3. Evaluierung

Norbert Fuhr

Perspektiven auf IR-Systeme

- Benutzer
- Käufer
- Manager
- Hersteller
- Entwickler
- ...

Fragestellungen an die Evaluierung

- Was kann ich ändern, um die Qualität eines Systems zu verbessern?
- Welche Art der Textrepräsentation ist am besten?
- Welches Retrievalmodell liefert die besten Ergebnisse?
- Welche Qualität weist ein System auf?
- Welches System ist besser?
- Welches System soll ich kaufen?
- Wie kann ich Qualität messen?
- Was bedeutet Qualität für mich?

Anforderungen an die Ergebnisse einer Evaluierung

Vorbemerkung:

IR-Experimente sind *stochastische Experimente* !

$\rightsquigarrow$ Ergebnisse sind statistische Messgrößen

Anforderungen:

Reliabilität (Zuverlässigkeit) Der Grad, in dem dieselbe Untersuchung im gleichen Kontext dieselben Ergebnisse liefert (Wiederholbarkeit)

$\rightsquigarrow$ ausreichende Dokumentation, hinreichend große Stichprobe

Validität Der Grad der Übereinstimmung zwischen einer Beobachtung und den 'tatsächlichen' Verhältnissen (Gültigkeit)

$\rightsquigarrow$ repräsentative Stichprobe

Arten von Evaluierungen

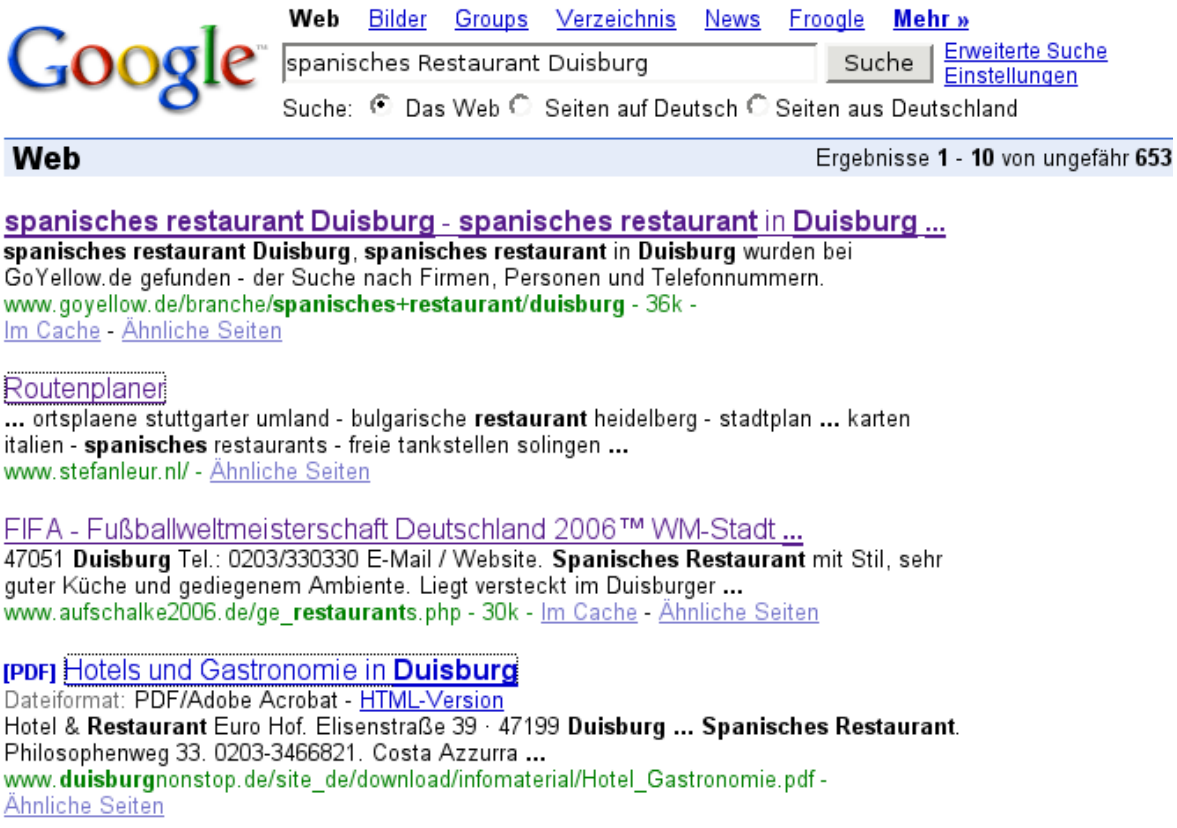
Formative Evaluierung zu Beginn der Systementwicklung: Festlegung von Funktionalität, Zielen und gewünschten Ergebnissen (keine Evaluierung im engeren Sinne)

Iterative Evaluierung begleitend zur Systementwicklung, als Basis für Entwurfsentscheidungen

Summative Evaluierung am Projektende: Gesamtbewertung des realisierten Systems

Komparative Evaluierung Vergleichende Evaluierung verschiedener Systeme

Qualitätskriterien



Wie gut ist die Antwort?

- Wie präzise?
- Wie vollständig?
- Wie schnell?

Effizienz — Effektivität

Effizienz

Nutzung der Systemressourcen für eine bestimmte Aufgabe
(Speicherplatz, I/O-Operationen, CPU-Zeit, Verweilzeit)

$$\text{Effizienz} \approx \frac{\text{Größe der Aufgabe}}{\text{Aufwand des Systems}}$$

Effektivität

Unterstützung des Benutzers (durch das System) bei seinem Anwendungsproblem

$$\text{Effektivität} \approx \frac{\text{Qualität der Lösung}}{\text{Aufwand des Benutzers}}$$

Qualität von Informationssystemen

Datenbanksysteme liefern stets korrekte und vollständige Antwort auf Anfragen

- im Sinne eines Beweisverfahrens
- i.a. *nicht* bezüglich der realen Welt

→ Betrachtung von Effektivität hier nicht sinnvoll

IR-Systeme können wegen Vagheit und Unsicherheit i.a.

- weder korrekte (alle gefundenen Dokumente relevant)
- noch vollständige (alle relevanten Dokumente)

Antworten liefern.
→ Effektivität als wichtiges Qualitätskriterium

Messung von Effektivität

$$\text{Effektivität} \approx \frac{\text{Qualität der Lösung}}{\text{Aufwand des Benutzers}}$$

a) Lösung fest vorgegeben
   - (z.B. Suche nach bestimmtem Dokument/Faktum)

   $\longrightarrow$ messe Zeitaufwand zum Finden der Lösung

b) Benutzeraufwand konstant
   - z.B. Benutzer formuliert nur eine Anfrage
   - z.B. feste Bearbeitungszeit

   $\longrightarrow$ messe Qualität der Lösung

im Folgenden nur Variante b) betrachtet:
$\rightsquigarrow$: Effektivität $\approx$ Retrievalqualität

Relevanz

("fiktive" Beziehung zwischen Anfragen und Dokumenten)
als Mittel zur Beurteilung von Retrievalalgorithmen

Annahmen

- Systemantwort ist eine Menge von Dokumenten
- Qualität des Dokuments hängt nur von der Anfrage ab

Probleme

- Systemantwort kann strukturiert sein
- Dokumente nicht unabhängig
- Keine einfache Beziehung zwischen Informationsbedürfnis (umgangssprachlich/subjektiv) und Anfrage (formal)

Arten von 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 (besser: Retrievalwert oder retrieval status value)

im folgenden:

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

Distributionen

abstrakte Darstellung von Retrievalantworten,
Grundlage für Bewertungsmaße
vom System berechnete Retrievalwerte:
$\{(d_1, 0.3), (d_2, 0.8), (d_3, 0.1), (d_4, 0.8),$
$(d_5, 0.8), (d_6, 0.6), (d_7, 0.3), (d_8, 0.1)\}$
$\longrightarrow$ Distribution von Dokumenten:

$$(\{d_2, d_4, d_5\}, \{d_6\}, \{d_1, d_7\}, \{d_3, d_8\})$$

Distribution: $(\{d_2, d_4, d_5\}, \{d_6\}, \{d_1, d_7\}, \{d_3, d_8\})$
Relevanzbeurteilung:
$\{(d_1, R), (d_2, R), (d_3, \bar{R}), (d_4, R), (d_5, R), \quad (d_6, \bar{R}), (d_7, R), (d_8, R)\}$
$\longrightarrow$ Distribution mit Relevanzurteilen

$$(\{d_2^+, d_4^+, d_5^+\}, \{d_6^-\}, \{d_1^+, d_7^+\}, \{d_3^-, d_8^+\})$$

Abstraktion von spezifischen Dokumenten
$\longrightarrow$ Äquivalenzklasse von Distributionen (im folgenden einfach Distributionen)

$$\Delta = (+ + +| - | + +| + -)$$

Standpunkte und Bewertungsmaße

Benutzerstandpunkte

mögliche Standpunkte eines Benutzers:
Er schaut die Rangliste der gefundenen Dokumente durch, bis

a) n Dokumente gesehen

b) n relevante Dokumente gesehen

c) n nicht relevante Dokumente gesehen

Bewertungsmaße

Wahl eines entsprechenden Bewertungsmaßes:
soll Präferenzen des Benutzers widerspiegeln

mögliche Maße für obige Standpunkte:

a) *n* Dokumente gesehen:
   # gesehene relevante Dokumente

b) *n* relevante Dokumente gesehen:
   # gesehene Dokumente

c) *n* nicht relevante Dokumente gesehen:
   # gesehene / # gesehene relevante Dokumente

Benutzer- vs. Systemstandpunkte

benutzerorientierte Maße
beziehen sich auf mögliches Verhalten und Präferenzen der Benutzer

systemorientierte Maße
entsprechen einer systemorientierten Sicht

- unabhängig von speziellen Benutzerstandpunkten
- streben „globale“ Bewertung der Distribution an
  (obige benutzerorientierte Maße betrachten jeweils nur die obersten Ränge)

Beispiel für systemorientiertes Maß

Korrelation zwischen Systemantwort Δ und idealer Distribution Δ'

S^+: # richtig angeordnete Paare
S^-: # falsch angeordnete Paare
$S_{\max}$: # richtig angeordnete Paare der optimalen Lösung

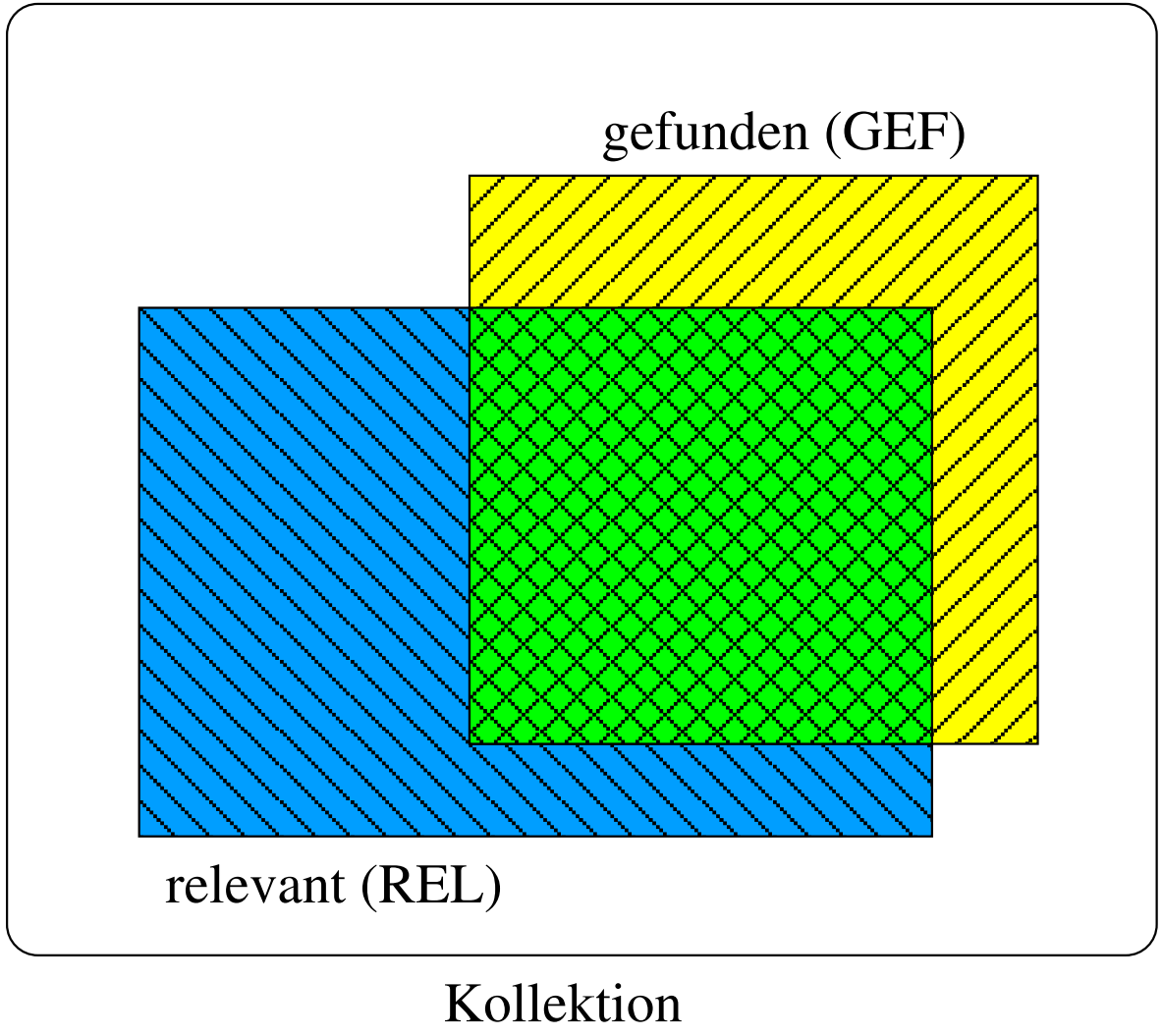
$$\varrho = \frac{S^+ - S^-}{S_{\max}}$$

$\Delta = (+ + +| - | + +| + -)$
$\Delta' = (+ + + + + + | - -)$

$$\varrho = \frac{8 - 3}{12} = \frac{5}{12}$$

Maße für boolesches Retrieval

Benutzerstandpunkt:
jeweils alle gefundenen Dokumente betrachtet



GEF: Menge der gefundenen Antwortdokumente

REL: Menge der relevanten Dokumente in der Datenbank

ALL: Menge aller Dokumente in der Datenbank

$$\text{Precision} \quad p = \frac{|REL \cap GEF|}{|GEF|}$$

$$\text{Recall} \quad r = \frac{|REL \cap GEF|}{|REL|}$$

$$\text{Fallout} \quad f = \frac{|GEF - REL|}{|ALL - REL|}$$

Probabilistische Interpretation der Retrievalmaße

$$\text{Precision} \quad p = \frac{|REL \cap GEF|}{|GEF|}$$

$\approx$ Wahrscheinlichkeit, dass (zufällig ausgewähltes) gefundenes Dokument relevant ist

$$\text{Recall} \quad r = \frac{|REL \cap GEF|}{|REL|}$$

$\approx$ Wahrscheinlichkeit, dass (zufällig ausgewähltes) relevantes Dokument gefunden wird

Anpassung der Maße an den Anwendungskontext

Web-Retrieval Benutzer schaut sich die erste Seite der Ergebnisse an (10 Dokumente):
Prec@10: Precision nach 10 Dokumenten/
Relevanz des ersten Dokumentes: Prec@1

Evaluierungsinitiativen (TREC, CLEF, INEX):

- Prec@5, Prec@10, Prec@30, Prec@100
- MAP (mean average precision): mittlere Precision nach jedem relevanten Dokument

Recall-Bestimmung

REL (=Gesamtzahl relevanter Dokumente) kann nur näherungsweise bestimmt werden

1. vollständige Relevanzbeurteilung einer Stichprobe
2. Document-source-Methode
3. Abgleich mit externen Quellen
4. Frageerweiterung
5. Pooling-Methode

- Nur die letzten beiden Methoden sind praktikabel
- aber: sie liefern nur untere Schranken für $|REL|$
- $\longrightarrow$ Recall-Werte im Allgemeinen zu optimistisch

$\longrightarrow$

Vollständige Relevanzbeurteilung einer Stichprobe

1. ziehe zufällige Stichprobe aus der Datenbasis
2. Relevanzbeurteilung aller Dokumente der Stichprobe

Beispiel:

- 10^6 Dokumente in der Datenbasis
- 100 relevante Dokumente
- 1000 Dokumente in der Stichprobe

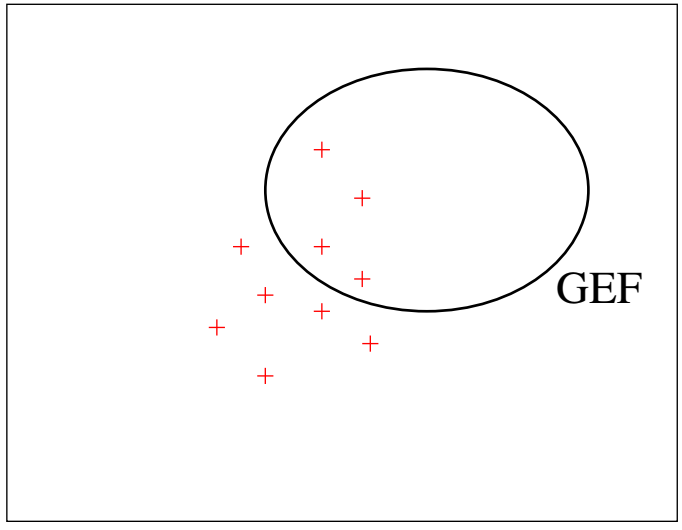
$\rightsquigarrow$: erwartete Anzahl relevanter Dokumente in der Stichprobe: 0,1
$\rightarrow$ für realistische Anwendungen kaum möglich

Document-source-Methode

1. Formuliere Anfrage zu gegebenem Dokument
2. Dokumente gefunden $\rightarrow$ r=1
   Dokument nicht gefunden $\rightarrow$ r=0
3. Schätze Recall als Mittelwert über größere Anzahl solcher Anfragen

(primär geeignet für Dokumentationssprachen als Textrepräsentation)
keine realen Anfragen!

Abgleich mit externen Quellen



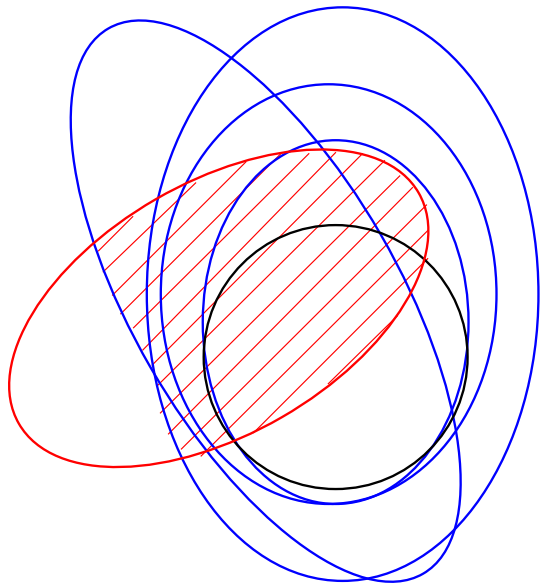
+ relevantes Dokument

- Wähle Anfragen, zu denen relevante Dokumente bekannt sind
- Schätze Recall als Anteil der gefundenen an den bekannten relevanten Dokumenten

sehr aufwändig

Frageerweiterung

Frage umformulieren, so dass (fast) alle relevanten Dokumente gefunden werden



gefundene Dokumente

relevante Dokumente

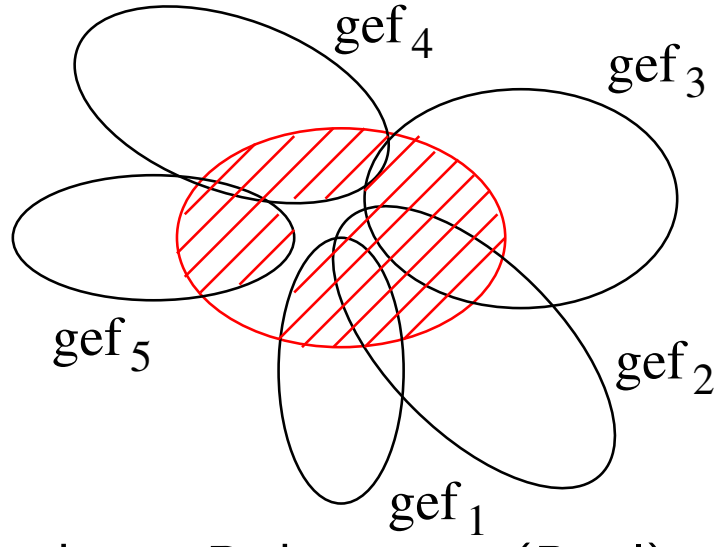
Frageerweiterung

$\rightsquigarrow$ relevante Dokumente in den zusätzlich gefundenen bestimmen

hoher Aufwand zur Fragereformulierung und Relevanzbeurteilung zusätzlicher Dokumente

Pooling-Methode (Retrieval mit mehreren Systemen)

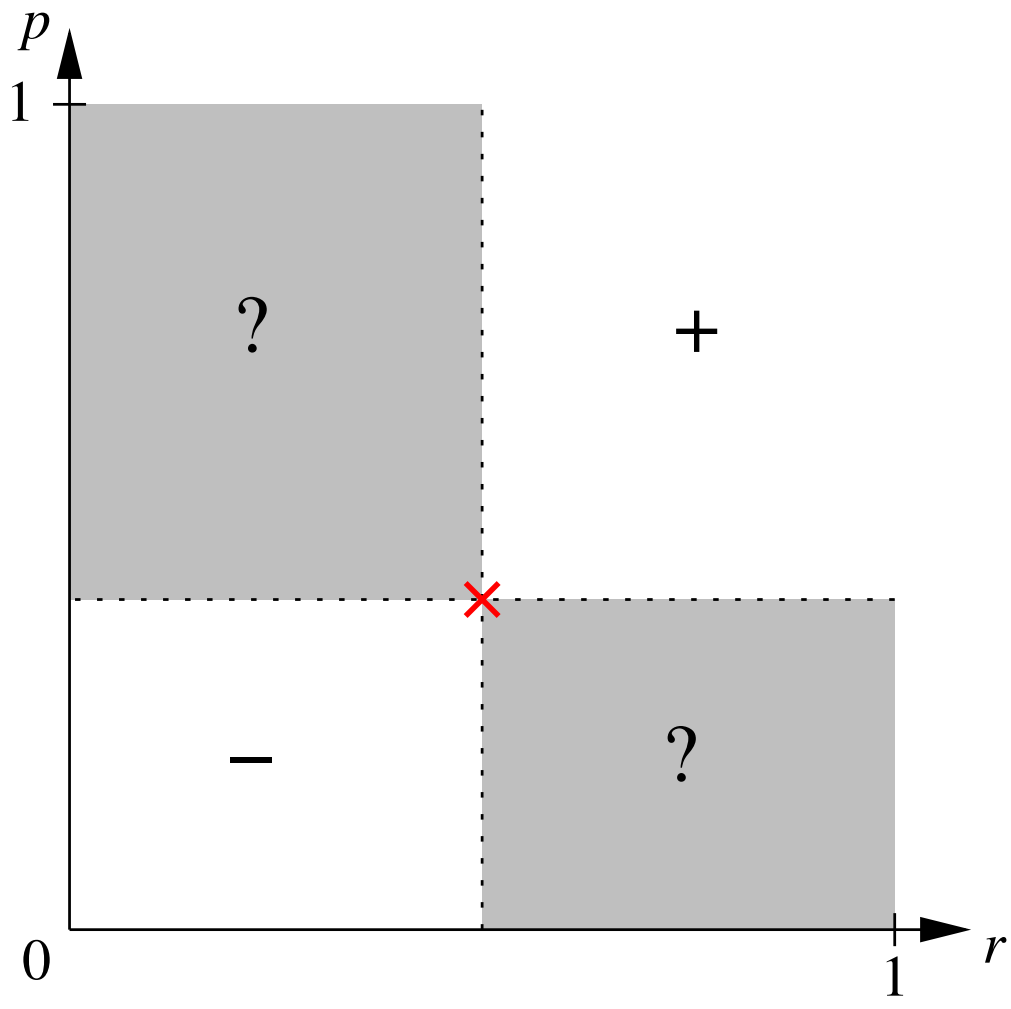
Prozessiere dieselbe Anfrage mit verschiedenen Systemen auf der gleichen Kollektion



- Bilde Vereinigungsmenge der gefundenen Dokumente (Pool)
- Relevanzbeurteilung aller Dokumente im Pool
- # relevante Dokumente im Pool $\approx$ # relevante Dokumente in der Kollektion

zusätzlicher Aufwand zur Beurteilung der Pools,
aber derzeit am meisten angewandte Methode

Vergleich von Recall-Precision-Paaren



F-Maß

Abbildung von (r, p)-Paar auf einzelnes Maß
(definiert Kurve zur Aufteilung des 'Unentschieden-Bereichs')

Grundidee:
harmonisches Mittel aus Recall und Precision

$$F = \frac{2}{\frac{1}{p} + \frac{1}{r}}$$

Unterschiedliche Gewichtung von Recall und Precision:
Gewichtungsfaktor β für Recall

$$F_\beta = \frac{1 + \beta^2}{\frac{1}{p} + \beta^2 \frac{1}{r}}$$

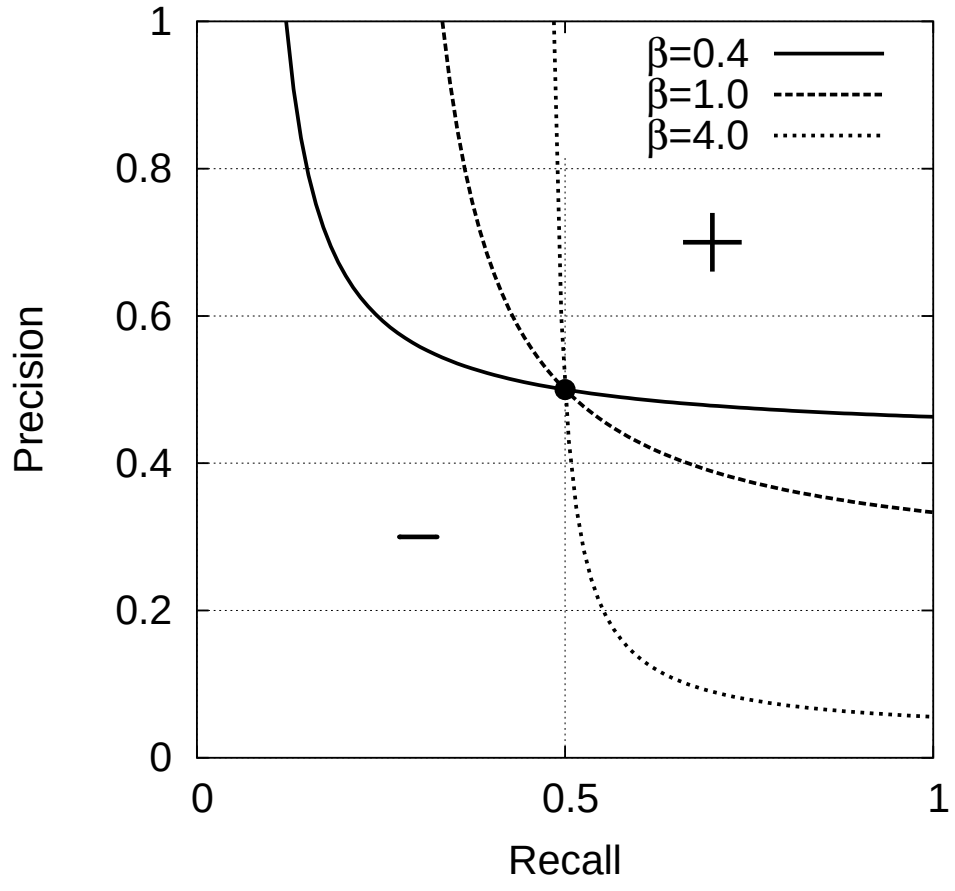
F_β-Maß mit $0 \leq \beta \leq \infty$

$$F_\beta = \frac{(\beta^2 + 1) \cdot p \cdot r}{\beta^2 \cdot p + r}$$

β: relative Wichtung des Recall

- $\beta = 0$: F-Maß entspricht Precision
- $\beta = \infty$: F-Maß entspricht Recall
- $\beta = 1$: Recall und Precision gleich stark gewichtet

Aufteilung von Recall-Precision-Punkten durch das F-Maß



Kostenmaße

als Alternative zu Recall und Precision, z.B. bei Filterung

Matrix mit Anzahl der Dokumente für jeden Fall:

	relevant	$\overline{\text{relevant}}$
gefunden	h_g^R	h_g^I
$\overline{\text{gefunden}}$	h_n^R	h_n^I

Gesamtkosten:

$C = C_g^R \cdot h_g^R + C_g^I \cdot h_g^I + C_n^R \cdot h_n^R + C_n^I \cdot h_n^I$

z.B. $C_g^R = C_n^I = 0,\ C_g^I = C_n^R = 1$

Beispiel: Spam-Filterung:

möglichst keine 'irrelevante' (nicht-Spam-) Mail sollte vom Filter selektiert werden

$\rightarrow h_g^I$ möglichst klein: $C_g^I \gg C_n^R$

Systemvergleich

IR-Experimente sind *stochastische Experimente*
$\rightsquigarrow$ Maße müssen für ausreichend große Menge von Anfragen berechnet werden

Vergleiche:

a) Frageweiser Vergleich zwischen zwei Systemen:
   Anzahl Fragen, bei denen A besser / B besser
   ($\rightarrow$ Vorzeichentest, Wilcoxon-Test, t-Test)

b) Mittelwerte von Recall und Precision über k Fragen
   - Makrobewertung (arithmetisches Mittel)
   - Mikrobewertung

Mittelwertbildung
Makrobewertung

$$p_M = \frac{1}{N} \sum_{i=1}^{N} \frac{|REL_i \cap GEF_i|}{|GEF_i|}$$

Frage-bezogen, Mittelwertbildung über eine Gruppe von Benutzern mit gleichem Standpunkt
(approximiert Erwartungswert für die Precision zu einer zufällig ausgewählten Anfrage)

Mikrobewertung

$$p_\mu = \frac{\sum_{i=1}^{N} |REL_i \cap GEF_i|}{\sum_{i=1}^{N} |GEF_i|}$$

Dokument-bezogen, realisiert systemorientierte Sicht
(approximiert Wahrscheinlichkeit, dass (zufällig ausgewähltes) gefundenes Dokument aus einer der N Anfragen relevant ist)
analog für Recall

Mikro-Precision: fehlende Monotonie

Monotonie:
Vergleichsaussage zu zwei Ergebnissen bleibt unverändert, wenn die Aussage durch Hinzufügen derselben Antwort zu beiden Resultaten unverändert bleibt.

$$p_\mu(\Delta_1) = \frac{1}{2}, \quad p_\mu(\Delta_2) = \frac{2}{5} \qquad p_\mu(\Delta) = \frac{2}{8}$$

$$p_\mu(\Delta_1) = \frac{1}{2} \quad > \quad \frac{2}{5} = p_\mu(\Delta_2), \text{ aber}$$

$$p_\mu(\Delta_1, \Delta) = \frac{3}{10} \quad < \quad \frac{4}{13} = p_\mu(\Delta_2, \Delta).$$

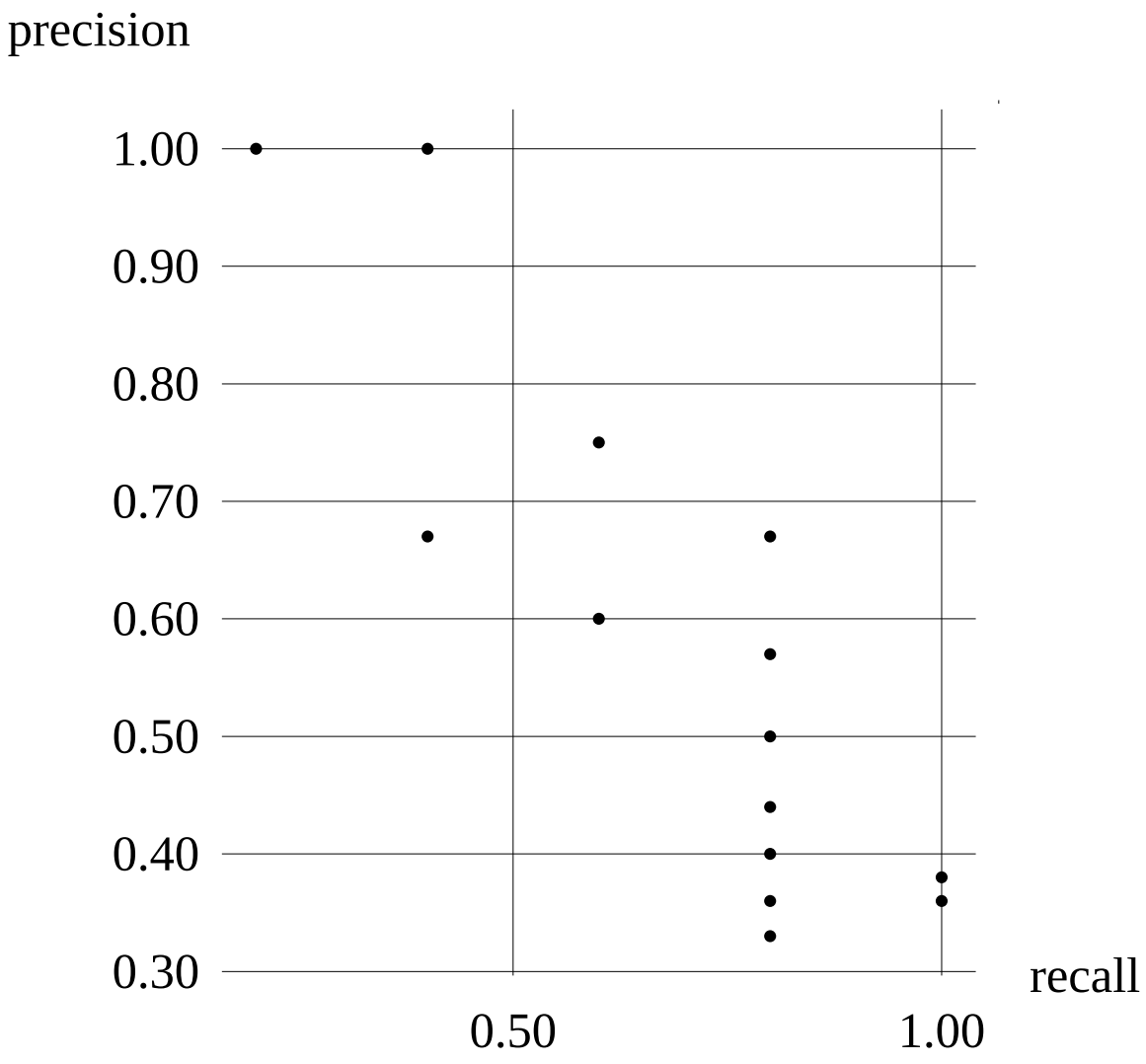
Lineare Ordnungen

Annahme:
Lineare Ordnung auf der Antwortmenge bzw. der Datenbank

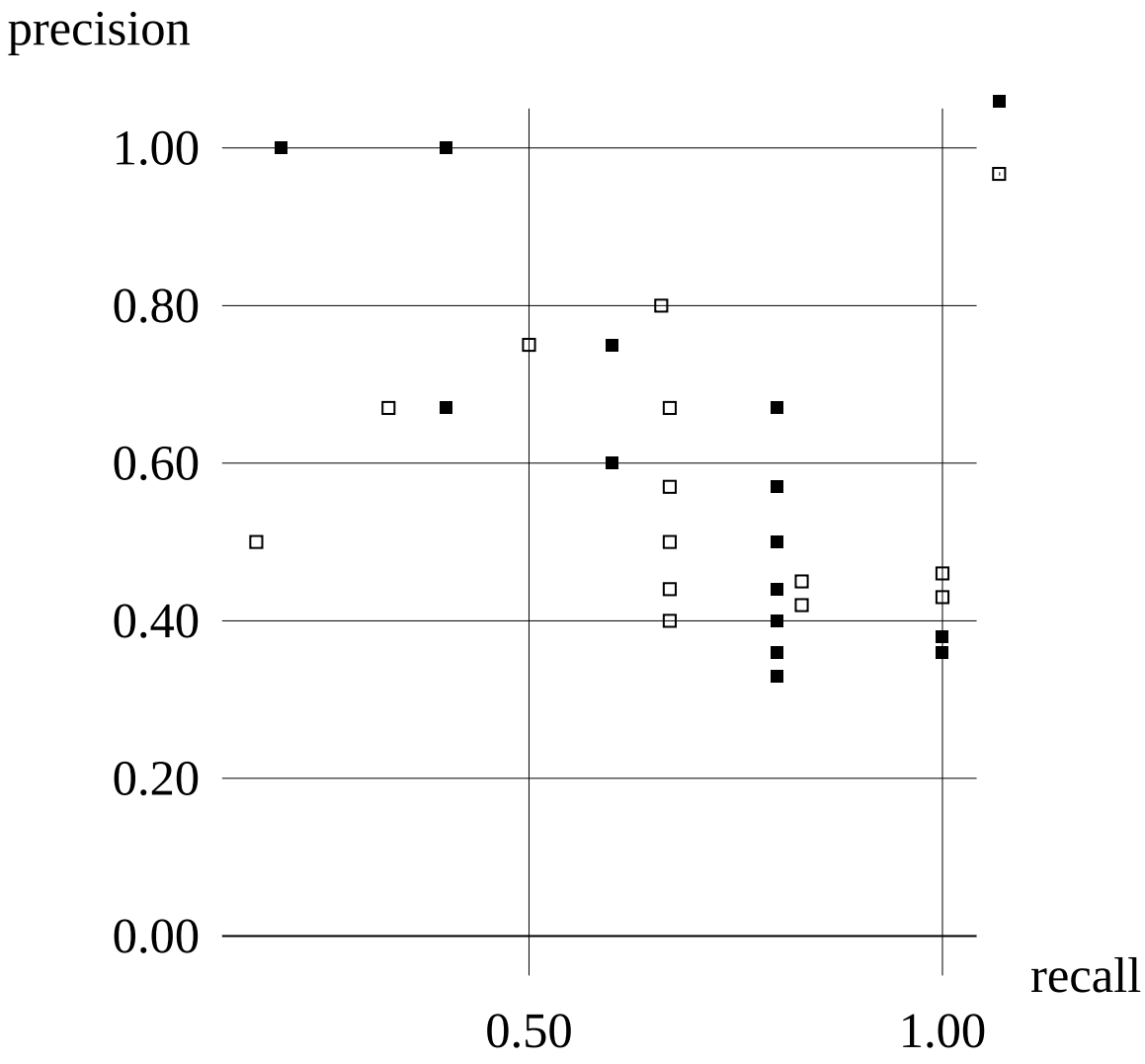
Recall und Precision (r, p) können für eine Anfrage nach jedem gefundenen Dokument bestimmt werden.

n	Dok-Nr.	x=rel.	Recall	Precision
1	588	x	0.2	1.00
2	589	x	0.4	1.00
3	576		0.4	0.67
4	590	x	0.6	0.75
5	986		0.6	0.60
6	592	x	0.8	0.67
7	984		0.8	0.57
8	988		0.8	0.50
9	578		0.8	0.44
10	985		0.8	0.40
11	103		0.8	0.36
12	591		0.8	0.33
13	772	x	1.0	0.38
14	990		1.0	0.36

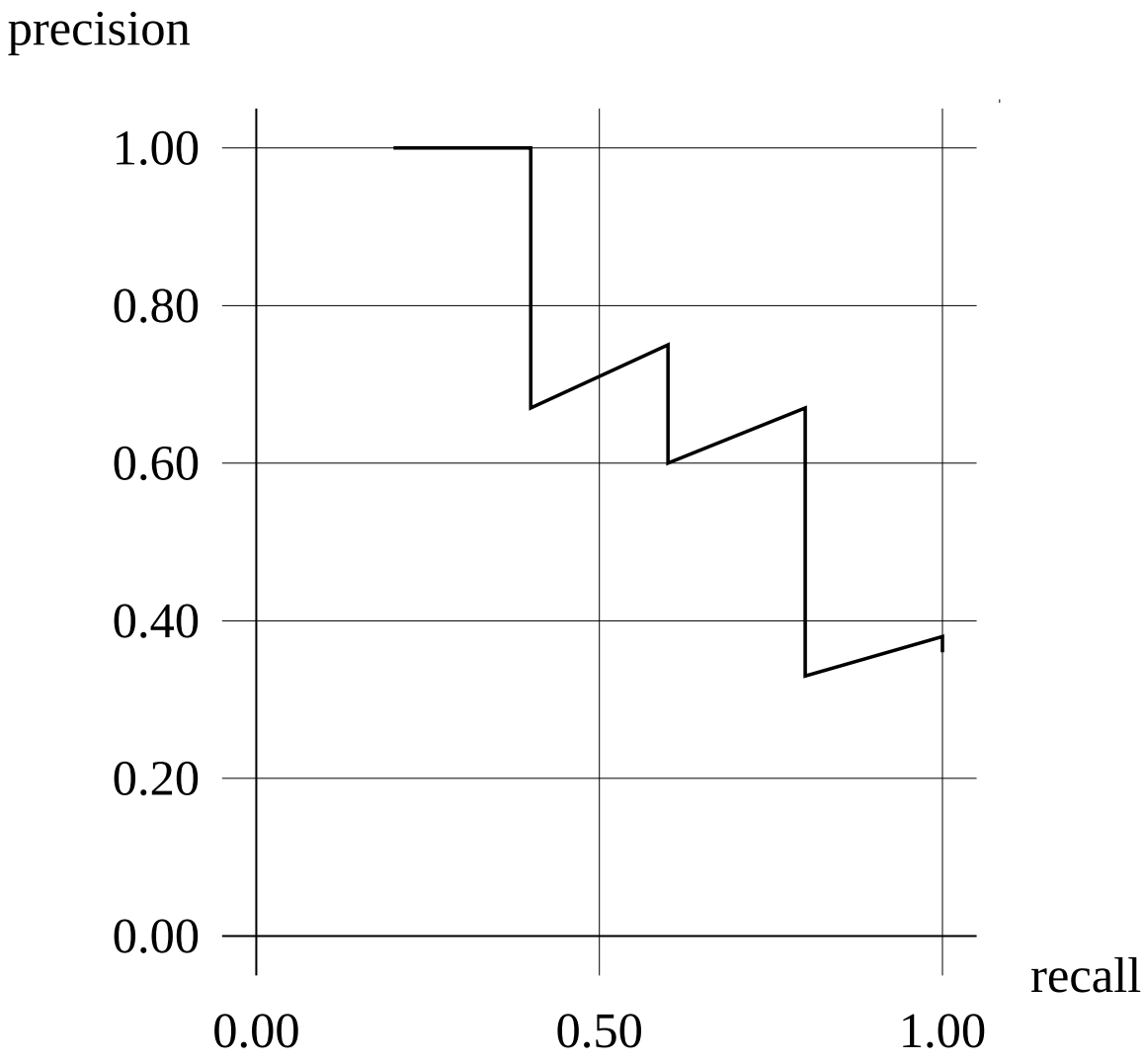
(r, p)-Punkte zum Beispielergebnis

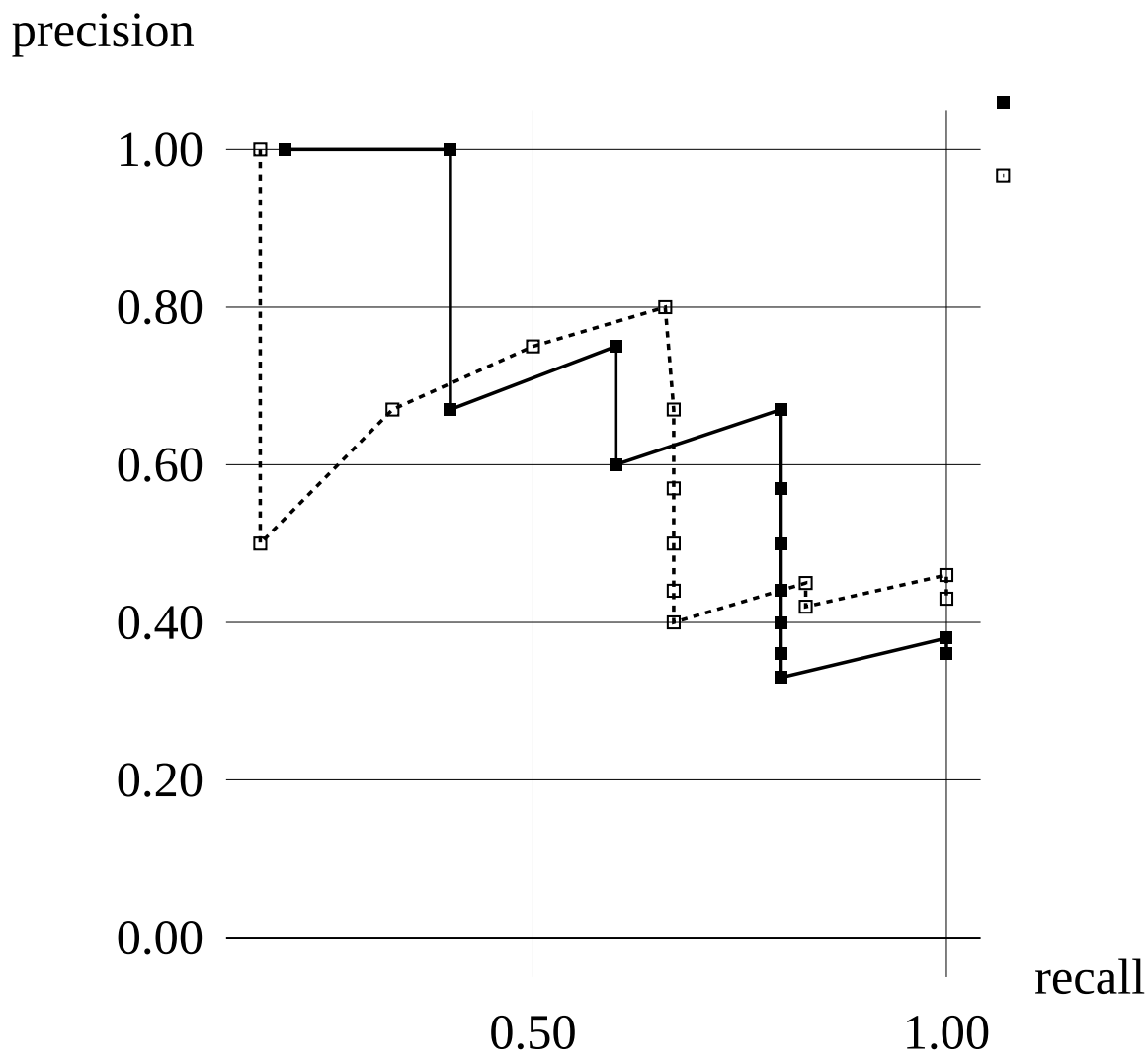


Vergleich zweier Ergebnisse



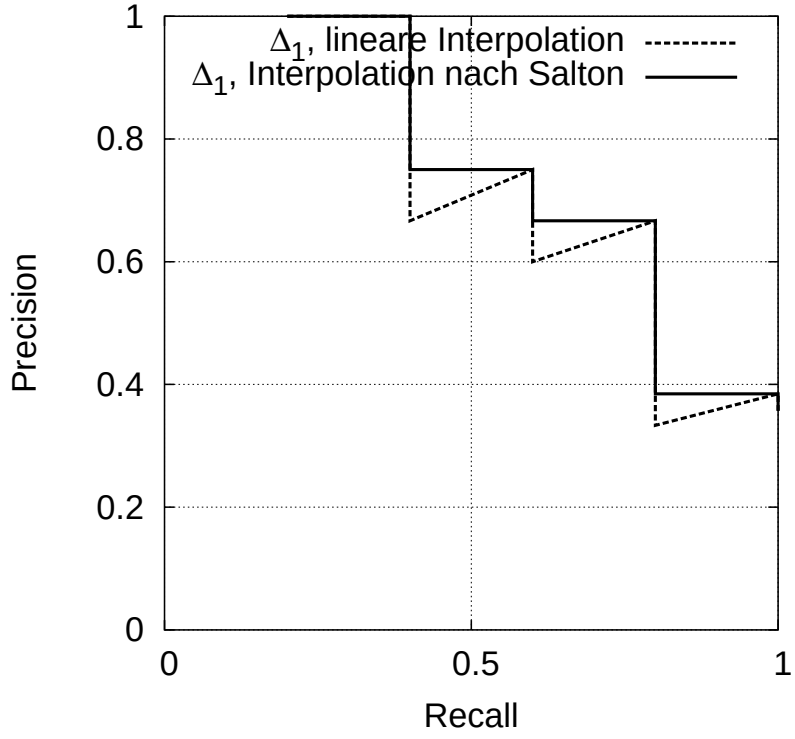
Geradenstücke zwischen den (r, p)-Punkten




precision
1.00
0.80
0.60
0.40
0.20
0.00
0.50
1.00
recall

Interpolation nach Salton

Annahme: Benutzer stoppt nur nach relevantem Dokument



Schwache Ordnungen

mehrere Dokumente mit demselben Retrievalgewicht/im selben Rang

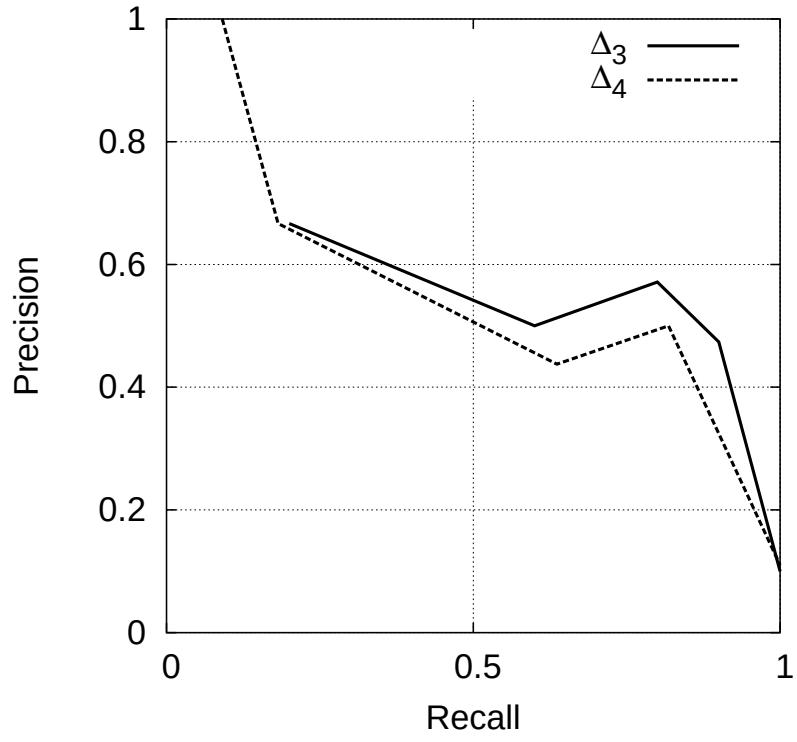
$$\Delta_3 = (++-|++++--------|++|+----|+-(80))$$
$$\Delta_4 = (+|+-|+++++-----------|++|++-(80))$$

Rangabhängige Recall- und Precision-Werte zu Δ_3 und Δ_4:

	Δ_3		Δ_4	
Rang	Recall	Precision	Recall	Precision
1	0.20	0.67	0.09	1.00
2	0.60	0.50	0.18	0.67
3	0.80	0.57	0.63	0.44
4	0.90	0.47	0.81	0.50
5	1.00	0.10	1.00	0.11

Recall-Precision-Graphen mit linearer Interpolation

zu Δ_3 und Δ_4



sinnvolle Interpolation?

Interpretation der Zwischenpunkte?

PRECALL

1. Berechne die Precision an jedem einfachen Recall-Punkt: $1/n \ldots (n-1)/n$ (bei n relevanten Dokumenten)
2. Erzwinge Monotonie ($p((h-1)/n) \leq p(h/n)$) durch Erhöhen von Precision-Werten.

Sei l_f der letzte berücksichtigte Rang. Die *ceiling* Interpolation ist dann definiert durch:

- j: Anzahl irrel. Dok. in Rängen $1 \ldots l_{f-1}$
- s: Anzahl rel. Dok., die aus Rang l_f gezogen werden
- r: Anzahl rel. Dok., in Rang l_f
- i: Anzahl irrel. Dok., in Rang l_f

$$PRECALL_{ceiling}(x) := \frac{\lceil x \cdot n \rceil}{\lceil x \cdot n \rceil + j + \frac{s}{r} i}$$

$$PRECALL_{ceiling}(x) := \frac{\lceil x \cdot n \rceil}{\lceil x \cdot n \rceil + j + \frac{s}{r} i}$$

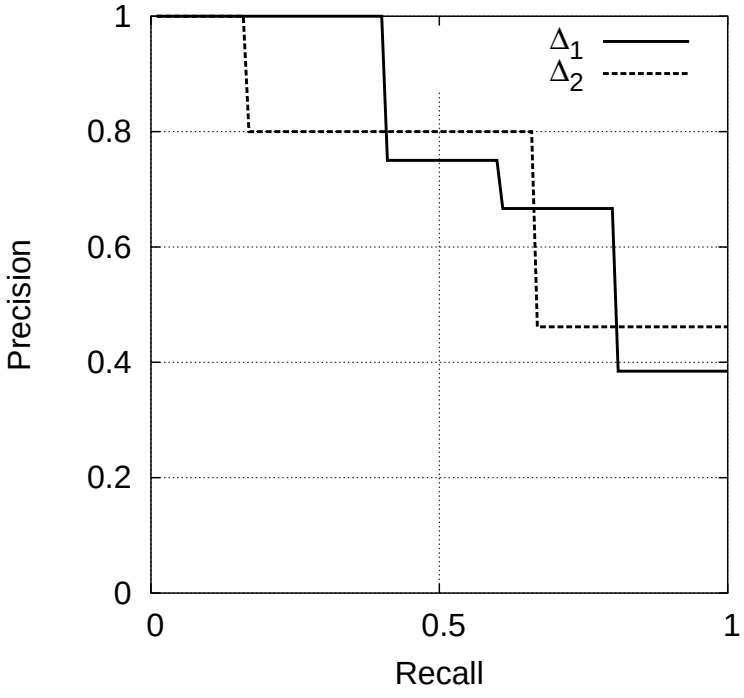
Beispiel

$$\Delta = (+ - -| + + + - - - - - - -)$$

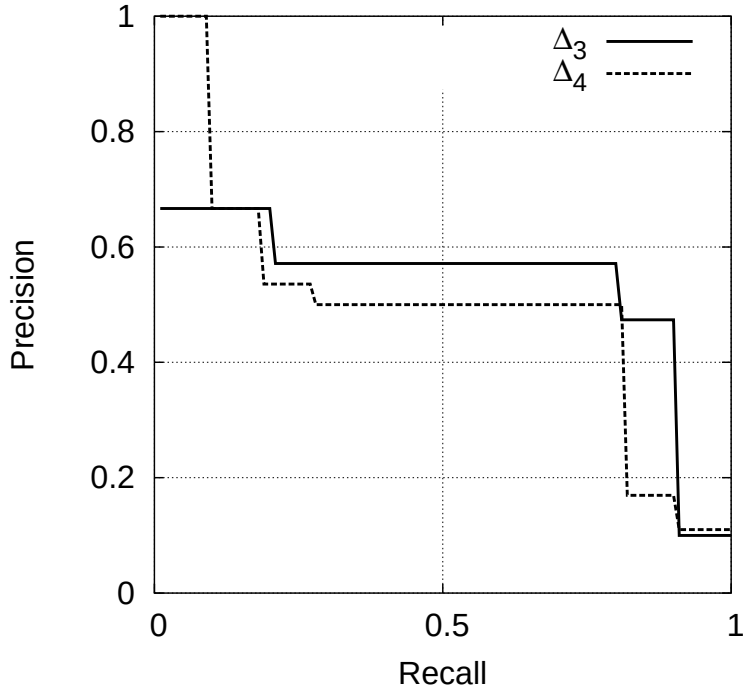
$$PRECALL(1/4) = \frac{1}{1 + 0 + \frac{1}{1} \cdot 2} = 1/3$$

$PRECALL_{ceiling}$

lineare Ordnungen



schwache Ordnungen



Einschub: Erwartete Suchlänge (esl)

Annahme:
Der Benutzer zieht aus einer Menge von r relevanten und i nicht relevanten Dokumenten solange Dokumente, bis er s relevante Dokumente hat.

esl_{NR} sei die zu erwartende Anzahl von nicht relevanten Dokumenten, die er bis dahin zieht.

$$esl_{NR} := \sum_{v=0}^{i} vP(v)$$

r # relevante Dokumente im aktuellen Rang

i # irrelevante Dokumente im aktuellen Rang

s gesuchte # rel. Dok. (im akt. Rang)

$X_{r,i,s}$ Gesamtzahl der gezogenen Dokumente

$E(X_{r,i,s}) =$ Erwartungswert für die aus dem aktuellen Rang zu ziehenden Dokumente, um s relevante zu bekommen

$$E(X_{r,i,s}) = \sum_v P(s-1 \text{ relevante aus } v-1) \cdot P(v.\ \text{Dok.} = s.\ \text{relevantes}) \cdot v$$

$$\begin{aligned} E(X_{r,i,s}) &:= \sum_{v=s}^{i+s} \frac{\binom{i}{v-s}\binom{r}{s-1}}{\binom{r+i}{v-1}} \frac{r-s+1}{r+i-v+1} v \\ &= \sum_{v=s}^{i+s} \frac{\binom{i}{v-s}\binom{r}{s}}{\binom{r+i}{v}} s \\ &= \sum_{v=s+1}^{i+s+1} \frac{\binom{i}{v-s-1}\binom{r}{s}}{\binom{r+i}{v-1}} s \end{aligned}$$

$$
\begin{aligned}
\frac{r+1}{(r+i+1)s} E(X_{r,i,s}) = & \\
& = \sum_{v=s+1}^{i+s+1} \frac{\binom{i}{v-s-1}\binom{r+1}{s}}{\binom{r+i+1}{v-1}} \frac{r-s+1}{r+i+1-v+1} \\
& = \sum_{v=s+1}^{i+s+1} P(X_{r+1,i,s+1} = v) \\
E(X_{r,i,s}) & = \frac{r+i+1}{r+1} s \\
esl_{NR} & = \frac{i}{r+1} s
\end{aligned}
$$

Abbruchkriterium: Anzahl der relevanten Dokumente

Bei schwach geordneten Antwortmengen sind zwei Benutzerstandpunkte denkbar. Der Benutzer zieht solange Dokumente aus dem höchsten noch nicht vollständig untersuchten Rang, bis er genug

- Dokumente (ND)
- relevante Dokumente (NR)

gesehen hat.

Wir wollen uns zunächst mit dem zweiten (plausibleren?) Vorgehen beschäftigen.

Da die Reihenfolge der Dokumente in einem Rang zufällig ist, müssen wir zu einer probabilistischen Definition von Precision übergehen.

Probability of Relevance (PRR)

Eine mögliche Definition von Precision ist die Wahrscheinlichkeit $P(rel|retr)$, daß ein untersuchtes Dokument relevant ist.

$$
\begin{aligned}
NR &: \text{Anzahl gewünschte rel. Dok.} \\
PRR &:= \frac{NR}{NR + esl_{NR}} \\
&= \frac{NR}{NR + j + \frac{s}{r+1} i}
\end{aligned}
$$

Wir interpolieren *intuitiv*, indem wir für NR reelle Zahlen zulassen.

$$
\begin{aligned}
PRR(x) &:= \frac{x \cdot n}{x \cdot n + esl_{x \cdot n}} \\
&= \frac{x \cdot n}{x \cdot n + j + \frac{s}{r+1} i}
\end{aligned}
$$

$$PRR(x) := \frac{x \cdot n}{x \cdot n + \frac{s}{r+1} i}$$

Beispiel

$$\triangle = (+ - - | + + + - - - - - - -)$$

$$PRR(1/4) = \frac{1}{1 + 0 + 1/(1+1) \cdot 2} = 1/2$$

Expected Precision

andere Möglichkeit, die Definition von Precision zu erweitern:
erwartete Precision: Erwartungswert der Precision.

$$EP_{NR} := \sum_{v} P(v)p(v) \quad v\text{: Anordnung}$$

Beispiel

$$\triangle = (+--|+++---------)$$

$$\begin{aligned} EP &= 1 \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{3} \cdot \frac{1}{3} \\ &= \frac{11}{18} \approx 0.611 \end{aligned}$$

Abbruchkriterium: Anzahl Dokumente

Erwartete Precision

PRR und fallen EP zusammen:

t_r : Anzahl relevante Dok. in Rängen $1 \ldots l_{f-1}$

k : Anzahl Dok., die aus Rang l_f gezogen werden

$$EP_{ND} = PRR_{ND} := \frac{1}{ND}\left(t_r + \frac{k}{r+i} \cdot r\right)$$

Erwarteter Recall

- Definiere *erwarteten Recall* ER_{ND} analog zu EP_{ND}:

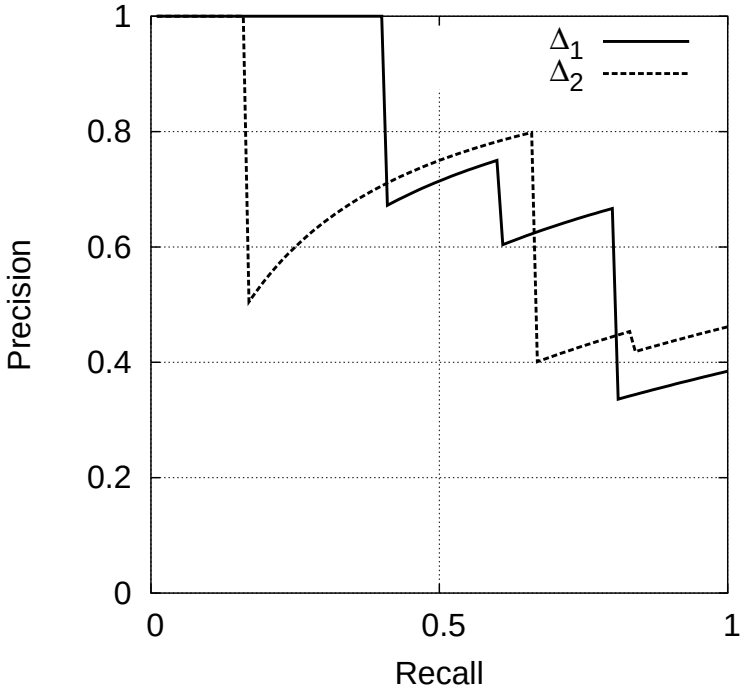
$$ER_{ND} := \frac{1}{n}\left(t_r + \frac{k}{r+i} \cdot r\right)$$

- Konstruiere den Graphen der Punkte $\{(ER(ND), EP(ND)\}$ Graph stimmt mit dem PRECALL-Graphen mit *intuitiver* Interpolation überein.

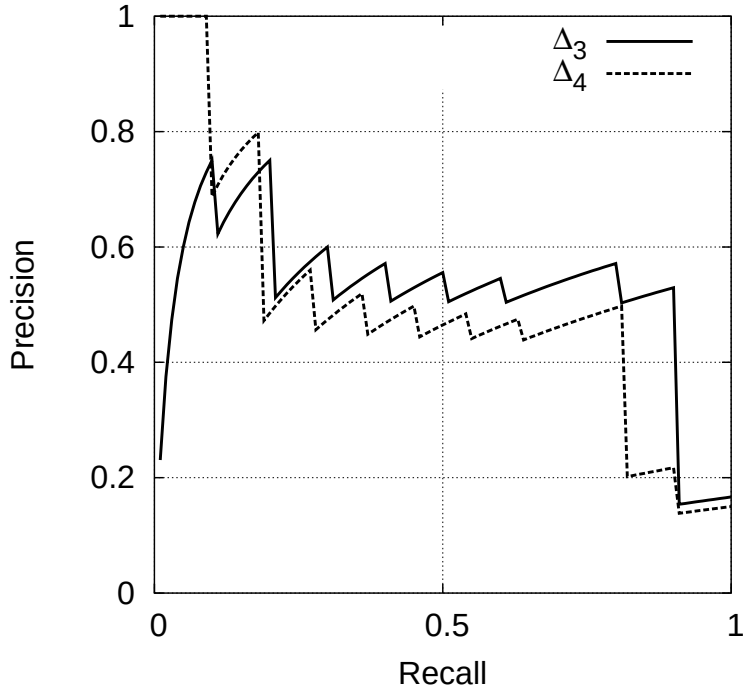
$$PRECALL_{intuitive}(x) := \frac{x \cdot n}{x \cdot n + j + (s/r) \cdot i}$$

Probability of Relevance (NR)

lineare Ordnungen

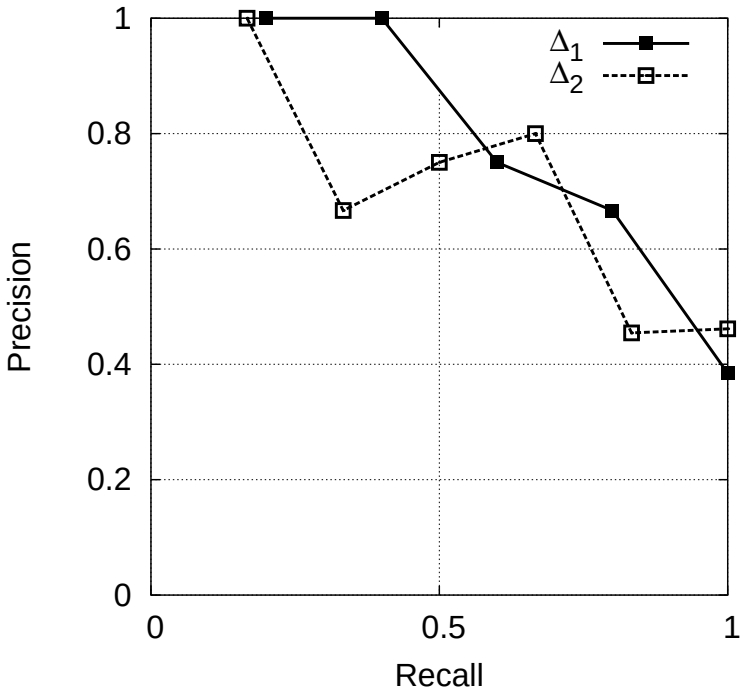


schwache Ordnungen

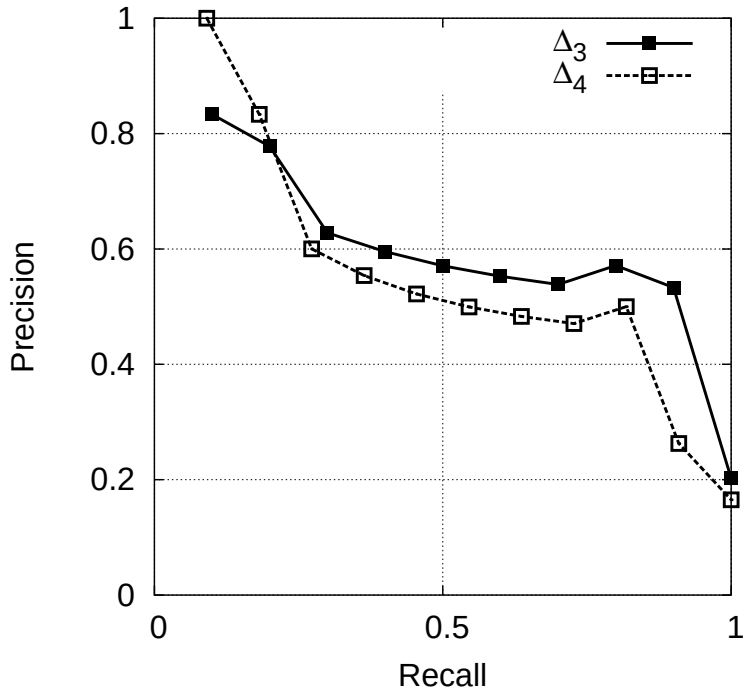


Expected Precision (NR)

lineare Ordnungen



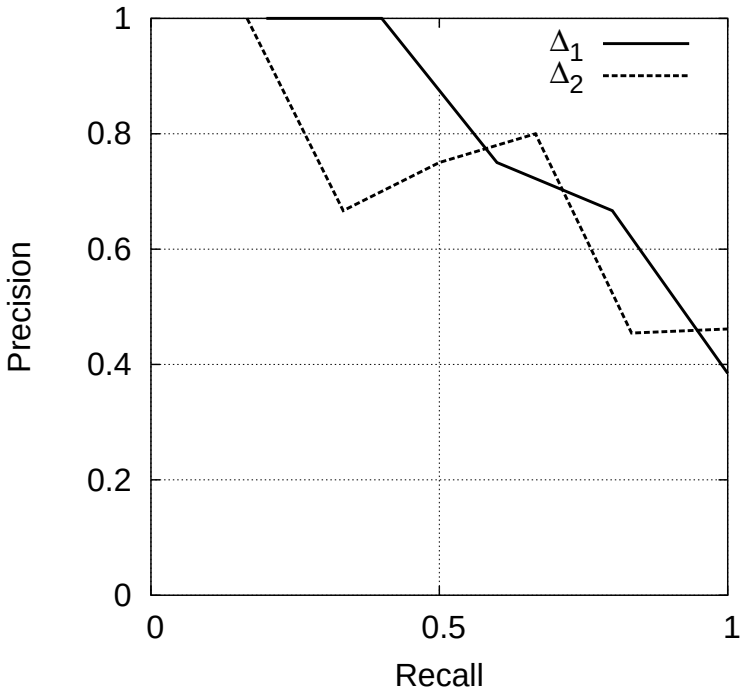
schwache Ordnungen



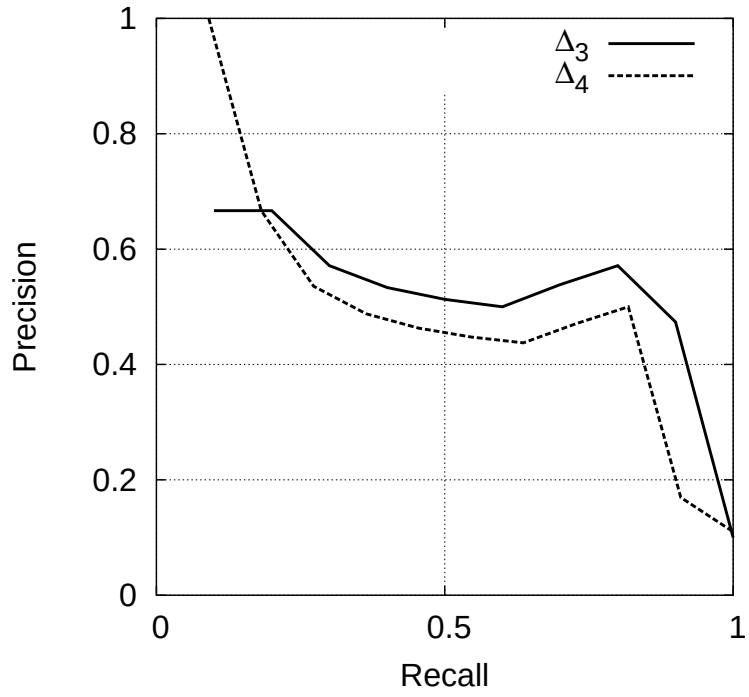
Benutzerstandpunkt *ND*

- es gilt: $EP_{ND} = PRR_{ND}$
- lineare Ordnung: kein Unterschied zu EP_{NR}

lineare Ordnungen



schwache Ordnungen



Evaluierungsinitiativen: TREC, CLEF, NTCIR und INEX

Standardumgebung für die Evaluierung von IR-Methoden:

- Dokumentkollektionen im Umfang praktischer Anwendungen (z.B. Newsticker-, Zeitungs-, Magazinartikel, Web-Kollektionen)
- vordefinierte Anfragen (*Topics*)
- verschiedene Aufgaben (*Tracks*)
- Relevanzbeurteilungen (Pooling-Methode)

TREC: Text REtrieval Conference

- Defacto-Standard
- seit 1992 (jährlich)
- Tracks: Adhoc, Routing, CLIR, Question Answering, Video, Speech, ...

CLEF: Cross-Language Evaluation Forum

- seit 2000 (jährlich)
- Schwerpunkt: multilinguales Retrieval mit europäischen Sprachen (Deutsch, Englisch, Französisch, Italienisch, Spanisch); Anfragen in vielen weiteren Sprachen
- Tracks: multilinguales, bilinguales, monolinguales (nicht Englisch) IR, ...

NTCIR: NACSIS Test Collection Project

- seit 1999 (jährlich)
- Schwerpunkt: multilinguales Retrieval mit asiatischen Sprachen (Chinesisch, Koreanisch, Japanisch)
- Tracks: multilinguales IR, Web-Retrieval, Patent-Retrieval, Question Answering, ...

INEX: Initiative for the Evaluation of XML Retrieval

- seit 2002 (jährlich)
- Kollektion: Vollständige Artikel aus Informatikzeitschriften / Wikipedia
- Schwerpunkt: Retrieval von (Teilen von) XML-Dokumenten unter Berücksichtigung der Dokumentstruktur
- Tracks: content-only queries, content-and-structure queries

TREC-Evaluierungsmaße

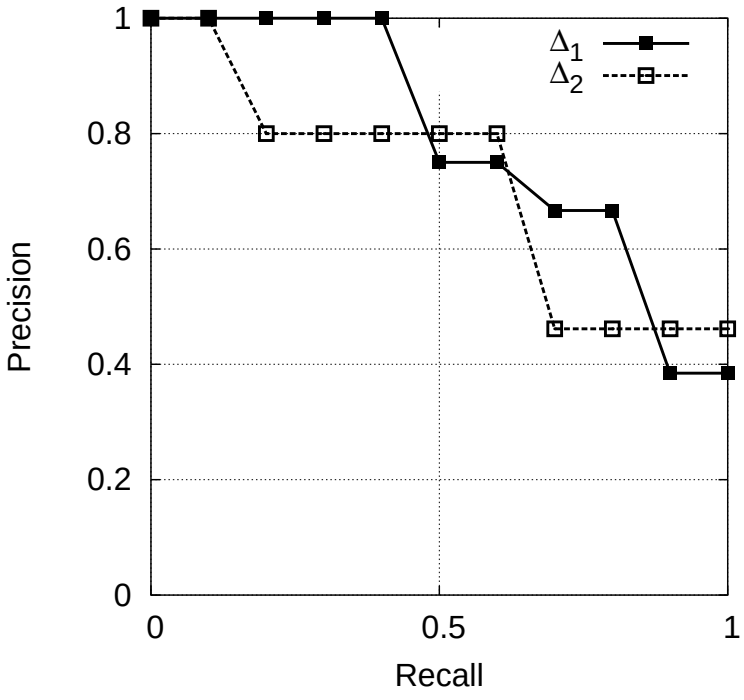
1. bestimme Precision für jeden Rang
2. Interpolation nach Salton
3. bestimme Precision für 11 Recall-Punkte $\{ 0, 0.1, 0.2, \ldots, 1 \}$
4. verbinde resultierende Punkte

Schwächen:

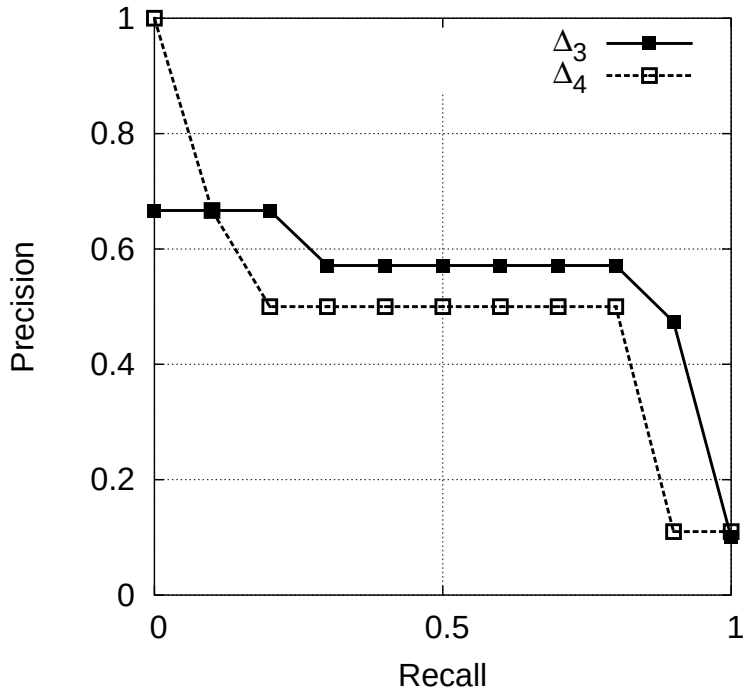
- bei Recall-Punkt 0 ist Precision eigentlich undefiniert
- schwache Ordnungen werden durch zufällige Anordnungen der Dokumente in lineare Ordnungen überführt

TREC-Evaluierungsmaße (2)

lineare Ordnungen



schwache Ordnungen



TREC-Evaluierungsmaße (3)

Benutzerorientierte Maße:

- Prec@5, Prec@10, Prec@30, Prec@100
  (jeweils als Makro-Mittelwert über alle Fragen)
- Mean reciprocal rank:
  Annahme, Benutzer ist nur an einem relevanten Dokument interessiert

  1. Bestimme Rang k des ersten relevanten Dokumentes
  2. Bilde Kehrwert $1/k$
  3. Mittele über alle Fragen

Systemorientiertes Maß: Mean average precision (MAP)

Mittelwert der Precision nach jedem relevanten Dokument einer Rangliste
(anschließend arithmetisches Mittel über alle Fragen)

Evaluierung von interaktivem Retrieval

Schwächen der gängigen Evaluierungen:

- nur Batch-Retrieval: Qualität bzgl. einmaliger Anfrage
- Relevance Feedback erlaubt auch nur Relevanzbeurteilung einiger Dokumente
- heutige IR-Systeme bieten oft reichhaltige Funktionalität: Reformulierung, Highlighting, Clustering, Browsing von Dokumenten / Termlisten, ...
- Ergebnisse aus TREC interactive track:
  Benutzer können bestimmte Schwächen der Systeme (z.B. schlechtes Ranking) leicht kompensieren

  $\rightarrow$ Ergebnisse aus Batch-Evaluierungen sind nicht auf interaktives Retrieval übertragbar

  $\rightarrow$ Notwendigkeit für Evaluierung von interaktivem Retrieval

Evaluierungsmethoden für interaktives Retrieval

- 'think aloud'-Protokolle
- Beobachtungsdaten, Log-Analyse
- Interviews
- Fehleranalyse
- Zeitbedarf zur Problembearbeitung
- Kosten-Nutzen-Analyse
- Fragebögen (Prä-, Post-)
- Simulationen