WWW — Wichtig, Wichtiger, am Wichtigsten

10. Algorithmus der Woche (PageRank)
Informatikjahr 2006

Ulrik Brandes Gabi Dorfmüller

Fachbereich Informatik & Informationswissenschaft
Universität Konstanz

9. Mai 2006

Suchmaschine Google

- Typische Anfrage an **Suchmaschine Google** (z.B. für ein Referat):

Suchmaschine Google

- Typische Anfrage an **Suchmaschine Google** (z.B. für ein Referat):

- Viele Treffer, aber meist die guten zuerst.

Suchmaschine Google

- Typische Anfrage an **Suchmaschine Google**
  (z.B. für ein Referat):

- Viele Treffer, aber meist die guten zuerst.
- **Wie geht das?**
  Wer entscheidet, welche Quellen ganz oben erscheinen?

PageRank

- **World Wide Web (WWW):**
  Netzwerk aus Milliarden von Dokumenten,
  vor allem *Web-Seiten*,
  die durch *Links* (Verweise) miteinander verknüpft sind.

PageRank

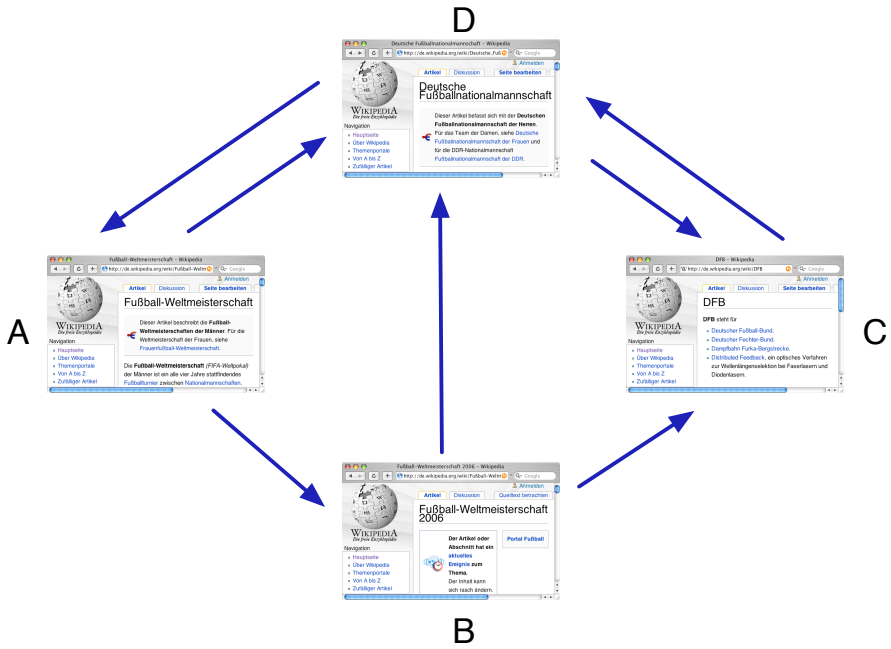
- **World Wide Web (WWW):**
  Netzwerk aus Milliarden von Dokumenten,
  vor allem *Web-Seiten*,
  die durch *Links* (Verweise) miteinander verknüpft sind.
- **Reihung der Ergebnisse:**
  Spezieller Algorithmus,
  der Dokumenten eine Relevanz (Wichtigkeit) zuordnet.

PageRank

- **World Wide Web (WWW):**
  Netzwerk aus Milliarden von Dokumenten,
  vor allem *Web-Seiten*,
  die durch *Links* (Verweise) miteinander verknüpft sind.
- **Reihung der Ergebnisse:**
  Spezieller Algorithmus,
  der Dokumenten eine Relevanz (Wichtigkeit) zuordnet.
- **PageRank:**
  Zentraler Teil des von Google verwendeten Algorithmus',
  beruht auf Auswertung der Links.

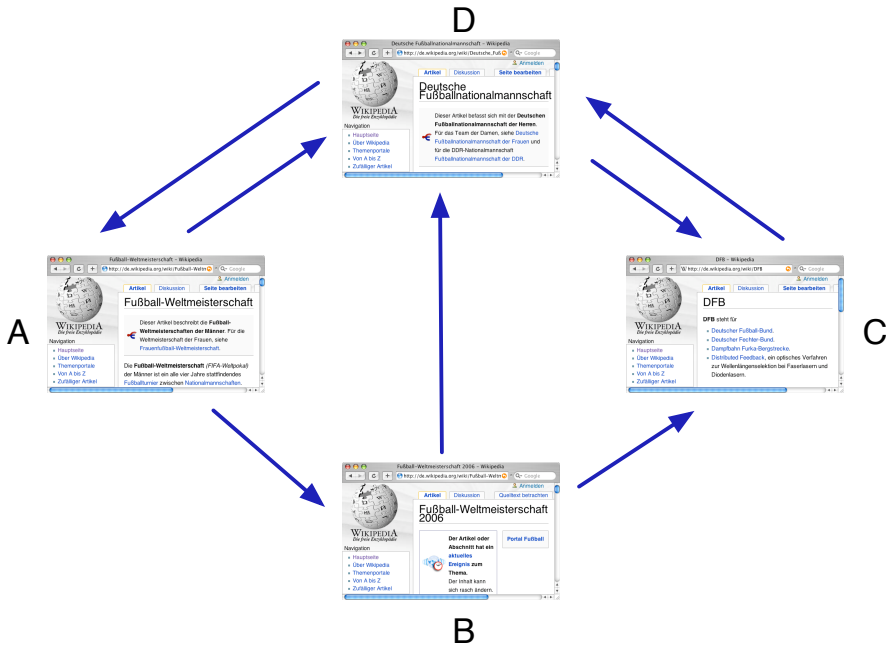
Internet-Surfen

- **Beispielnetzwerk:**



Internet-Surfen

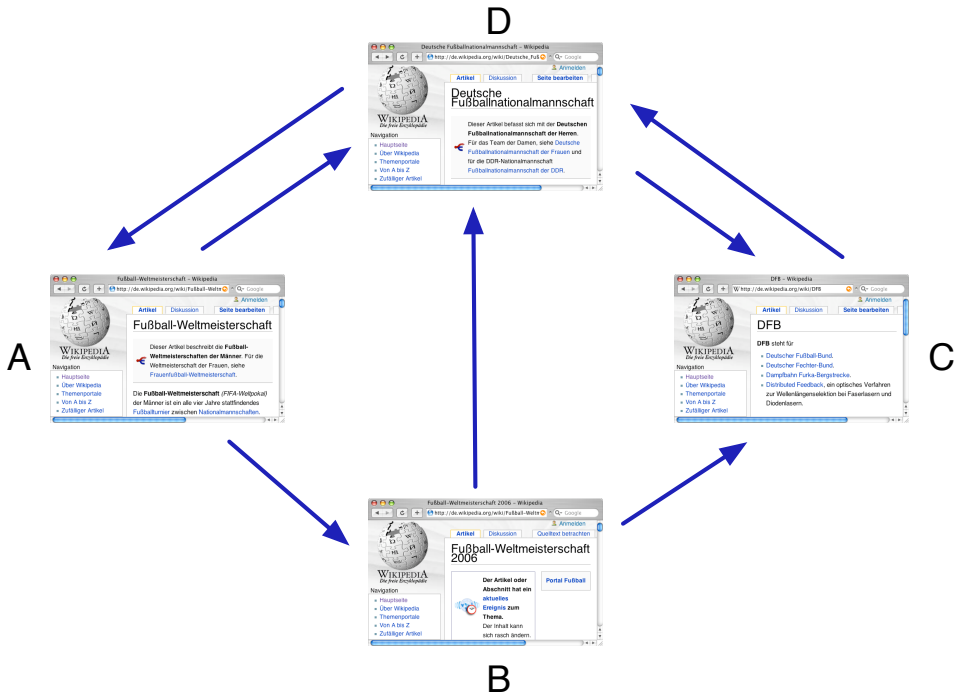
- **Beispielnetzwerk:**



- **Surfen im Internet:**
  - Start auf irgendeiner Seite
  - Klicken auf einen der in der Seite enthaltenen Links, um zur jeweils nächsten Seite zu gelangen

Experiment

Jede Person sucht sich im Beispielnetzwerk eine beliebige Seite aus. Von da bewegen sich alle durch das Netzwerk, dürfen dabei aber immer nur den Links folgen.



Nach einer halben Minute stoppen alle auf Kommando und merken sich die zuletzt besuchte Seite. Wie viele Personen sind auf welcher Seite stehen geblieben?

Erreichbarkeit

- Haben die Wenigsten auf Seite B angehalten?

Erreichbarkeit

- Haben die Wenigsten auf Seite B angehalten?
  Beobachtung: Verlinkung bestimmt Erreichbarkeit.

Erreichbarkeit

- Haben die Wenigsten auf Seite B angehalten?
  Beobachtung: Verlinkung bestimmt Erreichbarkeit.
- Erreichbarkeit einer Seite wird mit Wichtigkeit im WWW gleich gesetzt.

Erreichbarkeit

- Haben die Wenigsten auf Seite B angehalten?
  Beobachtung: Verlinkung bestimmt Erreichbarkeit.
- Erreichbarkeit einer Seite wird mit Wichtigkeit im WWW gleich gesetzt.
- Aber: Wie kommt man an solche Werte, ohne eine riesige Anzahl von Surfern loszugeschicken?

Berechnung

- **Annahme:** Bei mehreren möglichen Links wird von diesen einer zufällig gewählt.
- **Überlegung:** Wer auf Seite D ist, geht in einem von zwei Fällen zu A weiter. Erhalten daraus ein **Gleichungssystem** für die Besuchshäufigkeiten der Web-Seiten:

$$\begin{aligned} a &= \frac{1}{2} \cdot d & \qquad c &= \frac{1}{2} \cdot b + \frac{1}{2} \cdot d \\ b &= \frac{1}{2} \cdot a & \qquad d &= \frac{1}{2} \cdot a + \frac{1}{2} \cdot b + c \end{aligned}$$

Berechnung

- **Annahme:** Bei mehreren möglichen Links wird von diesen einer zufällig gewählt.
- **Überlegung:** Wer auf Seite D ist, geht in einem von zwei Fällen zu A weiter. Erhalten daraus ein **Gleichungssystem** für die Besuchshäufigkeiten der Web-Seiten:

$$\begin{aligned} a &= \frac{1}{2} \cdot d & \qquad c &= \frac{1}{2} \cdot b + \frac{1}{2} \cdot d \\ b &= \frac{1}{2} \cdot a & \qquad d &= \frac{1}{2} \cdot a + \frac{1}{2} \cdot b + c \end{aligned}$$

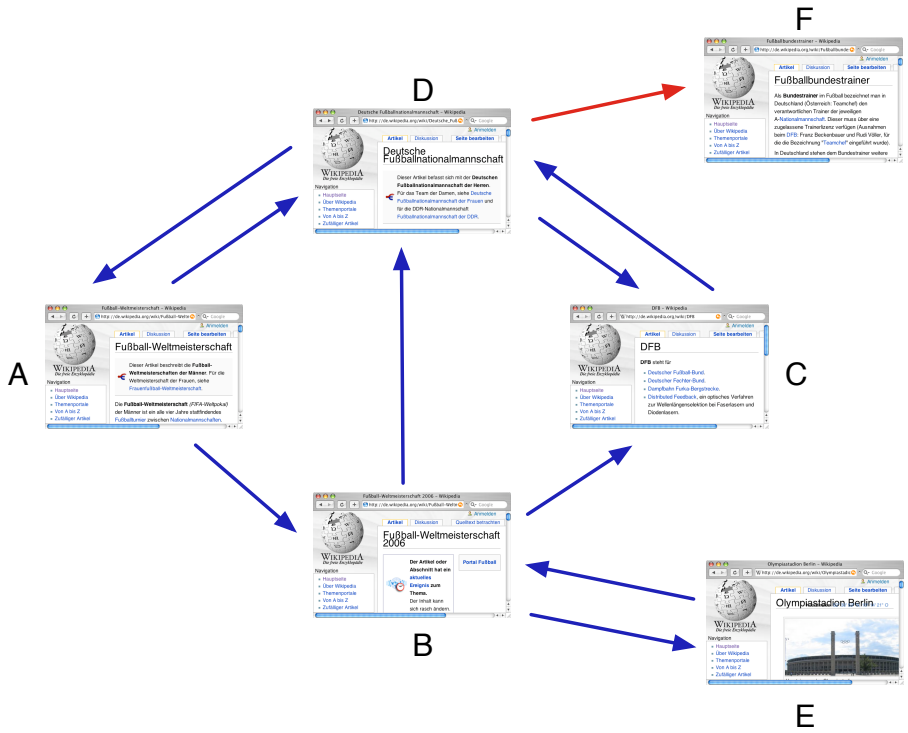
- Eine mögliche **Lösung** ist:

$$a = 4 \qquad b = 2 \qquad c = 5 \qquad d = 8$$

Jede andere Lösung ergibt sich durch Multiplikation aller vier Werte mit demselben Faktor.

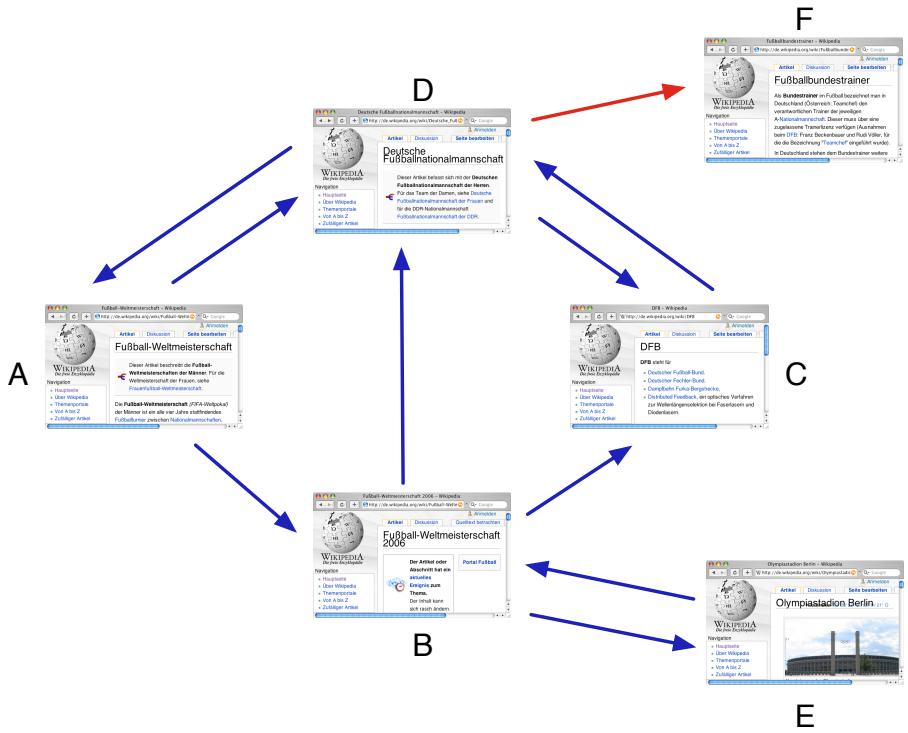
Sackgassen

- **Problematisches Beispiel:**



Sackgassen

- **Problematisches Beispiel:**



- **Sackgasse:** Es gibt Seiten, zu denen man irgendwann nicht mehr zurückkehren kann (Folge dem roten Link zu Seite *F*).
- Sackgassen führen zu Lösungen, die für die Sortierung ungeeignet sind. (Warum?)

Verbesserung

- **Reales Surf-Verhalten:** Wer keinen interessanten Link findet, ruft irgend eine andere Seite auf.
  (z.B. durch „Zurück“-Knopf, Favoriten oder direkte Eingabe der Adresse)

Verbesserung

- **Reales Surf-Verhalten:** Wer keinen interessanten Link findet, ruft irgend eine andere Seite auf.
  (z.B. durch „Zurück“-Knopf, Favoriten oder direkte Eingabe der Adresse)
- **Annahme:** In einem von fünf Fällen wird eine Seite nicht über einen Link erreicht, sondern direkt angesprungen. Dabei wird keine Seite bevorzugt.

Verbesserung

- **Reales Surf-Verhalten:** Wer keinen interessanten Link findet, ruft irgend eine andere Seite auf.
  (z.B. durch „Zurück“-Knopf, Favoriten oder direkte Eingabe der Adresse)
- **Annahme:** In einem von fünf Fällen wird eine Seite nicht über einen Link erreicht, sondern direkt angesprungen. Dabei wird keine Seite bevorzugt.
- Erhalten nun folgendes **Gleichungssystem**:

$$
\begin{aligned}
a &= \frac{4}{5}\cdot\left(\frac{1}{3}\cdot d\right) + \frac{1}{5}\cdot\frac{1}{6} & \qquad d &= \frac{4}{5}\cdot\left(\frac{1}{2}\cdot a + \frac{1}{3}\cdot b + c\right) + \frac{1}{5}\cdot\frac{1}{6}\\
b &= \frac{4}{5}\cdot\left(\frac{1}{2}\cdot a + e\right) + \frac{1}{5}\cdot\frac{1}{6} & \qquad e &= \frac{4}{5}\cdot\left(\frac{1}{3}\cdot b\right) + \frac{1}{5}\cdot\frac{1}{6}\\
c &= \frac{4}{5}\cdot\left(\frac{1}{3}\cdot b + \frac{1}{3}\cdot d\right) + \frac{1}{5}\cdot\frac{1}{6} & \qquad f &= \frac{4}{5}\cdot\left(\frac{1}{3}\cdot d\right) + \frac{1}{5}\cdot\frac{1}{6}
\end{aligned}
$$

Lösen des Gleichungssystems

- **WWW:** Gleichungssystem mit Millarden von Unbekannten und Gleichungen
  $\Rightarrow$ Auflösen und Einsetzen schaffen auch Computer nicht.

Lösen des Gleichungssystems

- **WWW:** Gleichungssystem mit Millarden von Unbekannten und Gleichungen
  $\Rightarrow$ Auflösen und Einsetzen schaffen auch Computer nicht.
- **Einfacher Algorithmus**
  zur schnellen Berechnung eines ausreichend guten Ergebnisses:

Lösen des Gleichungssystems

- **WWW:** Gleichungssystem mit Millarden von Unbekannten und Gleichungen
  $\Rightarrow$ Auflösen und Einsetzen schaffen auch Computer nicht.
- **Einfacher Algorithmus**
  zur schnellen Berechnung eines ausreichend guten Ergebnisses:
  - Beginn mit beliebigen Werten für eine Lösung (z.B. 1)

Lösen des Gleichungssystems

- **WWW:** Gleichungssystem mit Millarden von Unbekannten und Gleichungen
  $\Rightarrow$ Auflösen und Einsetzen schaffen auch Computer nicht.
- **Einfacher Algorithmus**
  zur schnellen Berechnung eines ausreichend guten Ergebnisses:
  - Beginn mit beliebigen Werten für eine Lösung (z.B. 1)
  - Für jede Unbekannte:
    Berechnung, was ihr richtiger Wert wäre, wenn alle anderen schon stimmten

Lösen des Gleichungssystems

- **WWW:** Gleichungssystem mit Millarden von Unbekannten und Gleichungen
  $\Rightarrow$ Auflösen und Einsetzen schaffen auch Computer nicht.
- **Einfacher Algorithmus**
  zur schnellen Berechnung eines ausreichend guten Ergebnisses:
  - Beginn mit beliebigen Werten für eine Lösung (z.B. 1)
  - Für jede Unbekannte:
    Berechnung, was ihr richtiger Wert wäre, wenn alle anderen schon stimmten
  - Wiederholung des letzten Schrittes mit den erhaltenen neuen Werten

Lösen des Gleichungssystems

- **WWW:** Gleichungssystem mit Millarden von Unbekannten und Gleichungen
  $\Rightarrow$ Auflösen und Einsetzen schaffen auch Computer nicht.
- **Einfacher Algorithmus**
  zur schnellen Berechnung eines ausreichend guten Ergebnisses:
  - Beginn mit beliebigen Werten für eine Lösung (z.B. 1)
  - Für jede Unbekannte:
    Berechnung, was ihr richtiger Wert wäre, wenn alle anderen schon stimmten
  - Wiederholung des letzten Schrittes mit den erhaltenen neuen Werten
  - Dasselbe nochmal;
    und nochmal;
    und so weiter und so fort.

Näherung

- Mit jedem Schritt werden die Werte ein bißchen besser.
  Wenn sich kaum noch was ändert, ist das ein Zeichen dafür,
  dass man schon nahe an der richtigen Lösung ist.

Näherung

- Mit jedem Schritt werden die Werte ein bißchen besser.
  Wenn sich kaum noch was ändert, ist das ein Zeichen dafür,
  dass man schon nahe an der richtigen Lösung ist.
- Für das letzte Beispiel ergibt sich:

	1. Schritt	2. Schritt	3. Schritt	...	21. Schritt	22. Schritt	...	Lösung
a	1.00000	0.30000	0.43333	...	0.08488	0.08478	...	0.08454
b	1.00000	1.23333	0.39333	...	0.11962	0.11950	...	0.11926
c	1.00000	0.56667	0.76222	...	0.11681	0.11668	...	0.11634
d	1.00000	1.50000	0.93556	...	0.19292	0.19292	...	0.19203
e	1.00000	0.30000	0.36222	...	0.06527	0.06523	...	0.06514
f	1.00000	0.30000	0.43333	...	0.08488	0.08478	...	0.08454

Test

Frage:
Wenn ich meine eigene *Homepage* von all meinen Freunden verlinken lasse, erscheint sie dann bei Google ganz oben?

Test

Frage:
Wenn ich meine eigene *Homepage* von all meinen Freunden verlinken lasse, erscheint sie dann bei Google ganz oben?
Antwort:
Das funktioniert nur, wenn die Seiten meiner Freunde selbst große Wichtigkeit im WWW haben — also eher nein.

Ausblick

- Es gibt viele Möglichkeiten, Wichtigkeit in einem Netzwerk zu definieren.

Ausblick

- Es gibt viele Möglichkeiten, Wichtigkeit in einem Netzwerk zu definieren.
- Es gibt viele Möglichkeiten, die Wichtigkeit beim Reihen der Treffer zu berücksichtigen.

Ausblick

- Es gibt viele Möglichkeiten, Wichtigkeit in einem Netzwerk zu definieren.
- Es gibt viele Möglichkeiten, die Wichtigkeit beim Reihen der Treffer zu berücksichtigen.
- Die Häufigkeit der Sprünge bei PageRank kann unterschiedlich festgelegt werden.

Ausblick

- Es gibt viele Möglichkeiten, Wichtigkeit in einem Netzwerk zu definieren.
- Es gibt viele Möglichkeiten, die Wichtigkeit beim Reihen der Treffer zu berücksichtigen.
- Die Häufigkeit der Sprünge bei PageRank kann unterschiedlich festgelegt werden.
- Die Auswahl der direkt angesprungenen Seiten kann beeinflusst werden.

Ausblick

- Es gibt viele Möglichkeiten, Wichtigkeit in einem Netzwerk zu definieren.
- Es gibt viele Möglichkeiten, die Wichtigkeit beim Reihen der Treffer zu berücksichtigen.
- Die Häufigkeit der Sprünge bei PageRank kann unterschiedlich festgelegt werden.
- Die Auswahl der direkt angesprungenen Seiten kann beeinflusst werden.
- Suchmaschinen benutzen noch viele, viele andere Algorithmen (einige waren oder sind noch Algorithmen der Woche, z.B. binäre Suche, String-Alignment)

Autoren und externe Links

Autoren:

- Prof. Dr. Ulrik Brandes:
  `http://www.inf.uni-konstanz.de/~brandes/`
- Dipl.-Math. Gabi Dorfmüller

Externe Links:

- PageRank-Eintrag der Wikipedia:
  `http://de.wikipedia.org/wiki/Pagerank`
- visone – Ein Programm mit dem man Netzwerke eingeben und z.B. PageRank berechnen kann:
  `http://www.visone.info/`