Crawling

Crawling - Ziele

- Vorhandensein einer Seite
- Inhalt einer Seite
- Links zu anderen Seiten
- Links von anderen Seiten
- Andere Seiten, auf die ähnlich gelinkt wird wie auf diese Seite
- Andere Seiten, die ähnliche Links haben
- Andere Seiten, die ähnlichen Inhalt haben
- ...
- Bereitstellung all dieser Information für weitere Anwendungen, wie Information Retrieval, Clustering, Nextlinks, etc.

Crawling - Regeln

- Bad bots vs. Good bots
  - robots.txt beachten!
  - Server nicht überlasten!
  - Eine Seite nur einmal herunterladen!

    (leider wird Aktualisierungsdatum nicht immer korrekt gesetzt)
  - Bilder und anderen Multimedia content nicht herunterladen
  - Linkfarmen ignorieren!
  - Nur das crawlen, was man benötigt (nicht im Englischen Web wildern, wenn man Deutsch haben will)
  - Stetig crawlen (Information verändert sich garantiert schneller als man crawlen kann)

Gutes schnelles Crawling

- Nicht Menge der Information ist das (größte) Problem, sondern Menge verschiedener Fehler und Fallen!
- Whitelist / Blacklist
- Datenverwaltung auf Crawlingserver ist ebenfalls Problematisch.
- Ein gefundener Link bereits heruntergeladen? Hash? Liste?
- Gutes Crawling benötigt aufwendige Logik bezüglich Linkfarmen, dynamischer URLs (damit nicht der gleiche Inhalt in 20 verschiedenen Sortierformen heruntergeladen wird), usw.
- Man kann nicht jede Seite erwischen, aber die wichtigen sollte man erwischen!
- Nur, wie erkennt man „wichtig“?

Crawling Strategien

- Breadth-first
- Depth-first
- Random-ordering
- Omniscient-ordering
- OPIC (one line page importance crawling (cash per page))
- Backlink-count
- Partial Page-rank
- Path-ascending crawl (einfach von hinten chunks im path weglassen)
- Deep web crawling (Daten hinter HTML-Formularen)
- Larger-sites-first
- Was unterscheidet in dieser Hinsicht eigentlich einen guten Crawl von einem schlechten?

Crawling - Konzepte

- Parallel Crawling

Independent

Each crawling process starts with its own set of seed URLs and follows links without consulting with other crawling processes.

Dynamic Assignment

There exists a central coordinator that logically divides the Web into small partitions and dynamically assigns each partition to a crawling process for download

Static Assignment

The Web is partitioned and assigned to each C-proc before they start to crawl

Crawling - Konzepte

- Parallel Crawling
  - Intra-site parallel crawler
  - Distributed Crawling

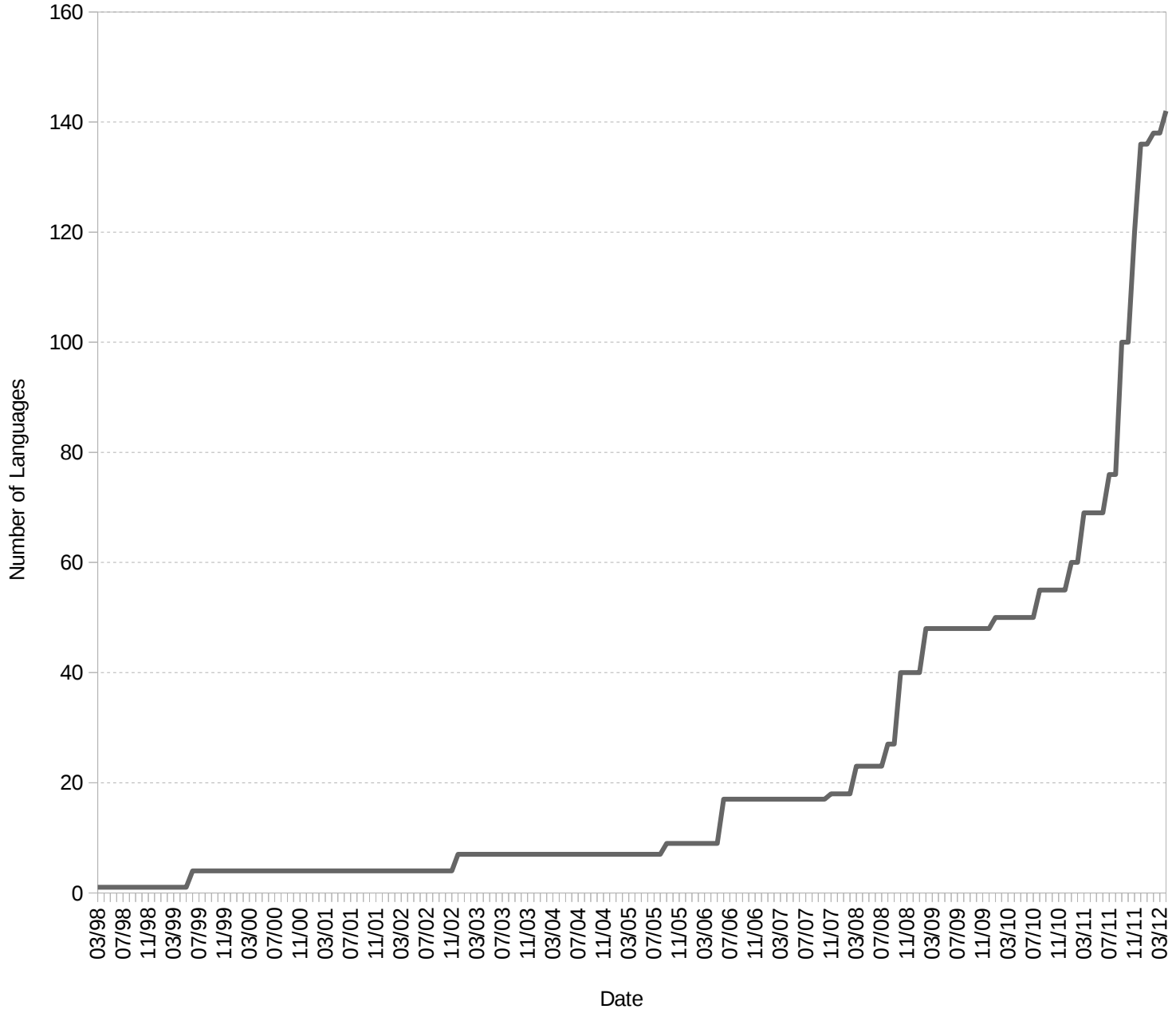
- Polite crawling

Crawling - Update Policy

- Uniform crawling
- Proportional policy
- Classes of Sites according their change-frequency

Crawling - Praktische Anwendung

Wortschatz


160
140
120
100
80
60
40
20
0
Number of Languages
03/98
07/98
11/98
03/99
07/99
11/99
03/00
07/00
11/00
03/01
07/01
11/01
03/02
07/02
11/02
03/03
07/03
11/03
03/04
07/04
11/04
03/05
07/05
11/05
03/06
07/06
11/06
03/07
07/07
11/07
03/08
07/08
11/08
03/09
07/09
11/09
03/10
07/10
11/10
03/11
07/11
11/11
03/12
Date

Sammeln von Text

- Hauptquelle: Internet
- Beispiele: „ganzes“ Internet (FindLinks), einzelne TLDs (Heritrix), einzelne Domains (UDHR, Bibeln, Wikipedia), News (RSS-News, AbyZnewslink+Heritrix)
  - Welche wünschenswerten Eigenschaften erfüllen die Quellen (Umfang, Kontinuität, Textqualität, Sachgebiete, Sprachen, Homogenität)?
  - Welche typischen Verarbeitungsschritte gibt es? Welche sind jeweils nötig?
  - Welche typischen Fehlerklassen werden auftreten?

Sammeln von Text

- Schritte:
  - URLs beschaffen
  - Webseiten herunterladen
  - HTML-Stripping
  - Spracherkennung
  - ---
  - Satzsegmentierung
  - Tokenisierung
  - Putzen etc.

Bootstrapping-Ansatz

- Texte einer Sprache sammeln
- Voraussetzung: Wortliste

- Vorgehen:
  - Bilden von Wort-Tupeln
  - Anfrage an Suchmaschine (API) senden
  - URLs sammeln
  - Webseiten herunterladen
  - Sprachseparierung

Bootstrapping-Ansatz

- Bsp.: Schwedisch
- 1. Wortliste
- Quellen:
  - UDHR
  - Watchtower
  - Bibeln

att	kan
och	om
som	ett
i	du
det	sig
en	eller
är	man
av	var
på	Det
för	här
de	jag
till	så
med	han
har	...
den	
inte	

Bootstrapping-Ansatz

- Bsp.: Schwedisch
- 2. Liste von Tupeln
  - Suchmaschine anfragen

strid vrede litteratur
inre inför bättre
felaktigt skapelse krets
maktlösa under trupper
ormar kroppen Hjälp
fortsätter art svaghet
snabba avhålla benägenhet
irriterad naturen följden
mördade frihet bergen
väderkvarnarna trevligt tyska
mera Portugal turister
universitetet föder förödande
berättelserna förvänta granskning
förhållande anhängare ren
Dina Dagen reagerade
Stress plats syster
via resultaten inför
närmare Report Resultatet
titlar nämndes händer
trygg ockultism barnet
tror stödde militärtjänst
hamna mådde utövar

Bootstrapping-Ansatz

- Bsp.: Schwedisch
- 3. Liste von URLs
  - Downloaden

http://www.eyebeep.se/
http://www.eye-do.se/synenpajobbet.html
http://www.eyedrop.net/bullegrims/nytt.asp
http://www.eyefeed.se/produktioner.htm
http://www.eyeline.se/?paged=4
http://www.eyeline.se/?tag=mat
http://www.eyenetsweden.se/page/45/2005.aspx
http://www.eyenetsweden.se/page/61/2008.aspx
http://www.eyescreamtattoo.com/faq.html
http://www.eyewitness.no/Teologi/johannes_evange.htm
http://www.ezcap.se/
http://www.ezdravlje.org/sr-Latn-BA/zdravlje-az/
http://www.ez-essays.com/free/2315.html
http://www.ezz.dk/132463-antal-lande-i-europa
http://www.ezz.dk/170898-ol-og-elefanter
http://www.ezz.dk/196239-norton-internet-security-fejl-ved
http://www.ezz.dk/200571-dr-licens
http://www.ezz.dk/232409-hvilke-medier-til-pioneer-dvd
http://www.ezz.dk/233779-udrensningskure
http://www.ezz.dk/235785-risengrod
http://www.ezz.dk/274615-creme-brulee-braender
http://www.ezz.dk/284785-danmarks-forste-frimaerke

Bootstrapping-Ansatz

- Bsp.: Schwedisch
- 4. Webseiten
  - 4.1 HTML-Stripping

<source><location>http://www.eyenetsweden.se/page/45/2005.aspx</location></source>
Personal
Forskningssekreterare Kristin Svensson har under året arbetat växelvis mellan 80 och 100 %. Två forskningssköterskor, Susanne Albrecht och Eva Wendel, har arbetat 80 %. Chefen för verksamheten, Mats Lundström, har arbetat mellan 80 och 90 % på enheten. Av denna arbetstid har EyeNet Sweden finansierat 25 %. Övrig tid har finansierats genom Landstinget Blekinge och enskilda projekt. Sammanlagt har de centrala anslagen till EyeNet Sweden finansierat 2,25 tjänster av 3,50 bemannade.

<source><location>http://www.ezz.dk/170898-ol-og-elefanter</location></source>
Kære gruppe
En svensk ven har sendt mig nedenstående anekdote, som han ønsker be- eller afkræftet. Kan nogle af jer bidrage?
Elefantöl. Vad har Carlsbergs bryggerier med elefanten att göra? Carl Jacobsen, död 1914, son till grundaren av Carlsbergs Bryggerier, och hans hustru hade svårt att få barn. Bl.a. förekom havandeskap med missfall. I misströstan över läkarvetenskapen köpte Jacobsen i samband med sin hustrus aktuella havandeskap en elefant.

Bootstrapping-Ansatz

- Bsp.: Schwedisch
- 4. Webseiten
  - 4.2 Sprachseparierung

<source><location>http://www.eyenetsweden.se/page/45/2005.aspx</location><language>swe</language></source>
Personal
Forskningssekreterare Kristin Svensson har under året arbetat växelvis mellan 80 och 100 %. Två forskningssköterskor, Susanne Albrecht och Eva Wendel, har arbetat 80 %. Chefen för verksamheten, Mats Lundström, har arbetat mellan 80 och 90 % på enheten. Av denna arbetstid har EyeNet Sweden finansierat 25 %. Övrig tid har finansierats genom Landstinget Blekinge och enskilda projekt. Sammanlagt har de centrala anslagen till EyeNet Sweden finansierat 2,25 tjänster av 3,50 bemannade.

<source><location>http://www.ezz.dk/170898-ol-og-elefanter</location><language>swe</language></source>
Kære gruppe
En svensk ven har sendt mig nedenstående anekdote, som han ønsker be- eller afkræftet. Kan nogle af jer bidrage?
Elefantöl. Vad har Carlsbergs bryggerier med elefanten att göra? Carl Jacobsen, död 1914, son till grundaren av Carlsbergs Bryggerier, och hans hustru hade svårt att få barn. Bl.a. förekom havandeskap med missfall. I misströstan över läkarvetenskapen köpte Jacobsen i samband med sin hustrus aktuella havandeskap en elefant.

Bootstrapping-Ansatz

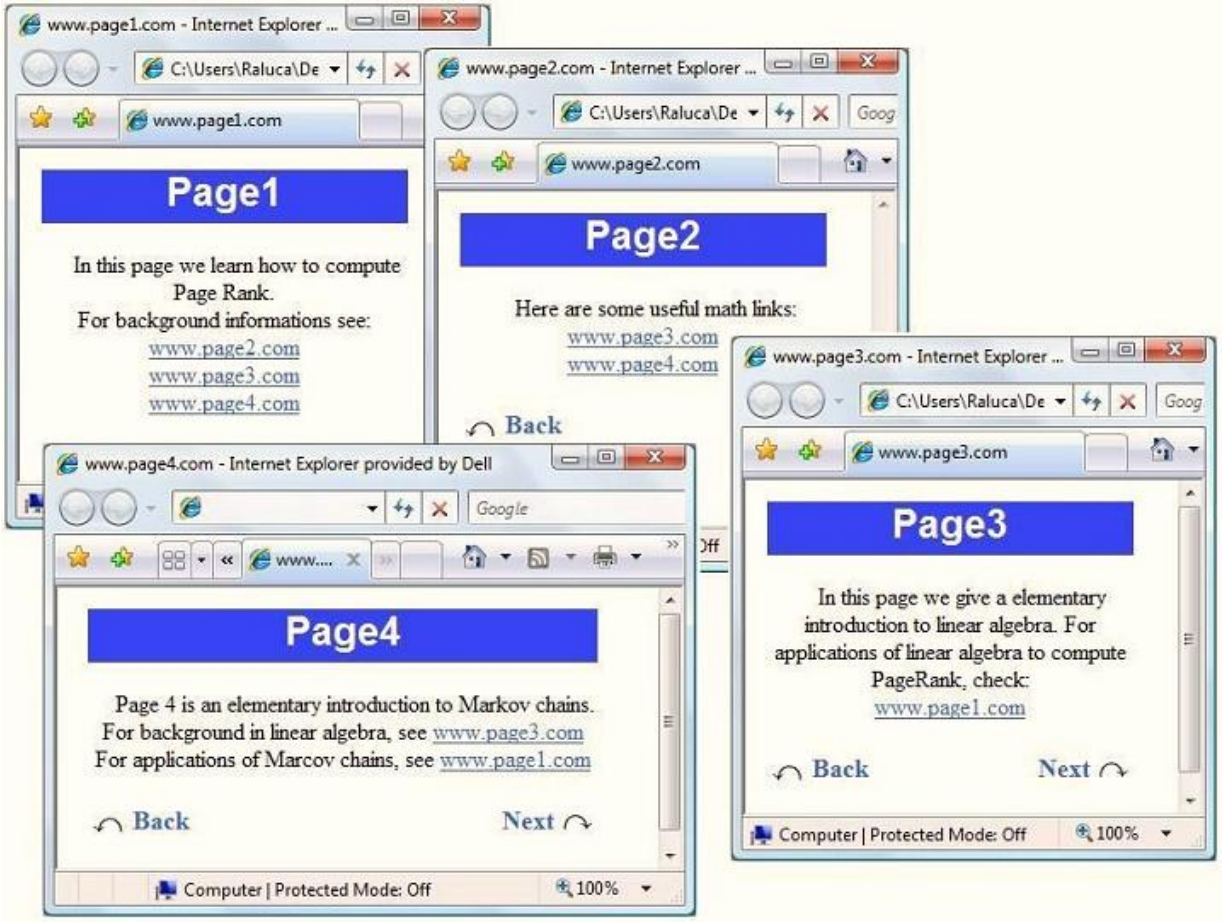
- Probleme:
  - Hoher Anteil anderer Sprachen wird heruntergeladen – „kleine Sprachen“
  - Gewünschte Dokumente fallen aus den von der API zurückgegebenen top-x Dokumenten heraus

-Rank

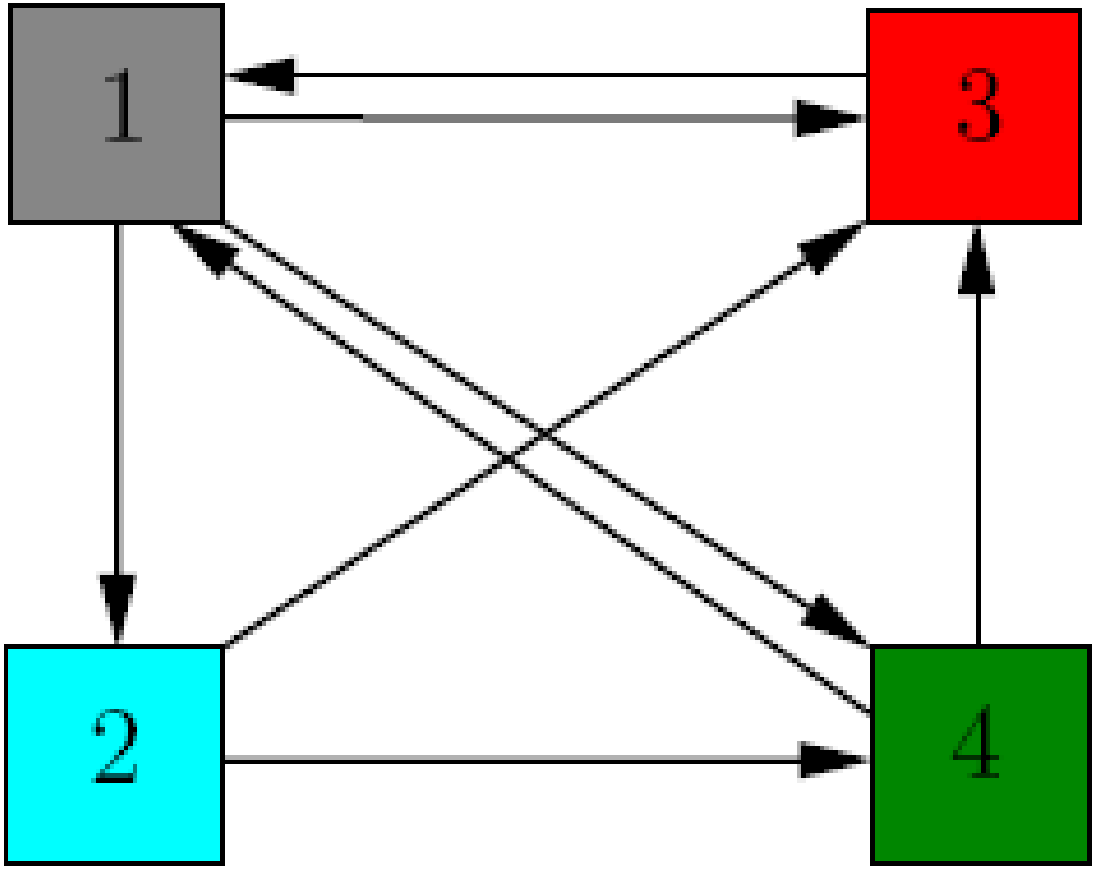
PageRank

- Ursprüngliche Idee (Brin & Page 98 „The anatomy of a large-scale hypertextual Web search engine “) ist, dass eine Seite umso wichtiger ist, je mehr Links darauf zeigen
- Verfeinerung: Je wichtiger eine Seite ist, umso wichtiger der Link.
- D.h. ein einziger Link einer wichtigen Seite kann mehr wert sein, als 1000 Links von unwichtigen Seiten (Random surfer)
- Das heisst aber auch, dass die Berechnung rekursiv ist!

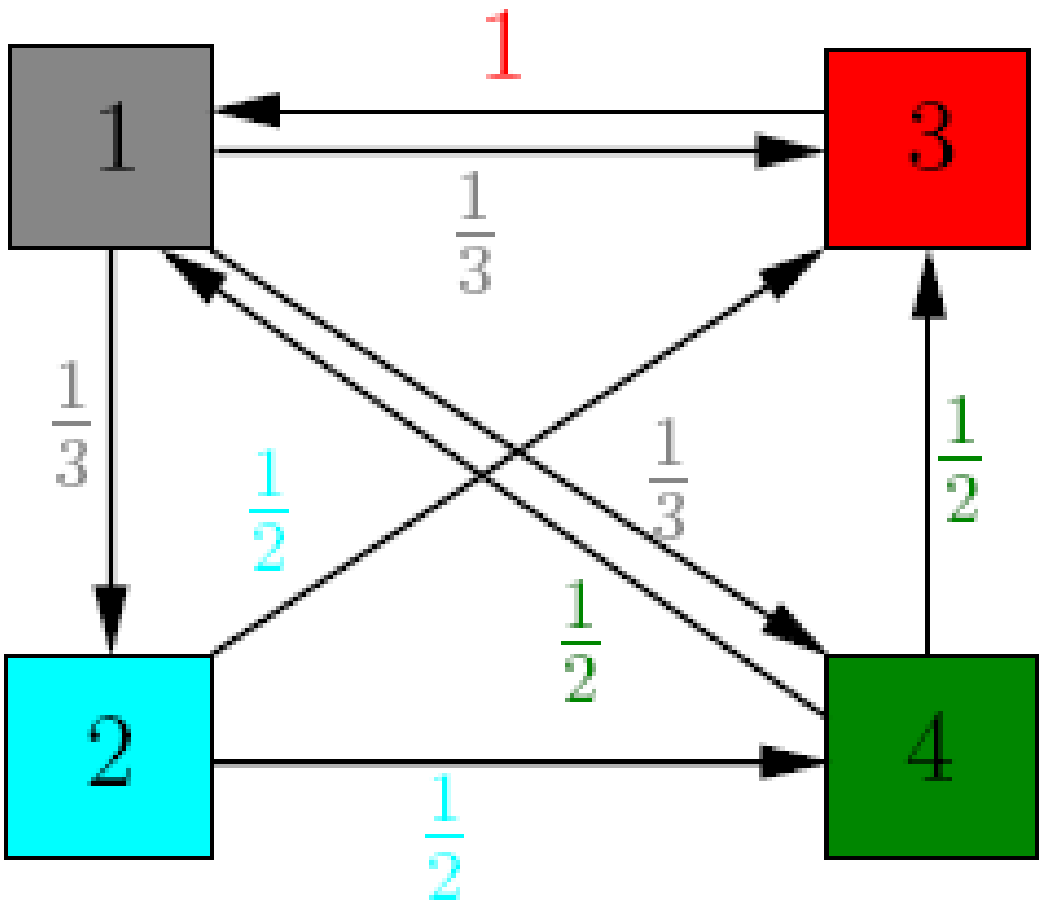
PageRank - Beispiel



PageRank - Beispiel



PageRank - Beispiel



PageRank - Beispiel

- Übergangsmatrix

$$\begin{bmatrix} 0 & 0 & 1 & \frac{1}{2} \\ \frac{1}{3} & 0 & 0 & 0 \\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{1}{2} & 0 & 0 \end{bmatrix}$$

PageRank - Beispiel

- Initialisierung

$$\prod(0) = \begin{pmatrix} 0{,}25 \\ 0{,}25 \\ 0{,}25 \\ 0{,}25 \end{pmatrix}$$

PageRank - Beispiel

$$\prod(0)=\begin{pmatrix}0{,}25\\0{,}25\\0{,}25\\0{,}25\end{pmatrix}$$

$$\begin{bmatrix}0 & 0 & 1 & \frac{1}{2}\\ \frac{1}{3} & 0 & 0 & 0\\ \frac{1}{3} & \frac{1}{2} & 0 & \frac{1}{2}\\ \frac{1}{3} & \frac{1}{2} & 0 & 0\end{bmatrix}$$

$$\pi_1 = 1\,\pi_3 + \frac{1}{2}\pi_4$$

$$\pi_2 = \frac{1}{3}\pi_1$$

$$\pi_3 = \frac{1}{3}\pi_1 + \frac{1}{2}\pi_2 + \frac{1}{2}\pi_4$$

$$\pi_4 = \frac{1}{3}\pi_1 + \frac{1}{2}\pi_2$$

$$\pi_1 + \pi_2 + \pi_3 + \pi_4 = 1$$

PageRank - Beispiel

$$\prod(0)=\begin{pmatrix}0{,}25\\0{,}25\\0{,}25\\0{,}25\end{pmatrix}$$

$$\prod(1)=\begin{pmatrix}1*0{,}25+\frac{1}{2}*0{,}25\\\frac{1}{3}*0{,}25\\\frac{1}{3}*0{,}25+\frac{1}{2}*0{,}25+\frac{1}{2}*0{,}25\\\frac{1}{3}*0{,}25+\frac{1}{2}*0{,}25\end{pmatrix}=\begin{pmatrix}\frac{3}{8}\\\frac{1}{12}\\\frac{1}{3}\\\frac{5}{24}\end{pmatrix}$$

$$\pi_1=1\,\pi_3+\frac{1}{2}\,\pi_4$$

$$\pi_2=\frac{1}{3}\,\pi_1$$

$$\pi_3=\frac{1}{3}\,\pi_1+\frac{1}{2}\,\pi_2+\frac{1}{2}\,\pi_4$$

$$\pi_4=\frac{1}{3}\,\pi_1+\frac{1}{2}\,\pi_2$$

PageRank - Beispiel

$$\prod(x) = \begin{pmatrix} 0{,}38 \\ 0{,}13 \\ 0{,}29 \\ 0{,}19 \end{pmatrix}$$

$$\pi_1 = 1\,\pi_3 + \frac{1}{2}\,\pi_4$$

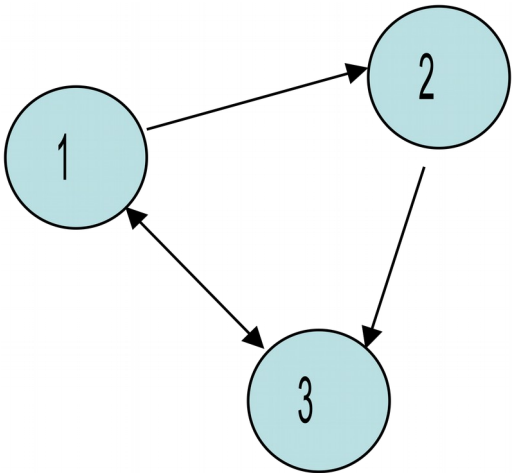
$$\pi_2 = \frac{1}{3}\,\pi_1$$

$$\pi_3 = \frac{1}{3}\,\pi_1 + \frac{1}{2}\,\pi_2 + \frac{1}{2}\,\pi_4$$

$$\pi_4 = \frac{1}{3}\,\pi_1 + \frac{1}{2}\,\pi_2$$

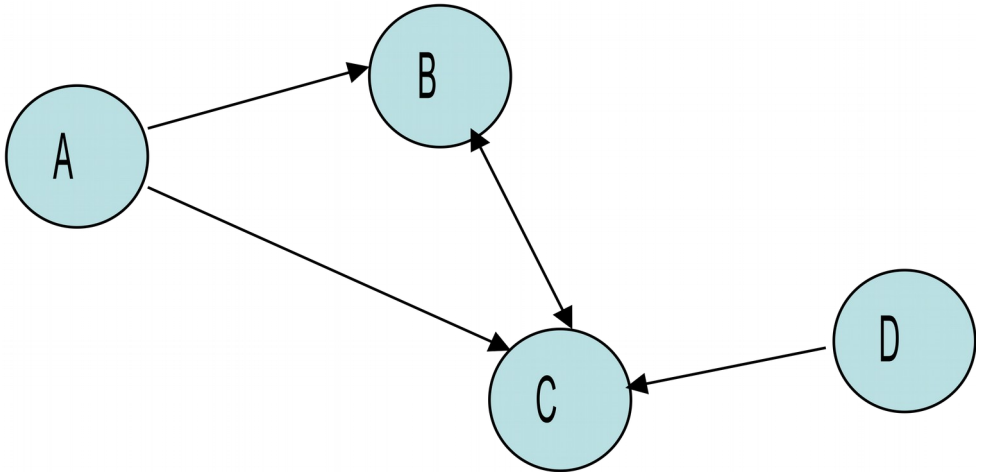
PageRank - Beispiel

- Nun bitte Sie:



PageRank

- Seiten ohne eingehenden Link: Was tun?



PageRank

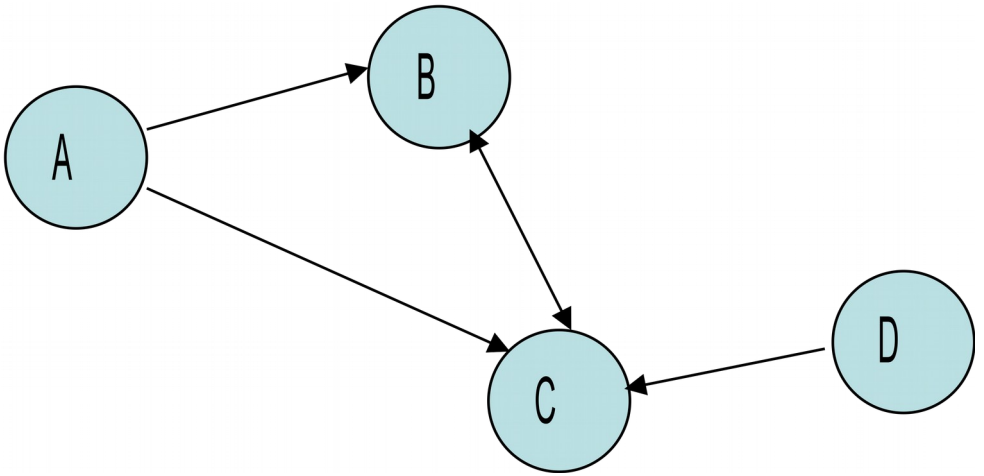
- Random Surfer:
  - In 85% der Fälle wird einem Link gefolgt
  - In 15% der Fälle wählen wir zufällig irgendeine URL

$$\pi_A = 0{,}15 * \frac{1}{4}$$

$$\pi_B = 0{,}15 * \frac{1}{4} + 0{,}85 * (\frac{1}{2}\pi_A + \pi_C)$$

$$\pi_C = 0{,}15 * \frac{1}{4} + 0{,}85 * (\frac{1}{2}\pi_A + \pi_B + \pi_D)$$

$$\pi_D = 0{,}15 * \frac{1}{4}$$



Danke

- Nächster Termin:
  - Ausschließlich Zusammenfassung / Wiederholung einiger Schwerpunkte
  - Wunschthemen Ihrerseits per Mail an teckart@informatik...