

# Evaluation von IR-Systemen

# Warum evaluieren?

- Evaluation ist der Schlüssel, um
  - **effektive** (Finden wir die richtigen Dokumente?) und
  - **effiziente** (Machen wir es schnell / mit hohem Durchsatz?)Suchmaschinen zu konstruieren.
- Messungen meist in kontrollierten **Laborexperimenten**
  - Auch User-Tests können vorgenommen werden (**User Satisfaction** / **Productivity**)
- Effektivität, Effizienz und **Kosten** hängen zusammen
  - z. B. wenn wir einen bestimmten Level von **Effektivität und Effizienz** erreichen wollen, wird dies die **Kosten der Systemkonfiguration** bestimmen
  - Effizienz- und Kostenziele können die Effektivität beeinflussen

# Evaluations-Korpora

Testkollektionen, die aus **Dokumenten**, **Anfragen** und **Relevanzbewertungen** bestehen, z. B.

- **CACM**: Titles and abstracts from the **Communications of the ACM** from 1958-1979. Queries and relevance judgments generated by **computer scientists**.
- **AP**: Associated Press **newswire documents** from 1988-1990 (from TREC disks 1-3). Queries are the title fields from TREC topics 51-150. Topics and relevance judgments generated by **government information analysts**.
- **GOV2**: **Web pages** crawled from websites in the .govdomain during early 2004. Queries are the title fields from TREC topics 701-850. Topics and relevance judgments generated by **government analysts**.

# Testkollektionen

- Eigenschaften der Dokumente

Kollektion	Anzahl der Dokumente	Größe	Durchschnittliche Anzahl Wörter/Dokument
CACM	3.204	2,2 MB	64
AP	242.198	0,7 GB	474
GOV2	25.205.179	426 GB	1.073

- Eigenschaften der Anfragen

Kollektion	Anzahl der Anfragen	Durchschnittliche Anzahl Wörter/Anfrage	Durchschnittliche Anzahl relevante Dokumente/Anfrage
CACM	64	13,0	16
AP	100	4,3	220
GOV2	150	3,4	180

# Beispiel zu TREC-Topic

Anfrage, die von Suchmaschine  
verarbeitet wird  
(potentiell mehrdeutig und schlecht  
formuliert, wie im richtigen Leben)

**<top>**

**<num>** Number: 794

**<title>** pet therapy

**<desc>** Description:

How are **pets or animals used in therapy for humans** and what are the benefits?

Hilfe für manuelle  
Relevanzbewertung der  
Ergebnisse

**<narr>** Narrative:

Relevant documents must include details of how pet- or animal-assisted therapy is or has been used. Relevant details include information about pet therapy programs, descriptions of the circumstances in which pet therapy is used, the benefits of this type of therapy, the degree of success of this therapy, and any laws or regulations governing it.

**</top>**

# Relevanzbewertungen

- Das Beschaffen von Relevanzbewertungen ist **ein teurer, zeitraubender Prozess**
  - **Wer** macht es?
  - Was sind die **Instruktionen**?
  - Wie hoch ist die **Übereinstimmung**?
- TREC Bewertungen
  - hängen von der evaluierten Aufgabe (Task) ab (z. B. **hoher Recall oder hohe Precision**)
  - i.d.R. **thematische Relevanz** betrachtet  
⇒ alle Dokumente, die thematisch relevant sind, werden positiv bewertet, auch wenn sie keine „neuen“ Infos enthalten
  - meist **binär** (aber auch „gestufte“ Relevanz möglich)
  - hohe **Übereinstimmung** zwischen verschiedenen Bewertern aufgrund des Narrative <narr>

# Poolbildung

**Erschöpfende Bewertung** aller Dokumente in einer großen Kollektion ist **nicht praktikabel**

- Poolbildungstechniken bei TREC verwendet
- **Beste  $k$  Ergebnisse** (für TREC  $50 \leq k \leq 200$ ) der Rankings, die **von verschiedenen Suchmaschinen** (oder Retrieval-Algorithmen) ermittelt wurden, werden zu einem Pool zusammengefügt
- **Duplikate** werden **entfernt**
- Dokumente werden den Relevanzbewertern in einer zufälligen Reihenfolge präsentiert

Erzeugt eine große Anzahl von Relevanzbewertungen für jede Anfrage, jedoch **immer noch unvollständig**

- **Problem:** Wenn eine Suchmaschine, die nicht im Pool berücksichtigt ist, ganz andere rel. Dokumente findet, wird sie schlecht abschneiden

# Anfragelogs

Wesentlich mehr Information  
als bei manuellen Relevanzurteilen

- Werden für das **Tunen und Evaluieren** von Suchmaschinen eingesetzt
  - Auch für andere Techniken, wie Suchbegriffsvorschläge
- Typische **Inhalte** der Anfragelogs
  - **Benutzeridentifikator** oder Identifikator für eine Benutzersitzung
  - **Anfrageterme** – genauso gespeichert, wie sie vom Nutzer eingegeben wurden
  - Liste der **Ergebnis-URLs**, ihrer Ränge auf der Ergebnisliste und ob sie **angeklickt** wurden
  - **Zeitstempel** – halten die Zeit von Benutzerereignissen wie Abschicken der Anfrage oder Klicks fest

AnonID	Query	QueryTime	ItemRank	ClickURL
8760	jojo the singer	2006-03-26 16:02:04	5	<a href="http://www.jojofan.com">http://www.jojofan.com</a>
8760	jennifer lopez	2006-03-26 16:05:29	4	<a href="http://www.allstarz.org">http://www.allstarz.org</a>
8760	jennifer lopez	2006-03-26 16:05:29	10	<a href="http://www.starpulse.com">http://www.starpulse.com</a>
8760	nicole richie	2006-03-26 17:28:58		
8760	free porn	2006-03-28 16:43:16		



# Anfrage logs

- Daten in Anfrage logs sind grundsätzlich privat und dürfen nicht veröffentlicht werden
- Anonymisierung der IP-Adresse alleine nicht ausreichend

## AOL releases search data on 500,000 users (updated)

The complete three-month search history of 500,000 AOL users was released by ...

NATE ANDERSON - 8/7/2006, 5:39 PM

AUG 7TH, 2006

## AOL Research Publishes 650,000 User Queries

Interested in users' online queries? Ever wanted to cluster similar users or mine their data? Wait no more, AOL's research team has published a huge data collection of 20,000,000 search queries from 650,000 users sampled over three months for the public to see, dig around and analyze.

## *A Face Is Exposed for AOL Searcher No. 4417749*

No. 4417749 conducted hundreds of searches over a three-month period on topics ranging from “numb fingers” to “60 single men” to “dog that urinates on everything.”

And search by search, click by click, the identity of AOL user No. 4417749 became easier to discern. There are queries for “landscapers in Lilburn, Ga,” several people with the last name Arnold and “homes sold in shadow lake subdivision gwinnett county georgia.”

It did not take much investigating to follow that data trail to Thelma Arnold, a 62-year-old widow who lives in Lilburn, Ga., frequently researches her friends' medical ailments and loves her three dogs. “Those are my searches,” she said, after a reporter read part of the list to her.

CNET > Tech Industry > AOL apologizes for release of user search data

## AOL apologizes for release of user search data

Search log information originally intended for use on new research site; company calls data posting a mistake.



Thelma Arnold's identity was betrayed by AOL records of her Web searches, like ones for her dog, ...

CNET > Tech Culture > AOL sued over Web search data release

## AOL sued over Web search data release

 AOL executive quits after posting of search data

By Tom Zeller Jr. The New York Times

Published: August 22, 2006

# Anfragelogs

- Klicks sind **keine Relevanzbewertungen**
  - obwohl die beiden korrelieren
  - **Verfälscht** durch Faktoren wie den Rang in der Ergebnisliste
- Man kann Klickdaten verwenden, um **Präferenzen zwischen Paaren von Dokumenten** vorherzusagen
  - angemessen für Aufgaben mit mehreren Relevanzleveln
  - ausgerichtet auf **Benutzerrelevanz**
  - verschiedene „Policies“ werden eingesetzt, um Präferenzen zu erzeugen
- **Andere Informationen** aus Toolbar, Cookie, ...
  - Wie lange wird eine Seite betrachtet?
  - Wird eine Seite gedruckt?
  - ...

# Beispiel für eine „Click Policy“

„Überspringe“ vorherige und folgende Ergebnisse

- **Klickdaten**

$d1$

$d2$

$d3(clicked)$

$d4$

Bereits sehr unklar, ob das wirklich gilt

- Erzeugte **Präferenzen**

$d3 > d2$

$d3 > d1$

$d3 > d4$

# Anfragelogs

- Klickdaten können auch **aggregiert** werden, um „**Noise**“ (Klicks, die nicht zu relevanten Dokumenten führen) zu entfernen
- Informationen über Klickverteilung
  - Können verwendet werden, um **Klicks** zu identifizieren, die **häufiger** auftreten als **erwartet**
  - hochgradig **korreliert mit Relevanz**
  - z. B. Verwendung der „Klickabweichung“, um Klicks für Präferenz erzeugungs-Policies zu filtern

# Filtern von „signifikanten“ Clicks

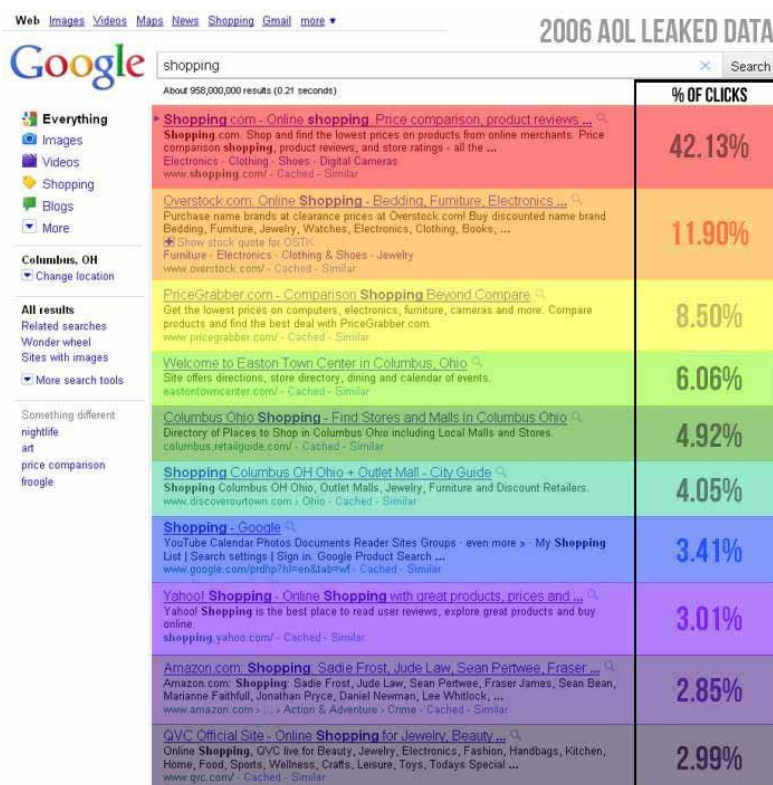
Klickabweichung (click deviation)  $CD(d,p)$  für ein Ergebnisdokument  $d$  auf Position  $p$ :

$$CD(d,p) = O(d,p) - E(p)$$

- $O(d,p)$ :
  - **Beobachtete (observed) Klickhäufigkeit** für das Dokument  $d$  an der Rangposition  $p$  über **alle Instanzen einer gegebenen Anfrage**
- $E(p)$ :
  - **Erwartete (expected) durchschnittliche Klickhäufigkeit** an einer Rangposition  $p$  über alle Anfragen

2006 AOL LEAKED DATA

$E(p)$



	% OF CLICKS
Shopping.com - Online shopping. Price comparison, product reviews ... Shopping.com. Shop and find the lowest prices on products from online merchants. Price comparison shopping, product reviews, and store ratings - all the ... Electronics - Clothing - Shoes - Digital Cameras www.shopping.com/ - Cached - Similar	42.13%
Overstock.com - Online Shopping - Bedding, Furniture, Electronics ... Purchase name brands at clearance prices at Overstock.com! Buy discounted name brand Bedding, Furniture, Jewelry, Watches, Electronics, Clothing, Books, ... Show stock quote for OSTK Furniture - Electronics - Clothing & Shoes - Jewelry www.overstock.com/ - Cached - Similar	11.90%
PriceGrabber.com - Comparison Shopping Beyond Compare ... Get the lowest prices on computers, electronics, furniture, cameras and more. Compare products and find the best deal with PriceGrabber.com. www.pricegrabber.com/ - Cached - Similar	8.50%
Welcome to Easton Town Center in Columbus, Ohio ... Site offers directions, store directory, dining and calendar of events. eastontowncenter.com/ - Cached - Similar	6.06%
Columbus Ohio Shopping - Find Stores and Malls in Columbus Ohio ... Directory of Places to Shop in Columbus Ohio including Local Malls and Stores. columbus.retailguide.com/ - Cached - Similar	4.92%
Shopping Columbus OH Ohio - Outlet Mall - City Guide ... Shopping Columbus OH Ohio, Outlet Malls, Jewelry, Furniture and Discount Retailers. www.discoveroutstown.com/ Ohio - Cached - Similar	4.05%
Shopping - Google ... YouTube Calendar Photos Documents Reader Sites Groups - even more > My Shopping List   Search settings   Sign in: Google Product Search ... www.google.com/prdp?hl=en&tab=ef - Cached - Similar	3.41%
Yahoo! Shopping - Online Shopping with great products, prices and ... Yahoo! Shopping is the best place to read user reviews, explore great products and buy online. shopping.yahoo.com/ - Cached - Similar	3.01%
Amazon.com: Shopping: Sadie Frost, Jude Law, Sean Pertwee, Fraser ... Amazon.com: Shopping: Sadie Frost, Jude Law, Sean Pertwee, Fraser James, Sean Bean, Marianne Faithfull, Jonathan Pryce, Daniel Newman, Lee Whitlock, ... www.amazon.com > ... > Action & Adventure > Crime - Cached - Similar	2.85%
QVC Official Site - Online Shopping for Jewelry, Beauty ... Online Shopping, QVC live for Beauty, Jewelry, Electronics, Fashion, Handbags, Kitchen, Home, Food, Sports, Wellness, Crafts, Leisure, Toys, Today's Special ... www.qvc.com/ - Cached - Similar	2.99%

# Maße für die Effektivität

- Recall und Precision (bereits in Kapitel 3)
- Durchschnittsbildung und Interpolation
- Konzentration auf die **Top-Dokumente**
- Nutzung von **Präferenzen zwischen Dokumenten**  
(Halbordnungen)

# False Negatives und False Positives

	Relevant	Nicht relevant
Gefunden	True Positives	False Positives
Nicht gefunden	False Negatives	True Negatives

- **True Positives (tp)**  
Gefundene relevante Dokumente
- **False Positives (fp)**  
Gefundene irrelevante Dokumente
- **True Negatives (tn)**  
Nicht gefundene irrelevante Dokumente
- **False Negatives (fn)**  
Nicht gefundene relevante Dokumente

# Precision und Recall

Die **Präzision** oder **Precision P** gibt an, wie groß der Anteil der korrekten Treffer an der gesamten Menge der gefundenen Dokumente ist. Sie ist definiert durch

$$P := \frac{tp}{tp + fp} = \frac{\text{Anzahl relevanter gefundener Dokumente}}{\text{Anzahl gefundener Dokumente}}$$

Die **Ausbeute** oder der **Recall R** gibt an, wie groß der Anteil der korrekten Treffer an der Menge der relevanten Dokumente ist. Er ist definiert durch

$$R := \frac{tp}{tp + fn} = \frac{\text{Anzahl relevanter gefundener Dokumente}}{\text{Anzahl relevanter Dokumente}}$$



# Bedeutung von Precision und Recall

- Je nach Aufgabenstellung unterschiedlich wichtig
  - **Recall-Orientierung:**
    - Wenn es wichtig ist, in jedem Fall alle relevanten Dokumente zu finden
    - Beispiel: Patent-Recherche
  - **Precision-Orientierung:**
    - wenn die Wahrscheinlichkeit, dass ein positives Ergebnis auch korrekt ist, wichtig ist
    - Beispiel: Alert

Fast immer stehen die Ziele Recall und Precision im Konflikt!

# F-Maß

- **Harmonisches Mittel** von Recall und Precision

$$F := \frac{1}{\frac{1}{2} \left( \frac{1}{R} + \frac{1}{P} \right)} = \frac{2RP}{(R + P)}$$

- Harmonische Mittel heben die **Bedeutung kleiner Werte** hervor,
- während **arithmetische Mittel** mehr von Ausreißern, die gewöhnlich **groß** sind, beeinflusst werden.

- Allgemeinere Form

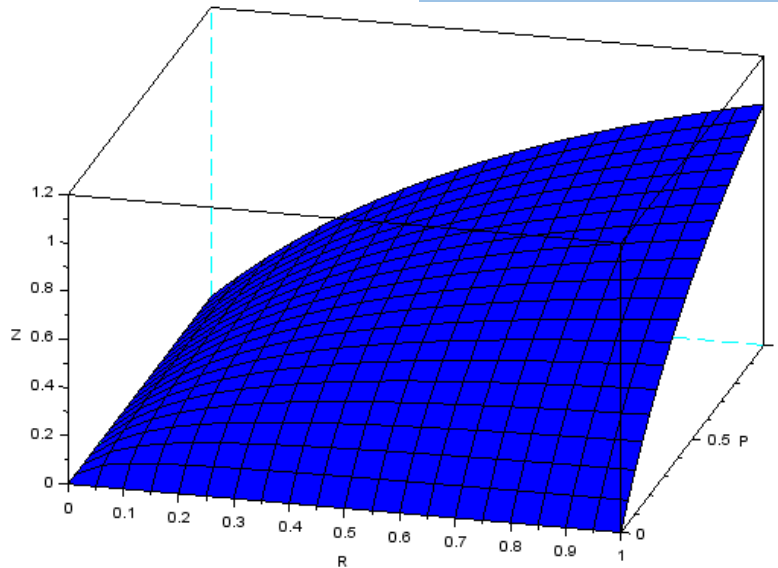
$$F_{\beta} := \frac{(\beta^2 + 1)RP}{R + \beta^2 P}$$

- $\beta$  ist ein Parameter, der über die **relative Bedeutung** von Recall und Precision entscheidet

# Mittelwerte

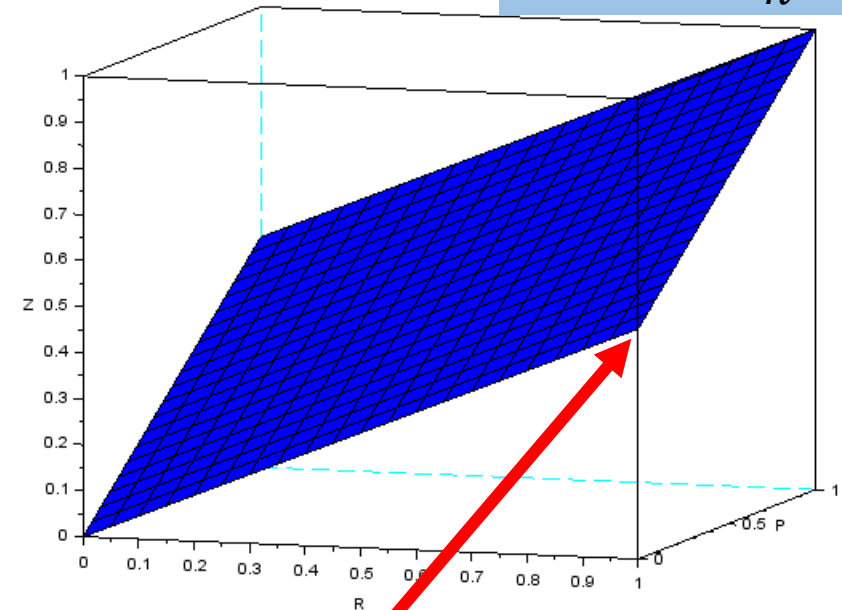
## Harmonisches Mittel

$$\bar{x}_{harm} = \frac{n}{\sum_{i=1}^n 1/x_i}$$



## Arithmetisches Mittel

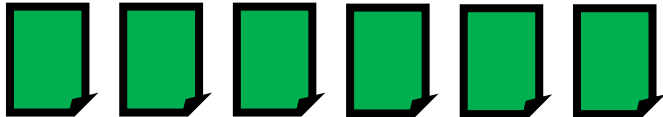
$$\bar{x}_{arith} = \frac{\sum_{i=1}^n x_i}{n}$$



Hier erhält man ein Mittel von ca. 0,5, wenn man die ganze Kollektion als Ergebnis liefert!

# Ranking Effektivität

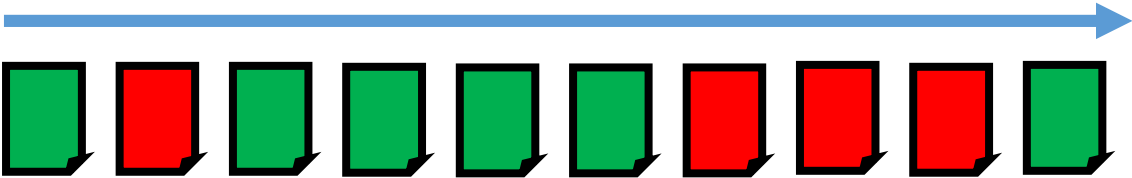
Problem:  
Recall und Precision eigentlich für Ergebnismengen, nicht für Rankings!



6 relevante Dokumente

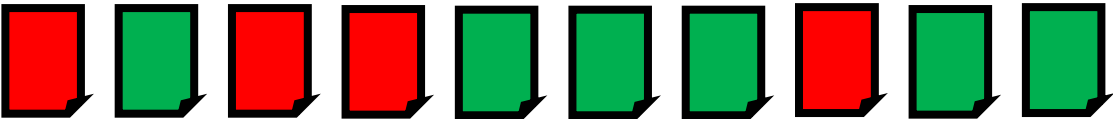
Ranking #1:

Werte nach jedem Rang



Recall	0,17	0,17	0,33	0,50	0,67	0,83	0,83	0,83	0,83	1,0
Precision	1,0	0,5	0,67	0,75	0,8	0,83	0,71	0,63	0,56	0,6

Ranking #2:



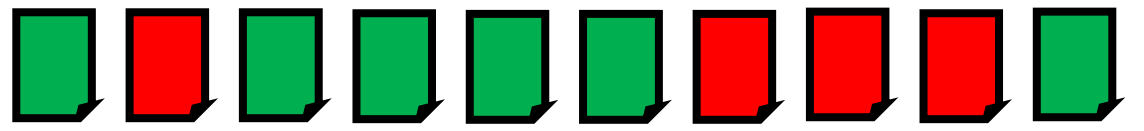
Recall	0,0	0,17	0,17	0,17	0,33	0,50	0,67	0,67	0,83	1,0
Precision	0,0	0,5	0,33	0,25	0,4	0,5	0,57	0,5	0,56	0,6

# Optionen zum Zusammenfassen eines Rankings

1. Berechnung von Recall und Precision an festgelegten Rang-Positionen
  - z. B. **Precision @ Rank 10**
  - Problem: dann wären Ranking #1 und #2 mit je 0,6 gleich gut
2. **Precision** wird an **Standard-Recall-Punkten** von 0,0 bis 1,0 im Abstand von 0,1 berechnet
  - **benötigt Interpolation**, da diese Recall-Werte i.d.R. nicht exakt auftreten
3. Bilden von **Durchschnittswerten über die Precision-Werte** der Rangpositionen, an denen ein relevantes Dokument abgerufen wurde

# Durchschnittliche Precision (AP)

Ranking #1:



Recall	0,17	0,17	0,33	0,50	0,67	0,83	0,83	0,83	0,83	1,0
Precision	1,0	0,5	0,67	0,75	0,8	0,83	0,71	0,63	0,56	0,6

Ranking #2:



Recall	0,0	0,17	0,17	0,17	0,33	0,50	0,67	0,67	0,83	1,0
Precision	0,0	0,5	0,33	0,25	0,4	0,5	0,57	0,5	0,56	0,6

Ranking #1  $(1,0 + 0,67 + 0,75 + 0,8 + 0,83 + 0,6) / 6 = 0,78$

Ranking #2  $(0,5 + 0,4 + 0,5 + 0,57 + 0,56 + 0,6) / 6 = 0,52$

Durchschnittsbildung für eine Anfrage über **mehrere Recall-Punkte**

# Benutzermodell für AP

## Wann ist eine Metrik sinnvoll?

Wenn sie (approximativ) misst, wie **zufrieden ein Benutzer mit bestimmten Erwartungen und Verhalten** mit dem System ist.

Benutzermodell für AP:

- Benutzer liest Ergebnisse vom Anfang zum Ende
- Benutzer **stoppt** bei **zufällig gewähltem relevantem Ergebnis** (gleichverteilt über alle relevanten Ergebnisse)
- Zufriedenheit = Präzision bis zu diesem Ergebnis
- **AP: Erwartungswert der Benutzerzufriedenheit**

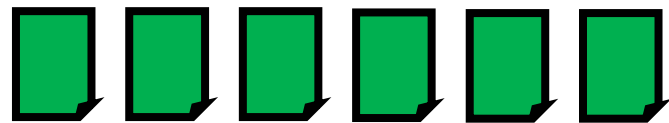
Stephen Robertson:

A new interpretation of average precision. [SIGIR 2008](http://dblp.org/rec/conf/sigir/Robertson08): 689-690

<http://dblp.org/rec/conf/sigir/Robertson08>

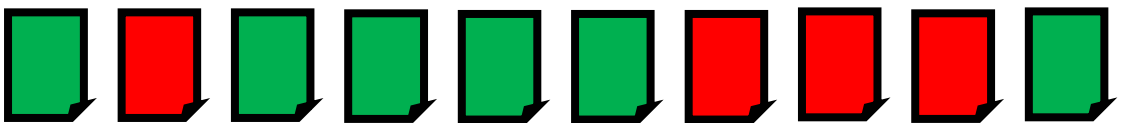
**AP evaluiert gleichzeitig verschiedene Retrievaltasks (recall-oriented und precision-oriented Tasks) und ist daher nicht ideal.**

# Durchschnittsbildung über mehrere Anfragen

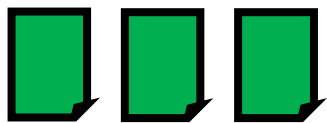


6 relevante Dokumente für Anfrage 1

System #1:

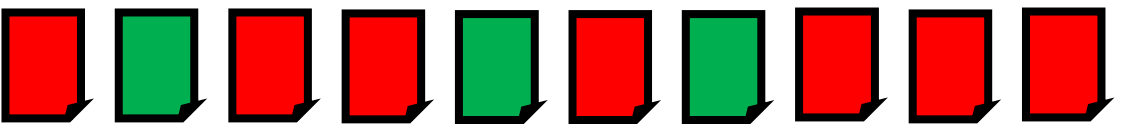


Recall	0,17	0,17	0,33	0,50	0,67	0,83	0,83	0,83	0,83	1,0
Precision	1,0	0,5	0,67	0,75	0,8	0,83	0,71	0,63	0,56	0,6



3 relevante Dokumente für Anfrage 2

System #1:



Recall	0,0	0,33	0,33	0,33	0,67	0,67	1,0	1,0	1,0	1,0
Precision	0,0	0,5	0,33	0,25	0,4	0,33	0,43	0,38	0,33	0,3

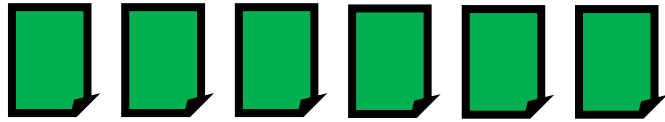
Betrachtung einer Anfrage reicht nicht, daher  
Durchschnittsbildung für ein System über **mehrere Anfragen**



# Durchschnittsbildung

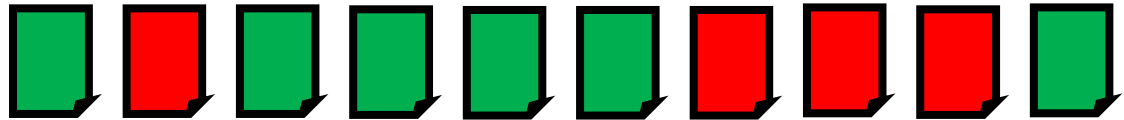
- **Precision**: für **eine Anfrage** an einem Recall-Punkt
- **Average Precision (AP)**: Mittelwertbildung **über die Recall-Punkte einer Anfrage**
- **Mean Average Precision (MAP)**:
  - Mittelwertbildung über **mehrere Anfragen**
  - Fasst **Rankings für mehrere Anfragen** zusammen, indem ein Durchschnitt über die mittleren AP-Werte gebildet wird
  - Sehr **verbreitetes Maß** in Forschungsliteratur
  - **Nimmt an**, dass der Nutzer daran interessiert ist, viele relevante Dokumente für jede Anfrage zu finden (= **Nutzerstandpunkt**)
  - Benötigt viele Relevanzbewertungen in der Textkollektion

# Mean Average Precision



6 relevante Dokumente für Anfrage 1

System #1:

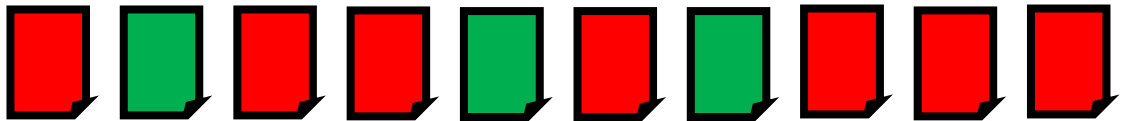


Recall	0,17	0,17	0,33	0,50	0,67	0,83	0,83	0,83	0,83	1,0
Precision	1,0	0,5	0,67	0,75	0,8	0,83	0,71	0,63	0,56	0,6



3 relevante Dokumente für Anfrage 2

System #1:



Recall	0,0	0,33	0,33	0,33	0,67	0,67	1,0	1,0	1,0	1,0
Precision	0,0	0,5	0,33	0,25	0,4	0,33	0,43	0,38	0,33	0,3

$$\text{AP für Anfrage 1} = (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6) / 6 = 0.78$$

$$\text{AP für Anfrage 2} = (0.5 + 0.4 + 0.43) / 3 = 0.44$$

$$\text{MAP} = (0.78 + 0.44) / 2 = 0.61$$

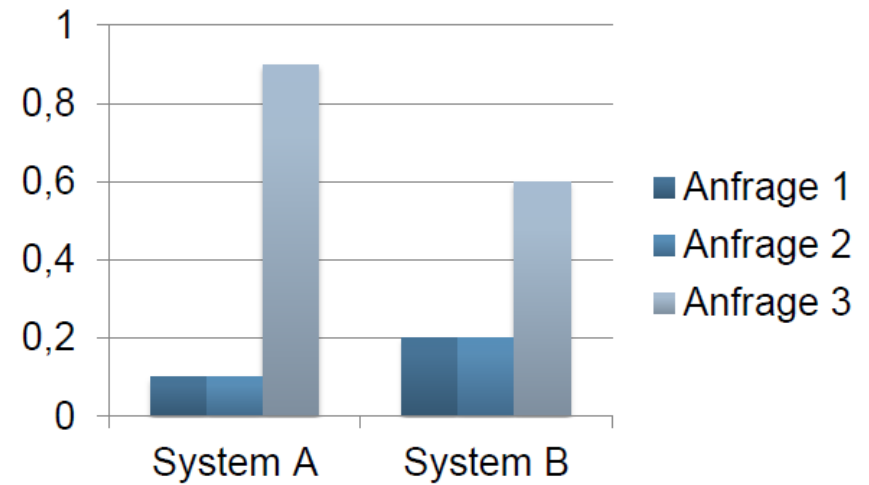
# Alternative zu MAP: GMAP

Legt ein höheres Gewicht auf Anfragen mit geringer  $AP$ .

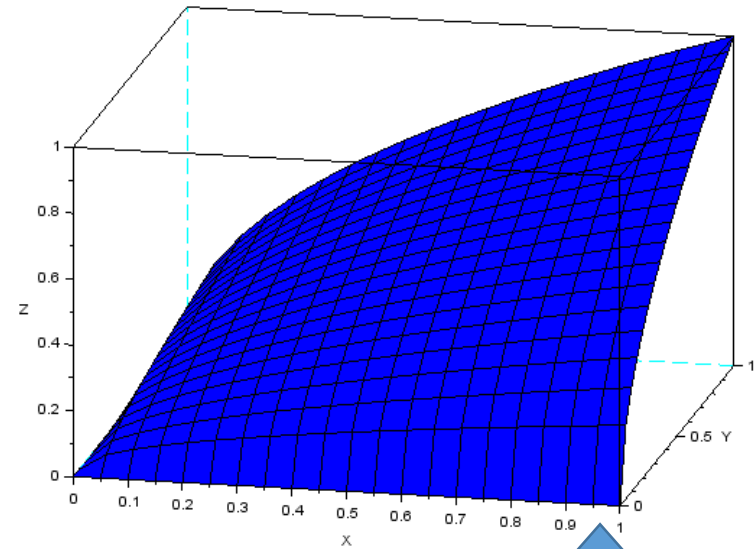
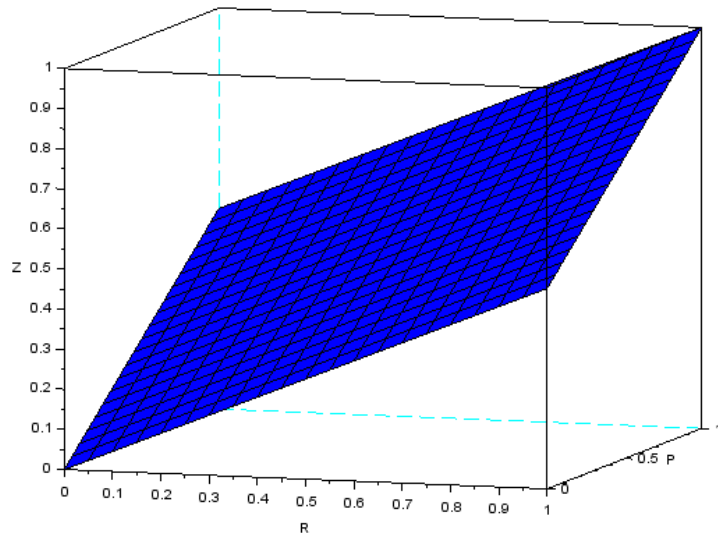
Bei  $n$  Anfragen:

$$GMAP := \sqrt[n]{\prod_{i=1}^n AP_i}$$

System A		System B	
Anfrage	AP	Anfrage	AP
1	0,1	1	0,2
2	0,1	2	0,2
3	0,9	3	0,6
GMAP	0,21	GMAP	<b>0,29</b>
MAP	<b>0,37</b>	MAP	0,33



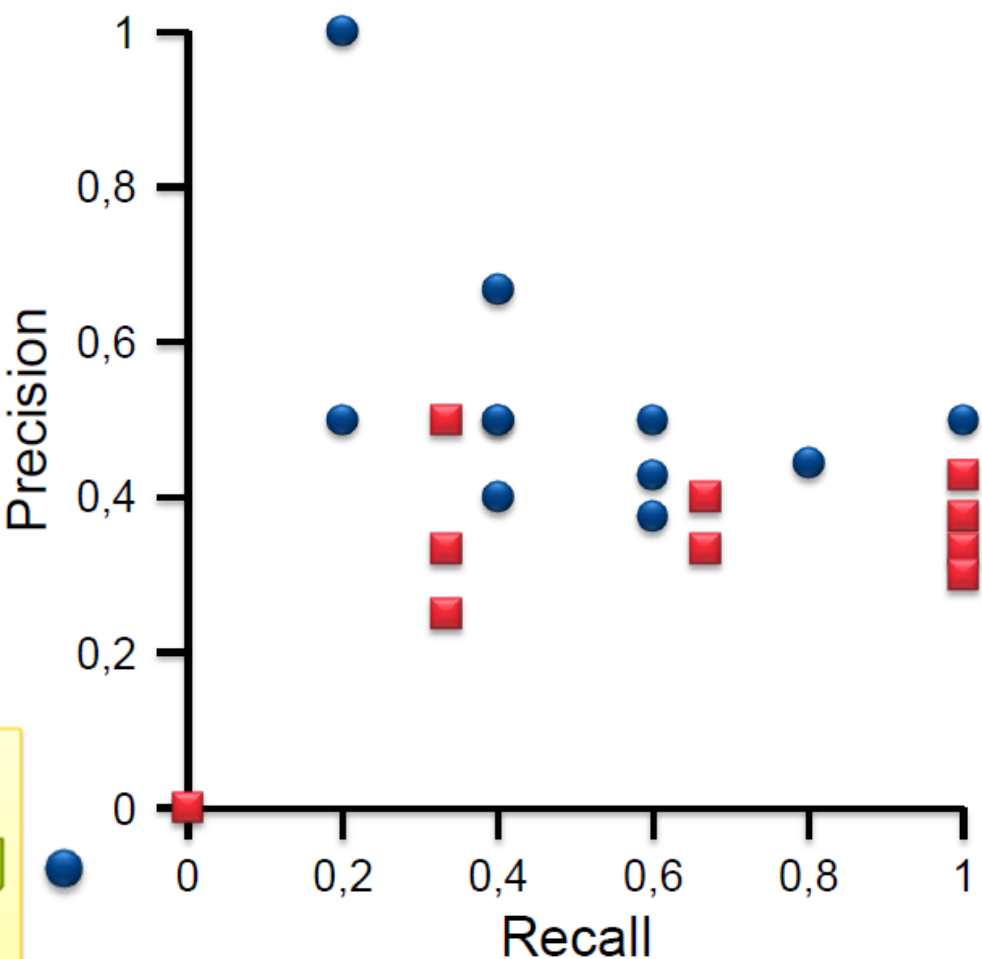
# Visualisierung: MAP vs. GMAP



Wenn einer der Werte sehr klein ist,  
kann dies durch einen anderen kaum  
ausgeglichen werden

# Recall-Precision-Graph

Recall-Precision-Graphen stellen ebenfalls nützliche Zusammenfassungen dar.



= die relevanten Dokumente für Anfrage 1

Ranking #1:

Recall:	0,2	0,2	0,4	0,4	0,4	0,6	0,6	0,6	0,8	1,0
Precision:	1,0	0,5	0,67	0,5	0,4	0,5	0,43	0,38	0,44	0,5

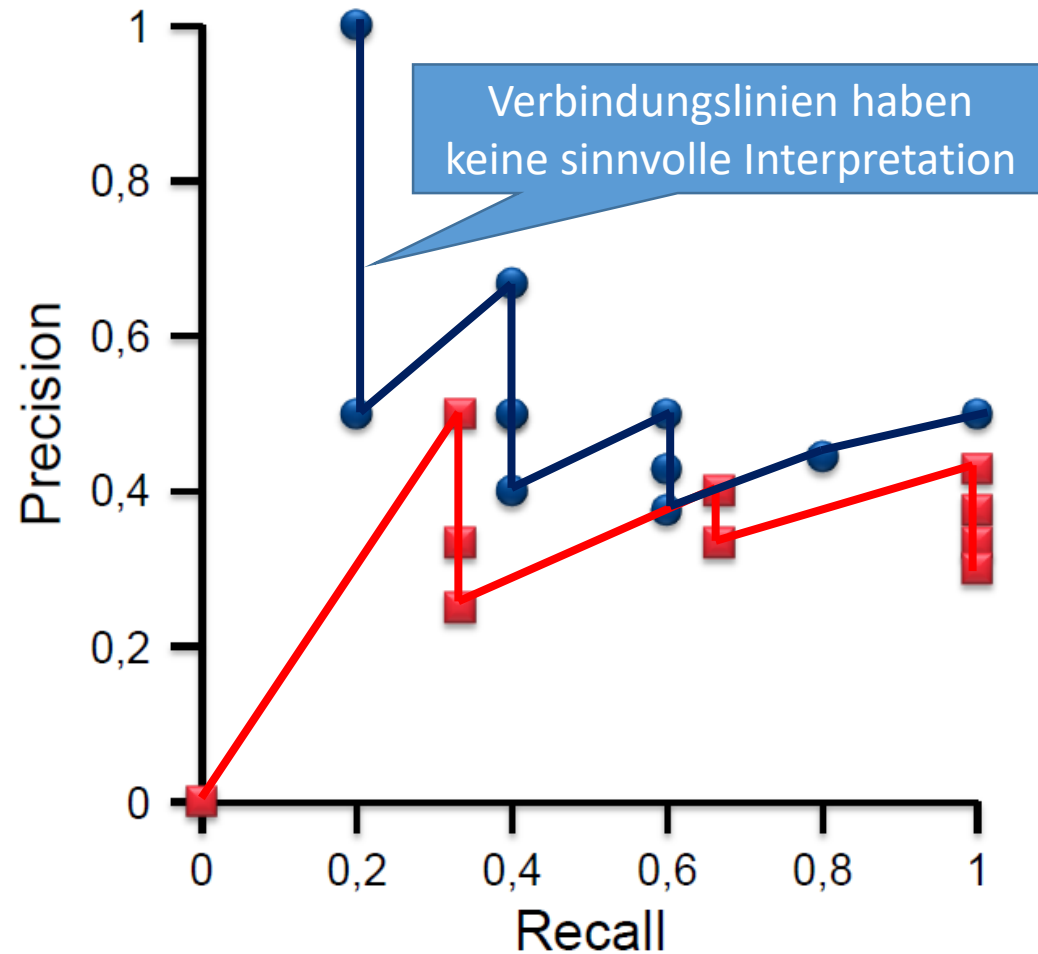
= die relevanten Dokumente für Anfrage 2

Ranking #2:

Recall:	0,0	0,33	0,33	0,33	0,67	0,67	1,0	1,0	1,0	1,0
Precision:	0,0	0,5	0,33	0,25	0,4	0,33	0,43	0,38	0,33	0,3

# Recall-Precision-Graph

**Recall-Precision-Graphen** stellen ebenfalls nützliche Zusammenfassungen dar.



# Interpolation

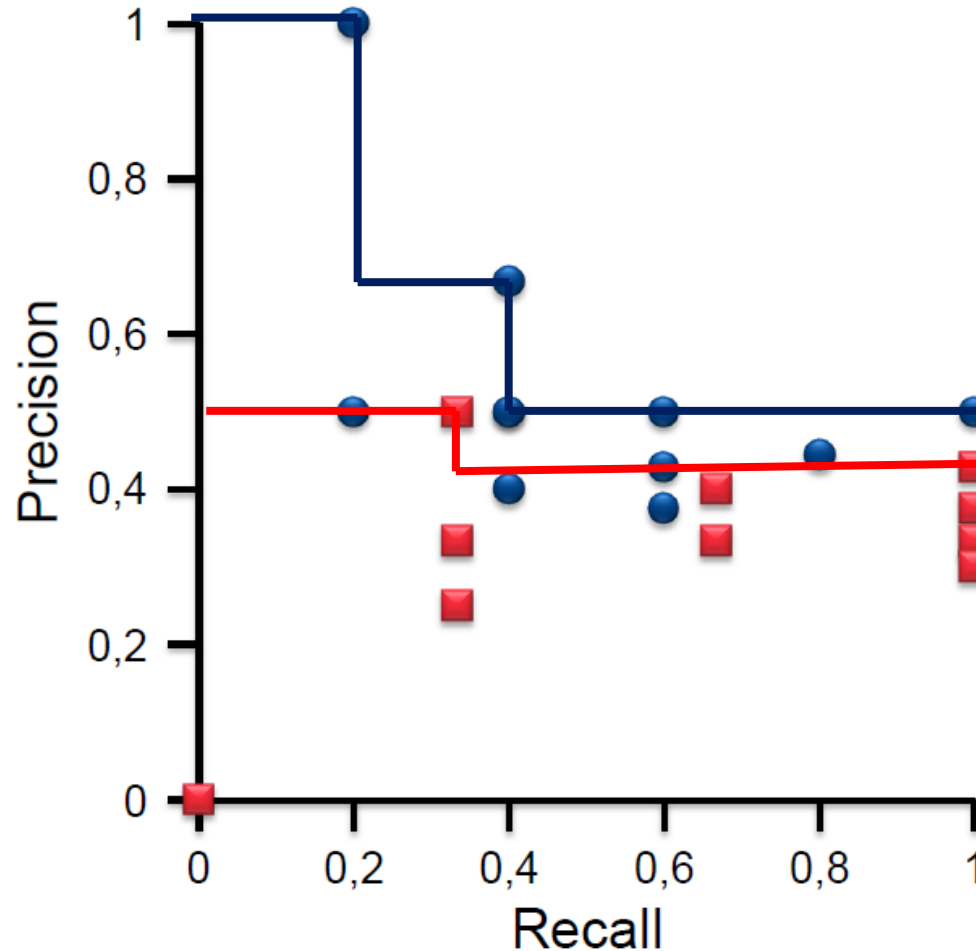
- Um einen **Durchschnittsgraphen** zu erzeugen, wird die **Precision an Standard-Recall-Punkten**  $R$  wie folgt berechnet

$$P(R) := \max\{P' : R' \geq R \wedge (R', P') \in S\}$$

- Wobei  $S$  die Menge von beobachteten  $(R, P)$  Punkten ist
- Dies definiert die Precision in jedem Recall-Punkt als **maximale Precision**, die in irgendeinem Recall-Precision-Punkt in einem **nicht kleineren Recall Level** beobachtet wird
  - Erzeugt eine **Stufenfunktion**
  - Definiert auch die Precision für Recall 0,0

# Interpolation: Ergebnis im Beispiel

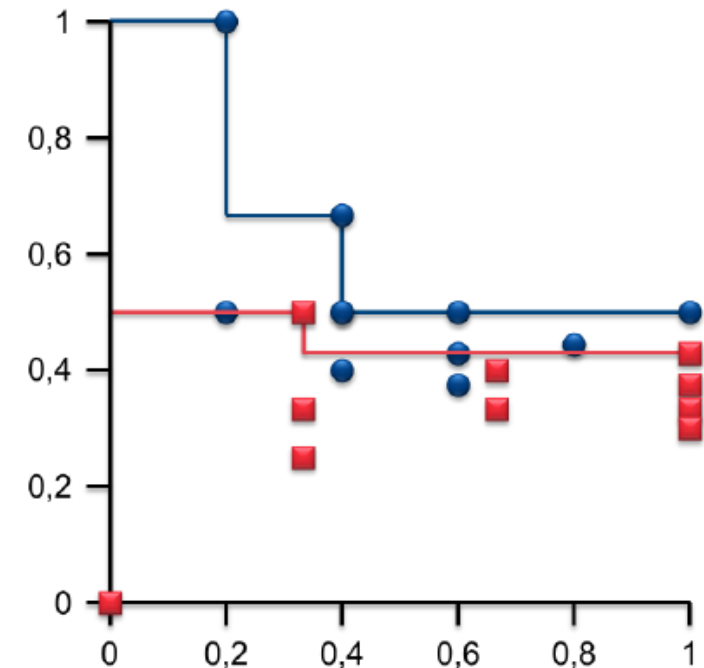
$$P(R) := \max\{P' : R' \geq R \wedge (R', P') \in S\}$$





# Durchschnittliche Precision an Standard-Recall-Punkten

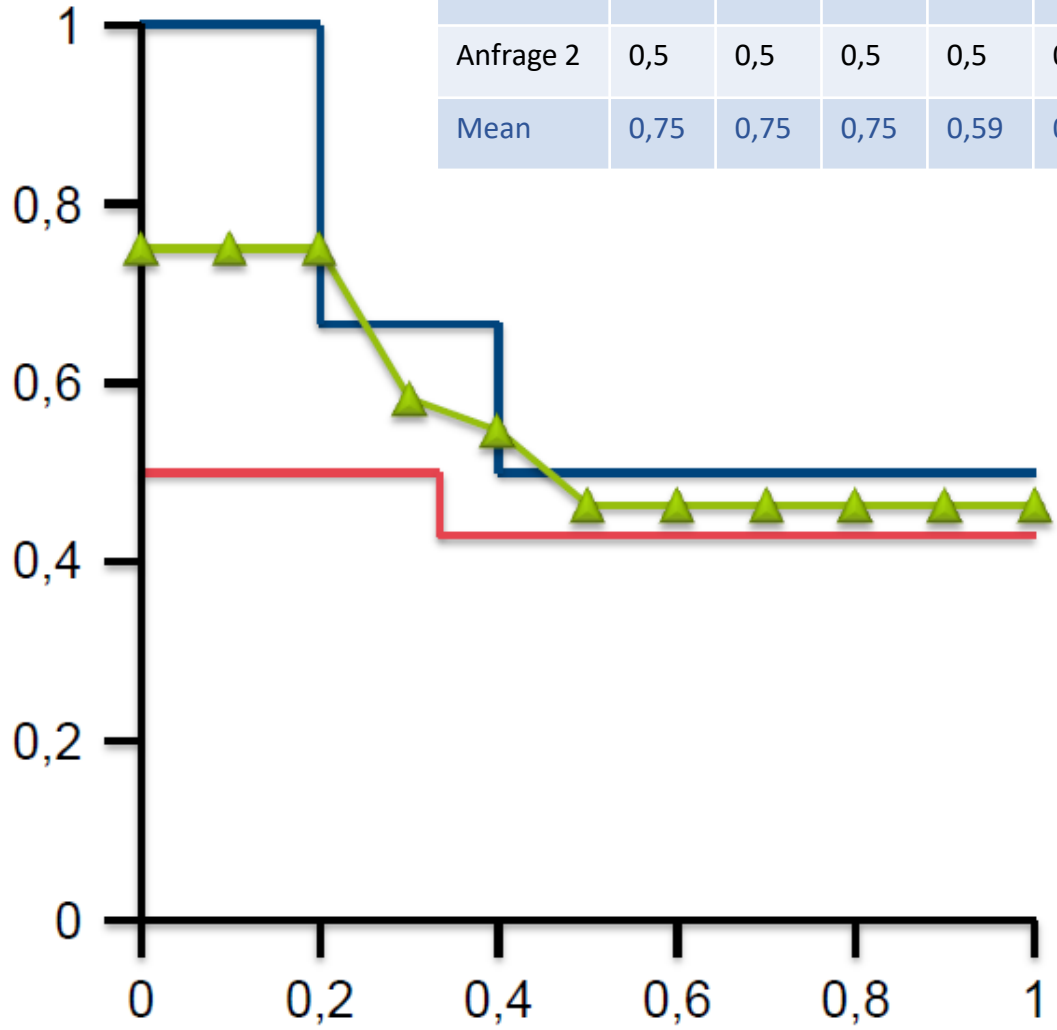
Der Recall-Precision-Graph wird gezeichnet, indem über mehrere Anfragen die durchschnittlichen Precision-Werte an Standard-Recall-Punkten berechnet werden:



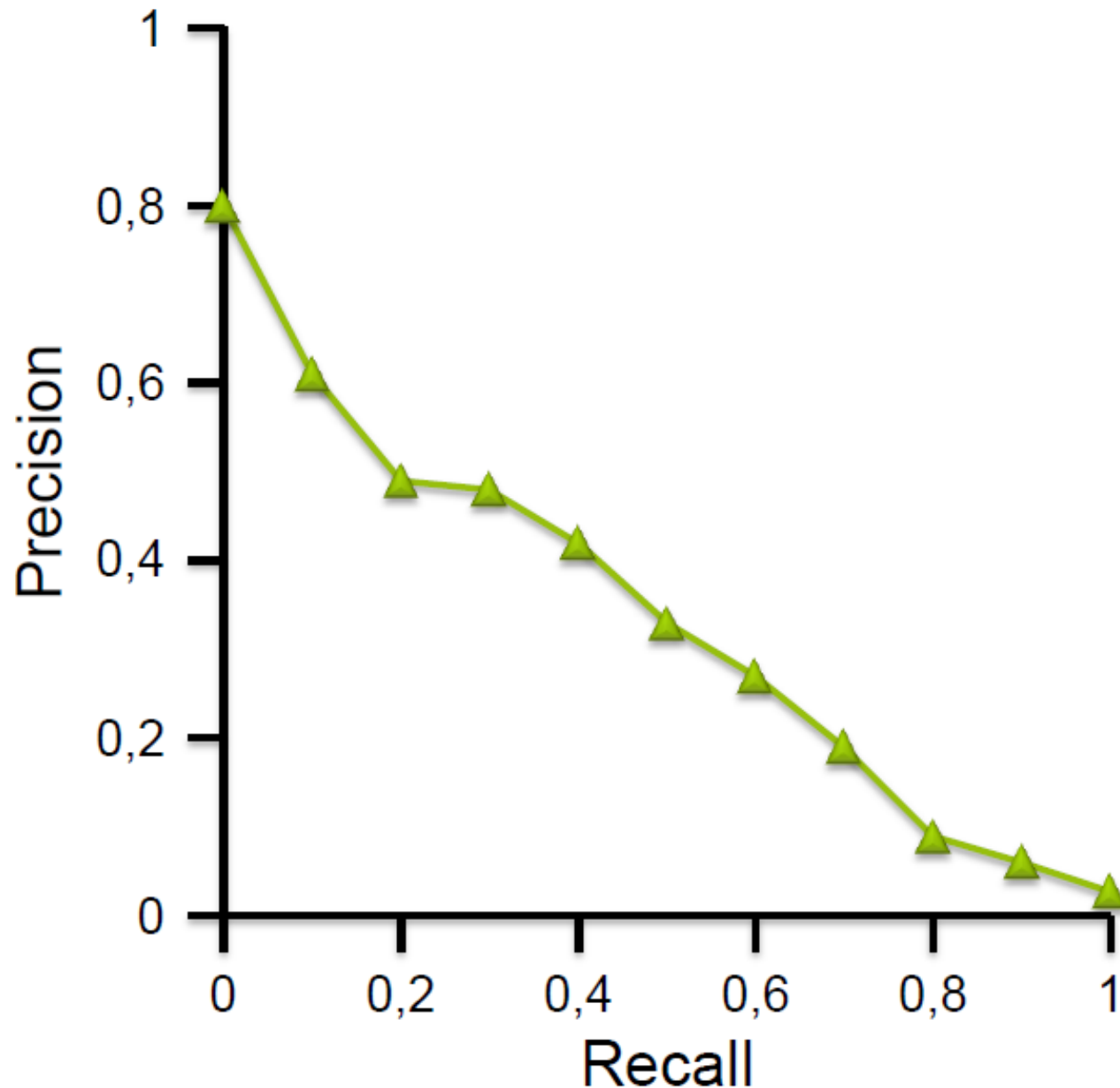
Recall	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
Anfrage 1	1,0	1,0	1,0	0,67	0,67	0,5	0,5	0,5	0,5	0,5	0,5
Anfrage 2	0,5	0,5	0,5	0,5	0,43	0,43	0,43	0,43	0,43	0,43	0,43
Mean	0,75	0,75	0,75	0,59	0,55	0,47	0,47	0,47	0,47	0,47	0,47

# Durchschnittlicher Recall-Precision-Graph

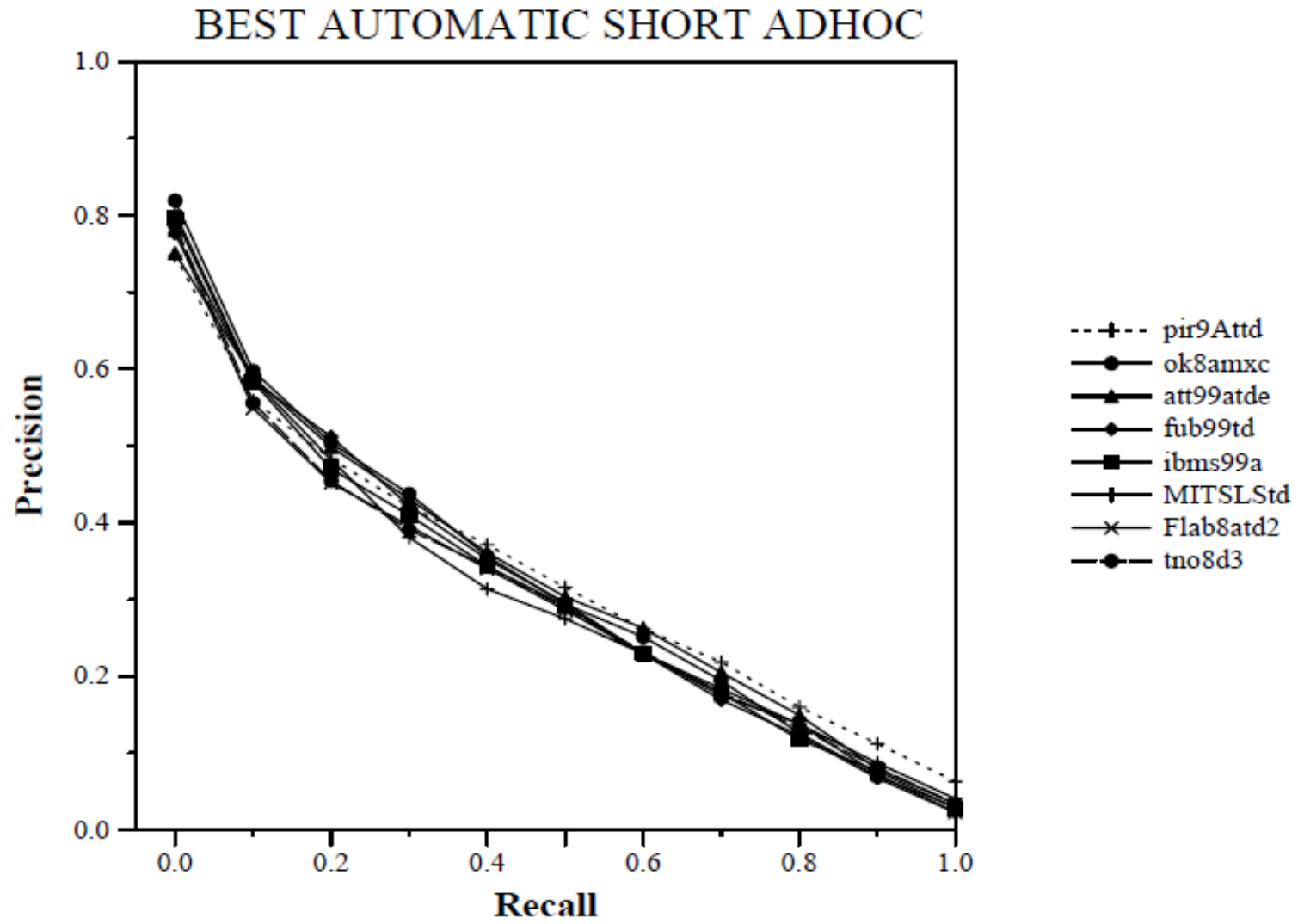
Recall	0,0	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1,0
Anfrage 1	1,0	1,0	1,0	0,67	0,67	0,5	0,5	0,5	0,5	0,5	0,5
Anfrage 2	0,5	0,5	0,5	0,5	0,43	0,43	0,43	0,43	0,43	0,43	0,43
Mean	0,75	0,75	0,75	0,59	0,55	0,47	0,47	0,47	0,47	0,47	0,47



# Typischer Graph für ein System mit 50 Anfragen

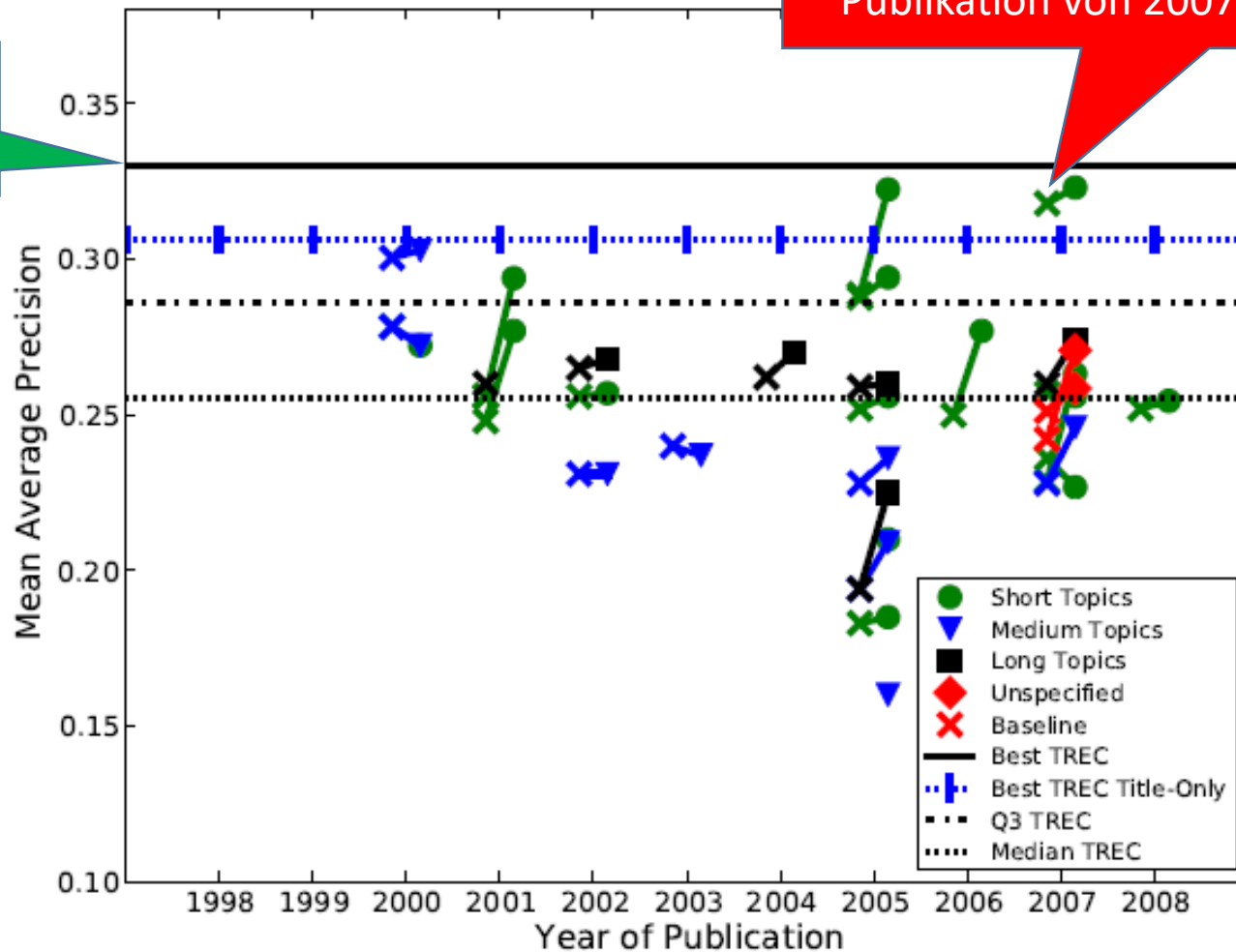


# Beispiel: TREC



Quelle: [Ellen M. Voorhees, Donna Harman:](#)  
Overview of the Eighth Text Retrieval Conference (TREC-8).  
<http://dblp.org/rec/conf/trec/VoorheesH99>

# Nutzen von Standard-Testcollections



Quelle: Timothy G. Armstrong, Alistair Moffat, William Webber, Justin Zobel:  
Improvements that don't add up: ad-hoc retrieval results since 1998. CIKM 2009: 601-610.  
<http://dblp.org/rec/conf/cikm/ArmstrongMWZ09>



# Konzentration auf die Top-Dokumente

- **Benutzer** tendieren dazu, nur den **obersten Teil der Ergebnisliste** anzusehen, um relevante Dokumente zu finden
- Einige Suchaufgaben führen nur zu einem relevanten Dokument
  - z. B. **navigierende Suche, Question Answering**

⇒ Recall kein angemessenes Maß

Stattdessen sollte gemessen werden, wie gut die Suchmaschine relevante Dokumente in Top-Rängen einstuft

# Fokus auf die besten Dokumente

- **Precision in Rang**  $R(P@R)$ 
  - $R$  typischerweise 5, 10 oder 20 ( **$P@20$** )
  - Einfach zu berechnen, einfache Mittelwertbildung, einfach zu verstehen
  - $P@R$  berücksichtigt aber die Verteilung innerhalb der Ränge nicht  
 genauso gut wie  mit  $P@5=0,2$
- **Reziproker Rang** (für Anfragen, bei denen es um ein relevantes Dokument geht)
  - Kehrwert des Ranges, an dem das **erste relevante Dokument** abgerufen wird
  - **Mean Reciprocal Rank** (MRR) ist der Durchschnitt der reziproken Ränge über einer Menge von Anfragen
  - Sehr empfindlich gegenüber der Rangposition
  - **Benutzermode**ll unklar: Ist wirklich der erste Rang soviel wichtiger als die anderen Ränge?

# Discounted Cumulative Gain

Verbreitetes Maß, um **Websuche** und verwandte Aufgaben zu evaluieren

Zwei Annahmen:

- **Hochrelevante Dokumente** sind nützlicher als nur marginal relevante Dokumente
- Je höher der Rang eines relevanten Dokuments, desto weniger nützlich ist es für den Nutzer, da es mit geringerer Wahrscheinlichkeit betrachtet wird



# Discounted Cumulative Gain

Verwendet **gestufte Relevanz** (graded relevance) als Maß für die Nützlichkeit oder den Gewinn, der durch Betrachtung eines Dokuments erreicht wird

- Der **Gewinn** wird beginnend mit den bestplatzierten Ergebnissen akkumuliert und kann bei **höheren Rängen reduziert** (discounted) werden.
- Typischer **Discount** ist  $\frac{1}{\log_2 \text{Rang}}$
- Mit Basis 2 ist der Discount  $\frac{1}{2}$  bei Rang 4 und  $\frac{1}{3}$  bei Rang 8

Rang	1	2	3	4	5	8	20	100	1000
Discount		1,0	0,63	0,50	0,43	0,33	0,23	0,15	0,10

# Discounted Cumulative Gain

- Sei  $rel_i$  die Relevanz des Dokumentes an Rang  $i$
- $DCG_p$  ist der gesamte **Gewinn**, der in einem bestimmten Rang  $p$  **akkumuliert** ist:

$$DCG_p := rel_1 + \sum_{i=2}^p \frac{rel_i}{\log_2 i}$$

- Alternative Formulierung mit mehr Gewicht für sehr relevante Ergebnisse:

$$DCG_p := \sum_{i=1}^p \frac{2^{rel_i} - 1}{\log_2(i + 1)}$$

Betonung auf dem Abruf hochrelevanter Dokumente

# Beispiel zu DCG

10 nach Rängen geordnete Dokumente bewertet gemäß einer Relevanzskala mit  $rel_i \in \{0,1,2,3\}$

Rang	1	2	3	4	5	6	7	8	9	10
Relevanz	3	2	3	0	0	1	2	2	3	0
Discounted gain DG	3	2/1	3/1,59	0	0	1/2,59	2/2,81	2/3	3/3,17	0
=	3	2	1,89	0	0	0,39	0,71	0,67	0,95	0
$\Rightarrow$ DCG =	3	5	6,89	6,89	6,89	7,28	7,99	8,66	9,61	9,61

# Normalisierter DCG

- Der **Mittelwert** über DCG-Werte wird über eine Menge von Anfragen in spezifischen Rängen gebildet:
  - z. B. DCG in Rang 5 ist 6,89 und in Rang 10 dann 9,61
- DCG-Werte werden oft normalisiert, indem die DCG-Werte in jedem Rang **mit den DCG-Werten für perfektes Ranking verglichen** werden
  - vereinfacht Durchschnittsbildung für Anfragen mit unterschiedlicher Anzahl an relevanten Dokumenten

# Beispiel zu NDCG

- *NDCG*-Werte=Division des eigentlichen Wertes durch idealen Wert
- $NDCG \leq 1$  bei jedem beliebigen Rang

Rang	1	2	3	4	5	6	7	8	9	10
Relevanz	3	2	3	0	0	1	2	2	3	0
Discounted gain DG	3	2/1	3/1,59	0	0	1/2,59	2/2,81	2/3	3/3,17	0
=	3	2	1,89	0	0	0,39	0,71	0,67	0,95	0
⇒ DCG =	3	5	6,89	6,89	6,89	7,28	7,99	8,66	9,61	9,61
Perfektes Ranking	3	3	3	2	2	2	1	0	0	0
Ideale DCG-Werte	3	6	7,89	8,89	9,75	10,52	10,88	10,88	10,88	10,88
⇒ NDCG =	1	0,83	0,87	0,76	0,71	0,69	0,73	0,8	0,88	0,88

# Verwenden von Präferenzen

- Wir haben bereits gesehen, wie **Nutzerpräferenzen** aus **Anfragelogs** abgeleitet werden können.
- Zwei Rankings, die durch Präferenzen beschrieben werden, können mit dem **Kendall's  $\tau$  Koeffizient** verglichen werden:
$$\tau := \frac{P - Q}{P + Q}$$
- $P$  ist die Anzahl an Präferenzen, die **übereinstimmen**, und  $Q$  ist die Anzahl derer, die **nicht übereinstimmen** (d.h. gegensätzlich sind)
- Für die Präferenzen, die von **binären Relevanzbewertungen** abgeleitet werden, kann **BPREF** verwendet werden

# BPREF

- Besonders wichtig für Anfragen mit unvollständigen Relevanzbewertungen
- Für eine Anfrage mit  **$R$  relevanten Dokumenten** werden nur die ersten  **$R$  als nicht relevant erkannten Dokumente** betrachtet

$$BPREF := \frac{1}{R} \sum_{d_r} \left( 1 - \frac{N_{d_r}}{\min(N, R)} \right)$$

- $N$  ist die Zahl der als nicht relevant erkannten Dokumente
- $d_r$  ist ein **relevantes Dokument** und
- $N_{d_r}$  ist die Anzahl der **ersten  $R$  nichtrelevanten Dokumente**, die vom System **höher eingestuft** wurden als  $d_r$
- Alternative Definition

$$BPREF := \frac{P}{P + Q}$$

# BPREF: Beispiel

$$BPREF := \frac{1}{R} \sum_{d_r} \left( 1 - \frac{N_{d_r}}{\min(N, R)} \right)$$

Rang	1	2	3	4	5	6	7	8	9	10
rel?	1	1	0	?	1	0	?	0	1	0
$1 - \frac{N_{d_r}}{R}$	1	1			0,75				0,25	

$$R=4$$

$$BPREF = 1/4 * (1 + 1 + 0,5 + 0) = 0,625$$

Von den ersten 4 als nicht relevant erkannten Dokumenten sind 4 höher gerankt als dieses Dokument

Ist die Relevanz dieser beiden Dokumente unbekannt, ergibt sich

$$BPREF = 1/4 * (1 + 1 + 0,75 + 0,25) = 0,75$$



# Effizienzmaße

Maß	Beschreibung
Verstrichene Indexierungszeit (elapsed index time)	Misst den Zeitverbrauch für die Erstellung eines Index auf einem bestimmten System
Prozessorzeit für Indexierung (indexing processor time)	Misst die vom Prozessor für die Indexierung benötigte Zeit in Sekunden. Diese Zeit entspricht der verstrichenen Indexierungszeit, jedoch werden I/O-Wartezeiten und Zeitgewinne durch parallel Verarbeitung nicht beachtet.
Anfragendurchsatz (query throughput)	Anzahl der pro Sekunde verarbeiteten Anfragen
Anfragelatenzzeit (query latency)	Die Zeit in Millisekunden, die der Nutzer nach Abschicken der Anfrage auf eine Antwort durchschnittlich wartet (arithmetisches Mittel, besser Median)
Temporärer Speicherplatz für Indexierung (indexing temporary space)	Speicherplatz, der während der Indexerstellung benötigt wird
Indexgröße (index size)	Speicherplatz, den der fertige Index benötigt

# Evaluierung von IR-Systemen

Tuning von Parametern

# Optimieren der Parameterwerte

- Retrievalmodelle enthalten oft **Parameter**, die **getunt** werden müssen, um die beste Leistung für verschiedene Datentypen und Anfragetypen zu erzielen
- Für Experimente:
  - Verwenden von (disjunkten) **Trainings- und Testdatenmengen**
  - Wenn wenige Daten vorhanden sind, wird **cross-validation** eingesetzt, wobei die Daten in  $K$  Submengen aufgeteilt werden
    - Man nimmt jeweils  **$K-1$  Mengen zum Trainieren** und die  **$K$ -te zum Testen** und **bildet das Mittel** über alle  $K$  Möglichkeiten
    - Idealerweise alle  $K$  Mengen gleichartig, insbesondere bezüglich Anzahl relevanter Dokumente
    - Spezialfall **Leave-one-out**: Eine Menge entspricht genau einem Topic einer Testkollektion (Nachteil: ungleich Zahl relevanter Dokumente)
  - Das Verwenden von disjunkten Trainings- und Testdaten vermeidet **overfitting**, wenn die Parameterwerte sich nicht auf andere Daten verallgemeinern lassen

# Finden der Parameterwerte

- Es werden viele Techniken eingesetzt, um optimale Parameterwerte für gegebene Trainingsdaten zu finden
  - Standardproblem im maschinellen Lernen
- Im IR wird der **Raum an möglichen Parameterwerten** oft **brute-force** exploriert
  - Erfordert eine **große Anzahl an Retrievaldurchgängen** mit kleinen Parameteränderungen (**parameter sweep**)

# Online-Tests

- **Bisher** Evaluierung und Optimierung **offline**
- Jetzt: Tests (oder auch Training/Optimierung) unter Verwendung von Echtzeitdaten von einer Suchmaschine
- **Vorteile:**
  - Echte Nutzer, weniger voreingenommen, große Mengen an Testdaten
- **Nachteile:**
  - Daten mit Noise behaftet
  - kann die „User Experience“ verschlechtern
- Oft **bei einem kleinen Anteil** (1 bis 5%) der Echtzeitdaten eingesetzt

# Online-Tests

## Google tests 'more results' mobile search interface and new search refinement buttons

Google told us, 'We constantly experiment with new search formats and experiences to deliver the best experience for our users.'

### How Netflix's A/B testing drives service's relentless subscriber growth

For example, Netflix will show a test audience two versions of a thumbnail for original series "Orange Is the New Black": one full of elements, the other only featuring star Tyler Schilling and a few text elements. The latter tested much better.

Google Analytics Solutions | Optimize

Test, adapt,  
personalize.

Discover the most engaging customer experiences with Google Optimize. Test different variations of your website and then tailor it to deliver a personalized experience that works best for each customer and for your business.

# Aber Vorsicht...

11. Januar 2016, 10:40 Uhr Bewusste Manipulation

## Wie Facebook seine Nutzer zu Versuchskaninchen degradiert



Facebook stellte mit absichtlichen Fehlern in der Android-App die Geduld seiner User auf (Foto: Lukas Schulze/dpa)

## GOOGLE ERZÜRNT USER MIT DESIGN EXPERIMENT

By Alex | Mai 10th, 2016 | Design, News

Für Studie

## Facebook manipuliert die Gefühle

Schlecht gelaunt nach einem Besuch auf Facebook? Das kann passieren, weil Facebook seine Nutzer manipuliert. Für eine Studie hat das Netzwerk die Laune mancher Nutzer verbessert - und andere Leute runtergezogen.

29.06.2014, von PATRICK BERNAU

f Teilen Twittern > Teilen E-mailen



### Suchmaschine: Google stellt Hintergrund-Experiment ein

11. Juni 2010, 12:22

f g+ > 39 POSTINGS

Benutzer beschwerten sich über farbenfrohe Bilder bei Besuch der Seite - Wer möchte kann das Feature aber nun nutzen

# Zusammenfassung

- Es gibt **kein Maß**, das für **jede beliebige Applikation** korrekt ist
  - Das gewählte Maß muss **angemessen für die Aufgabe** sein
  - Verwendung von **Kombinationen**
  - Zeigt **verschiedene Aspekte** der Effektivität des Systems
- Analyse der Ergebnisse individueller Anfragen
- Wichtig: **Nutzerstandpunkt!**



# Zusammenfassung

- **Analyse einzelner Anfragen** oft wichtiger als Durchschnittsbetrachtung!
  - Effektivität bei **leichten/schweren Anfragen**
- Kleine Unterschiede in Kennzahlen haben oft keinen Zusammenhang zum „**Nutzerempfinden**“