



Information Retrieval

07 - Evaluation von IR-Systemen

Philipp Schaer, Technische Hochschule Köln, Cologne, Germany

Version: 2021-05-04

“The word for modifying the atmosphere”

Google

the word for modifying the atmosphere

About 16,800,000 results (0.23 seconds)

[Terraforming - Wikipedia, the free encyclopedia](#)
en.wikipedia.org/wiki/Terraforming

... is the hypothetical process of deliberately **modifying** its **atmosphere**, tempera
... **The term** is sometimes used more generally as a synonym for planetary ...

[NASA - What's the Difference Between Weather and Climate?](#)

www.nasa.gov/mission_pages/noaa-n/climate/climate_weather.html

Weather is what conditions of the **atmosphere** are over a short period of time, and
When we talk about climate **change**, we talk about changes in long-**term** ...

[Frequent Questions - Science | Climate Change | U.S. EPA](#)

www.epa.gov/climatechange/fq/science.html

The term climate **change** is often used as if it means the same thing as **the term**
human activities that **change the atmosphere's** makeup (e.g., burning fossil ...

Bing

bing

the word for modifying the atmosphere

316,000,000 RESULTS

Showing results for **the word for modifying the air** too.
Do you want results only for **the word for modifying the atmospher**

[modify - definition of modify by the Free Online Dictionar](#)
www.thefreedictionary.com/modify ▾

(Linguistics / Grammar) Grammar (of a **word** or group of **words**) to
modify the temperature of; "attempt **the air**" syncopate - **modif**

[The | Define The at Dictionary.com](#)
dictionary.reference.com/browse/the ▾

(used before a **modifying** adjective to specify or limit its **modifying**
wrong road and drove miles out of his way. 9. (used to indicate on

[Modifying the Air Box Worth It? - Triumph Forum: Triu](#)
www.triumphrat.net/twins-technical-talk/192276-modifying-the-air

[NOTE: WHEN I SAID "MODIFYING" IN THE TITLE OF THIS THR
REMOVING THE AIR BOX. MY BAD.] I'm not into high performan

Themen der Veranstaltung

- Evaluation von IR-Systemen
- Relevanzbegriff
- IR-Evaluation nach dem Cranfield-Paradigma
- Evaluationsmaße
 - Recall, Precision für ungeranktes Retrieval
 - P@k, R@k, MAP, R-Precision für geranktes Retrieval

Was können wir messen?

Geschwindigkeit der Indexierung

- Anzahl der Dokumente pro Stunde

Geschwindigkeit der Suche

- Steigt die Suchzeit mit Größe des Index?

Ausdruckskraft der Anfragesprache

- Möglichkeit komplexe Informationsbedürfnisse auszudrücken
- Geschwindigkeit komplexer Anfragen

Verständlichkeit der Benutzeroberfläche

- usw.

Was sollten wir messen?

- All diese Kriterien können **gemessen** werden: Wir können also **Geschwindigkeit oder Größen quantifizieren**
- Ein Kriterium ist bisher aber vernachlässigt worden: Die **Zufriedenheit des Benutzers!**
 - Was ist das eigentlich?
 - Geschwindigkeit/Wartezeiten/Umfang des Index sind sicherlich Faktoren, die hier eine Rolle spielen...
 - Aber schnelle, jedoch nutzlose Antworten befriedigen nicht das Informationsbedürfnis eines Benutzers
 - Wir brauchen also einen Weg, die **Zufriedenheit eines Benutzers zu quantifizieren**

Zufriedenheit des Benutzers

Problem: Wer ist der Benutzer?

- Dies ist immer abhängig vom Einsatz und Anwendungsfall

Fall A: Websuchmaschine

- Benutzer findet, was er sucht und kommt wieder zurück zur Websuchmaschine → Anzahl der wiederkehrenden Nutzer

Fall B: eCommerce-Webseiten (z.B. Amazon etc.)

- Benutzer finden gewünschte Produkte → Anzahl der Käufe

Fall C: Firmenwebseiten

- Benutzer finden Firmeninformationen (z.B. Vertragsdokumente) → Gesparte Zeit, Aufwand

Fall X: ...

„Qualitätsmessung“

Fachgruppe IR / Gesellschaft für Informatik:

„... 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.“

Gebraucht werden:

- 1. Faktoren, die den Einfluss des Systems messen**
- 2. Anwendungsprobleme der Benutzer**
- 3. Unterstützung des Systems bei der Lösung**



Gesellschaft
für Informatik

Effektivität / Effizienz / Zufriedenheit

- **Effektivität** = Fähigkeit des Systems, die benötigte Information mit möglichst wenigen Nutzerressourcen (Zeit, Anstrengung) und ohne „Rauschen“ (Qualität der Ergebnismenge) anzubieten
 - Die „relevanten und nur die relevanten Dokumente finden.“
- **Effizienz** = Fähigkeit des Systems, die benötigten Ressourcen möglichst schnell (Rechnerzeit, Anfragezeit) und sparsam (Speicherplatz) finden zu können.
- **Nutzerzufriedenheit** = Fähigkeit des Systems, den Benutzer zufriedenzustellen.

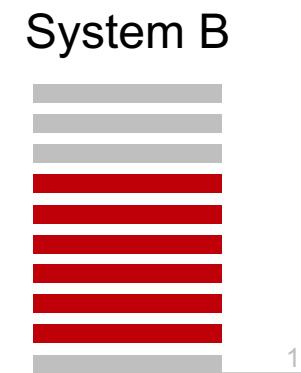
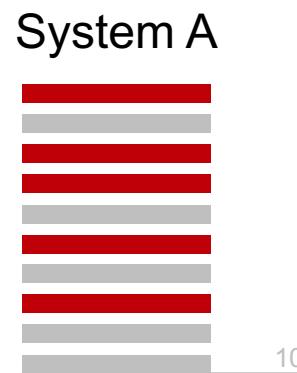
Systemfaktoren für Effektivität

- Auswahl der Dokumente
- Dokumentformate
- Dokumentaufbereitung
- Indexierungsmethode (IR-Algorithmus)
- Anfrageaufbereitung
- Darstellung der Ergebnisse

Anwendungsprobleme & Unterstützung

- Anwendungsprobleme = Anfragen an das System
- Unterstützung des Systems bei der Lösung = Ergebnislisten von Dokumenten

→ Vergleich der Ergebnislisten



Evaluationsszenarios im IR

Wie kann man die Effektivität von Systemen vergleichen?

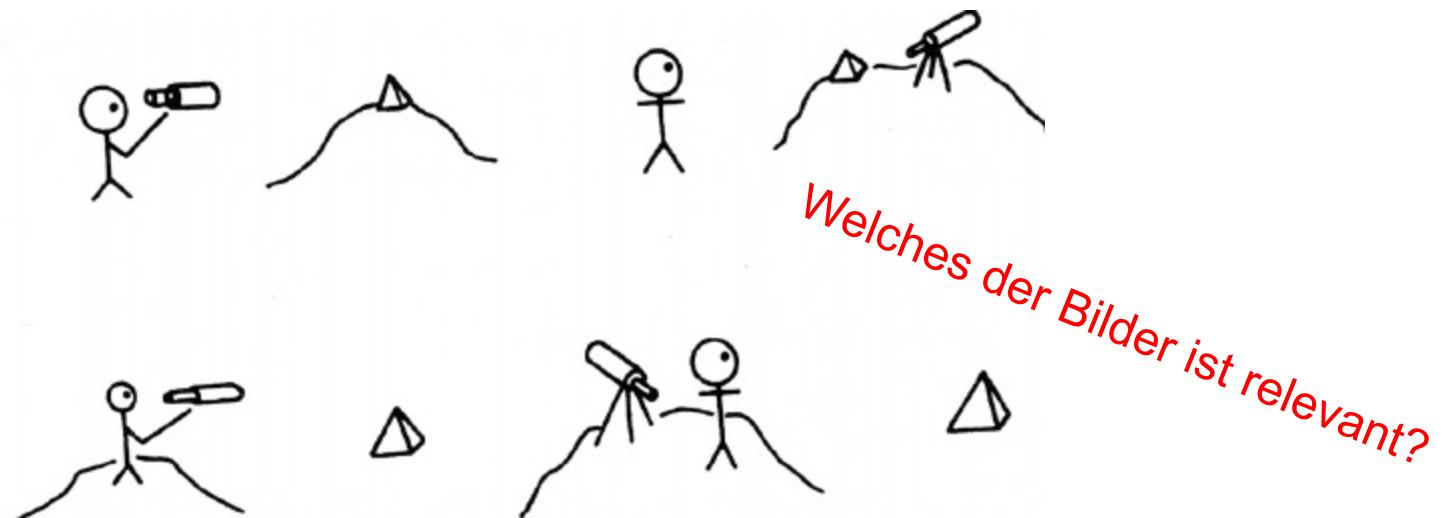
- **Gleiche Anfragen**
- **Gleiche Dokumentkollektion** (gleiche Chance, die „richtigen“ Dokumente“ zu finden)
- **Relevanzentscheidung für Dokumente** in Ergebnislisten

Wann ist ein Dokument „relevant“ auf eine Anfrage?

- Enthält alle Terme der Anfrage?
- Beantwortet den Informationsbedarf aus der Anfrage?
- Beantwortet das reale Informationsbedürfnis des Suchers?
- Ist neu für den Sucher?

Relevanz ist ein grundlegendes Problem

- „The man saw the pyramid on the hill with the telescope.“
- Viele Interpretationen dieses Satzes sind denkbar...



Relevanz und das Sprachproblem

Informationsbedürfnis 1 \neq Informationsbedürfnis 2
Anfrage \neq Informationsbedürfnis
Relevanz 1 \neq Relevanz 2
Relevanz $t(n)$ \neq Relevanz $t(n+1)$

Nutzer ist sich im Unklaren:

- über eigenen Informationsbedarf
- wie das Anliegen formuliert werden kann
- was abgesucht wird
- was erwartet wird
- ob relevante Ergebnisse gefunden wurden
- ob alle wichtigen Ergebnisse gefunden wurden

Unterschiedliche Ebenen der Relevanz

Situative Relevanz

- Tatsächliche Nützlichkeit einer Informationseinheit in einer konkreten Situation gemeint.

Subjektive Relevanz (bzw. Pertinenz)

- Die Nützlichkeit einer Informationseinheit für eine Person mit einem bestimmten Informationsbedürfnis.

Objektive Relevanz

- Keine subjektive Einschätzung einer Person, sondern mehrere Einschätzungen unterschiedlicher Personen (Abschwächung pers. Präferenzen)

Systemrelevanz

- Die algorithmische Berechnung der Relevanz durch ein IR-System. Hierbei handelt es sich nur um einen Schätzwert des Systems, der auch mit dem Begriff RSV (Retrieval Status Value) abgekürzt wird.

Relevanzentscheidungen

Relevanz ist subjektiv – wie kann sie objektiv bewertet werden?

- Nur „**topical relevance**“ (Dokument passt zu Fragestellung)
- **Objektivitätskriterium** (relevant für Informationsbedarf in Anfrage, nicht Nutzer)
- **Unabhängigkeitskriterium** (jedes Dokument wird separat betrachtet, unabhängig vom vorgehenden Dokumenten)
- **Mehrfachbewertung** durch verschiedene Personen (sogenannte Assessoren)
- Binärentscheidung oder abgestufte Entscheidungen möglich



Standardtests im IR

Ende 1950er: Cranfield Studien

- Cyril Cleverdon (Bibliothekar!)
- 1,400 Dokumente
- 279 Anfragen (Aufgabe: **ad-hoc Retrieval**)
- **Alle Dokumente für jede Anfrage relevanzbewertet** (abgestuft)
- 6 Studenten, 3 Monate – Überprüfung durch Wissenschaftler
- http://ir.dcs.gla.ac.uk/resources/test_collections/cran/

Alle drei Elemente (**Dokumente**, **Topics** und **Relevanzbewertungen**) zusammen bilden eine **Test-Kollektion** und können von Forschern genutzt werden um neue Systeme damit zu testen und zu evaluieren.

Standardtests im IR

Heute: **Evaluationsinitiativen**, die Dokumentkollektionen, Testanfragen und Relevanzbewertungen bereitstellen

- Verschiedene **Aufgabenstellungen**: Tracks (ad-hoc, Question Answering, Interactive, Filtering, Image, Speech, Legal...)
- 25+ Anfragen / Topics, damit Testergebnisse signifikant sind
- Dokumente: Anzahl abhängig von Aufgabe
(Web: 1 Mrd. Seiten oder mehr)
- Relevanzbewertungen über ein **Auswahlverfahren** (nicht alle Dokumente werden für jede Anfrage bewertet → **Pooling!**)

Das Pooling-Verfahren

Bei großen Evaluationskorpora/-kampagnen kann nicht jedes Dokument für jede Anfrage bewertet werden, es wird daher mit dem sog. **Pooling-Verfahren** gearbeitet:

- Jeder Teilnehmer/jedes System sendet **seine Top k** (z.B. k=20)
- Die Organisatoren tragen alle **Top k**-Ergebnisse zusammen (der Pool) und merken sich jeweils, wer welches Ergebnis geliefert hat
- **Dubletten** werden zusammengeführt
- Nun wird der **Pool Relevanz-bewertet** (in zufälliger Reihenfolge)
- Für jeden Teilnehmer/jedes System ist nun jedes Dokument bewertet

Das Pooling-Verfahren

System A

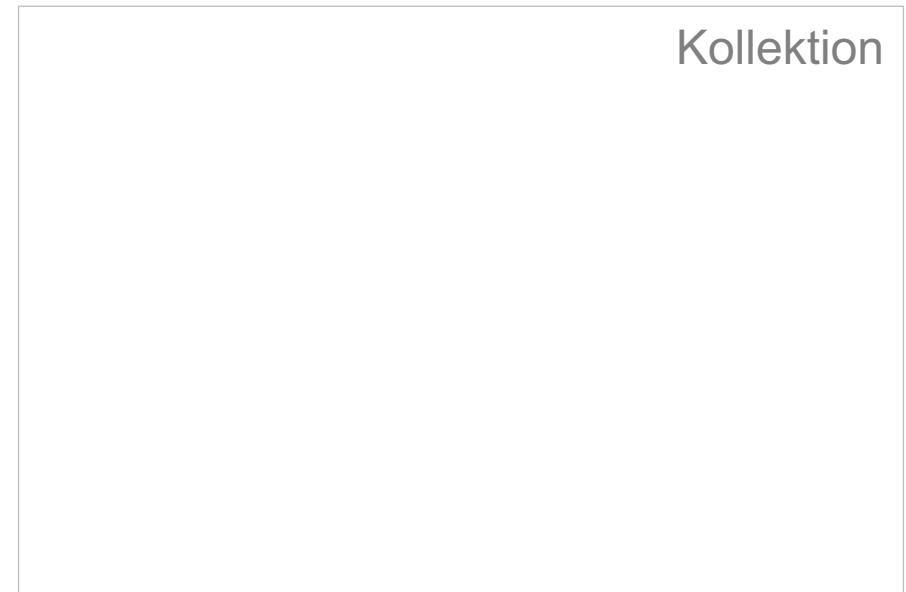


10

20

30

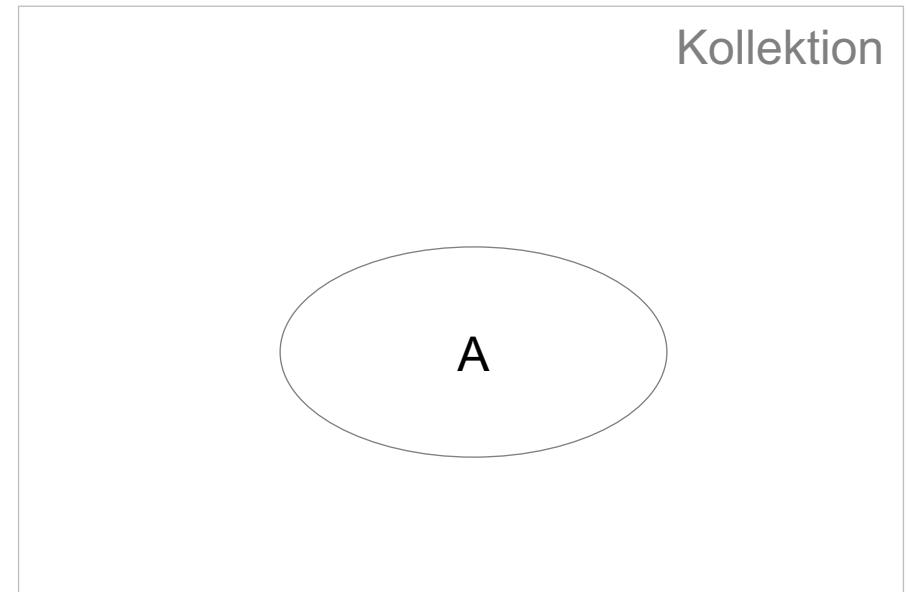
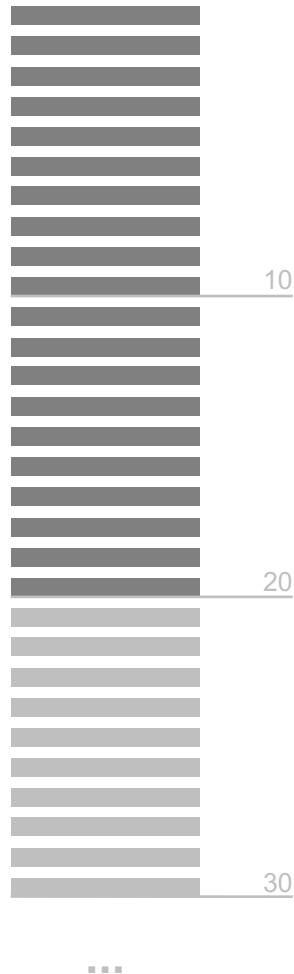
...



Kollektion

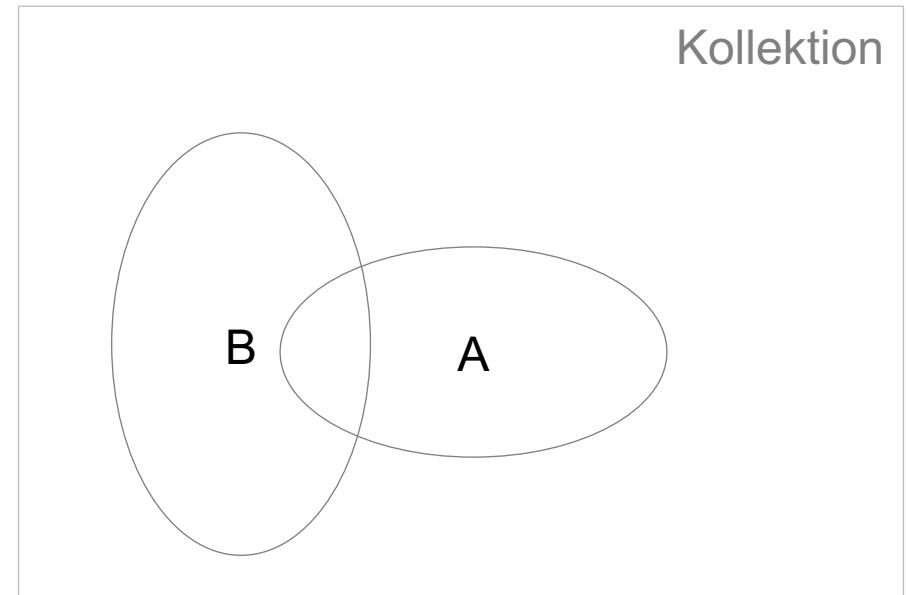
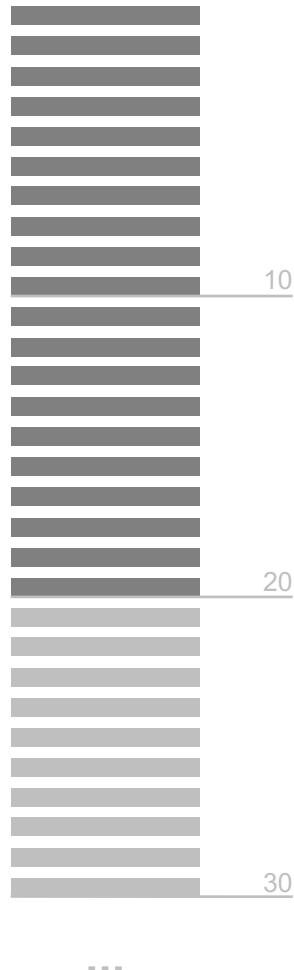
Das Pooling-Verfahren

System A



Das Pooling-Verfahren

System B

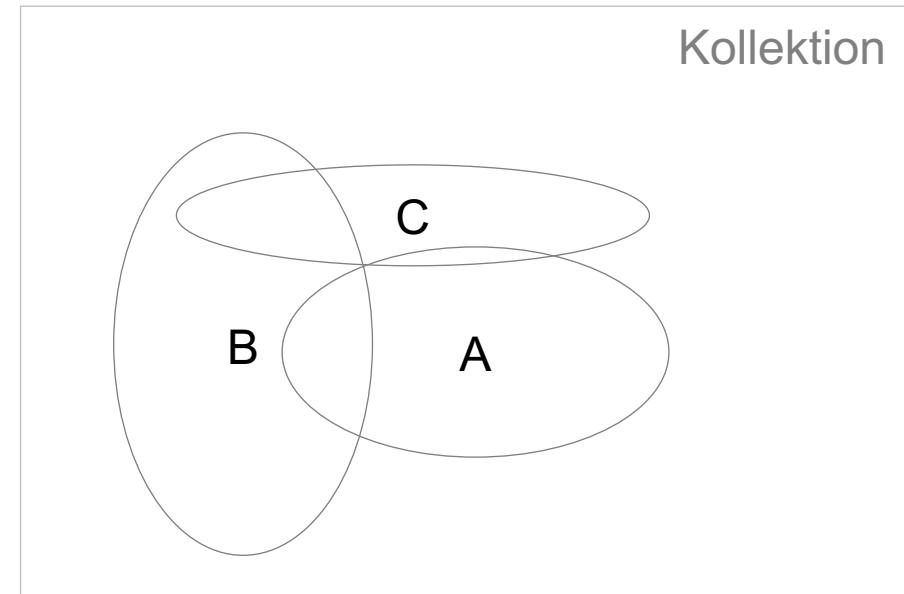


Das Pooling-Verfahren

System C



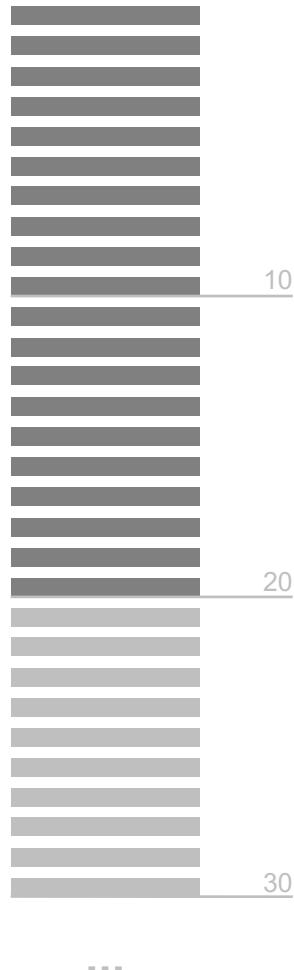
$k = 20$



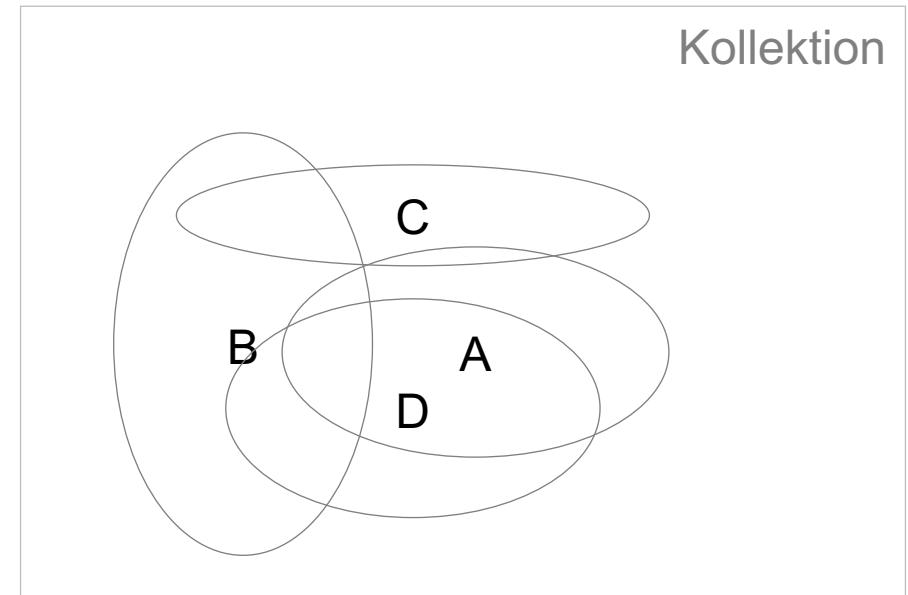
...

Das Pooling-Verfahren

System D

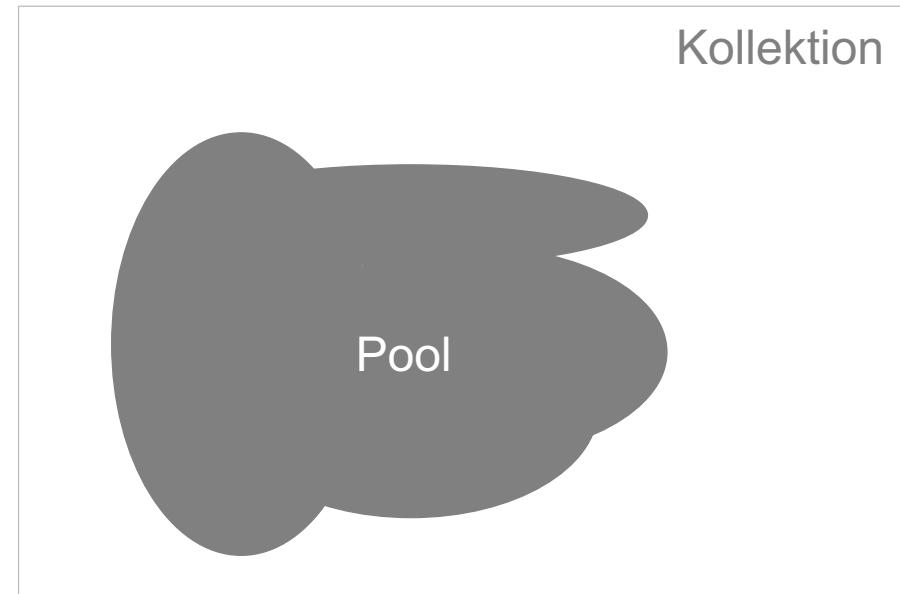


$k = 20$



Das Pooling-Verfahren

Wie viele Dokumente
müssen die Assessoren
max. bewerten?

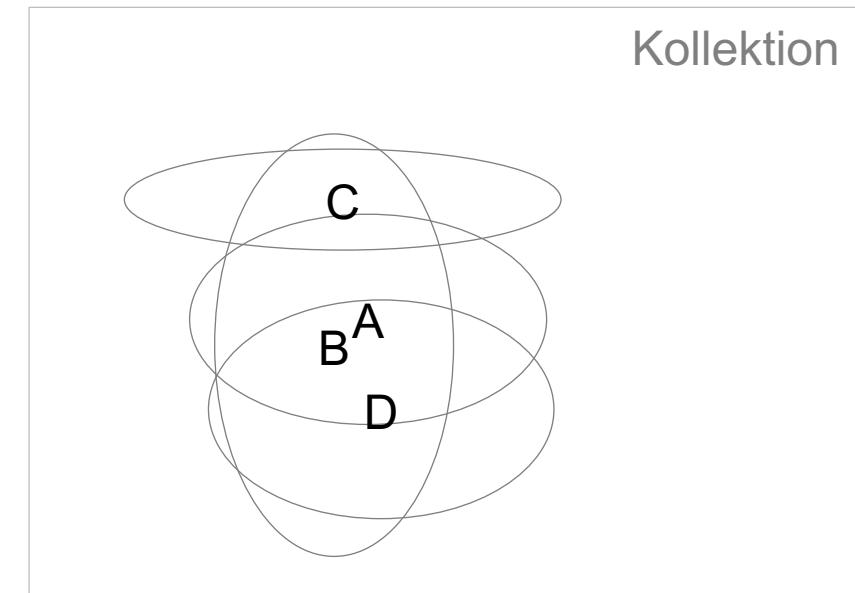
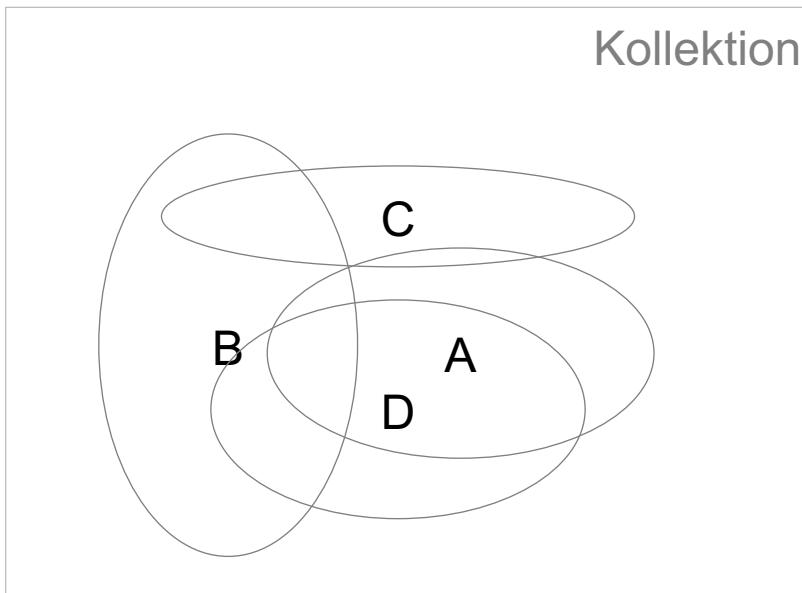


Die Annahme ist, dass durch die **Top k-Ergebnisse ein Großteil** der relevanten Dokumente **abgedeckt** wurde!

- Das ist aber natürlich nur ein Näherungswert...

Das Pooling-Verfahren

Welche System-Auswahl sollte beim Pooling bevorzugt werden?





The screenshot shows a web browser window for the NIST TREC-COVID website. The address bar displays "ir.nist.gov/covidSubmit/index.html". The page header includes the NIST logo, a back arrow, and a search icon. The main navigation menu at the top right contains links for "About", "Guidelines", "Data", "Tools", "Archive", and "Sign up!". Below the menu, the page title "TREC-COVID: Building a Pandemic Retrieval Test Collection" is displayed. A large section titled "TREC-COVID" discusses the challenge's purpose and impact. A prominent red banner in the center states: "TREC-COVID ist eine aktuelle, weltweite Retrieval-Kampagne, bei der es darum geht, Suchsysteme für aktuelle Daten zu COVID-19 zu bauen und zu evaluieren." At the bottom, two blue buttons offer "Learn more »" and "Round 2 Guidelines »".

< NIST

TREC-COVID: Building a Pandemic Retrieval Test Collection

About Guidelines Data Tools Archive Sign up!

TREC-COVID

Researchers, clinicians, and policy makers involved with the response to COVID-19 are constantly searching for reliable information on the virus and its impact. This presents a unique opportunity for the information retrieval (IR) and text processing communities to contribute to the response to this pandemic, as well as to study methods for quickly standing up information systems for similar future events. The results of the TREC-COVID Challenge will identify answers for some of today's questions while building infrastructure to improve tomorrow's search systems.

TREC-COVID ist eine aktuelle, weltweite Retrieval-Kampagne, bei der es darum geht, Suchsysteme für aktuelle Daten zu COVID-19 zu bauen und zu evaluieren.

annotation, known as relevance judgments, will then be used to score the submitted runs. After all rounds are complete, the final document and topic sets together with the cumulative relevance judgments will comprise a COVID test collection. The incremental nature of the collection will support research on search systems for dynamic environments.

[Learn more »](#)

[Round 2 Guidelines »](#)

»TH Köln Intern: Login 

Deutsch English

Technology Arts Sciences
TH Köln

Startseite Studium Forschung Weiterbildung Hochschule Internationales

Suchen 

Hochschule / Aktuelles / Nachrichten

- ◀ Hochschule
- ▼ Aktuelles
- ▶ Nachrichten
- Termine
- Standortentwicklung
- Presse und Kommunikation
- Amtliche Mitteilungen

Wo?

» Institut für
Informationswissenschaft

Ansprechpartner

 Prof. Dr. Philipp Schaer

@ philipp.schaer@th-koeln.de

TREC-COVID Challenge: Forschung zur zuverlässigen Informationssuche über COVID-19



In der aktuellen Pandemielage sind Forschende, medizinisches Fachpersonal und politische Entscheidungsträger ständig auf der Suche nach zuverlässigen Informationen über das Virus. Die konstante Flut an neuen Informationen muss daher durchsuchbar und handhabbar sein. Forschende des Information Retrieval (IR) und der automatischen

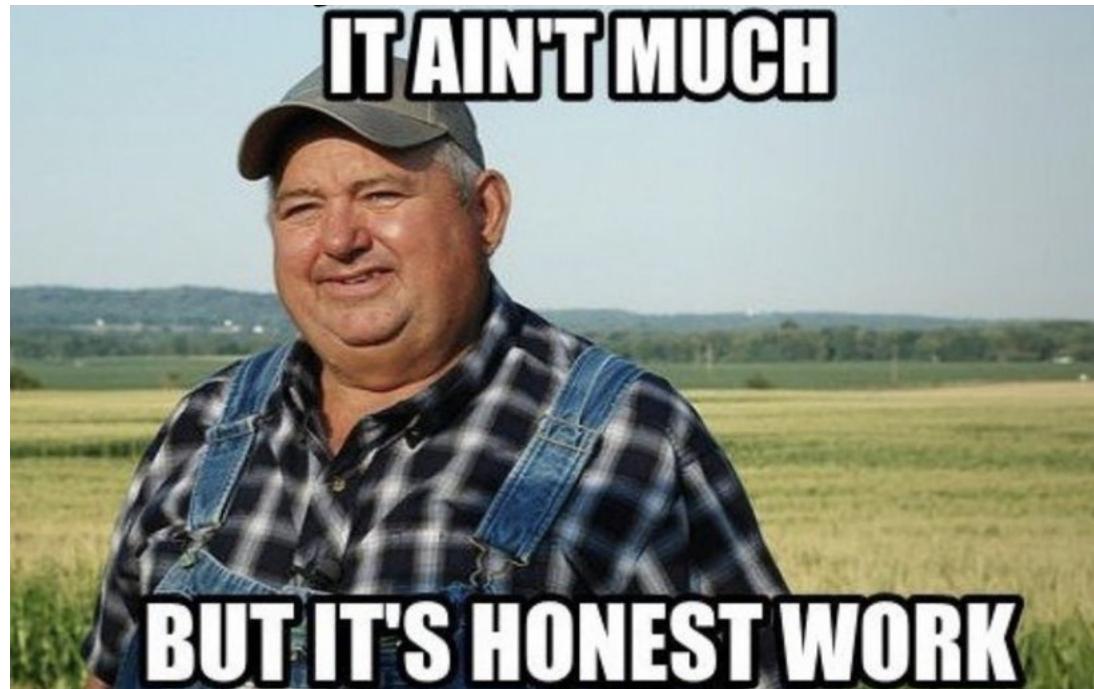
Textverarbeitung können hier einen wertvollen Beitrag leisten.

Zuverlässige Informationsstrukturen aufbauen

Es gilt nun, wissenschaftliche Informationssysteme so weiterzuentwickeln, dass sie mit den sich ständig ändernden Ereignissen und Erkenntnissen umgehen können. Hierzu organisiert das NIST, das US-amerikanische National Institute of Standards and Technology, zusammen mit dem Allen Institute for Artificial Intelligence die sogenannte TREC-COVID Challenge. Das Ziel dieser Challenge ist der Aufbau und die wissenschaftliche Bewertung von Infrastrukturen und Systemen zur Unterstützung der Su relevanten Informationen zu COVID-19. Im Rahmen dieser internationalen Challenge entwickeln Forschungsgruppen innovative Such- und Retrievalverfahren, die für eine zuverlässige Informationsversorgung genutzt werden können. Die Hoffnung ist, dass die Ergebnisse der Tf Challenge Antworten auf einige der Fragen von heute liefern und gleichzeitig eine Infrastruktur Verbesserung der Suchsysteme von morgen aufbauen.

Diese Webseite verwendet Cookies.
» Datenschutzhinweis

OK



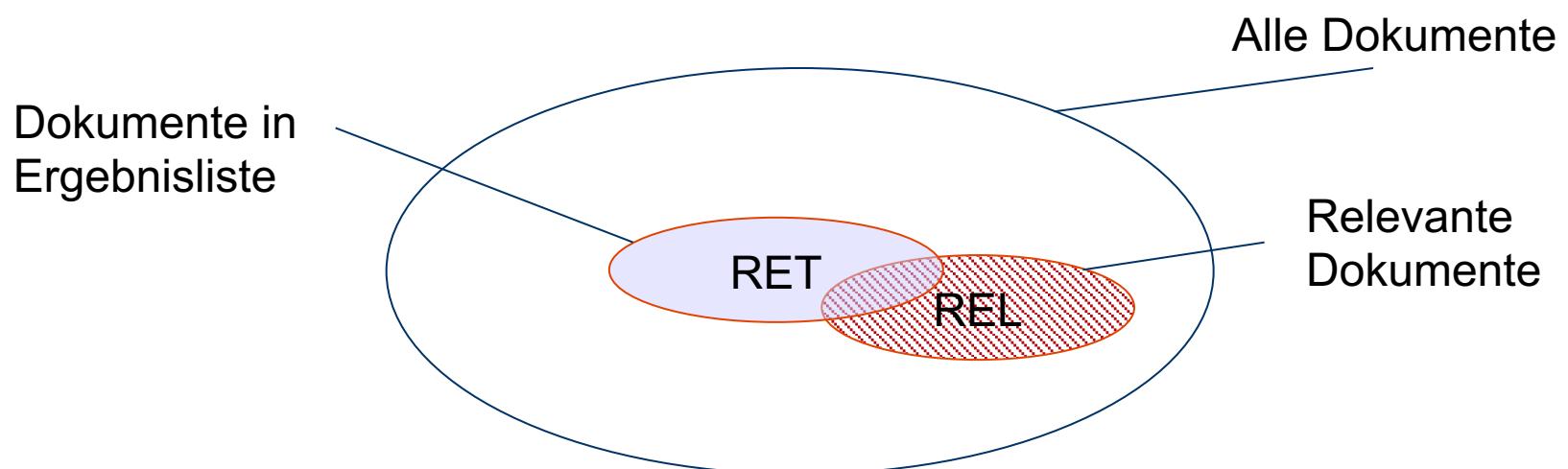
Maßzahlen für die Evaluation

- **Precision** (Treffergenauigkeit)

$$\mathcal{P} = \frac{|\text{RET} \cap \text{REL}|}{|\text{RET}|}$$

- **Recall** (Treffervollständigkeit)

$$\mathcal{R} = \frac{|\text{RET} \cap \text{REL}|}{|\text{REL}|}$$



Precision und Recall: Ein Beispiel

	Relevant	Nicht relevant
Gefunden	30	12
Nicht gefunden	14	44

Precision $P = 30 / (30 + 12) \approx 0,714$

Recall $R = 30 / (30 + 14) \approx 0,681$

Precision / Recall → F-Measure

Allgemein: „**Genauigkeit**“ vs. „**Vollständigkeit**“

Üblicher **Trade-off**: Recall hoch → Precision niedrig

- Auf welche Maßzahl sollte optimiert werden?
- Kombinierte Maßzahl: **F-Measure**

$$F = 2 * \frac{P * R}{P + R}$$

Precision und Recall im Ranked Retrieval

Bisher: Precision und Recall für **gesamte Dokumentmenge**, aber im Ranked Retrieval meist gar nicht bekannt.



- *Wieso kennen wir die gesamte Dokumentmenge nicht?*
- *Ist das ein Problem für die Berechnung der Precision?*
- *Ist das ein Problem für die Berechnung des Recall?*

Precision und Recall im Ranked Retrieval

Bei großen Dokumentmengen: **gerankte Liste!**

→ Die „relevantesten Dokumente“ sollten oben stehen!

- Precision / Recall gemessen nach einer festgelegten Anzahl von Dokumenten (z.B. nach 10 oder 20 Dokumenten)
- Interpolierte Precision: nach einem bestimmten Recall-Wert (z.B. nach 50% der relevanten Dokumente gesehen)

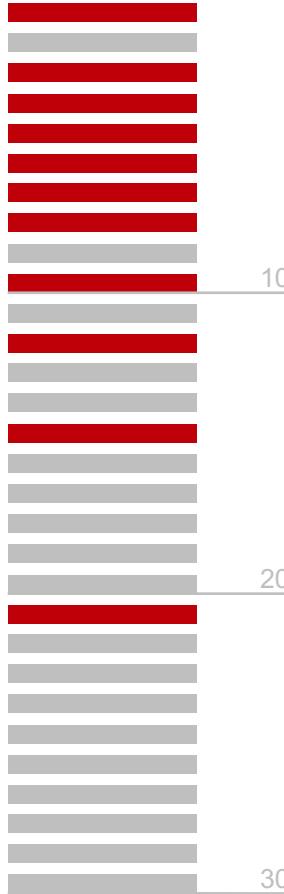
Precision-at-k

Wir wollen einheitliche Werte, die wir vergleichen können!

Precision könnte an festen Werten gemessen werden

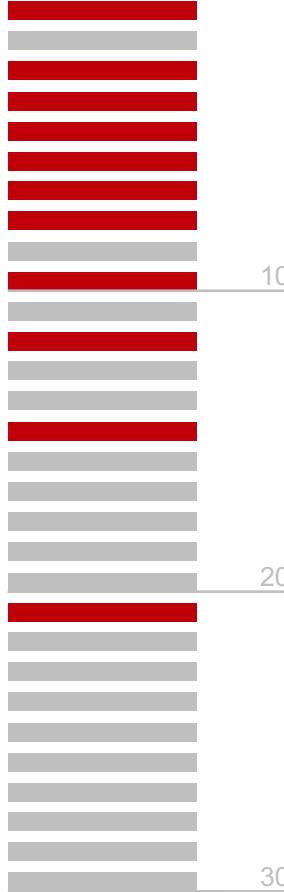
- **Precision-at-k:** Precision nach k Ergebnissen
- Sehr nah am **Nutzerbedürfnis**
(z.B. Precision-at-10: Treffer auf der ersten Ergebnisseite)
- Nachteil: Kann schlecht verglichen werden, da stark schwankend.

Precision-at-k



k	P@k
1	$1/1 = 1,0$
2	$1/2 = 0,5$
3	$2/3 = 0,67$
4	$3/4 = 0,75$
5	$4/5 = 0,8$
10	$8/10 = 0,8$
20	$10/20 = 0,5$
30	$11/30 = 0,37$

Precision-at-k und Recall-at-k



k	P@k	R@k
1	$1/1 = 1,0$	$1/20 = 0,05$
2	$1/2 = 0,5$	$1/20 = 0,05$
3	$2/3 = 0,67$	$2/20 = 0,1$
4	$3/4 = 0,75$	$3/20 = 0,15$
5	$4/5 = 0,8$	$4/20 = 0,2$
10	$8/10 = 0,8$	$8/20 = 0,4$
20	$10/20 = 0,5$	$10/20 = 0,5$
30	$11/30 = 0,37$	$11/20 = 0,55$

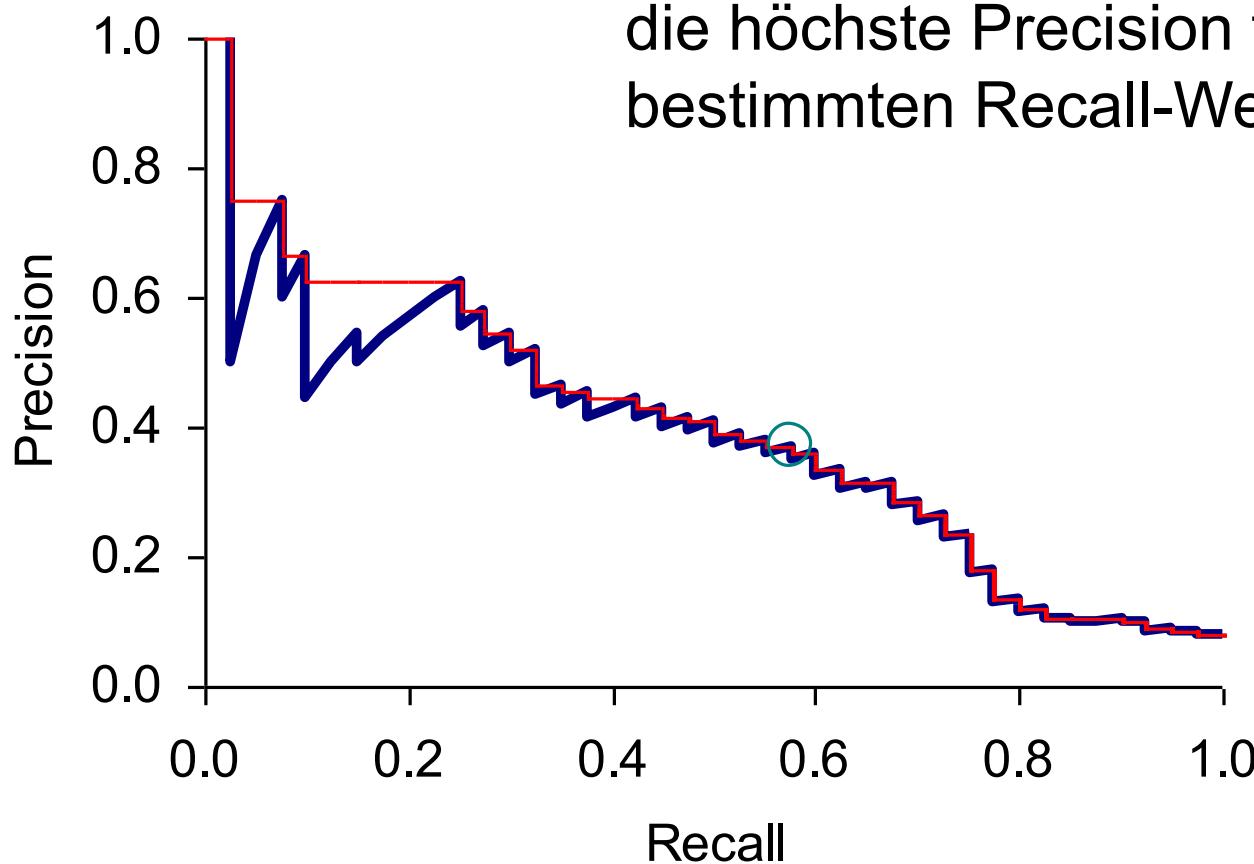
Wir gehen davon aus, dass es
20 relevante Dokumente gibt!

Precision-Recall-Graph

11-Punkt interpolierte Precision

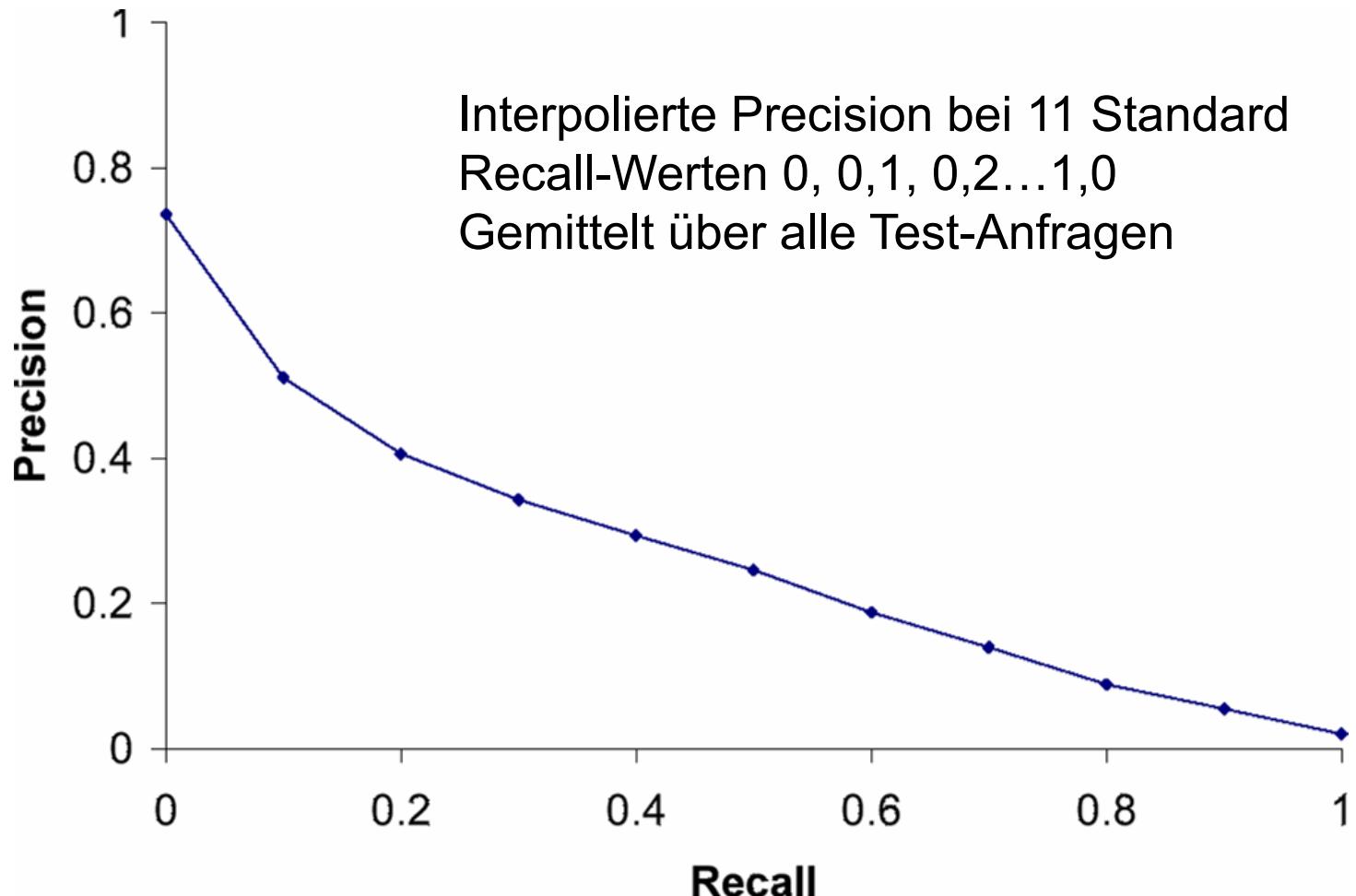
- Die Standardmaßzahl bei frühen TREC-Studien
- Es werden 11 Precision-Werte ermittelt für die Recall-Level 0, 0,1, 0,2...1 (in 0,1er-Schritten)
- Diese Werte können über Anfragen hinweg gemittelt werden

Precision-Recall-Kurve



→ Interpolierte Precision: immer die höchste Precision für einen bestimmten Recall-Wert

Average 11-point Precision-Recall-Graph



Average Precision (AP)

- **Standardwert** zum Vergleich
- **Robustheit** über verschiedenste Kollektionen
- **Ein Precision-Wert für alle Recall-Werte**
 - Für festgelegte Anzahl der Dokumente in Ergebnisliste (z.B. 1000)
 - Berechnung des Precision-Wertes nach jedem relevanten Dokument
 - Durchschnittswert über alle Precision Werte wird berechnet

Kann auch für mehrere Anfragen angewendet werden:

- Durchschnittswert über alle Anfragen wird berechnet
- Heißt dann **Mean Average Precision (MAP)**
- (eigentlich ein blöder Name...)

Beispiel für die AP-Berechnung

Zwei Systeme ranken Ergebnisse wie folgt:

(N = nicht-relevant, R = relevant; Sortierung von links nach rechts; es gibt nur zwei relevante Dokumente insgesamt)

- System 1: N R N N R
- System 2: R N R N N

Nach jedem relevanten Dokument wird nun die Precision bestimmt und anschließend durch die Anzahl der relevanten Dokumente geteilt.

- System 1: $(1/2 + 2/5) / 2 = 0,45$
- System 2: $(1/1 + 2/3) / 2 = 0,83$

Weitere Maßzahl: R-Precision

R-Precision = Precision nach der Anzahl von relevanten Dokumenten (normalisiert über Anzahl der relevanten Dokumente)

- R-Precision = 0,33
- „Precision an der Stelle 3“
- R-Precision soll einen faireren Vergleich ermöglichen, als immer P@k an einer fixen Stelle.

# Dok	Precision	Recall
1	1,0	0,33
2	0,5	0,33
3	0,33	0,33
4	0,5	0,66
5	0,6	1,0
6	0,5	1,0

Annahme: Es existieren nur **drei relevante Dokumente** für diese Anfrage in der Kollektion.

Außer Acht gelassen wurden...

Relevanz ist nicht zwangsläufig Nutzerzufriedenheit:

- „*For a web search engine, happy search users are those who find what they want. [...] Nevertheless, in general, we need to decide whether it is the end user's or the eCommerce site owner's happiness that we are trying to optimize. Usually, it is the store owner who is paying us.*“ (Manning et al. (2009), S. 155)

Mehrstufige Relevanzeinschätzungen

- nicht nur binär (relevanz / nicht-relevant)
- relevant, teilweise relevant, nicht-relevant etc.

Probleme mit Relevanz-Assessments

- Menschliche Relevanzbeurteiler (Judges, Rater, Assessoren)
- Sind „teuer“ (sprich: aufwendig, organisatorisch und zeitlich)
- Wie zuverlässig sind Assessoren?

Zusammenfassung IR-Evaluation

IR-Evaluationen mit Hilfe von Testkollektionen benötigen drei Komponenten

- Eine **gemeinsame Dokumentenmenge**
- Eine **Menge von Anfragen**, sogenannte **Topics**
(min. 25 besser 50)
- **Relevanzbewertungen**, gewonnen im Rahmen eines **Pooling**

Es lassen sich daraus unterschiedliche Maßzahlen bestimmen

- Precision („Genauigkeit“), P@n
- Recall („Vollständigkeit“), R@n
- kombinierte Maße wie F-Measure oder AP/MAP, Rprec, ...