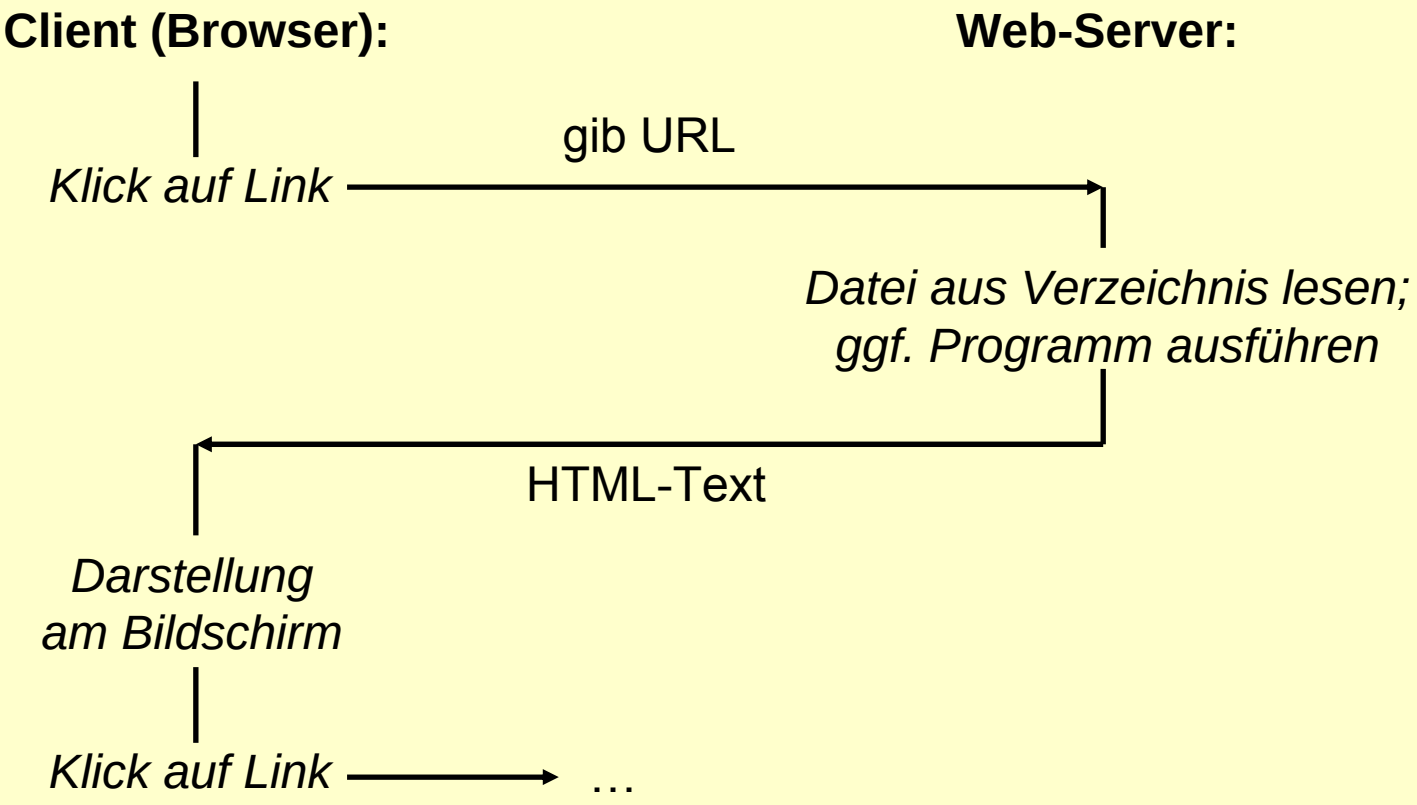
Alles nur Google?
Das Innenleben der Suchmaschinen

Prof. Dr. Klaus Meyer-Wegener
Friedrich-Alexander-Universität
Technische Fakultät
Institut für Informatik

1. Das World-wide Web (WWW)

- oft auch "Internet" genannt
  - aber das ist nur die Kommunikations-Infrastruktur
- Bestandteile:
  - **Server**: Rechner mit Internet-Anschluss, die auf Anfrage bestimmte Dateien (mit HTML-Inhalt) versenden
  - **Clients** (Browser): stellen solche Dateien am Bildschirm dar und richten bei einem Klick auf einen Link eine Anfrage an einen Server

Das World-wide Web (2)



Das World-wide Web (3)

- Jeder Rechner kann ein Server sein.
  - Es gibt inzwischen ca. 3 Millionen davon.
- Auf diesem Wege sind Unmengen von Dokumenten verfügbar.
  - nützliche wie weniger nützliche …
- Nachdem zunächst allein die **Navigation** (Klicken auf Links, "Hypertext") benutzt wurde, entstand sehr schnell der Bedarf, in diesen Dokumentmengen zu **suchen**.

Suche im WWW

- zwei grundsätzliche Verfahren:
  - **Directories** (Verzeichnisse, Link-Listen)
    - manuell gepflegt
    - Suche über hierarchische Klassifikationen
    - Hauptvertreter: Yahoo
  - **Suchmaschinen**:
    - automatisch ergänzt durch sog. Crawler
    - Suche über Schlagworte
    - Hauptvertreter: Google

Suchmaschinen

- 1993
  - MIT-Student Matthew Gray entwickelt den ersten Webcrawler: World Wide Web Wanderer
  - Martijn Koster entwickelt ALIWEB: indexiert Webseiten mit Metadaten nach Anmeldung
  - Suchmaschine **Excite** wird von sechs Studenten der Stanford University ins Netz gestellt
- 1994
  - Jerry Yang und David Filo entwickeln **Yahoo**: einen Verzeichnisdienst

Suchmaschinen (2)

- 1994
  - **WebCrawler** (Brian Pinkerton, Univ. of Washington) startet: Volltext-Indexierung
  - **Lycos** (Michael Mauldin, Carnegie Mellon): ausgefeiltere Ranking-Algorithmen
- 1995
  - **Altavista** startet: die erste Suchmaschine, die versucht, das gesamte WWW zu erfassen

Suchmaschinen (3)

- 1997
  - **AskJeeves**, **Northern Light**, **MetaGer** u.v.a.m. starten
- 1998
  - **Google** geht ans Netz

Quelle: W.Sander-Beuermann, Universität Hannover, RRZN, SuchmaschinenLabor, SuMa e.V.

Was machen Suchmaschinen?

- Das WWW ständig nach neuen Dokumenten absuchen
  - Analyse: Was steht in dem Dokument?
  - Stichworte: im Text vorkommende Worte
  - in einen großen **Index** aufnehmen
  - liefert bei Suche nach dem Wort alle Dokumente, die dieses Wort enthalten
- noch zu einfach …

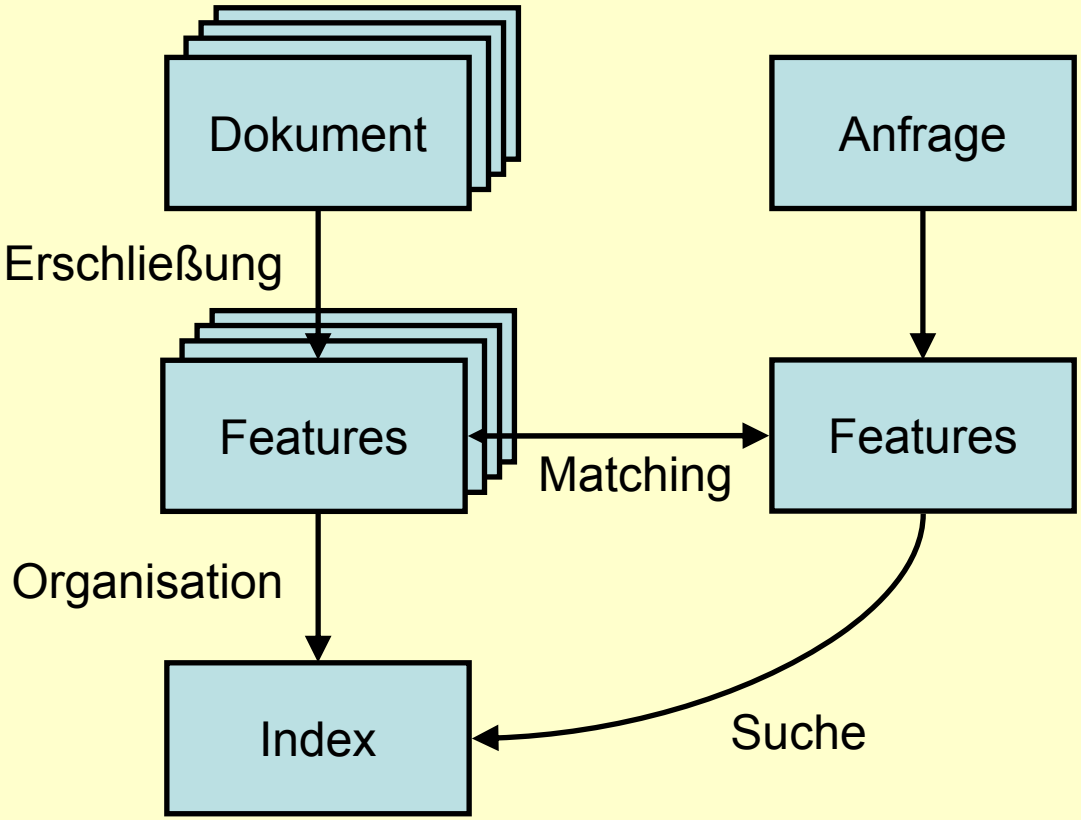
2. Die Technik im Hintergrund

- Suche nach Dokumenten gab es auch schon vor dem WWW.
- **Information-Retrieval-Systeme** (IRS)
  - deutsch: Informationsrecherche oder Informationswiedergewinnung

Information Retrieval

- semistrukturierte Daten:
  - vor allem Texte,
    aber auch Bilder, Tonaufnahmen, ...
  - Struktur (Titel, Autor, Kapitel, …) implizit
- unscharfe Suche
  - über Ähnlichkeit und Relevanz
- **Ranking** (Ordnung, Sortierung):
  - Dokumente passen mehr oder weniger
  - Anordnung nach (geschätzter) Relevanz
  - Ergebnis im Prinzip immer alle Dokumente

Überblick


Dokument
Anfrage
Erschließung
Features
Matching
Features
Organisation
Index
Suche

2.1 Dokumenterschließung

(= Vorbereitung für die Suche)

- **manuell**
  - (klassisches Bibliothekswesen)
  - **Schlagworte, Stichworte**
  - Kataloge
- **Thesaurus**
  - vorgegebene Menge von **Deskriptoren**
  - andere Wörter sind Nicht-Deskriptoren
    - verweisen auf zu verwendende Deskriptoren:

      `Datenbanksysteme, DBS: `*`use`*` Datenbanken`

Dokumenterschließung (2)

- **automatisch**
  - Extraktion von "**Features**" (Deskriptoren)
  - bei Texten zunächst Betrachtung der einzelnen Wörter
  - alle Verfahren arbeiten mit Liste von **Stoppworten:**
    - Das sind Wörter, die als Deskriptoren nicht in Frage kommen und sofort entfernt werden:
      und, oder, der, die, das, ein, eine usw.

Dokumenterschließung (3)

- **automatisch** (Forts.)
  - dann evtl. **Stammbildung** (Stemming):

    `learning` → `learn`

    `houses` → `house`

    `Häuser` → `Haus`

    - im Deutschen aufwändiger als im Englischen
  - evtl. Einsatz eines Thesaurus zur Auflösung von Synonymen

Dokumenterschließung (4)

- **automatisch** (Forts.)
  - heute allerdings meist automatische Thesaurus-Generierung
    - Vermeidung von Subjektivitäten
    - Angleichung an schleichende Bedeutungsänderungen:
      - z.B. Browser, surfen
  - gute Systeme machen Sonderbehandlung für Namen, Datumsangaben usw.

2.2 Index

- zu jedem Feature:
  - Liste der Identifikatoren sämtlicher Dokumente, denen das Feature zugeteilt wurde
- Eintrag (Zeile):

  ```
  (Feature i: Dok-Id 1, Dok-Id 2, ... )
  ```

- schneller Zugriff
  - ausgehend von den Features der Anfrage zu den Einträgen
- bei der Suche nach mehreren Features:
  - Schnittmenge der Listen von Dok-Ids

Index (2)

- zwei wichtige Faktoren noch ignoriert:
  - Reihenfolge der Features
  - Gewichtungen
- spezielle Form der Suche
  - Features sollen unmittelbar benachbart vorkommen

    `(Feature i` *`adjacent`* `Feature j)`
- zusätzliche Informationen in den Einträgen
  - zu jeder Dok-Id auch noch:

    `(Absatz-Nr., Satz-Nr., Wort-Nr.)`

Gewichtung

- Häufigkeit des Auftretens in einem Dokument
  - häufig vorkommende Features sind wichtiger für das Dokument
  - Gewicht drückt Grad der Übereinstimmung mit der Anfrage aus
- mit in den Einträgen speichern

  `(Feature i: (Dok-Id j, Gewicht j), … )`
  - normalisiert auf Intervall [0,1]

Gewichtung (2)

- auch noch ganze Dokument-Menge betrachten
  - wenn Feature in fast allen Dokumenten vorkommt: nicht gut geeignet für Suche, zu wenig differenzierend
- "gute" Features
  - kommen in wenigen Dokumenten häufig vor, aber kaum in den anderen Dokumenten

Gewichtung (3)

- außer der **Feature Frequency** ff_{ik}
  - Häufigkeit, mit der Feature i in Dokument k vorkommt
- auch noch **Document Frequency** df_i verwenden
  - Zahl der Dokumente, in denen Feature i vorkommt
- **Gewicht** W_{ik} eines Features i für ein Dokument k:

$$W_{ik} = ff_{ik} \times \log(N/df_j)$$

  (N = Zahl aller Dokumente)
  - ist proportional zur Feature Frequency und zur invertierten Document Frequency
  - ist bei $df_i = N$ null

Gewichtung (4)

- Darstellung als **Tabelle:**

	d_1	d_2	d_3	…	d_k	…	
f_1 = Haus	0	4	1		ff_{1k}		2
f_2 = Baum	2	0	0		ff_{2k}		1
f_3 = Mann	7	2	23		ff_{3k}		3
…					…		
f_i	ff_{i1}	ff_{i2}	ff_{i3}	…	ff_{ik}	…	df_i
…					…		

2.3 Anfrageverarbeitung

- Anfrage zunächst auch einfach Text
  - z.B. Sammlung von Wörtern
- Anfrage-Features nach *demselben* Verfahren ermitteln wie die Features der Dokumente
  - Stoppwortliste, Stammbildung usw.
- dann für jedes Dokument Gewicht bzgl. der Features der Anfrage ermitteln
  - genauer: nur die mit hohem Gewicht heraussuchen

Anfrageverarbeitung (2)

- weitere Verfahren des Vergleichs von Features:
  - Vektorraum-Modell
  - Probabilistisches Modell
- Allen ist gemeinsam, dass sie für jedes Dokument die Relevanz bzgl. der Anfrage abschätzen: in Form des **Retrieval Status Values** (RSV)
  - ein Maß für die Ähnlichkeit von Anfrage und Dokument

Anfrageverarbeitung (3)

- Ergebnis in allen Fällen **Rangliste**, absteigend sortiert nach berechnetem RSV
- Variante: Coordination Level Matching
  - Dokumente zuerst absteigend sortieren nach der Zahl der in ihnen vorkommenden Anfrage-Features, dann innerhalb einer Zahl nach RSV
  - populär, weil besser nachvollziehbar

Was macht Google?

- Stoppwortliste
- automatische Und-Verknüpfung
- Titel höher gewichtet als sonstiger Inhalt
  - überhaupt Position berücksichtigt
- **PageRank**:
  - Dokument hat höheres Gewicht, wenn andere Dokumente mit hohem Gewicht darauf verweisen

2.3 Bewertung von IRS

- Aufgabe:
  - Auffinden der relevanten Dokumente, klar
  - aber genauso auch:
    Nichtauffinden der nichtrelevanten Dokumente!
- quantifizieren mit Hilfe von
  Präzision und Ausbeute
  (Precision and Recall)

Bewertung von IRS (2)

  - F = Menge der gefundenen Dokumente
    R = Menge der relevanten Dokumente
    D = alle Dokumente

- **Präzision** = | F ∩ R | / | F |
  Anteil der *relevanten* Dokumente an den *gefundenen*
- **Ausbeute** = | F ∩ R | / | R |
  Anteil der *gefundenen relevanten* Dokumente an den *relevanten*
  - schwierig zu messen

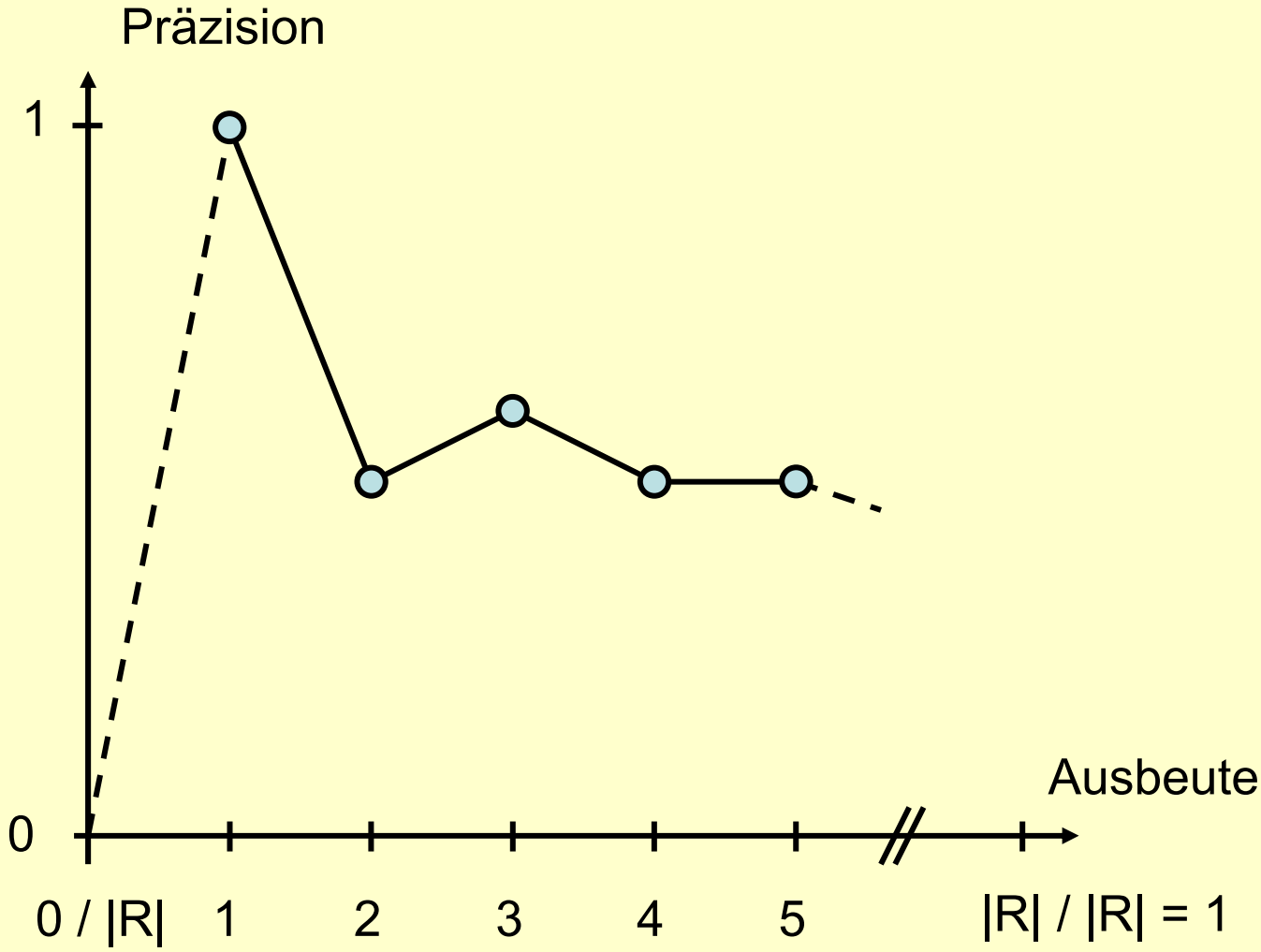
Bewertung von IRS (3)

- Vorgehensweise:
  - den ersten Eintrag der Ergebnisrangliste betrachten:
    - relevant: Präzision = 1, Ausbeute = 1 / |R|
    - nicht relevant: Präzision = 0, Ausbeute = 0
  - beim zweiten Eintrag:
    - relevant: Präzision um 1 erhöhen, Ausbeute um 1 / |R|
    - nicht relevant: Präzision halbieren, Ausbeute unverändert
  - usw.

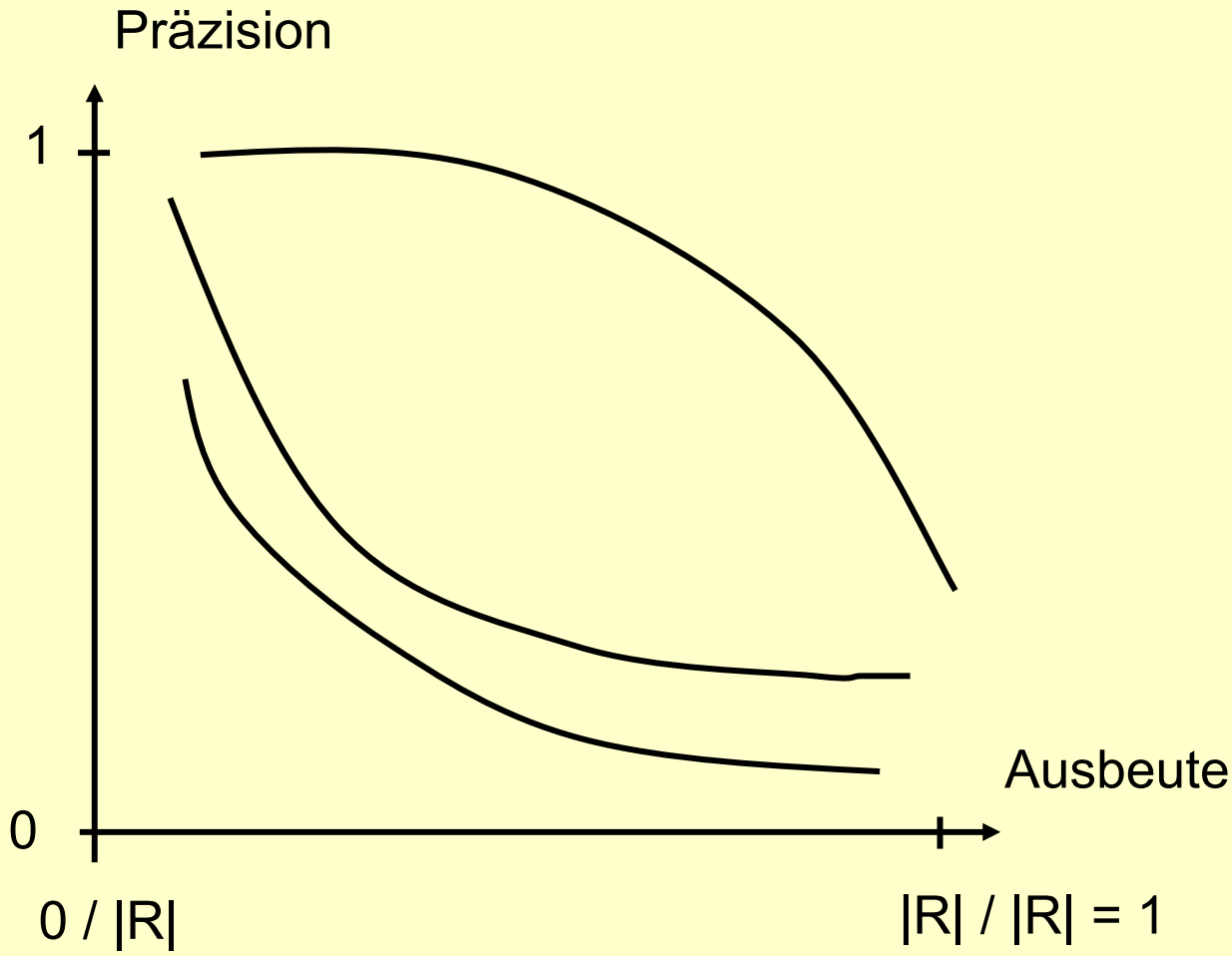
Beispiel

lauf. Nr.	R/N	Ausbeute	Präzision
1	R	**1** / |R|	**1/1**
2	R	2 / |R|	2/2
3	N	2 / |R|	2/3
4	N	**2** / |R|	**2/4**
5	R	**3** / |R|	**3/5**
6	R	4 / |R|	4/6
7	N	4 / |R|	4/7
8	N	**4** / |R|	**4/8**
9	R	5 / |R|	5/9
10	N	**5** / |R|	**5/10**

Beispiel (2)



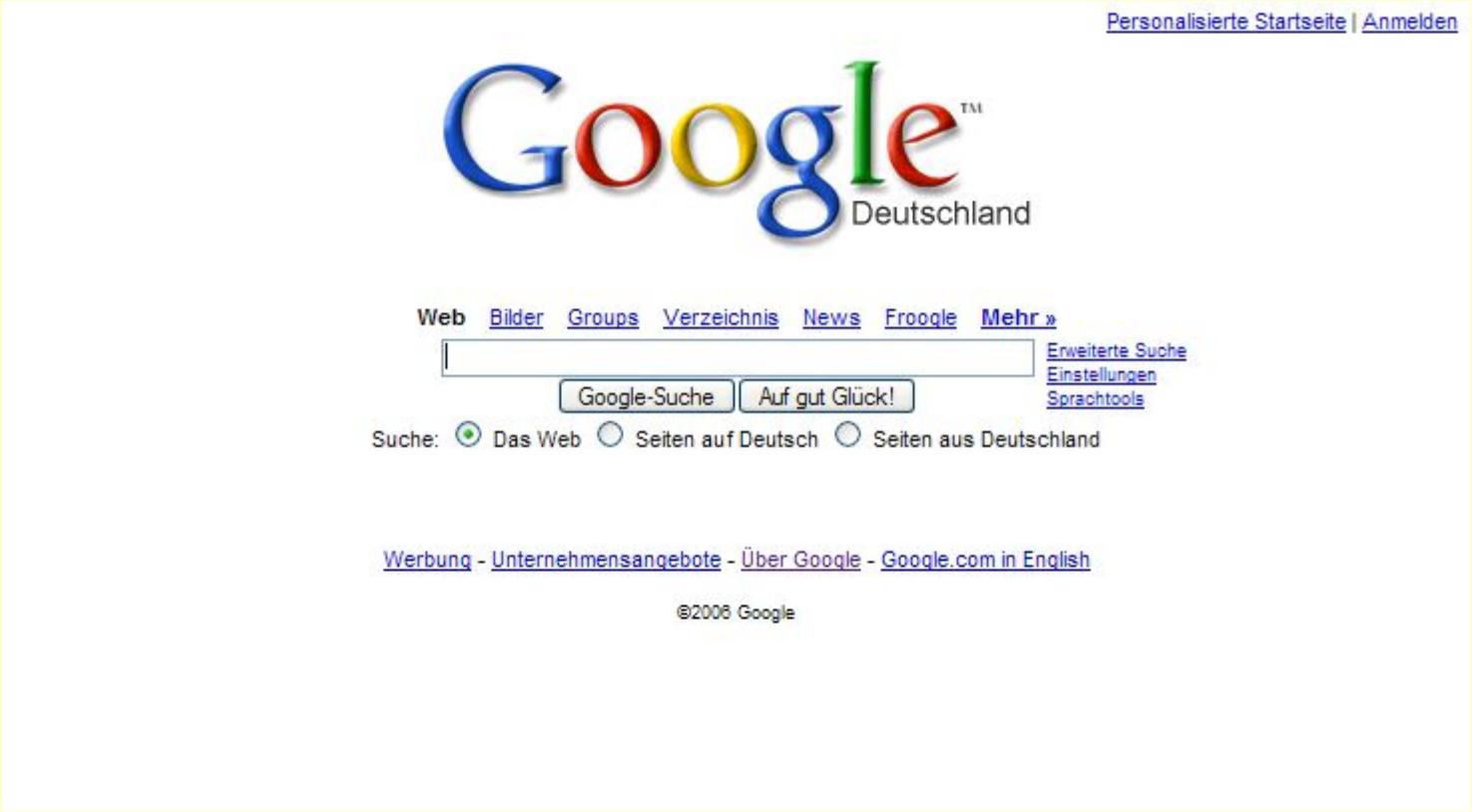
Präzisions-Ausbeute-Graph



Präzision oder Ausbeute?

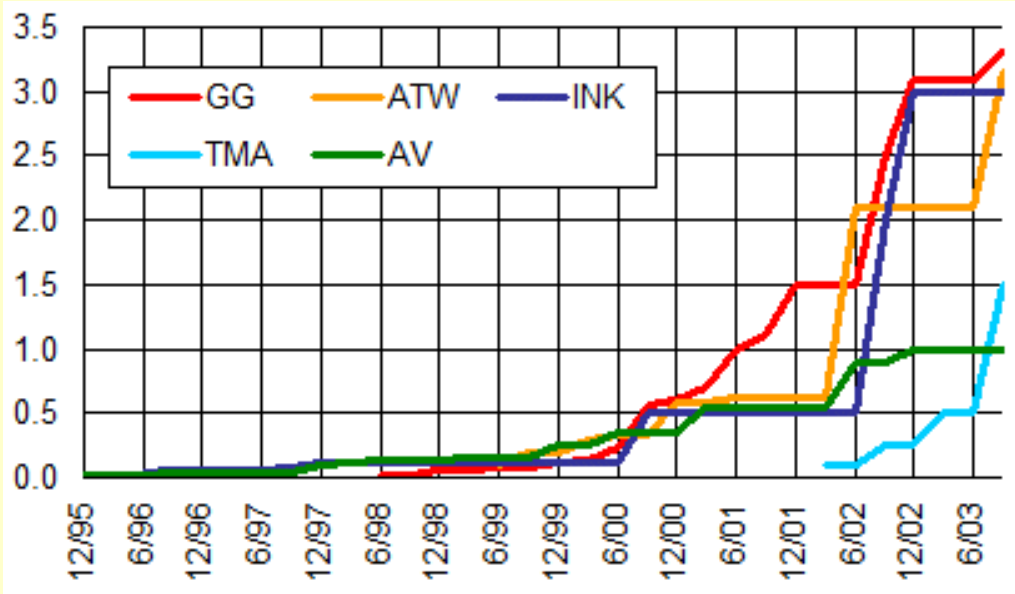
- Viele Anwender heute wollen Präzision.
  - gleich auf der ersten Seite viele relevante Ergebnisse
  - und wenig nichtrelevante
- Die Wissenschaft – und das Patentamt! – wollen Ausbeute.
  - Sie müssen "alles" finden.

3. Google


Personalisierte Startseite | Anmelden
Google™
Deutschland
Web Bilder Groups Verzeichnis News Froogle Mehr »
Erweiterte Suche
Einstellungen
Sprachtools
Google-Suche
Auf gut Glück!
Suche: Das Web Seiten auf Deutsch Seiten aus Deutschland
Werbung - Unternehmensangebote - Über Google - Google.com in English
©2006 Google

Google (2)

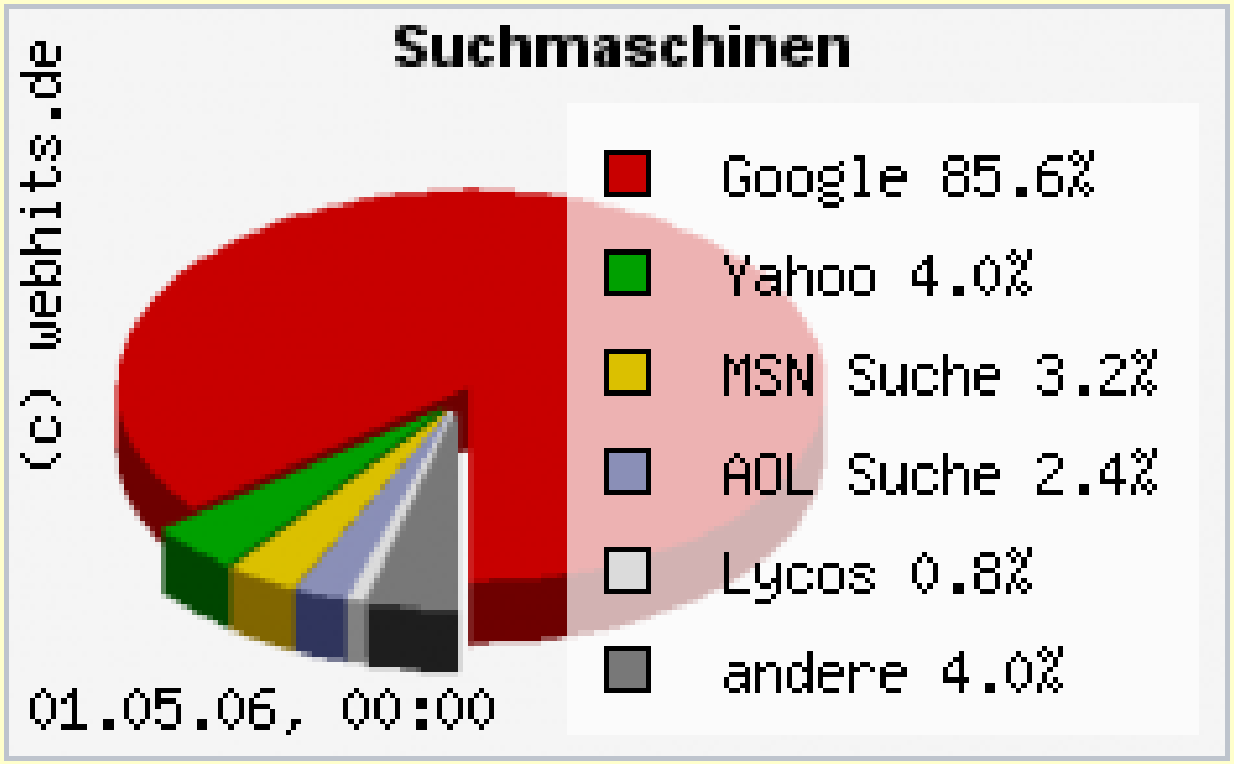
- hatte Ende 2003 4 Milliarden Dokumente in seinem Index. Heute sollen es bereits 8 Milliarden sein.



www.searchenginewatch.com

Google (3)

- ca. 85 % aller Suchen im WWW in Deutschland gehen über Google



Warum Google?

- einfach
  - minimalistische Web-Seite
- effektiv
  - PageRank
- effizient
  - schnelle Suche dank eines hochgradig parallelen Rechensystems
- umfassend

Probleme

- Google hat Macht:
  - Was man über Google nicht findet, existiert quasi nicht.
- Zensur?
  - In China darf Google bei Fragen nach bestimmten Suchworten ("Demokratie", "Menschenrechte") keine Antworten liefern.

4. Suchmaschinen allgemein

- sind längst ein Wirtschaftsfaktor
- Web-Dokumente werden so geschrieben, dass sie gefunden werden
  - unsichtbarer Text
- Ranking-Plätze werden gekauft
- besser (ehrlicher): "Sponsored Links"

5. Ausblick

- Google wird beobachtet
  - Das Vertrauen der Benutzer kann sehr schnell verspielt sein.
- Es gibt Konkurrenz
  - Und es wird ständig an Verbesserungen gearbeitet.
- Das WWW ist sehr schnelllebig.
  - Es existiert gerade erst 15 Jahre.
  - In fünf Jahren kann es wieder ganz anders sein.

Literatur

- Salton, G., and McGill, M.J., *Introduction to Modern Information Retrieval,* McGraw-Hill, New York 1983.
- Schäuble, Peter: *Multimedia Information Retrieval.* Kluwer Academic Publishers, 1998.
- Masermann, Ute, und Vossen, Gottfried: Suchmaschinen und Anfragen im World Wide Web. *Informatik-Spektrum* 21 (1998) 1, 9–15.
- Babiak, Ulrich: *Effektive Suche im Internet.* ISBN 3-89721-101-7.
- Franke, Thomas: *Gezielt suchen im Internet – Die verschiedenen Techniken.* ISBN 3-8155-7176-6.
- Gilster, Paul: *Suchen und Finden im Internet.* ISBN 3-446-18112-1.
- Jasper, Dirk: *Suchen und Finden im Internet – Tips für die erfolgreiche Online-Recherche.* ECON Computer-Taschenbuch 28138, ISBN 3-612-28138-0.
- u.v.a.m.

Literatur (2)

- David Vise ; Mark Malseed: *Die Google-Story*. Aus dem Amerikanischen von Bernd Rullkötter und Friedrich Griese. Murmann Verlag, Hamburg 2006, 304 Seiten, 19,90 Euro
- Sergey Brin ; Lawrence Page: The Anatomy of a Large Scale Hypertextual Web Search Engine. *Computer Networks* 30(1-7): 107-117 (1998)
- Sergey Brin ; Rajeev Motwani ; Larry Page ; Terry Winograd: What can you do with a Web in your Pocket. *IEEE Data Eng. Bull.* 21(2): 37-47 (1998)

Links

- www.searchenginewatch.com
  - Vergleiche, Hintergrundberichte, Tipps für Suche und Anzeigen
- http://metager.de/suma.html
  - Suchmaschinenlabor des RRZN, Universität Hannover
- www.webhits.de
  - insbesondere das Web-Barometer: Browser, Betriebssysteme, Suchmaschinen

Links (2)

- http://pr.efactory.de/d-index.shtml
  - ausführliche Darstellung von PageRank