Vorlesung
Information Retrieval
Wintersemester 04/05

25. November 2004

Dr. Jens E. Wolff

Institut für Informatik III
Universität Bonn

Tel. 02 28 / 73-45 31
Fax 02 28 / 73-43 82
jw@informatik.uni-bonn.de

Themenübersicht

- Evaluation
  - Begriffserklärung und Definition
  - Historie
  - Evaluationsarten
  - Bewertung der Effektivität

Begriffserklärung und Definition

Notwendigkeit der Evaluation

„Im Information Retrieval (IR) werden Informationssysteme in Bezug auf ihre Rolle im Prozeß des Wissenstransfers vom menschlichen Wissens-produzenten zum Informations-Nachfragenden betrachtet. Die Fachgruppe "Information Retrieval" in der Gesellschaft für Informatik beschäftigt sich dabei schwerpunktmäßig mit jenen Fragestellungen, die im Zusammenhang mit vagen Anfragen und unsicherem Wissen entstehen. ... Hierzu zählen neben Fragen mit unscharfen Kriterien insbesondere auch solche, die nur im Dialog iterativ durch Reformulierung (in Abhängigkeit von den bisherigen Systemantworten) beantwortet werden können; häufig müssen zudem mehrere Datenbasen zur Beantwortung einer einzelnen Anfrage durchsucht werden. ... Die Unsicherheit (oder die Unvollständigkeit) dieses Wissens resultiert meist aus der begrenzten Repräsentation von dessen Semantik (z.B. bei Texten oder multimedialen Dokumenten); ... **Aus dieser Problematik ergibt sich die Notwendigkeit zur Bewertung der Qualität der Antworten eines Informationssystems, wobei in einem weiteren Sinne die Effektivität des Systems in Bezug auf die Unterstützung des Benutzers bei der Lösung seines Anwendungsproblems beurteilt werden sollte.**“

(Gesellschaft für Informatik (GI) - Fachgruppe IR)

Dr. Jens E. Wolff
© 2004

Notwendigkeit der Evaluation

Will man diesem Anspruch gerecht werden, so lassen sich IR Systeme nur mit ungeheurem Aufwand beurteilen und vergleichen:

Repräsentative Auswahl von

- Anwendungsproblemen und
- Benutzenden, sowie
- isolierte Betrachtung der Lösung der Anwendungsprobleme.

→ im Allgemeinen nicht möglich!

Einflussfaktoren (1)

Nach Salton und McGill (1983) gibt es folgende Einflussfaktoren:

> Auswahl und Eingabe der Dokumente (*input policies*) bestimmen die

- Ausrichtung der Inhalte der Datenbank, die
- „Tiefe“ der Behandlung von Themen, die
- Aktualität der Dokumente und die
- Fehlerrate in den Dokumenten

Einflussfaktoren (2)

> Die Dokumentformate (*physical input form*) bestimmen,

- in welcher Form Dokumente gespeichert werden und
- was sie beinhalten können.
- Länge der Dokumente
- Titel
- Abstract oder
- ganze Artikel
- Bilder
- Tondokumente oder
- andere multimediale Dokumente
- die Repräsentation von Dokumenten
- Zugriff beim Suchen

Einflussfaktoren (3)

> Die Indexierungsmethode (*indexing language*) gibt die Repräsentation der Dokumente an.
Unterschiede z.B.:

- kontrolliertes Vokabular
- Regeln für die Indexierung
- wie spezifisch bzw. wie breit die Dokumente indexiert werden sollen.

> Indexierung von Hand: Indexierungsvorgang (*indexing operation*):

- Erfahrungen und Interessen,
- Motivation

der Indexierenden, beeinflussen die Konsistenz der Indexierung

Einflussfaktoren (4)

> Die Anfrageformulierung durch die Nutzenden:

- fachlichen Kenntnisstand der Anfragenden
- Vertrautheit mit dem IR System.

> Suche (*search operation*):
Präferenzen der Anfragenden:

- vollständige Abdeckung der Frage
- oder überschaubare Antwortmenge.

> Präsentation der Ergebnisse:
dargebotene Dokumente

- erfassen und
- beurteilen

→ **systematische Variation der Einflussfaktoren nicht möglich!**

Bewertungskriterien

Effizienz

Möglichst sparsamer Umgang mit Ressourcen wie Rechenzeit und Speicherplatz

Zeitkomplexität
Speicherkomplexität
empirisches Laufzeitverhalten
I/O-Aufkommen

→ Analyse der verwendeten Algorithmen und durch Benchmarktests

Bewertungskriterien

Effektivität

Fähigkeit des Systems, den Nutzenden die benötigte Information bei möglichst geringen Kosten an Zeit und Anstrengung anzubieten:

- möglichst alle (Recall)
- möglichst nichts anderes (Precision)

→ Test durch empirische Methoden

Relevanz

Zentrales Problem:

Bei der Beurteilung von Retrievalergebnissen muss die richtige Antwort bekannt sein, um die Antwort des Systems bewerten zu können.

Konkret:
Um zu überprüfen, ob die richtigen Dokumente gefunden werden, muß bekannt sein, welche Dokumente zu der Anfrage „gehören“.

Konstrukt der **Relevanz**:

Beziehung zwischen einer Anfrage und einem Dokument.

Relevanz

Salton und McGill (1983) zitieren Cuadra und Katter:

> *... relevance is the correspondence in context between an information requirement statement (a query) and an article (a document), that is, the extent to which the article covers the material that is appropriate to the requirement statement ...*

Relevanz

- Die Frage, wann ein Dokument „das Material abdeckt, das für die Anfrage angemessen ist“, ist eine neue Umschreibung des eigentlichen Problems des Information Retrieval.
- In der Praxis wird keine Definition von Relevanz benutzt, sondern es werden Personen gebeten, die Relevanz einzuschätzen.
- Man verwendet den intuitiven, umgangsprachlichen oder naiven Begriff, um Relevanz zu bestimmen.
- So gesehen sind Evaluationen von Retrieval Systemen Untersuchungen, die menschliches Verhalten simulieren wollen.

Relevanz - Definition

Die Relevanz eines Dokumentes für eine Anfrage ist eine Relation, r: D x Q $\rightarrow$ R wobei D = {d_1, ... ,d_m} die Menge der Dokumente, Q die Menge der Anfragen und R eine Menge von „Wahrheitswerten“, i.a. die Menge {0,1} ist.

Die Relation r wird im Allgemeinen durch Befragen von Experten zu konkreten Anfragen und Dokumentmengen ermittelt und als Tabelle oder in Form von Listen gespeichert.

Einschränkung durch diese Definition:

- die Relevanz eines Dokuments für eine Anfrage hängt lediglich von der Anfrage und dem Dokument ab
- keine Berücksichtigung früherer Dokumente oder des Wissensstandes des Nutzers

Arten der Relevanz

situative Relevanz
(tatsächliche) Nützlichkeit des Dokumentes in Bezug auf die Aufgabe, aus der heraus das Informationsbedürfnis entstanden ist

Pertinenz
subjektiv vom Benutzer empfundene Nützlichkeit des Dokumentes in Bezug auf das Informationsbedürfnis

objektive Relevanz
von neutralem Beobachter beurteilte Beziehung zwischen dem geäußerten Informationswunsch und dem Dokument

Systemrelevanz
von einem automatischen System geschätzte Relevanz des Dokumentes in Bezug auf die formale Anfrage

Arten der Relevanz

im folgenden:

- keine Unterscheidung zwischen objektiver Relevanz und Pertinenz
- Relevanzskala zweistufig (relevant / nicht relevant)

Historie

Epochen der IR-Evaluation

Evaluierung ist immer ein Thema im IR

- PHASE I (1950-80)
  - Cranfield I und II, viele individuelle Studien, Beginn von SMART, MEDLARS
- PHASE II (1980-90)
  - Standardisierungsbestrebungen im Testdesign, Festlegung und Operationalisierung evaluierbarer Systemeigenschaften, in Deutschland: PADOK, AIR, LIVE...
- PHASE III (1990-...)
  - TREC, GIRT, CLIR, CLEF

Gewonnene Erkenntnisse aus Phase I

- „Information processes need scientific study“ (Sparck Jones, 1981), Standardisierung der Experimente und Systembeschreibungen
- Vergleich manueller Verfahren mit automatischen oder halb-automatischen Alternativen
- zufällige, nicht interpretierbare und schwierig kommunizierbare Ergebnisse aufgrund individueller und heterogener Testgestaltung
- Einsicht, daß nur mittlere Performanzwerte erreichbar sind (sehr gut ist 40-60% Recall und Precision)
- Inverse Beziehung zwischen Recall und Precision wird als Gesetz formuliert (Cranfield II)

Gewonnene Erkenntnisse aus Phase I

- Kluft zwischen Retrievalpraxis und -forschung beginnt sich abzuzeichnen.
- „experimental work is hard work“

Gewonnene Erkenntnisse aus Phase II

- Testtypen: experiment vs. investigation (laboratory vs. real-world) [Roberston 1981]
- Erkenntnisse über statistisch wohlgeformte Evaluierungsgrundlagen
- Beginn der Diskussion über standardisierte Testkollektionen
- Theoretische Auseinandersetzung: Retrievaleffektivität und Meßtheorie [Van Rijsbergen 1981]
- Auseinandersetzung mit der Problematik von Bedeutung, *aboutness* und Relevanz [Belkin 1981]

Gewonnene Erkenntnisse aus Phase II

- Operationalisierung der IR-Evaluierung [Tague 1981]
- Mehrfachnutzung von Testkollektionen führt erstmals zu Vergleichbarkeit
- wichtigste Erkenntnis: je komplexer die Systeme, desto schwieriger die Evaluierung

Ab 2000

- TREC spezialisiert sich zunehmend auf Fragestellungen, die in den USA interessieren (Chinesisches, Arabisches IR)
- Europa übernimmt die Retrievalbewertung in den europäischen Sprachen (CLEF)
- CLEF: immer mehr Sprachen kommen hinzu
  - E, F, I, S, D als Standard
  - JP, R, NL, SW, FN, Thai etc.

Evaluationsarten

Evaluationsarten

- Rapid-Prototyping
- Simulationstest (Wizard-of-Oz-Experimente)
- Kontrollierte Experimente
- Untersuchungen
- Empirische Langzeitstudien
- Heuristische Evaluierung
- Managementmethoden wie
  - (Kritische) Erfolgsfaktorenanalyse
  - Benchmarking

Was wird evaluiert?

⇨ in Abhängigkeit davon die Wahl des Evaluierungs-werkzeugs

- Information-Retrieval-Systeme (IRS-Komponente)
- Benutzerschnittstelle bzw. Interaktionsparadigma? (z.B. Menü / Direktmanipulation / natürlichsprachlich (auch speech))
- Evaluierung innerhalb eines Interaktionsparadigmas? (z.B. natürliche Sprache)

Formen der Evaluation

Datenanalytische Verfahren:

- Vergleich: automatische Indexierung mit manuellem Pendant
- differenzierte Fehlerbewertung
  - wichtig sind die Fehler, die den Zugang zum Dokument verwehren

Statistische und qualitative Verfahren:

- Verbindung von Recall-Precision-Zahlen mit bestimmten Strategien
- z.B. Anzahl der Interaktionen
- Anzahl der Deskriptoren (enge und weite Anfragen) etc.

Bewertung der Effektivität

Effektivitätsmessung

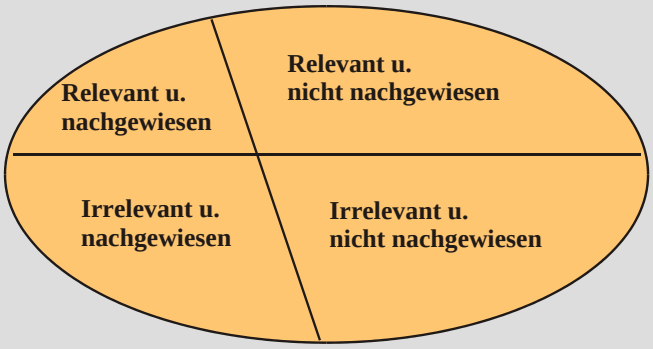
„ a measure of the ability of the system to retrieve relevant documents while at the same time holding back non-relevant ones“

[Van Rijsbergen 1979]

Klassische Effektivitätsmaße

Ausgangssituation:

- Bezüglich einer Anfrage liegt eine Treffermenge vor, die vom Benutzer als (im Einzelnen) relevant bzw. nicht relevant eingeschätzt wird.
- Zudem ist anzunehmen, daß der nicht in der Treffermenge enthaltene Teil der Kollektion ebenfalls in relevante und nicht-relevante Dokumente unterschieden werden kann, bildlich:



Precision und Recall - Definition

Sei D = {d_1, ..., d_m} eine Menge von Dokumenten, q $\in$ Q eine Anfrage und D_q die Menge der in D zur Anfrage q gefundenen Dokumente. Sei ferner r: D x Q $\rightarrow$ {0,1} eine Relevanzrelation und r_q: D $\rightarrow$ {0,1} ; $r_q(d) := r(d,q)$ die zur Anfrage q gehörende Relevanzfunktion. Dann heißt

$$P(q,D) := \frac{\left|D_q \cap r_q^{-1}(\{1\})\right|}{\left|D_q\right|}$$

Precision, **Präzision** oder **Genauigkeit** der Antwort auf die Anfrage q und

$$R(q,D) := \frac{\left|D_q \cap r_q^{-1}(\{1\})\right|}{\left|r_q^{-1}(\{1\})\right|}$$

Recall oder **Vollständigkeit** der Antwort auf die Anfrage q.

Precision und Recall

Precision:
Anteil der relevanten Dokumente unter den gefundenen Dokumenten

Recall:
Anteil der relevanten Dokumente, die gefunden wurden.

Optimal:

$D_q = r_q^{-1}(\{1\})$ alle relevanten Dokumente als Antwortmenge zurückgeliefert

- beide Maße sind gegenläufig

Precision und Recall - Extremfälle

1. Alle Dokumente werden zurückgeliefert $D_q = D$:

$$R(q,D) := \frac{\left|D \cap r_q^{-1}(\{1\})\right|}{\left|r_q^{-1}(\{1\})\right|} = \frac{\left|r_q^{-1}(\{1\})\right|}{\left|r_q^{-1}(\{1\})\right|} = 1$$

2. Nur ein einziges relevantes Dokument wird gefunden $d_r \in D$:

$$P(q,D) := \frac{\left|D \cap r_q^{-1}(\{1\})\right|}{\left|\{d_r\}\right|} = \frac{\left|\{d_r\}\right|}{\left|\{d_r\}\right|} = 1$$

- In der Regel werden die Antwortmengen aber zwischen diesen beiden Extremen liegen.

Precision und Recall

Verkleinerung der Antwortmenge durch spezifischere Anfrage:

- bessere Precision
- aber ein schlechterer Recall

Vergrößerung der Antwortmenge durch allgemeinere Anfrage:

- größerer Recall
- aber ein kleinere Precision

Eindeutige Aussagen darüber, ob ein System besser ist als ein anderes, können nur gemacht werden, wenn für das eine System sowohl der

- Precisionwert als auch der
- Recallwert

besser ist als bei dem anderen System.

Precision und Recall

<u>Bemerkungen</u>
Die Ermittlung des Recall-Wertes basiert auf der Bestimmung der Anzahl *aller* relevanten Dokumente im *gesamten* Datenbestand. Da dies oftmals nicht realistisch ist, werden Näherungsverfahren benutzt:

- vollständige Relevanzbeurteilung einer repräsentativen Stichprobe der Datenbank

$$\frac{\text{Anzahl relevanter Dok. in Stichprobe}}{\text{Größe der Stichprobe}} \approx \frac{r}{f_d}$$

- vollständige Relevanzbeurteilung der Ergebnis-vereinigung verschiedener Anfrageformulierungen

$$\text{Anzahl relevanter Dokumente in Vereinigung} \approx r$$

Precision und Recall

Bemerkungen (Forts.)

- Die Bestimmung beider Maße setzt voraus, daß die gesamte Antwortmenge „untersucht“ wurde. Dies ist typischerweise nicht der Fall!
- Stattdessen wird der Benutzer nur einige der ersten Dokumente in der geordneten Ergebnisliste (Ranking) betrachten.

Folge

- Die Maße Recall und Precision variieren je nach Größe der betrachteten Ergebnismenge.

Recall-Precision-Diagramm

Recall-Precision-Diagramme sind ein Standardverfahren zur Evaluation von IR-Systemen. Sie eignen sich gut zum Vergleich von unterschiedlichen Retrieval-Algorithmen.

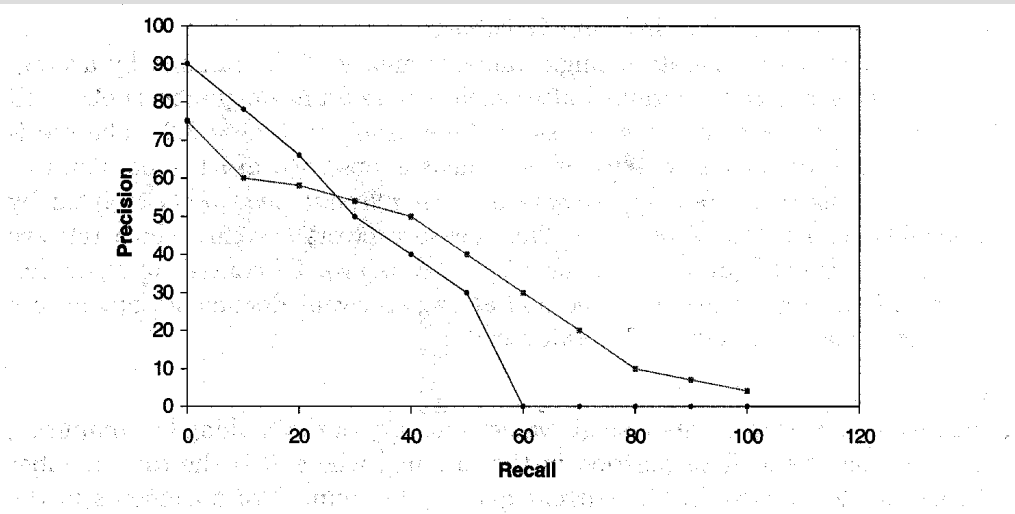


Figure 3.4 Average recall versus precision figures for two distinct retrieval algorithms.

Recall-Precision-Diagramm

Definition:
Sei $D_q = \{d_{s1},...,d_{sk}\}$ eine vollständig geordnete Antwortmenge $r_q^{-1}(\{1\})$, die Menge der zu der Anfrage q relevanten Dokumente und $(d_{t1}, ..., d_{tl})$ die Schnittmenge $D_q \cap r_q^{-1}(\{1\})$ in der Anordnung aus D_q. Dann bezeichnet die Folge $(R_i(q,D),P_i(q,D))_{i=1,...,l}$ mit

$$R_i(q,D) := \frac{\left|(d_{t_1},...,d_{t_i})\right|}{\left|r_q^{-1}(\{1\})\right|}$$

und

$$P_i(q,D) := \frac{\left|(d_{t_1},...,d_{t_i})\right|}{\left|(d_{s_1},...,d_{s_j})\right|}$$

mit $d_{sj} = d_{ti}$ das zu q gehörige **Recall-Precision-Diagramm**.

- Dieses Diagramm kann durch Punkte im Quadrat $[0,1]^2$ dargestellt werden.

Beispiel eines Recall-Precision-Diagramms

Es sei die folgende vollständig geordnete Antwortmenge gegeben, bei der ein R für ein relevantes Dokument und ein U für ein nicht relevantes Dokument steht.

RURRURRRUU	URRURRRUUU	URUUUURUUU	UUUUURUUUU
RURUUUUURU	UUURUUUURU	UUUUURUUUU	UUUUURUUUU
UUUUUUUURU	UUUUUUUURU	UUUUUUUURU	UUUUUUUUUU
UUUUURUUUU	UUUUUUURUU	UUUUUURUUU	UUUUUUURUU
UUUUUUUUUU	RUUUUUUUUU	UUUUUURUUU	UU...

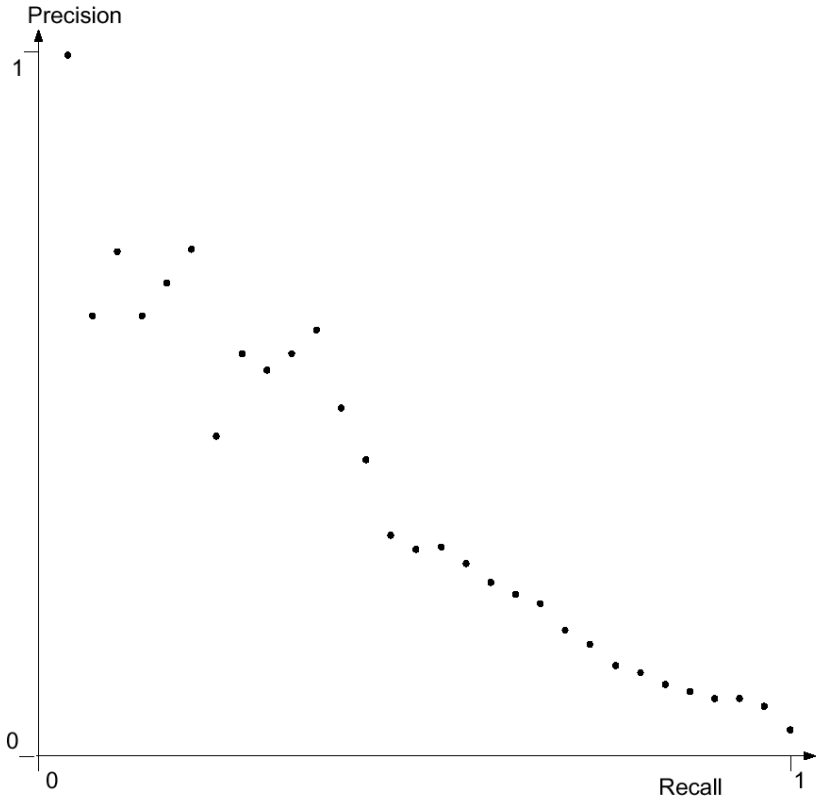
Beispiel eines Recall-Precision-Diagramms

Die Folge $(R_i(q,D),P_i(q,D))_{i=1,\ldots,30}$ sieht dann so aus:

$$\left(\frac{1}{30},\frac{1}{1}\right),\left(\frac{2}{30},\frac{2}{3}\right),\left(\frac{3}{30},\frac{3}{4}\right),\left(\frac{4}{30},\frac{4}{6}\right),\left(\frac{5}{30},\frac{5}{7}\right),$$
$$\left(\frac{6}{30},\frac{6}{8}\right),\left(\frac{7}{30},\frac{7}{12}\right),\left(\frac{8}{30},\frac{8}{13}\right),\left(\frac{9}{30},\frac{9}{15}\right),\left(\frac{10}{30},\frac{10}{16}\right),$$
$$\left(\frac{11}{30},\frac{11}{17}\right),\left(\frac{12}{30},\frac{12}{22}\right),\left(\frac{13}{30},\frac{13}{27}\right),\left(\frac{14}{30},\frac{14}{36}\right),\left(\frac{15}{30},\frac{15}{41}\right),$$
$$\left(\frac{16}{30},\frac{16}{43}\right),\left(\frac{17}{30},\frac{17}{49}\right),\left(\frac{18}{30},\frac{18}{54}\right),\left(\frac{19}{30},\frac{19}{59}\right),\left(\frac{20}{30},\frac{20}{66}\right),$$
$$\left(\frac{21}{30},\frac{21}{76}\right),\left(\frac{22}{30},\frac{22}{89}\right),\left(\frac{23}{30},\frac{23}{99}\right),\left(\frac{24}{30},\frac{24}{109}\right),\left(\frac{25}{30},\frac{25}{126}\right),$$
$$\left(\frac{26}{30},\frac{26}{138}\right),\left(\frac{27}{30},\frac{27}{147}\right),\left(\frac{28}{30},\frac{28}{158}\right),\left(\frac{29}{30},\frac{29}{171}\right),\left(\frac{30}{30},\frac{30}{187}\right)$$

Dr. Jens E. Wolff

Beispiel eines Recall-Precision-Diagramms



Dr. Jens E. Wolff

Recall-Precision-Diagramm

Bemerkung:

- Falls die Antwortmenge nicht vollständig geordnet ist, muß darauf geachtet werden, dass die Werte in der Folge nicht durch willkürliche Vergabe der Rangplätze beeinflusst werden.
- Wäre jeder Block eine Gruppe von Dokumenten mit gleicher Ähnlichkeit zur Anfrage, könnten die Dokumente innerhalb der Blöcke zufällig angeordnet werden.
- Dadurch würde das Precision-Recall-Diagramm beeinflusst.
- Es könnte das 14. Element der Folge auch $\left(\frac{14}{30}, \frac{14}{31}\right)$ oder $\left(\frac{14}{30}, \frac{14}{40}\right)$

Mittelwertbildung

Zur Bewertung eines Retrieval-Verfahrens wird nicht nur eine Anfrage, sondern eine Menge von Anfragen betrachtet. Die berechneten Recall/Precision-Werte werden dann über alle Anfragen gemittelt.

1. **Makrobewertung** oder der **nutzungsorientierte** Ansatz:

$$P_u(D) := \frac{1}{N}\sum_{i=1}^{N}\frac{\left|D_{q_i} \cap r_{q_i}^{-1}(\{1\})\right|}{\left|D_{q_i}\right|}$$

$$R_u(D) := \frac{1}{N}\sum_{i=1}^{N}\frac{\left|D_{q_i} \cap r_{q_i}^{-1}(\{1\})\right|}{\left|r_{q_i}^{-1}(\{1\})\right|}$$

Mittelwert gemäß des arithmetischen Mittels.

Mittelwertbildung

2. **Mikrobewertung** oder der **systemorientierte** Ansatz:

$$P_s(D) := \frac{\sum_{i=1}^{N} \left| D_{q_i} \cap r_{q_i}^{-1}(\{1\}) \right|}{\sum_{i=1}^{N} \left| D_{q_i} \right|}$$

$$R_s(D) := \frac{\sum_{i=1}^{N} \left| D_{q_i} \cap r_{q_i}^{-1}(\{1\}) \right|}{\sum_{i=1}^{N} \left| r_{q_i}^{-1}(\{1\}) \right|}$$

Mittelwert gemäß der Anzahl der beteiligten Dokumente (Anfragen mit wenigen relevanten Dokumenten spielen eine kleinere Rolle als solche mit vielen relevanten Dokumenten).

Mittelwertbildung – Makro- und Mikrobewertung

Beispiel von Salton und McGill (1983):
10 Veranstaltungen: 5 mit je 99 Studierenden und 5 mit je nur einer Person

Makrobewertungs - oder die „class level“ - Durchschnittsgröße:

$$\frac{5 \cdot 1 + 5 \cdot 99}{10} = 50$$

Mikrobewertungs - oder „student level“ - Durchschnittsgröße:

$$\frac{5 \cdot 99 \cdot 99 + 5 \cdot 1 \cdot 1}{5 \cdot 99 + 5 \cdot 1} = 98{,}02$$

Mittelwertbildung

- Mit Precision-Recall-Diagrammen können verschiedene Systeme nicht immer eindeutig verglichen werden.
- Nur wenn die Precision eines Systems für alle Recallwerte besser ist als die eines anderen, kann man sagen, dass dieses System besser ist als das andere.
- Ist für einen Recallwert die Precision des einen Systems höher und für einen anderen die des anderen, ist eine generelle Aussage, welches System besser ist, erstmal nicht möglich.
- Um Systeme in jedem Fall vergleichen zu können bzw. in eine Rangfolge bringen zu können, verwendet man häufig die **mittlere Precision**.

Mittelwertbildung

- Die mittlere Precision wird als Mittelwert der Precision-werte an einer fest vorgegebenen Menge von Recall-werten, z.B. den Recallwerten {0.1, 0.2, 0.3,...,0.9} oder {0.75, 0.5, 0.25}, berechnet

⇨ Daraus ergibt sich eine reelle Zahl, nach der verschiedene Systeme in eine Rangfolge gebracht werden können.

- Die Precisionwerte für die gewählten Recallwerte gegebenenfalls Interpolieren, da Anzahl der relevanten Dokumente ausschlaggebend
- oder den Precisionwert verwenden, bei dem der gesuchte Recall erstmal überschritten wird.

Interpolation - Beispiel

- Bezüglich einer Anfrage q sind 5 Dokumente relevant, die von einem Ranking-Verfahren auf den Plätzen 1, 2, 4, 6 und 13 plaziert werden.
- Sei r_j, $j \in \{0,1, \ldots ,10\}$ der Bezeichner für den j-ten Recall-Wert.
  Dann ist die interpolierte Precision für den j-ten Recall-Wert gegeben durch

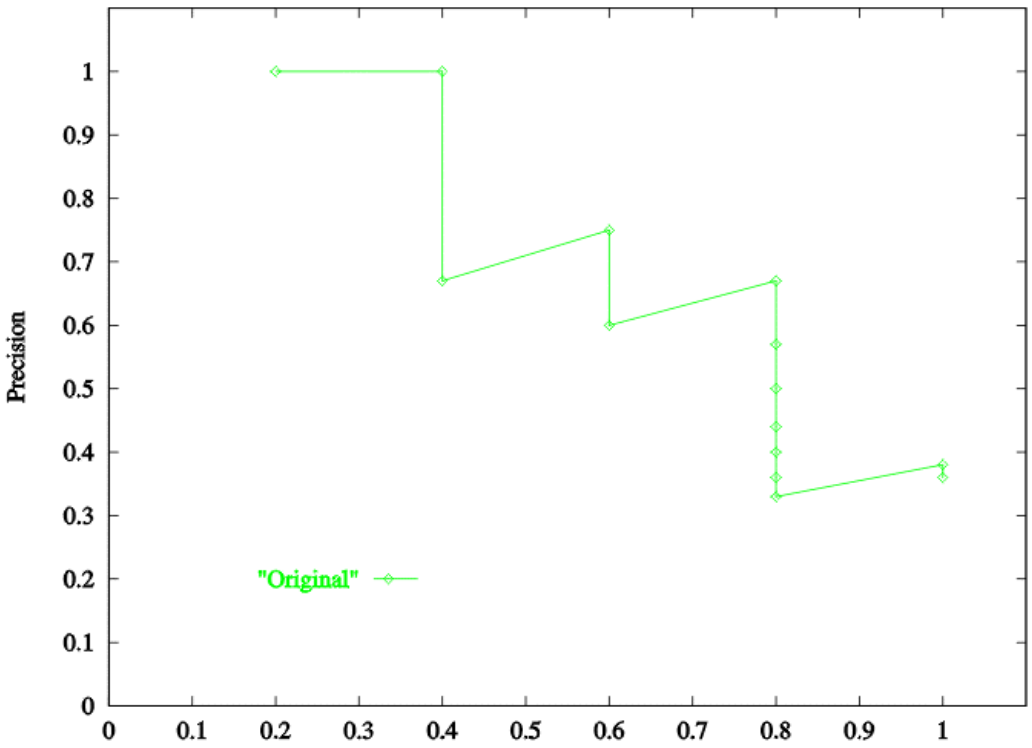
$$P(r_j) = \max_{r_j \le r \le r_{j+1}} P(r)$$

Interpolation - Beispiel

	Rang	Recall	Precision
•	1	0,2	1,0
•	2	0,4	1,0
	3	0,4	0,67
•	4	0,6	0,75
	5	0,6	0,6
•	6	0,8	0,67
	7	0,8	0,5
	8	0,8	0,57
	9	0,8	0,44
	10	0,8	0,4
	11	0,8	0,36
	12	0,8	0,33
•	13	1,0	0,38
	14	1,0	0,36

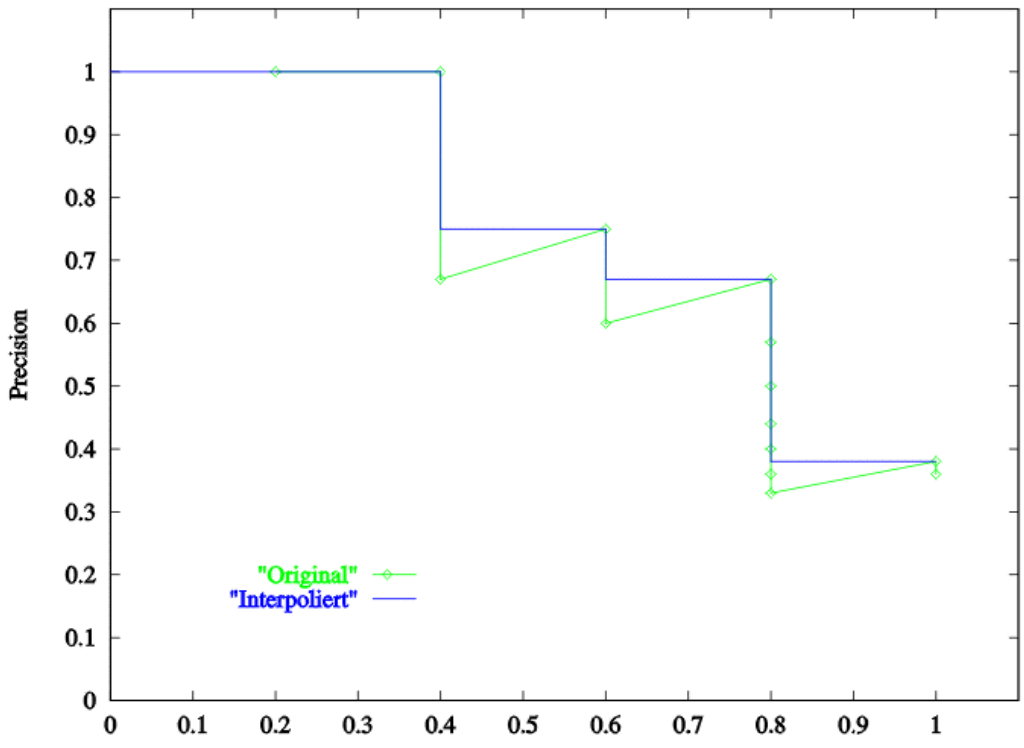
(r = 5, relevant •)

Interpolation - Beispiel



Dr. Jens E. Wolff

Interpolation - Beispiel



Dr. Jens E. Wolff

Weitere Effektivitätsmaße

- verschiedene Maße sind entwickelt worden, um in **einer** Meßgröße eine Systembewertung leisten zu können

Vorteile:

- Was sonst durch Precision und Recall ausgedrückt wird, kann dadurch in einem Maß zusammengefaßt werden.
- Unterschiedliche Benutzerpräferenzen lassen sich in das Maß mit einarbeiten. (Bevorzugung einer hohen Abdeckungsquote -> Recall bevorzugt vs. Bevorzugung einer möglichst auf das Ausgangsproblem beschränkten Treffermenge -> Precision bevorzugt).

E-Maß

Beispiel für ein eindimensionales Maß ist das E-Maß von Van Rijsbergen (1979):

$$E = 1 - \frac{1}{\alpha(1/P) + (1-\alpha)(1/R)}$$

α ist ein Parameter (in [0;1]), der die Gewichtung von Recall (R) und Precision (P) je nach Benutzergruppe erlaubt.

Auffälligkeiten beim E-Maß

- Falls entweder Recall oder Precision gleich 0 ist, ergibt sich für E immer der Wert 1 (d.h. das Maß kann nicht mehr zw. unterschiedlichen Recall bzw. Precision Werten differenzieren).
- Falls Recall und Precision einen gleichen Wert aufweisen, ergeben sich auch bei variierenden α-Werten keine unterschiedlichen E-Werte.

Maße ohne Bestimmung des Recall

Recall-precision measurements have not proved universally acceptable. Objections have been raised of a theoretical as well as practical nature. The most serious questions relate to the fact that recall, in particular, is apparently incompatible with the utility-theoretic approach to information retrieval, which forms the basis of a good deal of existing information retrieval theory. (Salton, 1992)

Da die Bestimmung des Recall eines der schwierigsten Probleme der Evaluierung darstellt, wurde versucht, Maße zu entwickeln, die eine Systembewertung ohne Rückgriff auf den Recall erlauben.

z.B. *Utility*-Maß nach Frei und Schäuble, 1991

Utility-Maß

- In die Bewertung gehen nur tatsächliche Benutzer-präferenzen bezüglich der Treffermenge (als Rankingbeurteilungen) ein.
- Das Maß errechnet über eine Anzahl von Anfragen einen Wert für die durchschnittlichen Unterschiede in den Rangordnungsbewertungen.
- Große Bedeutung für eine Reihe von Situationen in denen die Vollständigkeit (exhaustivity) der Suche nur ein nachrangiges Kriterium bildet.

Literatur

Literatur

- [Baeza-Yates & Ribeiro-Neto 99]
  - Baeza-Yates, R. und Ribeiro-Neto, B.: Modern Information Retrieval, Addison-Wesley, 1999. Kapitel 3.

- [Ferber 03]
  - Ferber, R.: Information Retrieval - Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. dpunkt.verlag, 2003. Kapitel 1.3.7. http://information-retrieval.de/