Prof. Dr. Michael Eisermann • Institut für Geometrie und Topologie

Tag der Wissenschaft 2016

Die Mathematik hinter Google

$$p_i = \frac{q}{N} + \sum_{j \to i} \frac{1-q}{\ell_j}\, p_j$$

www.igt.uni-stuttgart.de/eiserm • 18. Juni 2016

Begrüßung

Sehr geehrte Damen und Herren,

ich begrüße Sie beim Tag der Wissenschaft an der Universität Stuttgart! In diesem Rahmen darf ich Ihnen heute *die Mathematik hinter Google* erläutern und freue mich, dass Sie hier sind. Die Zusammenfassung sehen Sie auf der Titelseite; ihre Bedeutung wollen wir nun erkunden.

Kurz zu mir: Mein Name ist Michael Eisermann, seit 2009 bin ich in Stuttgart Professor am Fachbereich Mathematik. Meine Forschung und Lehre liegen hauptsächlich im Gebiet der Geometrie und Topologie. (Darüber werde ich heute nicht sprechen, ab Montag wieder gerne.)

Mein Vortrag zu Google dauert etwa 45 Minuten (von 13:30 bis 14:15). Anschließend darf ich kurz die Studiengänge der Mathematik vorstellen.

Sie dürfen gerne Fragen stellen, wenn Sie wollen gleich während des Vortrags, ich ermutige Sie hierzu, ansonsten auch gerne im Anschluss.

Einleitung

Als das World Wide Web Mitte der 1990er noch klein war, da genügte es, zu einer Suchanfrage einfach alle Treffer aufzulisten. Die Liste war noch kurz, der Nutzer konnte sie noch leicht selbst überblicken. Das Internet blieb jedoch nicht lange so klein und überschaubar…

Die Suchmaschine Google ist seit 1998 in Betrieb und dominiert seither den Markt. Sie wird ständig weiterentwickelt. Die meisten Optimierungen hütet Google als Firmengeheimnis, aber das ursprüngliche Grundprinzip ist veröffentlicht und genial einfach.

Bei vorherigen Suchmaschinen musste man endlose Trefferlisten durchforsten, bis man auf die ersten interessanten Ergebnisse stieß. Bei Google stehen sie auf wundersame Weise ganz oben auf der Liste. Wie ist das möglich? Die Antwort liegt (zu einem großen Teil) in folgender Formel. Google misst die Popularität p_i (*PageRank*) jeder Seite i durch folgendes Gleichungssystem:

$$\text{PageRank} \quad p_i = \frac{q}{N} + \sum_{j \to i} \frac{1-q}{\ell_j}\, p_j$$

Keine Angst, die Formel sieht nur auf den ersten Blick kompliziert aus. Ich werde sie anhand von Beispielen Schritt für Schritt erläutern. Wer sowas schon gesehen hat, weiß, dass es sich um eine besonders einfache Formel handelt, nämlich ein *lineares Gleichungssystem*, das keine Quadrate oder komplizierteres enthält. Schon die Formel von Pythagoras $a^2 + b^2 = c^2$ ist komplizierter.

Hier bezeichnet N die Gesamtzahl der Seiten. Der Parameter $q = 0.15$ ist die Sprunghaftigkeit. Jede Seite $j = 1, 2, \ldots, N$ hat jeweils ℓ_j ausgehende Links. Die Summe über $j \to i$ läuft über alle Seiten j, die auf die Seite i verlinken. Das schauen wir uns gleich konkret in Aktion an.

Computers before computers

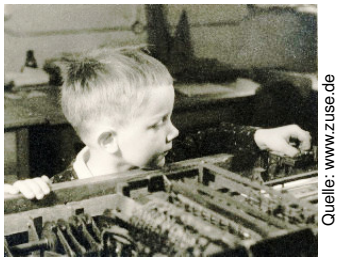
Quelle: www.computerhistory.org

In den 1930er Jahren gab es noch keine automatischen Rechenanlagen, doch der Bedarf war groß: Natur- und Ingenieurwissenschaften, Unternehmen und Behörden verarbeiten immer größere Datenmengen. Das war mühsame Handarbeit, und *Computer* waren damit befasste Angestellte. Das Photo zeigt einen *Computer Room* in Washington DC um 1920. Die hierzu genutzten mechanischen Hilfsgeräte bildeten später die Grundlage für elektronische Rechner. Wenn Sie Freude an alten (und inzwischen historischen) Rechenmaschinen haben, dann besuchen Sie doch gleich heute oder bei nächster Gelegenheit das Computermuseum der Informatik:

🙂 Tipp: Computermuseum, Universitätsstraße 38, Campus Vaihingen

Eine sehr kurze Geschichte des Internets

1941 elektronische Rechenmaschine Z3
1957 UdSSR starten *Sputnik*
1958 USA gründen *ARPA*
1969 ARPANet verbindet 4 Universitäten
1982 Protokoll TCP/IP vereinheitlicht Netze
1991 Webserver des CERN (WWW, HTML)
1993 Browser *Mosaic*, Expansion des WWW
1998 Google indiziert 26 Millionen Webseiten
2000 Google indiziert 1 Milliarde Webseiten
2008 Google sichtet 1 Billion Webseiten



Quelle: www.zuse.de

Konrad Zuses erster Rechner Z1 entstand 1937 in Berlin im elterlichen Wohnzimmer (sozusagen der erste Homecomputer). Er arbeitete noch mechanisch. Der Prototyp Z2 von 1939 nutzte Relais, ebenso die erste funktionsfähige Rechenmaschine Z3. Alle drei wurden im Krieg zerstört, es gibt jedoch Nachbauten. Das Bild von 1949 zeigt Konrad Zuses vierjährigen Sohn Horst vor der Z4.

Eine sehr kurze Geschichte des Internets

Elektronische, frei programmierbare Rechner gibt es seit den 1940er Jahren: zuerst die Zuse Z3 in Deutschland (1941, vor 75 Jahren!), wenig später Colossus in England (1943), Harvard Mark 1 in den USA (1944). Letztere dienen noch im Krieg zu Berechnungen und zur Dechiffrierung.

Im kalten Krieg starten die UdSSR 1957 den ersten künstlichen Erdsatelliten *Sputnik*. Die USA reagieren mit verstärkten Anstrengungen (Sputnik-Schock): Sie investieren ins Bildungsystem und gründen 1958 die Advanced Research Projects Agency (ARPA) für militärisch relevante Forschung und Entwicklung. Es folgt der Wettlauf ins All (*Space Race*): 1961 ist Yuri Gagarin der erste Mensch im All, 1969 ist Neil Armstrong der erste Mensch auf dem Mond.

Eines der ARPA-Projekte ist die Entwicklung eines Computernetzes zwischen Universitäten und Forschungslaboren. (Die Folklore besagt, dass dieses Netz im Falle eines nuklearen Angriffs wenig verwundbar sein sollte, andere Quellen bestreiten diese Zielsetzung.) Das ARPANet verbindet 1969 zunächst nur vier Computer (UCLA, Stanford, UCSB, UUtah). Parallel hierzu entstehen in den 1970 Jahre weitere Netze. Als gemeinsames Protokoll zur Datenübertragung wird 1974 das Transmission Control Program (TCP) entwickelt und 1982 das Internet Protocol (TCP/IP), das bis heute genutzt wird. Seit 1984 gibt es das Domain Name System (DNS).

1991 geht der erste Webserver online – am europäischen Forschungszentrum CERN; die Begriffe *World Wide Web* (WWW) und *Hypertext Markup Language* (HTML) entstehen. 1993 macht der erste Webbrowser *Mosaic* das WWW populär und löst den bis heute andauernden Boom aus.

Mit der rasanten Expansion des WWW werden neue Suchmechanismen dringend notwendig. Das Unternehmen Google wird 1998 gegründet und stellt die gleichnamige Suchmaschine zur Verfügung. Ihr Erfolg beruht vor allem auf der intelligenten Sortierung der Suchergebnisse.

Quiz – Schätzung der Größenordnung

„Data is the new oil.“

Wie viele Suchanfragen beantwortet Google im Jahr?

☐ 1 Mrd ☐ 10 Mrd ☐ 100 Mrd ☐ 1 Bio

Wieviel Gewinn machte Google im Jahr 2015? (Mrd USD)

☐ 12 Microsoft ☐ 17 Google ☐ 32 Exxon ☐ 53 Apple

Welches Unternehmen hat derzeit den höchsten Börsenwert?

☐ Apple 530 ☐ Google 500 ☐ Microsoft 390 ☐ Exxon 370

Quiz – Schätzung der Größenordnung

Den Slogan *„Daten sind das neue Öl.“* hört man oft von Marketingstrategen. Der Vergleich ist griffig, leider auch schnell abgegriffen, und man kann hierüber geteilter Meinung sein. Wir wollen uns in diesem Quiz erst einmal die Größenordnungen vor Augen führen.

Anfangs war der Internet noch klein und übersichtlich. Das Buch *The Internet – Complete Reference* rühmt sich größter Vollständigkeit und präsentiert hierzu sage und schreibe 750 Internet-Resourcen. Das war 1994. Die Lage hat sich seither dramatisch verändert.

Ins öffentliche Bewusstsein kam das Internet so langsam erst vor 20 Jahren. Der seither anhaltende Boom beruht vor allem darauf, das man damit Geld verdienen kann: Im und mit dem Internet werden bereits über 5% des BIP erwirtschaftet. Das ist enorm! Das Bruttoinlandsprodukt (BIP) der Bundesrepublik Deutschland betrug 2015 etwa 3 Billionen Euro.

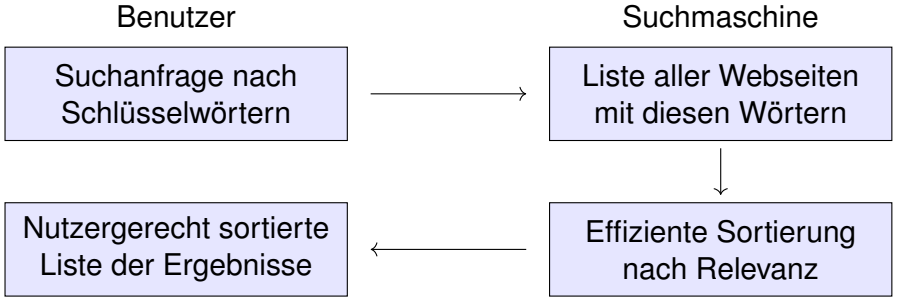
Nach eigener Auskunft beantwortet Google mehr als 1 Billion Suchanfragen pro Jahr, das sind 3 Milliarden pro Tag, also 2 Millionen pro Minute. Wie viele Webseiten es genau gibt, vermag niemand sicher zu sagen. Das liegt auch daran, das viele Seiten nicht mehr handgemacht sind, sondern von Computern generiert werden, z.B. `www.bahn.de`. Bei geeigneter Zusammenfassung und Zählweise kommt man auf etwa 1 Billion Websites.

Googles Gewinn nach Steuern betrug 2015 etwa 17 Mrd USD. Zum Vergleich: Apple 53, Exxon 32, Microsoft 12. Der baden-württembergische Landeshaushalt betrug etwa 44 Milliarden Euro.

Googles Börsenwert betrug diese Woche etwa 500 Mrd USD. (Seit der Neustrukturierung 2015 heißt die Muttergesellschaft *Alphabet*.) Zum Vergleich: Apple 530, Microsoft 390, Exxon 370.

Wie funktioniert eine Suchmaschine?

„Wo simmer denn dran? Aha, heute krieje mer de Suchmaschin. Wat is en Suchmaschin? Da stelle mer uns janz dumm. . . . “



Kernprobleme bei Suchmaschinen (und beim Datamining allgemein):

- Mathematik: Wie misst man Relevanz von Informationen?
- Informatik: Wie verarbeitet man enorme Datenmengen?
- Finanzstrategie: Wie verdient man mit einem Gratisprodukt?

Wie funktioniert eine Suchmaschine?

Stellen Sie sich eine riesige Bibliothek mit einer Billion Dokumenten vor. Einen Bibliothekar gibt es nicht; jeder darf Dokumente hinzufügen. Es gibt weder eine zentrale Redaktion noch einen gemeinsamen Katalog. Sie suchen nun dringend nach einer bestimmten Information. Da Sie ungeduldig sind, möchten Sie das Ergebnis innerhalb einer Sekunde. Das scheint unmöglich. . . und doch gelingt Suchmaschinen genau das!

Zunächst einmal sichtet jede Suchmaschine die vorhandenen Daten. Hierzu läuft unablässig im Hintergrund eine automatische Crawlsoftware (Spider, Webrobots oder kurz Bots genannt), die das Internet permanent durchforstet. Mit den angesteuerten Seiten geschieht zweierlei:

1. Die Suchmaschine speichert jede Seite im eigenen Rechenzentrum. Dabei gibt sie jeder katalogisierten Seite eine Nummer.
2. Sie erstellt einen Index, eine Liste von Schlagwörtern mit den Nummern aller Seiten, auf denen diese vorkommen.

Bei einer Suchabfrage schaut die Suchmaschine in ihrem vorbereiteten Index nach, auf welchen Seiten der gesuchte Begriff vorhanden ist. Für den Nutzer muss die Liste der Suchergebnisse dann „nur noch“ nach Relevanz sortiert werden, damit das Wichtigste ganz oben steht.

Chaos und Struktur: das Web als Graph

Die Seiten des World Wide Web haben einige Besonderheiten:

Dezentral: Viele unabhängige Autoren erzeugen Inhalte.

Heterogen: viele Informationen aber wenig Struktur

Syntax: *hypertext markup language* (HTML)

- Logische Struktur (Titel, Untertitel, Paragraphen, . . .)

```
<h1> Dies ist eine Überschrift. </h1>
<p> Dies ist ein Paragraph. </p>
```

- Erscheinungsbild (Schriftart, fett, kursiv, Farben, . . .)

```
<b> Dieser Text erscheint fett. </b>
<i> Dieser Text erscheint kursiv. </i>
```

- Querverweise / Links (Verweis von einer Seite auf eine andere)

```
<a href="http://www.igt.uni-stuttgart.de/eiserm/">
Hier geht's zur Homepage von Michael Eisermann. </a>
```

Chaos und Struktur: das Web als Graph

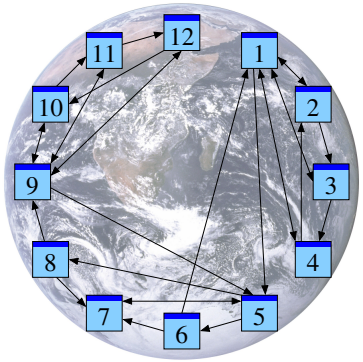
Klassische Texte sind von einem Autor und linear geschrieben: Ein Buch hat einen Anfang und ein Ende, typischerweise liest man es von vorne nach hinten in der Reihenfolge der Seiten. Meist gibt es zudem ein Inhaltsverzeichnis oder einen Index zum leichteren Nachschlagen.

(Ja, liebe Kinder, unsere Vorfahren konnten Texte mit hunderttausend Buchstaben am Stück lesen, ganz ohne Clicks und ohne Werbung. Man nannte das *Buch* und speicherte es auf *Papier*.)

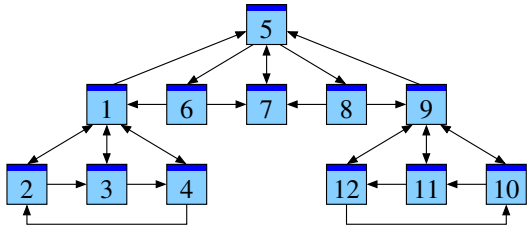
Webseiten bilden hingegen eine gänzlich andere Struktur. Niemand käme auf die Idee, das Internet von Anfang bis Ende durchzulesen: Es hat keine lineare Struktur, keine erste und keine letzte Seite, es ist zudem viel zu groß, und das meiste ist ohnehin uninteressant.

Die Webseiten verweisen gegenseitig aufeinander und bilden so einen *Hypertext*. Genau diese zusätzliche Struktur nutzen Suchmaschinen. Sergey Brin und Larry Page hatten 1998 die geniale Idee, aus der vorliegenden Linkstruktur ein Maß für die Popularität zu berechnen: Jeder Link $j \to i$ wird als Votum der Seite j für die Seite i gewertet. Das Stimmgewicht der Seite j ist ihre eigene Popularität.

Chaos und Struktur: das Web als Graph



Miniaturbeispiel des Web. Gegeben sind die Seiten $i = 1, \ldots, N$ und die Links $i \to j$.
Versuch einer hierarchischen Anordnung:



Eine Seite ist populär, wenn viele Seiten auf sie verweisen? Zu naiv!
Eine Seite ist populär, wenn viele populäre Seiten auf sie verweisen.

Das Web ist ein Graph aus Seiten $i = 1, \ldots, N$ und Links $i \to j$. Ein zufälliger Surfer folgt von der aktuellen Seite zufällig einem der Links.

Aufgabe: Berechnen Sie die Aufenthaltswkten. Konvergieren sie gegen ein Gleichgewicht? Wie schnell? Immer dasselbe, d.h. ist es eindeutig?

Chaos und Struktur: das Web als Graph

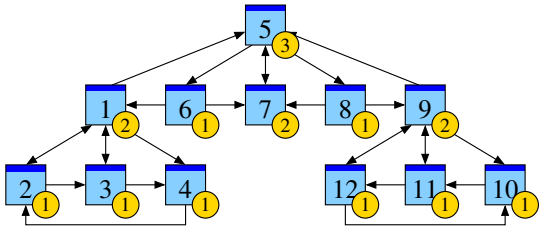
Die Webseiten verweisen gegenseitig aufeinander und bilden so einen *Hypertext*. Zur Illustration betrachten wir ein Miniaturbeispiel bestehend aus 12 Webseiten. Unter den Seiten $1, 2, 3, 4$ wird 1 am häufigsten zitiert. Die Seite 1 scheint daher besonders relevant oder populär. Gleiches gilt für $9, 10, 11, 12$ mit 9 an der Spitze. Die Struktur von $5, 6, 7, 8$ ist ähnlich mit 7 an der Spitze. Aber die Seiten 1 und 9, die wir schon als relevant erkannt haben, verweisen beide auf die Seite 5. Diese scheint daher wichtig und für die Suche besonders relevant.

Diese Anordnung war Handarbeit. Wie lässt sie sich automatisieren? Erster Versuch: Eine Seite ist populär, wenn viele Seiten auf sie verweisen. Diese Linkzählung ist zu naiv und anfällig für Manipulationen!

Zweiter Versuch: Eine Seite ist populär, wenn viele populäre Seiten auf sie verweisen. Das klingt zunächst zirkulär, lässt sich aber in eine einfache Gleichung (wie auf der Titelseite) fassen und lösen.

☺ Hätten Sie diese Aufgabe vor 1998 professionell gelöst, so wären Sie heute vermutlich Milliardär/in.

Irrfahrt eines zufälligen Surfers



Googles Heuristik: Aufenthaltswkt $\sim$ Popularität $\sim$ Relevanz

Aufgabe: Berechnen Sie die Aufenthaltswkten bei Start auf Seite 7.

	1	2	3	4	5	6	7	8	9	10	11	12
$t = 0$	.000	.000	.000	.000	.000	.000	1.00	.000	.000	.000	.000	.000
$t = 1$	.000	.000	.000	.000	1.00	.000	.000	.000	.000	.000	.000	.000
$t = 2$	.000	.000	.000	.000	.000	.333	.333	.333	.000	.000	.000	.000
$t = 3$	.167	.000	.000	.000	.333	.000	.333	.000	.167	.000	.000	.000
$t = 4$	.000	.042	.042	.042	.417	.111	.111	.111	.000	.042	.042	.042
$t = 5$	.118	.021	.021	.021	.111	.139	.250	.139	.118	.021	.021	.021
$\ldots$												
$t = 29$	.117	.059	.059	.059	.177	.059	.117	.059	.117	.059	.059	.059
$t = 30$	.117	.059	.059	.059	.177	.059	.117	.059	.117	.059	.059	.059

Irrfahrt eines zufälligen Surfers

Diese Definition der *Popularität* kann man als mathematische Gleichung formulieren und lösen. Ich erläutere dies lieber mit ein paar Beispielen. Besonders anschaulich ist die Betrachtungsweise des zufälligen Surfers.

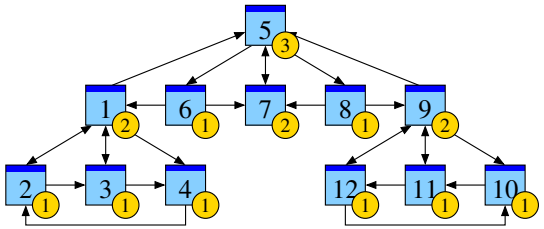
Wir beginnen unsere Reise durch das Internet auf irgendeiner Seite. In unserem Miniaturmodell mit nur 12 Seiten starten wir auf Seite 7.

Von Seite 7 führt nur ein Link weg, wir landen so im ersten Schritt sicher auf Seite 5. Hier führen genau drei Links weiter, und zwar nach 6, 7, 8. Wir folgen irgendeinem, jeweils mit Wkt $1/3$. Nach zwei Schritten sind wir demnach auf Seite 6 oder 7 oder 8, jeweils mit Wkt $1/3$.

Die Zahlen sind in der vorigen Tabelle angegeben. Von Hand ist die Rechnung mühsam, aber ein Computer kann sie schnell ausführen. Sie lässt sich ganz leicht programmieren. Wenn Sie es ausprobieren wollen: Es genügt eine Tabellenkalkulation, etwa *LibreOffice*.

Wir beobachten eine Diffusion: Die Aufenthaltswahrscheinlichkeiten konvergieren gegen eine stationäre Gleichgewichtsverteilung!

Irrfahrt eines zufälligen Surfers



Googles Heuristik: Aufenthaltswkt $\sim$ Popularität $\sim$ Relevanz

Aufgabe: Berechnen Sie die Aufenthaltswkten bei Start auf Seite 1.

	1	2	3	4	5	6	7	8	9	10	11	12
$t=0$	1.00	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
$t=1$	.000	.250	.250	.250	.250	.000	.000	.000	.000	.000	.000	.000
$t=2$	.375	.125	.125	.125	.000	.083	.083	.083	.000	.000	.000	.000
$t=3$	.229	.156	.156	.156	.177	.000	.083	.000	.042	.000	.000	.000
$t=4$	.234	.135	.135	.135	.151	.059	.059	.059	.000	.010	.010	.010
$t=5$	.233	.126	.126	.126	.118	.050	.109	.050	.045	.005	.005	.005
...												
$t=69$	.117	.059	.059	.059	.177	.059	.117	.059	.117	.059	.059	.059
$t=70$	.117	.059	.059	.059	.177	.059	.117	.059	.117	.059	.059	.059

Irrfahrt eines zufälligen Surfers

Zweites Beispiel: Wir beginnen unsere Reise diesmal auf Seite 1.
Hier führen genau vier Links weiter, nämlich auf die Seiten $2, 3, 4, 5$.
Wir folgen irgendeinem, jeweils mit Wkt $1/4$. Nach dem ersten Schritt sind wir demnach auf Seite 2 oder 3 oder 4 oder 5, jeweils mit Wkt $1/4$.
Von jeder dieser Seiten gehen wir jeweils schrittweise weiter.

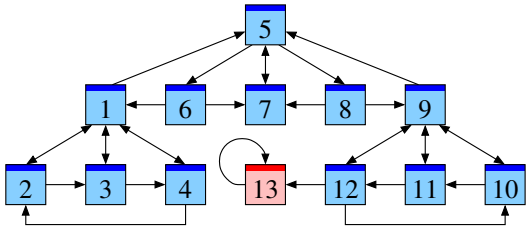
Die Zahlen sind in der obigen Tabelle angegeben. Von Hand ist die Rechnung mühsam, aber ein Computer kann sie schnell ausführen.

Wir beobachten eine Diffusion: Die Aufenthaltswahrscheinlichkeiten konvergieren gegen eine stationäre Gleichgewichtsverteilung!

Alles fließt. Zwei Beobachtungen sind bemerkenswert.
Erstens: Die Verteilungen konvergiert, egal wo wir starten.
Zweitens: Die Gleichgewichtsverteilung ist immer dieselbe!

Sie können dies gerne mit einer Tabellenkalkulation ausprobieren. Numerische Experimente wie diese sind in der Mathematik oft hilfreich. Anschließend kann man die gemachten Beobachtungen als allgemeinen Satz zusammenfassen und beweisen. So entsteht eine Theorie.

Vorsicht vor schwarzen Löchern!



Zum Schutz vor schwarzen Löchern nutzt Google folgendes Modell:

- Teleportation: Mit Wkt $q = 15\%$ beginnt der Surfer irgendwo neu. Anders gesagt, in jedem sechsten Schritt springt er irgendwo hin.
- Irrfahrt: Andernfalls folgt er zufällig einem Link der aktuellen Seite. Das ist die oben illustrierte Irrfahrt entlang der Links.

(Google verwendet noch zahlreiche weitere Verfeinerungen, aber diese sind geheim. Ich diskutiere daher hier nur das einfache Grundmodell.)

Vorsicht vor schwarzen Löchern!

Was passiert bei (Gruppen von) Seiten ohne ausgehende Links?

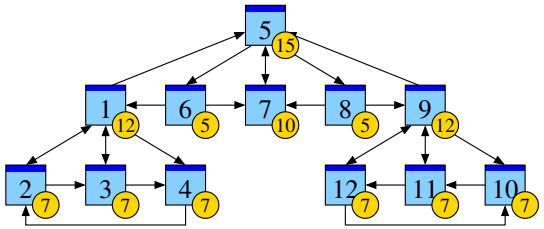
Im Beispiel links ist anschaulich klar: Unser Surfer landet früher oder später auf der Seite 13, wo er den Rest seines Lebens verbringt.
Hier ist unser Modell nicht realistisch! Wir müssen es verbessern.

Das verfeinerte Modell mit Teleportation löst dieses Problem auf erstaunlich einfache Weise. Die Aufenthaltswahrscheinlichkeiten sind ebenso leicht zu definieren und zu berechnen wie zuvor.

Die Konstante q können wir frei wählen. Bei $q = 0$ erhalten wir die Irrfahrt wie zuvor, ohne Teleportation. Bei $q = 1$ springt der Surfer willkürlich, ohne Ansehen der Links. Ein geeignete Wahl von q liegt zwischen 0 und 1. Zum Beispiel entspricht $q = 0.15$ dem Besuch von durchschnittlich etwa 7 aufeinanderfolgenden Seiten. Das entspricht ungefähr dem beobachteten Nutzerverhalten... und lässt sich empirisch anpassen.

Die so berechnete Wahrscheinlichkeitsverteilung entspricht recht gut der Nutzererwartung. Zudem ist sie recht robust gegenüber Manipulationen: Bösartig erzeugte Seiten (Webspam) bekommen wenig Gewicht.

Zufall und Notwendigkeit



Googles Heuristik: Aufenthaltswkt $\sim$ Popularität $\sim$ Relevanz

Aufgabe: Aufenthaltswkten bei Sprunghaftigkeit $q = 0.15$.

	1	2	3	4	5	6	7	8	9	10	11	12
$t=0$	1.00	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000	.000
$t=1$	.013	.225	.225	.225	.225	.013	.013	.013	.013	.013	.013	.013
$t=2$	.305	.111	.111	.111	.028	.076	.087	.076	.034	.020	.020	.020
$t=3$	.186	.124	.124	.124	.158	.021	.085	.021	.071	.028	.028	.028
$t=4$	.180	.105	.105	.105	.140	.057	.075	.057	.057	.040	.040	.040
$t=5$	.171	.095	.095	.095	.126	.052	.101	.052	.087	.042	.042	.042
...												
$t=29$	.120	.066	.066	.066	.150	.055	.102	.055	.120	.066	.066	.066
$t=30$	.120	.066	.066	.066	.150	.055	.102	.055	.120	.066	.066	.066

Zufall und Notwendigkeit

Wir beobachten eine Diffusion: Sie konvergiert gegen eine stationäre Gleichgewichtsverteilung! Ebenso beim Start in 1; sie konvergiert langsamer aber schließlich zum selben Gleichgewicht! Dank dieser Betrachtungsweise löst sich unser LGS sozusagen von allein!

Wir haben dies bislang nur auf einzelnen Beispielen betrachtet. Unser Modell können wir durch einen Algorithmus beschreiben, oder ebenso als einfache lineare Gleichungen zusammenfassen:

Diffusion $$p_i(t+1) = \frac{q}{N} + \sum_{j \to i} \frac{1-q}{\ell_j} p_j(t)$$

Gleichgewicht $$p_i = \frac{q}{N} + \sum_{j \to i} \frac{1-q}{\ell_j} p_j$$

☺ Unsere obigen Beobachtungen zur Konvergenz sind nicht bloß zufällig sondern beruhen auf mathematischen Gesetzmäßigkeiten. Diese kann man beweisen und darf sich darauf verlassen:

Aus dem Fixpunktsatz von Banach (1922) folgt sofort: Bei Sprunghaftigkeit $0 < q \leq 1$ gilt:
(1) Es gibt genau ein Gleichgewicht p. Dieses erfüllt $p_1, \dots, p_N > 0$ und $p_1 + \cdots + p_N = 1$.
(2) Für jede Anfangsverteilung konvergiert die Diffusion gegen die Gleichgewichtsverteilung p.
(3) Die Konvergenz ist mindestens so schnell wie die der geometrischen Folge $(1-q)^n \to 0$.

☺ Diese Gleichung hat viele interessante Aspekte und erlaubt mehrere nützliche Sichtweisen. Algebra: lineare Gleichung. Analysis: Fixpunktsatz, Konvergenz. Stochastik: Irrfahrt, Ergodizität. Geometrie: Gleichgewicht als harmonische Funktion. Numerik: Algorithmen für dünn besetzte Matrizen, Parallelisierung. Physik: Wärmeleitungsgleichung, Potentialgleichung, Spektrum.

Fazit

Webinhalte sind dezentral, heterogen und **wenig strukturiert**.
Eine Suchmaschine soll relevante Suchergebnisse **auflisten**.
Hierzu muss sie diese bewerten und nutzergerecht **sortieren**.

Aus Webseiten und Links berechnet Google ein Maß für die **Popularität**:

Diffusion $$p_i(t+1) = \frac{q}{N} + \sum_{j \to i} \frac{1-q}{\ell_j} p_j(t)$$

Gleichgewicht $$p_i = \frac{q}{N} + \sum_{j \to i} \frac{1-q}{\ell_j} p_j$$

Das lässt sich, wie oben illustriert, schnell und einfach berechnen:
Es entspricht ganz anschaulich der Irrfahrt eines zufälligen Surfers.
Einfach aber erfolgreich. Verfeinerungen bleiben Betriebsgeheimnis.

Vielen Dank für Ihre Aufmerksamkeit!

`www.igt.uni-stuttgart.de/eiserm/popularisierung/#Tag2016`

Fazit

Ich habe hier als mathematische Anwendung die Grundidee der Suchmaschine Google skizziert. Sie wurde 1998 veröffentlicht, fortlaufende Verfeinerungen bleiben Betriebsgeheimnis.

🕮 Sergey Brin, Larry Page: *The anatomy of a large-scale hypertextual web search engine.* Stanford University 1998, `infolab.stanford.edu/pub/papers/google.pdf`

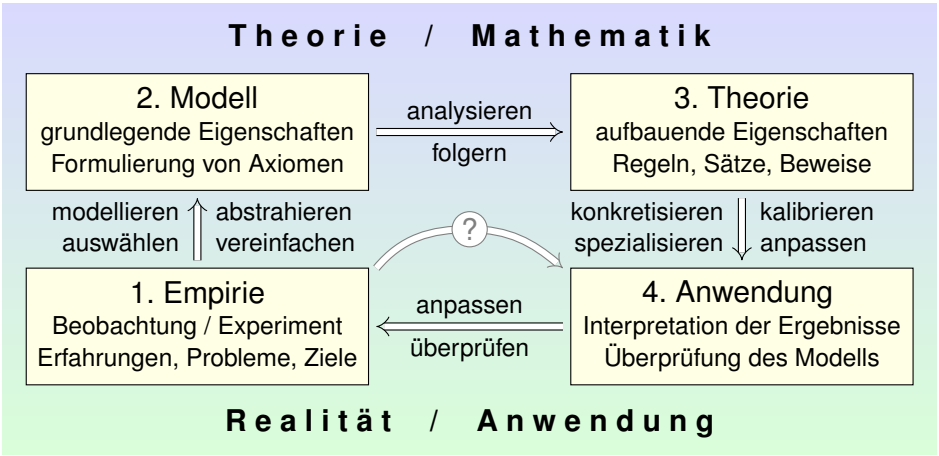
Die Besonderheit von Hypertext sind die gegenseitigen Links: Millionen von Autoren lesen gegenseitig ihre Webseiten, und ihre Bewertung schlägt sich in den Links nieder. Das Modell der Irrfahrt berechnet hieraus ein Maß der Popularität. Es beruht auf einem soliden mathematischen Fundament, das sich in der Praxis bewährt. Hauptargument für das Modell ist sein Erfolg: Die entstehende Sortierung scheint den Nutzererwartungen recht nahe zu kommen.

Zu Beginn sah sich Google rein deskriptiv: Wenn eine Seite relevant ist, dann steht sie oben auf der Liste. Ihr überwältigender Erfolg macht diese Suchmaschine normativ: Wenn eine Seite oben auf der Liste steht, dann ist sie relevant. Für kommerzielle Seiten ist die Optimierung inzwischen unerlässlich und zu einer eigenen Industrie geworden (*search engine optimization*, SEO).

🕮 Selbstdarstellung: `www.google.de/insidesearch/howsearchworks/thestory`, sowie `/webmasters/docs/einfuehrung-in-suchmaschinenoptimierung.pdf`

Offensichtliche Strategie: Viele Links anlocken, am besten von anderen populären Seiten, und selbst nur gut gewählte Links setzen. Somit verändert die Allgegenwart von Google spürbar das Verhalten der Autoren ... und damit die Grundannahme des Modells! Modelle und Werkzeuge werden ständig weiterentwickelt, leider wuchert auch der Webspam. Nutzer, Autoren und Suchmaschinen durchlaufen eine Art gemeinsame Evolution... Es bleibt spannend.

Wozu dient Mathematik?



Mathematik untersucht sowohl abstrakte Strukturen als auch konkrete Anwendungen. Dies sind keine Gegensätze, sondern sie ergänzen sich!

Sie erklärt und quantifiziert Zusammenhänge: Das ist ihr Nutzen!
Dank Abstraktion ist sie universell anwendbar: Das ist ihre Stärke!

Es gibt nichts Praktischeres als eine gute Theorie. (Immanuel Kant)

Wozu dient Mathematik?

Im Alltag wie in den Wissenschaften machen wir Beobachtungen und Erfahrungen (1). Damit wollen wir Probleme lösen (4). Wenn das direkt gelingt, dann ist alles gut. Meist ist der direkte Weg jedoch versperrt, entweder nicht zugänglich oder noch nicht erkennbar. Im Falle von Google ist unser Problem die automatisierte Berechnung der vermeintlichen „Relevanz“ einer Webseite.

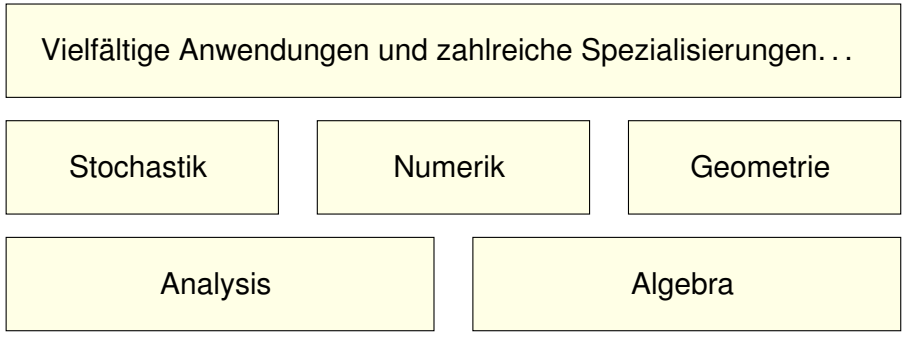
In solchen Fällen lohnt sich eine genauere und systematischere Untersuchung zur Problemlösung. Wir formulieren zunächst ein Modell (2); das dient zur Vereinfachung und zur Strukturierung. Im Falle von Google war das die Linkstruktur, die die Internetseiten untereinander verbindet. Abstrahieren heißt, Wesentliches von Unwesentlichem zu trennen. Das ist eine Kunst.

Hierauf aufbauend erkennen wir gewisse Muster und leiten weitere Eigenschaften ab. Im Beispiel haben wir numerische Experimente gemacht und Konvergenz beobachtet. Anschließend können wir dies als allgemeinen Satz zusammenfassen und beweisen. So entsteht eine Theorie.

Diese können wir nun auf das ursprüngliche spezielle Problem anwenden. Bei der Modellierung und Theoriebildung treffen wir Wahlen, eventuell geschickt oder auch ungeschickt. Wir müssen schließlich unsere Theorie den gegebenen Daten anpassen und ihre Anwendung überprüfen. Wenn das Problem damit gelöst ist, dann ist alles gut. Andernfalls durchlaufen wir den Kreis erneut (auf höherer Ebene): Wir wählen ein besseres Modell, untersuchen dies und wenden es an.

Im Beispiel war das erste Modell ohne Teleportation zwar sinnvoll aber noch sehr anfällig für Manipulationen. Wir haben es durch ein verfeinertes Modell mit Teleportation ersetzt. Auch dieses lässt sich noch verbessern... Die Firma Google durchläuft diesen Kreislauf systematisch: Ständig werden die Ergebnisse getestet und die Methoden verfeinert. Nur so gelingt’s.

Gebiete der Mathematik



Alles Leben ist Problemlösen. (Karl Popper)

Mathematik ist ein universelles Werkzeug und wird überall eingesetzt. Umfassende Informationen bietet die Seite `www.mathematik.de`.

Mathematiker/innen sind Generalisten im Problemlösen, ebenso kreativ wie systematisch, und in nahezu allen Bereichen einsetzbar.

Gebiete der Mathematik

Mathematik ist den meisten aus der Schule **un**bekannt: Sie bedeutet

- nicht (nur) Schulmathematik – sondern weit mehr,
- nicht (nur) Rechnen – sondern Verstehen,
- nicht (nur) Formeln – sondern Ideen.

Wir haben als Miniaturbeispiel die Mathematik hinter Google skizziert. Nahezu alle fünf Grundsäulen der Mathematik sind hieran beteiligt oder zumindest im Ansatz erkennbar.

Im ersten Studienjahr lernen alle Studierenden zunächst Analysis und Lineare Algebra: Das sind die Grundlagen der Mathematik und ihrer Anwendung in Natur- und Ingenieurwissenschaften.

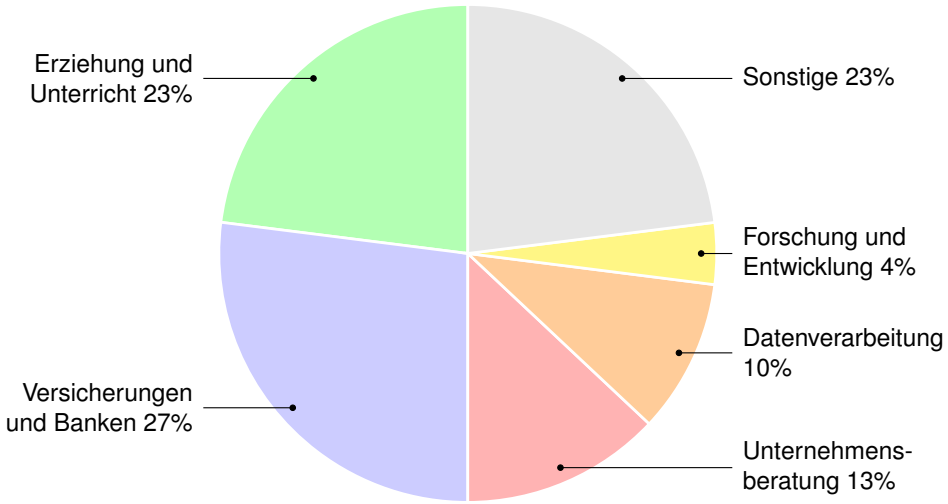
In der Analysis geht es um Konvergenz, Stetigkeit, Differential- und Integralrechnung. In der Algebra geht es um lineare Gleichungen und Matrizen, Vekorräume und lineare Abbildungen.

Die Stochastik beschäftigt sich mit Wahrscheinlichkeitsrechnung und Statistik. Diese sind in vielen Anwendungen sehr wichtig. Auch ich habe diesen Aspekt hier in den Vordergrund gerückt.

Die Numerik oder praktische Mathematik bearbeitet die zusätzlichen Fragen, die bei Umsetzung auf dem Computer entstehen: effiziente Algorithmen, ausreichende Rechengenauigkeit, etc.

Die Geometrie ist das älteste mathematische Teilgebiet. Sie ist nach wie vor sehr aktiv und lebendig und in nahezu allen Anwendungen relevant, von der industriellen Fertigung bis zur Relativitätstheorie. (Hiervon war in unserem Miniaturbeispiel bislang noch nicht die Rede.)

Tätigkeitsfelder und Berufsbild



Quelle: BfA, Stand: 2003

Interviews finden Sie unter `www.mathematik.de` → Mathematik im Beruf.

Tätigkeitsfelder und Berufsbild

Wer Mathematik studiert, qualifiziert sich für sehr viele interessante Tätigkeiten und hat erfahrungsgemäß konstant gute Berufsaussichten: geringe Arbeitslosigkeit, hohes Gehaltsniveau und Berufszufriedenheit.

Mathematiker/innen arbeiten in extrem vielfältigen Bereichen:
als Wissenschaftler/in in der Forschung oder Lehrer/in an der Schule,
in Versicherungen, Banken und Unternehmensberatungen,
in der öffentlichen Verwaltung z.B. bei statistischen Ämtern,
in der Medizin, Biotechnologie, Pharmaindustrie,
in Markt- und Meinungsforschungsinstituten,
in Entwicklungsabteilungen von Unternehmen (Konstruktion, Simulation, Optimierung), zum Beispiel im Maschinenbau oder der Fahrzeugtechnik,
im Informations- und Kommunikationssektor,
in der Softwareentwicklung.

Letzteres bedeutet nicht nur „Kalkulationsprogramme stricken“, sondern z.B. auch die Entwicklung von 3D-Visualisierungstools für Computerspiele und Filmanimationen. Auch hinter intelligenten Suchmaschinen im Internet stehen oft Mathematiker/innen.

Was braucht und was lernt man im Studium?

Freude an der Mathematik!

- Mut zum eigenen Denken
- Problemen auf den Grund gehen
- Logische Zusammenhänge verstehen

⚠ Gute Schulnoten sind ein erster Indikator, mehr nicht.

Wissenschaft braucht Leidenschaft!

- Präzision
- Ausdauer
- Frustrationstoleranz

⚠ Faustregel für jedes Studium: 20% Inspiration, 80% Transpiration

Eigenständigkeit!

- Relativ geringe Anwesenheitszeiten (ca. 20h/Woche VL+Ü)
- Faktor 2 bis 3 an eigener Arbeit (Nachbereitung, Übungen)
- Auch die Semesterferien werden Sie brauchen. (Prüfungen)

⚠ Studium ist nicht Schule!

Ist Mathematik für mich die richtige Wahl?

Mathematik in der Schule entspricht einem 50m–Lauf. Den schafft jeder, manche schnell, manche langsam, aber jeder kommt irgendwie ans Ziel. Das Studium hingegen ist eher ein Marathon: Ausdauer, Selbstdisziplin, Timing. Im Training muss man sich Stück für Stück aufbauen, notfalls auch mal durchquälen. Wer's geschafft hat, sagt stolz: Es lohnt sich!

Entscheiden Sie umsichtig. Vermeiden Sie naive Fehlschlüsse:

- „Als Kind spielte ich Blockflöte. Ich will Musiker/in werden.“
- „Bei Aufsätzen war ich gut. Ich will Schriftsteller/in werden.“
- „Rechnen mochte ich gerne. Ich will Mathematiker/in werden.“

Das ist nützlich, aber nicht ausreichend! Ein Mathematikstudium ist sehr anspruchsvoll und sehr lohnend. Es ist aber nicht jedem zu empfehlen:

- Werden Sie sich Ihrer eigenen Stärken und Schwächen bewusst.
- Informieren Sie sich (selbst)kritisch über mögliche Studiengänge.
- Sprechen Sie mit Studienberatern und Studierenden, . . .

Das Studium hat eine steile Progression. In drei Jahren lernen Sie eine erstaunliche Menge, aber es funktioniert, dank engagierter Dozenten und Tutoren, und vor allem dank Ihrer eigenen hochmotivierten Arbeit.