Wolfgang Sander-Beuermann:

Vortrag zur InetBib2000:

- Dortmund, 9.3.2000 -

"Neues und Megatrends bei Suchmaschinen"

Inhalt:

Vortragsunterlagen im WWW unter: http://metager.de/inetbib2000/

next

Who am I?

Wolfgang Sander-Beuermann, Dr.-Ing.

Universität Hannover

**Regionales Rechenzentrum für Niedersachsen, RRZN &
Lehrgebiet Rechnernetze und Verteilte Systeme, RVS**

Arbeitsgebiete:

- **Administration/Koordination WWW-Server Uni Hannover**
- **Suchmaschinen im Internet (u.a. Harvest, Meta-Maschinen, Level3)**
- **Projektleiter Suchmaschinen-Labor**

Adressen:

WWW: http://www.rrzn.uni-hannover.de/webadmin/

E-Mail: wsb@rrzn.uni-hannover.de

start

Definitionen:

- **Suchdienst: beliebiges Suchangebot im Internet (Oberbegriff)**

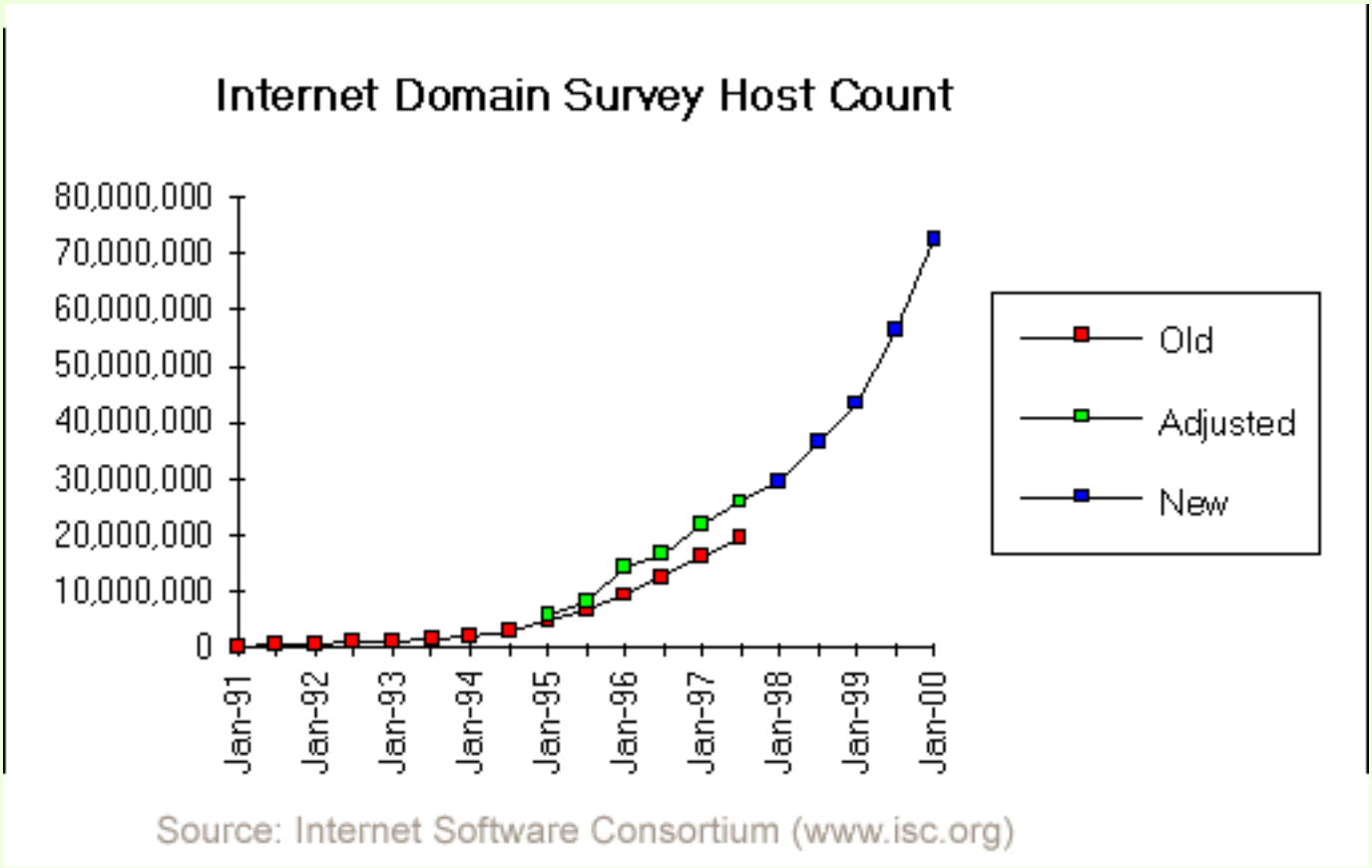
  - **Katalog/Liste: manuell erstelltes Suchangebot**
    - **Beipiele: Yanoff-Liste (veraltet), Karlsruhe Virtual Library, DINO**
  - Suchmaschine: automatisiert erstelltes Suchangebot
    - **Beipiele: Altavista, Crawler.de, DINO**
  - All-in-One-Formulare: ermöglichen Absuchen vieler Suchdienste nacheinander über einheitliche Eingabemaske
    - **Beispiele: CUSI, Klug-Suchen**
  - Meta-Suchmaschine / Metamaschine / MetaSucher / MetaCrawler / Multisearcher / ParallelSearcher:

    Suchangebot, welches (prinzipiell) beliebig viele Suchdienste parallel absucht und die Ergebnisse aufbereitet

    - **Beispiele: MetaCrawler, Highway61, MetaGer**

DER Megatrend

Die "Mutter aller Trends" im Internet: -> das *Wachstum*



Davon z.Zt. ca. 10 Mio Webserver

-> zu indexierendes Textvolumen: einige 10.000 GByte

start

Abdeckungsgrad der Suchmaschinen wird immer geringer

international:

April 98 - max.: Hotbot 34 %

Feb. 99 - max: NorthernLight 16 %

Quellen: NEC-Studies
Feb. 99: ca. 800 Mio. Webseiten, Wachstum: 1 Mio/d

deutschsprachig:

unbekannt - Studie läuft

Suchmaschinen: höhere Last von *zwei* Seiten:

- **höheres Indexierungsvolumen**
- **höhere Abfragezahlen (Fireball: ca. 1 Mio Queries/d)**

Inflation der Metasucher

Geringer Abdeckungsgrad Einzel-Maschine: -> Meta-Suche

2 Arten:

- **Client-based (Installation, Updates, Last-Mile Problem etc.)**
- **Server-based (hier betrachtet)**

Abdeckungsgrad:

- **international: (NEC-Study) -> bis zu 50%**
- **deutschsprachig: -> unbekannt (80%?)**

ca. 12 deutschspr. / 30 internat. Metasucher

---

scheinbar(!) techn. einfacher realisierbar - Qualität???

Meta-Suchmaschinen, Kriterien

Auszug aus: Proceedings Internet Summit of the InternetSociety, July 21-24, 1998, Genf, W.Sander-Beuermann, M.Schomburg "Internet Information Retrieval: The Further Development of Meta-Searchengine Technology"

1. **Parallele Suche (keine all-in-one Forms)**
2. **Ergebnis-Merging**
3. **Doubletten-Eleminierung**
4. **mindestens AND und OR Operatoren**
5. **Übernahme Kurzbeschreibung**
6. **Searchengine hiding**
7. **Möglichkeit vollständige Suche**

Meta-Searchengine	parallel	merge	noDouble	AndOr	descr.	hide	complete
metasearch.com	no	-	-	-	-	-	-
www.digiway.com/digisearch	yes	no	no	yes	yes	no	no
search.onramp.net	yes	yes	yes	no	no	yes	no
www.designlab.ukans.edu/profusion	yes	yes	yes	yes	yes	no	no
search.cyber411.com	yes	no	no	no	no	yes	no
search.metafind.com	yes	yes	yes	yes	no	partly	no
www.inference.com/infind	yes	partly	yes	yes	no	no	no
www.dogpile.com	yes	no	no	yes	no	yes	no
www.mamma.com	yes	yes	no	yes	yes	yes	no
guaraldi.cs.colostate.edu:2000/form	yes	no	no	yes	yes	yes	no
www.metacrawler.com	yes	yes	yes	yes	yes	yes	no
mesa.rrzn.uni-hannover.de	yes	yes	yes	yes	no	yes	no
meta.rrzn.uni-hannover.de	yes	yes	yes	yes	yes	yes	yes
www.highway61.com	yes	yes	yes	yes	yes	yes	yes

start

MetaGer: www.metager.de

- **bis zu 21 Suchdienste parallel absuchbar**
- **parallel dazu: lokale Resourcen werden abgesucht (QuickTips)**
- **konsistente Syntax (AND, OR, String)**
- **"verdeckt" Suchdienst-Unterschiede**
- **maximale Antwortzeit voreinstellbar**
- **Treffer nachlieferbar, solange Suchdienste liefern -> vollständig**
- **Link-Überprüfung optional, Aktualitätssortierung**
- **Ranking: die "besten Treffer" zuerst**

---

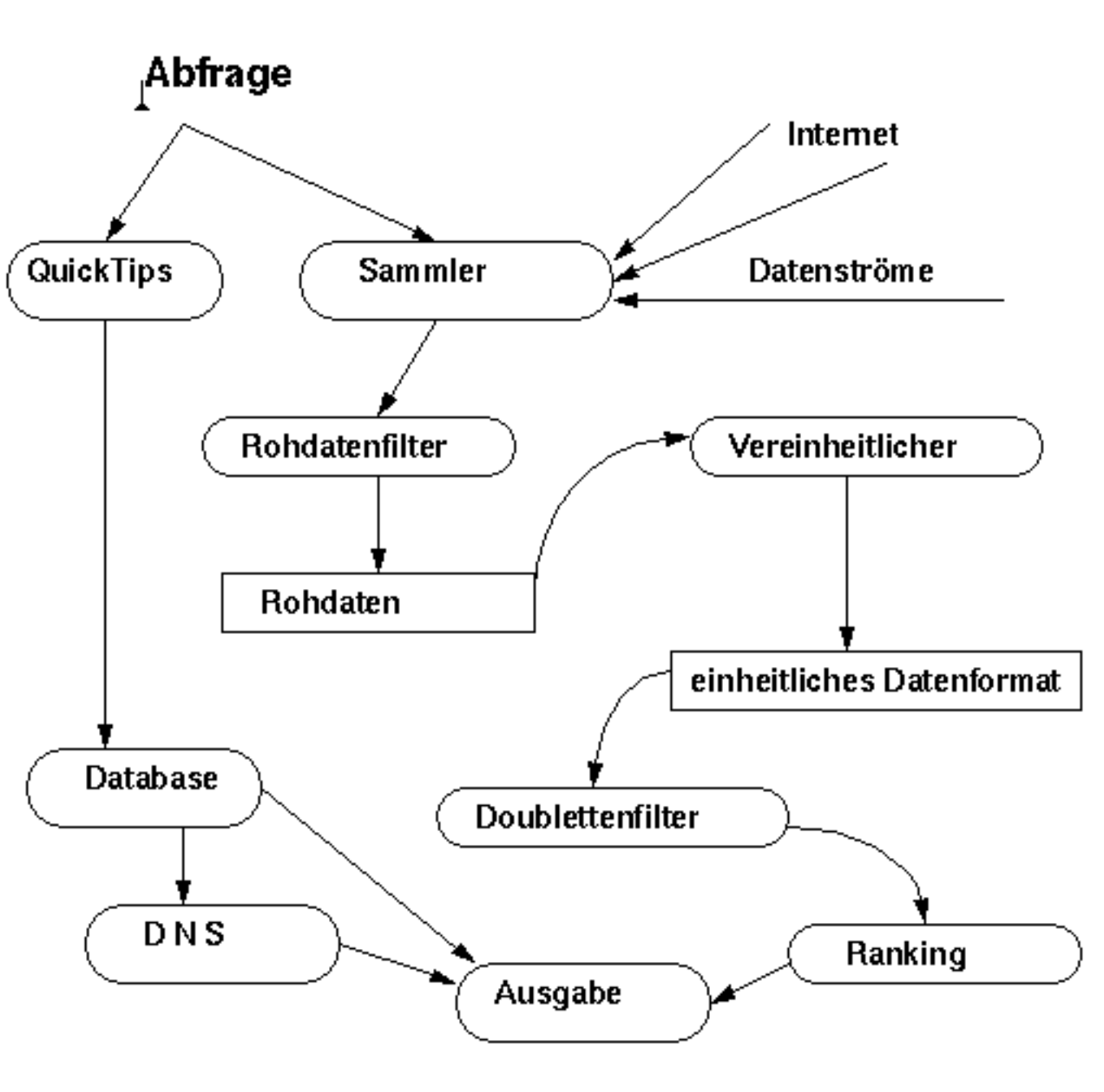
derzeit ca. 220.000 Abfragen/Queries pro Tag

-> Funktionsweise

Wie funktioniert MetaGer?

---

eigentlich ganz gut ;-) in grober Näherung so:



start

Zahl der "Bad-Guys" steigt, Attacken

Plausibilitätsansatz:

- **ca. 1 Promille aller Menschen haben "kriminelle Energie"**
- **ca. 500 Mio. Menschen "bevölkern" das Internet**

---

ca. 500.000 Menschen mit "krimineller Energie" im Internet

rapide Zunahme von Attacken:

1. **Ausnutzung von System-Fehlern/Schwächen**
2. **"Zumüllen", Denial-of-Service/DOS-Attacken**

"08. Februar 2000: Vandalen legen Yahoo lahm
Ein Alptraum wurde am Montag für Yahoo.com wahr: Durch einen Angriff war das Angebot vier Stunden lang nicht erreichbar. Saboteure hatten die Site mit gefälschten Anfragen überflutet"

-> Anti-Attacking-Programme

MetaTags

erfunden zur besseren inhaltlichen Erschliessung
`<meta name="keywords" content="...">`

rapide Steigerung in der Nutzung:

- **vor 1,5 a: ca. 5%**
- **derzeit : ca. 25 % (!)**

ABER: ca. 90% der MetaTag-Nutzer missbrauchen die Keywords
(Keyword-Spamming: Wortwiederholungen, irreführende Worte)

-> SPAM-Filter

Inhalts-Betrug steigt

Beispiel:
Suche auf T-Online Homepage SuMa nach -> Samuel Becket

"Treffer" Nr. 1 - http://home.t-online.de/home/strip/sexy_1.htm

WARNUNG: bei Anklicken -> per JavaSkript unzählige neue Fenster

HTML-Code der Seite:

```
<BODY BGCOLOR="#FFFFFF"

<FONT COLOR="#FFFFFF"> books star wars ccg star wars characters star wars chat star Diablo Dictionary Word Files ... Samuel Becket ... usw.
```

Härteste Betrugsmethode:

Nach Crawlen des Webservers durch SuMa:

kompletter Austausch des Inhaltes

-> Gegenmassnahmen

Massnahmen gegen Suchmaschinen-Betrug

"Page-Jacking"

- juristische, bisher nur USA bekannt
- SuMa-technische:
  - URL-Blacklists, wo ???
  - Real-Time-Prüfung der Treffer
    - aller Treffer: MetaGer2 -> unrealistisch
    - Real-Time-Prüfung einzelner Treffer:

      MetaGer: QuickCheck, QCheck:

      überprüft innerhalb von Sekunden:
      1. ob eine WWW-Seite existiert,
      2. ob die Seite Ihr/e Suchwort/e enthält, und versucht
      3. einen kurzen Text-Extrakt zu erstellen

kriminelle Inhalte: Filter?

2 Arten von Filtern:

1. **manuelle -> URL-Blacklists** (gibt es bisher nicht)
2. **automatische: Worterkennungen und Algorithmen**
   - **funktioniert nur bereichsweise** (z.B. Kinderpornographie)
   - **einen Automaten, der in der Lage wäre, kriminelle Inhalte zu erkennen, kann es nicht geben, solange es keinen Automaten gibt, der die Semantik eines Textes erkennt**

Massive juristische Probleme: unterschiedliche Rechtsnormen

"Schuster bleib bei Deinen Leisten"

"Jurist bleib in Deinem Lande"

---

oder: "schaffe eine globale Rechtsnorm" (Minimalkonsens)

Kommerzialisierung, Portale

Erkenntnis des Jahres 1999: man kann im Internet (doch) Geld verdienen,
v.a. mit Werbung

- **Umsätze mit Werbung 1999 in Deutschland: verdoppelt.**
- **Einnahmen ca. 150 Mio DM.**

Was macht Werbung im WWW so attraktiv?

1. **genaue Messbarkeit eines Erfolges (Klickrate)**
2. **zielgruppengenaue Werbung möglich (Keyword-Advertising)**

Top-Sites bauen zu "Portalen" aus:

möglichst viele Services unter einer Adresse

Suchmaschine + FreeMail + ShoppingCenter usw.

Das Potential: TKP, PIs

- **PIs: PageImpressions, AdImpressions, PageViews, Sichtkontakte**
- **TKP: Tausender-Kontakt-Preis: Anzahl DM pro 1000 PIs**
  - **derzeit ca. 5 - 50 DM für Bannerwerbung**
  - **50 - 300 DM für Keyword-Advertising (Suchmaschinen)**

Beispiel:
Website mit 100.000 PIs/Tag:
-> 100 x 25,-DM x Bannerzahl(=2) -> maximal um 5000 DM/Tag

-> Umsatz ca. 2 Mio/a

start

Reale Messgrössen, IVW

gemeinnütziger Verein: "Informationsgemeinschaft zur Feststellung der Verbreitung von Werbeträgern e.V." -> www.ivw.de

"von der Werbewirtschaft autorisiertes Zentralorgan zur Messung von PIs"

Site	PIs/Monat	BannerUmsatz p.a. Größenordnung geschätzt
Allesklar	2.898.430	2 Mio DM
Infoseek	46.955.765	28 Mio
Fireball	68.666.296	41 Mio
GMX	159.129.539	100 Mio

start

Zusammenfassung

1. **Das Internet-Wachstum ist nach-wie-vor ungebrochen**
2. **Die aktuellen Trends der Suchmaschinen sind eine Folge dessen:**
    1. **positive Folge: monetärer Gewinn, Investitionsbereitschaft, Innovationsbereitschaft wächst**
    2. **negative Folge: der "Müll der Informationsgesellschaft" wächst ebenso**

Die Notwendigkeit guter Filter/Suchmaschinen ist wichtiger als je zuvor