1 Einführung und Grundbegriffe

Information Retrieval (IR) ist die Technologie zum Suchen in Kollektionen (Korpora, Intranets, Web) schwach strukturierter Dokumente: Text, HTML, XML, ...

Darunter fällt auch:
- Text- und Strukturanalyse
- Inhaltserschließung und -repräsentation
- Gruppierung und Klassifikation
- Zusammenfassung
- Filtern und Personalisieren (z.B. von Nachrichten-"Feeds")
- „Routing" (Metasuche)

Globales Ziel:
Informationsbedürfnisse befriedigen - und dabei Beseitigung des Engpasses (teurer) intellektueller Zeit !

Einsatzgebiete der IR-Technologie

- Web-Suchmaschinen
  z.B. www.altavista.com, www.google.com
- Portale, Intranet-Suchmaschinen usw.
  z.B. www.yahoo.com, www.links2go.com, www.cnn.com
- Metasuchmaschinen, spezialisierte Suchmaschinen, ontologiebasierte Suchmaschinen
  z.B. www.metacrawler.com, www.profusion.com, www.vacationspot.com, www.mathsearch.com, ontobroker.aifb.uni-karlsruhe.de
- Digitale Bibliotheken
  z.B. webclient.alexandria.ucsb.edu, omnis.informatik.tu-muenchen.de
- Ähnlichkeitssuche auf Multimediadaten (Bilder, Videos, Musik)
  z.B. QBIC: wwwqbic.almaden.ibm.com, www.hermitagemuseum.org
- Ähnlichkeitssuche auf wissenschaftlichen Daten
  z.B. Petzidat: http://sothis.cs.uni-sb.de:7001/petzidat/plsql/pa_start.home

Schnittstellen von IR-Systemen

- **Ausgabe:**
  - Menge von Dokumenten, die Suchstring(s) enthalten: **Freitextsuche**
  - Menge inhaltlich relevanter Dokumente: **Inhaltssuche**
    - ungeordnete Menge: **Boolesches Retrieval**
  - nach Relevanz absteigend sortierte Rangliste: **Ranked Retrieval** (Ähnlichkeitssuche)

- **Eingabe:**
  - Keywords (positiv/negativ) (plus Phrasen, ganze Sätze)
  - (Boolesche) Ausdrücke über Keyword-Bedingungen
  - Strukturbedingungen (z.B. Tags, Links)
  - ontologisch basierte Bedingungen
  - Suchsprache (z.B. SQL mit interMedia)

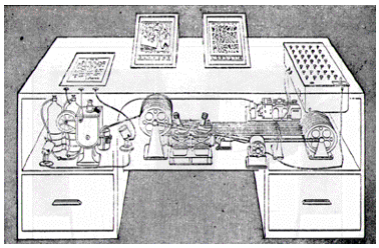
Beispiel: SQL in Oracle interMedia

Beispielanfrage:
```
Select URL, Content, Year From Docs
Where Year > 1995 And Category Like ,%drama%'
And Contains(Content, ,BT(king)', 1) > 10
And Contains(Content, ,SYN(traitor)|NT(traitor)', 2) > 0
Order By Score(1)*Score(2)
```

Weitere Operatoren (anhand von Beispielen):

~, &, |	Not, And, Or
NEAR (king, David,10)	höchstens 10 Wörter auseinander
king&David WITHIN Sentence	im selben Satz
!dog	ähnliche Aussprache (z.B. doc, dock)
$sing	gleicher Wortstamm (z.B. singer, sings, sang)
?apple	ähnliche Schreibweise(z.B. applet, apply)
NTP(computer)	narrower term partative (z.B. hard drive)
NTG(rodent)	narrower term generic (z.B. rat)
NTI(fairytale)	narrower term instance (z.B. Cinderella)
ABOUT(miracles by Jesus)	thematische Suche (verwendet intern selbst andere Operatoren)

Vannevar Bush's *Memex* (1945)



Collect all human knowledge into computer storage

⇒ will need IR + DBS + AI + ...

Size of today's and tomorrow's applications:

- Library of Congress: 20 TB books + 200 TB maps + 500 TB video + 2 PB audio
- Everything you see or hear: 1 MB/s * 50 years ≈ 2 PB

Problem: Inhaltserschließung

Umgang mit „unscharfen" Daten (und „unscharfen" Anfragen)
→ Dokumente werden typischerweise durch **Features** charakterisiert, z.B.:
- Wörter, Wortpaare oder Phrasen
- Worthäufigkeiten
- Anzahl eingehender Hyperlinks
- title, weitere Tags, Struktur von HTML- oder XML-Seiten
- Farbhäufigkeiten in Bildern (Bildmitte, oberer Rand, etc.)
- usw. usw.

→ Abbildung von natürlichsprachlichem Text auf Features:
- Behandlung von morphologischer Variation
- Behandlung von Synonymen und Polysemen (u.a. mittels Thesaurus)

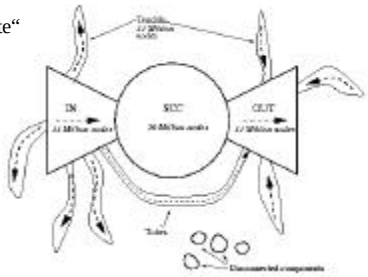
Problem: Effektivität (Retrieval-Güte)

Beispiel q: Chernoff theorem

AltaVista:	Fermat's last theorem . Previous topic Next topic ... URL: www-groups.dcs.st-and.ac.uk/~history/His..st_theorem.html
Northernlight:	J. D. Biggins - Publications. Articles on the Branching Random Walk http:/ / www.shef.ac.uk/ ~st1jdb/ bibliog.html
Lycos:	SIAM Journal on Computing Volume 26, Number 2 Contents Fail-Stop Signatures ... http://epubs.siam.org/sam-bin/dbq/toc/SICOMP/26/2
Excite:	The Official Web Site of Playboy Lingerie Model Mikki Chernoff http://www.mikkichernoff.com/
Google:	...strong convergence \cite{Chernoff}. \begin{theorem}\label{T1} Let... http://mpej.unige.ch/mp_arc/p/00-277
Yahoo:	Moment-generating Functions; Chernoff's Theorem; The Kullback-... http://www.siam.org/catalog/mcc10/bahadur.htm
Mathsearch:	No matches found.

Problem: Effizienz und Skalierbarkeit

Aktuelle „Landkarte“ des Webs:



!!! Aber:

Suchmaschinen überdecken das „Surface Web“:
1 Mrd. Dokumente, 20 TBytes

Die meisten Daten sind im „Deep Web“ hinter Portalen:
500 Mrd. Dokumente, 8 PBytes

Bewertung der Retrieval-Güte (Effektivität)

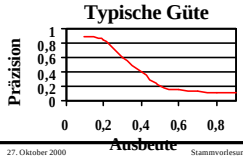
Fähigkeit, zu einer Anfrage **nur** relevante Dokumente zu liefern:

$$\textbf{\textit{Präzision (precision)}} = \frac{Anzahl\ relevanter\ Dokumente\ unter\ Top\ r}{r}$$

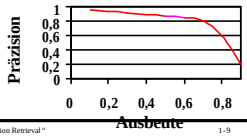
Fähigkeit, zu einer Anfrage **alle** relevante Dokumente zu liefern:

$$\textbf{\textit{Ausbeute (recall)}} = \frac{Anzahl\ relevanter\ Dokumente\ unter\ Top\ r}{Anzahl\ aller\ relevanten\ Dokumente}$$

Typische Güte



Ideale Güte



Bewertung eines IR-Systems

für eine Menge von n Anfragen q1, ..., qn
(z.B. den TREC Benchmark)

Makrobewertung = $\frac{1}{n}\sum_{i=1}^{n} Präzision(qi)$
(benutzerorientiert) der Präzision

Mikrobewertung = $\frac{\sum_{i=1}^{n} Anz.\ für\ qi\ relevanter\ \&\ gefundener\ Dok.}{\sum_{i=1}^{n} Anz.\ aller\ für\ qi\ gefundenen\ Dok.}$
(systemorientiert) der Präzision

analog für Ausbeute

Weitere Gütemaße

- Fähigkeit, möglichst wenige irrelevante Dokumente zu liefern:
  Fallout = $\frac{Anzahl\ irrelevanter\ Dokumente\ unter\ Top\ r}{Anzahl\ irrelevanter\ Dokumente\ insgesamt}$
- Präzision bei gegebener Ausbeute von x Prozent (z.B. x= 20, 40, 60, 80, 100 Prozent):
- **ROC** (Receiver Operating Characteristic): Recall (% true positives) als Funktion des Fallout (% false positives)
- Kombination von Präzision und Ausbeute durch das **F-Maß** (z.B. mit α=0.5: harmonisches Mittel): $F = \frac{1}{a\frac{1}{Präzision} + (1-a)\frac{1}{Ausbeute}}$

Weitere Maße wie accuracy, error, usefulness, coverage, novelty, etc.

Zusammenfassende Maßzahlen

- **Interpolierte durchschnittliche Präzision** einer Anfrage q mit Präzision p(x) bei Ausbeute x und Schrittweite Δ (z.B. 0.1): $\frac{1}{1/\Delta}\sum_{i=1}^{1/\Delta} p(i\Delta)$
- **Uninterpolierte durchschnittliche Präzision** einer Anfrage q mit Suchresultatsrangliste d1, ..., dm, relevanten Treffern $di_1, ..., di_k$ ($k \le m$, $i_j \le i_{j+1}$) und 100% Ausbeute (bei Top m): $\frac{1}{k}\sum_{j=1}^{k}\frac{j}{i_j}$
- **Precision-Recall-Breakeven-Point einer Anfrage q:** Punkt auf der Precision-Recall-Kurve p = f(r) mit p = r