Kapitel 11: Suchmaschinen

Literatur:

- RRZN Hannover: Suchen & Finden im Internet.
- Search Engine Watch. [http://searchenginewatch.com/]
- Peter Kent: Search Engine Optimization for Dummies, 2nd Ed. Wiley, 2006, ISBN 0-471-97998-8, 382 pages.
- Zoltán Gyöngyi, Hector Garcia-Molina: Spam: It's Not Just for Inboxes Anymore. Computer, October 2005 (Vol. 38, No. 10), pp. 28-34.
- Zoltán Gyöngyi, Hector Garcia-Molina: Link spam alliances. VLDB'2005: Proc. of the 31st Int. Conf. on Very Large Data Bases, pp. 517-528, 2005.
- Steve Kirsch: Interview: A Conversation with Matt Wells Queue, Volume 2, Issue 2 (April 2004), pp. 18-24, ACM, 2004.
- . . . (wird noch ergänzt)

Inhalt

Motivation (1)

- Im Januar 2005 wurde geschätzt, daß es im Web mehr als 11.5 Milliarden für Suchmaschinen relevante Seiten gibt.

  [http://www.cs.uiowa.edu/~asignori/web-size/]

- Die Frage nach der genauen Anzahl Seiten ist sinnlos: Ein Programm als Web-Server kann beliebig viele verschiedene Seiten auf Abruf ausliefern, das Web ist dann unendlich groß.

Motivation (2)

- Der Forschungsprototyp von Google hatte 1998 24 Millionen Seiten mit insgesamt 147 GB Daten.

  Brin/Page: The Anatomy of a Large-Scale Hypertextual Web Search Engine. In WWW-7, 1998. [http://www7.scu.edu.au/00/index.htm], [http://labs.google.com/papers.html]

- 2001 wurde geschätzt, daß das "Surface Web" ca. 1 Mrd. Seiten mit 19 TB Daten enthält.

  [http://www.brightplanet.com/technology/deepweb.asp]

- 2002 wurde geschätzt, daß es im Internet 3 Millionen öffentliche Web-Server gibt.

  [http://www.oclc.org/research/projects/archive/wcp/]

Motivation (3)

- Für viele Fragen gibt es im Web nützliche Informationen (gratis), aber man muß sie erst finden.

  Nicht alles ist korrekt, manches ist bewußte Fehlinformation.

- Hierzu werden häufig Suchmaschinen benutzt.

  Falls nicht bekannte URL, Einstieg über Verzeichnis, Portal, Datenbank (Amazon, Internet Movie Database), Raten von URLs.

- Informationsanbieter wollen gefunden werden (z.B. Angebote von Waren/Dienstleistungen).
- E-Commerce: 69.2 Mrd. $ Umsatz in den USA in 2004 (1.9% von allen Verkäufen). [US Census Bureau]

Bekannteste Suchmaschinen

- Google: 46% [43%] (in Deutschland 84.6%)

  Angaben von Nielson//NetRatings 11/2005. Angabe in [] von comScore Media Metrix, März 2006. Deutsche Angaben: `webhits.de`
  Einnahmen von Google 2004: 3.2 Mrd. $, Gewinn: 399.1 Mio $.
  250 Millionen Suchen pro Tag (Feb. 2003 nach SearchEngineWatch).

- Yahoo: 23% [28%] (in Deutschland 4.2%)

  Bis ca. Feb. 2004 Ergebnisse von Google übernommen. Im Dez. 2002 hat Yahoo Inktomi gekauft (235 Mio $), und hat jetzt einen eigenen Web Index. Yahoo im 2. Quartal 2004: 832 Mio $, Gewinn: 113 Mio $.

- MSN: 11% [13%] (in Deutschland: 4.6%)
- AlltheWeb, AltaVista, Ask Jeeves, Fireball, Gigablast, Hotbot, Lycos, Overture, Teoma, WiseNut.

Größe von Suchmaschinen (1)

- Abdeckung von Webseiten nach Untersuchung von Gulli/Signorini, WWW'05:
  - ◇ Google: 76.2%
  - ◇ Yahoo!: 69.3%
  - ◇ MSN: 61.9%
  - ◇ Ask/Teoma: 57.6%

  (bezogen auf die Webseiten, die von einer der vier Suchmaschinen geliefert wurden.)

  Man kann selbst einfache solche Tests machen: Man wähle seltene Suchbegriffe oder Kombinationen von Suchbegriffen und vergleiche die Ergebnisse.

Größe von Suchmaschinen (2)

- Manche Suchmaschinen geben an, wie viele Webseiten sie erfasst haben. Zwei Zählweisen:
  - ◇ Die URL ist erfasst (d.h. die Suchmaschine hat einen Link auf die Webseite gefunden).

    Aus dem Link-Text und der URL sind eventuell auch schon einige Schlüsselworte bekannt, d.h. die Seite könnte bei entsprechenden Anfragen schon geliefert werden.

  - ◇ Das Dokument selbst ist vollständig erfasst.
- Wegen der begrenzten Bandbreite besteht auch ein Konflikt zwischen Größe und Aktualität.

  Ebenso sollten Duplikate erkannt werden: Gleicher Inhalt, unterschiedliche URL. Dies würde die Größe verringern, wäre aber ein Vorteil.

Anfragen: Beispiel Google (1)

- `SQL Standard`: Liefere alle Seiten, die die Worte "SQL" und "Standard" enthalten.

  Beide Suchbegriffe müssen vorkommen, aber nicht unbedingt direkt hintereinander und nicht unbedingt in dieser Reihenfolge. Bei Google ist Groß-/Kleinschreibung egal. "ä" und "ae" werden gleich behandelt.

- `"SQL Standard"`: Die Worte müssen direkt hintereinander in dieser Reihenfolge stehen ("Phrase").

  In einer Phrase kann "*" für beliebige Worte verwendet werden.

- `Der SQL Standard`: "Der" wird ignoriert, weil es sehr häufig vorkommt (wenig spezifisch, "Stoppwort").

  Wenn man "+Der" schreibt, wird es nicht ignoriert.

Anfragen: Beispiel Google (2)

- `SQL -MySQL`: Seiten, in denen "SQL" vorkommt, aber nicht "MySQL".
- `SQL Oracle OR DB2`: Seiten, in denen "SQL" vorkommt, und Oracle oder DB2.
- `~Datenbank`: Auch Synonyme, Pluralform, etc. Z.B. werden auch Seiten mit "Database" gefunden.

  Leichte Schreibvarianten werden auch so gefunden. Google soll inzwischen ein "Stemming" durchführen (Wortstammbildung). Allerdings geben Anfragen nach Singular- und Pluralform eines Wortes unterschiedliche Ergebnisse. Ggf. mit "+" exakte Übereinstimmung verlangen.

Anfragen: Beispiel Google (3)

- `DB2 site:oracle.com`: Sucht nach Seiten in der Domain "oracle.com", auf denen "DB2" vorkommt.
- `link:http://www.informatik.uni-halle.de/~brass/`: Seiten, die auf meine Homepage verweisen.

  Google liefert hier nur drei Seiten, Yahoo 20.

- Weitere Schlüsselworte: `intitle:`, `allintitle:` (alle folgenden Worte), `inurl:`, `allinurl:`, `site:`.

  Andere Funktionen: `link:`, `related:` `cache:` `info:`, `define:`, `stocks:` Behauptet (im Netz): `inanchor:` `allinanchor:` `filetype:` `intext:` `numrange:` `pricerange:` `phonebook:` `rphonebook:` [http://www.googleguide.com/]

Anfragen: Beispiel Google (4)

- `Schema/Zustand`: Nur einige Sonderzeichen werden verstanden, andere gelten als Worttrenner.

  Z.B. liefert diese Anfrage auch Dokumente mit "Schema, Zustand" und "Schema: Zustand", es entspricht der Phrase `"Schema Zustand"`. Google ist in dieser Beziehung aber besser als viele andere Suchmaschinen, z.B. `C++` und `C#` werden verstanden.

- Weitere mögliche Einschränkungen (in erweiterter Suche): Bestimmtes Dateiformat, nur neue Seiten.

Such-Statistiken (1)

Suchthemen (2001, Excite):

24.7%	Wirtschaft, Waren, Reisen, Arbeitsstellen
19.7%	Menschen, Orte, Dinge
11.3%	nicht Englisch, unbekannt
9.6%	Computer, Internet
8.5%	Sex
7.5%	Gesundheit, Wissenschaft
6.6%	Freizeit, Unterhaltung
4.5%	Ausbildung, Geisteswissenschaften
3.9%	Gesellschaft, Kultur, Religion
2.0%	Staatliche Stellen
1.1%	Theater, Kunst

From E-Sex to E-Commerce: Web Search Changes [Computer, 3/2002].

Such-Statistiken (2)

- Anzahl Suchbegriffe pro Anfrage:

1	26.9%
2	30.5%
≥3	42.6%

Durchschnitt: 2.6

- Abgerufene Ergebnisseiten (je 10 URLs):

1	50.5%
2	20.3%
≥3	29.2%

Durchschnitt: 1.7

- Verwendung Boolescher Operatoren: 10%

Inhalt

Web Roboter (1)

- Eine Suchmaschine kann das Netz nicht für jede Anfrage neu durchsuchen, sondern läd sich alle erreichbaren Seiten lokal herunter.

  1998 hatte Google 24 Millionen Webseiten lokal kopiert, die komprimiert 53.5 GB Speicherplatz belegten (= 147.8 GB unkomprimiert).

- Das geschieht mit einem Programm ("Web Roboter", 'Crawler", "Spider"), das sich vollständig durch das Web "durchklickt".
- Ausgehend von bekannten URLs (z.B. explizit angemeldeten Seiten) werden die darin direkt oder indirekt referenzierten Seiten heruntergeladen.

Web Roboter (2)

- Beispiel (HTTP-Request von einem Web-Roboter):

```
GET /robots.txt HTTP/1.0
Host: www.informatik.uni-giessen.de
Accept: text/*
User-Agent: Slurp/si (slurp@inktomi.com;
        http://www.inktomi.com/slurp.html)
From: slurp@inktomi.com
```

- In der Datei “`robots.txt`” kann man angeben, ob Roboter auf dieser Webseite erwünscht sind, und welche Seiten sie ggf. herunterladen dürfen.

  Siehe: [http://www.robotstxt.org/wc/robots.html].
  Man kann nicht erzwingen, daß die Roboter sich auch daran halten.

Web Roboter (3)

- Web-Roboter in unserer Log-Datei (∼März 2006):
  - ◇ ConveraCrawler/0.9d (Excalibur/authoritiveweb)
  - ◇ iCCrawler (Intelligence Competence Center)
  - ◇ FAST-WebCrawler/3.8/Scirus (Scientific Inf.)
  - ◇ Googlebot/2.1
  - ◇ Yahoo! Slurp
  - ◇ Ask Jeeves/Teoma
  - ◇ msnbot/1.0
  - ◇ ZyBorg/1.0 (LookSmart/WiseNut)
  - ◇ Francis/2.0 (Neomo.de)

Web Roboter (4)

- Nach einiger Zeit müssen die Seiten neu besucht werden, um die lokale Kopie zu aktualisieren, falls sich die Website geändert hat.

  Man kann in den Log-Dateien des Servers sehen, wann eine Seite von einer Suchmaschine heruntergeladen wurde. Eine Suchmaschine wird die Seiten nicht gleich häufig besuchen: Z.B. besonders wichtige Seiten, oder Seiten, die sich in der Vergangenheit häufig geändert haben, werden in kürzeren Abständen besucht. Google: 1–7 Tage.

- Daher sind die Suchergebnisse nicht immer aktuell.

  Wenn man eine Seite gerade ins Netz gestellt hat, wird sie nicht sofort angezeigt (erst wenn der Roboter sie besucht hat). Umgekehrt werden gelöschte oder wesentlich geänderte Seiten noch eine Zeitlang angezeigt.

Web Roboter (5)

- Die akademische Version von Google (1998) benutzte vier Web Roboter, die jeweils etwa 300 Seiten gleichzeitig abfragten.

  [http://dbpubs.stanford.edu/pub/1998-8/de]

- Damit wurden 100 Seiten pro Sekunde (600 KB) heruntergeladen.

  Pro Tag also etwa 8 Millionen Dokumente (aber 100 Seiten/Sekunde war nur ein Spitzenwert). Damals hatten sie 322 Millionen URLs, aber nur 24 Millionen Seiten heruntergeladen (→ alle 40/3 Tage besuchen).

- Google berücksichtigt nur die ersten 101 KB von jeder Datei (ca. 120 KB für PDF).

Web Roboter (6)

- Seiten mit “?” in der URL werden eventuell von Suchmaschinen als kritisch angesehen.

  Zumindest bei mehr als einem Parameter. Es könnte ja beliebig viele URLs geben, zu denen Webseiten per Programm berechnet werden. Allzu dynamische Inhalte sind problematisch, da die Ergebnisse der Suchanfragen dann gar nicht mehr stimmen. Eventuell ist auch die Anzahl Seiten pro Domain/IP-Nummer begrenzt, die besucht werden.

- Um die Aktualität und den Nutzen bei begrenzter Bandbreite zu maximieren, werden nicht alle Seiten im gleichen Abstand besucht.

  Z.B. Seiten, die sich lange nicht mehr geändert haben, etwas seltener. Seiten mit hohem Pagerank oder die häufig in Anfrageergebnissen geliefert werden, etwas häufiger.

Web Roboter (7)

- URLs, die schon beim nächsten Besuch nicht mehr existieren, sind für Suchmaschinen ungünstig.

  Die Suchmaschine würde ihren Benutzern dann ja einen “broken Link” liefern. Natürlich kann sie das nicht vorher wissen. Es ist aber möglich, daß sie für eine Domain lernt, daß viele Seiten nur temporär existieren.

- Dies spricht gegen Sitzungsnummern in URLs.

  Die Suchmaschine besucht die Seiten einer Domain ja auch mit einer gewissen Verzögerung. Wenn sie versucht, Links, die sie vor einigen Tagen bekommen hat, zu verfolgen, und dann nur Fehlermeldungen bekommt, kann sie die Seiten nicht in den Index eintragen. Mindestens muß man in solchen Fällen implizit eine neue Sitzung öffnen und den eigentlichen Inhalt der Seite anzeigen. Suchmaschinen möchten aber auch nicht viele doppelte Seiten (mit unterschiedlichen URLs) haben. Auch Pagerank kann man so nicht aufbauen.

Index (1)

- Es wäre auch viel zu aufwendig, die lokalen Kopien der Seiten bei jeder Anfrage zu durchsuchen.
- Daher wird vorab eine Datenstruktur aufgebaut, in der zu jedem im Web vorkommenden Wort alle Dokumente verzeichnet sind, in denen das Wort vorkommt (Index).
- Dazu müssen Worte aus den Seiten extrahiert werden (z.B. nicht möglich in Bildern, bei von Skripten in der Seite erzeugten Texten).

  Frames sind ein anderes Problem: Inhalte unter der URL wechseln.

Index (2)

- Google hatte 1998 ein Lexikon (Wortliste, bildet Worte in interne Nummer ab) von 14 Millionen Worten (in 256 MB Hauptspeicher).

  Außerdem noch eine Liste von sehr seltenen Worten in einer Datei.

- Das Lexikon enthält für jedes Wort einen Verweis auf einen Eintrag im “Inverted Index”.

  Um Google skalierbar zu machen, ist der Index auf “Barrels” aufgeteilt, die jeweils ein bestimmtes Intervall von Dokumentnummern abdecken. Der Index für die 24 Millionen Dokumente war 41 GB groß.

Index (3)

- Im Inverted Index steht zu jedem Wort eine Liste von Dokumentnummern, die das Wort enthalten, zusammen mit den Positionen in dem Dokument.

  Für jedes Wortvorkommen wurden 2 Byte benötigt. Damit lassen sich nur 4096 Positionen unterscheiden, da auch Zusatzinformation wie die Fontgröße dargestellt wurde.

- Möglichkeiten, die Treffer für ein Wort zu sortieren:
  - ◇ Nach Dokumentnummern (gut für Phrasen)
  - ◇ Nach Wichtigkeit (gut für Ranking, s.u.).

    Lösung in akademischer Version von Google: Zwei Listen für zwei verschiedene Wichtigkeitsstufen (Vorkommen in Titel/Hyperlink vs. alle anderen). Innerhalb jeder Liste nach Dokumenten.

Index (4)

DOKUMENTE		
DokID	URL	···
1000	http://www.xy.com/	···
1001	http://www.abc.de/	···

LEXIKON	
WortID	Wort
100	Datenbank
101	Zustand

INDEX			
WortID	Vorkommen		
	DokID	Position	Attribute
100	1000	3	groß
	1000	27	
	1001	15	
101	1001	16	

Hardware von Google

- Mehrere Tausend relativ billige PCs.

  Laufen unter Linux mit eigenen Erweiterungen. Im Jahr 2000 waren es 1000 Zwei-Prozessor PCs.

- Es geht ca. einer pro Tag kaputt.

  Einmal hat es offenbar auch einen Brand in einem der Rechenzentren von Google gegeben. Google lief ohne Unterbrechung weiter.

- Eigenes Dateisystem (Google File System), das Parallelität und Ausfallsicherheit bringt.
- Eigenes Programmiermodell für Parallelität und Skalierbarkeit (Map Reduce).

Inhalt

Ergebnis-Reihenfolge (1)

- Für viele Suchbegriffe gibt es Tausende von Seiten, die den Suchbegriff enthalten.
- Nicht alle sind gleich nützlich.
- Benutzer von Suchmaschine schauen sich meist nur die ersten 10–20 Treffer der Anfrage an.
- Daher ist die Reihenfolge, in der die Ergebnisseiten angezeigt werden, wichtig ("Ranking").

  Benutzer möchten, daß für Ihr Informationsbedürfnis möglichst nützliche Seiten ganz oben stehen. Informations-Anbieter möchten ihre Seiten ganz oben haben.

Ergebnis-Reihenfolge (2)

- Ranking-Kriterien, die sich nur auf den Inhalt/Text der Seite beziehen ("on-page factors"):
  - ◇ Wo kommt der Suchbegriff vor?

    Vorkommen in Titel/Überschrift/am Anfang sind wichtiger. Eventuell auch Meta-Tags (besonders Meta-Tag und normaler Text).
  - ◇ Fontgröße/Art im Vergleich zum Rest der Seite.
  - ◇ Kommt der Suchbegriff mehrfach vor?

    Wie häufig im Verhältnis zur Länge der Seite? Zwei Vorkommen in einem kurzen Dokument sind besser als zwei Vorkommen in einem langen Dokument ("Keyword Density").
  - ◇ Falls die Anfrage aus mehreren Suchbegriffen besteht, kommen diese nah beieinander vor?

Ergebnis-Reihenfolge (3)

- Weitere Ranking-Kriterien:
  - ◇ Macht die Seite einen gepflegten Eindruck?

    Z.B. nicht viele broken Links, regelmäßige Änderungen.
  - ◇ Kommt der Suchbegriff in der Domain vor?

    Oder im Pfad auf dem Server? Wie kompliziert/lang ist die URI? Wie tief steht die Datei im Verzeichnisbaum?
  - ◇ Gibt es in der gleichen Website (Domain) noch mehr relevante Seiten?
  - ◇ Klicken Nutzer die Seite im Suchergebnis an?
- Moderne Suchmaschinen berücksichtigen viele Kriterien, und berechnen z.B. ein gewichtetes Mittel.

Manipulationen (1)

- Für Anbieter von Waren/Dienstleistungen im Web ist es wichtig, unter den ersten 10 zu erscheinen.
- Daher wird häufig versucht, das Ranking künstlich zu verbessern (Suchmaschinen-Spam).

  Nicht nur die Anbieter selbst versuchen es, sondern viele Web-Shops haben "Affiliate Programs", bei denen sie Partner für Zugriffe über die Web-Seiten dieser Partner bezahlen. Die Partner haben häufig nichts anderes als Webseiten, die über Suchmaschinen Kunden für den eigentlichen Anbieter anlocken.
- Im Februar 2006 wurde BMW aus Google wegen unerlaubter Tricks vorübergehend ausgeschlossen.

  [http://www.heise.de/newsticker/meldung/69264]

Manipulationen (2)

- Sehr viele Wiederholungen des Suchbegriffs.

  Für menschliche Leser unsichtbar gemacht durch weiße Schrift auf weißem Grund.

- Seiten, die nur den Suchbegriff enthalten (mehrfach), mit einem Verweis auf die eigentlich zu besuchene Seite (“Bridge Page”, “Doorway Page”).

  Es gibt dann eventuell eine automatische Weiterleitung auf die eigentliche Seite. Manchmal liefert der Webserver an eine Suchmaschine eine ganz andere Seite aus als an einen normalen Benutzer (“Cloaking”).

- Seiten, die beliebte Suchbegriffe enthalten, ohne Beziehung zum eigentlichen Inhalt der Seite.

Manipulationen (3)

- Es herrscht ein ständiger Kampf zwischen den Manipulatoren (“Search Engine Optimizers”) und den Betreibern der Suchmaschinen.

  Wettbewerb: Wer kommt für “Nigritude Ultramarine” auf Platz 1?

- Google soll seinen Ranking-Algorithmus alle zwei Monate ändern. Selbstverständlich sind die genauen Details der Algorithmen geheim.

- Risiko: Suchmaschinenbetreiber verwenden auch Verfahren (ebenfalls geheim), um schwarze Schafe herauszufinden und aus dem Index zu entfernen.

Manipulationen (4)

- Gegenmaßnahmen:
  - ◇ Text mit wenig Farbkontrast zum Hintergrund wird erkannt, wenn übliche HTML-Befehle.

    Aber Suchmaschinen führen kein JavaScript auf den Seiten aus.
  - ◇ Zu hohe Dichte des Schlüsselworts wird erkannt.

    AltaVista konnte schon lange nur bis zwei zählen: Noch mehr Vorkommen brachten nichts. Manche SEOs empfehlen eins in 6–7 Nicht-Stoppworten.
  - ◇ Seiten mit automatischen Verweisen werden aus dem Index ausgeschlossen.

    Wieder geht das nur für die übliche Methode (“Meta-Tag mit Refresh”). JavaScript Anweisungen werden nicht analysiert.

Inhalt

PageRank Verfahren (1)

- Idee: Seiten, auf die viele andere Seiten verweisen, sind wichtiger als solche, auf die nur wenige andere Seiten verweisen.
- Problem: Man kann sich mit einem Programm sofort beliebig viele Seiten gleichen Inhalts erzeugen lassen, die auf eine bestimmte Seite verweisen.
- Verweise von wichtigen (selbst häufig zitierten) Seiten sollten mehr wert sein als Verweise von Seiten, die niemand anders wichtig findet.

PageRank Verfahren (2)

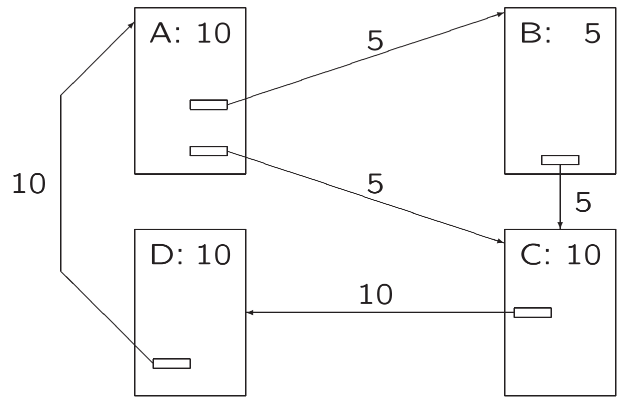
- PageRank von Lawrence Page, Sergey Brin, 1998.
  [http://dbpubs.stanford.edu:8090/pub/1999-66]
- Gleichungssystem:

$$P(s) = \frac{(1-d)}{N} + d * \sum_{i=1}^{n} P(v_i)/A(v_i)$$

  ◇ $P(s)$: PageRank der Seite s
  ◇ $v_1, \ldots, v_n$: Alle Seiten, die auf Seite s verweisen
  ◇ $A(v_i)$: Anzahl ausgehender Links in Seite v_i
  ◇ d: Dämpfungsfaktor (z.B. 0.85).
  ◇ N: Anzahl aller Seiten im Netz.

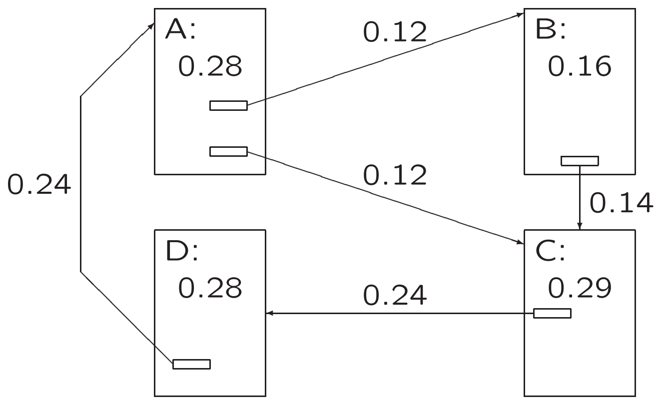
PageRank Verfahren (3)

Beispiel (ohne Dämpfungsfaktor):



PageRank Verfahren (4)

Beispiel (mit Dämpfungsfaktor $d = 0.85$, $\frac{(1-d)}{N} = 0.04$)



PageRank Verfahren (5)

```
/* Berechnung für Beispielgraph: */
d = 0.85;
N = 4;
A = 1; B = 1; C = 1; D = 1;
for(i = 1; i <= 100; i++) {
        A_neu = (1-d)/N + d * D;
        B_neu = (1-d)/N + d * A/2;
        C_neu = (1-d)/N + d * (A/2 + B);
        D_neu = (1-d)/N + d * C;
        A = A_neu; B = B_neu; C = C_neu; D = D_neu;
        printf("A=%f, B=%f, C=%f, D=%f\n",
                A, B, C, D);
}
```


PageRank Verfahren (6)

- Man kann sich den PageRank einer Seite mit dem Google-Toolbar anzeigen lassen.

  Es wird offenbar ein Logrithmus des errechneten Werts angezeigt (?)

- Der PageRank einer Seite ist völlig unabhängig vom Suchbegriff.
- Daher wird er mit einem konventionellen Ranking-Wert kombiniert.

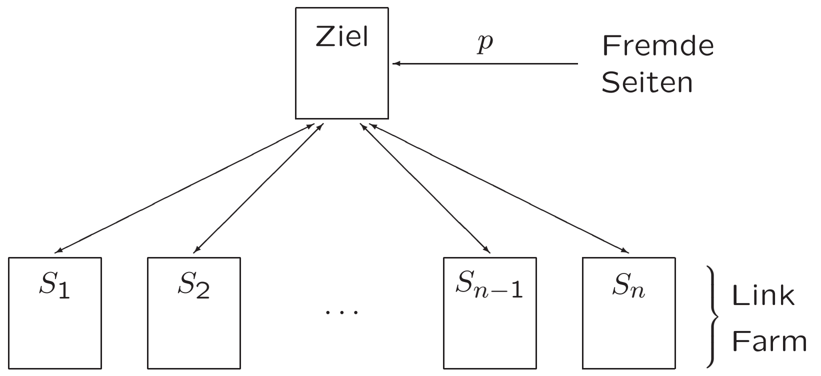
  Es werden aber ohnehin nur Seiten ausgesucht, die alle Suchbegriffe enthalten. Daher würde es auch schon Sinn manchen, unter diesen einfach die Seite mit größtem PageRank zuerst zu liefern. Das geschieht aber nicht, Google beachtet Häufigkeit, Position und Nähe von Suchbegriffen.

PageRank: Manipulation (1)

- Man braucht viele Verweise auf die eigene Seite, möglichst von hoch gerankten Web-Seiten.
- Also legt man künstliche Webseiten an, automatisch und in großer Zahl ("Link-Farm"), mit sinnlosen Wortkombinationen oder von anderen Seiten recyceltem Inhalt.

  Marc Najork: Detecting Spam Web Pages
  [http://www2.sims.berkeley.edu/courses/is141/f05/schedule.html]
  Gyöngyi/Garcia-Molina: Link Spam Alliances. In VLDB, 2005.
  [http://infolab.stanford.edu/~zoltan/publications.html]

PageRank: Manipulation (2)



$$\text{Pagerank(Ziel)} = \frac{1}{1-d^2}\left(\frac{dp + (1-d)(dn+1)}{N}\right)$$

PageRank: Manipulation (3)

- Die Struktur auf der vorigen Folie ist optimal für eine Link-Farm mit n Seiten.
- "Some of the more evil Webmasters will purchase hundreds of IPs supporting thousands of domains hosting millions of randomly generated pages. So you get this entirely artificial web community that boosts itself to the top of the results. . . . Banning the IP addresses is not enough because they move their domains around on a monthly basis."

  [Interview Matt Wells from Gigablast in ACM Queue, April 2004].

PageRank: Manipulation (4)

- Wenn eine Domain mit ehemals gutem Inhalt aufgegeben wird, kann man sie kaufen, und unter den hoch gerankten Web-Adressen Verweise auf die eigene Seite eintragen.

  Die Links auf die URL verschwinden ja nicht sofort.
- Eventuell gelingt es, Links in Gästebücher, Blogs, Foren, offene Directories, etc. einzutragen.
- Es gibt auch Link-Austausch-Programme.

"Google Bombs"

- Gibt man bei Google als Suchbegriff "Miserable Failure" ein, erhält man als erstes die Biographie von Gorge W. Bush auf dem Web-Server des Weißen Hauses.
- Diese Seite enthält den Suchbegriff nicht, aber es gibt inzwischen ziemlich viele Links mit dem Text "Miserable Failure", die auf die Seite verweisen.
- Google ordnet (häufige?) Worte in Links der Seite zu, und die Seite selbst hat einen hohen PageRank-Wert.

Hilltop Verfahren (1)

- Entwickelt von Krishna Bharat und George A. Mihaila (2000).

  [http://www.cs.toronto.edu/~georgem/hilltop/]
- Für Anfragen, auf die sehr viele Seiten passen.
- Soll von Google mitverwendet werden.

  Google hat das Patent 2003 gekauft.
- Basiert auf "Experten-Seiten" ("Link Directories"), das sind Seiten, die auf viele Seiten verweist, mit denen der Autor der Expertenseite vermutlich keine Geschäftsbeziehungen unterhält.

Hilltop Verfahren (2)

- Vermutete Geschäftsbeziehung:
  - ◇ Ähnliche Domain

    Das letzte nicht-generische Stück stimmt überein, z.B. www.ibm.com und software.ibm.co.uk.
  - ◇ Ähnliche IP-Nummer

    Unterschied nur in letzten 8 Bit.
  - ◇ Indirekt über mehrere solche Verbindungen.

Hilltop Verfahren (3)

- Nun werden Experten-Seiten ausgesucht, die sich am ehesten mit dem Suchbegriff befassen.

  Es werden Vorkommen im Titel (16 Punkte), in einer Überschrift (6 Punkte), und im Verweis-Text (1 Punkt) gezählt. Dabei werden Phrasen (Titel, Überschriften, Verweistexte) bevorzugt, die möglichst genau mit den Suchbegriffen übereinstimmen (möglichst alle enthalten und möglichst nichts sonst).
- Nach diesem Ranking der Experten werden die 200 besten Experten für den Suchbegriff bestimmt.

Hilltop Verfahren (4)

- Es werden dann nur Seiten ausgewählt, auf die zwei Experten verweisen, die nicht untereinander oder zu der Seite vermutete Geschäftsbeziehungen haben.
- Dann wird für jeden Verweis in einem Experten-Dokument ein Gewicht berechnet aus
  - ◇ dem Ranking der Experten und
  - ◇ der Anzahl der passenden Phrasen im Dokument, die sich auf den Verweis beziehen.

    Begriffe in einem Verweistext beziehen sich nur auf diesen einen Verweis. Begriffe in einer Überschrift beziehen sich auf alle Verweise bis zur nächsten Überschrift gleicher oder höherer Stufe.

Hilltop Verfahren (5)

- Die Ranking-Werte der Verweise auf eine Seite werden addiert zum Ranking der Seite.

  Bei Verweisen von Experten, die möglicherweise Geschäftsbeziehungen zu einander haben, wird nur der höhere Werte berücksichtigt.
- Falls es nicht Seiten gibt, die Verweise von zwei unabhängigen Experten haben, liefert dieses Verfahren nichts (funktioniert nur bei häufigen Begriffen).
- Google wird in diesem Fall vermutlich seinen alten Algorithmus einsetzen.

Einfluß der Suchmaschinen

- Seiten mit hohem PageRank werden über die Suchmaschinen gefunden, und bekommen noch mehr Links.
- Neue, gute Seiten werden über die Suchmaschinen nicht gefunden und bekommen nur wenig Links.
- Es wurde vorgeschlagen, die Änderung des PageRank-Wertes über die Zeit (Ableitung) in die Bewertung einer Seite mit einzubeziehen.

  Bei neuen Seiten ist die relative Änderung der Anzahl eingehender Links größer. Siehe: Cho/Roy/Adams: Page Quality: In Search of an Unbiased Web Ranking. In: SIGMOD'2005.

Inhalt

Keyword-Liste (1)

- Zunächst muß man eine Liste aller Suchbegriffe erstellen, für die man gefunden werden möchte.
  - ◇ Man beginne mit den offensichtlichen Begriffen.
  - ◇ Häufig hat man bei spezifischeren Phrasen und Kombinationen mehr Chancen.

    Viele seltenere Phrasen in der Summe so gut wie häufiger Begriff.
  - ◇ Man kann reale Anfragen anschauen, die gegebene Suchbegriffe enthalten, und so auf Ideen für Kombinationen kommen.

    Z.B. bei Yahoo Pay per click, Google. Kommerzielles Werkzeug für Entwicklung/Bewertung von Suchbegriffen: "`wordtracker.com`".

Keyword-Liste (2)

- Entwicklung einer Liste von Suchbegriffen, Forts.:
  - ◇ Keyword Tags der Konkurrenz.

    Produktnamen/Markenzeichen der Konkurrenz sind gefährlich.
  - ◇ Häufige Tippfehler.
- Es lohnt sich nicht, viel Arbeit in die Optimierung für ein Wort zu stecken, das selten gesucht wird.
- Wenn es zu einem Begriff sehr viele Seiten gibt, wobei viele den Begriff schon im Titel haben, wird es schwierig werden.

  Auch teure Anzeigenpreise für diesen Begriff sind ein Indiz.

Optimierung von Seiten (1)

- Die Suchmaschine muß die Seiten lesen können. Z.B. schlecht:
  - ◇ Text nur in Bildern.
  - ◇ Navigation nur über Javascript.
- Der Titel des Dokuments (`title`) und insbesondere sein Beginns wird hoch gerankt. Hier sollten wichtige Schlüsselworte stehen, nicht "Willkommen".

  `title` und die `meta`-Tags für `description` und `keywords` sollten in der HTML-Datei auch ziemlich am Anfang stehen, und nicht erst nach langen Stylesheets, Javascript, u.s.w.

Optimierung von Seiten (2)

- Der Text von Überschriften (`h1`, ...) wird auch hoch gerankt.

  Es wäre ungeschickt, für Überschriften in seinem Text nicht die `h*`-Elemente zu verwenden, sondern sich auf anderem Wege (z.B. mit CSS) etwas zu basteln, was wie eine Überschrift aussieht.
- Wenn einzelne Worte fett/kursiv geschrieben sind (`em`, `strong`), werden sie auch etwas höher gerankt.
- Auch Worte in Listen (`ul`, `ol`) sollen höher gerankt werden.
- Bilder sollten einen `alt`-Text haben.

Optimierung von Seiten (3)

- Der Text von Links ist wichtig. Links, die auf eine Seite verweisen, sollten die wichtigen Schlüsselworte der Seite enthalten.

  Mit "click here" Links verschenkt man diese Chance
- Jede Seite sollte auch ausgehende Links haben.
- Seiten sollten nicht sehr lang sein. Es ist besser, viele kurze als wenige lange Seiten zu haben.

  Optimal sollen 100 bis 250 Worte sein, 1000 Worte sollen aber auch noch kein Problem sein. Man kann jede Seite nur für einen oder wenige Suchbegriffe optimieren (z.B. eine Phrase).

Optimierung von Seiten (4)

- Wichtige Suchbegriffe sollten mehrfach auf der Seite auftauchen.

  Man achte auf die Schlüsselwort-Dichte. Man vermeide alles, was zu sehr nach SPAM aussieht.
- Die Suchbegriffe sollten auch in Domain- und/oder Dateinamen (URL) auftauchen.

  Man beachte, daß Unterstrich "_" und Bindestrich "-" ganz anders behandelt werden: Der Unterstrich ist Teil eines Wortes, der Bindestrich trennt Worte. Damit ist der Bindestrich meistens nützlicher (das Dokument wird auch gefunden, wenn nach den einzelnen Worten oder der Phrase gesucht wird). Man übertreibe es aber nicht mit ganz langen Dateinamen/URLs.

Optimierung von Seiten (5)

- Man sollte Frames eher vermeiden.

  Zumindest muß man berücksichtigen, daß die Suchmaschine vermutlich den Inhalt eines einzelnen Frames liefern wird, nicht die ganze Seite. Damit fehlt dem Benutzer der Kontext. Links auf bestimmte Zustände der Seite mit Frames gehen auch nicht.

- Die Länge des HTML Quelltextes sollte in einem vernünftigen Verhältnis zu den tatsächlich angezeigten Zeichen stehen.

  Sehr viel JavaScript oder Stylesheet Information vor dem eigentlichen Text sollte vermieden werden (durch Auslagerung der Scripte und Stylesheets in eigene Dateien). Durch Cut und Paste von Word soll auch viel Müll (Zusatzinformation für Formatierung) in die Seiten kommen. Man schaue sich den Quelltext an.

Links/Pagerank (1)

- Man achte auf die Eintragung in manuelle Verzeichnisse, z.B. das Open Directory Project.

  [www.dmoz.org] bzw. [dir.google.com].

- Man kann Sponsor z.B. beim W3C werden, um von dort einen Link zu bekommen.
- Es gibt Link-Tausch Programme.

  Man achte darauf, zu wem man linkt. Wenn man auf Spam-Seiten verweist, kann man für die "schlechte Nachbarschaft" mit bestraft werden. Außerdem wird jeder versuchen, möglichst wenig Pagerank aus seiner Site "abfließen" zu lassen. Links können auch so verfasst werden, daß Google sie nicht beachtet.

- Man kann Links kaufen.

Links/Pagerank (2)

- Eventuell kann man in Blogs, Foren, Gästebüchern Links hinterlassen.
- Man kann schauen, von wo es Links auf die Seiten der Konkurrenz gibt, und den jeweiligen Autor bitten, auch einen Link auf die eigene Seite zu setzen.
- Man frage Geschäftspartner (z.B. Referenz).
- Teilnahme an Wettbewerb für gute Webseiten.
- Man mache Werbung (gedruckt, Online).
- Man gebe Web-Software gegen Links ab.

Google Sitemap (1)

- Google und inzwischen alle großen Suchmaschinen verstehen Dateien mit Informationen für Suchmaschinen, die bei der Suchmaschine registriert werden müssen (Beispiel siehe nächste Folie).
- Damit kann man u.U. steuern, wie häufig eine Suchmaschine bestimmte Seiten besucht, und welche Seite sie bei mehreren Treffern einer Site anzeigt.
- Die Suchmaschine muß sich aber natürlich nicht an die Angaben halten.
- Google bietet auch interessante Statistiken.

Google Sitemap (2)

```
<?xml version="1.0" encoding="UTF-8"?>
<urlset xmlns=
      "http://www.sitemaps.org/schemas/sitemap/0.9">
   <url>
      <loc>http://www.example.com/</loc>
      <lastmod>2005-01-01</lastmod>
      <changefreq>monthly</changefreq>
      <priority>0.8</priority>
   </url>
   ...
</urlset>
```


SEO: Schlußbemerkung

- Es gibt keine schnellen, kostenlosen Tricks, die eine Nummer 1 Position auf Dauer garantieren.

  Dann würde jeder sie verwenden, und es gibt aber nur eine Nummer 1 Position.

- Man muß schon genügend Zeit und Aufwand in interessanten Inhalt investieren.

  Eventuell kann man Inhalt auch kaufen. Es wird empfohlen, pro Tag eine neue Seite zu schreiben.

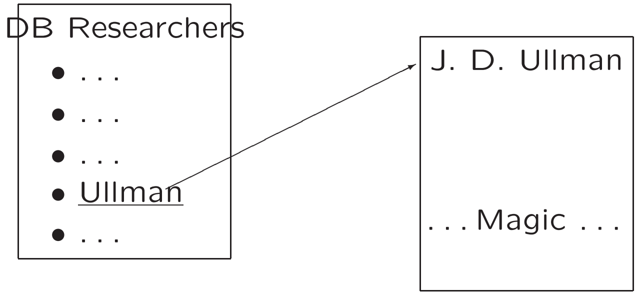
- Der Pagerank ändert sich nicht sofort.

  Er wird nur von Zeit zu Zeit neu berechnet. Google soll neue Seiten auch eine Zeitlang als kritisch einstufen ("google sandbox").

Inhalt

Suchmaschinen: Grenzen (1)

- Anfragen können sich nur auf einzelne Knoten beziehen, und nicht auf Teilgraphen des Web.



Suchmaschinen: Grenzen (2)

- Es gibt keine Möglichkeit, Wissen über das Web abzuspeichern und wiederzuverwenden.
- Informationen, die nur über Web-Formulare erreichbar sind, fehlen im Index.
- Keine Mechanismen zur Datenextraktion für (semi-) strukturierte Daten (Tabellen, Listen, XML).
- Keine beliebigen Anfragen, z.B. „Was ist die Seite in der Domain `uni-halle.de`, auf die die meisten Verweise zeigen?“

Web-Anfragesprachen

- Man nehme eine DB-Anfragesprache
  und erweitere sie um Möglichkeiten zum Zugriff auf
  - WWW-Seiten
  - Verweise zwischen Webseiten (Hyperlinks)
  - Web Indexe (Suchmaschinen)
- Beispiele:
  - WebSQL (Univ. Toronto, Mendelzon et.al.)
  - W3QS (Technion Haifa, Konopnicki/Shmueli)
  - WebLog (Concordia Univ., Lakshmanan et.al.)
  - F-Logic Web Interface (U. Freiburg, Lausen et.al.)

Ausblick

- Es werden immer neue Algorithmen gebraucht, um neue Manipulationen zu bekämpfen.
- Bessere Benutzerschnittstellen: Ziel des Benutzers verstehen, Verfeinerung der Anfrage
- Verstecktes Web (Datenbanken)
- WWW-Anfragesprachen
- Semantisches Web (XML, Metadaten, Ontologien)
- Folksonomies (geteilte Bookmarks/Anmerkungen)