

PageRank

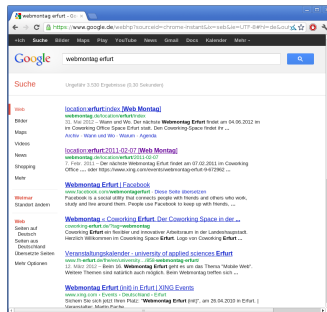
- WebMontag Erfurt -

Sascha Grau

Technische Universität Ilmenau

10. Dezember 2012

Ranking von Dokumenten



Ziel
(Relevanz-)Ordnung über vorgegebener Menge von Dokumenten.

Web-Ranking bis 1998

```
<html>
<head>
  <title>
    [Web Montag Erfurt]
  </title>
</head>
<body>
  <h1>Web-Montag EF</h1>
  <p>Hello World!</p>
</body>
</html>
```

Rein stichwort-basierte Suche

Marktführer:  altavista
THE SEARCH COMPANY

- 1 Indizierung von Text und HTML-Tags
- 2 Reduktion auf Wortstämme
- 3 Textgewichtung und Textnormalisierung (Überschriften, Textlänge)
- 4 häufigkeitsbasierte Auswertung

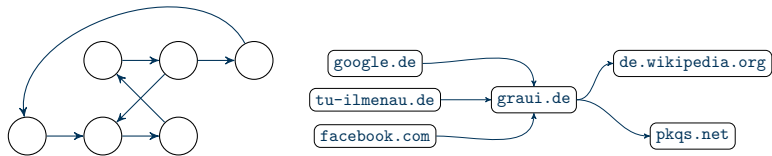
Ein akademisches Projekt



PageRank (S. Brin, L. Page, 1998)

- **Larry Page** und **Sergey Brin** als Doktoranden in Stanford
- Ziel: Ordnung wissenschaftlicher Veröffentlichungen
- 1996: 'BackRub' Suche auf <http://www.stanford.edu>
- 1998: Studienabbruch, Gründung Google Inc.

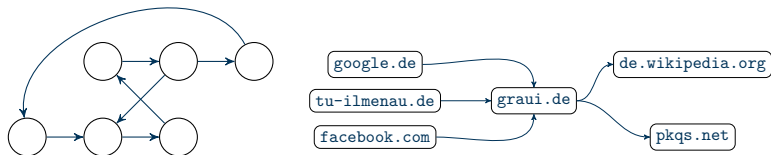
Die Idee von PageRank



Grundidee

- Ranking berücksichtigt zusätzlich **Link-Struktur** der Dokumente.
- Linksetzung als *Empfehlung* der Zielseite
- Links *wichtiger* Seiten sind *wichtiger*

Die Idee von PageRank



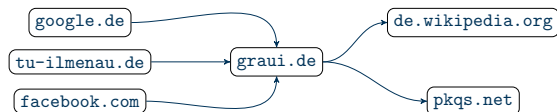
Grundidee

- Ranking berücksichtigt zusätzlich **Link-Struktur** der Dokumente.
- Linksetzung als *Empfehlung* der Zielseite
- Links *wichtiger* Seiten sind *wichtiger*

Erhebung der Link-Daten

Automatisiertes Crawling ("GoogleBot"), Filterung: robots.txt

Rang und Ranking

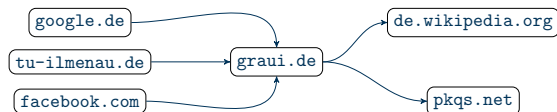


Rang	Seite
0.3	google.de
0.21	de.wikipedia.org
0.2	facebook.com
0.1	tu-ilmenau.de
0.001	graui.de
0.001	pkqs.net

Rang

- Jedes Dokument v besitzt *Rang* $pr[v] \in \mathbb{Q}$
- Ranking ist Sortierung der Dokumente nach ihrem Rang
- Dokumente verteilen eigenen Rang **gleichmäßig** auf alle Nachfolger

Rang und Ranking



Rang	Seite
0.3	google.de
0.21	de.wikipedia.org
0.2	facebook.com
0.1	tu-ilmenau.de
0.001	graui.de
0.001	pkqs.net

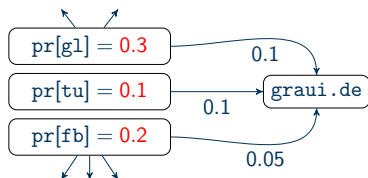
Rang

- Jedes Dokument v besitzt *Rang* $pr[v] \in \mathbb{Q}$
- Ranking ist Sortierung der Dokumente nach ihrem Rang
- Dokumente verteilen eigenen Rang **gleichmäßig** auf alle Nachfolger

Wie kommt Rang anfänglich in das System?

Initialisierungsvektor \mathbf{e} mit Einträgen $\mathbf{e}_v \in \mathbb{Q}$

PageRank Berechnung (1)



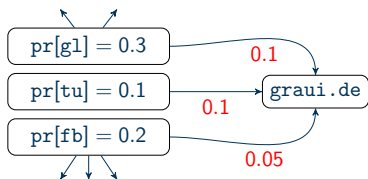
$$\text{pr}[gr] = 0.85 \cdot (0.1 + 0.1 + 0.05) + 0.15 \cdot \mathbf{e}_{gr}$$

$$\text{pr}[v] = d \sum_{w|w \rightarrow v} \frac{\text{pr}[w]}{\text{out}(w)} + (1 - d)\mathbf{e}_v$$

Was ist gegeben?

- Dokumentenstruktur als **Graph**
- Initialisierungsvektor \mathbf{e}
- Dämpfungsfaktor d mit $0 < d < 1$

PageRank Berechnung (1)



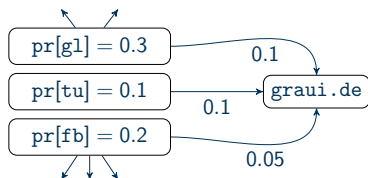
$$\text{pr}[gr] = 0.85 \cdot (0.1 + 0.1 + 0.05) + 0.15 \cdot e_{gr}$$

$$\text{pr}[v] = d \sum_{w|w \rightarrow v} \frac{\text{pr}[w]}{\text{out}(w)} + (1 - d)e_v$$

Was ist gegeben?

- Dokumentenstruktur als Graph
- Initialisierungsvektor e
- Dämpfungsfaktor d mit $0 < d < 1$

PageRank Berechnung (1)



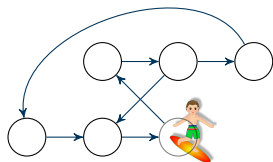
$$\begin{aligned} pr[gr] &= 0.85 \cdot (0.1 + 0.1 + 0.05) + 0.15 \cdot e_{gr} \\ &= 0.2125 + 0.15 \cdot e_{gr} \end{aligned}$$

$$pr[v] = d \sum_{w|w \rightarrow v} \frac{pr[w]}{\text{out}(w)} + (1 - d)e_v$$

Was ist gegeben?

- Dokumentenstruktur als Graph
- Initialisierungsvektor e
- Dämpfungsfaktor d mit $0 < d < 1$

Interpretationen der PageRank-Formel



$$\text{pr}[v] = d \sum_{w|w \rightarrow v} \frac{\text{pr}[w]}{\text{out}(w)} + (1 - d)\mathbf{e}_v$$

$$\text{mit } \sum_{v \in V} \mathbf{e}_v = 1$$

Der Zufallssurfer

- ... startet mit Wahrscheinlichkeit \mathbf{e}_v auf Seite v
- ... entscheidet mit Wkt. $(1 - d)$ dass er seine Reise in v beendet
- ... oder wählt gleichverteilt Nachfolgerseite und springt dorthin
- ... schmolzt falls kein Nachfolger existiert

Der PageRank von v ist die Wkt. dass Zufallssurfer *freiwillig* in v hält.

PageRank Berechnung (2)

PageRank ist Lösung lin. Gleichungssystems

$$(I - dM^T) \text{pr} = (1 - d)\mathbf{e}$$

mit $M = (m_{uv})$ und $m_{uv} = \begin{cases} 1/\text{out}(u) & , \text{ falls } u \rightarrow v \text{ existiert} \\ 0 & , \text{ sonst.} \end{cases}$

- **Herausforderung:** $\geq 7.620.000.000$ Webseiten im Google-Index
- Direkte Lösung: Gauss'sches Eliminationsverfahren
- Iterative Lösung: (Jakobi / Gauss-Seidel-Verfahren)
 - ▶ Beginn mit beliebiger Initialisierung
 - ▶ Rundenweise Adaption der Rang-Werte über PageRank-Formel
 - ▶ Mit $0 < d < 1$ ist Konvergenz beweisbar
 - ▶ hervorragend parallelisierbar

PageRank Berechnung (2)

PageRank ist Lösung lin. Gleichungssystems

$$(I - dM^T) \text{pr} = (1 - d)\mathbf{e}$$

mit $M = (m_{uv})$ und $m_{uv} = \begin{cases} 1/\text{out}(u) & , \text{ falls } u \rightarrow v \text{ existiert} \\ 0 & , \text{ sonst.} \end{cases}$

- Herausforderung: $\geq 7.620.000.000$ Webseiten im Google-Index
- Direkte Lösung: Gauss'sches Eliminationsverfahren
- **Iterative Lösung: (Jakobi / Gauss-Seidel-Verfahren)**
 - ▶ Beginn mit beliebiger Initialisierung
 - ▶ Rundenweise Adaption der Rang-Werte über PageRank-Formel
 - ▶ Mit $0 < d < 1$ ist Konvergenz beweisbar
 - ▶ hervorragend parallelisierbar

Personalisierung (1)

Interpretationen Zufallssurfer

Initialisierungsvektor \mathbf{e} bestimmt Startpunkt

⇒ \mathbf{e} zu a priori Auf- und Abwertung von Webseiten einsetzbar

- generell: beliebiger Vektor \mathbf{e} funktioniert
- nur Verhältnis der Einträge ist relevant

- Beispiel:
$$\mathbf{e}_{\text{de.wikipedia.org}} = 0.1$$
$$\mathbf{e}_{\text{webmontag.de}} = 0.05$$
$$\mathbf{e}_{\text{bing.de}} = 0.00000001$$



Personalisierung (1)

Interpretationen Zufallssurfer

Initialisierungsvektor \mathbf{e} bestimmt Startpunkt

⇒ \mathbf{e} zu a priori Auf- und Abwertung von Webseiten einsetzbar

- generell: beliebiger Vektor \mathbf{e} funktioniert
- nur Verhältnis der Einträge ist relevant

- Beispiel:
$$\mathbf{e}_{\text{de.wikipedia.org}} = 0.1$$
$$\mathbf{e}_{\text{webmontag.de}} = 0.05$$
$$\mathbf{e}_{\text{bing.de}} = 0.00000001$$



Aber:

- PageRank-Berechnung pro individualisiertem Vektor \mathbf{e} extrem aufwendig
- so nur zentralisierte Gewichtung, keine wirklich Individualisierung


Personalisierung (2)

Eine praktische Eigenschaft

PageRank ist lineare Abbildung des Personalisierungsvektors \mathbf{e} , d.h. es gilt

$$\begin{aligned}c \cdot \text{pr}_{\mathbf{e}}[v] &= \text{pr}_{c \cdot \mathbf{e}}[v], \\ \text{pr}_{\mathbf{e}_1}[v] + \text{pr}_{\mathbf{e}_2}[v] &= \text{pr}_{\mathbf{e}_1 + \mathbf{e}_2}[v]\end{aligned}$$

Thematische Individualisierung

- PageRank-Vorbereitung für verschiedene Individualisierungsvektoren $\mathbf{e}_1, \dots, \mathbf{e}_k$
- z.B. thematische Vektoren: News, Unterhaltung, Sport, Linux 
- Nutzer wählt Themenbereiche $T \subseteq \{1, \dots, k\}$
- gewichtete Kombination ergibt PageRank:

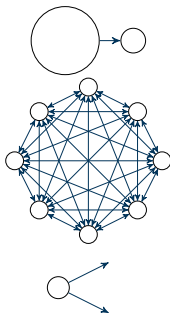
$$\text{pr}[v] = \sum_{i \in T} c_i \cdot \text{pr}_{\mathbf{e}_i}[v]$$

Einflussnahme auf eigenen PageRank

Direkte Einflussnahme schwierig, da PageRank von **eingehenden** Links bestimmt wird.

PageRank-Manipulation

- eingehende Links von Seiten mit hohem Rang (Link-Tausch, unfreiwillig: Kommentar-Spam)
- hohe Zahl eingehender Links und enge Vermaschung (Link-Farming, Klone)
- aber auch: selbst Links setzen (Google erteilt Mali auf Senken)



Abwertung gemäß Google Qualitätsrichtlinien

- strukturelle Analyse
 - ▶ Link-Farming, sehr hohe Ausgangsgrade
 - ▶ inhaltsarme aber suchthemenoptimierte Eingangsseiten die auf Hauptseite verlinken
- inhaltliche Analyse
 - ▶ generierte Inhalte (Übersetzungen, automatische Mash-Ups, ...)
 - ▶ versteckter Text oder Links
 - ▶ Kopien hochwertiger Inhalte (Wikipedia)
 - ▶ unnatürlich hohe Frequenz von Schlüsselworten (Listen, Wiederholungen)
- Verhaltensanalyse
 - ▶ versch. Content für Suchmaschinen und User (Redirect, Browserschranke)
 - ▶ manipulierte “rich snippets”
 - ▶ Phishing, Malware, ...

Fazit

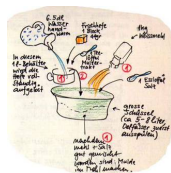
Rezept für solide Positionierung im Ranking

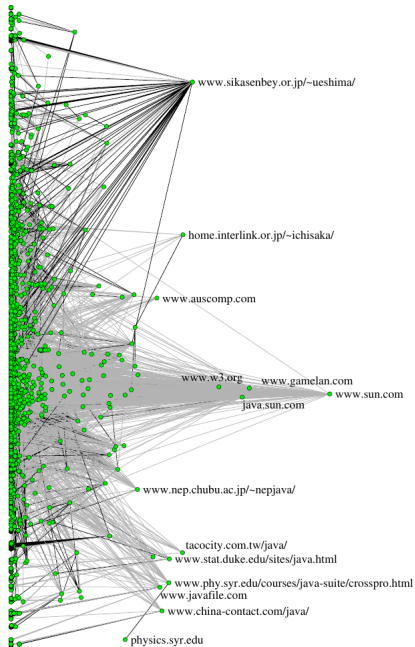
PageRank allein:

- häufige Verlinkung
- Verlinkung durch einflussreiche Seiten
- gezieltes und bedachtes Setzen eigener Links

Zusätzlich durch Google:

- gute, *eigene* Inhalte
- richtige Stichworte
- Nutzung der Google-Werkzeuge (z.B. rich snippets)
- wenig "schmutzige" Tricks





Vielen Dank für eure
Aufmerksamkeit.

Literatur



Ulrik Brandes

Visual Ranking of Link Structures.

In *Journal of Graph Algorithms and Applications*, 7(2):191–201, 2003.



Sergey Brin and Lawrence Page.

The anatomy of a large-scale hypertextual web-search engine.

In *Proc. of the 7th World Wide Web Conference (WWW7)*, 1998.



Michael Brinkmeier.

PageRank revisited.

In *ACM Trans. Internet Techn.*, 6(3), S. 282-301, 2006



Taher H. Haveliwala.

Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search.

In *IEEE Trans. Knowl. Data Eng.*, 15(4):784–796, 2003.