# Evaluierung der Güte von Information-Retrieval-Systemen

Viele Folien in diesem Abschnitt sind eine deutsche Übersetzung der Folien von Raymond J. Mooney (http://www.cs.utexas.edu/users/mooney/ir-course/).

# Warum Systemevaluierung?

- Es gibt viele Retrievalmodelle/Algorithmen/Systeme.
- Welches ist das Beste?
- Welches ist die beste Komponente für:
  - Ranking-Funktion (Skalarprodukt, Kosinus, …)?
  - Termselektion (Entfernen von Stopwörtern, Stemming…)?
  - Termgewichtung (TF, TF-IDF,…)?
- Wie weit muss ein Anwender in einer geordneten Liste nach unten gehen, um einige/alle relevanten Dokumente zu finden?

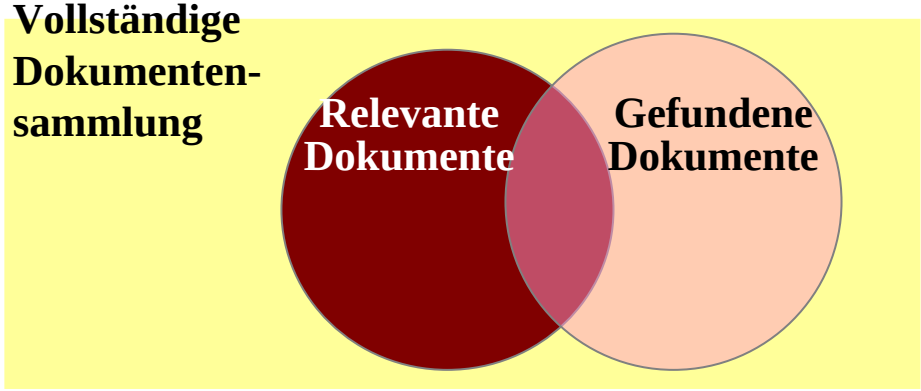
# Schwierigkeiten bei der Evaluierung von IR Systemen

- Die Bewertung der Effektivität steht in engem Bezug zur ***Relevanz*** von gefundenen Elementen.
- Relevanz ist nicht typischerweise binär, sondern eine stetige Größe.
- Selbst wenn Relevanz binär ist, kann es schwierig sein, eine Beurteilung abzugeben.
- Relevanz aus menschlicher Sicht ist:
  - subjektiv: hängt von der spezifischen Beurteilung des Anwenders ab.
  - situativ: bezieht sich auf die aktuellen Bedürfnisse des Anwenders.
  - kognitiv: hängt von der menschlichen Wahrnehmung und dem Verhalten ab.
  - dynamisch: verändert sich im Laufe der Zeit.

# Manuell indizierte Korpora (Gold-Standard)

- Gegeben sei ein Korpus von Dokumenten.
- Sammle eine Liste von Anfragen für diesen Korpus.
- Ein oder mehrere menschliche Experten kennzeichnen (labeln) zu jeder Anfrage die relevanten Dokumente.
- Man geht (der Einfachheit halber) typischerweise von binären Relevanz-Beurteilungen aus.
- Erfordert für große Korpora mit vielen Anfragen erheblichen menschlichen Aufwand.

# Precision und Recall




| | gefunden | nicht gefunden |
|---|---|---|
| irrelevant | gefunden & irrelevant | nicht gefunden & irrelevant |
| relevant | gefunden & relevant | nicht gefunden aber relevant |

$$Recall = \frac{Anzahl\ der\ relevanten\ gefundenen\ Dokumente}{Gesamtzahl\ der\ relevanten\ Dokumente}$$

$$Precision = \frac{Anzahl\ der\ relevanten\ gefundenen\ Dokumente}{Gesamtzahl\ der\ gefundenen\ Dokumente}$$

# Precision und Recall

- Precision
  - Gibt den Anteil der relevanten Dokumente bzgl. aller Dokumente in der Ergebnismenge wieder.
- Recall
  - Misst die Fähigkeit des Suchverfahrens, ***alle*** relevanten Dokumente im Korpus zu entdecken.

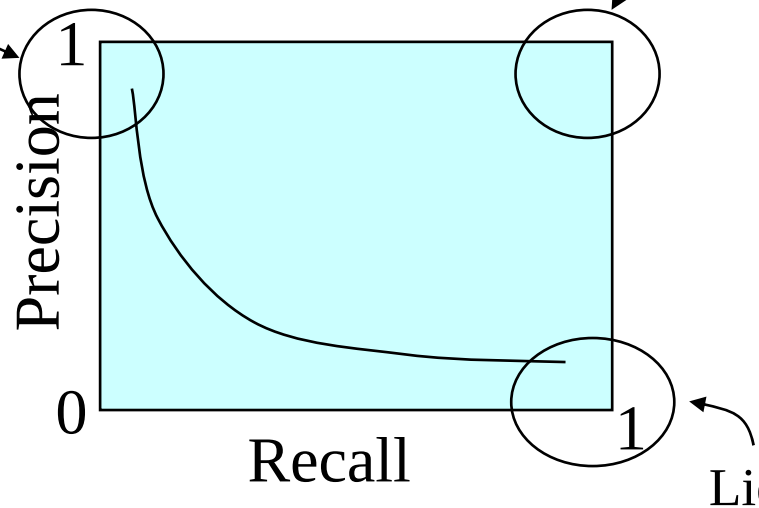
# Bestimmen des Recalls ist schwierig

- Die Gesamtzahl der relevanten Elemente ist häufig nicht verfügbar:
  - Nimm einen Ausschnitt der Datenbank und führe eine Relevanzbeurteilung anhand dieser Elemente durch.
  - Wende verschiedene Retrievalalgorithmen auf die gleiche Datenbank und für die gleiche Anfrage an. Die Vereinigung aller relevanten Elemente über alle Algorithmen wird als die Menge aller relevanten Elemente angesehen.

# Kompromiss zwischen Recall und Precision

Liefert die meisten relevanten Dokumente,
aber verfehlt viele nützliche.

Ideales Ergebnis




Liefert die meisten relevanten Dokumente, aber enthält zu viele irrelevante Elemente.

# Berechnung von Recall/Precision

- Bilde für die gegebene Anfrage die Rankingliste.
- Das Abschneiden der Liste an unterschiedlichen Schwellwerten führt zu verschiedenen Mengen von gefundenen Dokumenten und demzufolge zu verschiedenen Recall/Precision-Ergebnissen.
- Markiere jedes Dokument der Rankingliste, das gemäß dem Gold-Standard relevant ist.
- Berechne für jedes relevante Dokument das Recall-/Precision-Paar, dass beim Abschneiden der Liste an dieser Stelle entsteht.

# Berechnung von Recall/Precision: Ein Beispiel

| n | doc # | relevant |
|---|---|---|
| 1 | 588 | x |
| 2 | 589 | x |
| 3 | 576 | |
| 4 | 590 | x |
| 5 | 986 | |
| 6 | 592 | x |
| 7 | 984 | |
| 8 | 988 | |
| 9 | 578 | |
| 10 | 985 | |
| 11 | 103 | |
| 12 | 591 | |
| 13 | 772 | x |
| 14 | 990 | |

Gesamtzahl relevanter Doks. = 6
Prüfe für jeden neuen Recall-Punkt:

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/2=1

R=3/6=0.5; P=3/4=0.75

R=4/6=0.667; P=4/6=0.667

Ein relevantes Dokument fehlt. 100% Recall wird nie erreicht

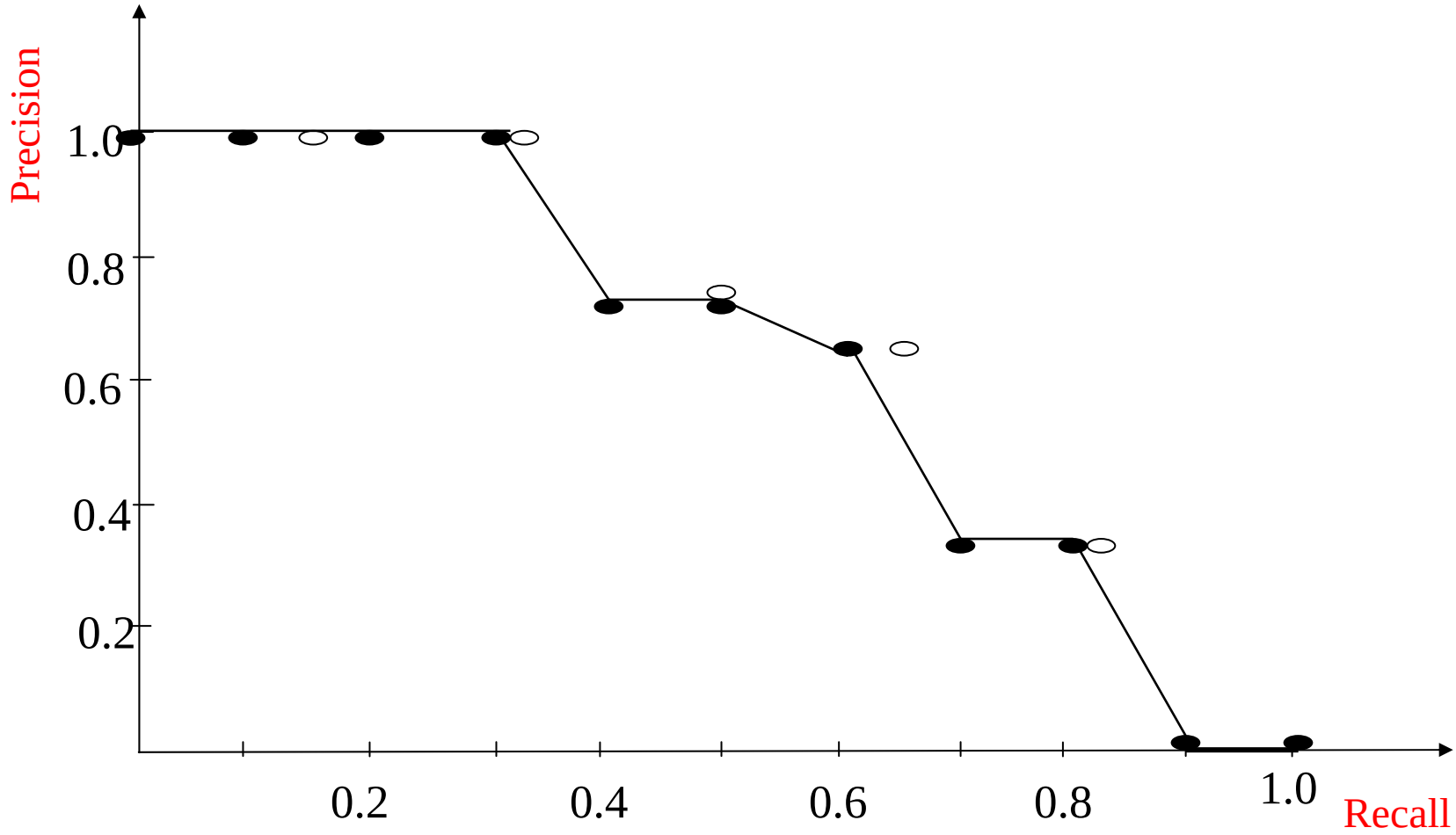
R=5/6=0.833; p=5/13=0.38

# Interpolation der Recall/Precision-Kurve

- Interpoliere den Precision-Wert für jeden *Standard Recall-Level*:
  - $r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
  - $r_0 = 0.0, r_1 = 0.1, \ldots, r_{10}=1.0$
- Die interpolierte Precision am *j*-ten Standard Recall-Level ist die maximal bekannte Precision bei jedem Recall-Level zwischen dem *j*-ten und (*j* + 1)-ten Level:

$$P(r_j) = \max_{r_j \le r \le r_{j+1}} P(r)$$

# Interpolation der Recall/Precision Kurve: Ein Beispiel

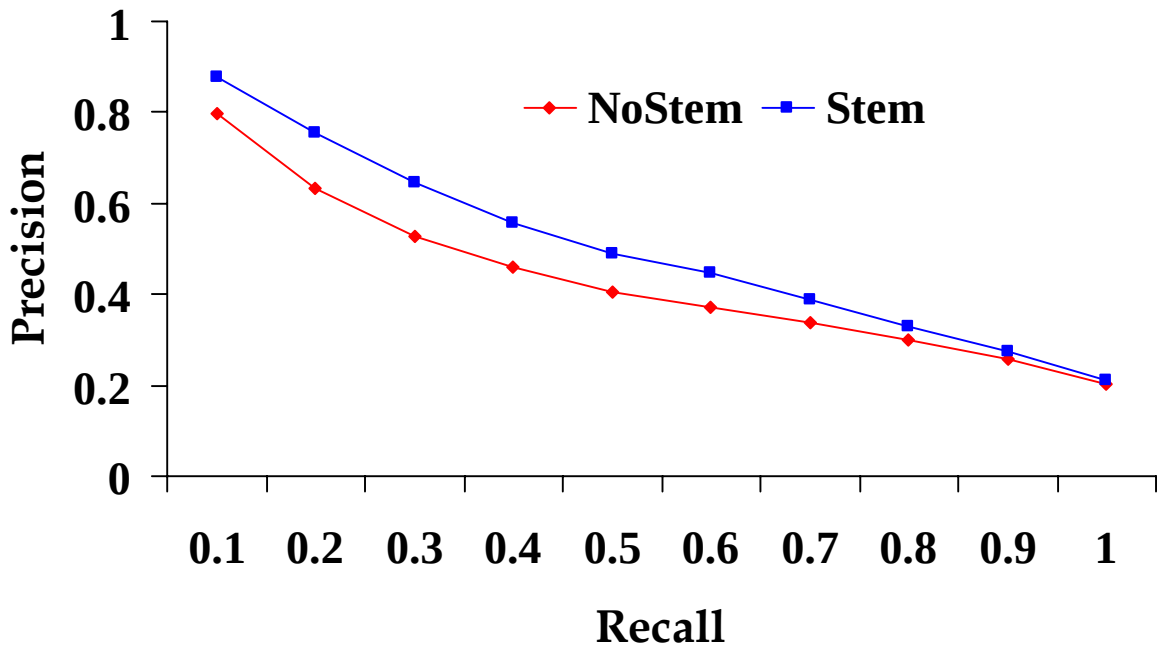


# Durchschnittliche Recall/Precision-Kurve

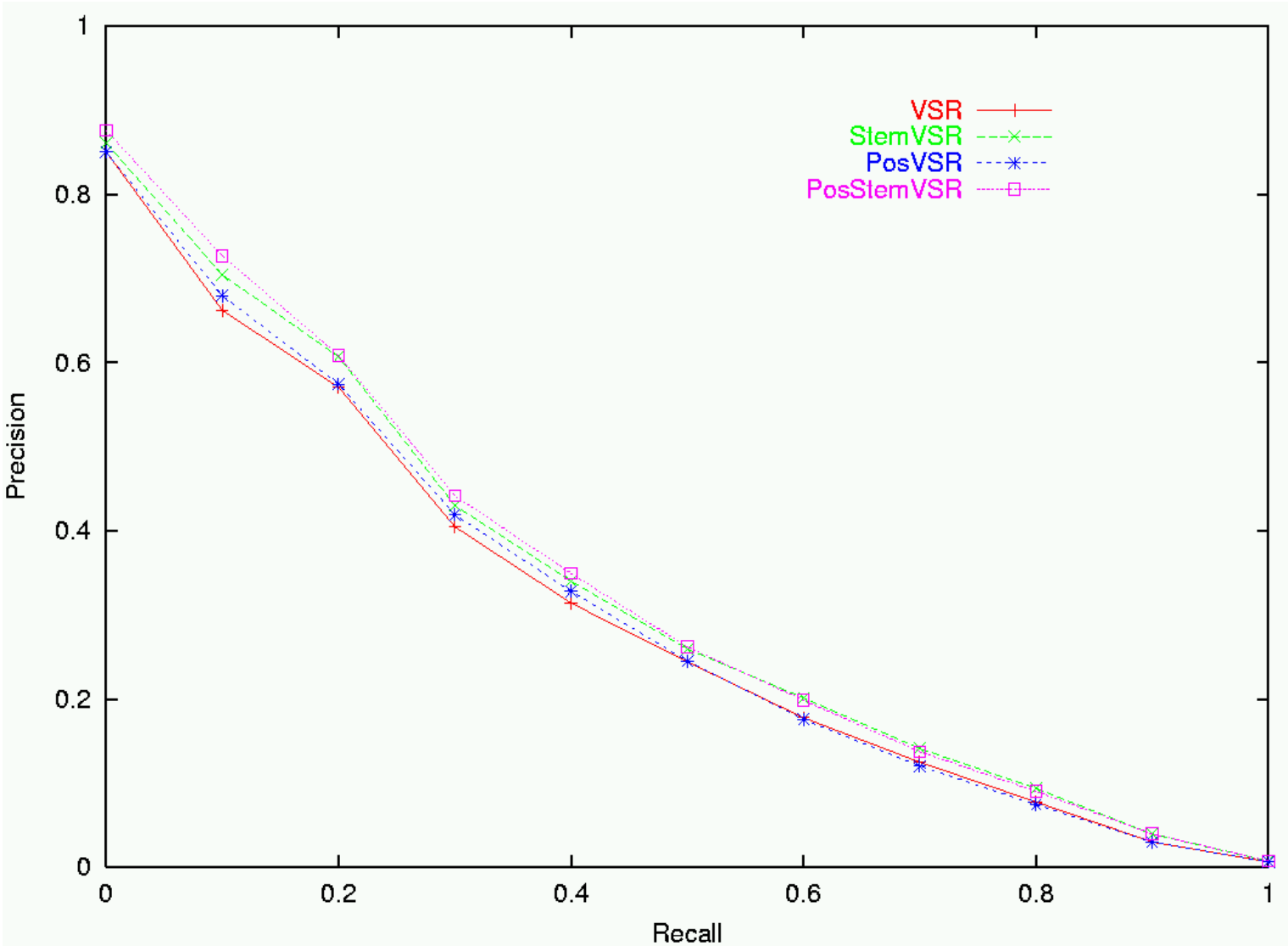
- Zeigt die durchschnittliche Performance für eine große Menge von Anfragen an.
- Berechne für jeden Standard-Recall-Level den Durchschnitt der Precisionwerte über alle Anfragen.
- Zeichne die Precision/Recall-Kurve, um die Gesamtsystemleistung für einen Dokumenten-Korpus bei den auf ihn angewandten Anfragen zu bewerten.

# Vergleiche zwei oder mehrerer Systeme

Die Kurve, die am nächsten an der oberen rechten Ecke des Graphen liegt, zeigt die beste Leistung an.



# Beispiel RP-Kurve für Cystic Fibrosis-Korpus



# R- Precision

| n | doc # | relevant |
|---|---|---|
| 1 | 588 | x |
| 2 | 589 | x |
| 3 | 576 | |
| 4 | 590 | x |
| 5 | 986 | |
| 6 | 592 | x |
| 7 | 984 | |
| 8 | 988 | |
| 9 | 578 | |
| 10 | 985 | |
| 11 | 103 | |
| 12 | 591 | |
| 13 | 772 | x |
| 14 | 990 | |

- Precision an der R-ten Position im Ergebnisranking einer Anfrage, die R relevante Dokumente hat.

R = # der relevanten Doku. = 6

R-Precision = 4/6 = 0.67

# F-Maß

- Ein Leistungsmaß, das sowohl Recall als auch Precision berücksichtigt.
- Harmonisches Mittel von Recall und Precision:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

- Verglichen mit dem arithmetischen Mittel müssen beide Werte hoch sein, damit das harmonische Mittel hoch ist.

# E-Maß (parametrisiertes F-Maß)

- Eine F-Maß-Variante, die eine Gewichtung von Precision vs. Recall erlaubt:

$$E = \frac{(1+\beta^2)PR}{\beta^2 P + R} = \frac{(1+\beta^2)}{\frac{\beta^2}{R}+\frac{1}{P}}$$

- Wert für β kontrolliert den Kompromiss:
  - β = 1: gleichmäßige Gewichtsverteilung zwischen Precision und Recall (E=F).
  - β > 1: mehr Gewicht auf Precision.
  - β < 1: mehr Gewicht auf Recall.

# Ausfallrate (Fallout Rate)

- Probleme sowohl mit Precision als auch mit Recall:
  - Anzahl irrelevanter Dokumente in der Sammlung wird nicht berücksichtigt.
  - Recall ist undefiniert, wenn es kein relevantes Dokument in der Sammlung gibt.
  - Precision ist undefiniert, wenn kein Dokument gefunden wird.

$$Fallout = \frac{Anz.\ gefundener\ \ nicht\ relevanter\ Elemente}{Gesamtzahl\ \ nichtrelevanter\ Elemente\ \ in\ der\ Sammlung}$$

# Subjektive Relevanzmaße

- *Novelty Ratio*: Der Anteil gefundener und vom Anwender als relevant beurteilter Elemente, deren er sich zuvor nicht bewusst war.
  - Fähigkeit, ***neue*** Informationen zu einem Thema zu finden.
- *Coverage Ratio*: Der Anteil relevanter Elemente, die in den gesamten relevanten Dokumenten gefunden wurden, die der Anwender vor der Suche ***kannte***.
  - Relevant, wenn der Anwender Dokumente lokalisieren möchte, die er zuvor gesehen hat (z.B., der Budgetbericht für das Jahr 2000).

# Andere zu berücksichtigende Faktoren

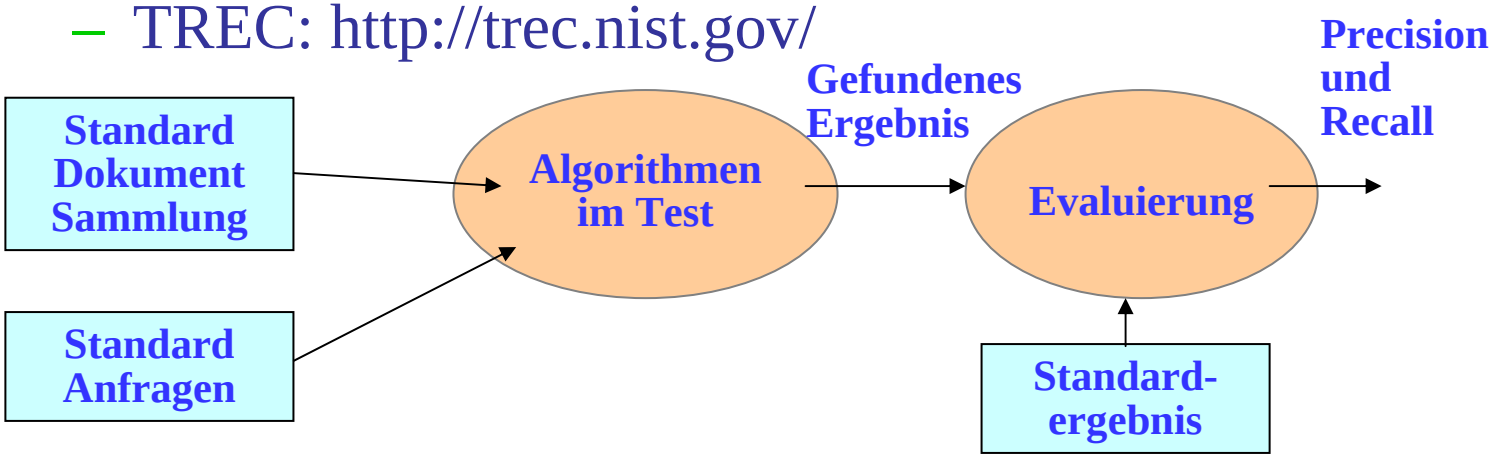
- *Anwenderaufwand*: Arbeit, die vom Anwender zu erledigen ist: Formulieren von Anfragen, Leiten der Suche und Selektion des Outputs.
- *Antwortzeit*: Zeitintervall zwischen Absetzen einer Anwenderanfrage und der Präsentation der Systemantworten.
- *Form der Präsentation*: Einfluss der Ausgabeform einer Suchanfrage auf die Fähigkeit des Anwenders, das gefundene Material zu verwenden.
- *Umfang der Sammlung*: Ausmaß, in dem jegliche/alle relevanten Elemente im Dokumentencorpus enthalten sind.

# Experimentelles Setup für Benchmarking

- Eine ***analytische*** Leistungsevaluierung ist für Dokument-Retrieval-Systeme schwierig, da viele Eigenschaften wie Relevanz, Verteilung der Worte etc. nur schwer präzise zu beschreiben sind.
- Leistung wird durch ***Benchmarking*** gemessen. D.h. die Retrieval-Effektivität eines Systems wird anhand *einer gegebenen Liste von Dokumenten*, *Anfragen* und *Relevanzbeurteilungen* evaluiert.
- Leistungsdaten gelten nur für die Umgebung, in der das System evaluiert ist.

# Benchmarks

- Eine Benchmark-Sammlung umfasst:
  - Eine Liste von Standarddokumenten und Anfragen/Themen.
  - Eine Liste relevanter Dokumente für jede Anfrage.
- Standard-Sammlungen für traditionelles IR:
  - Smart collection: ftp://ftp.cs.cornell.edu/pub/smart
  - TREC: http://trec.nist.gov/



# Benchmarking – die Probleme

- Leistungsdaten gelten nur für einen speziellen Benchmark.
- Der Aufbau eines Testdatensatzes (benchmark corpus) ist eine schwierige Aufgabe.
- Web-Testdatensätze sind gerade in der Entwicklung.
- Testdatensätze mit nicht-englischen Dokumenten oder mehr als einer Sprache sind gerade in der Entwicklung.

# Frühe Testsammlungen

- Erste Experimente basierten auf der SMART-Sammlung, die ziemlich wenige Dokumente enthält. (ftp://ftp.cs.cornell.edu/pub/smart)

| Sammlung Name | Anzahl Dokumente | Anzahl Anfragen | Größe (Mbytes) |
|---|---|---|---|
| CACM | 3,204 | 64 | 1.5 |
| CISI | 1,460 | 112 | 1.3 |
| CRAN | 1,400 | 225 | 1.6 |
| MED | 1,033 | 30 | 1.1 |
| TIME | 425 | 83 | 1.5 |

- Verschiedene Forscher haben unterschiedliche Testsammlungen und Evaluierungstechniken angewendet.

# Die TREC Benchmark

- TREC: Text REtrieval Conference (http://trec.nist.gov/)
  Stammt aus dem TIPSTER-Programm, das
  die Defense Advanced Research Projects Agency (DARPA)
  sponsert.
- Wurde 1992 zu einer jährlichen Konferenz, mitgesponsert vom National Institute of Standards and Technology (NIST) und DARPA.
- Den Teilnehmern wurde zum Trainieren und Testen der Systeme Teile einer Standardliste von Dokumenten und THEMEN (wovon Anfragen abzuleiten sind) in verschiedenen Stadien gegeben.
- Die Teilnehmer legen die P/R-Werte für den endgültigen Dokument- und Anfrage-Korpus vor und präsentieren ihre Ergebnisse bei der Konferenz.

# Die TREC-Ziele

- Schaffen einer gemeinsamen Grundlage für den Vergleich verschiedener IR-Techniken.
  - Gleiche Dokumenten- und Anfrageliste und gleiche Evaluierungsmethoden.
- Teilen von Ressourcen und Erfahrungen bei der Entwicklung des Benchmarks.
  - Hauptsponsoring durch die amerikanische Regierung, um große Benchmark-Sammlungen zu entwickeln.
- Förderung der Beteiligung von Industrie und Wissenschaft.
- Entwicklung neuer Evaluierungstechniken, besonders für neue Anwendungen.
  - Retrieval, Routing/Filtering, nicht-englische Dokumente, web-basierte Sammlung, Fragenbeantwortung (question answering).

# TREC: Vorteile

- Riesige Datensätze (verglichen mit ein paar MB in der SMART Collection).
- Relevanzbeurteilung wird zur Verfügung gestellt.
- In ständiger Entwicklung mit Unterstützung der U.S.-Regierung.
- Große Beteiligung:
  - TREC 1: 28 Papers 360 Seiten.
  - TREC 4: 37 Papers 560 Seiten.
  - TREC 7: 61 Papers 600 Seiten.
  - TREC 8: 74 Papers.

# TREC-Aufgaben

- Ad hoc: Neue Fragen werden zu einem statischen Datensatz gestellt.
- Routing: Die gleichen Fragen werden gestellt, aber neue Informationen werden gesucht (neue Zeitungsausschnitte, Library Profiling).
- Neue Aufgaben wurden nach TREC 5 hinzugefügt
  - interaktiv,
  - vielsprachig (multilingual),
  - Verarbeitung natürlicher Sprache,
  - Mischen mehrerer Datenbanken,
  - Filtering,
  - sehr große Korpora (20 GB, 7.5 Millionen Dokumente),
  - Fragenbeantwortung (question answering).

# Eigenschaften der TREC-1-Sammlung

- Sowohl lange als auch kurze Dokumente (von ein paar hundert zu mehr als tausend unterschiedlichen Termen in einem Dokument).

- Testdatensätze bestehen aus:

| | | |
|---|---|---|
| WSJ | Wall Street Journal articles (1986-1992) | 550 M |
| AP | Associate Press Newswire (1989) | 514 M |
| ZIFF | Computer Select Disks (Ziff-Davis Publishing) | 493 M |
| FR | Federal Register | 469 M |
| DOE | Abstracts from Department of Energy reports | 190 M |

# Musterdokument (mit SGML)

<DOC>
<DOCNO> WSJ870324-0001 </DOCNO>
<HL> John Blair Is Near Accord To Sell Unit, Sources Say </HL>
<DD> 03/24/87</DD>
<SO> WALL STREET JOURNAL (J) </SO>
<IN> REL TENDER OFFERS, MERGERS, ACQUISITIONS (TNM)
MARKETING, ADVERTISING (MKT) TELECOMMUNICATIONS,
BROADCASTING, TELEPHONE, TELEGRAPH (TEL) </IN>
<DATELINE> NEW YORK </DATELINE>
<TEXT>
John Blair & Co. is close to an agreement to sell its TV station advertising representation operation and program production unit to an investor group led by James H. Rosenfield, a former CBS Inc. executive, industry sources said. Industry sources put the value of the proposed acquisition at more than $100 million. ...
</TEXT>
</DOC>

# Musteranforderung (mit SGML)

<top>
<head> Tipster Topic Description
<num> Number: 066
<dom> Domain: Science and Technology
<title> Topic: Natural Language Processing
<desc> Description: Document will identify a type of natural language processing technology which is being developed or marketed in the U.S.
<narr> Narrative: A relevant document will identify a company or institution developing or marketing a natural language processing technology, identify the technology, and identify one of more features of the company's product.
<con> Concept(s): 1. natural language processing ;2. translation, language, dictionary
<fac> Factor(s):
<nat> Nationality: U.S.</nat>
</fac>
<def> Definitions(s):
</top>

# TREC-Eigenschaften

- Sowohl Dokumente als auch Anforderungen enthalten viele verschiedene Arten von Informationen (Felder).
- Generierung der formalen Anfragen (Boolesche, Vektorraum, etc.) liegt in der Verantwortung des Systems.
  - Ein System kann sehr gut beim Anfragen und Ranking sein, aber wenn es aus der Anforderung dürftige Anfragen erzeugt, wird seine endgültige P/R dürftig sein.

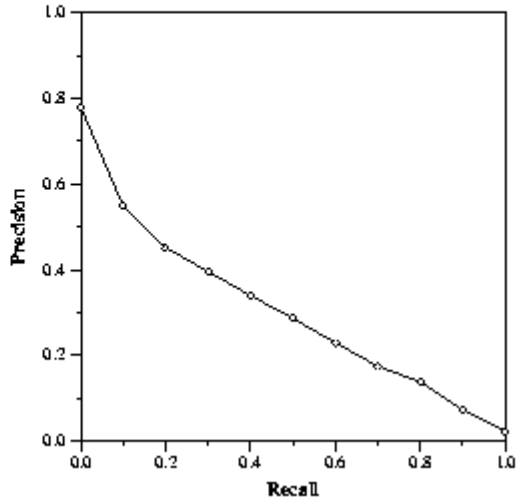
# Evaluierung

- Zusammenfassende Tabellenstatistiken: Anzahl Themen, Anzahl gefundener Dokumente, Anzahl relevanter Dokumente.
- Durchschnittlicher Recall-Precision: Durchschnitts-Precision bei 11 Recall-Levels (0 bis 1 in 0.1er-Schritten).
- Dokumentbasierter Durchschnitt: Durchschnitts-Precision wenn 5, 10, .., 100, … 1000 Dokumente gefunden werden.
- Histogramm über die durchschnittliche Precision: Unterschied zwischen der R-Precision für jedes Thema und der durchschnittlichen R-Precision aller Systeme zu diesem Thema.

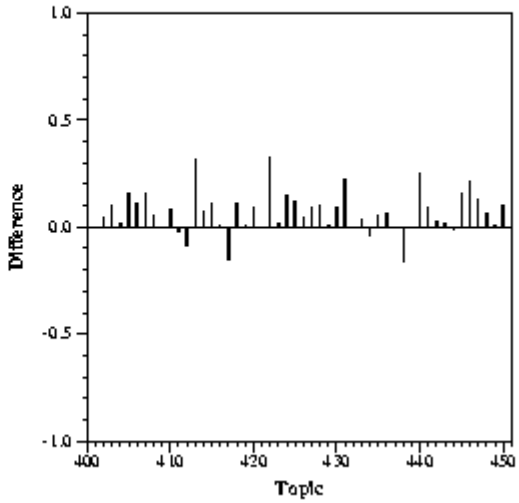
| Summary Statistics | |
|---|---|
| Run Number | Flab8atd2 |
| Run Description | Automatic, title + desc |
| Number of Topics | 50 |
| **Total number of documents over all topics** | |
| Retrieved: | 50000 |
| Relevant: | 4728 |
| Rel ret: | 2990 |

| Recall Level Precision Averages | |
|---|---|
| Recall | Precision |
| 0.00 | 0.7796 |
| 0.10 | 0.5490 |
| 0.20 | 0.4517 |
| 0.30 | 0.3954 |
| 0.40 | 0.3397 |
| 0.50 | 0.2863 |
| 0.60 | 0.2291 |
| 0.70 | 0.1745 |
| 0.80 | 0.1381 |
| 0.90 | 0.0720 |
| 1.00 | 0.0224 |
| **Average precision over all relevant docs** | |
| non interpolated | 0.2930 |

| Document Level Averages | |
|---|---|
| | Precision |
| At 5 docs | 0.5480 |
| At 10 docs | 0.4880 |
| At 15 docs | 0.4587 |
| At 20 docs | 0.4200 |
| At 30 docs | 0.3887 |
| At 100 docs | 0.2490 |
| At 200 docs | 0.1777 |
| At 500 docs | 0.1011 |
| At 1000 docs | 0.0598 |
| **R Precision (precision after R docs retrieved (where R is the number of relevant documents))** | |
| Exact | 0.3203 |



Recall-Precision Curve



Difference from Median in Average Precision per Topic

# Cystic Fibrosis (CF) Sammlung

- 1.239 Zusammenfassungen von medizinischen Zeitungsartikeln in CF.
- 100 Informationsanforderungen (Anfragen) in Form kompletter englischer Fragen.
- Relevante Dokumente, die von vier Medizinexperten auf einer Skala von 0-2 bestimmt und bewertet wurden:
  - 0: Nicht relevant.
  - 1: Marginal relevant.
  - 2: Sehr relevant.

# CF Dokumentenfelder

- MEDLINE Zugangsnummer
- Autor
- Titel
- Quelle
- Wesentliche Themen
- Untergeordnete Themen
- Zusammenfassung (oder Auszug)
- Verweise auf andere Dokumente
- Zitate zu diesem Dokument

# Beispiel CF-Dokument

AN 74154352
AU Burnell-R-H. Robertson-E-F.
TI Cystic fibrosis in a patient with Kartagener syndrome.
SO Am-J-Dis-Child. 1974 May. 127(5). P 746-7.
MJ CYSTIC-FIBROSIS: co. KARTAGENER-TRIAD: co.
MN CASE-REPORT. CHLORIDES: an. HUMAN. INFANT. LUNG: ra. MALE.
SITUS-INVERSUS: co, ra. SODIUM: an. SWEAT: an.
AB A patient exhibited the features of both Kartagener syndrome and
cystic fibrosis. At most, to the authors' knowledge, this
represents the third such report of the combination. Cystic
fibrosis should be excluded before a diagnosis of Kartagener
syndrome is made.
RF 001 KARTAGENER M BEITR KLIN TUBERK 83 489 933
002 SCHWARZ V ARCH DIS CHILD 43 695 968
003 MACE JW CLIN PEDIATR 10 285 971
…
CT 1 BOCHKOVA DN GENETIKA (SOVIET GENETICS) 11 154 975
2 WOOD RE AM REV RESPIR DIS 113 833 976
3 MOSSBERG B MT SINAI J MED 44 837 977
…

# Beispiele für CF-Anfragen

QN 00002
QU Can one distinguish between the effects of mucus hypersecretion and infection on the submucosal glands of the respiratory tract in CF?
NR 00007
RD 169 1000 434 1001 454 0100 498 1000 499 1000 592 0002 875 1011

---

QN 00004
QU What is the lipid composition of CF respiratory secretions?
NR 00009
RD 503 0001 538 0100 539 0100 540 0100 553 0001 604 2222 669 1010 711 2122 876 2222

NR: Anzahl relevanter Dokumente
RD: Relevante Dokumente

Ratingcode: Vier 0-2 Ratings, eines von jedem Experten