
Wie funktioniert das Ranking bei modernen Suchmaschinen?

Prof. Dr. Clemens Cap

Universität Rostock

www.internet-prof.de

Aufgabe des Ranking

User sehen sich typischerweise die ersten 10 – 40 Treffer an
Gutes Ranking bestimmt also, ob man gesehen wird

Interesse der Autoren: Hohes Ranking

Interesse der Search Engine: Hohes Ranking = Gute Seite (für den User !)

Früher: Naives Zählen von Suchworten

Heute: Linktopologische Analyse

Hypothese: Nur gute Sites verlinken auf gute Sites

Frage: Was kann man daraus machen?

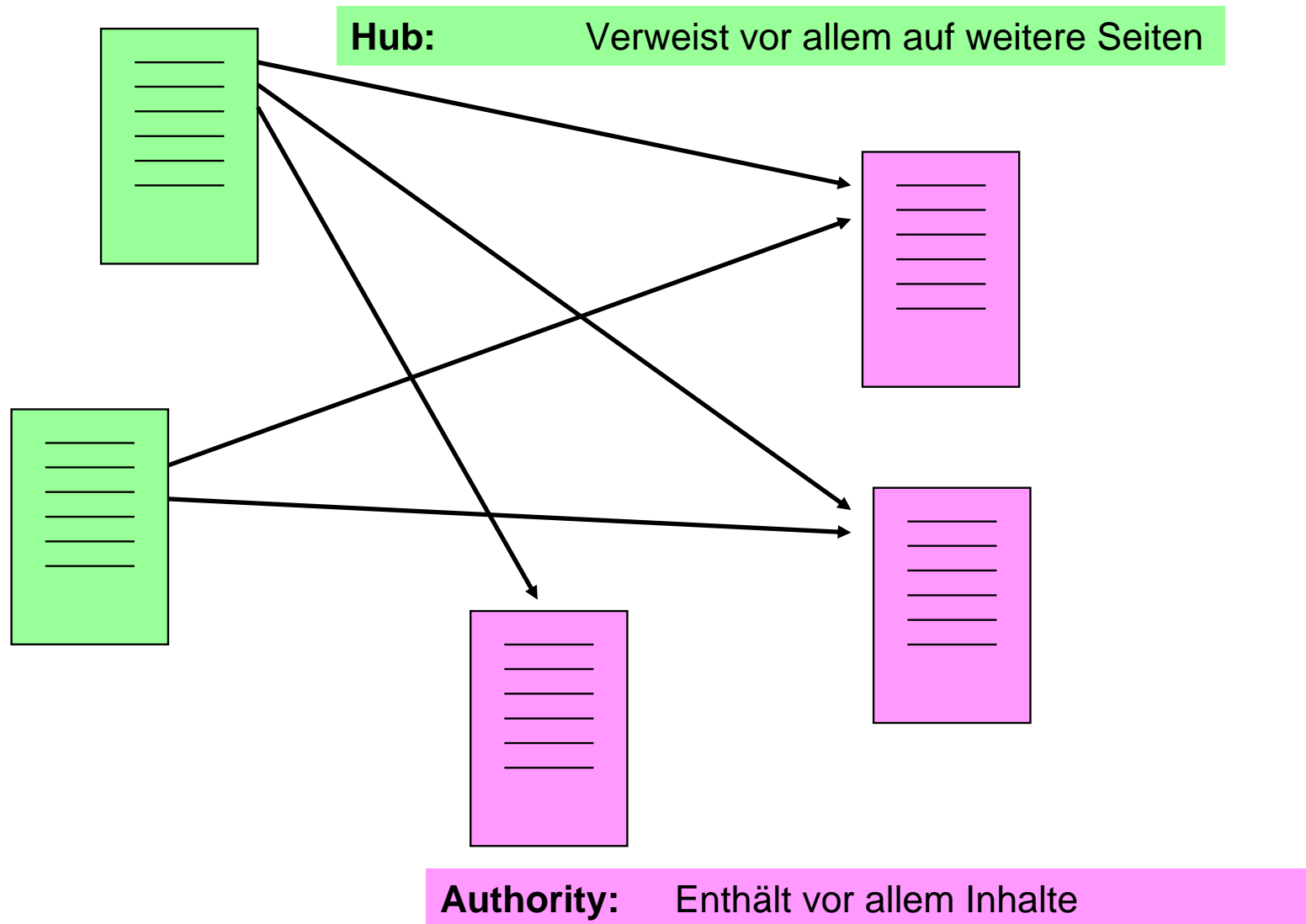
HITS-Algorithmus

Pagerank (Kern des Google Mechanismus)

HITS Algorithmen

Hypertext Induced Topic Selection

Hubs und Authorities



Hubs und Authorities

Grundkonzept

- **Hub:** Verweist vor allem auf weitere Seiten
- **Authority:** Enthält vor allem Inhalte

Zirkuläre Definition

- Eine Seite ist ein guter Hub, wenn sie auf gute Autoritäten verweist
- Eine Seite ist eine gute Autorität, wenn sie auf gute Hubs verweist

Frage: Kann diese zirkuläre Definition zu einem Algorithmus führen ?

Exkurs: Ableitung des HITS

Jede Seite hat einen Hub und einen Authority Wert

- Hub: Verweist auf gute Seiten
- Authority: Wird als gut angesehen, da viele darauf verweisen

$$h_i = \varphi \sum a_k \quad a_i = \alpha \sum h_k$$

h_i verweist auf a_k *h_k verweist auf a_i*

$$h_i = \varphi \sum_{k=1}^n A_{ik} a_k \quad a_i = \alpha \sum_{k=1}^n A_{ik}^t h_k$$

$$\vec{h} = \alpha \varphi A A^t \vec{h}$$

Eigenwert / Eigenvektor
Problem

Praktische Berechnung

Problem

- Matrix nie statisch bekannt, da laufend weiter gecrawlt wird
- Komplette Eigenwert Berechnung zu aufwendig
- Daher laufendes Iterationsverfahren

Vorgehen

- Start mit randomisierten Werten
- Schritt für Schritt abarbeiten der zirkulären Definition
- Bei Erweiterung des Graphen Vektoren erweitern und weiterrechnen

Qualität versus Quantität Problem

Problem

- Hub der auf **viele schlechte** Authorities zeigt wird besser gewertet als ein Hub der auf **sehr wenige gute** Authorities zeigt

Lösung 1: Hub Averaging

- Hub Wert in Richtung des Mittelwerts aller vom Hub referenzierten Autoritäten korrigieren

Lösung 2: Hub Average-Thresholding

- Bei Bestimmung des Authority Wertes nur jene Hubs berücksichtigen, deren Hub-Wert über dem Mittelwert aller verweisenden Hub-Werte steht

Lösung 3: Authority Best-n Thresholding

- Bei Bestimmung des Hub Wertes nur die besten n Autoritäten berücksichtigen

Beachte: Kombinationsmethoden sind oft schlechter (sic !) als die Einzelmethoden

Pagerank

Pagerank

Ranking Mechanismus von Google

- Daher auch zentrale Bedeutung

Mögliche Zugänge zum Pagerank

- | | |
|-----------------|----------------------|
| 1. Heuristisch | Einfache Vorstellung |
| 2. Stochastisch | Random Surfer Modell |
| 3. Eigenvektor | Sichert Konvergenz |

Anzeige in der
Google Toolbar



Pagerank

Heuristischer Zugang

Pagerank: (Erster) Ranking Mechanismus von Google

$$PR(A) = (1 - d) + d \left(\sum \frac{PR(T_i)}{C(T_i)} \right)$$

PR(X): Pagerank von Seite X

d: Dämpfung und Ausgleichsfaktor

C(X): Gesamtzahl der Links, die auf Seite X zeigen

Summenbildung über alle Seiten T_i , von denen ein Link auf die Seite A führt

Pagerank

Stochastischer Zugang

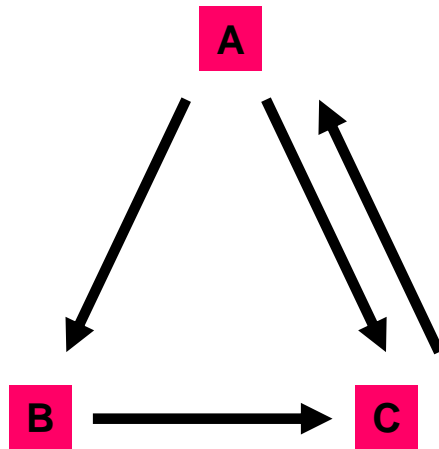
Random Surfer Modell

- **Einsprung:** Springe an beliebiger Stelle (gleichverteilt) ins Web
- **Fortsetzung:** Wähle den Fortsetzungslink gleichverteilt aus allen möglichen Optionen aus
- **Ferner:** Wenn kein Fortsetzungslink, dann beliebiger Sprung
- **Ferner:** Mit bestimmter Wahrscheinlichkeit erfolgt ein Abbruch des Surfens und ein erneuter Zufallssprung (sog. Abbruchwahrscheinlichkeit)
- **Pagerank der Seite ist Wahrscheinlichkeit**, den Surfer auf dieser Seite zu finden
- **Beachte:** Erfordert Normierung der Pagerank Werte so, daß Summe der Wahrscheinlichkeiten über das ganze Web 1 ergibt

Random Surfer Modell erlaubt ebenso die Aufstellung der Gleichungen !

Pagerank

Stochastischer Zugang



$$PR(A) = \frac{1}{3} + \frac{2}{3} PR(C)$$

$$PR(B) = \frac{1}{3} + \frac{2}{3} (PR(A) / 2)$$

$$PR(C) = \frac{1}{3} + \frac{2}{3} (PR(A) / 2 + PR(B))$$

Lösung der Gleichungssystems

- Klassischer Solver –oder –
- Iterationsverfahren

Beachte

- Pagerank genügt Eigenvektor Gleichung

Praxis

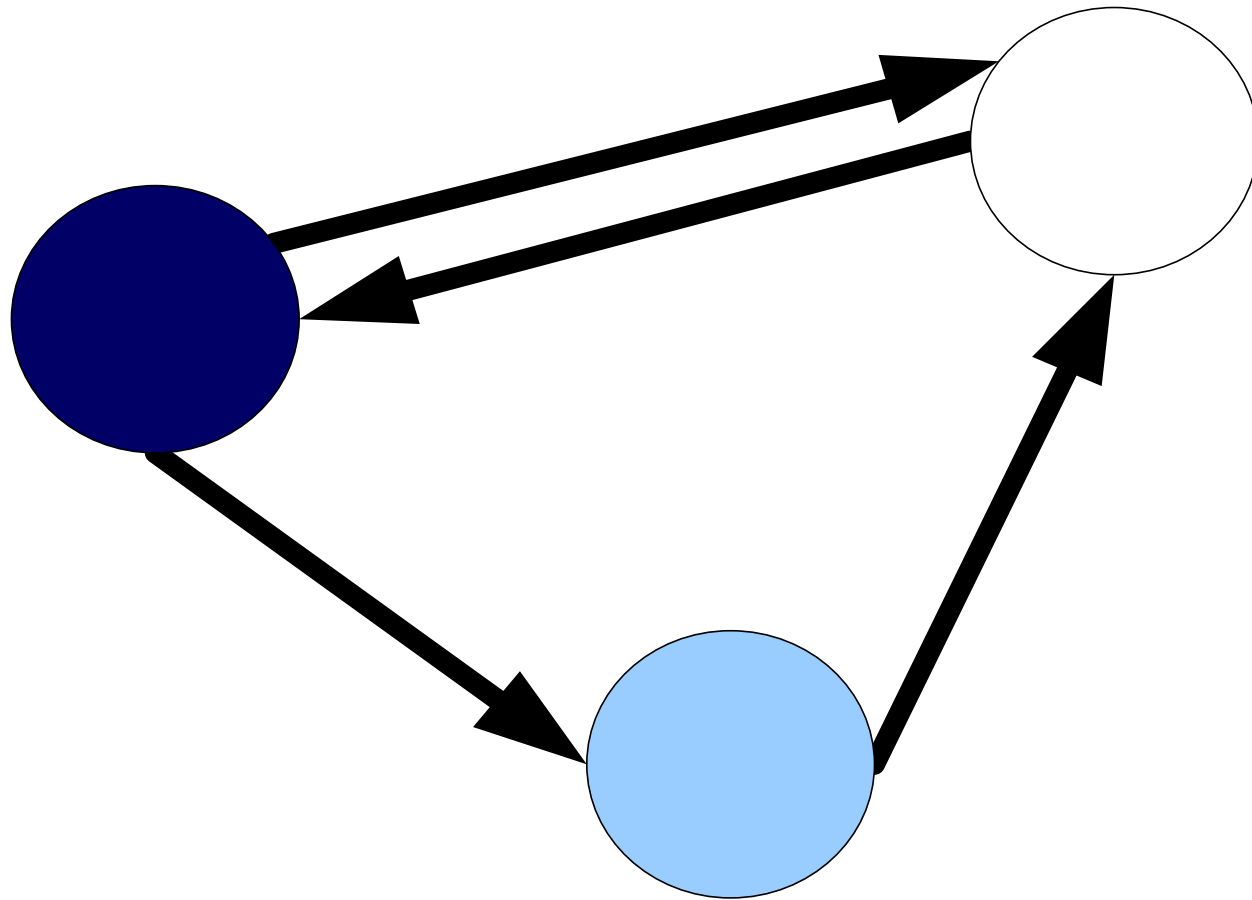
- Kenne Web nicht komplett
- Nur inkrementelle Erweiterung
- Daher nicht klassischer Solver
- Iterationsverfahren hilft
- Parallel zur Iteration Modifikation der Matrix der Gleichung

Exkurs: Iteration des Pagerank

Quelle: Jan Korves, Münster

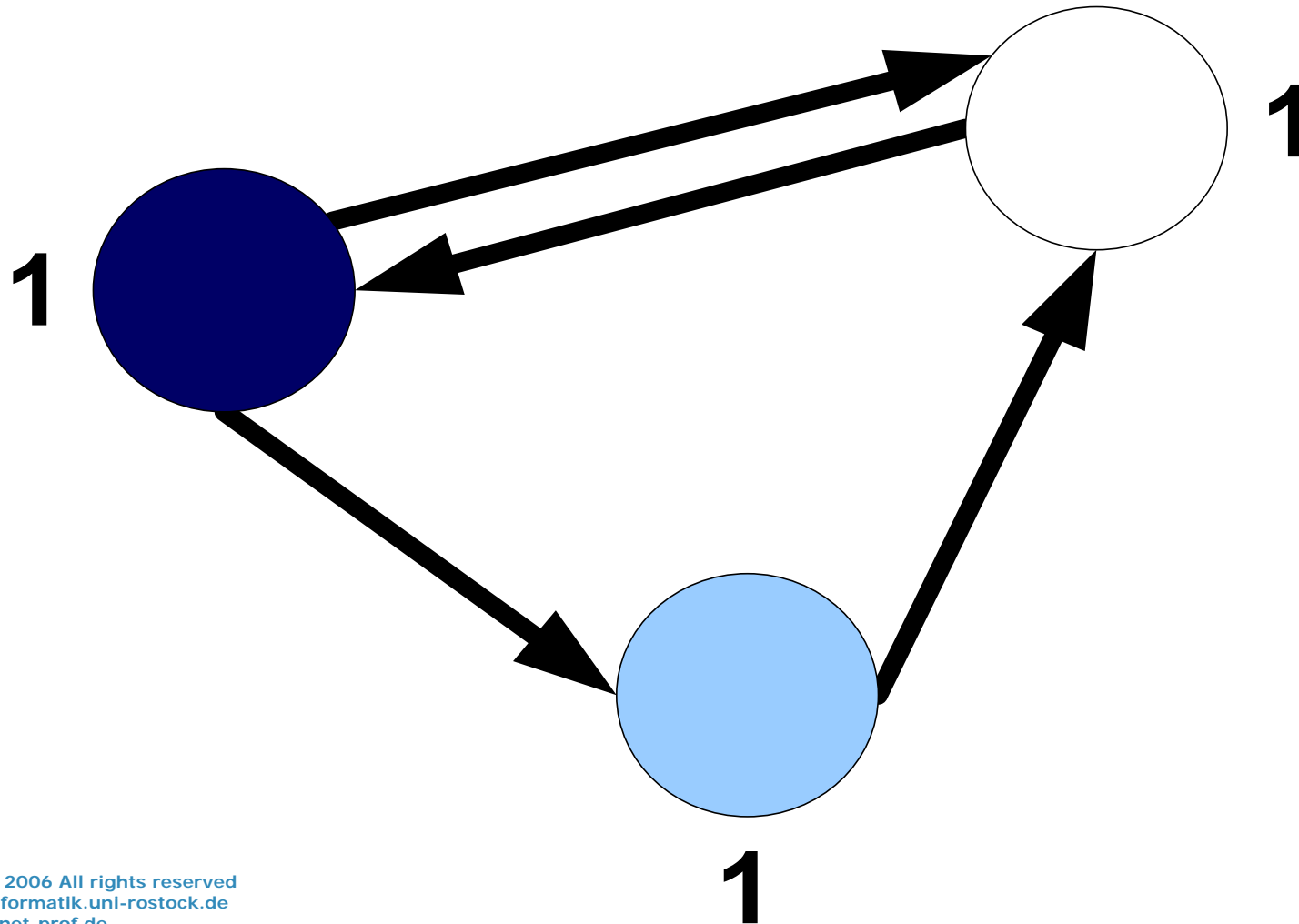
PageRank-Beispiel

- Beispielgraph mit 3 Seiten und 4 Links:



PageRank-Beispiel

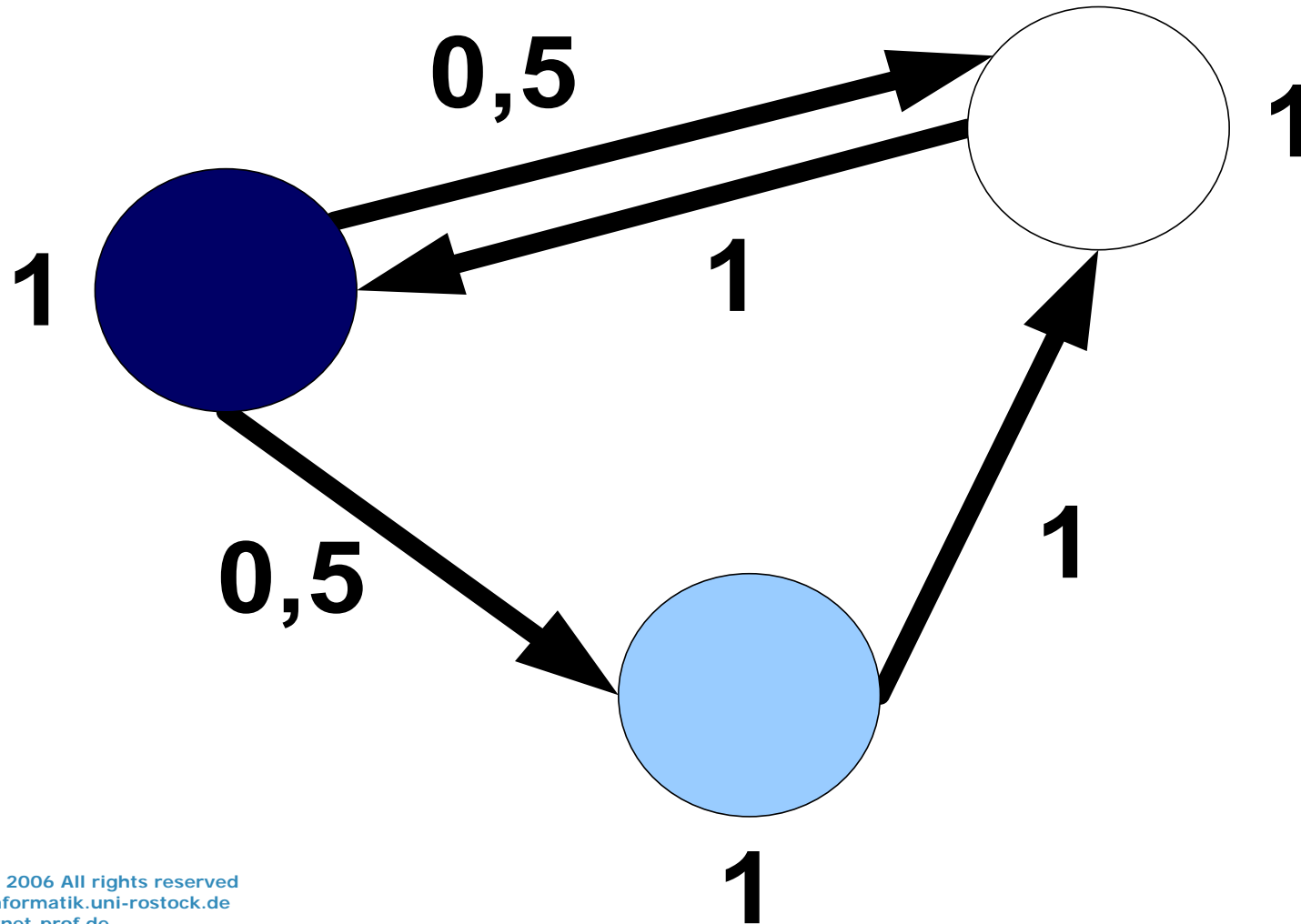
- Initialisierung des PageRank aller Seiten



PageRank-Beispiel

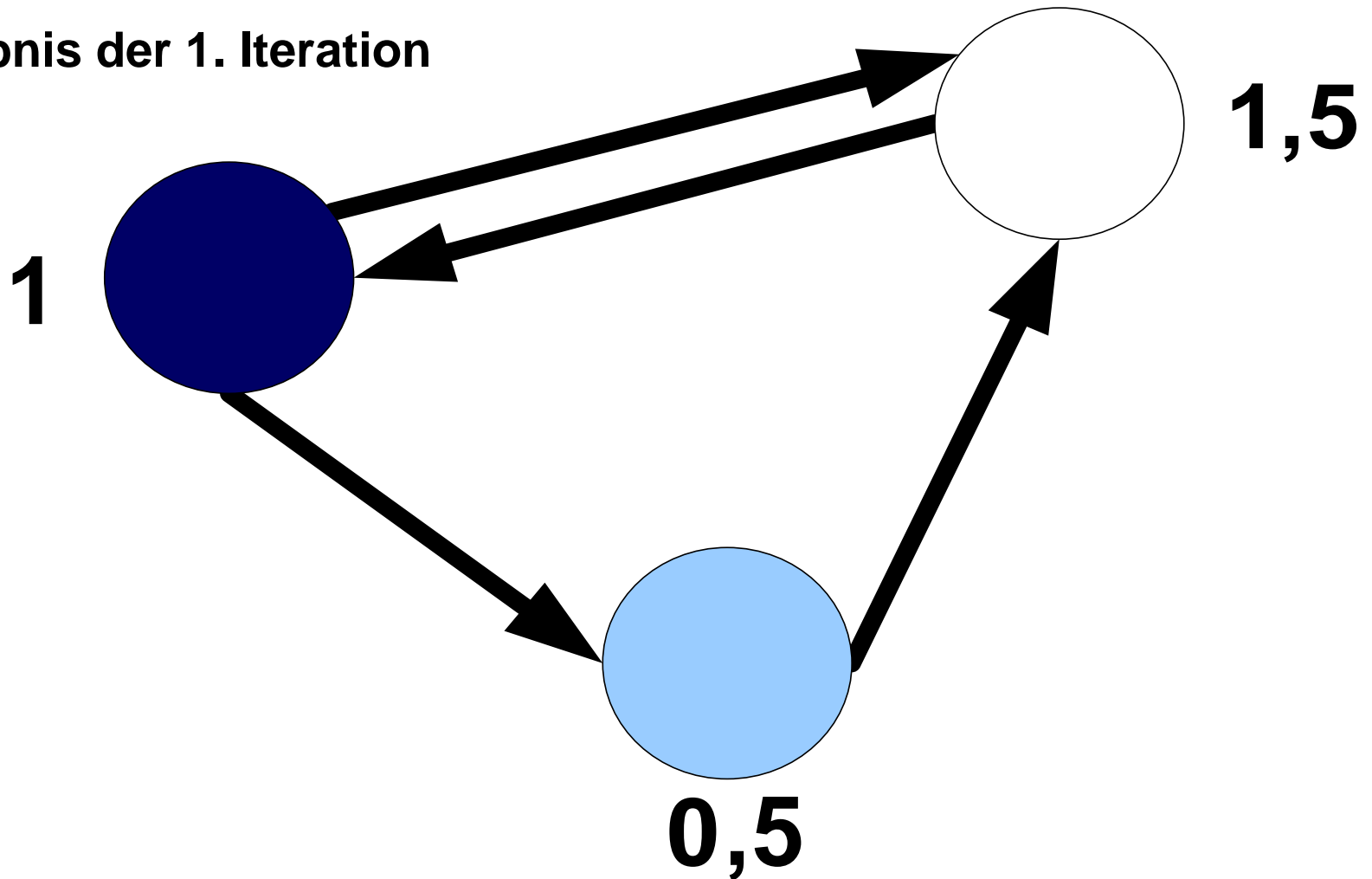
Beachte: Keine Dämpfung

- 1. Iteration: Aufteilung des PageRank über die Links



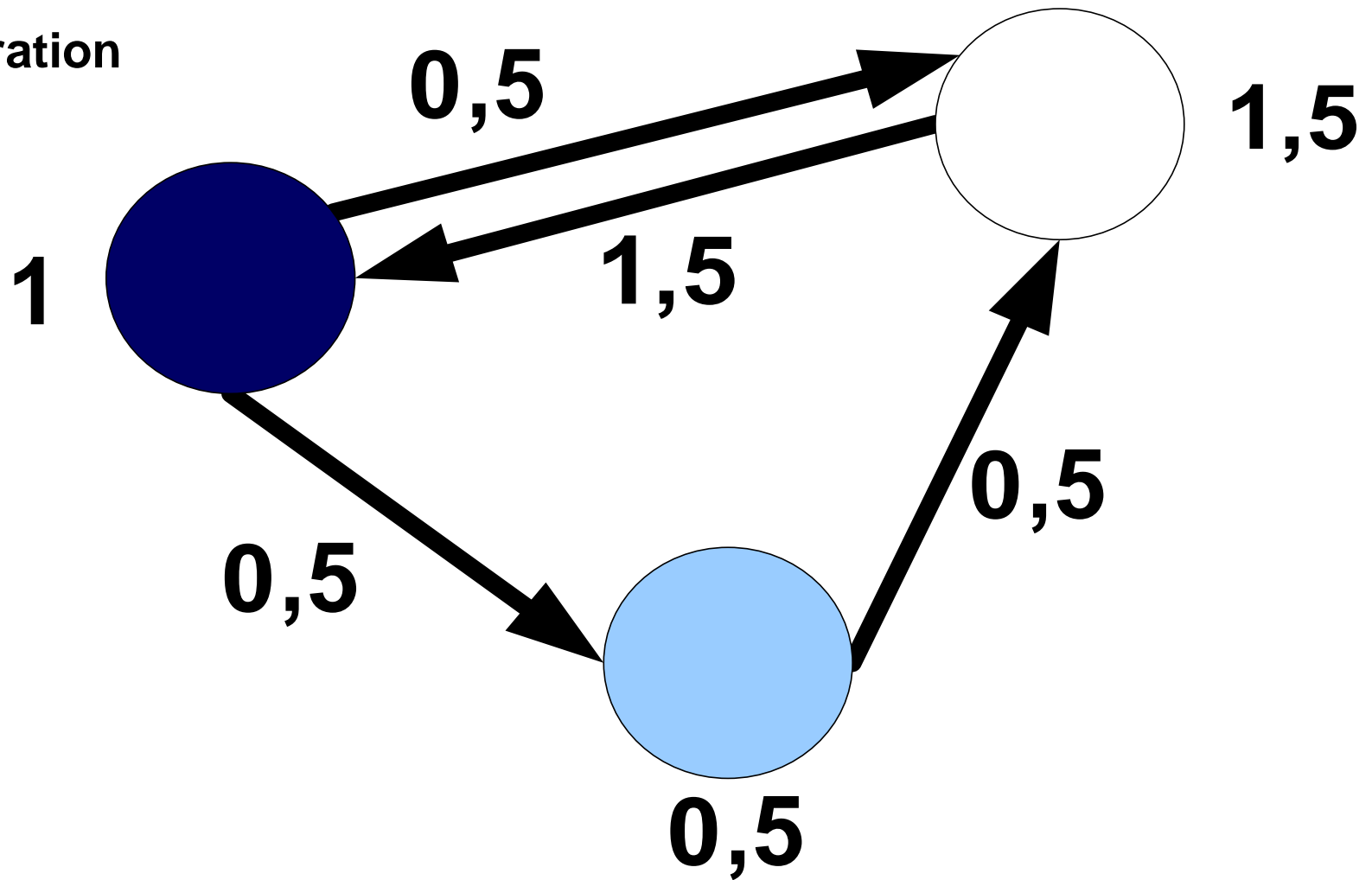
PageRank-Beispiel

- Ergebnis der 1. Iteration



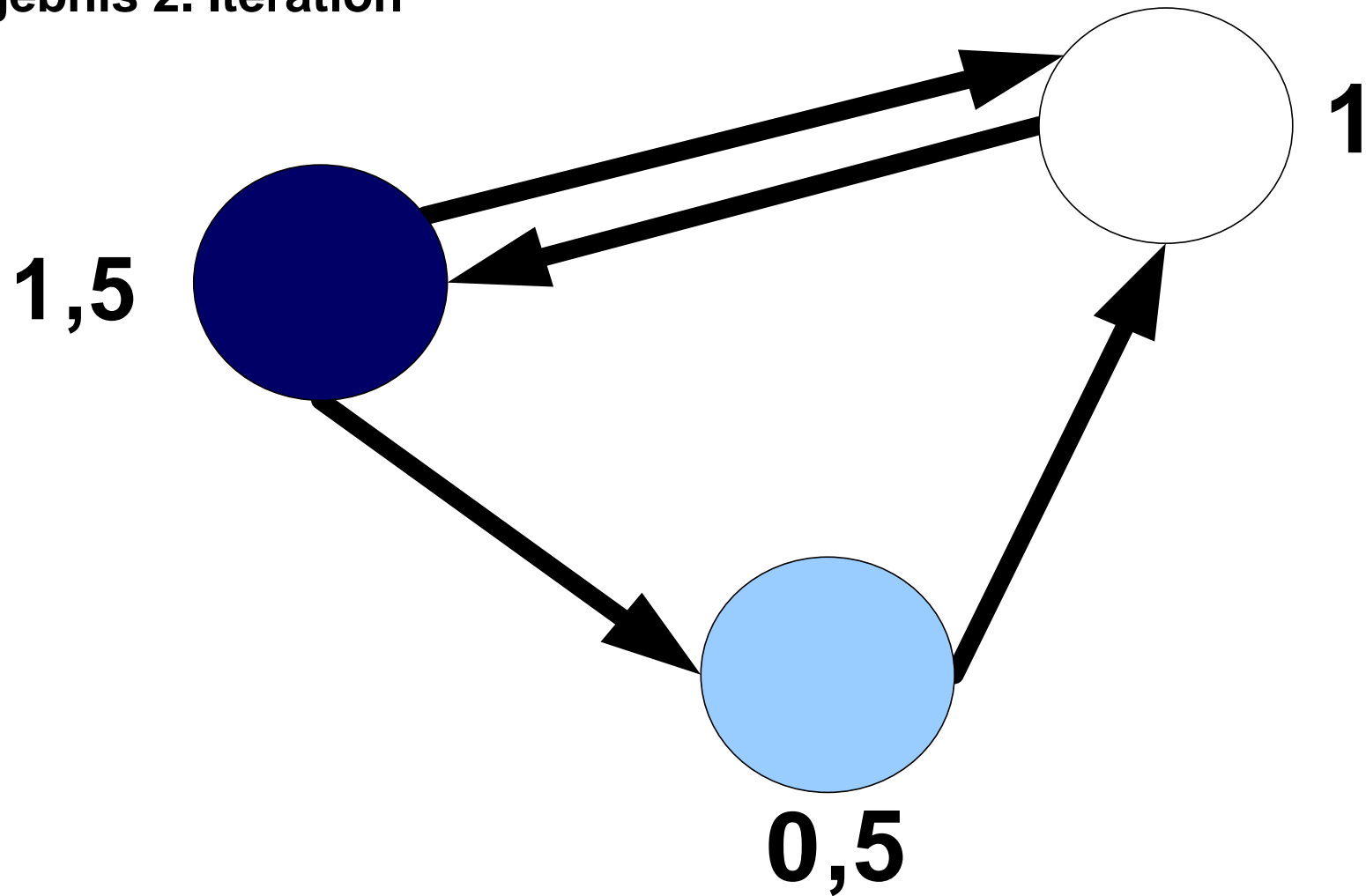
PageRank-Beispiel

- 2. Iteration



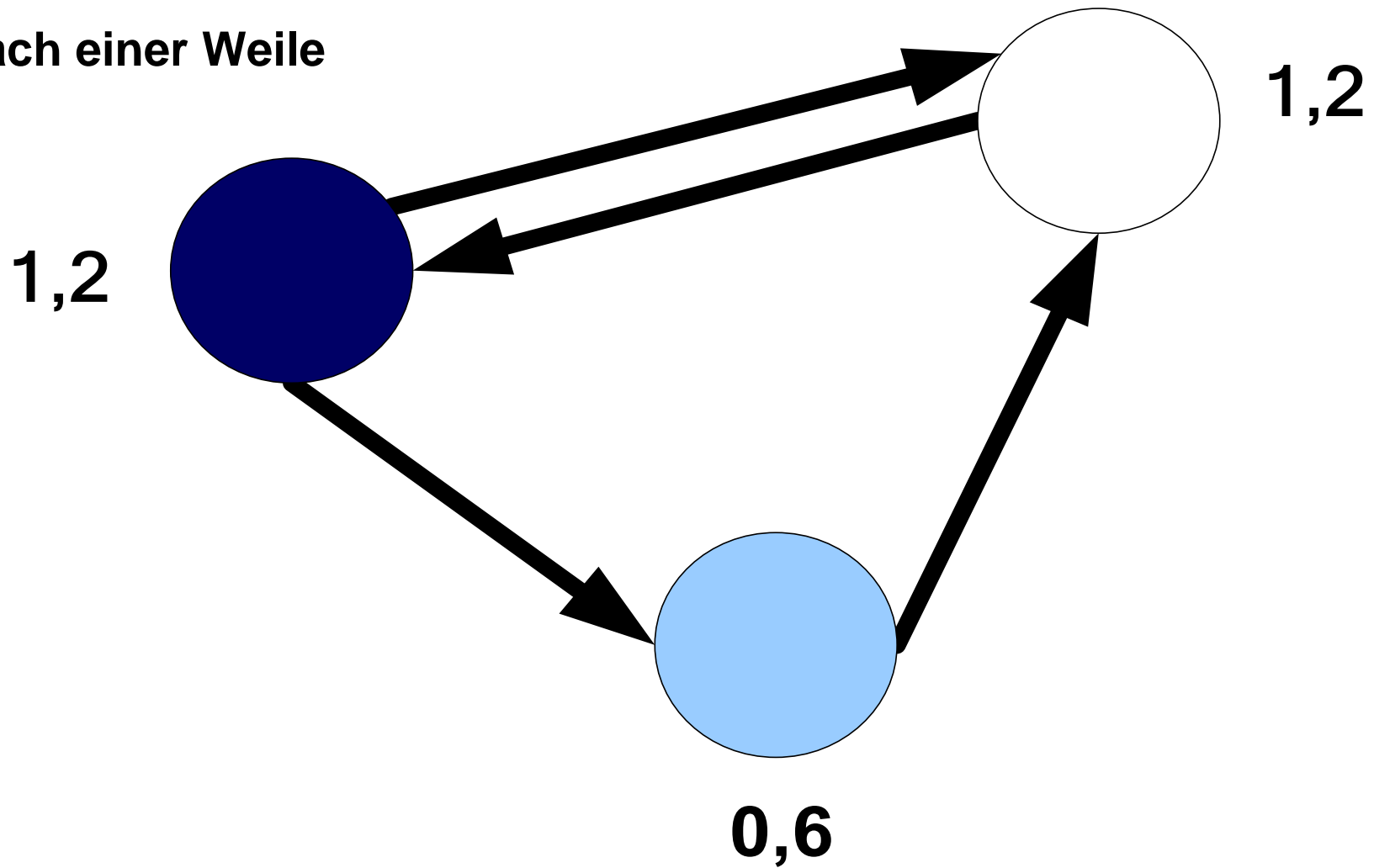
PageRank-Beispiel

- Ergebnis 2. Iteration



PageRank-Beispiel

- Nach einer Weile



Pagerank Bewertung

Eigenschaft:	Beurteilt die Topologie des Netzes
Robust:	Schwer von einem einzelnen zu spammen
Anwendung:	Auch für andere Netze geeignet (Bsp: Soziale Netze, Freundschaften)

Probleme

- Ideal zum **Ranking von Seiten**, aber nicht ideal zum **Ranking von Suchanfragen**
Bsp: Seite hat mit Suchanfrage fast nichts zu tun, ist aber insgesamt gut verlinkt
Fazit: Sehr hoher Pagerank
Ausweg: Themenspezifischer Pagerank
- Es gibt Ansätze zum Spammen, wenn Sites **kollaborieren**

Fazit

- Wichtig ist, daß möglichst viele Seiten,
- die am besten noch wenig abgehende Links haben
- auf mich verweisen

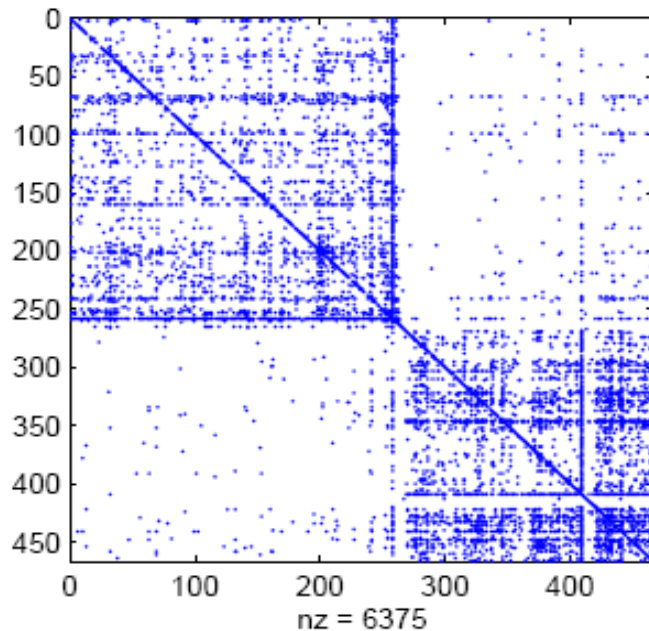
Pagerank

Variante: Blockrank

Web hat eine lokale Struktur

Beispiel

- 1. Quadrant Stanford Domäne
- 3. Quadrant Berkeley Domäne
- Diagonale zeigt stark ausgeprägte Blockstruktur



<http://www.internet-prof.de>

Vorgehensweise

- Berechnung des Pagerank für alle Seiten einer Domäne
- Gewichtung des lokalen Pageranks nach der Bedeutung der Domäne
- Lokale Pageranks als Einheit betrachten

Vorteile

- Schnellere Berechnung des Pagerank
- Lokale Pageranks rechnen Dominanz anderer Domänen "heraus"

Kritik linktopologischer Ansätze

Topic Drift

Bsp 1: Topic Drift durch irrelevante Dokumente

- Eine Seite kann sehr gut verlinkt, für die Suche thematisch aber bedeutungslos sein
- Bsp: Google Ad Words verweisen (zu vielen Themen) alle auch auf Google
Google gute Seite, aber **nicht zum Thema "Autokauf"**
- **Ausweg:** Content-Analyse, zur Feststellung, daß Seite wenig mit bestimmtem Thema zu tun hat

Bsp 2: Topic Drift bei eng verbundenen Communities

- Bei komplexer Suche mit mehreren Themenzentren
wird die besser vernetzte Community unfair bevorzugt
- Anfrage "Jaguar Auto" **driftet ins Thema Auto** ab und
zeigt nicht gute Treffer zum Thema Auto und unkonventionelles Haustier

Automatische Links

Automatisch gesetzte Links

- Reflektieren keine menschliche Seitenbewertung
- Erzeugen keinen Aufwand, sind daher tendentiell häufiger als manuelle Links
- Bsp: Google Ads verlinken alle auch auf Google Homepage
 - Das ist keine Wertschätzung für Google als gute Seite für fast alle Themen
 - Ferner: Phänomen des "Topic Drift durch irrelevante Dokumente" (s. oben)

Nachteil

- Die Links tragen stark zu jeder Art linktopologischer Analyse bei
- Die Links können auch themenrelevant sein
- Sie reflektieren aber nicht menschliche Wertschätzung

Reinforcing Relationships

Problematik

- Gegenseitig stark aufeinander verlinkende Sites treiben Bewertung hoch
- Werden ohne weitere Maßnahmen nicht erkannt

Ergebnis

- Bewertung führt zu "Link"-Tourismus
- Linkst Du mich dann link ich Dich

Ausweg

- Weitere Heuristiken sind erforderlich

Auswege

Content Analysis

Textbasierte Inhaltsanalyse

Idee

- Zähle in einem Dokument d die wichtigsten k Wörter nach der Häufigkeit
- Ggf: Gewichte noch mit der Häufigkeit des jeweiligen Wortes generell
Generell = In der Sprachlandschaft bzw. im Internet
- Erhalte daraus einen Vektor im k -dimensionalen Raum
- **Frage 1:** Ist dieser Vektor nahe zum Vektor eines anderen Dokuments ?
- **Frage 2:** Ist dieser Vektor nahe zu einem Referenzvektor einer Thematik ?

Benötige noch: Distanzmaß zwischen Vektoren des k -dimensionalen Raumes

Idee: Winkel zwischen den Vektoren (sog. Cosinus Maß)

Cosinus Maß

$$\cos \angle(x, y) = \frac{\sum_{i=1}^k x_i y_i}{\sqrt{\sum_{i=1}^k x_i^2} \sqrt{\sum_{i=1}^k y_i^2}}$$

Anwendung der Textbasierten Inhaltsanalyse

Dokumenten Ähnlichkeit

- Zur ersten Beurteilung von Dokumentenähnlichkeit

Pruning

- Entfernung fernliegender Dokumente vor dem Ranking (HITS, Pagerank)

Modifikation von HITS und Pagerank

- Ähnlichkeit mit Referenzvektoren als Faktor an die Werte in linktopologischen Algorithmen dranmultiplizieren

Topic Detektion

- "Zeitungssente" oder "Füttern am See"
- Vergleich mit repräsentativen Referenzvektoren zum Gebiet