

Information Engineering 1: Information Retrieval

Kapitel 6

Web Search

M. Braschler

(basierend u.a. auf Material von J. Savoy)

"Themenkarte" Websearch

- Fallstudie CHerCHer
- Das Web in Zahlen
- Suche im Web
- Entwicklung und Anatomie von Suchmaschinen
- Spidering
- Indexierung Webseiten / Anfrageverarbeitung
- Evaluation
- Spreading Activation (SA)
- Indegree
- PageRank
- HITS (Kleinberg Model)
- Link-Analyse
- Appendix: Small Data

Lernziel Kapitel

- Web Search ist der Use-Case in IR, den alle kennen (Google...)
- Sie sollen verstehen,
 - .. inwiefern die bereits behandelte Theorie in diesem Umfeld zum Zuge kommt
 - .. was die Besonderheiten des Web Search sind
 - .. inwieweit die speziellen Erkenntnisse auf andere Umfelder übertragen werden können

Fallstudie: CHerCHer

- Das fiktive Schweizer Suchmaschinenunternehmen CHerCHer betreibt eine Suchmaschine, die vollständig aus Werbeeinnahmen finanziert werden soll. Die CHerCHer-Technologie basiert auf preiswerter PC-Hardware. Ein CHerCHer-PC ist fähig, Indizes mit 100 Millionen Webseiten zu verarbeiten. Durch Clustering der PC's wird es möglich, enorme Datenmengen verarbeiten zu können. Neben den Such-PC's ist eine Anzahl von PC's vorhanden, die sich ausschliesslich um gepufferte Antworten von gebräuchlichen Anfragen kümmern.
- Der CHerCHer-Spider crawlt das gesamte Web ab. Die folgenden Folien zeigen das voraussichtliche Budget und Unterhaltskosten für das kommende Jahr.

Ursprüngliches Beispiel von D. Hawking, CSIRO zitiert nach J.Savoy.

Technische Daten von CHerCHer

- Indexgrösse: 30 Milliarden Webseiten
- Durchschnittliche Seitengrösse: 15 Kbytes
- Ertrag pro Anfrage: 0.25 Cents
- Durchschnittliche Anzahl Anfragen pro Tag: 20 Millionen
- Spitzenwert Anzahl Anfragen pro Tag: 100 Millionen
- Durchschnittliche Dauer einer vollständigen Anfrage: 0.5 sek.
- Dauer um eine gepufferte Anfrage zurückzuliefern: 0.001 sek. (Cluster)
- Anteil von Anfragen, die gepuffert werden: 35%
- Kosten für einen Standard-PC: €300 (jährliche Leasingkosten)
- Netzwerkkosten: €20 per Terabyte
- Spidering-Budget: €1 Million
- Fixkosten (z.B. Gehälter, Miete, Ferrari für CEO): €2.5 Millionen

Probleme / Fragen

- Frage 1: Wie viel Netzwerkkosten fallen bei einem vollständigem Spidern an?
 - $30 \cdot 10^9 \text{ Webseiten} \cdot 15 \text{ Kb} = 450 \cdot 10^9 \text{ Kb} \rightarrow 450 \text{ TB} = \text{€}9000$
- Frage 2: In was für einem Intervall kann gespideret werden um innerhalb des Budgets zu bleiben?
 - $1 \text{ Million} / 9000 = 111/\text{Jahr}$ oder ca. alle 3 Tage
- Frage 3: Wie viele PC's werden gebraucht, um einen Cluster aufzubauen, der die gesamte Datenmenge verarbeitet werden kann?
 - $1 \text{ PC} = 100 \text{ Mill Webseiten}$, Total 30 Mrd Webseiten $\rightarrow 300 \text{ PC/Cluster}$
- Frage 4: Wie viele PC's werden gebraucht um mit ungepufferten Anfragen in Spitzenzeiten umgehen zu können?
 - $100\text{M} \cdot 65\% \cdot 0.5 \text{ sec} = 32.5 \text{ Msec} \rightarrow 377 \text{ Tage für 1 Cluster}$
Wir brauchen 377 Clusters = $377 \cdot 300 \text{ PC's} = 113,100 \text{ PC's}$
 - Google Schätzung schon im Jahr '06: bis >450,000 PCs

Probleme / Fragen

- Frage 5: Wie viele PC's werden für die gepufferten Anfragen in Spitzenzeiten gebraucht?
 - $100M \cdot 35\% \cdot 0.001 \text{ sec} = 35 \text{ Ksec} \rightarrow 9.7 \text{ h für 1 Cluster}$
wir brauchen 1 Cluster = 300 PCs
- Frage 6: Wie hoch sind die Hardwarekosten?
 - $113,400 \text{ PC} \cdot € 300 = € 34M$
- Frage 7: Wie hoch ist der voraussichtliche Erlös?
 - $€0.0025 \cdot 20 \cdot 10^6 = €50K/day \rightarrow 360days = €18M$
- Frage 8: Was kostet der Strom?
 - 113,400 Server à 8 kWh pro Tag schlucken 27 GWh/Monat: kostet irgendwo in der Grössenordnung von 4-5 Millionen CHF (Stand 2021)
- Frage 9: Wie hoch ist der voraussichtliche Verlust für dieses Jahr?
 - $(34M + 1M + 4M + 2.5M) - 18M = €-23.5M$

Probleme / Fragen

- Frage 10: Was wird der Gewinn/Verlust für CHerCHer sein, wenn folgenden Vorschläge umgesetzt werden?
- Verwende grössere und teurere (€3000 pro Jahr) PC's für gepufferte Anfragen, um den Anteil der gepufferten Anfragen auf 50% zu steigern. (Aber: realistisch?)
 - $100M \cdot 50\% \cdot 0.001 \text{ sek.} = 50 \text{ Ksek.} \rightarrow 13.89 \text{ h für 1 Cluster}$
 wir brauchen 1 Cluster = 300 PCs $\rightarrow \text{€}0.9 \text{ M}$
 $100M \cdot 50\% \cdot 0.5 \text{ sek.} = 25 \text{ Msek.} \rightarrow 290 \text{ Tage für 1 Cluster}$
 wir brauchen $290 \cdot 300 \text{ PCs} = 87,000 \text{ PCs} \rightarrow \text{€}26.1 \text{ M}$
 Anfrageverarbeitungs-kosten: $\text{€}26.1 + 0.9 = \text{€} 27M (-7M)$
- Einführung einer Anfrageoptimierung, die die Dauer einer Anfrageverarbeitung auf 0.33 sek. kürzt.
 - $100M \cdot 65\% \cdot 0.33 \text{ sek.} = 21.45M\text{sek.} \rightarrow 249 \text{ Tage}$
 wir brauchen $249 \cdot 300 \text{ PCs} = 74,700 \text{ PCs}$
 Kosten $\rightarrow \text{€}22.41 \text{ M (anstatt €}33.93M \rightarrow -\text{€}11.5M)$

Probleme

- Frage 11: Was könnten die Motivationen für CHerCHer sein um:
 - die Qualität der Suchresultate zu verbessern?
 - häufigere Aktualisierungen des Indexes vorzunehmen?

Das Web in Zahlen..

- Anzahl der Internet-User

61 mio. in 1996

147 mio. in 1998

604 mio. in 2002

1,802 mio. in 2010

2,405 mio. in 2012

4,930 mio. in 2020/Q3

- Weltweite Verteilung der Internet-User (Penetration in %)

728M Europa (87%)

468M Lateinamerika (72%)

2556M Asien (60%)

632M Afrika (47%) (war nur 9% im Jahr 2010!)

333M Nordamerika (90%)

185M Naher/mittlerer Osten (71%)

29M Ozeanien (68%)

(www.internetworldstats.com)

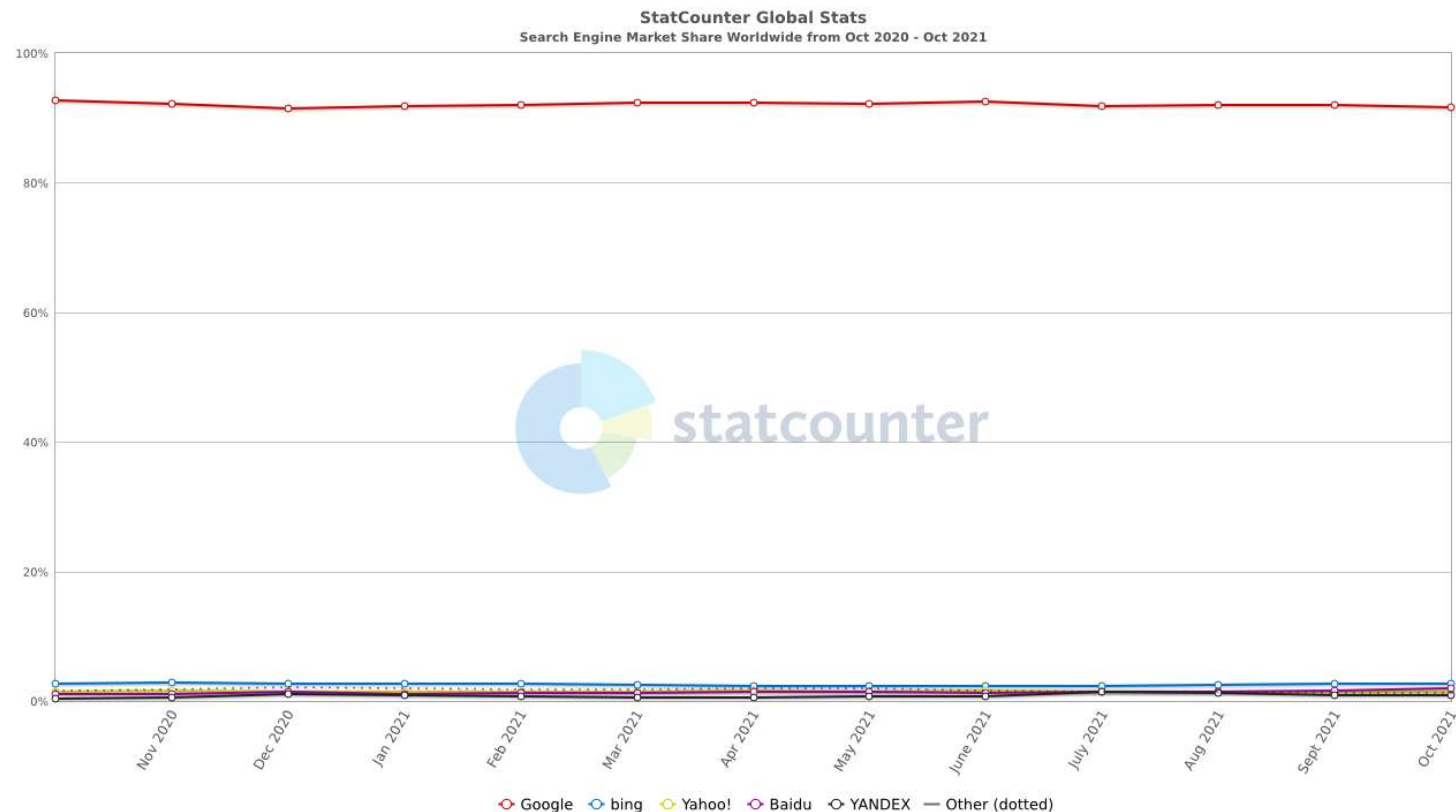
Das Web in Zahlen...

- Volumen der verfügbaren Daten
 - 5% “surface web”
 - 95% “deep web” (dynamisch)
 - unbekannt: “dark web”
- Schätzung von 2015: Google-Index des “surface web” ~14.5 Milliarden Seiten
- Schätzungen variieren dramatisch, oft zwischen 5 und 60 Milliarden Seiten

Marktanteil

Globaler Marktanteil Google Websuche: ~92%

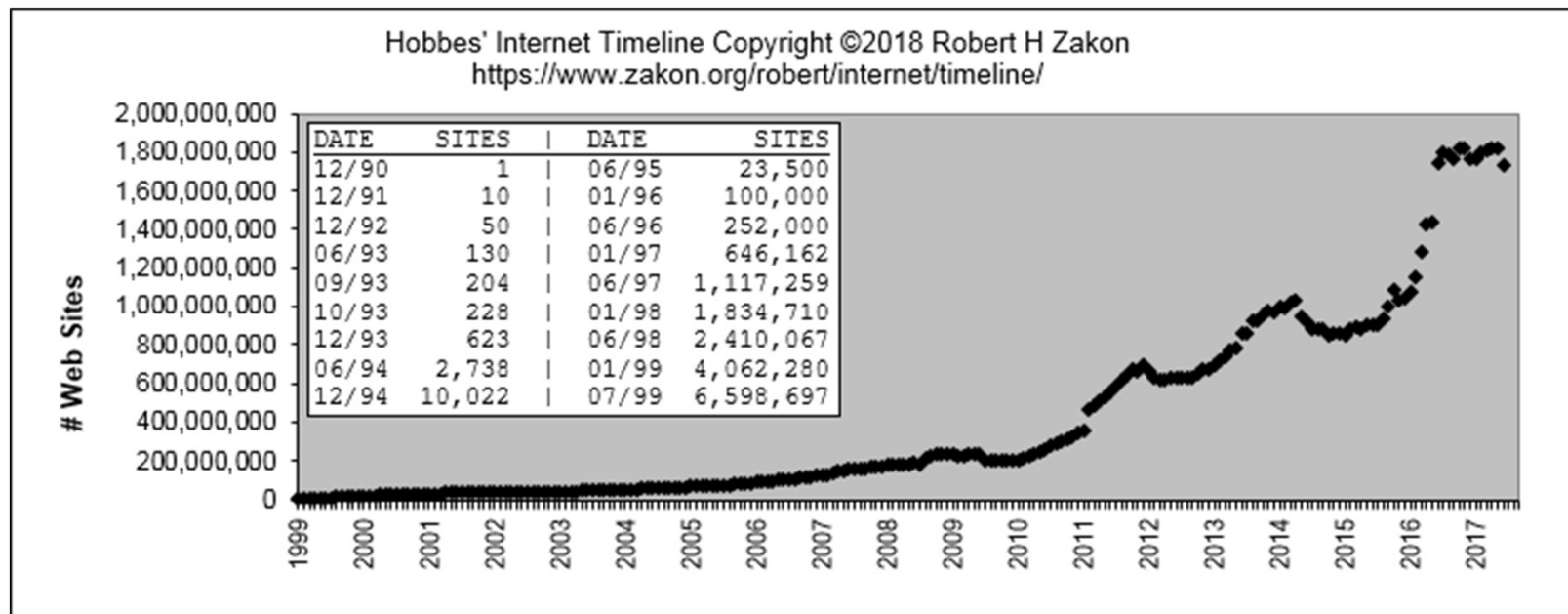
Aber: Marktführer in RU: Yandex, in CN: Baidu, in JP: Yahoo



Das Web in Zahlen...

Anzahl der Websuchen pro Tag (2019)

→ Google verarbeitet 3.5 Milliarden Suchen/Tag
(internetlivestats, 2019)



Das Web in Zahlen...

- Zusammensetzung des “surface web”

Typ	% (totale Dateigrösse)
Bilder	23.2%
HTML	17.8%
PHP	13.0%
Adobe PDF	9.2%
Videos	4.3%
Komprimiert/Archiv	3.7%
Audio	2.3%
Programme	1.4%

Nicht nur Google!

- Verschiedene spezialisierte Suchdienste
 - Allgemein
 - Nachrichten
 - Shopping
 - Für Kinder
 - Fachspezifisch (Medizin, öff. Behörden, Juristisch, QA, Reisen, ...)
 - Bilder/audio/video
 - Metasearch
 - Länderspezifisch
 - Eigene Suchmaschine für eine Website
 - Enterprise Search (Web+Emails+Memos+...)

Nicht nur Google!

- ABER: Überschätzen Sie Web Search nicht!
- Benutzer gehen vermehrt direkt zur gesuchten Websites, als über den Umweg eines Suchdienstleisters.
- IDC behauptet, dass 70% aller Anfragen direkt bei den gesuchten Websites abgesetzt werden (→ Bedeutung Enterprise Search)

(Quelle: "Who Owns the Web? Guess Again", Frank Gens. 21/03/2007, <http://blogs.idc.com/ie/?p=92>)

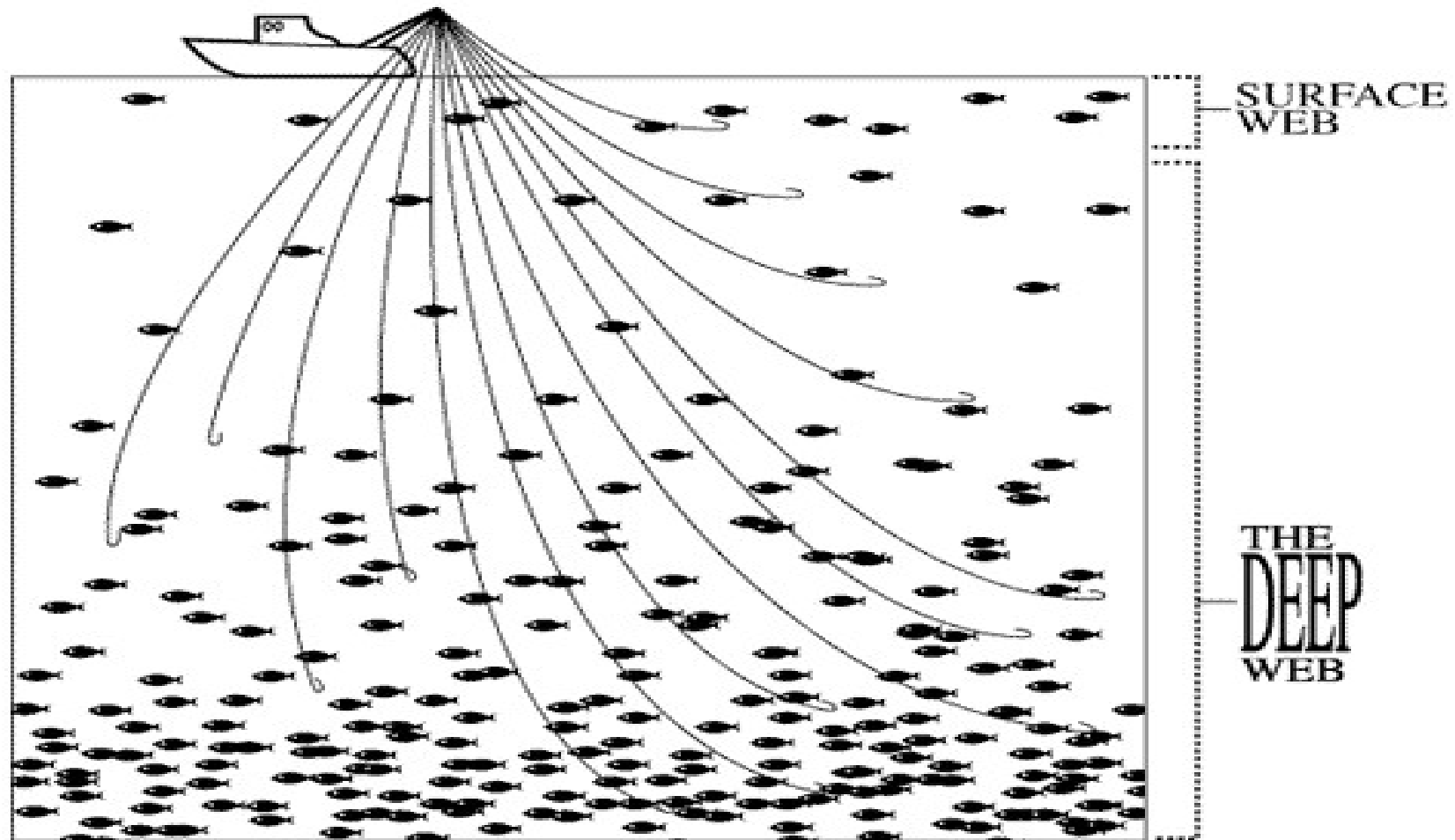
Klassisches IR

- Korpus:
 - Wohldefinierte, statische Dokumentensammlung
 - Zentral, an einem bekannten Ort, gespeichert
- Sammlung ist homogen, typischerweise von hoher Qualität
- Ziel: Dokumente zurückgeben, die Informationen enthalten, welche relevant sind in Hinblick auf das Informationsbedürfnis des Nutzers
- Klassische Definition von Relevanz
- Für jede Query Q und jedes Dokument D der Sammlung existiert nur ein Relevanzscore $RSV(Q, D)$
- Der Score ist ein Durchschnitt über alle User U und Kontexte C
- Wir optimieren also den $RSV(Q, D)$ statt $RSV(Q, D, U, C)$
- → User und Kontext ignoriert, akzeptabel in einem kontrollierten Setting

Suche im Web

- Korpus: Das öffentliche zugreifbare Web besteht aus statischem und dynamischen Inhalt. Die Daten im Web sind sehr unbeständig (40% monatliche Änderung), unstrukturiert, teilweise von schlechter Qualität (Spam!) und/oder heterogen.
- Leute kommen auch zu Google oder Bing um im Web zu navigieren (schwierig den Anbieter direkt zu finden)
- Ziel: Finde qualitativ hoch stehende Resultate (nicht unbedingt Dokumente!), die für den Benutzer relevant sind.
- Resultat:
 - Statische Seiten (Dokumente) z.B. Texte, mp3, Photos, Videos ...
 - Resultat ist nicht ausbeuteorientiert!
 - Ggf. Probleme mit "dynamischen Seiten": diese werden bei einer Anfrage durch den Benutzer generiert. Die Daten werden dynamisch aus einer Datenbank geladen und dargestellt. → „Deep Web“

Deep Web



Quelle: Rapporto tra Deep Web e Surface Web, Silvia Panzavolta, <http://www.indire.it/content/index.php?action=read&id=1303>

Suche im Web

- Bedürfnis (Einteilung nach A. Broder, damals bei Altavista)
 - **informationell**
 - man will etwas lernen (~40%) z.B. "Was ist das Semantic Web?"
 - **navigational**
 - man will zu einer gewissen Webseite (~25%) z.B. "Tschechien Bahn"
 - **transaktional**
 - man will etwas tun (~35%)
 - Zugriff auf einen Service z.B. "Wetter in Zürich"
 - Downloads, z.B. "Oberflächenfotos vom Mars"
 - Shop, z.B. "iTunes"
 - **Graubereiche**
 - Finde einen guten Hub, z.B. "Automiete in Seattle"
 - Erkundungssuche "see what's there"

Entwicklung von Suchmaschinen

- **Erste Generation** – nur einfache Webseiten, Textdateien
 - Worthäufigkeit, Sprache
 - AltaVista, Lycos, Excite
- **Zweite Generation** – „off-page“, web-specific data
 - Hyperlink- oder Verbindungsanalyse
 - Click-through data (Angeklickte Resultate)
 - Ankertext (wie Leute auf diese Webseite verweisen)
 - Google (1998) mit PageRank-Algorithmus
- **Dritte Generation** – Beantwortet “das Bedürfnis hinter der Anfrage” (immer noch im Fluss, die heutigen Systeme setzen diese Idee zunehmend um) (nimmt stärkere Rücksicht auf die Unterscheidung informationell/navigational/-transaktional)

Anatomie von Suchmaschinen

■ **Spider** (Crawler oder Roboter) – bildet den Korpus

- Sammelt die Daten rekursiv
 - Für jede bekannte URL, hole die Webseiten, parse diese und extrahiere die neuen URL's.
- Zusätzliche Daten von direkten Einträgen und anderen Quellen.
- Verschiedene Suchmaschinen haben unterschiedliche Grundsätze – kleine Übereinstimmung unter den Korpora

■ Der **Indexer** – verarbeitet die Daten (inverted files)

- Verschiedene Grundsätze, welche Worte indexiert oder gestemmt werden, Phrasenunterstützung, Grossschreibung, Unicodeunterstützung etc.

■ **Anfrageverarbeitung** – akzeptiere Anfrage und liefere Resultat zurück

- Front end – macht Anfragereformulierung – Stemming, Grossschreib-Regeln, Boolean-Optimierung, Zusammensetzung, etc.
- Back end – findet passende Dokumente und rangiert diese

Spidering

- Starte mit einer umfassenden Menge von URLs, von welchen die Suche (Spidering) gestartet wird (seeds S_0).
- Speichere die Dokumente in D und die Hyperlinks in E . Dabei ist sowohl D als auch E ein eigenständiger Datenbehälter.
- Während dem Crawling wird eine Liste Q von URLs intern unterhalten.
- Wir extrahieren eventuell eine URL von Q oder fügen eine URL Q hinzu (Funktionen `Dequeue()` und `Enqueue()`)
- Wir bezeichnen
 - u, v als eine URL
 - $d(u)$ die zugehörige Webseite

Spidering

- Unser einfacher Webcrawler-Algorithmus:

Simple-Crawler (S_0, D, E)

$Q \leftarrow S_0;$

while $Q \neq \emptyset$ **do**

$u \leftarrow \text{Dequeue}(Q);$

$d(u) \leftarrow \text{Fetch}(u);$

Store ($D, (d(u), u);$

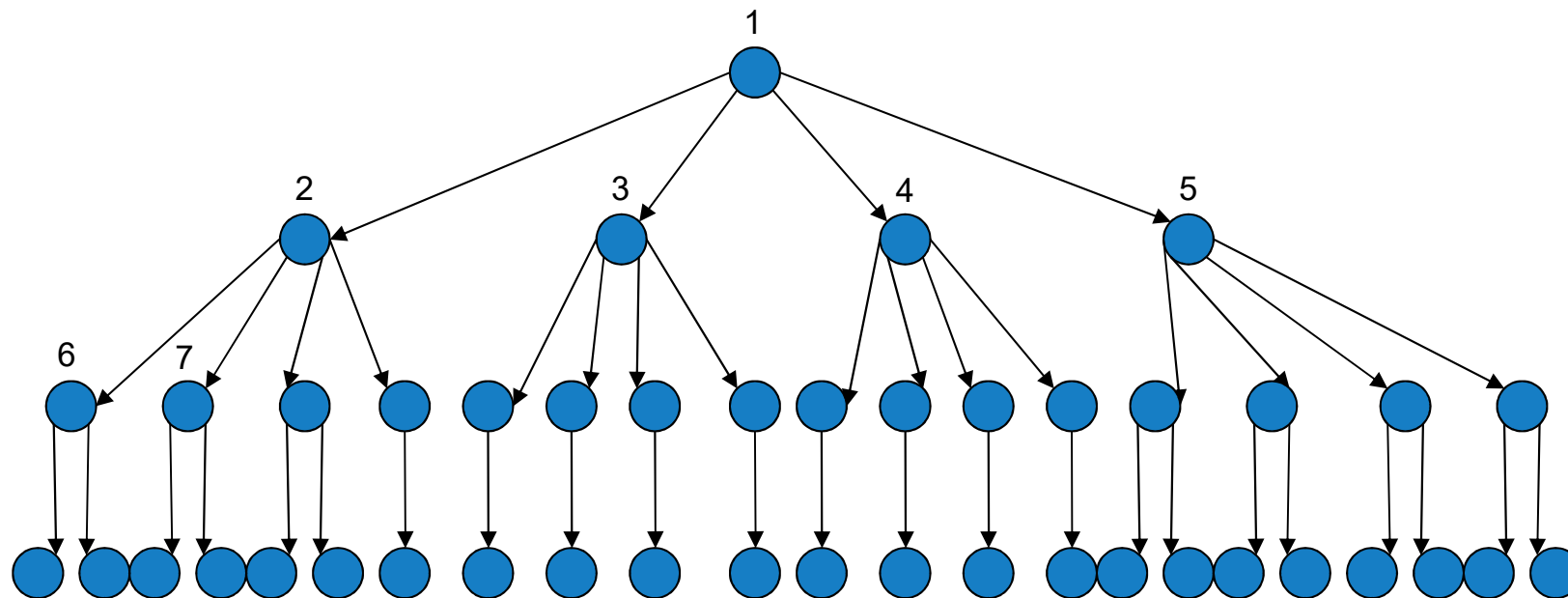
$L \leftarrow \text{Parse}(d(u));$

for each v **in** L **do**

Store ($E, (u, v);$

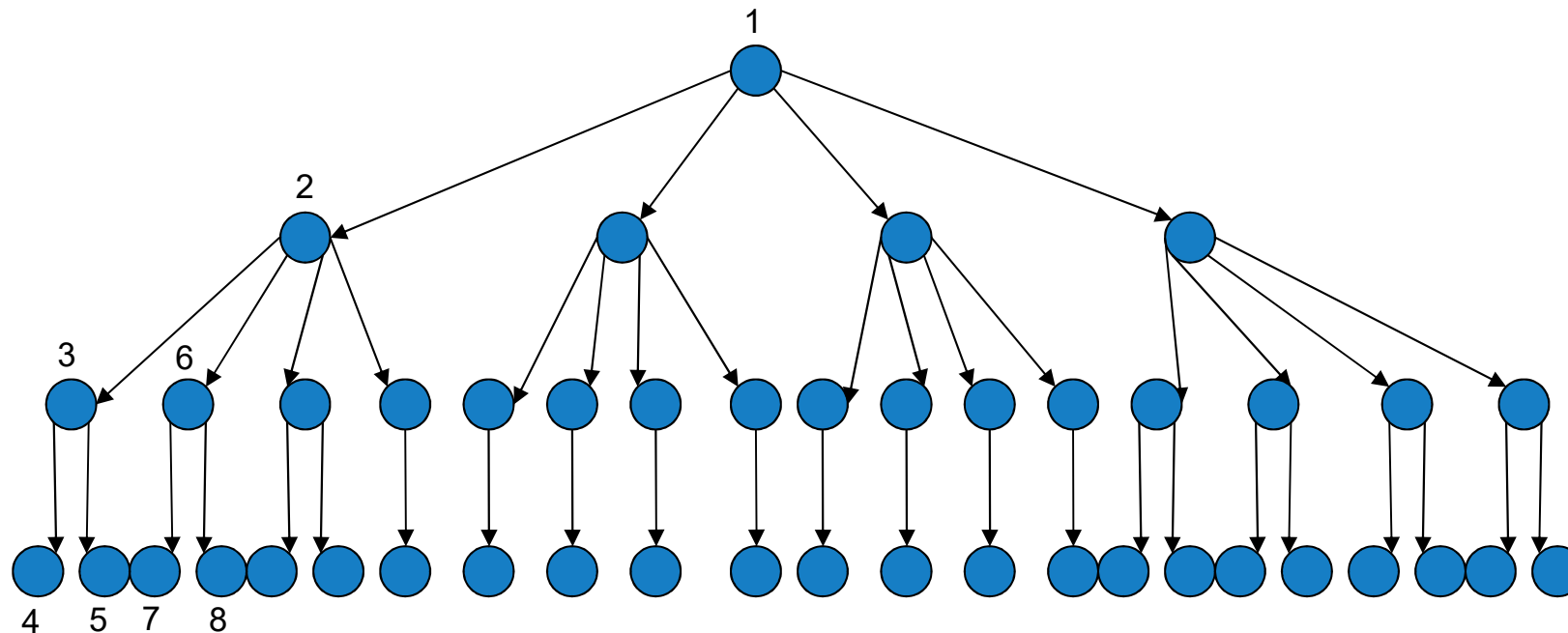
if $\neg((v \in D) \vee (v \in Q))$ **then** Enqueue(Q, v);

Breitensuche (Breadth-first)



Spidering

Tiefensuche Suche (Depth-first)



Spidering (Erforschungsstrategie)

- Tiefen- und Breitensuche sind Suchalgorithmen um einen Knoten in einem Graphen zu suchen.
- Tiefensuche erfordert Speicher nur für die Tiefe (d) mal den Verzweigungsgrad (b) ($O(bd)$). → Algorithmus benötigt aber zu viel Zeit, um nur einem Zweig nachzugehen (Wie tief kann ein Ast sein?).
- Breitensuche erforscht von der Rootwebseite gleichmässig nach aussen. Dieser Algorithmus erfordert Speicher für alle Knoten des Graphen von früheren Ebenen ($O(b^d)$).
- Kompromisslösung nötig
- Wie neue Links zu Q hinzugefügt (Enqueue()) und extrahiert (Dequeue()) werden hängt von der jeweiligen Suchstrategie ab.
 - FIFO („first in first out“ → wird am Ende von Q angehängt) → Breitensuche
 - LIFO („last in first out“ → wird am Anfang von Q angehängt) → Tiefensuche

Spidering (Erforschungsstrategie)

- Heuristisches Anordnen von Q ergibt einen fokusorientierten Crawler, der dadurch die Suche auf interessante Seiten ausrichtet.
- Der Spider kann auf eine bestimmte Webseiten begrenzt werden
 - Lösche Links zu anderen Seiten von Q
- Der Spider kann auf bestimmte Ordner begrenzt werden (z.B. Homepagesuche)
 - Lösche Links, die nicht im festgelegten Ordner liegen.
- Der Spider kann auf bestimmte Domänen begrenzt werden (Region, Land, Sprache), unter Annahme von genauen Themenbeschreibungen oder eine Menge von vorgegebenen und interessanten Webseiten .
 - Sortiere Q nach Ähnlichkeit (z.B. Kosinus). Als Sortierkriterien kann die ganze Webseite und/oder Ankertext des Themas verwendet werden.
 - Themenzuordnung durch Verfolgung und Entdeckung

Spidering (Erforschungsstrategie)

- Beobachten aller „in-degree“ und „out-degree“-Werte jeder Webseite.
- Beispiel:
 - Sortiere Q so, dass populäre Seiten (mit vielen in-coming Links) bevorzugt werden.
 - Sortiere Q so, dass zusammenfassende Seiten (mit vielen out-going Links) bevorzugt werden.
 - Verwende den Google-Algorithmus → PageRank
- Aber: Problem mit Start-Set S0 ("seeds") der URLs. Wie wählen?

Spidering (Praxis)

■ Praktische Betrachtungen

- Die Objekte D und E müssen auf einer Disk gespeichert werden (Grösse, Recovery,...)
- Wir müssen die Schleife abrechnen bevor $Q = \emptyset$ terminiert (Zeit, Speichereinschränkungen)
- Die Zeit, um ein Dokument herunterzuladen, ist unbestimmt (der Crawler kann nicht stur abwarten → gleichzeitiges herunterladen ist die übliche Lösung)
- Indiziere alle neuen Dokumente sofort (nicht trivial, da Behinderung des Spidering- und Suchprozesses)
- Eventuell wird den Webseitenbetreibern direkt erlaubt, ihre Webseiten anzugeben, die gespider werden sollen.

Spidering (Praxis)

■ Praktische Betrachtungen

- Das Web enthält viel redundante Information, und der Linkgraph enthält Zyklen
- URLs müssen in kanonische Form gebracht werden, um die Zyklen zu erkennen (trailing /, /index.html, ...)
- Duplizierte Seiten, Mirrors, ...
- Near-Duplicates (interessantes Forschungsproblem – das Erkennen von Near-Duplicates ist inhärent ein quadratischer Aufwand. Alternative: Fingerprints)

Spidering (Praxis)

■ Praktische Betrachtungen

- Crawler sollten respektvoll mit den verfügbaren Serverressourcen umgehen, und die Restriktionen der Webseitenbetreibern beachten. (Roboterausschluss, zwei Möglichkeiten)
 - Robots exclusion protocol (www.robotstxt.org): Umfasst Restriktionen, die für gewisse Verzeichnisse gesetzt werden können.
 - Roboter META-Tag: Individuelle Dokumenten-Tags, um Indexierung oder Verfolgung auszuschliessen.
- Webseitenadministratoren fügen ein “robots.txt”-File in das Rootverzeichnis ihrer Webseiten (z.B. www.foo.bar/robots.txt)
- Probleme mit Links, die mit JavaScript, eingebettetem Flash etc. erstellt worden sind.

Spider (Anker)

- Extrahiere Ankertext (zwischen `<a>` und ``) für jeden Out-Link
- Normalerweise beschreibt der Ankertext das Dokument sehr gut, auf welches es zeigt (→ Schlüsselworte der Webseite, Ultra-Kurz-Zusammenfassung).
- Füge Ankertext hinzu, damit die Zielwebseite mit zusätzlichen Schlüsselworten versehen wird:
 - `Software-Gigant`
 - `IBM`
- Sehr effektiv, um relevante Information zu finden (wird u.a. von Google verwendet).
- Nicht immer hilfreich («Click here», «forward»)
- Google bietet teilweise Suchresultate, die nur aufgrund des Ankertexts zustande gekommen sind.
- → «Google Bombing»

Indexierung Webseiten

- Die exakte Indexierungsstrategie ist ein Geschäftsgeheimnis und variiert von Unternehmen zu Unternehmen. (Was erwarten Sie aber für eine Tendenz?)
- Extremfall: Man kann die Beschreibung jeder Seite auf die Kollektion der zugehörigen Ankertexte beschränken.
 - Riesige Reduktion der Dokumentenmenge (z.B. NTCIR-5 Webtrack von 1.5TB auf 80MB)
 - Geeignet, um Webseiten zu indexieren, die keinen Text enthalten (z.B. Multimedia-Daten)
- SPAM ist ein Problem, sowohl beim Spidern, als auch beim Indizieren.
- !!! Es gibt Probleme bei der Speicherung des Index (muss verteilt werden – Konsequenzen?)

Indexierung Webseiten

- HTML-Struktur kann genutzt werden
 - <TITLE>-Tag
 - <H1>...<H5>, , , <I>,
 - <META>-Tags
-
- Wird nicht konsistent genutzt!

Anfrageverarbeitung

- Grosse Anzahl von Anfragen (in Millionen oder Milliarden pro Tag)
- Eine kleine Anzahl von häufigen Anfragen (80/20 Regel)
 - Periodische Themen (“chat”, “britney spears”)
 - Kurzlebige Themen (“tsunami”) → Kann mit “vorberechneten” Resultaten abgedeckt werden. Auch bei politisch sensiblen Resultaten (→ gerichtlich relevante, Zensur?)
 - Denken Sie daran: Suchresultate können grossen Einfluss ausüben
- Wenn die nächste Resultatseite angefragt wird (sehr häufige Anfragen): vorberechnen
- Falls die Anfrage gesucht werden muss → Cluster von PC's
- Kann als Spezialfall von verteiltem IR angesehen werden (Probleme?)

Evaluation

- Intuition: Das Wachstum des Web macht es immer schwerer, relevante Seiten zu finden (Ausschuss nimmt zu)
- Aber: NEIN, siehe Suchparadox. Die “Precision@1” nimmt zu, wenn das Volumen zunimmt.
- (Einschränkung: es gibt bedeutend weniger Aussagen zum Thema “Ausbeute” – das ist auch ein Artefakt der verwendeten Testkollektionen)

WT10g

- Der erste “substantielle” Web-Korpus, 1.69M dicht verlinkte Seiten (TREC 9/10)
- 50 Topics/Queries, aus Search Engine-Logs abgeleitet
 - Irrelevant, relevant, hochgradig relevant
 - Die Betrachtung von “hochgradig relevant” ist wichtig [Gord99, Voor01]
- 145 Homepage-Suchanfragen (in TREC-10)
- Aber: gross genug? Genügend Links?

Evaluation mit WT10g

- 50 klassische Topics, aus SE-Logs
 - <title> Vikings in Scotland?
<desc> What hard evidence proves that the Vikings visited or lived in Scotland?
<narr> A document that merely states that the Vikings visited or lived in Scotland is not relevant. A relevant document must mention the source of the information, such as relics, sagas, runes or other records from those times.
 - <title> halloween?
<desc> When, where, and how did Halloween evolve?
<narr> A relevant document will discuss the origin of Halloween and the original customs of Halloween. Modern day trick-or-treating stories are not relevant.
- 145 Homepage-Suchen
 - “Information Technology Institute” (www.iti.gov.sg)
 - “Digital realms” (www.drealms.co.uk)

Erkenntnisse

Homepage-Suchen (145 Anfragen in TREC-10)

Okapi system	MRR
+ SMART stemmer	0.261
No stemming	0.274
No stemming + proximity	0.367
+ URL length	0.653
Nostem+prox+URL	0.693

J. Savoy, Y. Rasolofo: Report on the TREC-10 Experiment: Distributed Collections and Entrypage Searching. Proceedings TREC'10, Gaithersburg (MD), 2002, 586-595

.GOV-Kollektion

- .GOV corpus, 1.25M Seiten, Crawl der .gov-Domain, erstellt für TREC 11/12
- 50 Standardsuchen
- 150 Site-Suchen
- 50 Anfragen für "topic distillation" ("Dossier zusammenstellen")
- Nicht wirklich repräsentativ, aber interessant
- Realistischer Web-Graph

Evaluation (.GOV)

- 50 Standardanfragen, aus SE-Logs
 - <title> intellectual property
<desc> Find documents related to laws or regulations that protect intellectual property
<narr> Relevant documents describe legislation or federal regulations that protect authors or composers from copyright infringement, or from piracy of their creative work. These regulations may also be related to fair use or to encryption.
- 150 Site-Suchen
 - “US agriculture changes 20th century”
 - “US passport renewal”
 - “white house west wing history”

Evaluation (.GOV)

- 50 Anfragen für “topic distillation” (TREC-12, 2003)
 - <title>cotton industry
<desc> Where can I find information about growing, harvesting cotton and turning it into cloth?
 - Ein Dossier zusammenstellen (nicht *alle* Antworten)
 - Nicht nur eine einzelne Site (→ Site-Suche)
- Erfolgsfaktoren
 - Web-Struktur (<title>, Meta)
 - Ankertext
 - Indegree
 - URL (Länge)

ClueWeb – 1 Milliarde Dokumente

- Experimente in Hinblick auf Big Data-Phänomene werden erst mit sehr grossen Testdatensets möglich.
- ClueWeb mit ~1 Milliarde Dokumenten/Webseiten (multilingual)
- Topics aus Logs von Websuchdiensten hergeleitet
- Zum Teil andere Form der Evaluation, da keine klassischen Relevance-Assessments

Suchmaschinen (Schlussfolgerung)

- Spidering ist ein wichtiger Aspekt
- Ranking ist der Schlüssel (eine gute Antwort in den Top 5/10)
- Verwendung von verschiedenen Merkmalen → präzisionsorientiert!
 - Webseiteninhalt (+tags)
 - Meta tags + <title>
 - Ankertext
 - Hyperlink-Struktur
 - URL (Länge, Text, Struktur)
- Stemming / Stoppwortliste verbessert nicht immer das Resultat im gewünschten Sinne
- Die Geschwindigkeit ist ein Schlüsselement
- Wirtschaftlichkeit! → Wie teuer darf ein Suche sein?

Suche und Hyperlink

- Spreading Activation (SA)
- Indegree & PageRank
- Hub und Autorität

Spreading Activation (SA)

- Mit dieser Methode:

- Der Grad der Ähnlichkeit zwischen D_i and Q , (bezeichnet mit $\text{Sim}(D_i, Q)$), wird durch eine bestimmte Anzahl von Kreisläufen (normalerweise 1) zu den verlinkten Dokumenten propagiert. Dazu wird ein Faktor λ verwendet. Normalerweise werden nur die Top r Dokumente angepasst.

$$RSV(D_i, Q) = \text{Sim}(D_i, Q) + \lambda \cdot \sum_{j=1}^k \text{Sim}(D_j, Q)$$

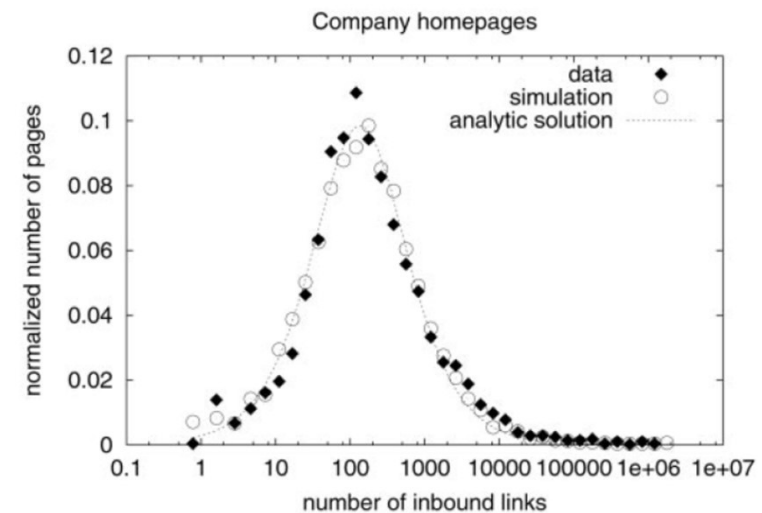
- Der Wert von lambda kann abhängen von:

- Hyperlinktyp
- Dokumenttyp
- Hyperlink-Richtung
- Ähnlichkeit zwischen Q and D_j

- Quelle: J. Savoy, J. Picard: Retrieval effectiveness on the web. Information Processing & Management, 2001, 37(4), 543-569

Indegree

- Indegree ist die Anzahl von einkommenden Links zu einem gegebenen Dokument.
- Indegree als Qualitätseinflussfaktor ist schwach:
 - Indegree kann von Spammern missbraucht werden
 - Nicht alle „Vaterseiten“ sind gleich wertvoll (einige Links sind prestigeträchtiger als andere)
 - Navigationslinks vs. Inhaltliche Links



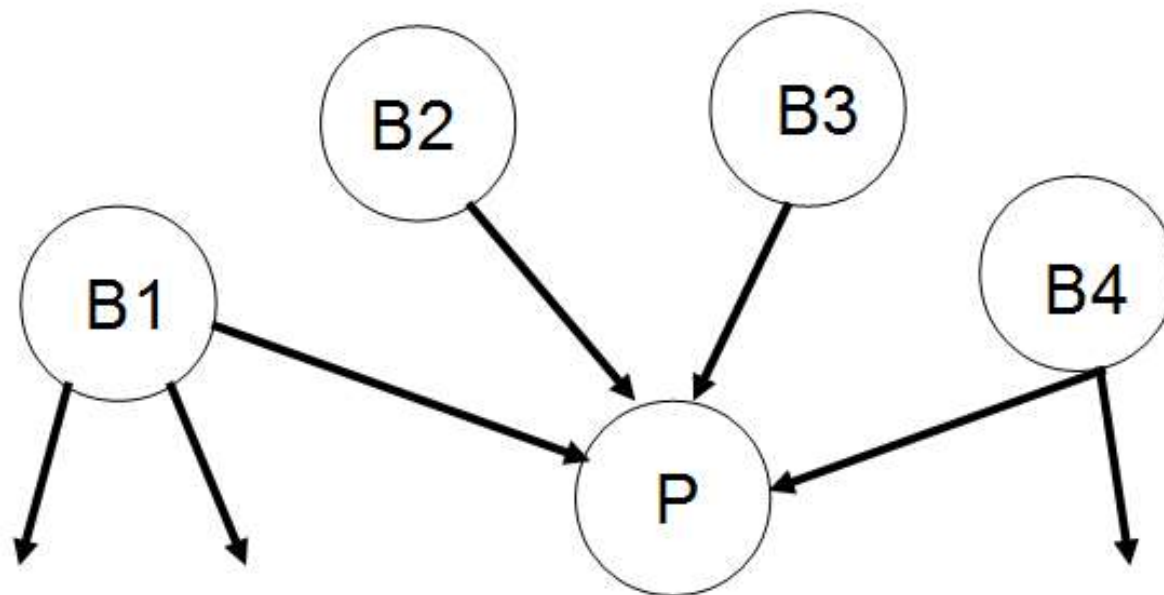
PageRank

- Zu Beginn ist der Surfer auf einer zufälligen Webseite
 - Bei jedem Schritt kommt der Surfer mit einer Wahrscheinlichkeit d zu einer zufällig gewählten Webseite (z.B. Wahrscheinlichkeit von einem zufälligen Schritt = 0.15)
 - Oder zu einem zufällig gewählten Nachfolger der aktuellen Seiten mit der Wahrscheinlichkeit $1-d$ (z.B. Wahrscheinlichkeit der Verfolgung eines zufälligen outlink = 0.85)
- PageRank einer Webseite = Wahrscheinlichkeit, dass ein Surfer sich zu einem beliebigen Zeitpunkt auf dieser Webseite befindet.
- Quelle: Brin S., Page L., The anatomy of a large-scale hypertextual web search engine, Proceedings of the WWW7, Amsterdam, Elsevier, 107-117, 1998.

PageRank

Idee:

- Qualität von P: $Q(P) = Q(B1)/3 + Q(B2) + Q(B3) + Q(B4)/2$



$$PR^{c+1}(D_i) = (1-d)\frac{1}{n} + d \left[\frac{PR^c(D_1)}{C(D_1)} + \dots + \frac{PR^c(D_m)}{C(D_m)} \right]$$

$C(D_1)$: Anzahl Outlinks von D_1 . Faktor $1/n$ ist umstritten.

Gedanken zu PageRank

- Berechnung rekursiv
- Kann offline berechnet werden
- Anfangs Page-Rank aller Knoten = 1
- Konvergiert gegen "Random Surfer"-Wahrscheinlichkeiten
- Durchaus anfällig auf Spam

Fundamentaler Nachteil?

- Muss theoretisch nach jeder Änderung neu berechnet werden

Kleinberg Model

- HITS (Hypertext Induced Topic Selection) vorgeschlagen von Kleinberg
- Eine Webseite kann angesehen werden als:
 - Ein Hub (zeigt/verweist auf gute Quellen)
 - Eine Autorität (besitzt den Inhalt)
- Von einer Ursprungsmenge R , ziehen wir allen Nachbarn in Betracht
- Von dieser erweiterten Menge können wir die „hubness“ und „Autorität“ jedes einzelnen Knoten (Webseite) bestimmen.
- Quelle: Kleinberg J., Authoritative sources in a hyperlinked environment, Journal of the ACM, 46(5), 1999, 604-632.

HITS

- In diesem Algorithmus erweitern wir die Ursprungsmenge um sowohl die Kinder ($ch[v]$) als auch die Eltern ($pa[v]$) miteinzubeziehen.

DefineGraph (R, σ)

$S \leftarrow R$;

for each v **in** R **do**

$S \leftarrow S \cup ch[v]$;

$P \leftarrow pa[v]$;

if $|P| > \sigma$ **then**

$P \leftarrow$ arbitrary subset of P of size σ ;

$S \leftarrow S \cup P$;

return S ;

- Mit den Knoten, die im zurück gelieferten Graphen vorkommen, können wir die Werte vom Hub und der Autorität berechnen.

HITS

■ Hub- und Autoritätswerte

$$A^{c+1}(D_i) = \sum_{D_j \in \text{parent}(D_i)} H^c(D_j)$$

$$H^{c+1}(D_i) = \sum_{D_j \in \text{children}(D_i)} A^c(D_j)$$

- $A^c(D_i)$: Autoritätswert von Webseite D_i nach c Iterationen
 - $H^c(D_i)$: Hubwert von Webseite D_i nach c Iterationen
 - Wir starten mit Gleichverteilung $H^0(D_i)=1$ und $A^0(D_i)=1$
- Wir berechnen beide Werte iterativ (z.B. werden nach 5 Iterationen die Werte schon relativ stabil)
- Die Werte werden nach jeder Iteration normalisiert, so dass die Summe der Quadrate = 1 (siehe auch Praktikum)

Page Rank vs. Kleinberg/HITS

■ Page Rank:

- kann vorausberechnet werden (rel. effizient)
- berechnet nur eine Form von "Popularität"

■ HITS:

- auf Ranglistenauszug
- berechnet Autoritäten und Hubs
- einfach zu berechnen, aber ineffizient

■ Merke: auf kleinen Kollektionen Effektivität kaum bewiesen

Link-Analyse

- Anzahl von Webseiten eines kompletten Webauftrittes
- Anzahl von Hyperlinks einer gegebenen Webseite oder eines Webauftrittes
- Bowtie Model

Link-Analyse

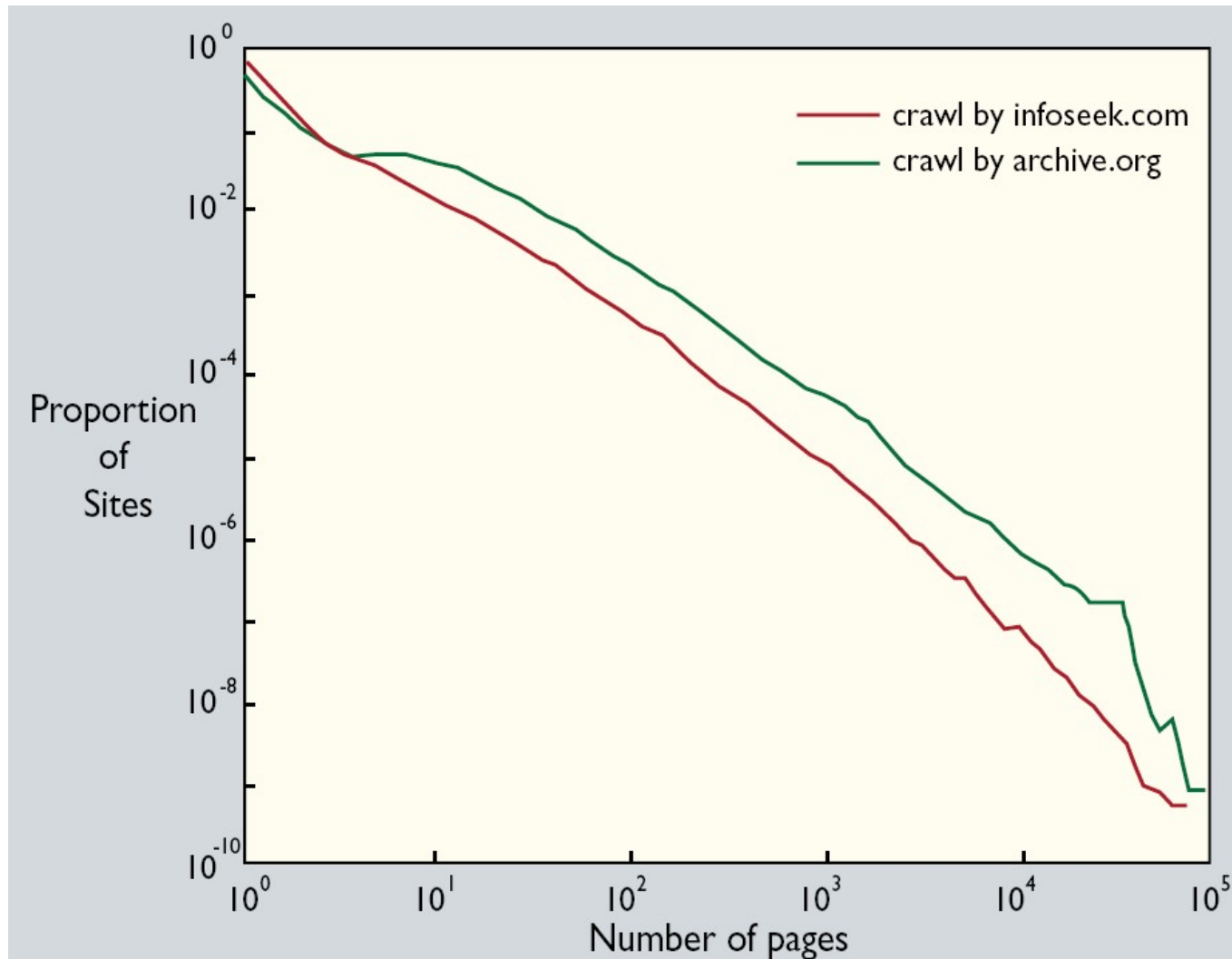
- Die Anzahl von Webseiten eines gegebenen Webauftritts wird als n angenommen. $P(n)$ bezeichnet die Wahrscheinlichkeit, dass der Webauftritt n Webseiten hat

$$P(n) = b \cdot n^{-\beta}$$

with $b > 0$ and $\beta > 1$

- Folgt "Power-law". (Zipf's Gesetz ist auch „power-law“)
- Quelle: Adamic L.A., Huberman B.A., The web's hidden order, Communications of the ACM, 44(9), 2001, 55-59.

Link-Analyse



Link-Analyse

- Link-Analyse: einzelne Seiten
- $P(k)$ kennzeichnet die Wahrscheinlichkeit, dass eine einzelne Seite im Web k Hyperlinks hat (indegree oder inlinks oder outdegree oder outlinks)

$$P(k) = b \cdot k^{-\gamma}$$

with $b > 0$ and $\gamma > 1$

- γ_{in} und γ_{out} können verschieden sein.
- Quellen:
 - Albert R., Jeong H., Barabási A.L., Diameter of the world-wide web, Nature, 401(6749), 1999, 130-130.
 - Broder A., Kuwar R., Maghou F., Raghavan P., Rajagopalan S., Stata R., Tomkins A., Wiener J., Graph structure in the web, Proceedings of the WWW9, Computer Networks, vol. 33(1-6), 2000, 309-320.
 - Huberman B.A., The laws of the web, patterns in the ecology of information, Cambridge, The MIT Press, 2001.

Link-Analyse

- Links-Analyse: Web Sites
- $P(k)$ kennzeichnet die Wahrscheinlichkeit, dass die Site k Hyperlinks hat (inlinks oder outlinks)

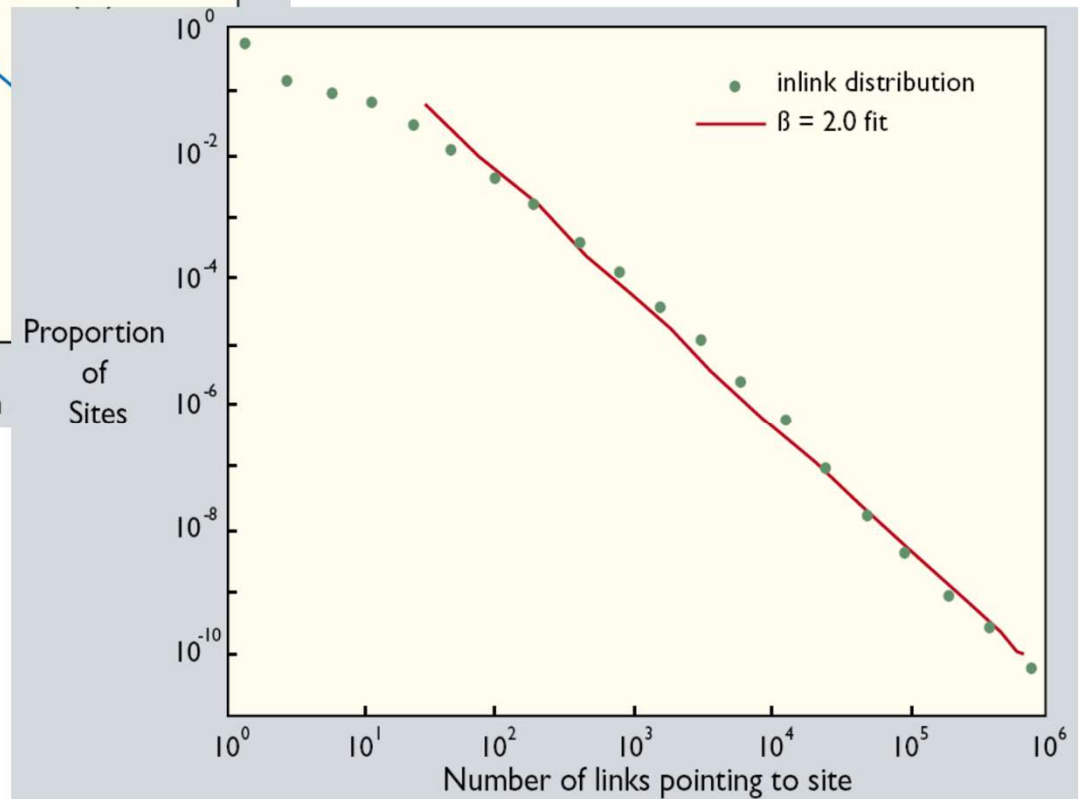
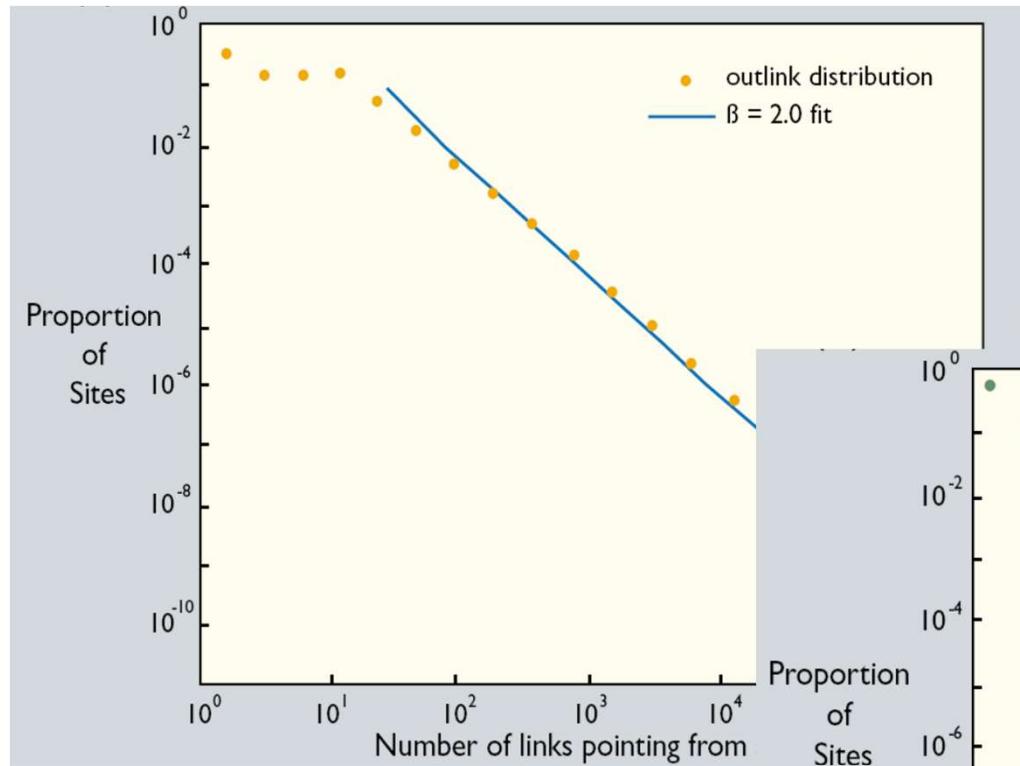
$$P(k) = b \cdot k^{-\gamma}$$

with $b > 0$ and $\gamma > 1$

with $\gamma = a + c \cdot \log(n_d)$

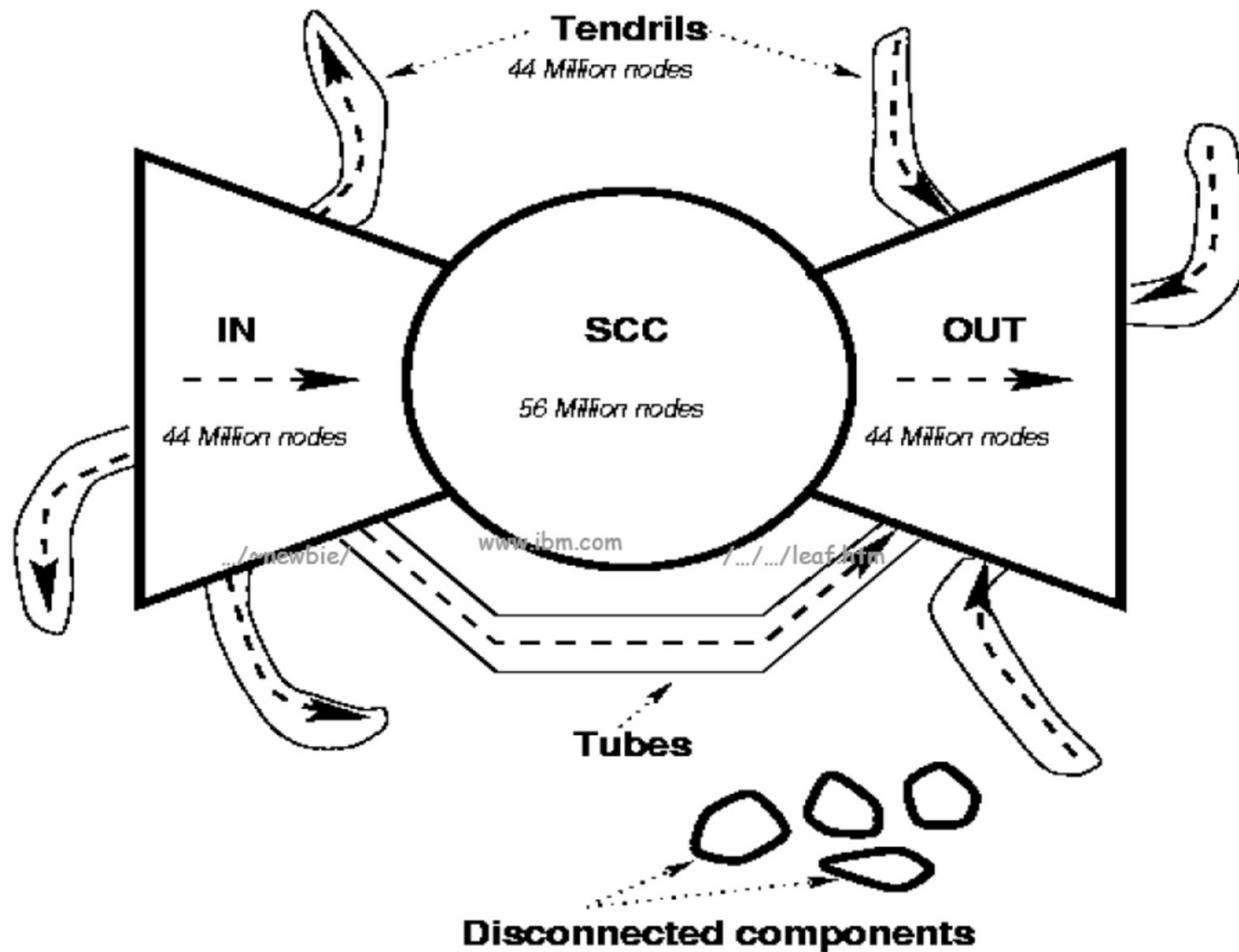
- γ erhöht sich mit der Anzahl von einzelnen Seiten innerhalb der Domäne (n_d)
- Quelle: Bharat K., Chang B., Henzinger M., Ruhl M., Who links to whom: mining linkage between web sites, Proceedings of IEEE ICDM-01, 51-58, 2001

Link-Analyse



Link-Analyse

Web Topologie



Quelle: Broder, R. Kumar, F. Maghoul, P. Raghavan, S. Rajagopalan, R. Stata, A. Tomkins and J. Wiener, "Graph Structure in the Web", Proc. WWW9 conference, 309-320, May 2000. See also: <http://www9.org/w9cdrom/160/160.html>

Frage



- Wir stellen uns die Frage: Inwiefern modelliert PageRank die Charakteristiken des realen Webs?

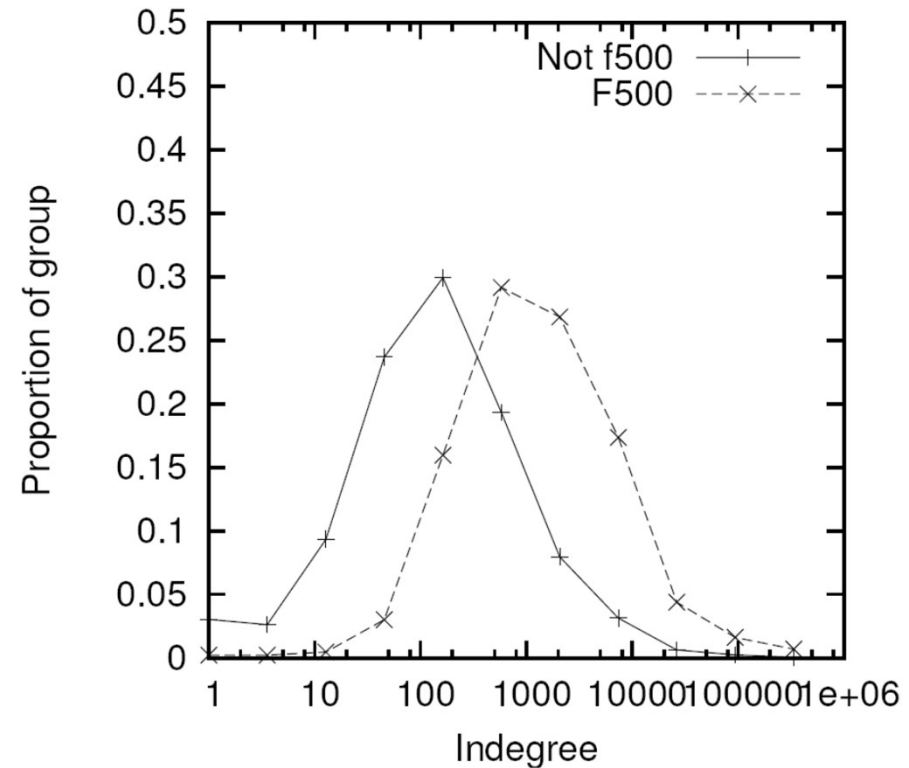
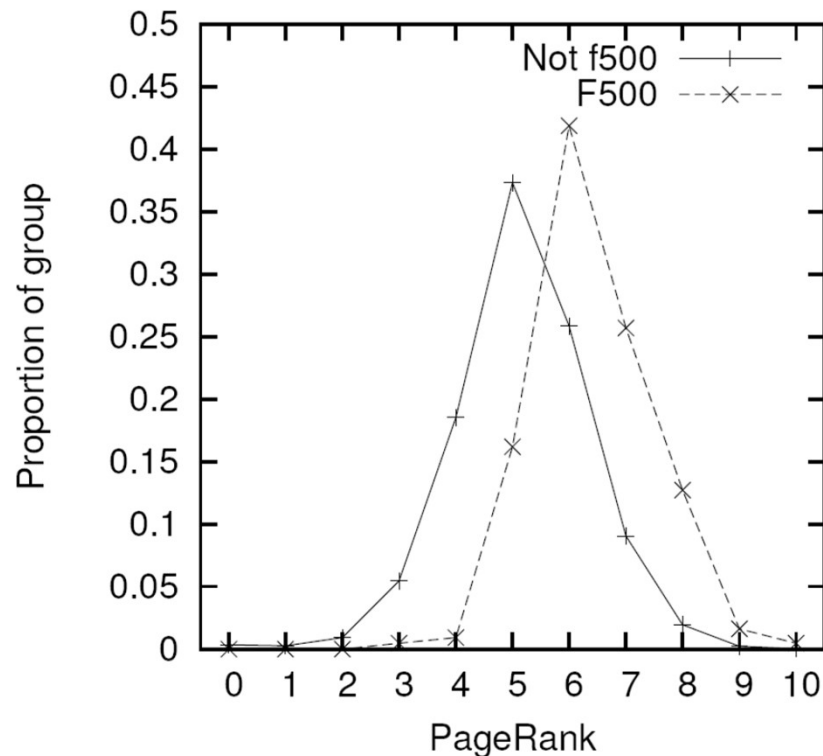
PageRank & Indegree

- Die Verlinkung ist wichtig für das Ranking der SE („wichtige, qualitative hoch stehende“ Seiten vor relevanten, aber weniger "wertvollen" Seiten)
- Liefert die Anzahl von eingehenden Links (indegree) die gleiche Information wie PageRank?
- Studie von 5370 Unternehmens-Webauftritten (2003)
- Starke Korrelation zwischen Indegree und PageRank (0.767, Pearson r)
- Quelle: Upstill T., Craswell N., Hawking D., Predicting fame and fortune: PageRank or Indegree?, Proceedings of 8th Australasian Document Computing Symposium (ADCS), Canberra, 2003.

Verzerrungen im PageRank

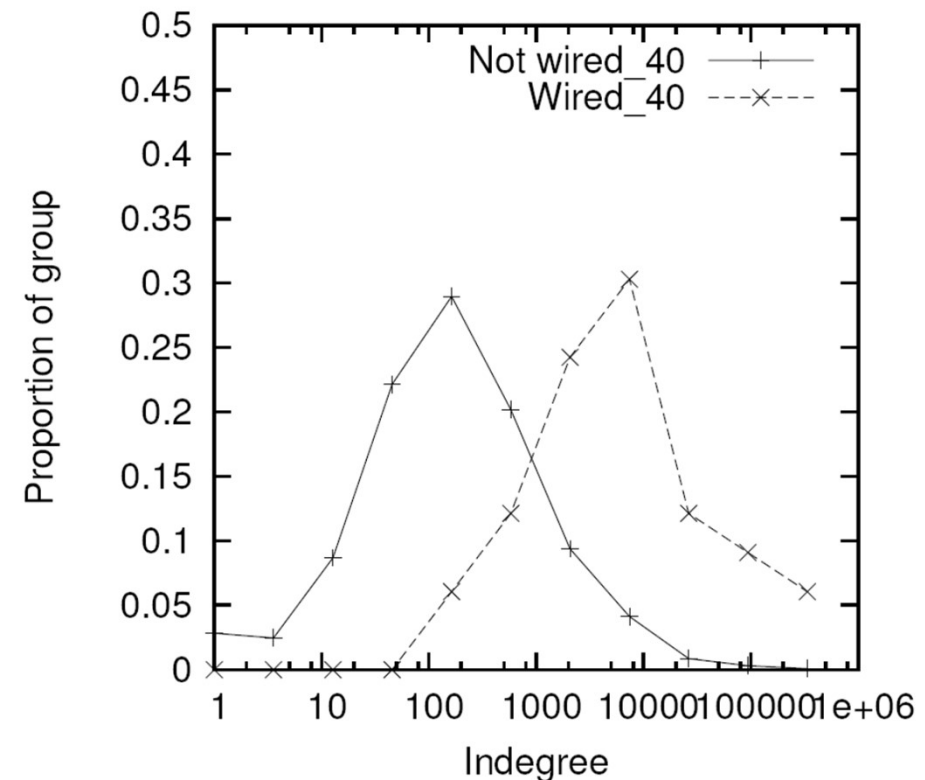
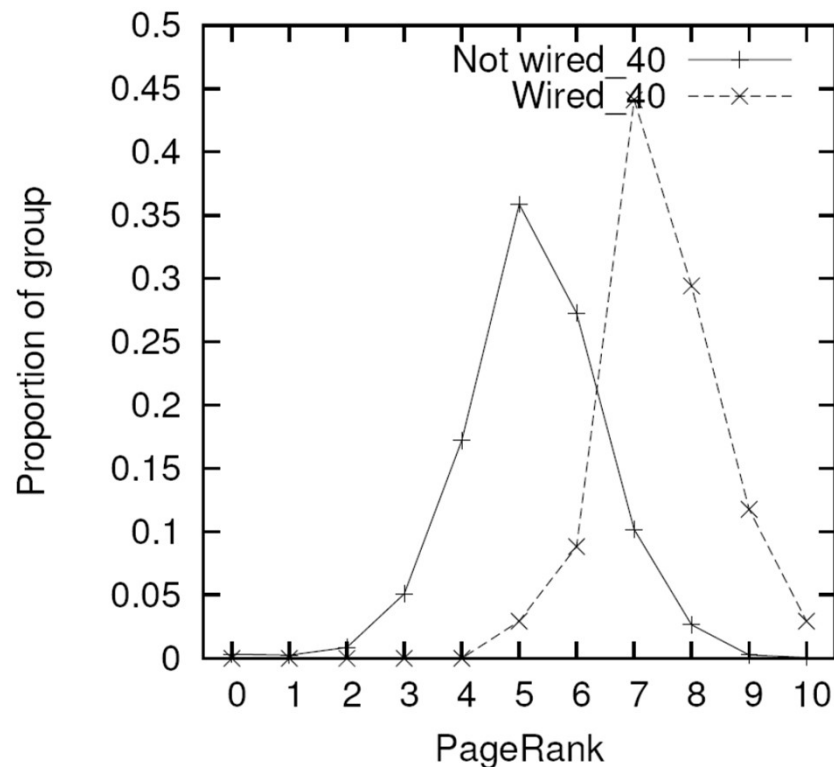
- Es gibt einige (reale oder vermutete) Verzerrungen im PageRank
- Homepages werden begünstigt
- In Richtung zu den großen, berühmten Firmen
 - Mit grossen Einnahmen (Fortune 500): +1 pt
 - Beliebte, bewunderte Unternehmen (Most Admired): +1pt
 - Technologieorientierte: +1pt
 - Mit bekannten Marken: +2 pt
 - In der “new economy” (Wired 40): +2 pt
 - Kein Vorteil zugunsten von US-Unternehmen

PageRank & Indegree



Quelle: Upstill T., Craswell N., Hawking D., Predicting fame and fortune: PageRank or Indegree?, *Proceedings of 8th Australasian Document Computing Symposium (ADCS)*, Canberra, 2003.

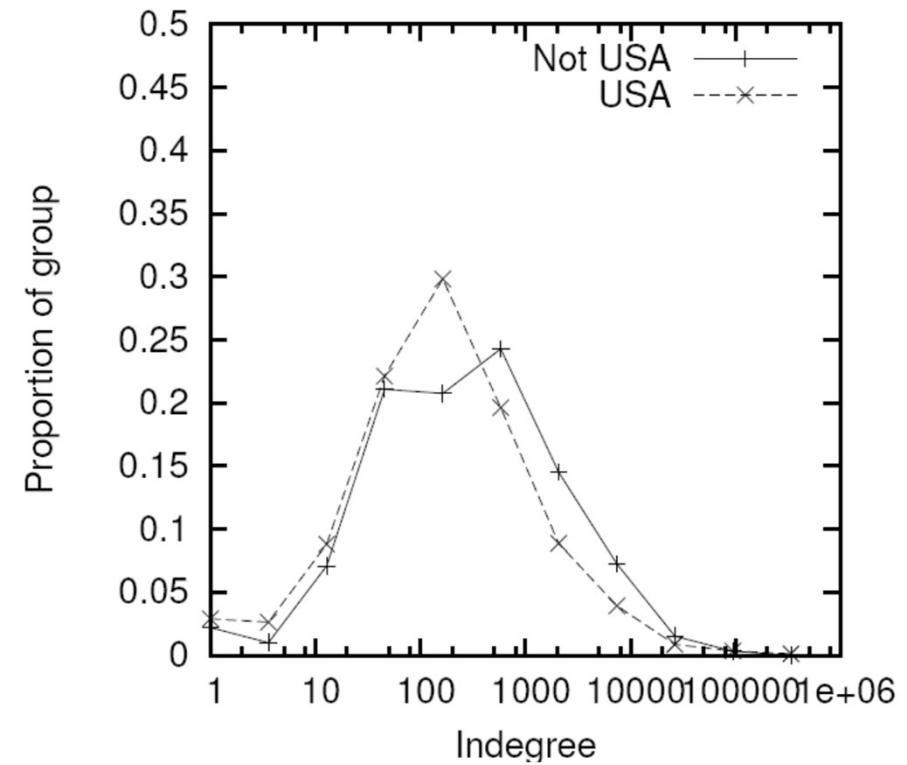
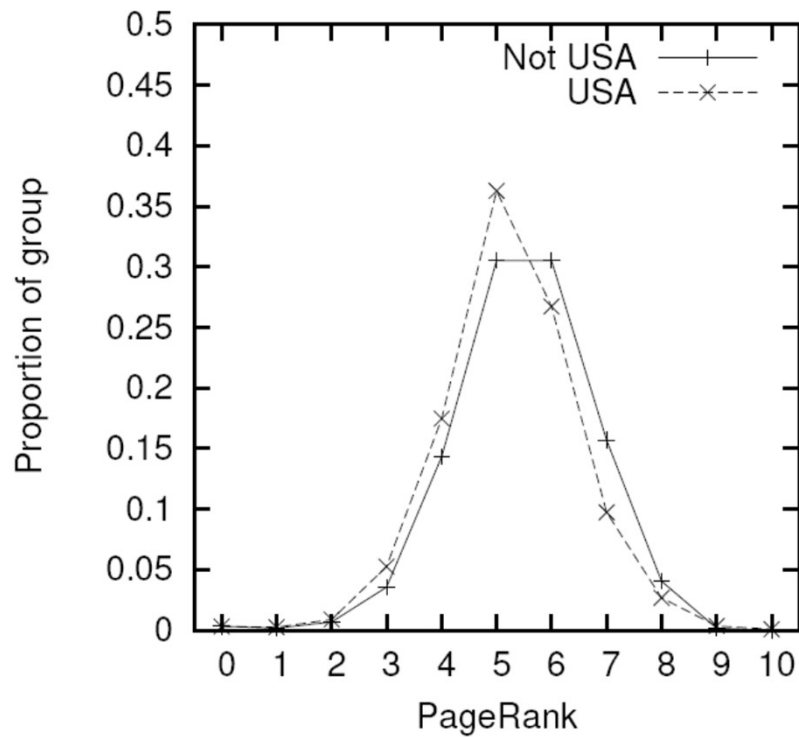
PageRank & Indegree



Bemerkung: Technologie-Leader sind proportional “überrepräsentiert”.

Quelle: Upstill T., Craswell N., Hawking D., Predicting fame and fortune: PageRank or Indegree?, *Proceedings of 8th Australasian Document Computing Symposium (ADCS)*, Canberra, 2003.

PageRank & Indegree



Fallstudie: Yahoo Search

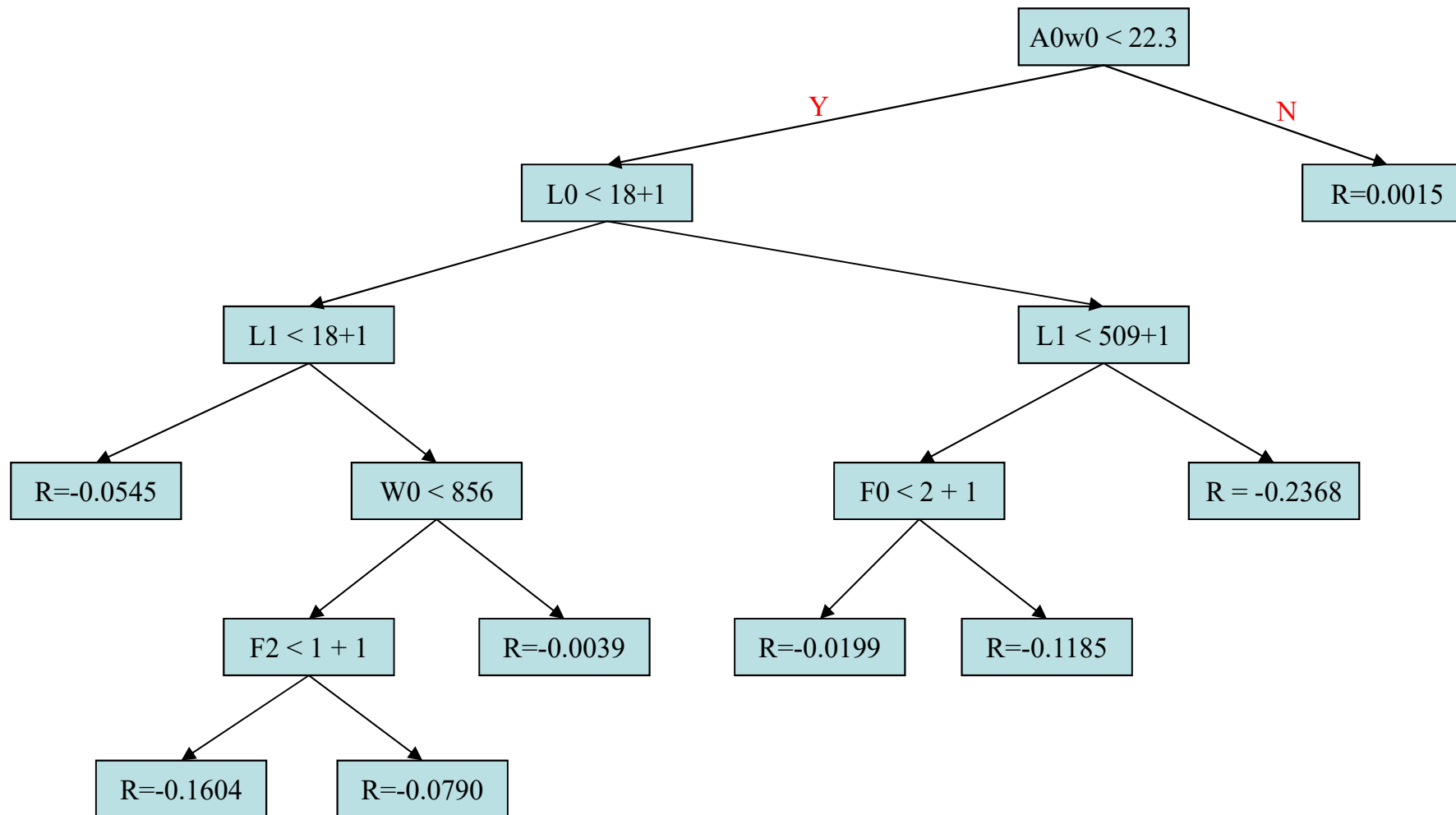
■ Gewichtungsmerkmale

- A0 - A4 anchor text score per term
- W0 - W4 term weights
- L0 - L4 first occurrence location (encodes hostname and title match)
- SP spam index: logistic regression of 85 spam filter variables (against relevance scores)
- F0 - F4 term occurrence frequency within document
- DCLN document length (tokens)
- ER Eigenrank
- HB Extra-host unique inlink count
- ERHB $ER \cdot HB$
- A0W0 etc. $A0 \cdot W0$
- QA Site factor – logistic regression of 5 site link and url count ratios
- SPN Proximity
- FF family friendly rating
- UD url depth

Quelle: Search Quality, Jan Pedersen, 10.09.2007,

http://courses.ischool.berkeley.edu/i141/f07/lectures/pedersen_search_quality07.pdf

Entscheidungsbaum Yahoo Search



Quelle: Search Quality, Jan Pedersen, 10.09.2007,

http://courses.ischool.berkeley.edu/i141/f07/lectures/pedersen_search_quality07.pdf

Schlussfolgerung

- Das Internet hat spezifische Eigenheiten
- Es gibt auch andere wichtige Datenbehälter als das Internet
- Gelegenheiten für verschiedene Erforschungen
- Die Technologien der Suchmaschinen sind nicht auf IR limitiert.
- Einfachheit ist die Richtlinie
- Die Entwicklung / Gründe
 - Wenig grosse Webauftritte / viel kleine Webauftritte (power-law)
 - Schnelles Wachstum (Google, Amazon, eBay)
 - Neue Entwicklungen (MP3, Webblog, skype)

Quellen

- Jacques Savoy: Kurs Web Search
- R. Baeza-Yates, B. Ribeiro-Neto: Modern Information Retrieval. Addison-Wesley, Reading (MA), 2011
- P. Baldi, P. Frasconi, P. Smyth, Modeling the Internet and the Web. John Wiley & Sons, Chichester (UK), 2003
- <http://trec.nist.gov/> (see « Publications » and the Web track)

Institut für angewandte Informationstechnologie InIT

Appendix: Small Data

Was ist Big Data?

Oft zitiert werden im Zusammenhang mit Big Data die 3 V's:

1. Volume
2. Velocity
3. Variety

Manchmal ergänzt um ein 4tes V:

4. Veracity

Es geht also nicht nur um den Aspekt «gross»

Wann ist «gross» interessant?

- Eine Analogie aus der Physik in Sachen Geschwindigkeit:
 - «Was passiert, wenn ein Baseball mit (annähernd) Lichtgeschwindigkeit geworfen wird»? (siehe <https://what-if.xkcd.com/1/>)
 - (Teil der) Antwort: hier versagt zum Beispiel alles, was wir über Aerodynamik wissen: der Ball bewegt sich *schneller* als die Luftmoleküle, die daher mit den Atomen des Balls *verschmelzen*.
- Total neue Effekte treten auf, wenn gewisse Grenzen überschritten werden

Die Take-Away-Message für IR

- «Gross» per se ist aber nicht interessant
- Manche Problem werden nicht «schwieriger», nur weil die Datenmenge gross ist.
 - Beispiel: Streichen Sie alle Namen mit Buchstaben «a» aus der folgenden Liste:
 - Anna, Kurt, Bernie, Amalia, Laura, Jasmin, Thomas, Peter
 - Die Aufgabe wäre zwar zeitraubender, aber kaum intellektuell fordernder, wenn wir stattdessen 10 Millionen Namen verarbeiten würden
- Das «Hello World»-Analog von MapReduce ist «wordcount» → dies illustriert, dass hier die Verarbeitung an sich im Zentrum steht, nicht das Problem

Die Take-Away-Message für IR

- «Gross» per se ist also auch für IR nicht interessant.
- Die Text Retrieval Conference (TREC) wurde 1992 gegründet, um eine neue Grössendimension an Textkollektionen zu erschliessen. So wurden Kollektionen im Umfang von ca. 1 Gigabyte zusammengestellt. (Harman 1992)
- Diese Grösse war damals ein ganz wesentlicher Teil der Challenge für eine Teilnahme an den Experimenten. Sozusagen also «Big Data für IR» in den frühen 90er-Jahren.
- Heute ist diese Kollektionsgrösse problemlos im Rahmen von Studentenexperimenten verarbeitbar, ohne irgendwelche speziellen Massnahmen.

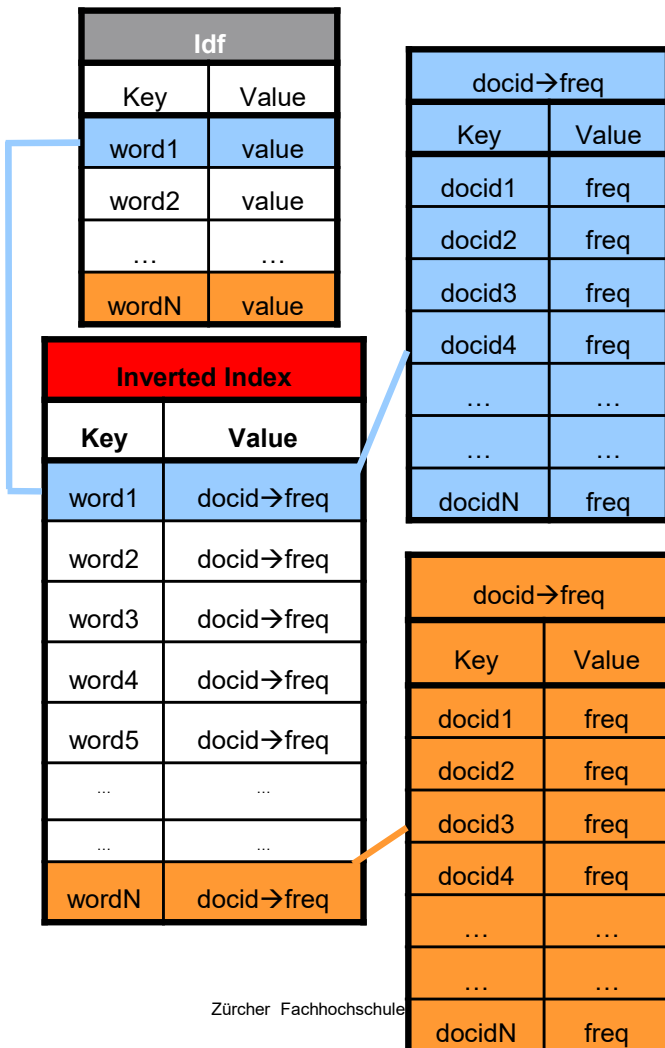


«Was für Implikationen hat es für die Effizienz eines IR-Systemes, wenn die Kollektion wächst»?

Architektur minimales IR-System

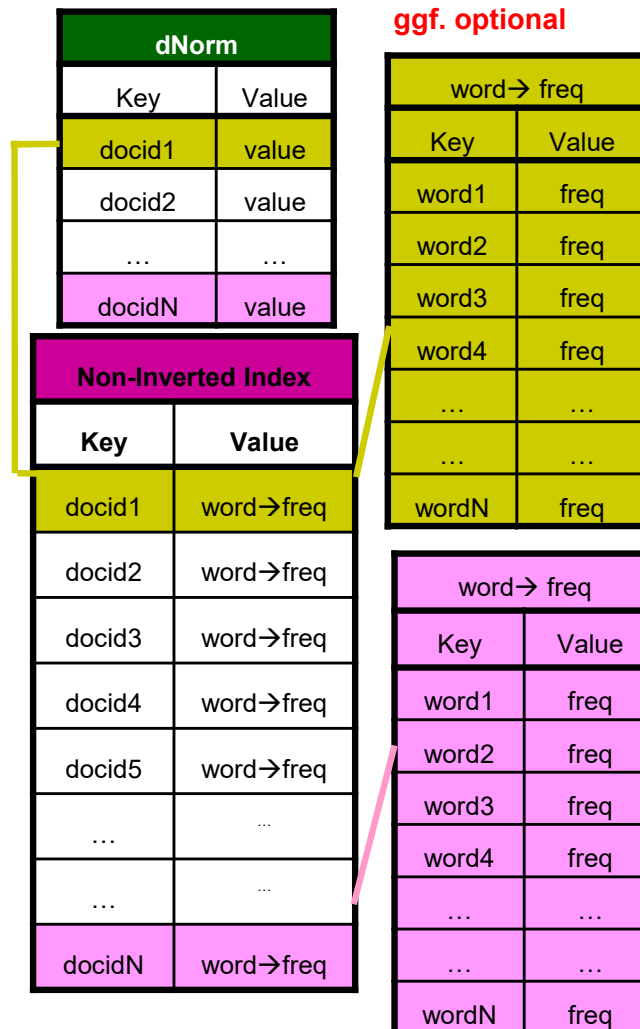
Invertierter Index

Quelle: Dokumente



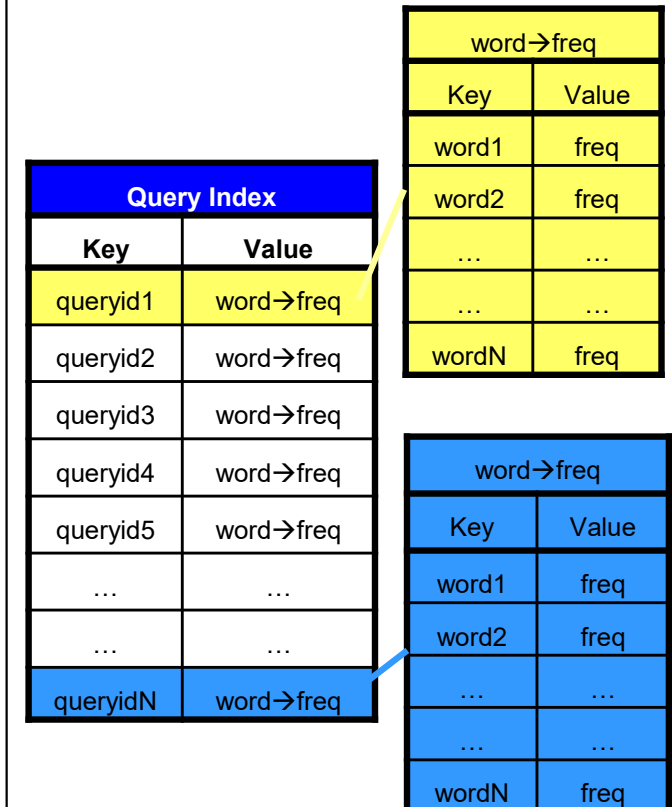
Nicht-Invertierter Index

Quelle: Dokumente



Anfrageindex

Quelle: Anfrage



Architektur minimales IR-System (cont.)

accumulator

accu	
Key	Value
doc 1	acc. value
doc 2	acc. value
doc 3	acc. value
doc 4	acc. value
...	...
...	...
doc N	acc. value

Der Akkumulator summiert das Produkt von idf und Termhäufigkeit für jedes Word, das im entsprechenden Dokument vorkommt.

dNorm

dNorm	
Key	Value
docid1	value
docid2	value
docid3	value
docid4	value
docid 5	value
...	...
...	...
docidN	value

Dokumentenorm „dNorm“ wird für alle Dokumente vorberechnet

idf

idf	
Key	Value
word1	value
word2	value
word3	value
word4	value
word5	value
...	...
...	...
wordN	value

Der idf (Inverse document frequency) wird für alle Wörter in allen Dokumenten vorberechnet, ggf. auch für Anfrageterm mit $df=0$.

Gewichtungsformel RSV

$$a_{i,j} := ff(\varphi_i, d_j) * idf(\varphi_i)$$

$$b_i := ff(\varphi_i, q) * idf(\varphi_i)$$

$$RSV(q, d_j) := \frac{\sum_{\varphi_i \in \Phi(q) \cap \Phi(d_j)} a_{i,j} * b_i}{\sqrt{\sum_{\varphi_i \in \Phi(d_j)} a_{i,j}^2} * \sqrt{\sum_{\varphi_i \in \Phi(q)} b_i^2}}$$

%accu

%dNorm

\$qNorm

■ Wobei:

- RSV = retrieval status value
- ff = feature frequency
- idf = inverse document frequency
- d = document
- q = query
- φ = term

Effizienzüberlegungen zu grossen IR-Kollektionen

- Grundsätzlich stützen sich IR-Systeme auf einen invertierten Index (Term → Dokument)
- Der Zugriff erfolgt per Hashtabelle, mit dem Term als Schlüssel (→ Konsequenzen für Zugriffsgeschwindigkeit? Konsequenzen für Zugriffsmöglichkeiten?)
- Ein potentieller Bottleneck ist die Sortierung von Ranglisten (Aufwand Sortierung $O(n \log n)$)
- Was passiert, wenn unser Index die realistische Festplattengrösse sprengt?

Exkurs «Distributed IR»

- Die Globalstatistiken des Invertierten Index sind das Hauptproblem
- Betrachtet man zum Beispiel die tf.idf-Kosinus-Formel,

$$a_{i,j} := ff(\varphi_i, d_j) * idf(\varphi_i)$$

$$b_i := ff(\varphi_i, q) * idf(\varphi_i)$$

$$RSV(q, d_j) := \frac{\sum_{\varphi_i \in \Phi(q) \cap \Phi(d_j)} a_{i,j} * b_i}{\sqrt{\sum_{\varphi_i \in \Phi(d_j)} a_{i,j}^2} * \sqrt{\sum_{\varphi_i \in \Phi(q)} b_i^2}}$$

so wird offensichtlich, dass die Berechnung der df, resp. der idf globale Informationen benötigt.

- Das Problem lässt sich nicht so einfach umgehen → Probleme mit Merging etc.



«Was für Implikationen hat es für die Effektivität eines IR-Systemes, wenn die Kollektion wächst»?

Was macht IR denn schwierig? (Repetition)

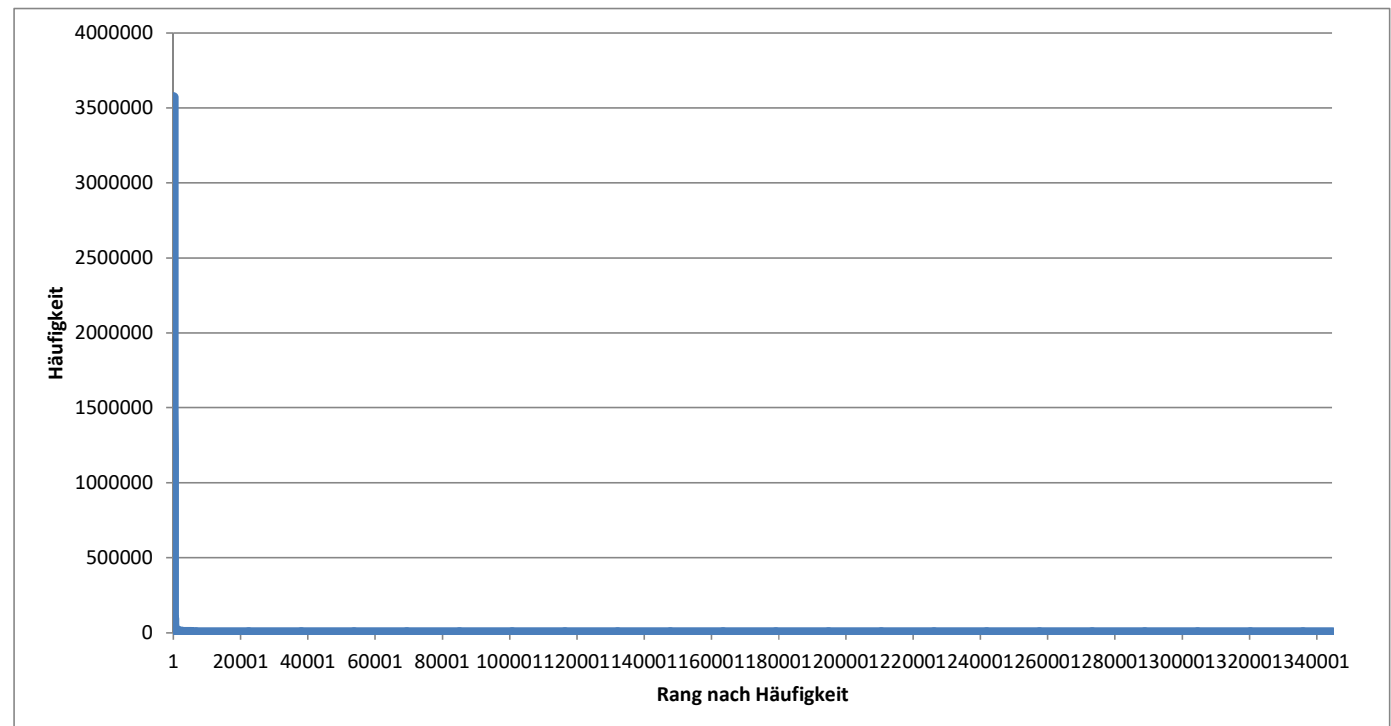
- Sprachliche Phänomene:
 - Synonyme, Homonyme, Umschreibungen, Metaphern
 - Morphologie, Wortbildung, Wortformen
 - Akronyme, Schreibfehler
- Subjektivität der Relevanz
 - Persönliche Präferenzen
 - Zeitliche Effekte (Lerneffekt, Bedeutungsverschiebungen)
 - Autorität von Quellen
- Ungenügende Verbalisierung/Codierung von Informationsbedürfnissen
 - Suchparadox: ich habe eine Frage, muss aber «die Antwort erraten»
- Gewichtung verschiedener Begriffe beim Ranking

uvm

Eigenschaften von natürlichsprachigem Text (Repetition)

- Wörter/Terme sind Zipf'sch verteilt, d.h.: wenige Wörter (welche?) sind sehr häufig, viele Wörter sind sehr selten
- Beispiel LA Times (TREC-Daten):

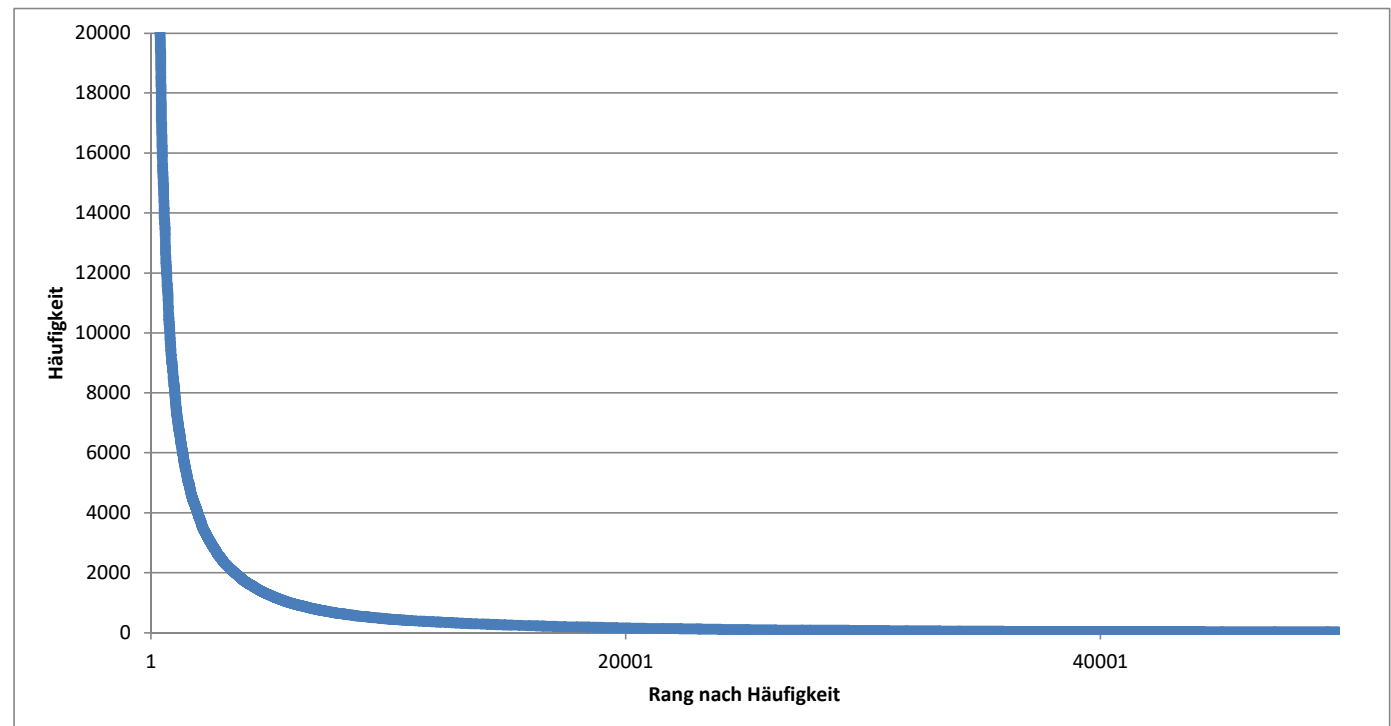
Sie sehen, dass Sie nichts sehen...



Eigenschaften von natürlichsprachigem Text

- Wörter/Terme sind Zipf'sch verteilt, d.h.: wenige Wörter (welche?) sind sehr häufig, viele Wörter sind sehr selten
- Beispiel LA Times (CLEF-Daten):

Starker Zoom



Konsequenzen der Zipf'schen Verteilung

- Die Anzahl der neuen, bisher noch nicht gesehenen Wörter* (Wortformen, Terme) pro neuem Dokument sinkt mit steigender Kollektionsgrösse.
 - Anzahl Terme in 1 Monat LA Times: ca. 88'000
 - Anzahl Terme in 6 Monaten LA Times: ca. 226'000 (d.h., ~38k/Mt)
 - Anzahl Terme in 12 Monaten LA Times: ca. 344'000 (d.h., ~29k/Mt)
- Viele der Wörter (Wortformen, Terme) sind äusserst selten
 - ca. 187'000 der ca. 344'000 Wörter in einem Jahr LA Times kommen nur einmal vor («hapax legomenon»)
 - Weitere ca. 30'000 Wörter kommen nur 2mal vor
 - Wenn wir in die Teilkollektion schauen (1 Monat), dann sind ca. 34k Wörter nur einmal enthalten – ABER: >18k dieser Wörter kommen noch mindestens ein weiteres mal in den «restlichen» 11 Monaten vor!

*Wörter ist hier nicht im linguistischen Sinn gemeint, sondern als «unique character string»

Wortformen/Morphologie

- Fürs Retrieval wären eigentlich «Konzepte» interessant – Matching sollte möglich sein, unabhängig von Wortformen.
- Typischer Ansatz im IR: Morphologische Analyse, resp. Stemming
- Wir können per Stemming abschätzen, inwiefern der Formenreichtum mit der Grösse der Kollektion zunimmt.
- In der LA Times-Kollektion findet man:
 - Durchschnittlich 1.36 Wortformen pro Stem in einem Auszug von 1 Monat Umfang
 - Durchschnittlich 1.68 Wortformen pro Stem (+24%) in der vollen Kollektion (1 Jahr)
- Es wird also wahrscheinlicher, ein beliebiges «Konzept» mit *irgendeiner* der möglichen Wortformen zu matchen
- Ähnliche Effekte sind zu erwarten für Synonyme o.ä.

Effektivitätssteigerung

- Hypothese: Die Bedeutung verschiedener «IR-Mittel», die zur Effektivitätssteigerung verwendet werden, nimmt mit zunehmender Kollektionsgrösse ab.
 - Zahlen für Stemming gemischt, aber: im Szenario, in welchem Stemming am «wichtigsten» ist (kurze Queries, kleine Kollektion), ist mit Abstand der grösste Effekt zu messen (T-Queries^{**}: P@10 +27% vs +8%)
- Hypothese: Qualität der Gewichtung nimmt zu:
 - Blind Relevance Feedback: T-Queries^{**}: 3%* vs. 23%*, TDN-Queries^{**}: kein Effekt* vs. +10%*
 - Untersuchungen zum Effekt besserer df/idf-Werte sind noch in Arbeit, Hawking & Robertson sind skeptisch

* Average Precision-Werte, ** Vergleich 1 Monat vs. 1 Jahr LATimes

Effektivitätssteigerung

- Am bekanntesten und am weitesten diskutiert ist die Auswirkung auf «Precision@10» oder «Precision@20»-Werte.
- Diese sind bedeutend besser für *grosse* Kollektionen
- Untersucht z.B. durch Hawking & Robertson 2003: «on collection size and retrieval effectiveness»
- Zusammengefasst: einerseits ein Artefakt des Masses $P@n$, andererseits eine Folge des Verhaltens eines Systems: wenn ein System für z.B. Recall = 10% eine Präzision von 40% liefert, dann wird die Präzision an einem fixen Cut-off-Point mit grösserer Kollektion steigen.
- Flapsig: mehr relevantere Dokumente → einfacher etwas zu finden



«Was sind nun Best Practices für IR auf Small Data»?

Gross vs. Klein

- Google listet >80 Mio. Treffer für «Bier»
- Wie kommen diese zustande?
- Was ist überhaupt mit der Anfrage gemeint?

80'000'000:1

Stiftungschweiz.ch – ein Portal für die Suche nach gemeinnützigen Stiftungen

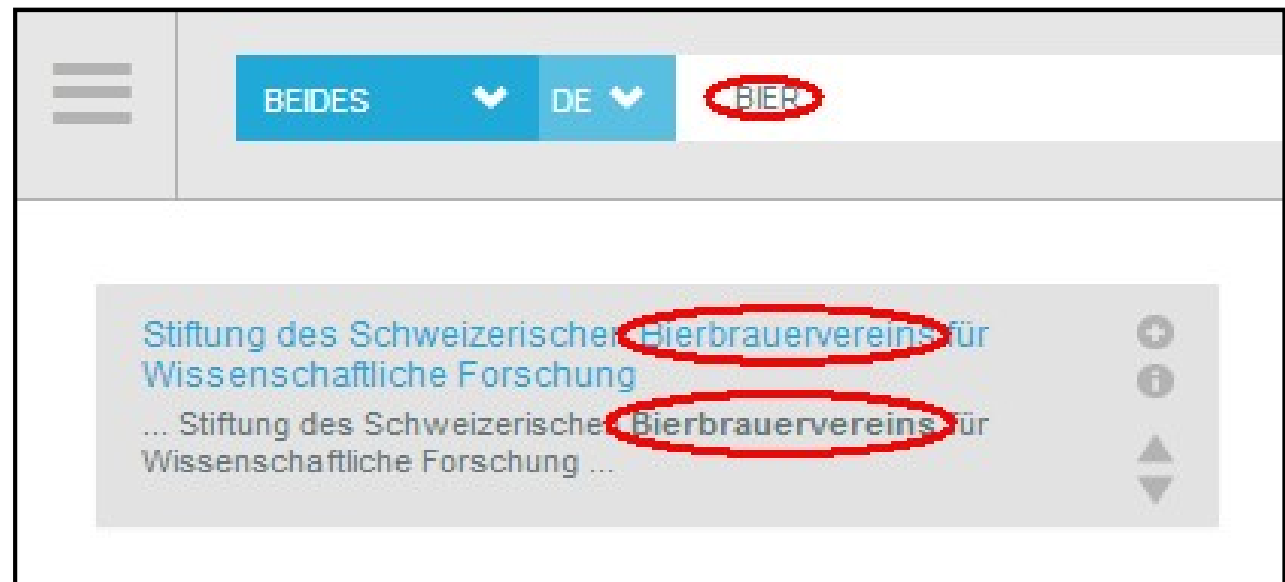
Das grösste Portal in diesem Gebiet – mit 13,000 Stiftungen

Für die Frage «Bier»: 0 Resultate, oder...?

Jeder Treffer zählt! Einsatz von Mitteln wie Stemming und Decomponing erhöht Chance auf Matches.

Experimentell:

Expansion in Sekundärquellen (z.. B. Anreicherung aus dem Web).



Dialog mit der Applikation

Es gilt zu gewichten. Anfrage: «Alkohol»:

Geht es um wissenschaftliche Fragen?

Geht es um soziale Fragen?

Geht es um Hilfe für Betroffene?

«Mein» Suchresultat muss persönliche

Präferenzen berücksichtigen können!

Die Suche in Stiftungsschweiz.ch lernt auf

Wunsch via Nutzerbewertungen

(Relevance Feedback)



Beachten Sie in diesem Kontext den Unterschied Relevance Feedback \leftrightarrow

Blind Relevance Feedback

Die Suche erlaubt auch fast beliebig lange Anfragen.

Nutzer können ganze Webseiten oder Anträge kopieren

Bessere Anfragen

Nutzer können initial keine optimalen Anfragen formulieren – hierzu müssten Sie das Resultat kennen! → Paradox

Der Dialog Nutzer/Applikation ist so zu gestalten, dass iterativ die Lösung gefunden werden kann

Gaps sind zu überwinden:

Laiensprache <-> Fachterminologie

Ausbildung/Skills <-> Stellenanforderungen

Nachfrage <-> Angebot

etc.

Fazit: Massnahmen IR auf Small Data

«Best Practices»:

- Anreiz schaffen für längere Anfragen
- Anfrage künstlich verlängern (Thesauri, Web-Anreicherung etc. – Achtung: evtl. Probleme mit Blind Relevance Feedback!)
- Zusätzliche Informationen von den Usern einholen (explizites Feedback, Profile etc.)
- Stemming/Decompounding
- Dokumente anreichern (Thesauri, Web-Anreicherung etc.)
- Nutzung von Information aus mehreren Modalitäten