Suchmaschinen im Internet

Stefan Brass
Martin-Luther-Universität Halle-Wittenberg

(Professor für Datenbanken, Certified Oracle8 DBA)

Forschungsgebiete: Deduktive Datenbanken,
Erkennung semantischer Fehler in SQL

Inhalt

Das Internet (1)

- Weltweiter Verbund von Rechnernetzen.
- Es wird geschätzt, daß 2005 ca. 1 Milliarde Menschen Zugang zum Internet hatten.

  Davon 22% in Nord Amerika, 21% Europa, 29% Asien/Pazific, 9% Japan, 20% Rest der Welt. [http://www.yahoo.client.shareholder.com /downloads/2006AnalystDay.pdf] verweist auf World Bank, CIA Factbook [http://www.cia.gov/cia/publications/factbook/], etc.
  Siehe auch: [http://www.internetworldstats.com/stats.htm].

- An das Internet sind (vermutlich) mehr als 395 Millionen Rechner angeschlossen.

  [http://www.isc.org/ds/], Januar 2006. Gezählt wurden Rechner, deren IP-Nummer in das DNS eingetragen ist.

Das Internet (2)

- Rechner werden über IP-Nummern identifiziert.
- Zu einem gegebenen Rechnernamen aus der Webadresse wird durch Anfrage an das DNS (Domain Name System) eine IP-Nummer bestimmt:

```
nslookup www.informatik.uni-halle.de
    Name: www.informatik.uni-halle.de
    Address: 141.48.3.149
```

- Zu einer Domain kann man durch Abfrage der `whois`-Datenbank Ansprechpartner herausfinden.

  [http://www.uwhois.com], [http://www.denic.de/de/whois/]. Beim Universal Whois (`uwhois`) kann man auch nach IP-Nummern suchen, sonst ist für unseren Bereich [http://www.ripe.net/whois/] zuständig.

Das Internet (3)

tracert www.informatik.uni-halle.de:

(per Modem über MSN eingewählt, siehe auch [http://www.teltarif.de])

1	96 ms	srv3.hlle.mediaWays.net
2	96 ms	62.53.166.92
3	110 ms	rmws-hlle-de04-gigaet-0-0-0-2.nw.mediaways.net
4	109 ms	rmws-brln-de11-so-0-1-1-0.nw.mediaways.net
5	137 ms	rmwc-brln-de02-gigaet-5-3-0.nw.mediaways.net
6	124 ms	rmwc-frnk-de02-so-0-0-0-0.nw.mediaways.net
7	124 ms	rmwc-frnk-de01-so-1-0-0-0.nw.mediaways.net
8	123 ms	rmws-frnk-de07-pos-6-0.nw.mediaways.net
9	124 ms	ir-frankfurt2-po5-0.x-win.dfn.de
10	138 ms	cr-frankfurt1-po8-2.x-win.dfn.de
11	124 ms	cr-leipzig1-po4-0.x-win.dfn.de
12	*	Request timed out.

Das Internet (4)

- Auf einem Rechner können viele Dienstprogramme laufen, nicht nur ein Webserver.

  Z.B. DNS-Server, EMail-Server, Datenbank-Server, `ssh`-Server.

- Das Web ist nur eine Anwendung des Internets.
- Zu welchem Programm man sich verbinden will, wird über Portnummern unterschieden, der Web-Server wartet meist auf Port 80 (oder 8080).

  Der Client (z.B. der Web Browser) bekommt auf seinem Rechner auch eine Portnummer zugewiesen. Aktuelle Netzwerkverbindungen kann man sich mit `netstat` auflisten lassen.

Das Internet (5)

- Web Browser (Internet Explorer, Mozilla, etc.) und Web Server unterhalten sich über das HyperText Transfer Protocol (HTTP).
- Der Browser schickt eine Anforderung (Request) für eine Web-Seite (oder Bild-Datei, etc.), und bekommt sie mit einigen Zusatzdaten in einer Antwort (Response).

  [http://www.ericgiguere.com/tools/http-header-viewer.html], [http://web-sniffer.net/], [http://www.rexswain.com/httpview.html]

Das Internet (6)

Beispiel für HTTP-Request (von Internet Explorer):

```
GET /index.html HTTP/1.1
Accept: image/gif, image/x-xbitmap, image/jpeg,
        image/pjpeg, application/vnd.ms-powerpoint,
        application/vnd.ms-excel,
        application/msword, */*
Referer: http://www.informatik.uni-giessen.de/.../c3.html
Accept-Language: en-us,de;q=0.5
Accept-Encoding: gzip, deflate
User-Agent: Mozilla/4.0
        (compatible; MSIE 5.5; Windows 98; Win 9x 4.90)
Host: wega.informatik.uni-giessen.de:8080
Connection: Keep-Alive
(Leerzeile)
```

Das Internet (7)

Beispiel für HTTP-Response (von Apache):

```
HTTP/1.1 200 OK
Date: Thu, 16 Nov 2000 18:52:10 GMT
Server: Apache/1.3.12 (Unix)
Last-Modified: Mon, 08 May 2000 09:22:58 GMT
ETag: "60304-46b-39168772"
Accept-Ranges: bytes
Content-Length: 1131
Connection: close
Content-Type: text/html

<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.0 Frameset//EN"
        "http://www.w3.org/TR/REC-html40/frameset.dtd">
<html><head><title>Meine Web-Seite</title>
...
```


Das Internet (8)

- Das Web besteht aus Dokumenten, die über Verweise (Links) verknüpft sind:

```
Web-Standards finden sich beim
<a href="http://www.w3.org/">W3C</a>.
```

- 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.

Das Internet (9)

- 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/]

Das Internet (10)

- 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]

Inhalt

1. Das Internet
2. Abfragen: Leistung von Suchmaschinen
3. Komponenten und Datenstrukturen
4. Ranking I: Seiten-lokale Verfahren
5. Ranking II: PageRank, Hilltop, etc.
6. Ausblick

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

- 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.)

Anfragen (1)

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

  Beide müssen vorkommen, aber nicht unbedingt direkt hintereinander und nicht unbedingt in dieser Reihenfolge. Für Google ist die Groß- und 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 (2)

- `SQL -MySQL`: Seiten, in denen “SQL” vorkommt, aber nicht “MySQL”.
- `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.
- Weitere Schlüsselworte: **`intitle: inurl: inanchor: intext: related: numrange: pricerange: filetype:`**

Anfragen (3)

- `~Datenbank`: Auch Synonyme, Pluralform, etc. Z.B. werden auch Seiten mit “Database” gefunden.

  Leichte Schreibvarianten werden auch so gefunden, aber Google führt kein “Stemming” durch (Wortstammbildung). Ggf. mit “+” exakte Übereinstimmung verlangen.
- `Schema/Zustand`: Nur einige Sonderzeichen werden verstanden, andere gelten als Worttrenner.

  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%
$\geq$3	42.6%

  Durchschnitt: 2.6
- Abgerufene Ergebnisseiten (je 10 URLs):

1	50.5%
2	20.3%
$\geq$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).

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

1. Das Internet
2. Abfragen: Leistung von Suchmaschinen
3. Komponenten und Datenstrukturen
4. Ranking I: Seiten-lokale Verfahren
5. Ranking II: PageRank, Hilltop, etc.
6. Ausblick

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 in Meta-Tags (Infos speziell für Suchmaschinen).
  - ◇ 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?

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 ("Coaking").
- 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)

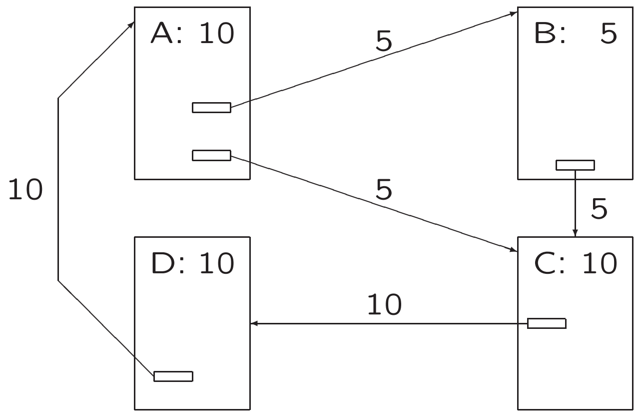
- PageRank von Lawrence Page, Sergey Brin, 1998.
  [http://dbpubs.stanford.edu:8090/pub/1999-66]
- Gleichungssystem:

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

  ◇ $P(s)$: 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.

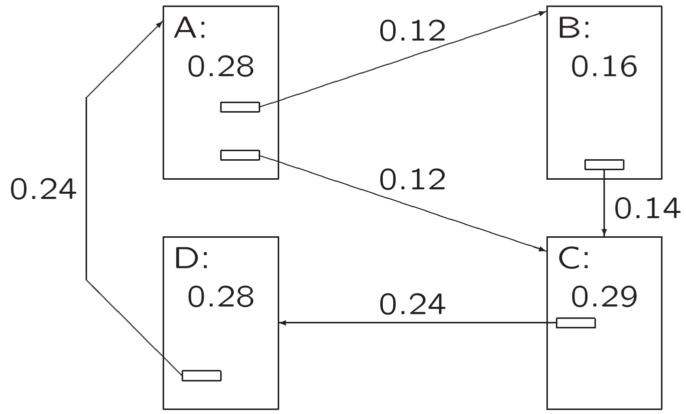
PageRank Verfahren (3)

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.

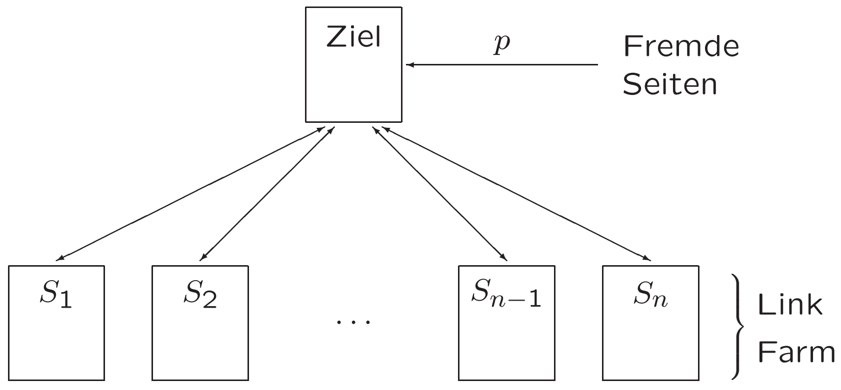
  Es werden aber ohnehin nur Seiten ausgesucht, die alle Suchbegriffe enthalten. Daher würde es auch schon Sinn manchen, unter diesen einfach die Seite mit größtem PageRank zuerst zu liefern. Das geschieht aber nicht, Google beachtet Häufigkeit, Position und Nähe von Suchbegriffen.

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

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)