Hans HUMENBERGER, Wien

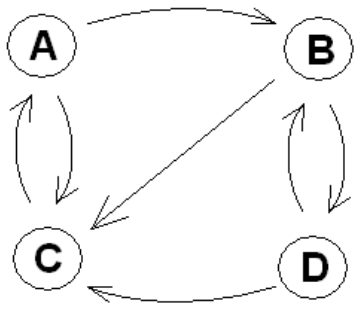
Das PageRank-System von Google – eine aktuelle Anwendung im Mathematikunterricht

„Mehrstufige Prozesse“ (in elementarer Form) gehören in manchen deutschen Bundesländern zum möglichen Lehrstoff in der Oberstufe, in Österreich leider nicht. In der Tat ist es ein Gebiet, in dem der *Vernetzungsgedanke* besonders gut verwirklicht werden kann: Lineare Algebra (Übergangsmatrizen), Stochastik (Wahrscheinlichkeiten), Analysis (Grenzwerte).

Dieser Beitrag soll eine Anregung für eine elementare und besonders aktuelle Anwendung zum Thema „Mehrstufige Prozesse“ sein. Bei Google erhebt sich die interessante Frage: Wie kommt Google eigentlich zu einer Reihung der „Liste“, wie schafft es Google, dass *wichtige* Seiten zum gesuchten Thema (bzw. Begriff) am *Anfang* der Liste stehen?

Durch einfache Modellannahmen (nicht als selbständige Modellierungsaufgabe für Schülerinnen und Schüler gedacht!) gelingt es, zu einem Resultat mit erstaunlicher Tragweite zu kommen. Natürlich ist der in Wirklichkeit bei Google verwendete Algorithmus komplizierter als hier dargestellt, eine wesentliche Kernidee ist aber ganz einfach zu beschreiben.

Die Suchmaschinen beginnen damit, mit einem so genannten *spider* oder *webcrawler* (spezielles Computerprogramm) das WWW zu „durchforsten“: Auf welchen Seiten ist zum gesuchten Begriff etwas zu erfahren, wo kommt er vor? Ziel dieser umfangreichen Suchtätigkeit ist es, eine möglichst gute „Momentaufnahme“ der Inhalte und der Vernetzungsstruktur (welche Seite ist mit welcher verlinkt?) des WWW in Bezug auf den Suchbegriff zu erhalten. Im Prinzip entsteht dadurch ein „gerichteter Graph“ der Art von nebenstehender Abbildung: Hier sind es der Einfachheit halber nur 4 Internetseiten (die auf die durch Pfeile angegebene Weise verlinkt sind), in Wirklichkeit sind dies oft mehrere 100.000 oder gar Millionen!



Modellannahme 1: Alle Links auf einer Seite werden mit jeweils derselben relativen Häufigkeit bzw. Wahrscheinlichkeit benutzt, so dass bei n ausgehenden Pfeilen bei allen das „Gewicht“ $1/n$ stehen soll (deswegen oben gar nicht extra dazu geschrieben).

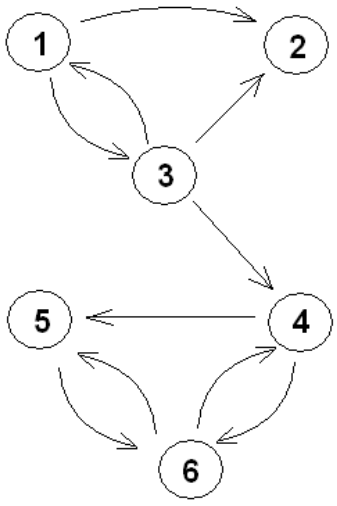
Wie soll nun die Wichtigkeit einer Seite gemessen werden?

Man kann sich dazu z. B. Folgendes vorstellen:
Sehr viele User nutzen dieses Netz (gerichteter Graph): Welcher Anteil davon wird sich – à la longue – 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 eine Möglichkeit, die Wichtigkeit einer Seite zu beschreiben. Wir müssen also die langfristige Verteilung („Grenzverteilung“) der User auf die Seiten A, B, C, D bestimmen.

Die relativen Anteile der Seiten in den **Grenzverteilungen** können als Maß für ihre jeweilige Wichtigkeit herangezogen werden, wobei die Bestimmung von Grenzverteilungen (Markoff-Ketten) auf mehrere Arten möglich ist (EXCEL, CAS; hier aus Platzgründen nicht näher ausgeführt[1]). Nach dem berühmten Satz von Markoff ist dies dann besonders einfach, wenn z. B. die zugehörige „Übergangsmatrix“ keine Nullen, sondern nur *positive* Elemente enthält[2]. Insofern liegt es auch nahe nach Modellierungen zu suchen, so dass die Übergangsmatrix eben nur positive Elemente enthält. Diese Modellannahmen brauchen aber nicht vom Himmel zu fallen, sondern können alle leicht inhaltlich nachvollzogen und interpretiert werden.

Ein etwas komplizierteres Beispiel

Ein kleines Netzwerk aus 6 Internetseiten habe nebenstehende Verlinkungsstruktur. Die Übergangsmatrix können wir aus dem Übergangsgraphen ablesen:



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

In Spalte *i* stehen dabei die Übergangswahrscheinlichkeiten $\boxed{i} \to \boxed{j}$.

[1] Für eine ausführlichere Version sei auf Humenberger 2009 verwiesen.
[2] Die Grenzmatrix *G* hat in diesem Fall eine besonders einfache Gestalt, sie besteht nämlich aus *identischen Spalten*. Die Grenzverteilung ist dann durch eine solche Spalte der Grenzmatrix gegeben und *eindeutig* (unabhängig von der Anfangsverteilung).

Hier gehen von ② keine Pfeile aus („Sackgasse“, 2. Spalte hat nur Nullen). So kann das natürlich nicht bleiben, es bieten sich mehrere Auswege an: Z. B. könnte man die zweite Null von oben durch eine 1 ersetzen (d. h. wenn man nach ② kommt, so bleibt man auch bei ②, „Ende“; dies würde bedeuten im Übergangsgraphen einen Pfeil von ② zu sich selbst zu ergänzen). Wir wählen aber eine andere Möglichkeit:

Modellannahme 2: Wenn man beim Surfen auf eine Seite ohne weiterführende Links kommt („Sackgasse“), so kehrt man zurück zur Liste und wählt zufällig eine der anderen Seiten auf der Liste: alle mit derselben Wahrscheinlichkeit $1/n$ (bei n möglichen Seiten).

Für die Übergangsmatrix U bedeutet dies im obigen Beispiel, dass alle Nullen in der zweiten Spalte durch 1/6 ersetzt werden → U^*.

Dadurch auf den Plan gerufen: Auch wenn die Seite keine Sackgasse ist, kann es doch vorkommen, dass jemand nicht den Links auf dieser Seite folgt, sondern eben zur Liste zurückkehrt und eine andere Seite einfach anklickt. Mit Wahrscheinlichkeit α mögen die User irgendwelchen Links auf der jeweiligen Seite folgen, mit Wahrscheinlichkeit $1-\alpha$ zur Liste zurückkehren und neu einsteigen, d. h. eine beliebige andere Seite (mit Wahrscheinlichkeit $1/n$) 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^* gegeben. Beim *Neueinsteigen* muss die nächste Verteilung durch den Vektor $(1/n,\ldots,1/n)^t$ gegeben sein, d. h. die Übergangsmatrix wird in diesem Fall $\begin{pmatrix} 1/n & \cdots & 1/n \\ \vdots & & \vdots \\ 1/n & \cdots & 1/n \end{pmatrix}$ lauten, denn: $\begin{pmatrix} 1/n & \cdots & 1/n \\ \vdots & & \vdots \\ 1/n & \cdots & 1/n \end{pmatrix} \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_n \end{pmatrix} = \begin{pmatrix} 1/n \\ \vdots \\ 1/n \end{pmatrix}$.

Dabei beschreibt $\vec{v}_n = (v_1,\ldots,v_n)^t$ eine beliebige Verteilung ($v_i \geq 0$, Summe = 1). Insgesamt ergibt sich für die neue Übergangsmatrix T durch Gewichten der beiden Fälle bzw. Übergangsmatrizen mit α bzw. $1-\alpha$:

$$T = \underbrace{\alpha \cdot U^*}_{\substack{\text{Mit W' } \alpha \text{ den} \\ \text{Links folgen}}} + \underbrace{(1-\alpha) \cdot \begin{pmatrix} 1/n & \cdots & 1/n \\ \vdots & & \vdots \\ 1/n & \cdots & 1/n \end{pmatrix}}_{\text{Mit W' } (1-\alpha) \text{ neu einsteigen}}$$

Die Übergangsmatrix 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.

Wie groß soll der Wert von α 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 obiges Beispiel ergibt sich mit $\alpha = 0{,}85$ für die Reihenfolge der Wichtigkeit der einzelnen Seiten $6 \rightarrow 5 \rightarrow 4 \rightarrow 2 \rightarrow 3 \rightarrow 1$, wie man aus der Grenzverteilung ablesen kann.

Fazit: Man kann auf elementare Art und Weise den Kern des PageRank-Algorithmus von Google, der es in den Neunzigerjahren überlegen machte und einen beträchtlichen Teil des Erfolges von Google ausmacht, beschreiben und analysieren. Die einfachen Modellannahmen sind dabei so gewählt, dass man (→ Satz von Markoff) zu einer eindeutigen Grenzverteilung kommen muss. Diese Grenzverteilung kann in der Praxis nicht in geschlossener Form berechnet werden, sondern nur *näherungsweise* mit *iterativen* Methoden, da es sich bei den Matrizen bzw. Vektoren meist um Dimensionen von mehreren Hunderttausend bis Millionen handelt.

Das Potential dieses Themas im Schulunterricht in Stichworten

- Ein spannendes und aktuelles Phänomen wird analysiert
- Realitätsbezug: jede/r verwendet Google
- Motivation, Verblüffung: Mit welch *elementaren* Ideen ist etwas „Weltbewegendes“ auf die Beine zu stellen und viel Geld zu verdienen. Bestätigung, dass grundlegende Ideen bedeutungsvoll sind.
- Wenige Voraussetzungen: Matrizen und Vektoren
  2-stufige Prozesse können zur *Einführung* der *Matrizenmultiplikation* genommen werden oder als *zusätzliche sinnvolle Anwendung*.
- Sinnvoller Computereinsatz: EXCEL, CAS
- Gute Vernetzungsmöglichkeit: Stochastik, Lineare Algebra, Analysis

Literatur

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

Humenberger, H. (2009). Das PageRank-System von Google – eine aktuelle Anwendung von Markoff-Ketten und großen linearen Gleichungssystemen. Erscheint in *mathe matiklehren*

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