiten sind, ergibt sich dadurch kein großer Unterschied: Man ersetzt die Einträge der zweiten Spalte (Nullen) nun jeweils durch 1/6 (allgemein durch $\frac{1}{m}$, wenn es sich um $m$ Webseiten, also um einen gerichteten Graphen mit $m$ Knoten, d. h. um eine $m \times m$-Matrix handelt). So kann man also auch dann eine stochastische Übergangsmatrix $U_1$ erreichen, wenn es in der Netzstruktur Sackgassen gibt.

Was wäre, wenn es eine Seite gäbe, auf die kein Link verweist (nur Links von ihr weg)? Wäre dies ähnlich schlimm wie eine Sackgasse? Und auch ohne eine „Sackgasse", kommt es vor, dass man nicht den Links auf dieser Seite folgt, sondern eben zur Liste zurückkehrt und eine andere Seite einfach anklickt.

### 3. Modellannahme

Mit Wahrscheinlichkeit $\alpha$ mögen die User irgendwelchen Links auf der jeweiligen Seite folgen, mit Wahrscheinlichkeit $1 - \alpha$ zur Liste zurückkehren und neu einsteigen, d. h. eine beliebige Seite (mit Wahrscheinlichkeit $\frac{1}{m}$) anklicken.

Wie kann man nun dieses Szenario mathematisch beschreiben? Wie sieht die dann zugehörige, neue Übergangsmatrix $T$ aus? Wenn man den Links auf der Seite folgt, ist die Übergangsmatrix durch $U_1$ gegeben. Wie muss die Übergangsmatrix $U_2$ im Falle des Neueinstiegs lauten (Schüleraufgabe)? Herauskommen muss ein Spaltenvektor mit Einträgen $\left(\frac{1}{m}\right)$.

Insgesamt ergibt sich also für die neue Übergangsmatrix $T$ durch Gewichten der beiden Fäl-

le bzw. Übergangsmatrizen mit den Faktoren $\alpha$ bzw. $1-\alpha$:

$T = \alpha \cdot U_1 + (1-\alpha) \cdot U_2$

$\alpha \cdot U_1 \triangleq$ mit Wahrscheinlichkeit $\alpha$ den Links folgen

$(1-\alpha) \cdot U_2 \triangleq$ mit Wahrscheinlichkeit $(1-\alpha)$ neu einsteigen

Es ist sehr leicht einzusehen (Schüleraufgabe): Weil $U_1$ und $U_2$ stochastische Matrizen sind, ist auch $T$ stochastisch.

Die Matrix $T$ hat *nur positive Einträge*, keine Nullen mehr. Nach dem Satz von Markoff liegt mit der Übergangsmatrix $T$ also sicher jene gewünschte und besonders einfache Situation vor, in der es eine eindeutige und von der Startverteilung unabhängige Grenzverteilung gibt. Und diese Grenzverteilung kann dann die gewünschte Reihung der Seiten angeben, ihre Wichtigkeit messen („PageRank").

Wie groß soll der Wert von $\alpha$ gewählt werden? Es ist bekannt, dass Google lange Zeit $\alpha = 0{,}85$ gewählt hat. Möglicherweise ist Google aber in der Zwischenzeit von diesem Wert abgewichen.

Für das Beispiel in Kasten 3 ergibt sich für die Grenzverteilung das lineare Gleichungssystem $T \cdot \vec{v} = \vec{v}$ und mit $\alpha = 0{,}85$ berechnet das CAS auf 4 Nachkommastellen:

$$\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \\ v_5 \\ v_6 \end{bmatrix} = \begin{bmatrix} 0{,}0517 \\ 0{,}0737 \\ 0{,}0574 \\ 0{,}1999 \\ 0{,}2686 \\ 0{,}3487 \end{bmatrix}$$

Dieses ergibt sich auch, wenn man eine hohe Matrixpotenz von $T$ berechnet und eine Spalte davon nimmt. Auch mit EXCEL könnte man zu diesem Ergebnis kommen. Nach diesem Ergebnis wären also die Seiten in absteigender Reihenfolge ihrer Wichtigkeit:
Seite 6 → Seite 5 → Seite 4 → Seite 2 → Seite 3 → Seite 1.

### Explizite Lösung (Formel)

Durch Einsetzen von $T = \alpha U_1 + (1-\alpha) U_2$ in $T \cdot \vec{v} = \vec{v}$ und Verwenden der Matrixschreibweise kann diese Gleichung noch ein wenig umgeschrieben werden, so dass sich sogar eine explizite Formel für $\vec{v}$ ergibt ($I$ bezeichnet dabei die $m$-dimensionale Einheitsmatrix):

$$\alpha \cdot U_1 \cdot \vec{v} + (1-\alpha) \cdot U_2 \cdot \vec{v} = I \cdot \vec{v}$$

$$\text{und mit } U_2 \cdot \vec{v} = \begin{bmatrix} 1/m \\ \vdots \\ 1/m \end{bmatrix} \Rightarrow$$

$$(\alpha \cdot U_1 - I) \cdot \vec{v} = (\alpha - 1) \cdot \begin{bmatrix} 1/m \\ \vdots \\ 1/m \end{bmatrix}$$

$$\Rightarrow \vec{v}$$

$$= (\alpha - 1) \cdot (\alpha U_1 - I)^{-1} \cdot \begin{bmatrix} 1/m \\ \vdots \\ 1/m \end{bmatrix}$$

Man kann zeigen, dass die Matrix $\alpha U_1 - I$ „nichtsingulär" ist, so dass sich für $\vec{v}$ dabei immer eine eindeutige Lösung ergibt. Dies folgt auch aus dem Satz von Markoff.

Die explizite Lösung ist aber für die Google-Praxis ungeeignet. Dort sind es ja sehr große lineare Gleichungssysteme (z. B. $m = 100\,000$ oder mehr) und dabei kann leider nicht mit der expliziten Lösung („Formel") gerechnet werden, denn das Invertieren einer Matrix kann man nur bei relativ kleinem $m$ in vernünftiger Zeit bewerkstelligen. (Noch dazu gibt es ja sehr viele Anfragen an Google pro Sekunde). Man kommt in der Praxis vielmehr iterativ zu einer (*Näherungs-)Lösung*, wobei die gesuchten Werte $\vec{v}_i$ in diesem Zusammenhang die Parameter für die Wichtigkeit einer Seite, die PageRank-Werte sind.

Diese Probleme sollen hier aber gar nicht im Vordergrund stehen. Es geht hier primär um die elementaren Überlegungen, die zur (stochastischen) Übergangsmatrix $T$ mit nur positiven Einträgen führen, so dass es nach dem Satz von Markoff sicher eine von der Startverteilung unabhängige Grenzverteilung gibt. Diese Ideen sind gleichermaßen einfach und genial! Sie garantieren, dass das Verfahren „immer funktioniert".

Die drei einfache Modellannahmen (siehe oben) haben hier eine enorme Wirkung. Natürlich kann dieses Modellieren kein selbständiges Modellieren durch die Schülerinnen und Schüler sein, sie lernen dadurch aber ein ganz aktuelles Stück „angewandter Mathematik" kennen.

## Der Algorithmus in der Praxis

Insgesamt muss man sagen, dass der Algorithmus von Google in der Praxis natürlich komplizierter abläuft als hier dargestellt (es fließen da noch viele andere Nebenbedingungen ein), aber die prinzipielle hinter dem Suchalgorithmus von Google steckende mathematische Idee ist eigentlich eine sehr elementare.

Es ist einerseits verblüffend, mit welch einfacher zugrunde liegender Idee so viel Geld zu machen ist, wie es den Gründern von Google gelang. Andererseits liegt darin wieder einmal mehr eine wohltuende Bestätigung, dass grundlegende mathematische Ideen sehr wichtig werden – hier für die Millionen User und wirtschaftlich gesehen für die Begründer dieser Firma und für die zugehörigen Aktionäre.

Dies soll die Leistung der beiden Google-Gründer in keiner Weise schmälern. Die konkrete Umsetzung dieser Idee in ein Programm, das diese Internetseiten-Reihung auch bei 100 000 oder noch mehr Seiten in akzeptabel kurzer Zeit erledigt, ist eine höchst schwierige Aufgabe und eine tolle Leistung!

Bei Google wurden schon 2005 etwas über 8 Milliarden URLs durchsucht und etwas über 1 Milliarde Bilder, heute mit Sicherheit schon viel mehr. Google bedient in jeder Sekunde sehr viele Anfragen in über 100 „Domains" und Sprachen, und alle wollen ihr Ergebnis sofort, d. h. auf Knopfdruck ohne zu warten.

Es wurde und wird eine Antwortzeit von höchstens einer halben Sekunde als Richtwert angestrebt. Diese schnelle Lieferung von Ergebnissen hat auch frühzeitig zur Popularität von Google beigetragen, die Konkurrenz hat sich mit der Anfragebeantwortung oft mehr Zeit gelassen und ist dadurch eindeutig ins Hintertreffen geraten. Mittlerweile beschäftigt Google eine ganze Schar hervorragender Programmierer, aber die ersten Schritte dürften die beiden Gründungsherren schon weitgehend alleine bewerkstelligt haben – Hut ab!

Es hat sich bei Internetrecherchen sogar schon das zugehörige Verbum „googeln" eingebürgert, auch im Englischen spricht man von „to google". Wenn jemand zu einem bestimmten Begriff jemand anderen fragt, so hört man oft: „Hast du diesen Begriff schon *gegoogelt*, um mehr darüber zu erfahren?"

### Literatur


Büchter, A./Henn H.-W.(²2007): Elementare Stochastik. – Springer, Berlin-Heidelberg.

Chartier, T. P.( 2006): Googling Markov. – In: The UMAP Journal, Heft 27, 1, S. 17–30.

Humenberger, H. (2002a): Der PALIO – das Pferderennen von Siena als Ausgangspunkt für Modelle von Auswahlprozessen und als Einstieg zum Thema Markoff-Ketten. – In: Stochastik in der Schule 22, 2, S. 2–13.

Humenberger, H. (2002b): Der PALIO – das Pferderennen von Siena. – In: *mathematik lehren* Heft 113 (August 2002), S. 58–62.

Lambacher-Schweizer (2001): Lineare Algebra mit analytischer Geometrie, Leistungskurs. – Klett, Stuttgart.

Langville, A. N. (2006): Google's PageRank and Beyond: The Science of Search Engine Rankings. – Princeton University Press, Princeton.

Meyer, D. (1998): Markoff-Ketten. – In: Mathematik in der Schule 36, 12, S. 661–670 und S. 675–680.

Stern: http://www.stern.de/wirtschaft/unternehmen/:Das-Google-Duo-Die-Internetstars-Page-Brin-/523515.html (Stand 24. 03. 2009).

Wills, R. S. (2006): Google's PageRank: The Math Behind the Search Engine. – In: The Mathematical Intelligencer 28, 4, S. 6–11.

Wirths, H. (1997): Markoff-Ketten – Brücke zwischen Analysis, linearer Algebra und Stochastik. – In: Mathematik in der Schule 35, 11, S. 601–606 und S. 611 – 613.


*Hans Humenberger, Wien*