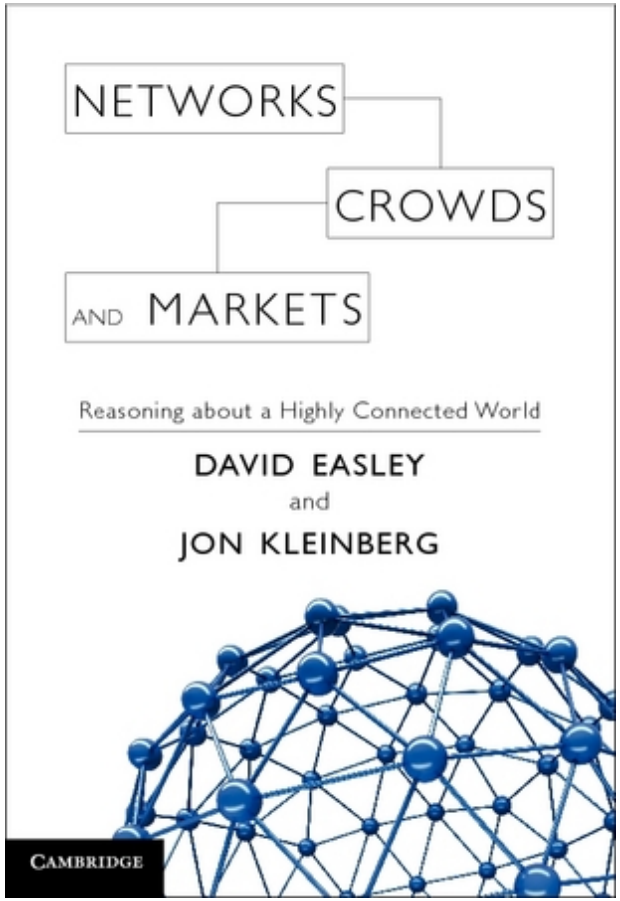
Link Analysis and Web Search

Jan Benedikt Führer

TECHNISCHE UNIVERSITÄT DARMSTADT



Gliederung

TECHNISCHE UNIVERSITÄT DARMSTADT

Motivation

Link-Analyse mit Hubs und Authorities

PageRank

Anwendung innerhalb des WWW

Anwendungen außerhalb des Internets

Gliederung

TECHNISCHE UNIVERSITÄT DARMSTADT

Motivation

Link-Analyse mit Hubs und Authorities

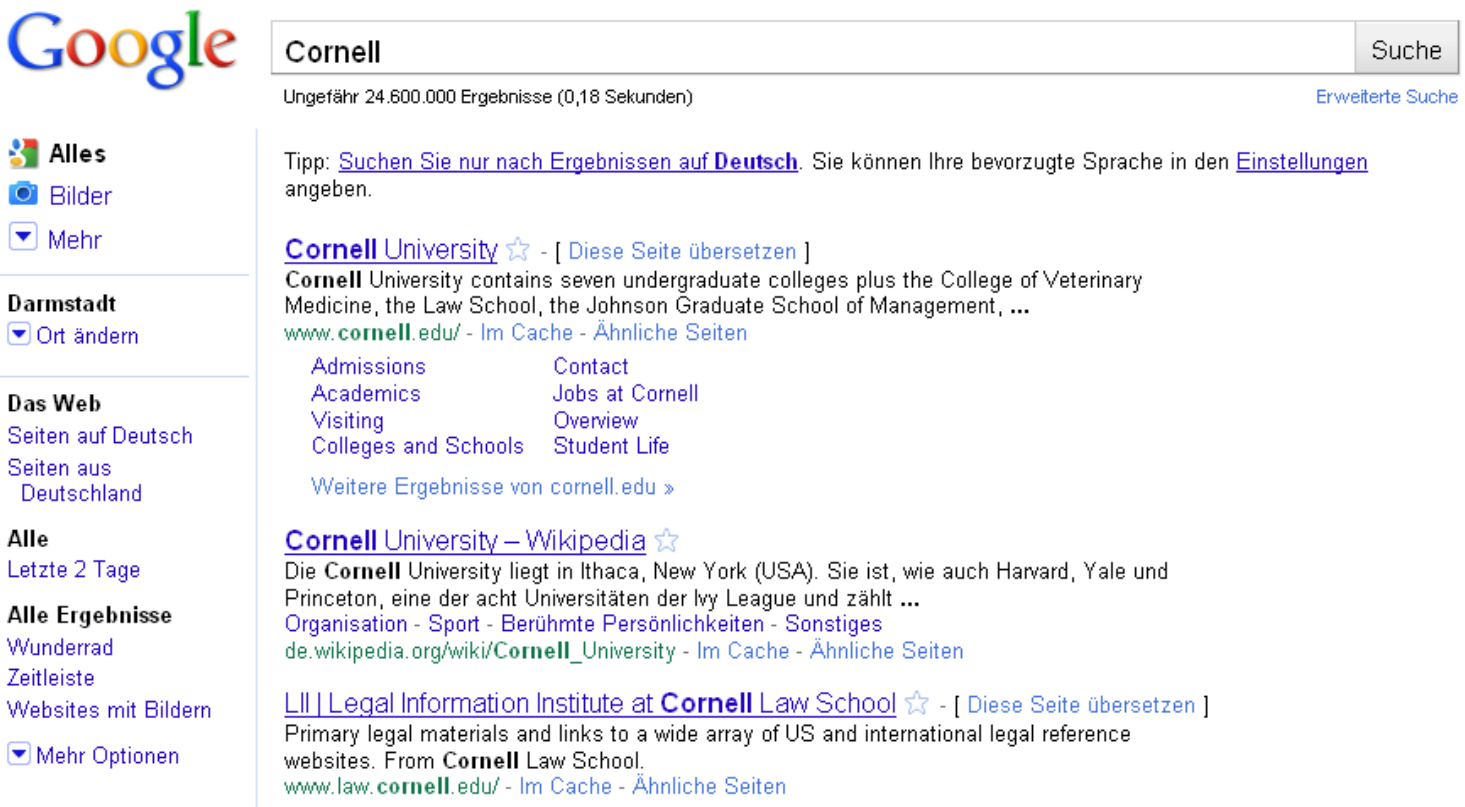
PageRank

Anwendung innerhalb des WWW

Anwendungen außerhalb des Internets

Motivation

TECHNISCHE UNIVERSITÄT DARMSTADT



- Warum?

Das Ranking-Problem

TECHNISCHE UNIVERSITÄT DARMSTADT

- Suchen nach Inhalten schwieriges Problem für Computer
- Fachgebiet des Information Retrieval seit 1950/60

- Problem: Schlüsselwörter $\neq$ komplexe Information

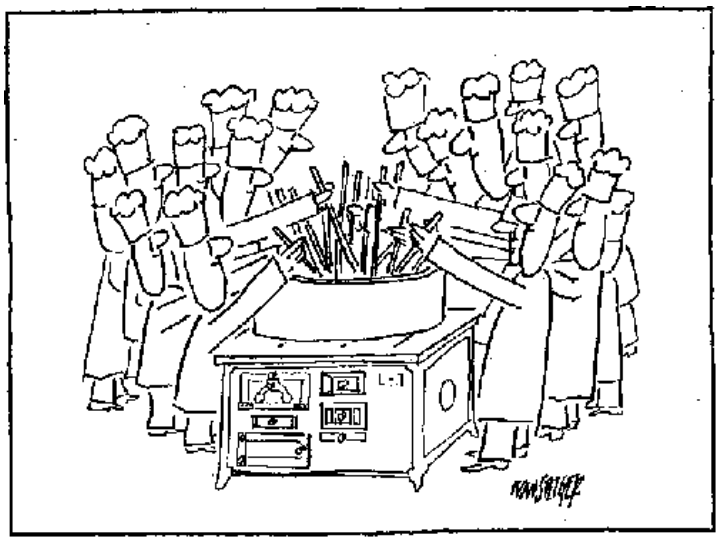
$\Rightarrow$ Synonymie: Z.B.: Sahne, Rahm und Schlagobers meinen das gleiche

$\Rightarrow$ Mehrdeutigkeit: Z.B.: Jaguar = Tier/Auto/Mac OS/...

Information Retrieval & Das WWW

TECHNISCHE UNIVERSITÄT DARMSTADT

- Bis 1980: Dokumente von Experten für Experten (Bibliothekare, Anwälte, etc.)
- Festgelegter Stil und Vokabular $\Rightarrow$ Effektive Abfragen
- Web: Exponentielle Zunahme der Benutzer und Benutzertypen
- Der Mangel wird zum Überfluss an Information
- Filtern wird zur Hauptaufgabe



Das WWW ist dynamisch

TECHNISCHE UNIVERSITÄT DARMSTADT

- 9/11: „World Trade Center“

⟺

World Trade Center – Wikipedia ☆
Das **World Trade Center** [wɜːldˈtɹeɪdˌsɛntə] (deutsch Welthandelszentrum), abgekürzt **WTC**, war ein Gebäudekomplex aus sieben Gebäuden in New York City. **...**
de.wikipedia.org/wiki/**World_Trade_Center** - Im Cache - Ähnliche Seiten

One World Trade Center
7 World Trade Center (alt)

Weitere Ergebnisse von wikipedia.org »

|| **World Trade Center** || ☆ - [Diese Seite übersetzen]
World trade center website for information on lower Manhattan and leasing at the **world trade center**.
www.**wtc**.com/ - Im Cache - Ähnliche Seiten

World Trade Center Bremen ☆
Die BIG-Gruppe ist Ihr Partner in allen Fragen zur Landesentwicklung und Wirtschaftsförderung in Bremen. Wirtschaftsförderung aus einer Hand - von A bis Z!
www.**wtc**-bremen.de/ - Im Cache - Ähnliche Seiten

World Trade Center Dresden :: Home ☆
Das Haus informiert über Veranstaltungen, Geschäfte, Restaurants, Kultur und Mieter.
www.**wtc**-dresden.de/ - Im Cache - Ähnliche Seiten

Gliederung

TECHNISCHE UNIVERSITÄT DARMSTADT

Motivation

Link-Analyse mit Hubs und Authorities

PageRank

Anwendung innerhalb des WWW

Anwendungen außerhalb des Internets

Voting anhand von In-Links

TECHNISCHE UNIVERSITÄT DARMSTADT

- Ergebnis basiert nicht nur auf seiteninternen Features
- Verweise kodieren die Relevanz (= Authority) einer Website zu einem Thema
- Bsp.: www.cornell.edu wird von vielen relevanten Websites referenziert

Problem

- Unterschiedliche Bedeutungen (Off-Topic, Kritik, bezahlte Werbung, usw.)
- Annahme: Viele In-Links $\widehat{=}$ Hohe Authority

Ein einfacher Algorithmus

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Sammle eine große Stichprobe an relevanten Ergebnissen (textbasiert)
2. Lasse die Ergebnisse über ihre Linkstruktur „wählen“

⇒ Funktioniert gut bei eindeutigen Suchanfragen

⇒ Problematisch bei Mehrdeutigkeiten

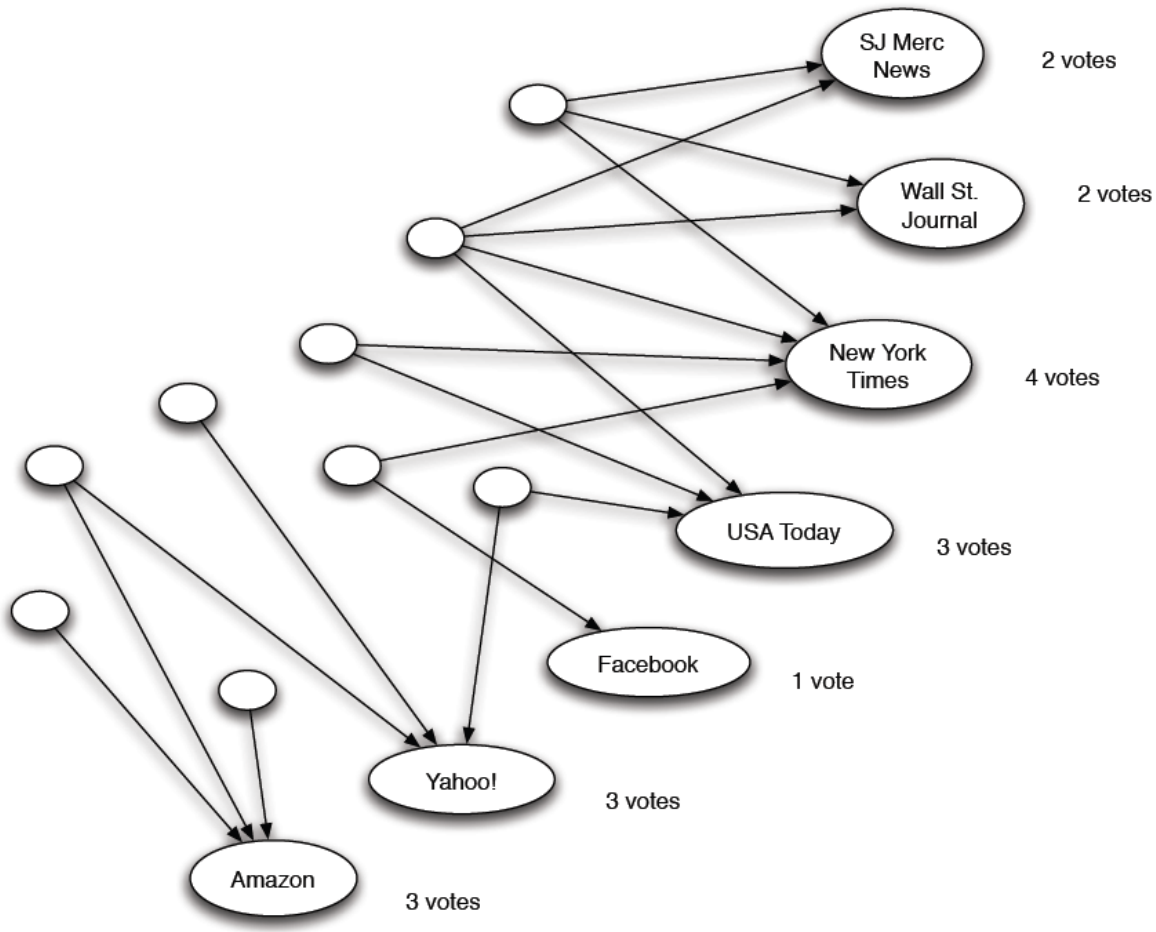
List-Finding

TECHNISCHE UNIVERSITÄT DARMSTADT

- Was passiert bei Abfragen wie „Zeitung“?
- Keine eindeutige „beste“ Antwort
- Stattdessen Auflistung der bekanntesten Zeitungen
- Ergebnis des Voting-Algorithmus:
  - ☺ Websites prominenter Zeitungen
  - ☹ Websites mit unabhängig hohen In-Links (Yahoo, Facebook, Amazon)

List-Finding

TECHNISCHE UNIVERSITÄT DARMSTADT



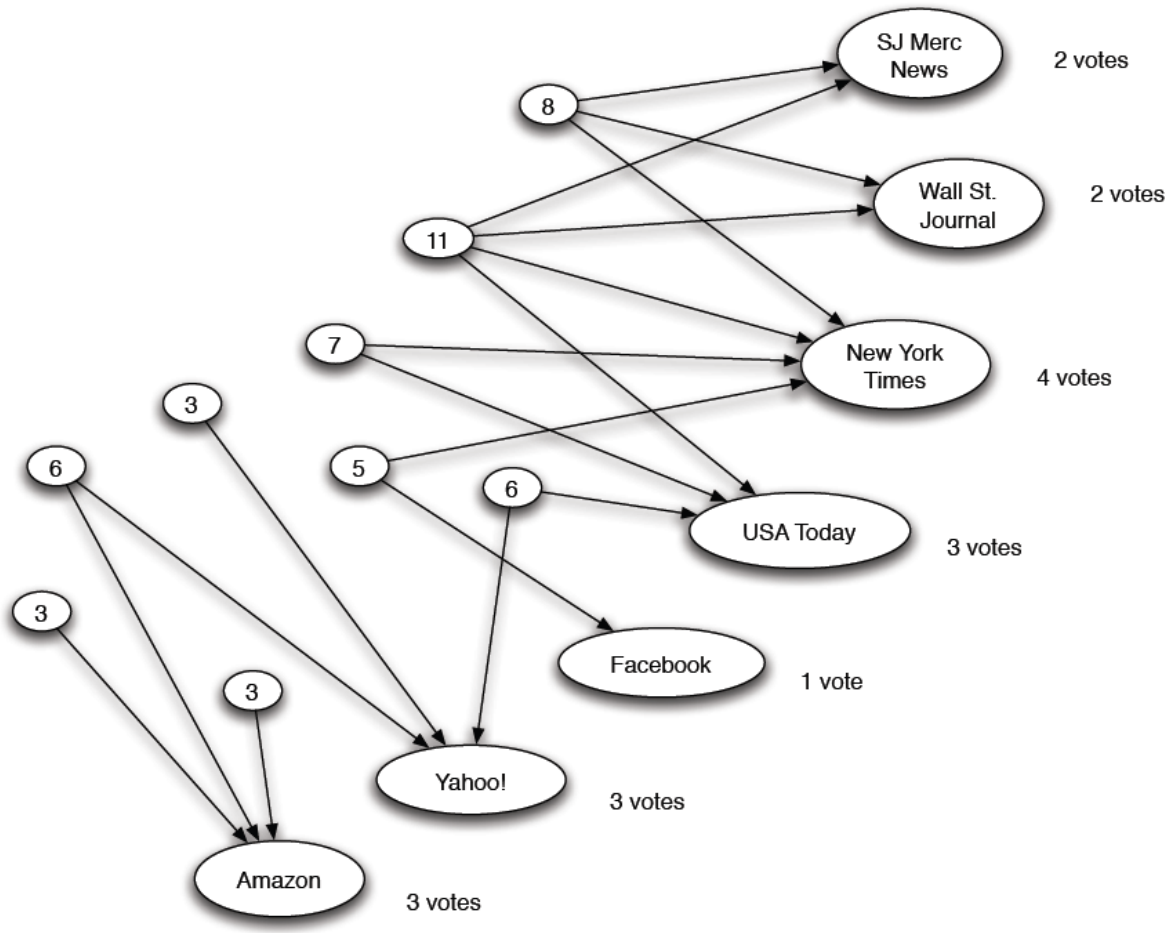
List-Finding

TECHNISCHE UNIVERSITÄT DARMSTADT

- Referenzen bilden ein zu simples Maß
- Weitere Informationen in Link-Struktur enthalten
- Linkverzeichnisse sammeln Ergebnisse mit hoher Relevanz
- Listenwert(s) = Summe der Stimmen aller Websites für die s gestimmt hat

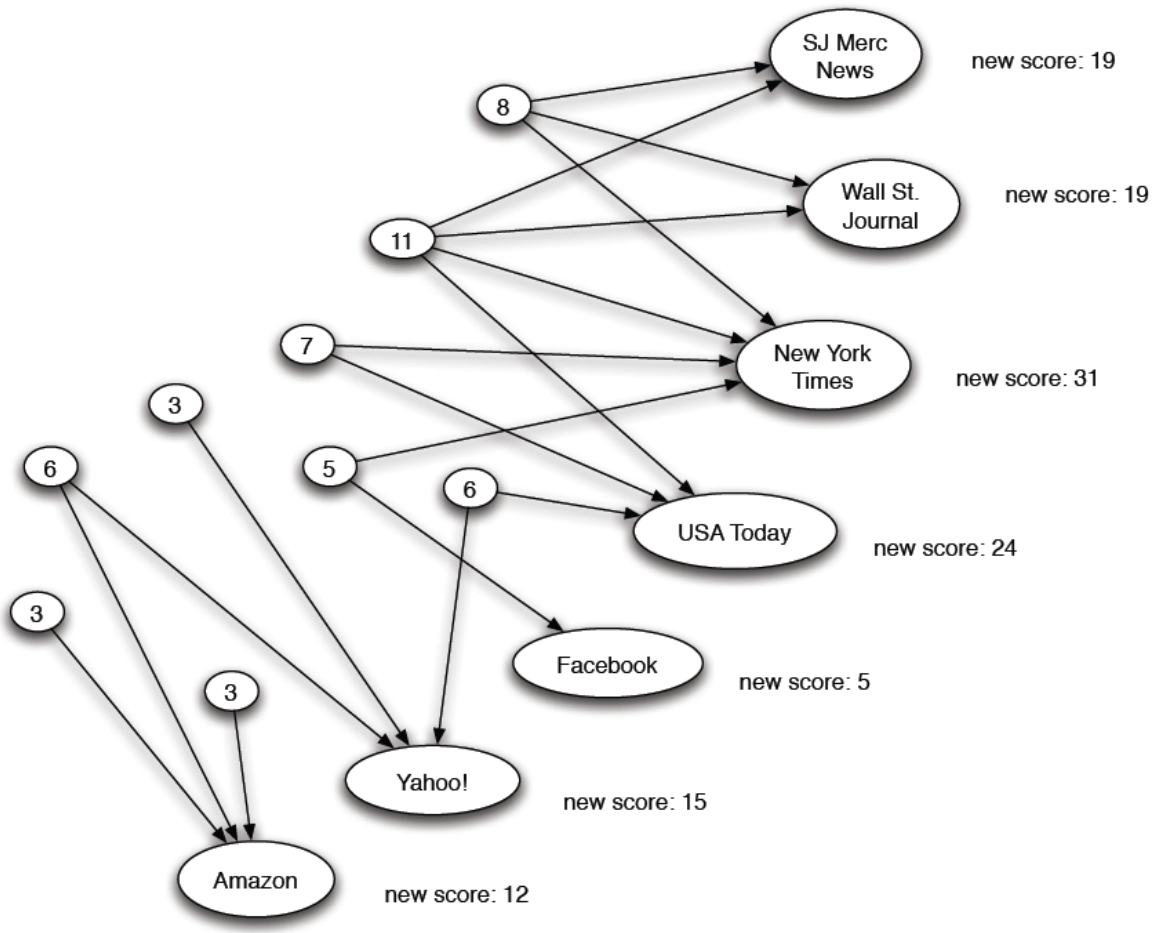
List-Finding

TECHNISCHE UNIVERSITÄT DARMSTADT



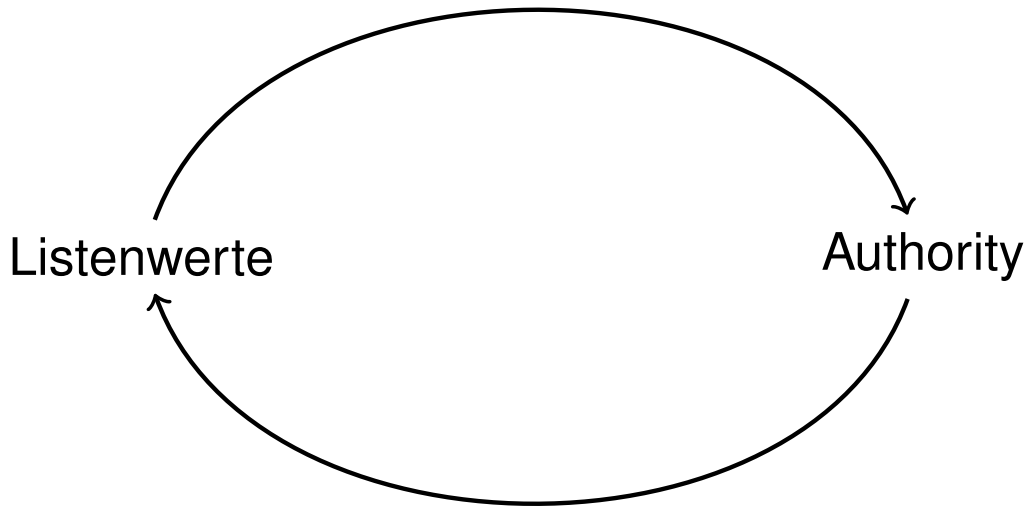
The Principle of Repeated Improvement

TECHNISCHE UNIVERSITÄT DARMSTADT



The Principle of Repeated Improvement

TECHNISCHE UNIVERSITÄT DARMSTADT

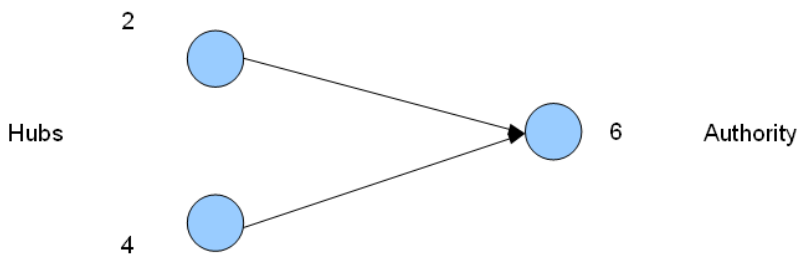


Hubs & Authorities

TECHNISCHE UNIVERSITÄT DARMSTADT

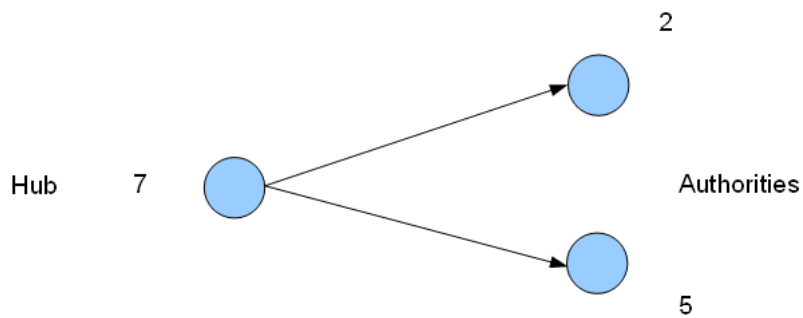
- *Authorities:* Häufig referenzierte Websites
- *Hubs:* Websites mit hohem Listenwert
- Berechne auth(p) und hub(p) für jede Website p

1. Schritt: Voting



Authority Update Rule

2. Schritt: List-Finding



Hub Update Rule

HITS-Algorithmus

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Initialisiere *auth*(*p*) und *hub*(*p*) für jede Website p mit 1
2. Wähle Anzahl von Schritten *k*
3. Sequenz von *k* Authority- & Hub-Updates
   - Authority Update Rule
   - Hub Update Rule
4. Normalisierung

$\Rightarrow$ Werte konvergieren für $k \rightarrow \infty$ (Equilibrium)

$\Rightarrow$ Unabhängig von Initialisierung

Gliederung

TECHNISCHE UNIVERSITÄT DARMSTADT

Motivation

Link-Analyse mit Hubs und Authorities

PageRank

Anwendung innerhalb des WWW

Anwendungen außerhalb des Internets

Page-Rank

TECHNISCHE UNIVERSITÄT DARMSTADT

- Erinnerung: Websites spielen verschiedene Rollen $\Rightarrow$ Authorities & Hubs
- Direktes Referenzieren dominant in nicht-kommerziellen Abfragen
- Bsp: Wissenschaftliche Literatur, Universitäten, Regierungen, etc.
- Voting & Principle of Repeated Improvement
- Knoten verschicken kontinuierlich „Authority“ über das Netzwerk
- Gewichtung $\widehat{=}$ PageRank
- Intuitiv: Flüssigkeit fließt durch das Netzwerk

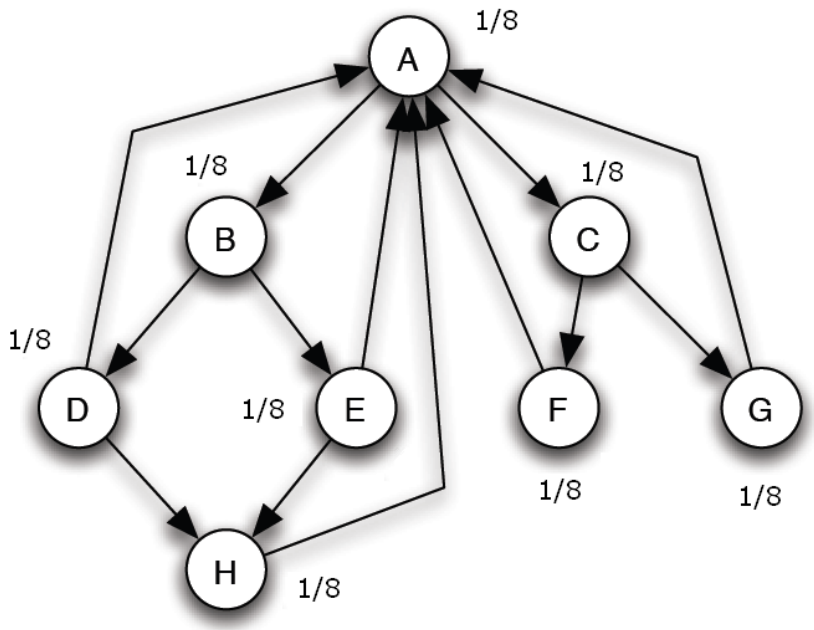
Algorithmus

TECHNISCHE UNIVERSITÄT DARMSTADT

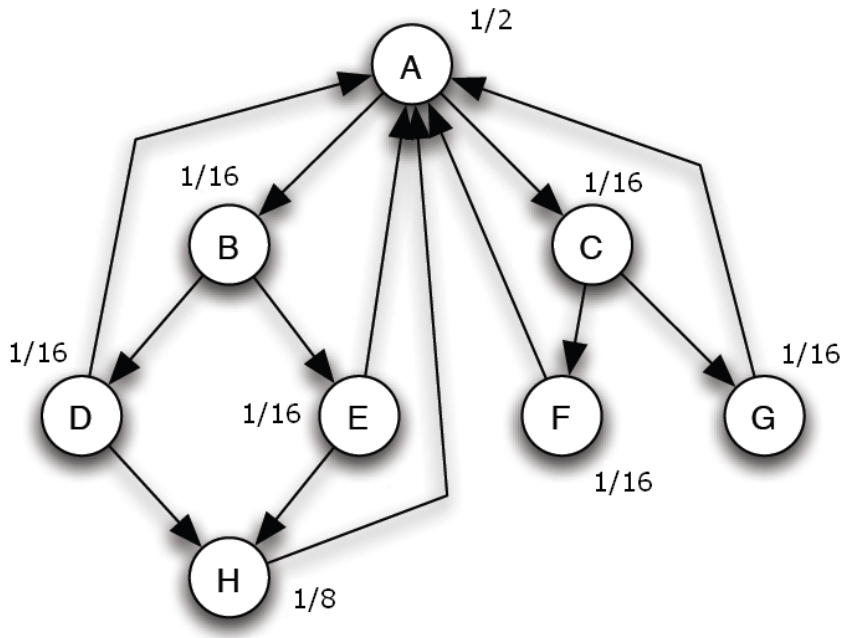
1. Initialisiere alle n Knoten mit $1/n$
2. Wähle eine Anzahl von Schritten k
3. Sequenz von k Updates gemäß der Basic Page Rank Update Rule:
    4. Verschicke PageRank gleichmäßig über alle ausgehenden Links.
    5. Aktualisiere PageRank auf die Summe der Werte aller eingehenden Links.

Beispiel

TECHNISCHE UNIVERSITÄT DARMSTADT



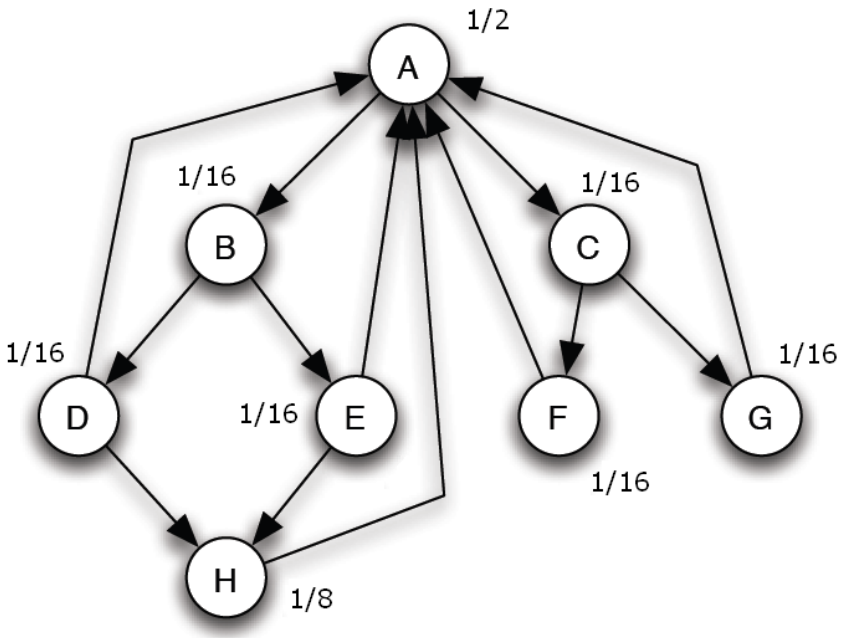
Initialisierung



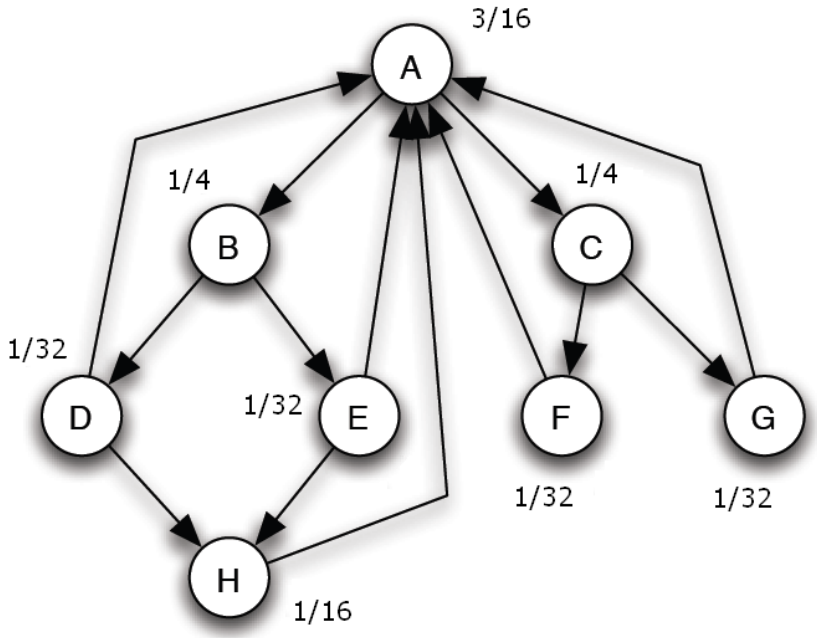
1. Schritt

Beispiel

TECHNISCHE UNIVERSITÄT DARMSTADT



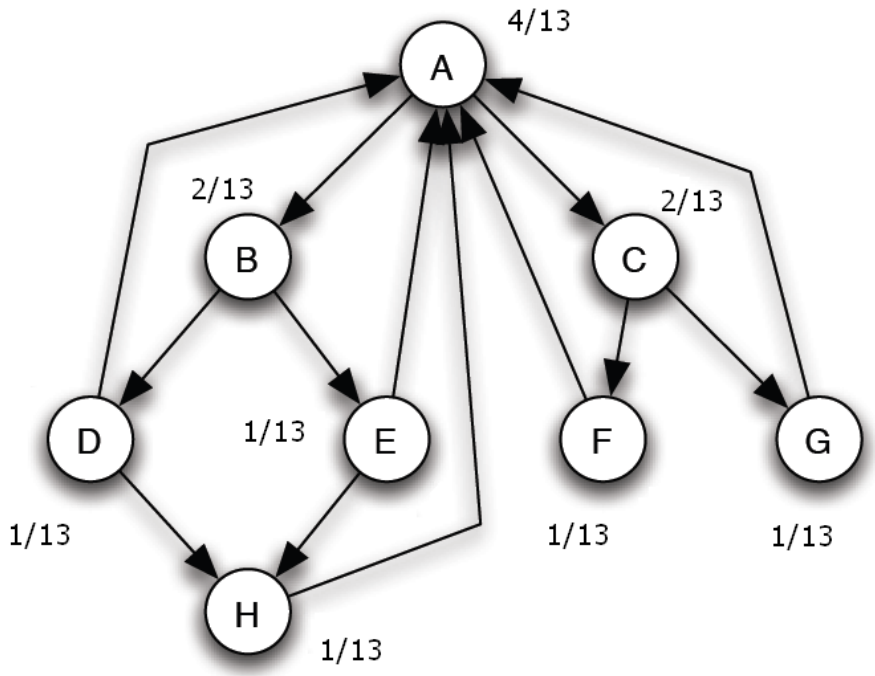
1. Schritt



2. Schritt

Equilibrium

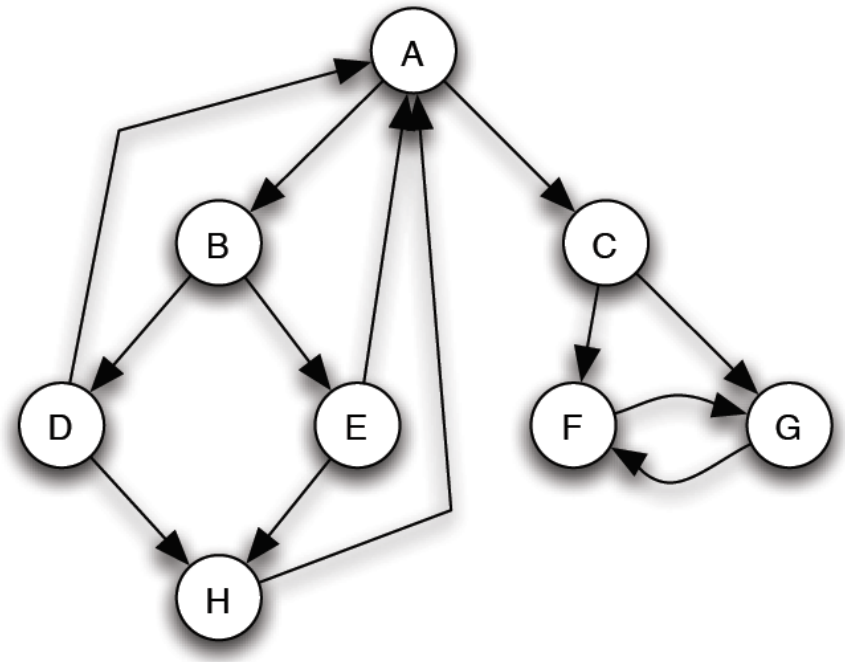
TECHNISCHE UNIVERSITÄT DARMSTADT



- PageRanks konvergieren für $k \to \infty$
- Eindeutige Grenzwerte bei stark verbundenen Netzwerken

Problem: Leaks

TECHNISCHE UNIVERSITÄT DARMSTADT



- Erinnerung: PageRank ~Flüssigkeit
- Ausgleichsprozess: Verdunsten von Wasser & Regen

Scaled Page Rank Update Rule

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Basic PageRank Update Rule
2. Skaliere alle PageRanks mit Scaling Factor $s \in (0, 1)$
3. Verteile die verbleibenden $1 - s$ Einheiten gleichmäßig über alle Knoten des Netzwerks

- Erhält totalen PageRank des Netzwerks
- Grenzwerte eindeutig aber von s abhängig!
- Unempfindlicher für Strukturänderungen

Random Walks

TECHNISCHE UNIVERSITÄT DARMSTADT

- Äquivalente Definition
- Person surfed zufällig auf Websites innerhalb eines Netzwerks

Behauptung

Die Wahrscheinlichkeit sich nach *k* Schritten auf Website X zu befinden, entspricht dem PageRank von X nach *k* Applikationen der Basic PageRank Update Rule.

Gliederung

TECHNISCHE UNIVERSITÄT DARMSTADT

Motivation

Link-Analyse mit Hubs und Authorities

PageRank

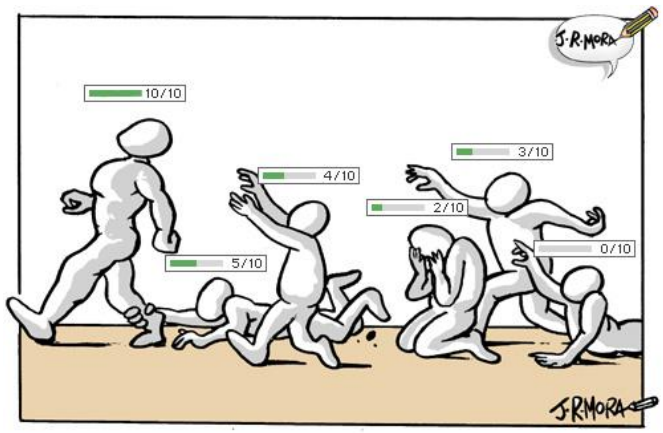
Anwendung innerhalb des WWW

Anwendungen außerhalb des Internets

Anwendung innerhalb des WWW

TECHNISCHE UNIVERSITÄT DARMSTADT

- Integrale Rolle in Ranking-Funktionen moderner Suchmaschinen
- Z.B.: Google, Yahoo, Bing, Ask, usw.
- Stetige Weiterentwicklung bei strenger Geheimhaltung



Links, Text und Benutzerdaten

TECHNISCHE UNIVERSITÄT DARMSTADT

- In der Praxis: Linkstruktur **und** Text $\Rightarrow$ *Anchor Text*
- Bsp.: „Ich bin Student an der Cornell University“
- Bisherige Methoden leicht zu erweitern

$\Rightarrow$ Gewichtung entsprechend der Güte des Anchor Texts

- Evtl. weitere Features: Benutzerdaten

Ein bewegliches Ziel

TECHNISCHE UNIVERSITÄT DARMSTADT

- Erinnerung Spieltheorie: Die Welt reagiert auf den Benutzer
- Mit der Entwicklung der Websuche führt zu neuen Geschäftsmodellen
- Update der Ranking-Funktionen ~Hurricane ?
- Erstellung von Websites mit dem Ziel hohe Rankings zu erreichen

„Web search is a new kind of information retrieval application in that the documents are actively behaving badly.“

Ein bewegliches Ziel

TECHNISCHE UNIVERSITÄT DARMSTADT

- Entstehung der SEO (Search Engine Optimization)
- Konsequenzen:
  1. Die „perfekte“ Ranking-Funktion ist ein bewegliches Ziel
  2. Strenge Geheimhaltung
  3. Bezahlte Werbung

Gliederung

TECHNISCHE UNIVERSITÄT DARMSTADT

Motivation

Link-Analyse mit Hubs und Authorities

PageRank

Anwendung innerhalb des WWW

Anwendungen außerhalb des Internets

Zitationsanalyse

TECHNISCHE UNIVERSITÄT DARMSTADT

- Existiert weitaus länger als das Internet
- Garfield's Impact Factor:

$$\frac{\text{Zahl der Zitate im Bezugsjahr auf die Artikel der vergangenen zwei Jahre}}{\text{Zahl der Artikel in den vergangenen zwei Jahren}}$$

$\Rightarrow$ Voting mit Hilfe von In-Links

- 1970: Gewichtung von Zitierungen $\Rightarrow$ Principle of Repeated Improvement
- *Influence Weights* ~PageRank

U.S. Supreme Court Citations

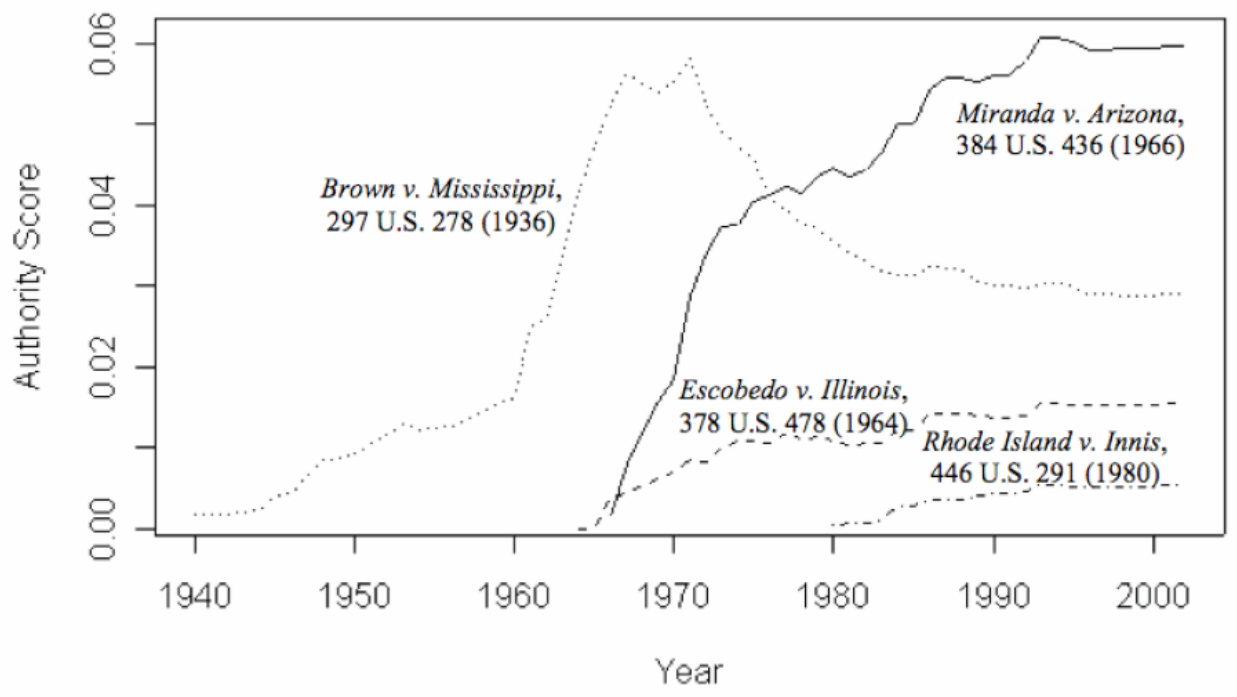
TECHNISCHE UNIVERSITÄT DARMSTADT

- US-Rechtssystem basiert auf der Auswertung von Präzedenzfällen
- Analyse der Verweisstruktur von Rechtsentscheiden kann Schlüsselfälle aufzeigen
- Bsp.: Anwendung des HITS-Algorithmus auf Urteile der letzten 20 Jahre
- Ergebnis: Übereinstimmung mit Expertenmeinungen (sogar besser?)

Authority-Dynamik

TECHNISCHE UNIVERSITÄT DARMSTADT

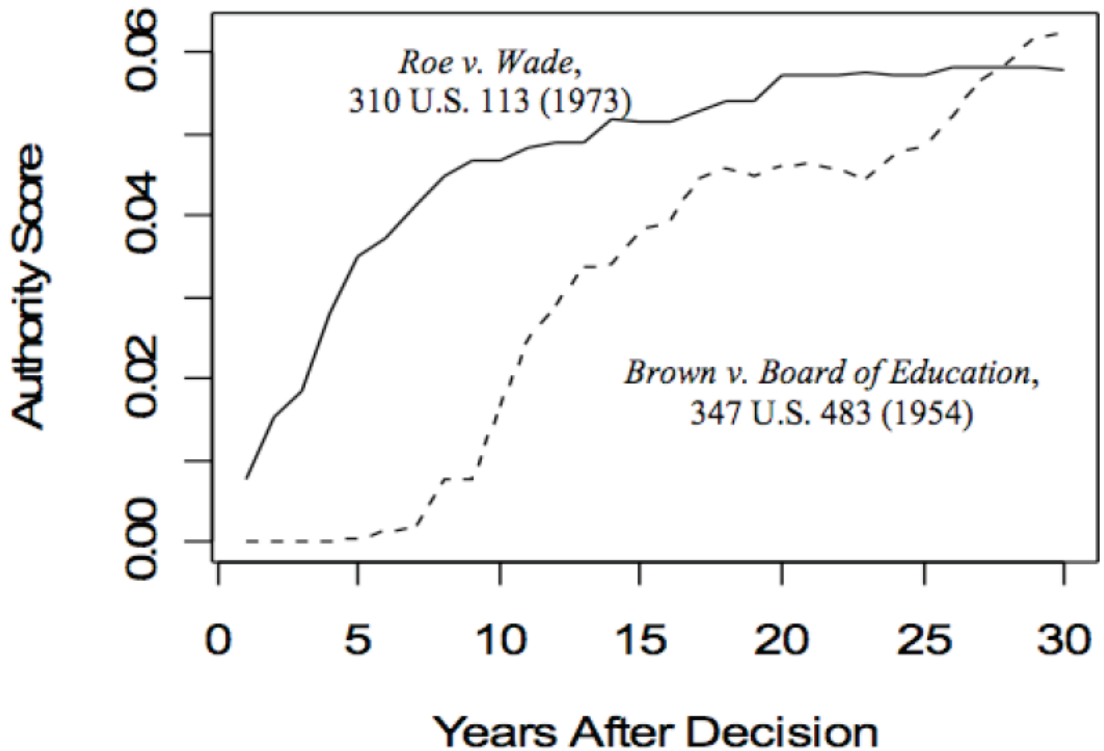
- Analyse von Schlüsselfällen des 5. Zusatzartikels der US-Verfassung



Authority-Dynamik

TECHNISCHE UNIVERSITÄT DARMSTADT

- Analyse von Schlüsselfällen des 5. Zusatzartikels der US-Verfassung



TECHNISCHE
UNIVERSITÄT
DARMSTADT

Vielen Dank

Quellen

TECHNISCHE UNIVERSITÄT DARMSTADT

Inhalt

David Easley, Jon Kleinberg: Networks, Crowds, and Markets: Reasoning About a Highly Connected World, 2010, Kap. 14: Link Analysis and Web Search

Bilder

David Easley, Jon Kleinberg: Networks, Crowds, and Markets: Reasoning About a Highly Connected World, 2010, Kap. 14: Link Analysis and Web Search

`https://www.vile-netzwerk.de/static_pages/politik_archiv/archiv/archiv/www.gemeinsamlernen.de/vile-netzwerk/Regionalgruppen/nord/projekte/Archiv/europa/images/eu_koeche.gif`

`http://ohs-image.ohiohistory.org/images/about/pr/ctm/2002.jpg`

`http://de.toonpool.com/user/611/files/pagerank_241905.jpg`