Die Mathematik hinter Google

Michael Eisermann

Institut für Geometrie und Topologie, Universität Stuttgart

Tag der Wissenschaft, 30. Juni 2012

www.igt.uni-stuttgart.de/eiserm/popularisation/#Tag2012

Sehr geehrte Damen und Herren, ich begrüße Sie!

Mein Name ist Michael Eisermann und ich bin seit drei Jahren Professor für Mathematik an der Universität Stuttgart. Heute möchte ich Ihnen die mathematische Grundidee hinter der Suchmaschine Google erklären.

Der Vortrag richtet sich an interessierte Laien. Mathematik steht im Titel, und Mathematik will ich Ihnen bieten anhand vieler Beispiele auf Schulniveau (einer idealisierten Schule ;-).

Sie können mich während des Vortrags gerne unterbrechen und nachfragen. Ich möchte Sie hierzu ermutigen! Ich stehe Ihnen auch nach dem Vortrag gerne für Fragen zur Verfügung. Ebenso können Sie den Vortrag nochmal in Ruhe lesen, Sie finden ihn auf meiner Webseite.

Wie Sie wissen hat die Mathematik eine lange, reiche Geschichte. Einen winzigen Teil davon sieht man bereits in der Schule. Manche empfinden das als langweilig und anwendungsfern — zu unrecht. Die Mathematik ist Grundlage für Naturwissenschaft und Technik: Sie ist eine moderne Schlüsseltechnologie. Einen kleinen Aspekt hiervon will ich heute vorstellen.

Viele Dinge unseres täglichen Lebens würden ohne Mathematik völlig anders aussehen oder gar nicht existieren. Das betrifft auch und ganz besonders das inzwischen allgegenwärtige Internet, zum Beispiel Datenkompression, Kryptographie oder eben Suchmaschinen.

„Heute krieje mer de Suchmaschin. Wat is en Suchmaschin? Da stelle mer uns janz dumm.“

Übersicht

1. Einführung
   - Was leistet eine Suchmaschine?
   - Das Unternehmen Google Inc.
   - Das Internet ist ein Graph!

2. Wie bestimmt man Relevanz von Internetseiten?
   - Erstes Modell: naive Zählung
   - Zweites Modell: gewichtete Zählung
   - Drittes Modell: rekursive Zählung

3. Mathematische Analyse
   - Der zufällige Surfer auf Irrfahrt im Internet
   - Googles Erfolgsmodell von 1998: PageRank
   - Warum funktioniert PageRank?

Was leistet eine Suchmaschine?

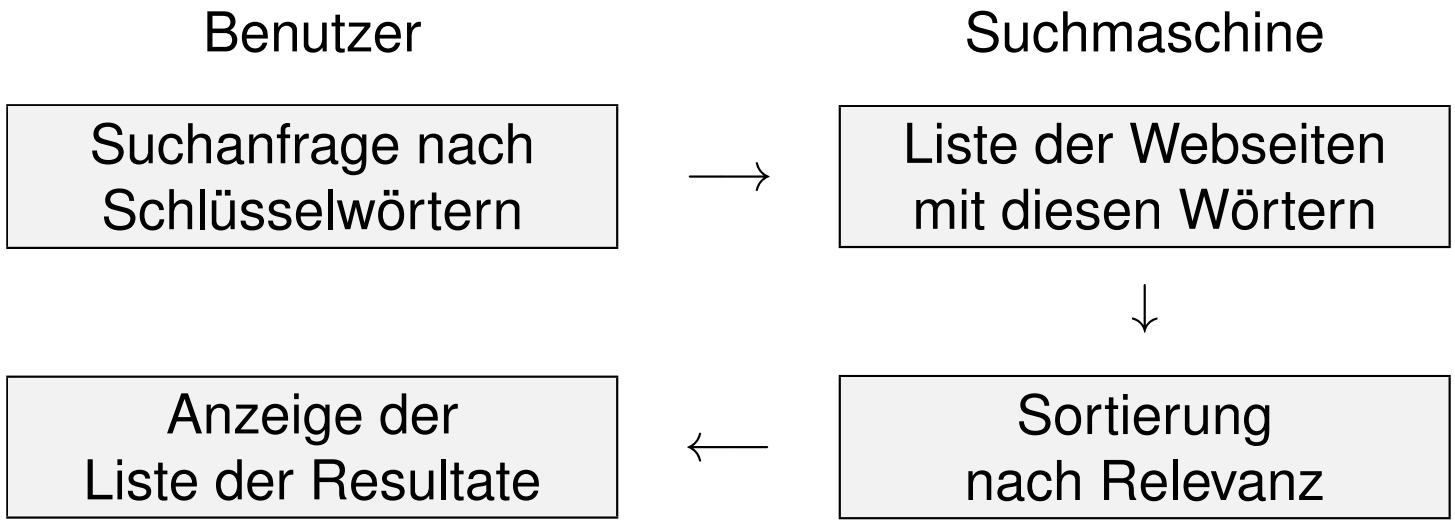
Stellen Sie sich eine riesige Bibliothek mit über hundert Milliarden Dokumenten vor. Einen Bibliothekar gibt es nicht. Jeder darf Dokumente hinzufügen und muss niemanden darüber informieren. Sie suchen nun dringend nach bestimmten Informationen, und da Sie ungeduldig sind, möchten Sie das Ergebnis innerhalb von Sekunden. Wie soll das gehen? Zunächst scheint das völlig unmöglich… und doch gelingt Suchmaschinen im Internet genau das!

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

1. Erstens speichert die Suchmaschine eine Kopie der Seite im hauseigenen Rechenzentrum. Dabei gibt sie jeder katalogisierten Seite eine Nummer.
2. Zweitens wird ein Index erstellt: Dieser ist wie das Schlagwortregister eines Buches eine lange Liste von Wörtern und dazu die Nummern der 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 die Nutzer ebenfalls wichtig: Die Liste der Suchergebnisse muss dann nach Relevanz sortiert werden, damit das Wichtigste ganz oben steht.

Was leistet eine Suchmaschine?



Kernprobleme:

1. Mathematik: Wie bestimmt man die Relevanz?
2. Informatik: Wie verarbeitet man enorme Datenmengen?
3. Finanzstrategie: Wie verdient man an einem Gratisprodukt?

In diesem Vortrag werde ich nur die erste Frage diskutieren und hierzu einige mathematische Modelle diskutieren. Die Ideen sind einfach aber schließlich doch durchschlagend.

Vom Studienprojekt zum weltweiten Unternehmen

1996 Studienprojekt von Page und Brin

September 1998 Unternehmensgründung

Seit 2000 Finanzierung durch Werbung

August 2004 Börsengang

Geschäftsjahr 2011:

Umsatz	38 Mrd. USD	(Fortune Platz 73)
Gewinn	10 Mrd. USD	(Fortune Platz 18)
Börsenwert	210 Mrd. USD	(Fortune Platz 7)
Mitarbeiter	mehr als 33 000	
Computer	vermutlich 900 000	

Zum Vergleich die Top 5 (Geschäftszahlen des Jahres 2011 in Mrd. USD):
Umsatz: Exxon 453, Wal-Mart 447, Chevron 246, ConocoPhillips 237, General Motors 150.
Gewinn: Exxon 41, Chevron 27, Apple 26, Microsoft 23, Ford Motor 20.
Börsenwert: Apple 568, Exxon 405, Microsoft 270, IBM 241, Chevron 211.

Neben Suchmaschine zahlreiche weitere Dienste

Neben viel Lob auch Kritik: Monopol, Datengier

Ziel dieses Vortrags

Die Suchmaschine Google ist seit 1998 in Betrieb und dominiert seither den Markt. Ihre Stärke liegt in der intelligenten Sortierung der Suchergebnisse. 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 der folgenden Formel, die ich hier erklären will.

Erklärung des PageRank–Algorithmus von Google. Kurzfassung:

$$m_i = \frac{c}{n} + \sum_{j \to i} \frac{1-c}{\ell_j} m_j$$

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

📖 Sergey Brin, Larry Page (1998)
The anatomy of a large-scale hypertextual web search engine.
`http://infolab.stanford.edu/pub/papers/google.pdf`

📖 Kurt Bryan, Tanya Leise (2006)
The $25,000,000,000 eigenvector: the linear algebra behind Google.
`http://www.rose-hulman.edu/~bryan/google.html`

Chaos und Struktur im Internet

Dezentral:
Viele unabhängige Benutzer erzeugen und verändern Inhalte.

Heterogen:
Das Internet bietet viele Informationen. . . aber wenig Struktur.

Gemeinsame Syntax: *hypertext markup language* (HTML)

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

```
<h1> Dies ist der Titel. </h1>
<p> Dies ist ein Paragraph. </p>
```

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

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

- 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>
```

Zeitleiste zur Entwicklung des Internets

1957 UdSSR starten den ersten künstlichen Erdsatelliten *Sputnik*

1958 USA gründen Advanced Research Projects Agency (ARPA)

1960ff Visionen/Theorie zu Netzwerken, Datenpaketen, HyperText

1969 ARPANet (zunächst UCLA, Stanford, UCSB, UUtah)

1972 erste Verwendung von Email und Telnet

1974 Transmission Control Program (TCP)

1982 Internet Protocol (TCP/IP)

1984 Domain Name System (DNS)

1991 erster Webserver am CERN (WWW, HTML)

1993 Webbrowser Mosaic, Expansion des WWW

1998 Google indiziert 26 Millionen Webseiten

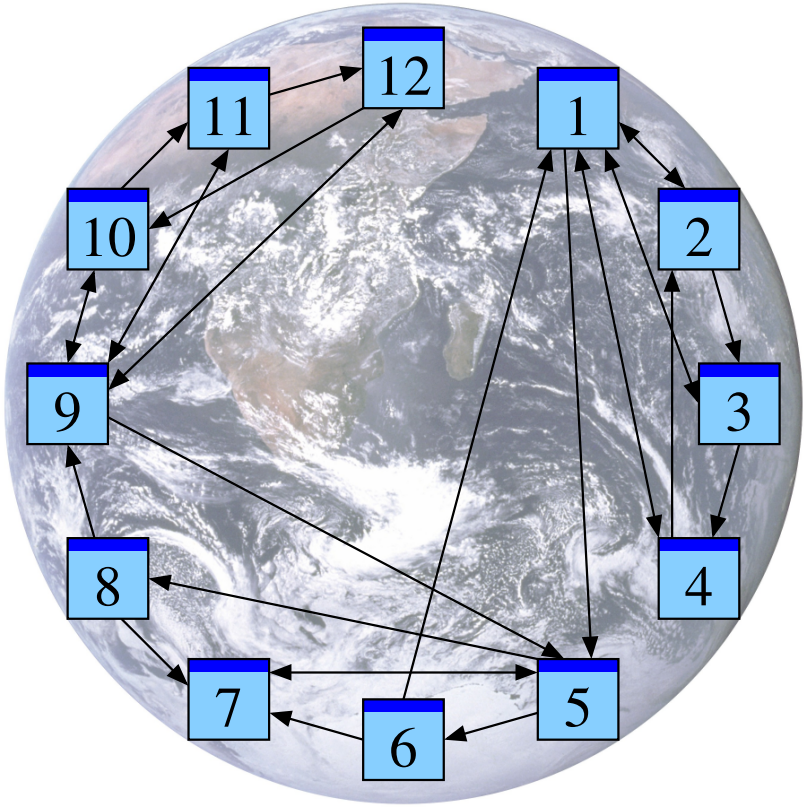
2000 Google indiziert 1 Milliarde Webseiten

2008 Google sichtet eine Billion Webseiten

Das Internet ist ein Graph!

Die Links enthalten Information. Diese wollen wir auswerten.

Miniaturbeispiel:



Hier $1 \to 2,3,4,5$; $2 \to 1,3$;
$3 \to 1,4$; $4 \to 1,2$; $5 \to 6,7,8$;
$6 \to 1,7$; $7 \to 5$; $8 \to 7,9$;
$9 \to 5,10,11,12$; $10 \to 9,11$;
$11 \to 9,12$; $12 \to 9,10$.

Umgeschrieben $2,3,4,6 \to 1$;
$1,4 \to 2$; $1,2 \to 3$; $1,3 \to 4$;
$1,7,9 \to 5$; $5 \to 6$; $5,6,8 \to 7$;
$5 \to 8$; $8,10,11,12 \to 9$;
$9,12 \to 10$; $9,10 \to 11$;
$9,11 \to 12$.

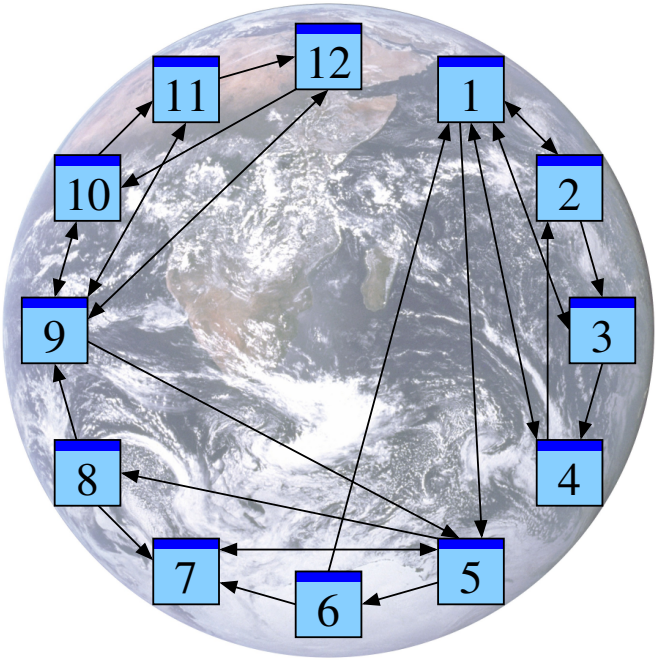
Das beschreibt die Rohdaten. Wie kann man hieraus eine Bewertung der Seiten gewinnen?

Wie kann man diese Information nutzen?

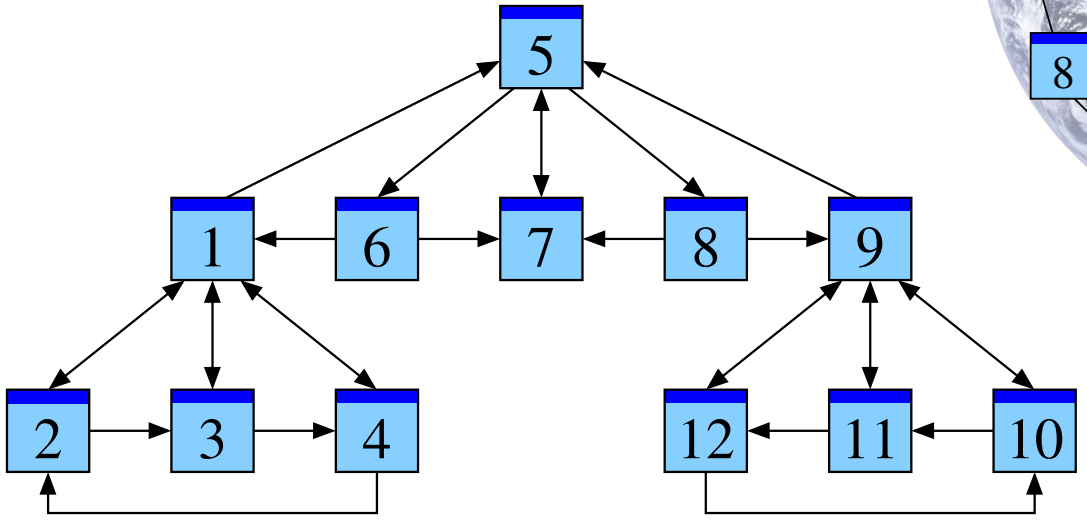
Wie kann man das Internet hierarchisieren?

Die Seiten nach Wichtigkeit sortieren?

Betrachten wir unser Miniaturbeispiel. . .



Ein möglicher Vorschlag:



Diese Anordnung war Handarbeit. . . Ist sie plausibel?

Lässt sie sich automatisieren? Nach welchen Regeln?

Begründung der vorgeschlagenen Anordnung

Wir betrachten den Graphen $1 \to 2,3,4,5;$ $2 \to 1,3;$ $3 \to 1,4;$
$4 \to 1,2;$ $5 \to 6,7,8;$ $6 \to 1,7;$ $7 \to 5;$ $8 \to 7,9;$
$9 \to 5,10,11,12;$ $10 \to 9,11;$ $11 \to 9,12;$ $12 \to 9,10.$

Die oben vorgeschlagene Anordnung ist zunächst willkürlich.
Die folgenden Argumente sollen ihre Plausibilität belegen.

Unter den Seiten P_1, P_2, P_3, P_4 wird P_1 am häufigsten zitiert und bildet eine Art gemeinsame Wurzel. Die Seite P_1 scheint daher für die Suche besonders relevant.

Gleiches gilt für $P_9, P_{10}, P_{11}, P_{12}$ mit P_9 an der Spitze.

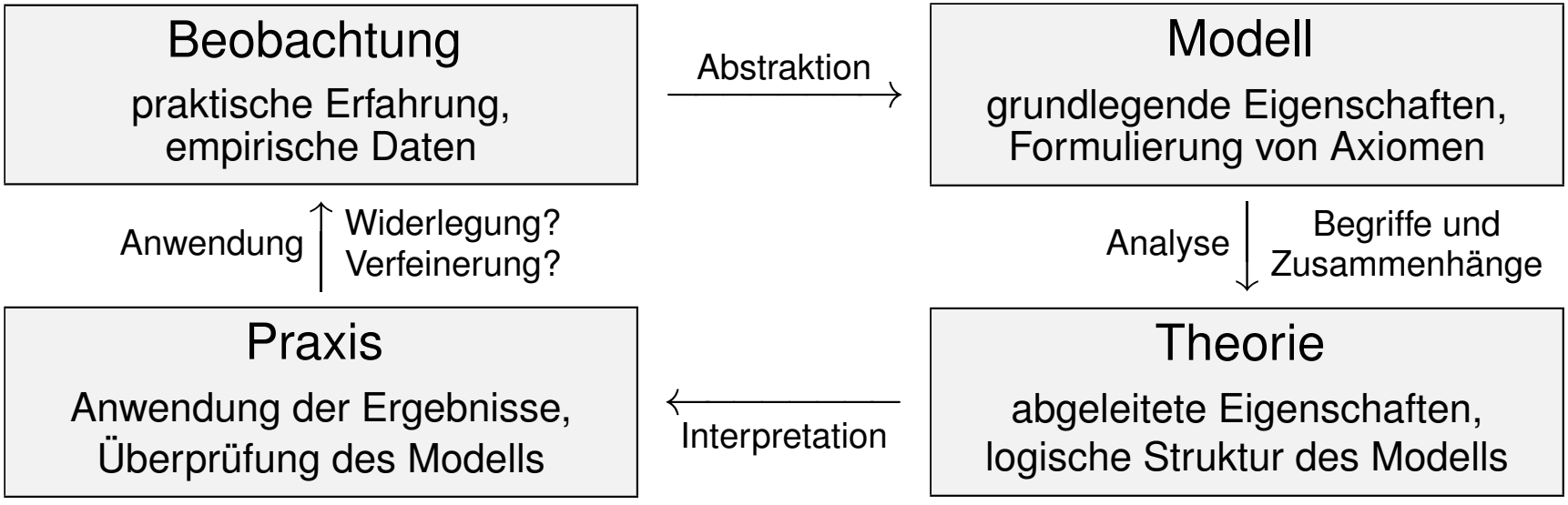
Die Struktur von P_5, P_6, P_7, P_8 ist ähnlich mit P_7 an der Spitze.

Aber die Seiten P_1 et P_9, die wir schon als relevant erkannt haben, verlinken beide auf P_5. Man kann daher vermuten, dass die Seite P_5 für alle wesentlich ist, und für die Suche besonders relevant.

Was bedeutet Modellierung?

Konkretes Problem:
Wie bestimmt man die Relevanz von Webseiten?

Modellierungskreislauf:
Typische Wechselwirkung zwischen Theorie und Praxis:



Erstes Modell: naive Zählung

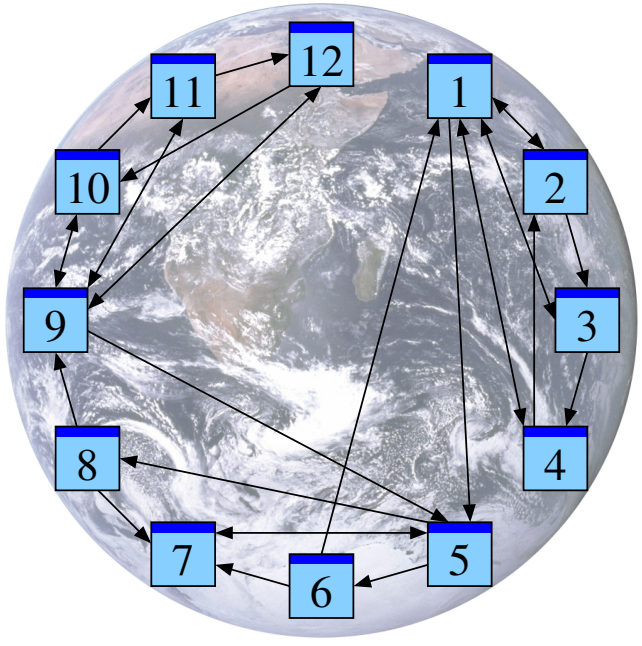
Heuristik: *Wenn eine Seite wichtig ist, dann bekommt sie viele Links.*

Naive Umkehrung: *Bekommt eine Seite viele Links, so ist sie wichtig.*

Als Maß für die Wichtigkeit einer Seite P_i könnten wir demnach die eingehenden Links ($j \to i$) zählen. Schreibweise:

$$m_i = \sum_{j \to i} 1$$

Beispiel:

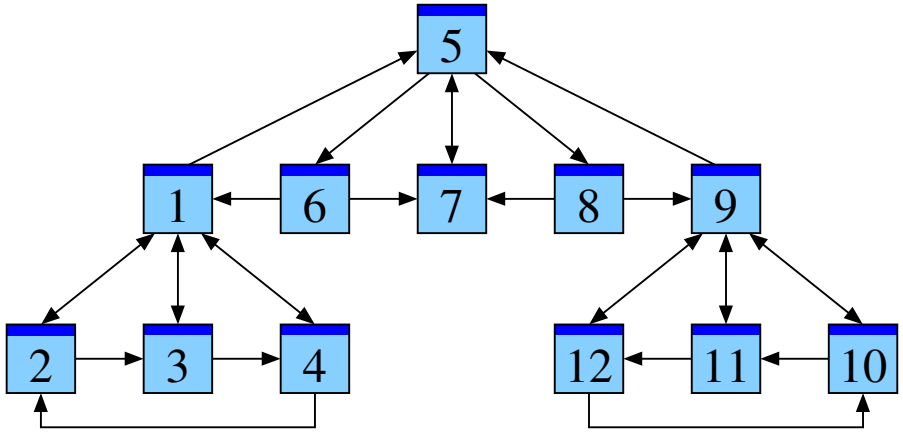


$$2,3,4,6 \to 1; \quad 1,4 \to 2; \quad 1,2 \to 3;$$
$$1,3 \to 4; \quad 1,7,9 \to 5; \quad 5 \to 6;$$
$$5,6,8 \to 7; \quad 5 \to 8; \quad 8,10,11,12 \to 9;$$
$$9,12 \to 10; \quad 9,10 \to 11; \quad 9,11 \to 12.$$

$$m_1 = m_9 = 4$$
$$m_5 = m_7 = 3$$
$$m_2 = m_3 = m_4 = 2$$
$$m_{10} = m_{11} = m_{12} = 2$$
$$m_6 = m_8 = 1$$

Erstes Modell: Formulierung als Matrix

Darstellung als Matrix $A = (a_{ij})$ mit $a_{ij} = \begin{cases} 1 & \text{falls } j \to i, \\ 0 & \text{sonst.} \end{cases}$



$$A = \begin{pmatrix}
0 & 1 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
1 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 1 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 & 1 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0
\end{pmatrix}$$

Spaltensumme = Anzahl der ausgehenden Links.

Zeilensumme = Anzahl der eingehenden Links.

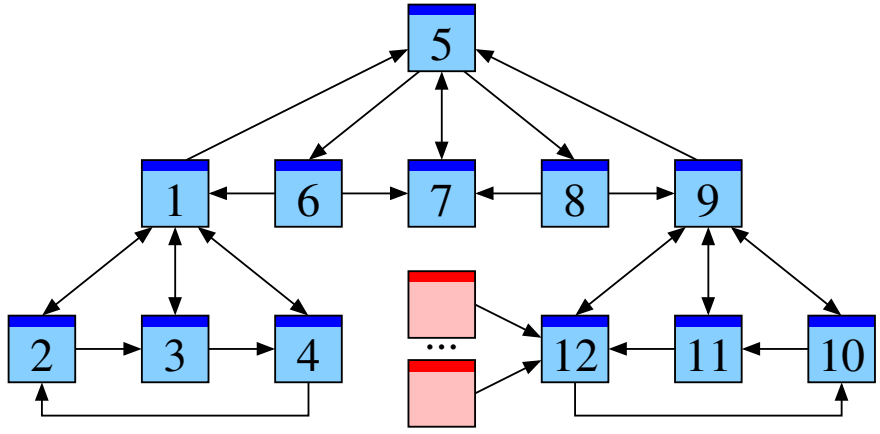
Es fällt auf, dass diese Matrix *dünn besetzt* ist, das heißt viele Nullen enthält. In realistischen Anwendungen auf das Internet ist dies noch frappierender: Es gibt Milliarden Internetseiten, aber eine typische Seite Seite verweist nur auf wenige andere mit ein paar Dutzend Links.

- ☺ Leicht zu definieren und zu berechnen
- ☹ Wenig treffsicher, entspricht nicht der Nutzererwartung
- ☹ Wenig robust, leicht zu manipulieren durch „Linkfarmen“

Erstes Modell: Anfälligkeit für Linkfarmen

Das Internet ist dezentral organisiert: Viele unabhängige Benutzer erzeugen und verändern Inhalte. Jeder verfolgt seine eigenen Interessen, zum Beispiel die Optimierung seiner Seiten.

Die Anzahl der eingehenden Links lässt sich leicht manipulieren:



Auch Links von sinnlosen Seiten erhöhen die Wertung

$$m_i = \sum_{j \to i} 1.$$

Die Platzierung einer Seite verbessert sich, wenn möglichst viele Seiten auf sie verlinken. Durch Einfügen sinnloser Seiten lässt sich daher die Platzierung willkürlich verbessern.

Zweites Modell: gewichtete Zählung

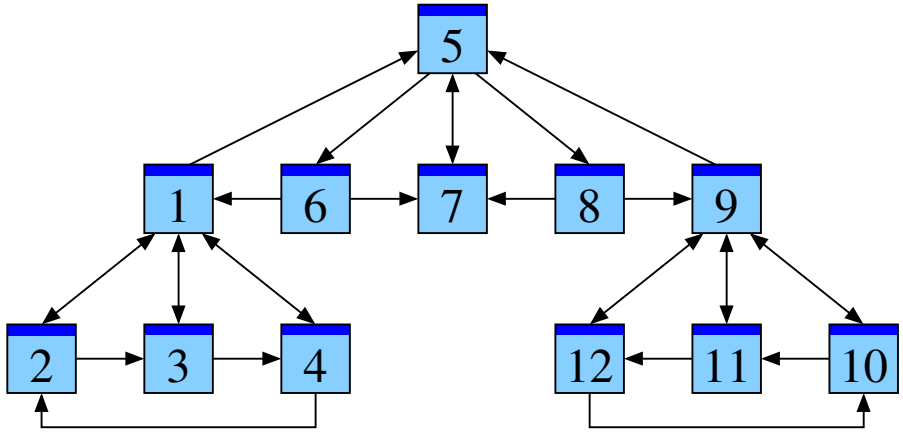
Heuristik: *Ein Link $j \to i$ ist ein Votum der Seite P_j für die Seite P_i.*

Wir geben zunächst jeder Seite P_j dasselbe Stimmgewicht 1.
Bei ℓ_j ausgehenden Links teilt sich dieses in ℓ_j gleiche Teile.

Als Abstimmungsergebnis für die Seite P_i erhalten wir

$$m_i = \sum_{j \to i} \frac{1}{\ell_j}.$$

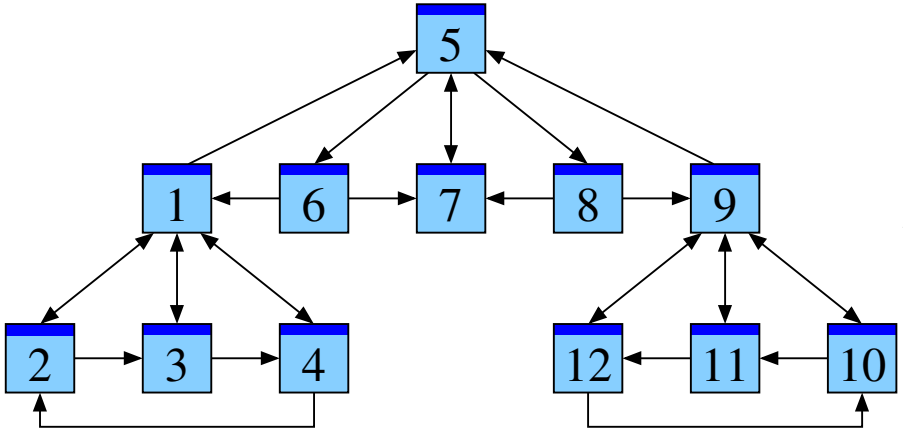
Beispiel:



$$m = \begin{pmatrix} 2 \\ 3/4 \\ 3/4 \\ 3/4 \\ 3/2 \\ 1/3 \\ 4/3 \\ 1/3 \\ 2 \\ 3/4 \\ 3/4 \\ 3/4 \end{pmatrix}$$

Zweites Modell: Formulierung als Matrix

Darstellung als Matrix $A = (a_{ij})$ mit $a_{ij} = \begin{cases} \frac{1}{\ell_j} & \text{falls } j \to i, \\ 0 & \text{sonst.} \end{cases}$



$$A = \begin{pmatrix}
0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \frac{1}{4} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{2} & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & 0 & 0 & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{2} & 0 &
\end{pmatrix}$$

Jede Spaltensumme ist gleich 1. Die i–te Zeilensumme ergibt m_i.

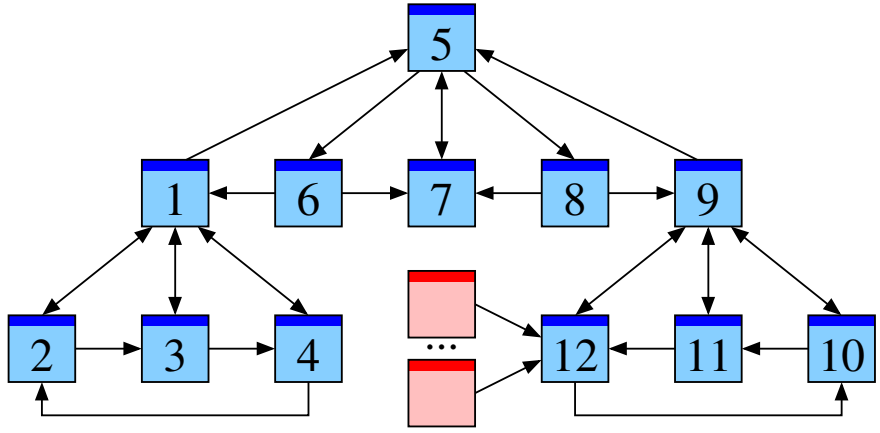
Wie zuvor ist auch diese Matrix dünn besetzt, das heißt, sie enthält viele Nullen.

- ☺ Leicht zu definieren und zu berechnen
- ☹ Entspricht noch nicht der Nutzererwartung
- ☹ Leicht zu manipulieren durch „Linkfarmen“

Zweites Modell: Anfälligkeit für Linkfarmen

Das Internet ist dezentral: Viele unabhängige Benutzer erzeugen und verändern Inhalte.

Auch die gewichtete Anzahl der Links lässt sich leicht manipulieren:



Auch Links von sinnlosen Seiten erhöhen die Wertung

$$m_i = \sum_{j \to i} \frac{1}{\ell_j}.$$

Die Platzierung einer Seite verbessert sich, wenn möglichst viele Seiten auf sie verlinken.
Durch Einfügen sinnloser Seiten lässt sich daher die Platzierung willkürlich verbessern.

Drittes Modell: rekursive Zählung

Heuristik:
Eine Seite ist wichtig, wenn viele wichtige Seiten auf sie verlinken.

Ein Link $j \to i$ ist ein Votum der Seite P_j für die Seite P_i.
Wir geben jeder Seite P_j ihr eigenes Stimmgewicht, nämlich m_j.

Als Abstimmungsergebnis für die Seite P_i erhalten wir dann

$$m_i = \sum_{j \to i} \frac{1}{\ell_j} m_j.$$

Zuerst scheint diese Gleichung zirkulär: Zur Berechnung von m_i (links) müssen wir alle m_j kennen (rechts). Nüchtern betrachtet ist dies aber nur eine lineare Gleichung.

Ich schreibe sie für unser Miniaturbeispiel einmal aus (nächste Seite). Hier handelt es sich um 12 lineare Gleichungen mit 12 Unbekannten. So etwas haben die meisten von Ihnen schon in der Schule gesehen — allerdings etwas kleiner, etwa 2 Gleichungen mit 2 Unbekannten.

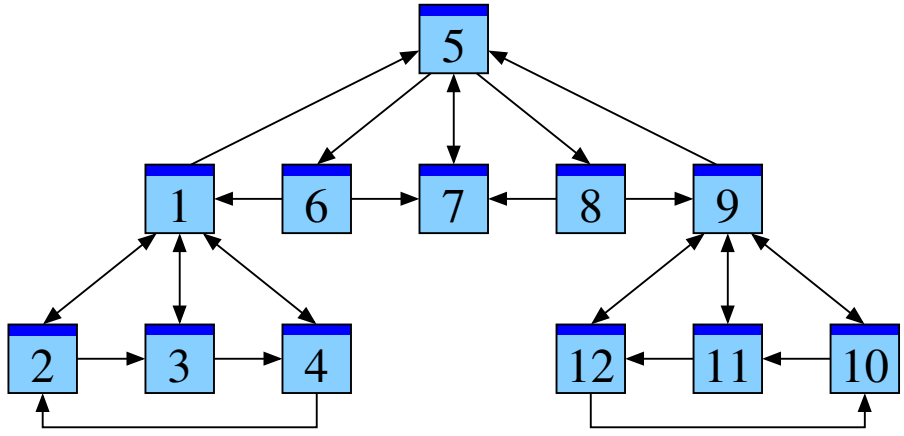
Im Prinzip ist unser Gleichungssystem nicht komplizierter, nur größer.

Die Lineare Algebra erklärt uns, wie man die gesuchten m_i ausrechnen kann. (Die Lineare Algebra ist ein mächtiges Universalwerkzeug. Deshalb lernt man die Anfänge schon in der Schule. Zu voller Blüte gelangt sie dann im naturwissenschaftlich-technischen Studium.)

Selbst in unserem Miniaturbeispiel ist dieses Gleichungssystem nicht ganz leicht zu lösen. (Probieren Sie es!) Aber eine gegebene Lösung ist leicht zu prüfen. (Probieren Sie auch das!)

Drittes Modell: Formulierung als lineare Gleichung

Beispiel:



$$
\begin{array}{lcccccccccccc}
m_1 &=& . & +\frac{1}{2}m_2 & +\frac{1}{2}m_3 & +\frac{1}{2}m_4 & . & +\frac{1}{2}m_6 & . & . & . & . & . \\
m_2 &=& \frac{1}{4}m_1 & . & . & +\frac{1}{2}m_4 & . & . & . & . & . & . & . \\
m_3 &=& \frac{1}{4}m_1 & +\frac{1}{2}m_2 & . & . & . & . & . & . & . & . & . \\
m_4 &=& \frac{1}{4}m_1 & . & +\frac{1}{2}m_3 & . & . & . & . & . & . & . & . \\
m_5 &=& \frac{1}{4}m_1 & . & . & . & . & . & +m_7 & . & +\frac{1}{4}m_9 & . & . \\
m_6 &=& . & . & . & . & +\frac{1}{3}m_5 & . & . & . & . & . & . \\
m_7 &=& . & . & . & . & +\frac{1}{3}m_5 & +\frac{1}{2}m_6 & . & +\frac{1}{2}m_8 & . & . & . \\
m_8 &=& . & . & . & . & +\frac{1}{3}m_5 & . & . & . & . & . & . \\
m_9 &=& . & . & . & . & . & . & . & +\frac{1}{2}m_8 & . & +\frac{1}{2}m_{10} & +\frac{1}{2}m_{11} \\
m_{10} &=& . & . & . & . & . & . & . & . & +\frac{1}{4}m_9 & . & . \\
m_{11} &=& . & . & . & . & . & . & . & . & +\frac{1}{4}m_9 & +\frac{1}{2}m_{10} & . \\
m_{12} &=& . & . & . & . & . & . & . & . & +\frac{1}{4}m_9 & . & +\frac{1}{2}m_{11}
\end{array}
$$

Ein Lob auf die Lineare Algebra!

Drittes Modell: Formulierung als Matrix

Mit unserer Matrix A schreibt sich diese Gleichung kurz $m = Am$.

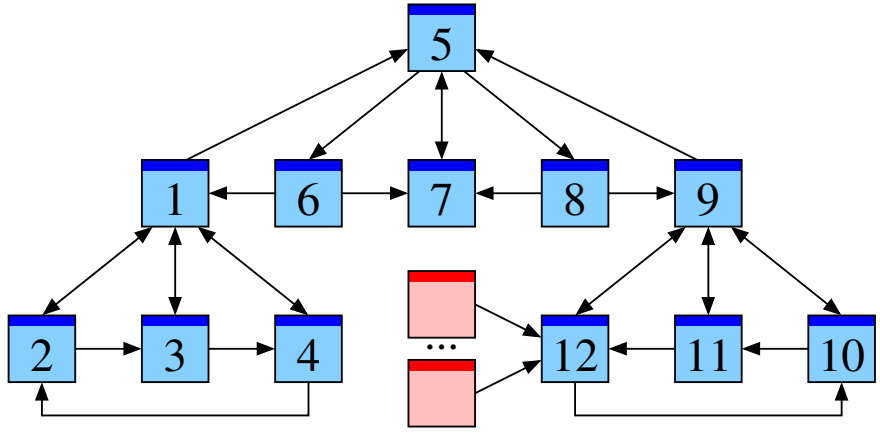
$$A = \begin{pmatrix}
0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & 0 & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
\frac{1}{4} & 0 & 0 & 0 & 0 & 0 & 1 & 0 & \frac{1}{4} & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & \frac{1}{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{2} & 0 & \frac{1}{2} & \frac{1}{2} & \frac{1}{2} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & 0 & \frac{1}{2} \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & \frac{1}{2} & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{1}{4} & 0 & \frac{1}{2} & 0
\end{pmatrix}, \qquad m = \begin{pmatrix} 2 \\ 1 \\ 1 \\ 1 \\ 3 \\ 1 \\ 2 \\ 1 \\ 2 \\ 1 \\ 1 \\ 1 \end{pmatrix}, \qquad \frac{1}{17} m \approx \begin{pmatrix} .117 \\ .059 \\ .059 \\ .059 \\ .177 \\ .059 \\ .117 \\ .059 \\ .117 \\ .059 \\ .059 \\ .059 \end{pmatrix}.$$

- ☹ Leicht zu definieren aber nicht ganz so leicht zu berechnen
  Grundlegende Fragen: Existiert eine Lösung? Ist sie eindeutig?
  Wie kann man sie berechnen? Wie geht das möglichst effizient?
- ☺ Entspricht oft recht gut der Nutzererwartung
- ☺ Nicht mehr ganz so leicht zu manipulieren

Drittes Modell: Robustheit gegenüber Linkfarmen

Das Internet ist dezentral: Viele unabhängige Benutzer erzeugen und verändern Inhalte.

Das dritte Modell ignoriert Linkfarmen:



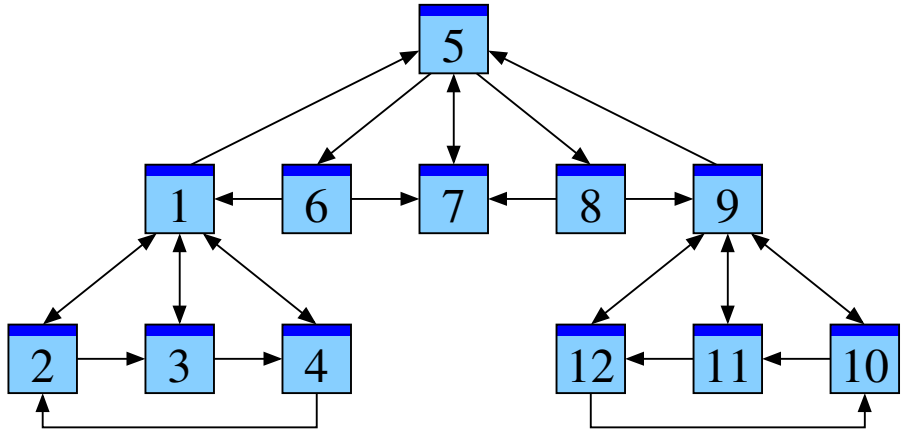
Sinnlose Seiten erhalten Gewicht 0 und erhöhen nicht die Wertung

$$m_i = \sum_{j \to i} \frac{1}{\ell_j} m_j .$$

Das Gewicht jedes Links ist proportional zum Gewicht der aussendenden Seite. Sinnlose Seiten ändern nichts: Sie erhalten das Gewicht 0 und tragen nichts zur Berechnung bei. Wir nehmen hierbei an, dass keine sinnvolle Seite auf eine sinnlose Spamseite verlinkt.

Irrfahrt im Internet

Wir folgen einem „zufälligen Surfer“ im Internet.



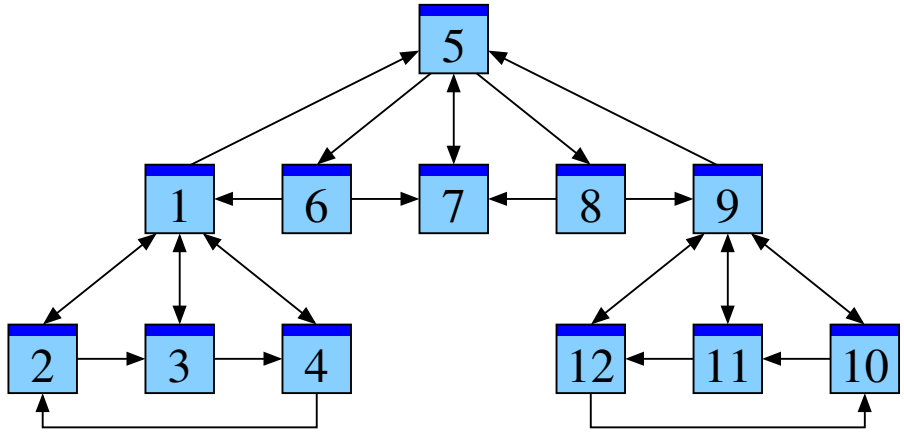
Entwicklung der Aufenthaltswahrscheinlichkeit bei Start auf Seite P_7:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{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
...												
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

Dies ist eine Art Diffusion. Sie konvergiert gegen eine Gleichgewichtsverteilung.
Dies ist genau die Lösung, die wir oben gefunden haben (normiert auf Summe 1).

Irrfahrt im Internet

Wir überprüfen diese Beobachtung in einem zweiten Durchgang.



Entwicklung der Aufenthaltswahrscheinlichkeit bei Start auf Seite P_1:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{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

Wir finden erneut eine Diffusion, und diese konvergiert zum selben Gleichgewicht!
Dank dieser Betrachtungsweise löst sich die Gleichung sozusagen von allein!

Zufall und Notwendigkeit

Auch der Zufall gehorcht Gesetzen: Man kann damit rechnen!

Zur Zeit t ist unser Surfer auf der Seite P_j mit Wahrscheinlichkeit p_j.
Er wählt zufällig einen der ℓ_j ausgehenden Links.
Also folgt er dem Link $j \to i$ mit Wahrscheinlichkeit $\frac{1}{\ell_j} p_j$.

Zur Zeit $t+1$ landet er auf der Seite P_i mit Wahrscheinlichkeit

$$p'_i = \sum_{j \to i} \frac{1}{\ell_j} p_j .$$

Wenn man die Aufenthaltswahrscheinlichkeiten zur Zeit t kennt, so kann man hieraus leicht die Aufenthaltswahrscheinlichkeiten zur Zeit $t+1$ berechnen. Mit dieser einfachen Gleichung habe ich die obigen Beispiele ausgerechnet. Insbesondere finden wir folgendes:

Jede Gleichgewichtsverteilung erfüllt demnach die Gleichung

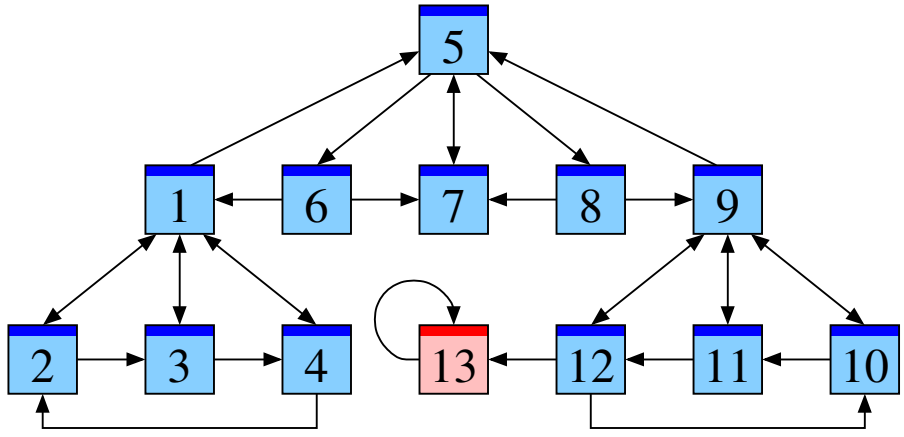
$$m_i = \sum_{j \to i} \frac{1}{\ell_j} m_j .$$

Ein Lob auf die Wahrscheinlichkeitsrechnung!

Wir sehen dieselbe Gleichung plötzlich aus einem völlig neuen Blickwinkel.
Dies werden wir im Folgenden nutzen, sie zu verstehen und zu lösen…

Vorsicht vor schwarzen Löchern!

Was wenn es (Gruppen von) Seiten ohne ausgehende Links gibt?



Anschaulich ist klar: Unser Surfer landet früher oder später auf der Seite P_{13}, wo er den Rest seines Lebens verbringt. Hier ist unser Modell nicht realistisch!

Die Gleichgewichtsverteilung können wir sofort ausrechnen:

$$m = (0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1).$$

Es handelt sich um ein System von 13 Gleichungen und 13 Unbekannten. Dies im Kopf zu lösen ist im Allgemeinen wohl kaum möglich. Dennoch ist uns dies soeben gelungen, da wir die Bedeutung der Gleichung verstehen.

PageRank: Googles Erfolgsmodell

Zum Schutz vor schwarzen Löchern nutzt Google folgendes Modell:

1. Teleportation: Mit Wahrscheinlichkeit c beginnt der Surfer neu auf irgendeiner der n Seiten des Internet. $(0 \leq c \leq 1)$
2. Irrfahrt: Andernfalls, mit Wahrscheinlichkeit $1 - c$, folgt der Surfer einem Link der Seite P_j, zufällig ausgewählt wie zuvor.

Die Aufenthaltswahrscheinlichkeiten berechnen sich also zu

$$p'_i = \frac{c}{n} + \sum_{j \to i} \frac{1-c}{\ell_j} p_j .$$

Jede Gleichgewichtsverteilung erfüllt demnach die Gleichung

$$m_i = \frac{c}{n} + \sum_{j \to i} \frac{1-c}{\ell_j} m_j .$$

Dieses Modell vereint mehrere Vorzüge:

- Leicht zu definieren und zu berechnen
- Entspricht recht gut der Nutzererwartung
- Robust gegenüber Manipulationen

PageRank: Interpretation der Konstanten c

Die Konstante c können wir frei wählen. Zwei Extremfälle:

- Bei $c = 0$ erhalten wir die Irrfahrt wie zuvor, ohne Teleportation.
- Bei $c = 1$ springt der Surfer willkürlich, ohne Ansehen der Links.

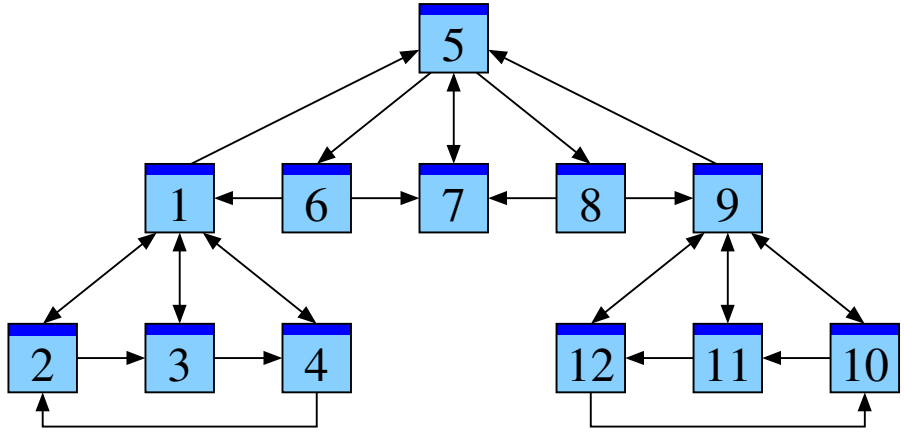
Ein geeignete Wahl von c liegt zwischen 0 und 1.

Es handelt sich um ein Bernoulli–Experiment: Der Kehrwert $\frac{1}{c}$ ist die mittlere Anzahl der besuchten Seiten bevor der Surfer neu beginnt.

Zum Beispiel entspricht $c = 0.15$ dem Besuch von durchschnittlich etwa 7 aufeinanderfolgenden Seiten. Das entspricht ungefähr dem beobachteten Nutzerverhalten. . . und lässt sich empirisch anpassen.

Bemerkung: Auch die PageRank-Gleichung lässt sich mit Hilfe einer Matrix schreiben: Die Teleportation führt zu einem Grundbetrag c/n in jeder Spalte plus $(1 - c)$ mal die alte Spalte. Die neue Matrix ist allerdings nicht mehr dünn besetzt. Für die praktische Berechnung ist es daher günstiger, die Teleportation gesondert zu behandeln, und dann die Irrfahrt wie zuvor.

PageRank: schnelle Diffusion zu einem Gleichgewicht

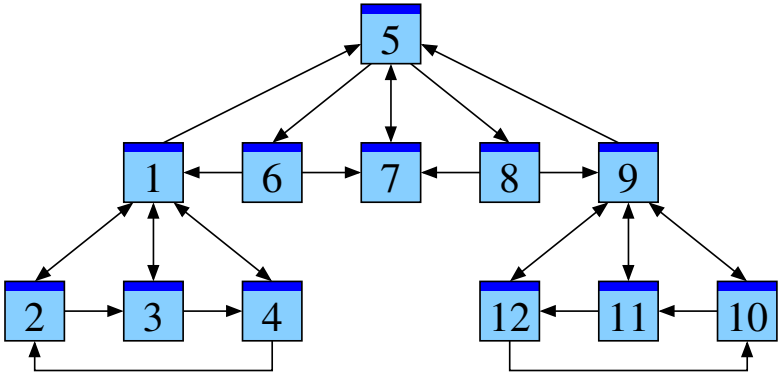


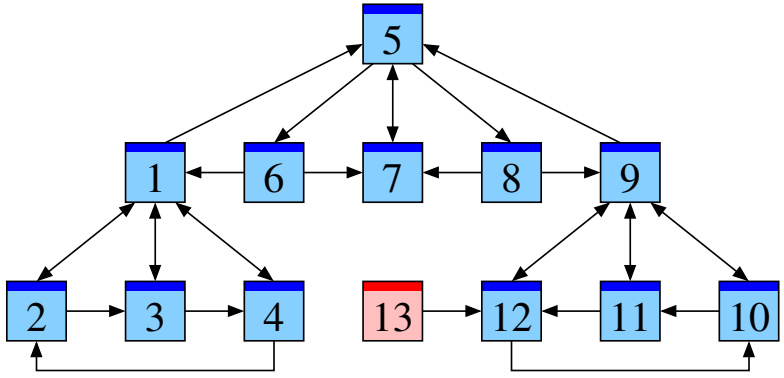
Entwicklung der Aufenthaltswahrscheinlichkeit für $c = 0.15$:

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{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

Ein Lob auf die Numerik!

PageRank: Robustheit gegenüber Manipulationen





Im ersten Beispiel haben wir:

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}
.120	.066	.066	.066	.150	.055	.102	.055	.120	.066	.066	.066

Einfügen einer Spamseite ändert hieran wenig:

P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9	P_{10}	P_{11}	P_{12}	P_{13}
.113	.062	.062	.062	.145	.053	.097	.053	.126	.071	.069	.077	.012

Warum funktioniert PageRank?

Wie bestimmt man die Relevanz von Webseiten?

Dies ist zuerst ein Problem der Modellierung:

- Welche Daten sind gegeben? Was will man hieraus extrahieren?
- Mathematik liefert eine Sprache zur Formulierung der Modelle.
- Rechnungen erlauben die Modelle zu testen und zu verfeinern.

Ein geeignetes Modell können wir dann genauer untersuchen:

- Existiert immer eine Lösung? Ist sie eindeutig?
- Wie kann man sie berechnen? möglichst effizient?

Für realistische Anwendungen ist die Effizienz wesentlich.
Google indiziert mehrere Milliarden Webseiten!

Warum funktioniert PageRank?

Entwicklung der Aufenthaltswahrscheinlichkeit:

$$p'_i = \frac{c}{n} + \sum_{j \to i} \frac{1-c}{\ell_j} p_j. \tag{1}$$

Gleichung für die Gleichgewichtsverteilung:

$$m_i = \frac{c}{n} + \sum_{j \to i} \frac{1-c}{\ell_j} m_j. \tag{2}$$

Satz ($\Leftarrow$ Fixpunktsatz von Banach, 1922)

Wir betrachten einen endlichen Graphen und $0 < c \leq 1$*. Dann gilt:*

1. *Die Gleichung* (2) *hat genau eine Lösung* m*.*
   Diese erfüllt $m_1, \ldots, m_n > 0$ *und* $m_1 + \cdots + m_n = 1$*.*
2. *Für jede Anfangsverteilung konvergiert die Diffusion* (1)
   gegen die Gleichgewichtsverteilung m*.*
3. *Die Konvergenz ist mindestens so schnell wie die Konvergenz der geometrischen Folge* $(1-c)^n$ *gegen* 0*.*

Ein Lob auf die Analysis!

Zusammenfassung

Eine Suchmaschine muss Ergebnisse nicht nur **auflisten** sondern möglichst nutzergerecht **sortieren**.

- Google analysiert Webseiten und Links als einen **Graphen**.
- Jeder Link ist ein Votum; PageRank bestimmt so die **Popularität**.
- Der Fixpunktsatz garantiert Lösung und schnelle **Berechnung**.

Genial einfach! Der erste Erfolg von Google beruhte auf

- diesen mathematischen Werkzeugen zur **Grundlegung**,
- informatischen Werkzeugen zur effizienten **Umsetzung**,
- einer geschickten Geschäftsstrategie zur **Finanzierung**.

Seither werden Modelle und Werkzeuge ständig weiterentwickelt...
Leider wuchert auch der Webspam... Es bleibt spannend.

Vielen Dank für Ihre Aufmerksamkeit!

`www.igt.uni-stuttgart.de/eiserm/popularisation/#Tag2012`

Ist das Modell PageRank plausibel?

Die Besonderheit von Hypertext sind die gegenseitigen Links.

- Die Grundannahme von PageRank ist, dass jeder Autor einer Webseite nur Links einfügt, die er für sinnvoll hält.
- Somit lesen und bewerten Millionen von Autoren gegenseitig ihre Webseiten, und ihr Urteil schlägt sich in den Links nieder.
- Das Modell der Irrfahrt berechnet aus der gegenseitigen Bewertung ein globales Maß der Popularität.

Diese Argumente gilt es zu diskutieren und empirisch zu testen...

Hauptargument für das Modell ist sein Erfolg: Die entstehende Sortierung scheint den Nutzererwartungen nahe zu kommen.

Ist das Modell PageRank deskriptiv oder normativ?

Zu Beginn sah sich Google rein deskriptiv:
Wenn eine Seite wichtig ist, dann steht sie oben auf der Liste.

Sein überwältigender Erfolg macht Google normativ:
Wenn eine Seite oben auf der Liste steht, dann ist sie wichtig.

Für kommerzielle Seiten ist die Optimierung inzwischen unerlässlich und eine Wissenschaft für sich (*search engine optimization*, SEO).

Offensichtliche Strategie: Viele Links anlocken, am besten von anderen wichtigen Seiten, und selbst nur gut gewählte Links setzen.

Somit verändert die Allgegenwart von Google das Verhalten der Autoren . . . und damit die Grundannahme des Modells PageRank!

So gesehen befinden sich Nutzer, Autoren und Suchmaschinen in einer komplizierten Evolution. Auch dies bleibt spannend!