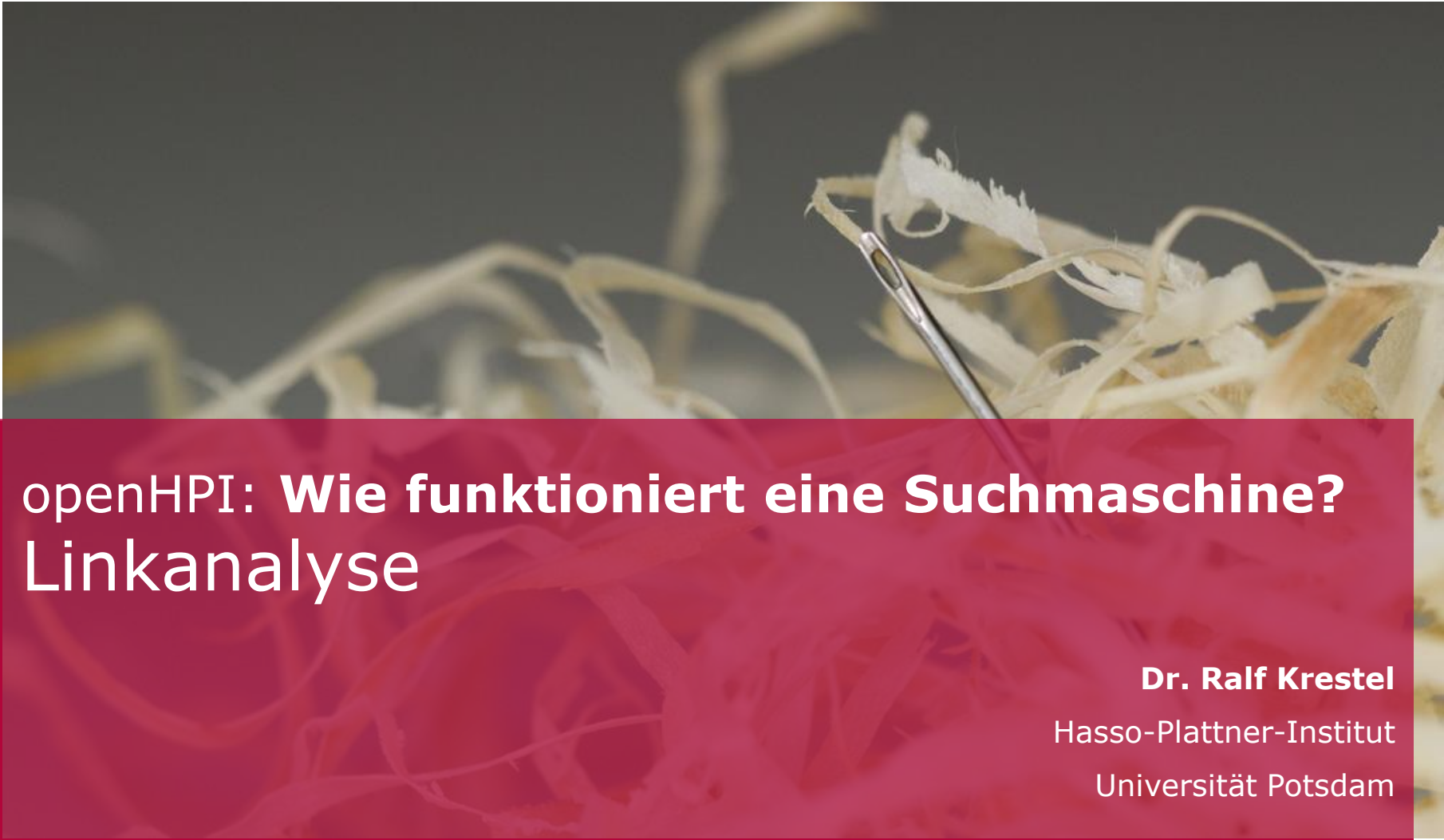




openHPI: **Wie funktioniert eine Suchmaschine?** Linkanalyse

Dr. Ralf Krestel
Hasso-Plattner-Institut
Universität Potsdam

- Klassisches Information Retrieval
 - Der Wert eines Dokuments in Bezug auf eine Anfrage beruht **nur** auf dem **Inhalt** des Dokuments (**intrinsisch**)
 - Dokumente sind in sich geschlossene Einheiten, welche **qualitativ** hochwertig und **vertrauenswürdig** sind
- Moderne Websuche
 - Qualität und Vertrauenswürdigkeit bei **Webseiten** nicht gegeben
 - Einbeziehen **extrinsischer** Faktoren hilft. Annahme:
 - **Links** sind Empfehlungen
 - **Ankertext** kann als Beschreibung der verlinkten Seite angesehen werden
 - Anwendung von Methoden zur Analyse von **sozialen Netzwerken**

Linkanalyse

Links sind Empfehlungen

- Annahme 1:
 - Ein **Link** ist ein **Indikator für Qualität** und/oder Beliebtheit
 - Man könnte auch sagen ein Link ist ein demokratisches Votum

Twitter Considers Premium Version After 11 Years As a Free Service (reuters.com) **64**

Posted by BeauHD on Thursday March 23, 2017 @09:03PM from the gotta-pay-the-bills-somehow dept.

Apple Explores Using An iPhone, iPad To Power a Laptop (appleinsider.com) **57**

Posted by BeauHD on Thursday March 23, 2017 @08:45PM from the two-in-one dept.

According to the U.S. Patent and Trademark Office, Apple has filed a patent for an "[Electronic accessory device](#)." It describes a "thin" accessory that contains traditional laptop hardware like a large display, physical keyboard, GPU, ports and more -- [all of which is powered by an iPhone or iPad](#). The device powering the hardware would fit into a slot built into the accessory. AppleInsider reports:

#selfie



RETWEETS 902 LIKES 1,643

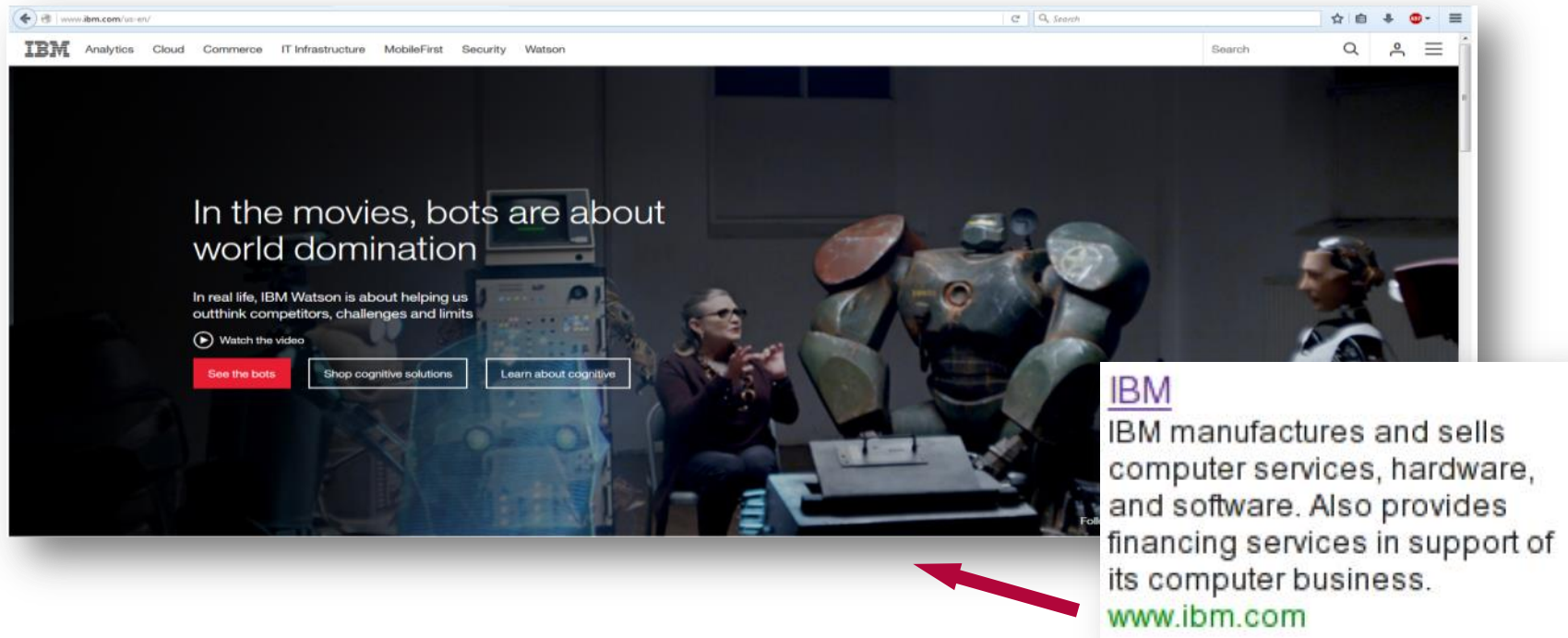
3:24 AM - 8 Oct 2014

121 902 1.6K

Linkanalyse

Ankertexte sind Beschreibungen

- Annahme 2:
 - Der **Ankertext** (oder der Text drum herum) **beschreibt die verlinkte Seite**



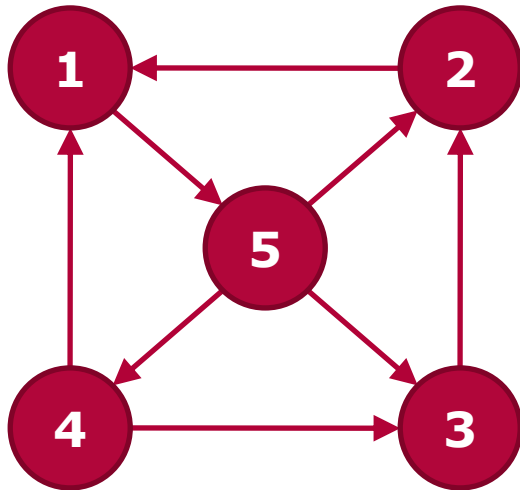
- Ankertext ist sehr hilfreich, um Webseiten zu charakterisieren
 - Google-Bombe: "miserable failure" verlinkte auf President Bush

- Beide Annahmen treffen offensichtlich nicht immer zu
 - Aber beides anzunehmen führt zu viel besseren Websuchergebnissen als es nicht anzunehmen
- Wir konzentrieren uns auf Annahme 1: „**Links** sind Ausdruck von **Qualität**“
 - Zwei populäre Algorithmen, die auf dieser Annahme beruhen
 - HITS (Kleinberg 1999)
 - PageRank (Page et al. 1998)
- Idee bei **HITS**:
 - Für **jede Anfrage**, berechne für **jede Webseite** einen **Hub-Score** und einen **Authority-Score**
 - **Hub**: Eine gute **Sammlung von Links** zu relevanten Webseiten
 - **Authority**: Eine **relevante Webseite**
- Idee bei **PageRank**:
 - Berechne einen **Prestigewert** pro Seite **unabhängig** von Anfragen

- Das **Random-Surfer-Modell** modelliert eine Person, die im Web surft mit Hilfe eines sogenannten **Random Walk**:
 - In 90% aller Fälle klickt der Random-Surfer auf einen beliebigen Hyperlink auf der Webseite auf der er sich gerade befindet
 - In 10% der Fälle gibt der Surfer eine zufällige URI ein
 - PageRank=Die langfristige Besucherrate jeder Seite
- Sehr grobes Modell eines Webnutzers, aber nützlich
 - Passt gut zu der Idee von Seeley $p(u) = \alpha \sum_{v \rightarrow u} p(v)$
 - Eine Seite, auf der langfristig viele Benutzer vorbei kommen, hat ein hohes Prestige
 - Jede Webseite erhält einen Prestigewert unabhängig von der Anfrage
 - Entspricht nicht ganz der Realität:
 - Niemand folgt Links zufällig, sondern meist themenorientiert
 - Wie surft man zu einer zufälligen Seite?
 - Was ist mit dem „Zurück“-Knopf und Lesezeichen?

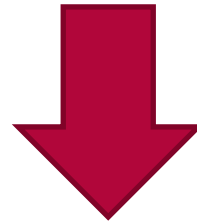
- Eine etwas formale Version:
 1. Starte auf einer zufälligen Seite
(alle Seiten haben die gleiche Wahrscheinlichkeit)
 2. Werfe eine Münze, die mit der Wahrscheinlichkeit λ Zahl zeigt
 3. Falls die Münze Kopf zeigt und die aktuelle Seite Links enthält:
 - Folge zufällig einem der Links
 - Fahre mit Punkt 2. fort
 4. Falls die Münze Zahl zeigt oder die aktuelle Seite keine Links enthält
 - Surfe zu einer zufälligen Seite
 - Fahre mit Punkt 2. fort

■ Beispiel:



Adjazenzmatrix:

A	1	2	3	4	5
1					1
2	1				
3		1			
4	1		1		
5		1	1	1	

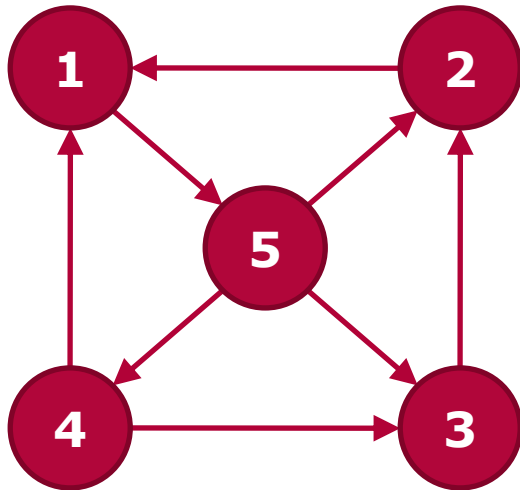


Setze $\lambda = 0.25$

Übergangsmatrix:

T	1	2	3	4	5
1	0.05	0.05	0.05	0.05	0.75 + 0.05
2	0.75 + 0.05	0.05	0.05	0.05	0.05
3	0.05	0.75 + 0.05	0.05	0.05	0.05
4	0.375 + 0.05	0.05	0.375 + 0.05	0.05	0.05
5	0.05	0.25 + 0.05	0.25 + 0.05	0.25 + 0.05	0.05

■ Beispiel:



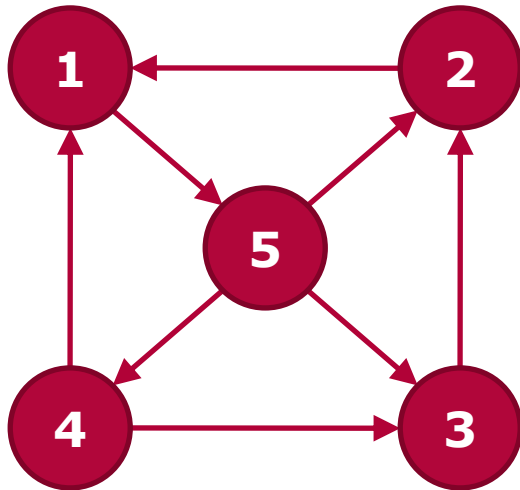
Übergangsmatrix:

T	1	2	3	4	5
1	0.05	0.05	0.05	0.05	0.8
2	0.8	0.05	0.05	0.05	0.05
3	0.05	0.8	0.05	0.05	0.05
4	0.425	0.05	0.425	0.05	0.05
5	0.05	0.3	0.3	0.3	0.05

■ Falls der Surfer zum Zeitpunkt t auf Seite 3 ist

- Dann wird er/sie zum Zeitpunkt $t+1$ mit einer Wahrscheinlichkeit von 5% auf Seite 1 sein
- Mit 80% auf Seite 2
- Mit jeweils 5% auf Seite 3,4,5

■ Beispiel:



Übergangsmatrix:

T	1	2	3	4	5
1	0.05	0.05	0.05	0.05	0.8
2	0.8	0.05	0.05	0.05	0.05
3	0.05	0.8	0.05	0.05	0.05
4	0.425	0.05	0.425	0.05	0.05
5	0.05	0.3	0.3	0.3	0.05

	1	2	3	4	5
$t = 0$	1	0	0	0	0

■ Simulation:

- Falls wir auf Seite 1 starten, was ist die Wahrscheinlichkeit nach t Schritten auf Seite i zu landen?

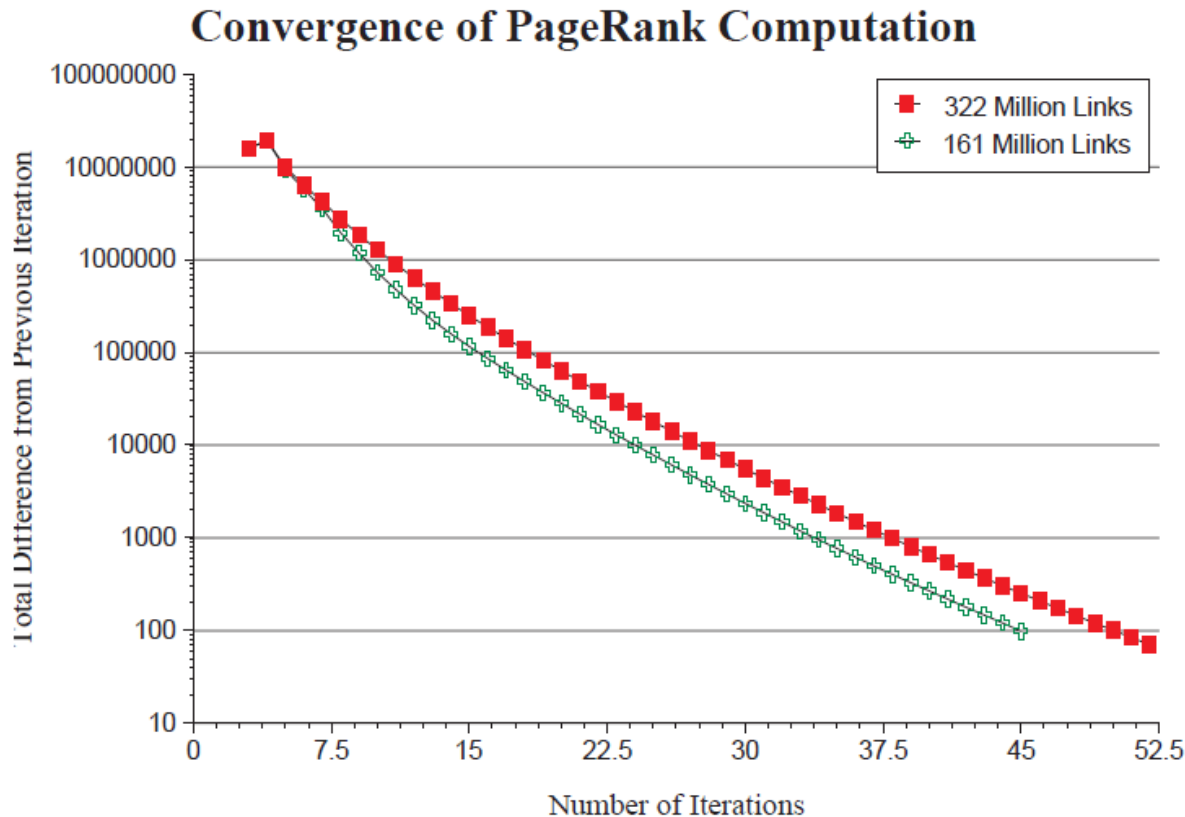
Der Wahrscheinlichkeitsvektor scheint zu konvergieren...

- PageRank wurde von Larry Page und Sergey Brin, den Gründern von Google, 1996 an der Universität Stanford entwickelt
- PageRank wurde als US-Patent mit der Nummer 6285999 patentiert:
 - "Method for node ranking in a linked database"
 - Auch die Berechnung etc. sind patentiert
 - Das Patent ging an die Stanford Universität (nicht an Google)
 - Google hat exklusive Nutzungsrechte
 - Stanford hat 1.8 Millionen Aktienanteile von Google für die Nutzungsrechte bekommen
 - Stanford hat diese Anteile 2005 für 336 Millionen Dollar verkauft



- Wie kann man den PageRank nun berechnen?
- Einfachste Art: **Potenzmethode**
 - Beruht auf mathematischen Erkenntnissen von **Eigenwerten** und **Eigenvektoren** von **Übergangsmatrizen**
- Konkreter Algorithmus:
 1. Starte mit einem beliebigen initialen Vektor b_0
 2. Setze $i = 0$
 3. Setze $b_{i+1} = A \cdot b_i$
 4. Setze $b_{i+1} = b_{i+1}/|b_{i+1}|$, normalisiere b_{i+1} auf Einheitslänge
 5. Setze $i = i + 1$
 6. Falls nicht konvergiert, gehe zu 2.

- Man kann zeigen, dass die Potenzmethode zum Eigenvektor von A für den betragsmäßig größten Eigenwert konvergiert
- Wie viele Iterationen sind notwendig?
 - Nicht viele...insbesondere, da wir kein perfektes Ergebnis benötigen



- Wie berechnet man den PageRank für Billionen von Seiten?
 - Verteilt auf sehr viele Maschinen
 - Details nennt Google nicht
- Nochmal zum Nachlesen der Details des Algorithmus
<https://webworkshop.net/pagerank.html>
- Suchmaschinenmythos:
„PageRank ist die wichtigste Komponente beim Ranking von Google“
- Die Wahrheit:
 - PageRank nur **einer von sehr vielen** Rankingfaktoren
 - Gerüchte besagen gar, dass PageRank in seiner originalen Form gar nicht mehr eingesetzt wird
 - **Varianten** aber höchstwahrscheinlich im Einsatz
 - **Link-Spam** größte Herausforderung