Mit Mathematik zum Milliardär

Ehrhard Behrends, FU Berlin

Vor einiger Zeit ist GOOGLE an die Börse gegangen, die Gründer Sergej Brin und Lawrence Page gehören seit diesem Tag zu den reichsten Männern der Welt.

Wer es Ihnen nachmachen wollte, müsste sich natürlich zuerst einen riesigen Computer kaufen und dann einen Katalog aller Webseiten dieser Welt erstellen: Es sind so an die 20 Milliarden. Zu jeder Seite sollte es natürlich einen Index derjenigen Begriffe geben, die dort interessant sind. Das ist sicher recht zeitaufwändig, aber für ein Team von begabten Programmierern ist das keine unüberwindliche Herausforderung: Schließlich kann man das Suchen ja auch an Computer delegieren.

Mal angenommen, diese Aufgaben sind zur Zufriedenheit erledigt. Leider kann man damit noch keine attraktive Suchmaschine anbieten. Der Grund ist die gewaltige Größe des Internets. Denn wenn eine konkrete Anfrage kommt (suche nach allen Internetseiten, in denen man „USA“ und „Hurrikan“ findet), so ist es kein großes Problem, alle Seiten zusammenzustellen, in denen diese beiden Begriffe vorkommen. Die Frage ist jedoch, wie man sie denn präsentieren soll. In der Regel gibt es nämlich Hunderttausende bis mehrere Millionen von Treffern. Niemand hat die Geduld, die alle zu sichten, vielmehr möchte man die „wichtigen“ Seiten zum Thema als erstes angeboten bekommen. Wer hin und wieder „googelt“, weiß, dass GOOGLE das Problem bemerkenswert gut löst, denn unter den ersten wenigen Dutzend Angeboten findet man in der Regel das, was man sucht.

Das Geheimnis ist die richtige Definition von „wichtig“. Bei GOOGLE besteht die Grundidee darin, die Wichtigkeit einer Seite dadurch zu messen, dass viele wichtige Seiten darauf verweisen. Damit ist folgendes gemeint. Wenn wir die Internetseiten mit 1, 2, ... durchnummerieren und die Wichtigkeit der einzelnen Seiten mit W_1, W_2, ... bezeichnen, so soll zwischen diesen Zahlen eine Abhängigkeit bestehen. Wenn etwa die Seite 5 auf die Seite 2 verlinkt und Seite 5 insgesamt drei weiterführende Links hat, so „erbt“ Seite 2 den dritten Teil der Wichtigkeit von Seite 5. Vielleicht verweist auch Seite 7 auf Seite 2, und das liefert (wenn von Seite 7 zehn Links weiterführen) den Anteil „W_7 geteilt durch 10“. Mal angenommen, das wäre alles: Niemand anders verweist auf Seite 2. Dann führt das zur Gleichung

$$W_2 = W_5/3 + W_7/10.$$

Für die meisten Webseiten ergeben sich kompliziertere Bedingungen, insgesamt erhält man jedoch ein Gleichungssystem von 20 Milliarden Gleichungen für die 20 Milliarden Unbekannten W_1, W_2, ...

Schulmathematik hilft da leider nicht weiter, für viele sind schon 2 Gleichungen mit 2 Unbekannten das Schwierigste, was sie jemals kennen gelernt haben.
Doch auch für Profis ist das Problem eine Nummer zu groß, auch wenn – wie bei einigen Optimierungsproblemen - Systeme mit einigen Hunderttausend oder gar einigen Millionen Unbekannten vorkommen.

Ein anderer Weg führt aber zum Ziel, das passende Stichwort heißt „Zufallsspaziergang“. Man stelle sich einen Internetsüchtigen Surfer vor, der – zum Beispiel – auf der Homepage der WELT startet. Dort sucht er sich durch Zufall einen der verfügbaren Links aus und klickt sich auf die entsprechende Seite. Da angekommen, werden wieder die Links gesichtet, einer wird zufällig ausgewählt und – Klick! – geht die Reise weiter.
Auf diese Weise ergibt sich ein Streifzug durch die Welt des WWW, bei dem logischerweise die „wichtigen“ Seiten häufiger besucht werden als andere. Bemerkenswerter Weise ist sogar so, dass die relativen Häufigkeiten des Besuchs die weiter oben aufgestellten Gleichungen erfüllen. Kurz: Die Wichtigkeit einer Seite kann dadurch gemessen werden, dass man misst, in welchem Prozentsatz seiner Zeit unser Surfer dort anzutreffen ist.

Eine konkrete Berechnung scheint nun aber auch nicht viel leichter als das Ausgangsproblem zu sein. Wenn man es ganz genau nimmt, stimmt das auch, aber eine näherungsweise Lösung (bei der die Wichtigkeiten dann bis auf – z.B. – fünf gültige Dezimalen feststehen) ist in einigen Stunden Rechenzeit zu haben.

Damit ist die Suchmaschine voll funktionsfähig, denn wenn man die Wichtigkeiten hat, ist alles einfach: Suche alle Seiten, die „USA“ und „Hurrikan“ enthalten und gib sie in der Reihenfolge der Wichtigkeit aus.

Das ist die erste Annäherung an das wirkliche Verfahren, die Feinheiten sind kompliziert und so geheim wie das Rezept von Coca-Cola. Als Beispiel für eine notwendige Verfeinerung kann man etwa darauf hinweisen, dass unser Zufalls-Surfer ein Problem bekäme, wenn er auf eine Seite ohne weiterführende Links geraten würde. Um das zu vermeiden, bekommt er den Rat, bei jedem Surfschritt mit einer gewissen Wahrscheinlichkeit p die gerade vorhandenen Links zu ignorieren und einfach bei irgendeiner zufällig gewählten Seite im Netz weiterzumachen. (Das wäre für unsereinen sicher schwierig zu realisieren, mit einem Verzeichnis aller möglichen Internetseiten ist es aber sogar praktisch machbar.) GOOGLE verwendet, sagt man, die Wahrscheinlichkeit p=15 Prozent: Das scheint ein erfolgreicher Erfahrungswert zu sein. Auch arbeitet GOOGLE ständig daran, etwas gegen das „GOOGLE-bombing“ zu unternehmen. Das ist die Strategie, seine eigene Seite dadurch aufzuwerten, dass man die Anzahl der darauf verweisenden Links künstlich erhöht.

Die Konkurrenten sind natürlich auch nicht faul. Es wird fieberhaft daran gearbeitet, neue Ansätze und Berechnungsverfahren für das Problem der „Wichtigkeit“ von Webseiten zu finden. Für verschiedene Nutzer und verschiedene Kombinationen von Suchbegriffen kann ja „wichtig“ etwas ganz anderes bedeuten. Doch ist es dann fraglich, ob die angeforderte Auswahl auch – wie bei GOOGLE – innerhalb von Sekundenbruchteilen zur Verfügung steht.