Netzbasierte Informationssysteme (WS 2004/05)
World Wide Web Erschließung und Struktur

Robert Tolksdorf
Freie Universität Berlin
Institut für Informatik
Netzbasierte Informationssysteme
mailto:tolk@inf.fu-berlin.de
http://www.robert-tolksdorf.de
http://nbi.inf.fu-berlin.de

Überblick

- Crawling
- Web Größe und Struktur
- Deep Web

Crawling

Information Discovery

- Lynch, C. (1995). Networked Information Resource Discovery: An Overview of Current Issues (Invited paper). IEEE Journal on Selected Areas of Communications, 13(8):1505–1522:

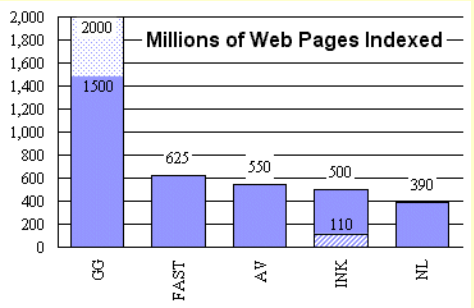
  "information discovery is a complex collection of activities that can range from simply *locating a well-specified digital object on the network* through lengthy iterative research activities which involve the *identification of a set of potentially relevant networked information resources*, the *organization and ranking resources in this candidate set*, and the *repeated expansion or restriction of this set* based on characteristics of the identified resources and exploration of specific resources."

Web Information Discovery

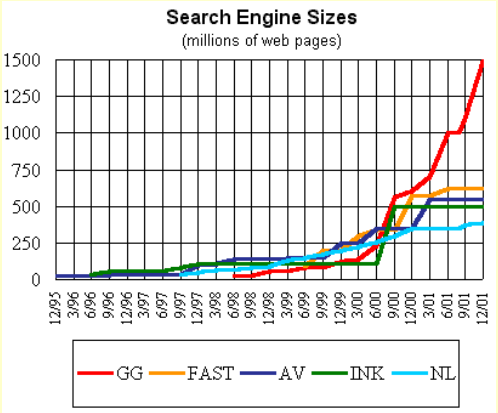
- Das Web ist
  - Verteilt
  - Dezentral organisiert
  - Dynamisch
- Resource Discovery Problem:
  Wo sind Informationsquellen von Interesse
- Lösungsidee für das Web:
  - Automatisches Navigieren über Seiten
  - Indexierung der gefundenen Seiten
  - *Crawler* (auch *Spider*, *Robot*, *Worm* etc.)

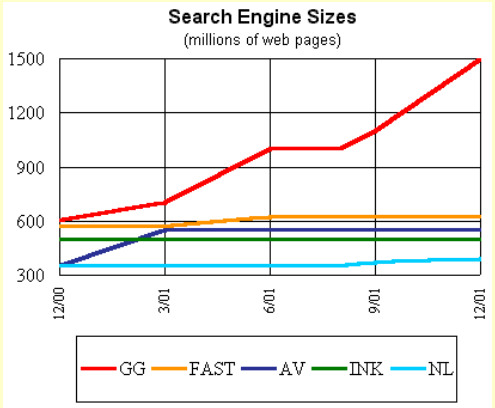
WebCrawler

- Eines der ersten Systeme: *WebCrawler* [Pinkerton94]
- Zwei Funktionen
  - Indexierung des Web
  - Automatische Navigation nach Bedarf
- WebCrawler in 94:
  - 50000 Dokumente von 9000 Quellen indexiert
  - 6000 Anfragen täglich
  - Updates wöchentlich
- Suchmaschinen heute: [Searchenginewatch]



Größenverhältnisse heute

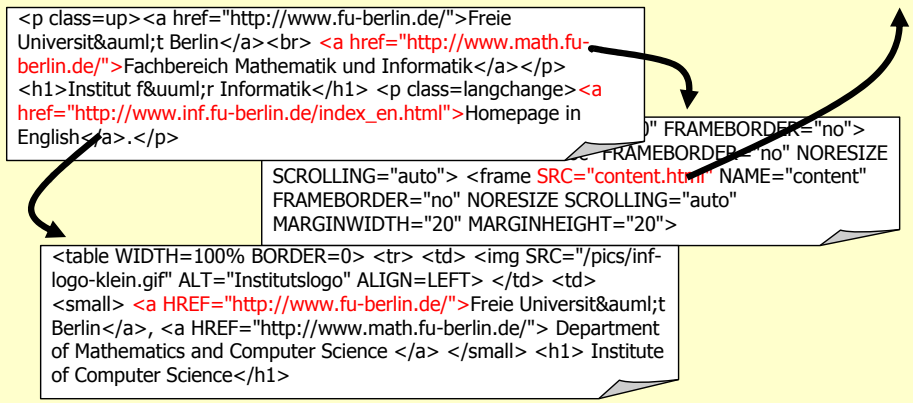




- Google 10/02: 2500m
- Google 11/04: 4200m
- GG=Google, FAST=FAST, AV=AltaVista, INK=Inktomi, NL=Northern Light
  "Sizes are as reported by each search engine and as of Dec. 11, 2001"

Crawling Algorithmus

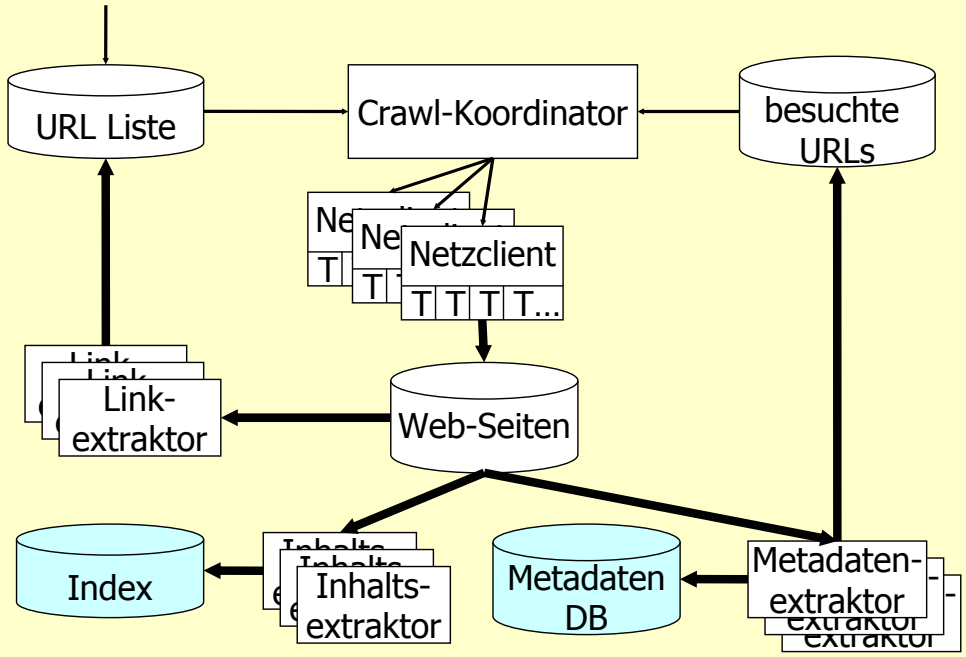
- Das Web als traversierbarer Graph von Seiten die über Links als Kanten verbunden sind
  - <a>, <link>, <meta>, <img>, <object>, <frameset>
  - FTP-Server, Adressen in nicht-HTML Dokumenten
  - ..



Crawling Algorithmus

1. URL-Liste initial füllen
2. Nehme URL aus Liste und teste
   - schon besucht?
   - passender Medientyp (html/ps/pdf/gif/...)?
   - andere Kriterien (Ort/...)?
3. hole Seite
4. extrahiere URLs und schreibe sie in URL-Liste
5. extrahiere und indexiere Seiteninhalt
6. extrahiere und speichere Metadaten
7. gehe nach 2

Einfache Architektur



Designoptionen

- Wie wird die URL Liste abgearbeitet?
  - Depth-First, LIFO Schlange:
    "Enge" Suche in die Tiefe einzelner Sites
  - Breadth-First, FIFO Schlange:
    "Breite" Suche über viele Sites, übliches Vorgehen
  - Breadth-First pro Site, FIFO-Schlagen
    Nicht mehr beliebig, aber "breit" genug

Designoptionen

- Crawl-Koordinator
  - Schon gesehen?
  - Eigenschaften der URL
    - aus .de?
  - Verarbeitbarer Filetyp?
    - HTML
    - PDF, Postscript, Word
    - Excel?
    - MP3?
  - Serverzugriff zurückstellen?
    - Kurz vorher schon zugegriffen?
    - Schon zu viel von Server geholt?
    - Zu tief von / entfernt?

Designoptionen

- Netzzugriffe
  - Wieviele Zugriffe parallel?
  - Welche Timeouts?
  - Umgang mit Fehlern
  - Verteilte Zugriffe?
- Erste Google-Versionen ca. 1997/8 (http://google.stanford.edu):
  - 3 Netzclients
  - je ca. 300 Verbindungen
  - mit 4 Clients ca. 100 Web Seiten/Minute crawlbar (144000/Tag, 6944 Tage für 1 Milliarde Seiten = 19 Jahre)
  - ca. 600Kb / Sekunde Netzlast

Designoptionen

- Link Extraktion
  - Welche Links verfolgen?
    - Art (<a>, <link>, <meta>, <img>, <object>, <frameset> etc)
    - Relevanz der Seite die sie enthält?
    - Entfernung von /
    - Angenommener Medientyp
    - Ankertext?

Designoptionen

- Inhaltsextraktion
  - Welche Teile des Inhalts indexieren?
    - Überschriften
    - Nur Ankertexte
    - Titel
    - Gesamtdokument

Designoptionen

- Metadatenermitteln
  - Welche Metadaten speichern?
    - Titel
    - Besucht
    - <meta> Tag
    - Klassifikation?

Diverse Probleme

- Relative Pfade in URLs / Gleichheit von URLs?
  Hamlet
- Framesets
- Unterschiedliche URLs für dieselbe Seite
  Sitzungs-IDs, dynamisch erzeugte Pfade
- Errechnete Links ("Next year" auf einem Kalender)
- Dynamische Seiteninhalte (Javascript etc.)
- Fehlerhafte Seiten
- Transportprobleme durch Netz
- Transportprobleme durch Größe

Crawling aus Server-Sicht

Crawler Last

- Crawler erzeugen Last beim Server
  - Verarbeitung der Anfragen
  - Auslieferung der Ergebnisse
- "Freundliche" Crawler versuchen das zu vermeiden
  - Keine fortlaufenden Anfragen zum Indexieren einer gesamten Site auf einen Schlag
  - Beachtung des Robot Exclusion Protokolls
  - Beachtung der <meta>-Tags zum Steuern von Robotern

Robots Exclusion Protokol

- Definiert einen Mechanismus mit dem ein Server festlegt, ob er von einem Crawler besucht werden will
- Daten /robots.txt auf Server
- http://www.inf.fu-berlin.de/robots.txt:

```
# robots.txt for http://www.inf.fu-berlin.de/
User-agent: *
Disallow:   /tec/net/
Disallow:   /tec/rechner/
Disallow:   /tec/software/packages/
Disallow:   /cgi-bin/
User-agent: MOMspider/1.00
Disallow:   /cgi-bin/
Disallow:   /tec/software/packages/
```

robots.txt

- `User-agent:` bezeichnet den Roboter, für die die folgenden Regeln gelten sollen
  - Namen wie (s. http://www.robotstxt.org/wc/active.html)
    - `Googlebot`
    - `Grapnel/0.01 Experiment`
    - `InfoSeek Robot 1.0`
  - Platzhalter * für alle Roboter
- Bezeichnet jeweils einen Teil der Dokumentenraums, der nicht besucht werden soll
  - Eintrag
    `Disallow:  /tec/net/`
  - `http://www.inf.fu-berlin.de/tec/net` soll nicht besucht werden

robots.txt

- Alle Roboter ausschließen:
```
User-agent: *
Disallow: /
```
- Einzelne Roboter ausschließen:
```
User-agent: Roverdog
Disallow: /
```
- Einzelne Seiten schützen:
```
User-agent: googlebot
Disallow: cheese.htm
```
- Nur einen Crawler zulassen:
```
User-agent: WebCrawler
Disallow:
User-agent: *
Disallow: /
```


<meta>-Element

- Das HTML <meta>-Tag kann ebenfalls zur Roboter-Steuerung genutzt werden
```
<html>
 <head>
  <meta name="robots"
        content="noindex,nofollow">
  <title>...</title>
 </head>
```
- Verbreitung bei Robots unklar

<meta>-Element

- `index`: Diese Seite soll indexiert werden
- `noindex`: Diese Seite soll nicht indexiert werden
- `follow`: Die Links dieser Seite weiterverfolgen
- `nofollow`: Die Links dieser Seite nicht weiterverfolgen
- `all = index,follow`
- `none = noindex,nofollow`

- Keine Möglichkeit, Verhalten für bestimmte Crawler zu bestimmen
- Kein Zugriff auf `robots.txt` notwendig

Masse (Maße) des Web

Nach: Andrei Broder, Ravi Kumar, Farzin Maghoul, Prabhakar Raghavan, Sridhar Rajagopalan, Raymie Stata, Andrew Tomkins, Janet Wiener. Graph structure in the Web. Proc. 9th International World Wide Web Conference, 2000.

Grundlage

- Analyse der Struktur des Web
- Grundlagen
  - Daten von AltaVista
  - Repräsentation des Web-Graphen als Datenbank von URLs und Links
  -

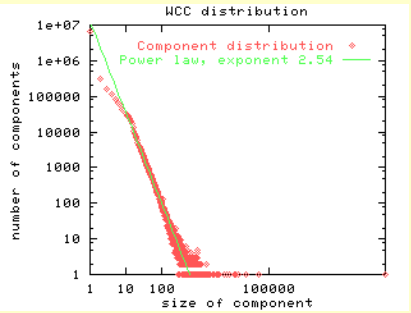
	Datum	URLs	Links
Crawl1	Mai 99	203m	1466m
Crawl2	Oct 99	271m	2130m

Power Laws

- Power Laws / Potenzgesetze beschreiben in verschiedenen Gebieten Verhältnisse zwischen Variablen:
  - Ökonomie (Pareto 1897)
  - Literaturanalyse (Yule 1944)
  - Soziologie (Zipf 1949)
  - Natur: Lawinenstärke
  - Web Charakteristiken
- Form: $y \propto x^a$ für festes a>1
- Monotone strukturlose Verteilung
- Verhältnis ändert sich nicht entlang der Größenskalen -> Skalenfreiheit
- Tritt als Phänomen an verschiedenen Stellen bei Web-Massen auf (Topologie, Nutzerverhalten etc) auf

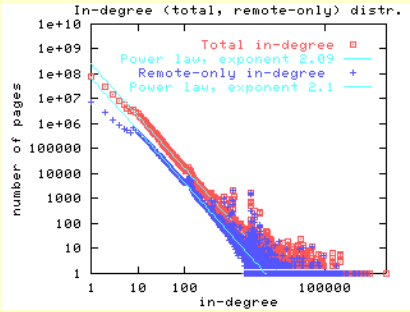
Komponenten im ungerichteten Graphen

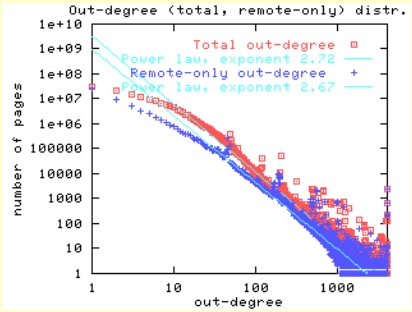
- Ungerichteter Graph (V,E) mit Kanten als {u,v}
- Pfad: $(u,u_1), (u_1,u_2), \ldots (u_k,v)$, {u,v}⇒(u,v),(v,u)
- Komponente: Menge von Knoten, so dass für Knoten u und v im Graphen ein Pfad von u nach v existiert
- Eine große Komponenten mit 186m Knoten (91%)
- Verteilung der Größen der Komponenten hat Powerlaw mit $\frac{1}{n^{2,54}}$



Messung in- und out-Degree

- Web: Gerichteter Graph (V,E), Knoten V und Kanten E, Kante ist Paar (u,v) als Verbindung von u nach v
- in-degree: $|\{(u,v_1)\ldots(u,v_k)\}|$, out-degree: $|\{(v_1,u)\ldots(v_k,u)\}|$
- Anteil der Seiten mit in-degree i proportional zu $\frac{1}{i^{2,1}}$
- Anteil der Seiten mit out-degree i proportional zu $\frac{1}{i^{2,72}}$



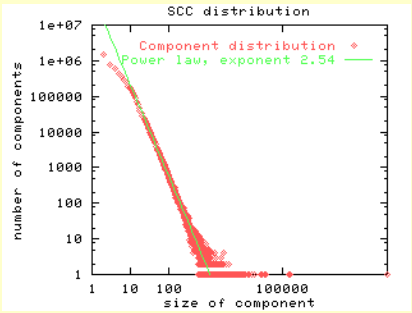


Komponenten im ungerichteten Graphen

- Hubs:
  Seiten, auf die viele verweisen (hoher in-degree)
- Autoritäten:
  Seiten, die auf viele verweisen (hoher out-degree)
- Sind Hubs und Autoritäten für die großen Komponenten verantwortlich?
- Links auf Seiten mit hohem in-degree entfernen (>5): Große Komponente mit Größe 59m Seiten
- Das Web ist auch ohne Hubs und Autoritäten gut verknüpft

Komponenten im gerichteten Graphen

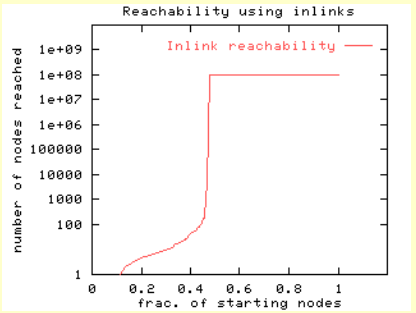
- Stark verbundene Komponente (SCC): Knotenmenge, so dass für alle u, v ein Pfad von u nach v existiert
- Eine große Komponente mit 56m Knoten (28%)
- Andere Komponenten deutlich kleiner
- Powerlaw für Größen der Komponenten mit $\frac{1}{n^{2,54}}$

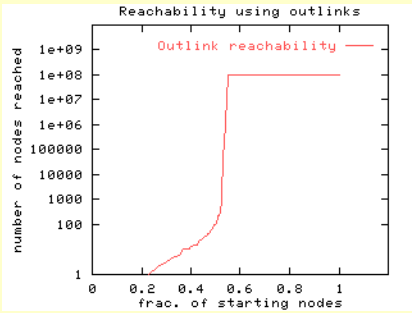


- Wo sind die restlichen 72% der Seiten?

Traversierungsmessung

- Breadth-first search (BFS): Von einem Knoten aus alle erreichbaren Knoten in Schichten nach Pfadlänge ordnen. Pfadlänge ∞ bei nicht erreichbaren Knoten
- BFS mit zufälligem Startknoten in beiden Richtungen:
  - Entweder: Ende des Algorithmus nach wenigen Knoten (<90 Knoten in 90% der Fälle)
  - Oder: Explosion zu einer Abdeckung von ca. 100m Knoten

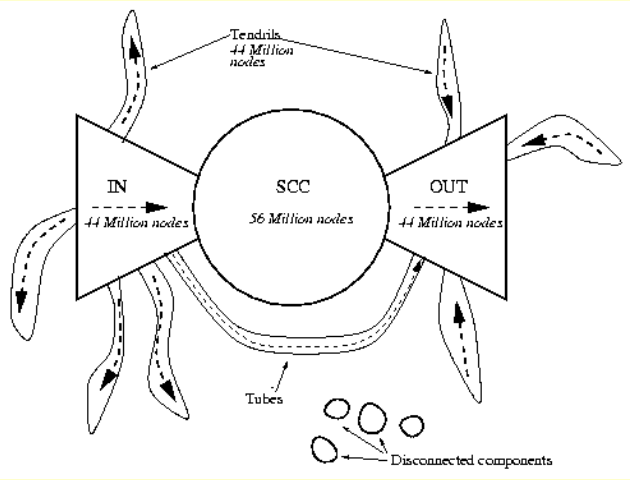




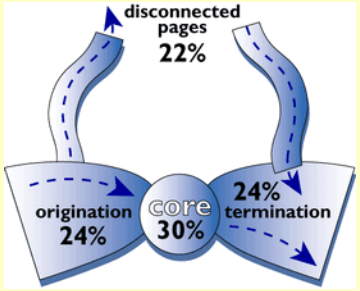
Ermittelte Struktur

- Startpunkte für BFS, die "vorwärts" explodieren sind entweder in SCC oder in einer Menge IN
- IN: Es existiert für jeden Knoten ein Pfad nach SCC
- Startpunkte für BFS, die "rückwärts" explodieren sind entweder in SCC oder in einer Menge OUT
- OUT: Es existiert für jeden Knoten ein Pfad von SCC
- Zusätzlich:
  - TENDRILS aus IN ohne SCC zu erreichen
  - TENDRILS nach OUT ohne aus SCC zu kommen
  - TUBES von IN nach OUT
  - DISCONNECTED ohne Verbindung

Struktur des Web



"Bow tie":



Region	SCC	IN	OUT	Tendrils	Disc.	Total
Grösse	56463993	43343168	43166185	43797944	16777756	203549046
Anteil	28%	21%	21%	22%	8%	100%

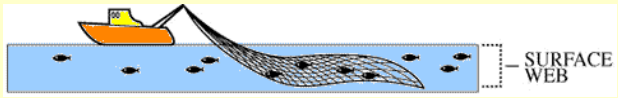
Weitere Masse

- Erreichbarkeit:
  - zwischen zwei zufällig gewählten Knoten existiert nur mit einer Wahrscheinlichkeit von 25% ein Pfad
- Durchmesser:
  - Durchmesser eines Graphen: Maximum aller kürzesten Pfade über alle Paare (u,v)
  - Durchmesser von SCC > 28
- Entfernungen:
  - Entfernung zwischen zwei Knoten ohne Berücksichtigung der Richtung von Links: 6,83
  - "Vorwärts", entlang Out-links: 16,18
  - "Rückwärts", entlang In-links: 16,12
  - Beides nur falls ein Pfad existiert (75% der Fälle nicht)

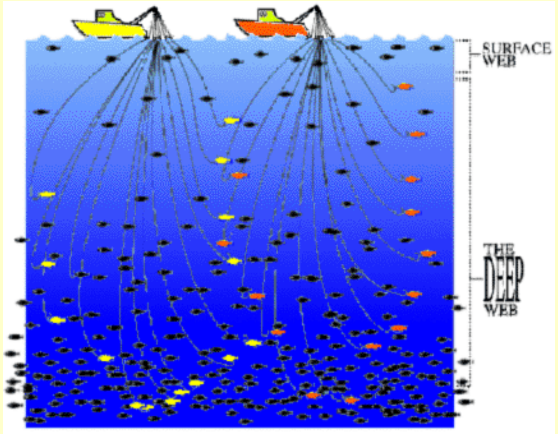
"Deep Web" Problematik

Nach: Michael K. Bergman. The Deep Web: Surfacing Hidden Value. The Journal of Electronic Publishing August, 2001 Volume 7, Issue 1 und http://www.brightplanet.com/deepcontent/tutorials/DeepWeb

"Deep Web"-Argumentation



- Traversierung des Web über Links führt nur zu einem Bruchteil der Informationen
- "Deep Web" wird von Datenbankinhalten gebildet
- Umfang 400-500 mal größer als "normales" Web
- 500Mrd Dokumente vs. 1 Mrd Dokumente
- Zugriff aber nur durch Datenbank-anfragen möglich



Deep Web Studie

- 100 Sites analysiert
  - Schätzung der enthaltenen Datensätze oder Dokumente
  - Abfrage von Stichprobe von 10 Dokumenten zu Größenabschätzung durch Mittelwertbildung
  - Indexierung und Klassifizierung des Suchformulars
- Größenschätzung
  - Nachfrage bei Betreibern
  - Aussagen auf Site
  - Aussagen über Site in anderen Berichte
  - Zahlen bei Suchantworten, z.B. Treffer für "NOT sfgjsljffjd"
  - Ausschluss aus Untersuchung
- Schätzung: Durchschnittlich 74,4 MB pro Site

Größenschätzung Sites des Deep Web

Name	Type	Web Size (GBs)
National Climatic Data Center (NOAA)	Public	366,000
NASA EOSDIS	Public	219,600
National Oceanographic (combined with Geophysical) Data Center (NOAA)	Public/Fee	32,940
Alexa	Public (partial)	15,860
...	...	...
Subtotal Public and Mixed Sources		**673,035**
DBT Online	Fee	30,500
Lexis-Nexis	Fee	12,200
Dialog	Fee	10,980
Genealogy - ancestry.com	Fee	6,500
ProQuest Direct (incl. Digital Vault)	Fee	3,172
...	...	...
Subtotal Fee-Based Sources		**75.469**
Total		**748,504**

Anzahl von Sites des Deep Web

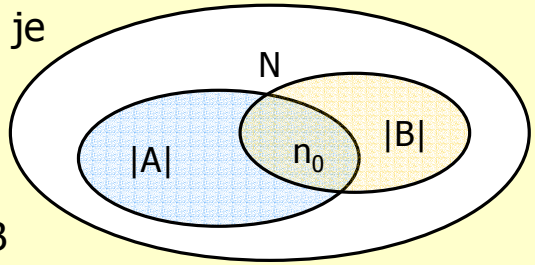
- Manuell und teilweise automatisch unterstützt:
  - 53220 URL-Hinweise aus anderen Sites
  - 45732 ohne Duplikate
  - 43348 nochzugängige
  - 17579 anscheinend suchbare
  - 13,6% davon nicht suchbar

Schätzung Anzahl der Sites

					DB A			Tot Est Deep Web
DB A	**A no dups**	**DB B**	**B no dups**	**A+ B**	**Uniq.**	**DB Fract.**	**DB Size**	**Sites**
Lycos	5,081	Internets	3,449	256	4,825	0.074	5,081	68,455
Lycos	5,081	Infomine	2,969	156	4,925	0.053	5,081	96,702
Internets	3,449	Infomine	2,969	234	3,215	0.079	3,449	43,761

- Schätzung: Ca. 100000 Deep Web Sites

Overlap analysis: Gesucht N - Größe des Deep Web



- n_A, n_B Abdeckung durch je eine Suchmaschine / ein Verzeichnis
- n_0 Überlappung
- |A|, |B|: Größe von A, B
- p(A): Wahrscheinlichkeit, Seite von A gefunden wird
- $p(A \cap B)=p(A)*p(B)$
- $|A|=N*p(A)$, $|B|=N*p(B)$, $|A \cap B|=N*p(A \cap B)$

- $N=|A|*|B|/|A \cap B|$

- Da Verzeichnisse nicht zufällig: Untere Genze

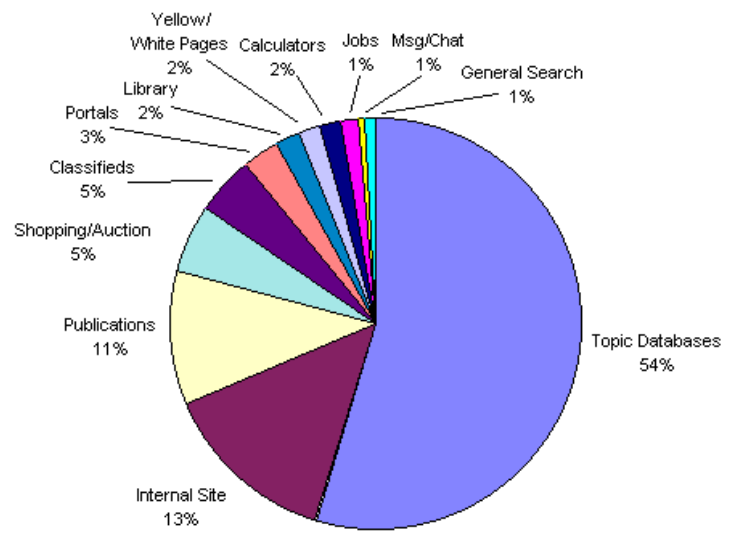
Inhaltsanalyse

- Inhaltsüberprüfung durch Anfragen aus 20 Gebieten
- Typanalyse durch Handauswertung von 700 Sites

Inhaltsklassifikation

Agriculture	2.7%	Law/Politics	3.9%
Arts	6.6%	Lifestyles	4.0%
Business	5.9%	News, Media	12.2%
Computing/Web	6.9%	People, Companies	4.9%
Education	4.3%	Recreation, Sports	3.5%
Employment	4.1%	References	4.5%
Engineering	3.1%	Science, Math	4.0%
Government	3.9%	Travel	3.4%
Health	5.5%	Shopping	3.2%
Humanities	13.5%	Law/Politics	3.9%

Site-Klassifikation



Vergleiche

- Deep Web: 7500 Terabytes, Web: 19 Terabytes
- Deep Web: 550 Mrd Docs, Web: 1 Mrd Docs
- Mehr Traffic auf Deep Web Sites (50%)
- Mehr Wachstum im Deep Web
- Deep Web Sites mehr inhaltliche Tiefe und weniger inhaltliche Breite
- 95% des Deep Web frei zugänglich

- Probleme:
  - Intention der Deep Web Studie
  - Erschließung?

Literatur

- Brian Pinkerton. Finding What People Want: Experiences with the WebCrawler. Second International World-Wide Web Conference: Mosaic and the Web, Chicago, IL, October 17--20 1994. http://www.ncsa.uiuc.edu/SDG/IT94/Proceedings/Searching/pinkerton/WebCrawler.html
- www.searchenginewatch.com
- The Web Robots Pages. www.robotstxt.org