

PD Thomas Mandl
Informationswissenschaft
Universität Hildesheim
mandl@uni-hildesheim.de

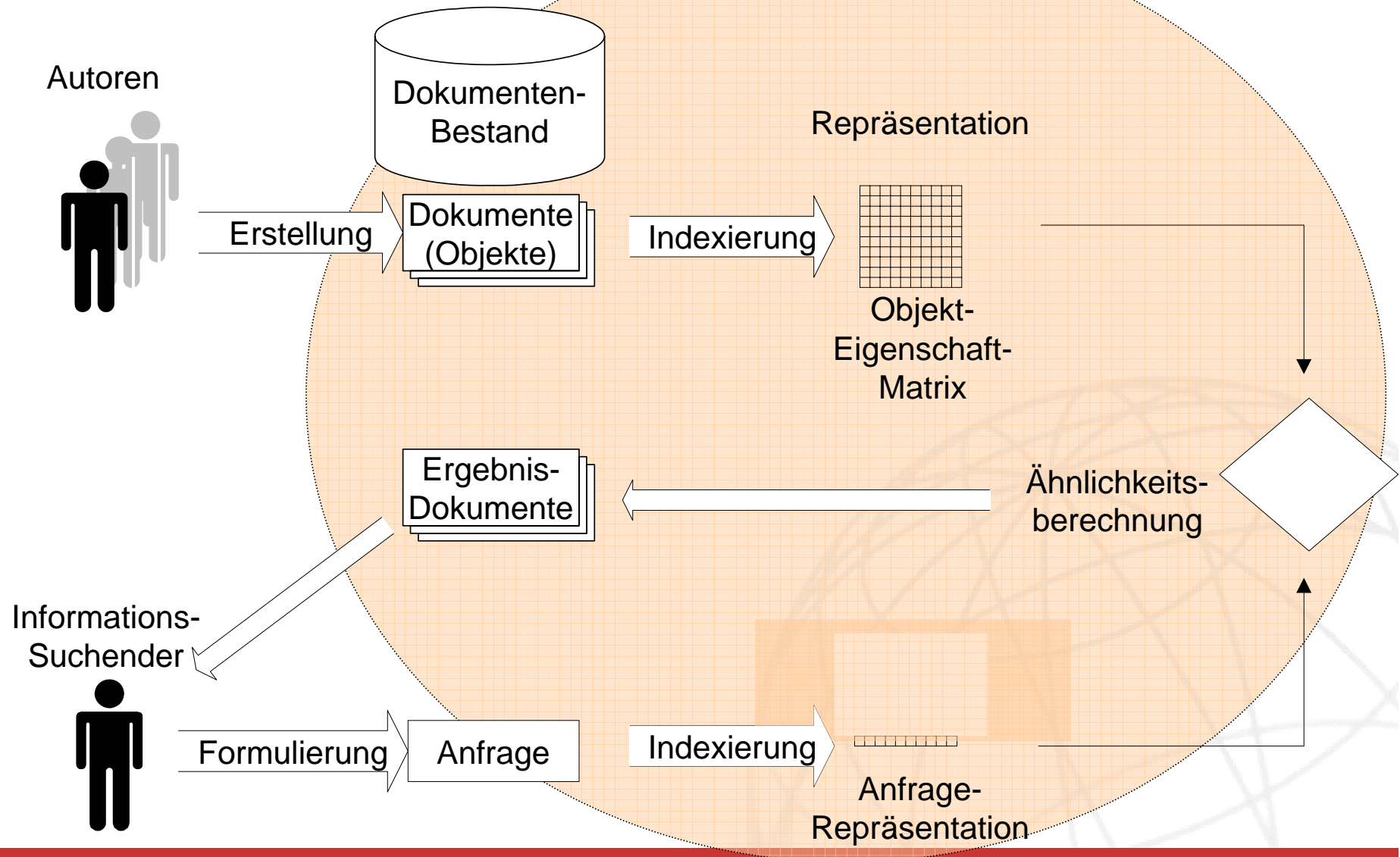
Workshop Realistische
Evaluierungsansätze für
P2PIR-Systeme
Leipzig 29.2.2008

Evaluierung von Informationssuche und P2P Retrieval



- P2P ?
- Evaluierung IR
 - Evaluierungsinitiativen
- Vier Thesen zur P2P Evaluierung

Evaluierung IR



„There must be some fundamental understanding of what it means to be good and what it means to be better“
(Bollmann/Cherniavsky 1983,3)



Cranfield-Paradigma der Evaluierung im Information Retrieval

- Objektive Relevanz wird von neutralem Beobachter beurteilt
- Beziehung zwischen dem geäußerten Informationswunsch und dem Dokument
- Keine individuelle und subjektive Relevanzbewertung im Kontext
- Bis heute Testaufbau aller Evaluierungsinitiativen im Information Retrieval (TREC, CLEF, NTCIR, INEX, ...)

- „TREC is a new ballgame for IR research and development“ (Sparck Jones 1994)
- Evaluierungsinitiative des National Institute of Standards and Technology (NIST) in den USA
- 1992: TREC-1 (Proceedings 1993)



NIST

National Institute of Standards and Technology

Cross-Language Evaluation Forum

Forschung zu cross- und multi-lingualen Information Retrieval Systemen



EU Förderung: DELOS NoE for Digital Libraries

Testumgebung

Systementwicklung

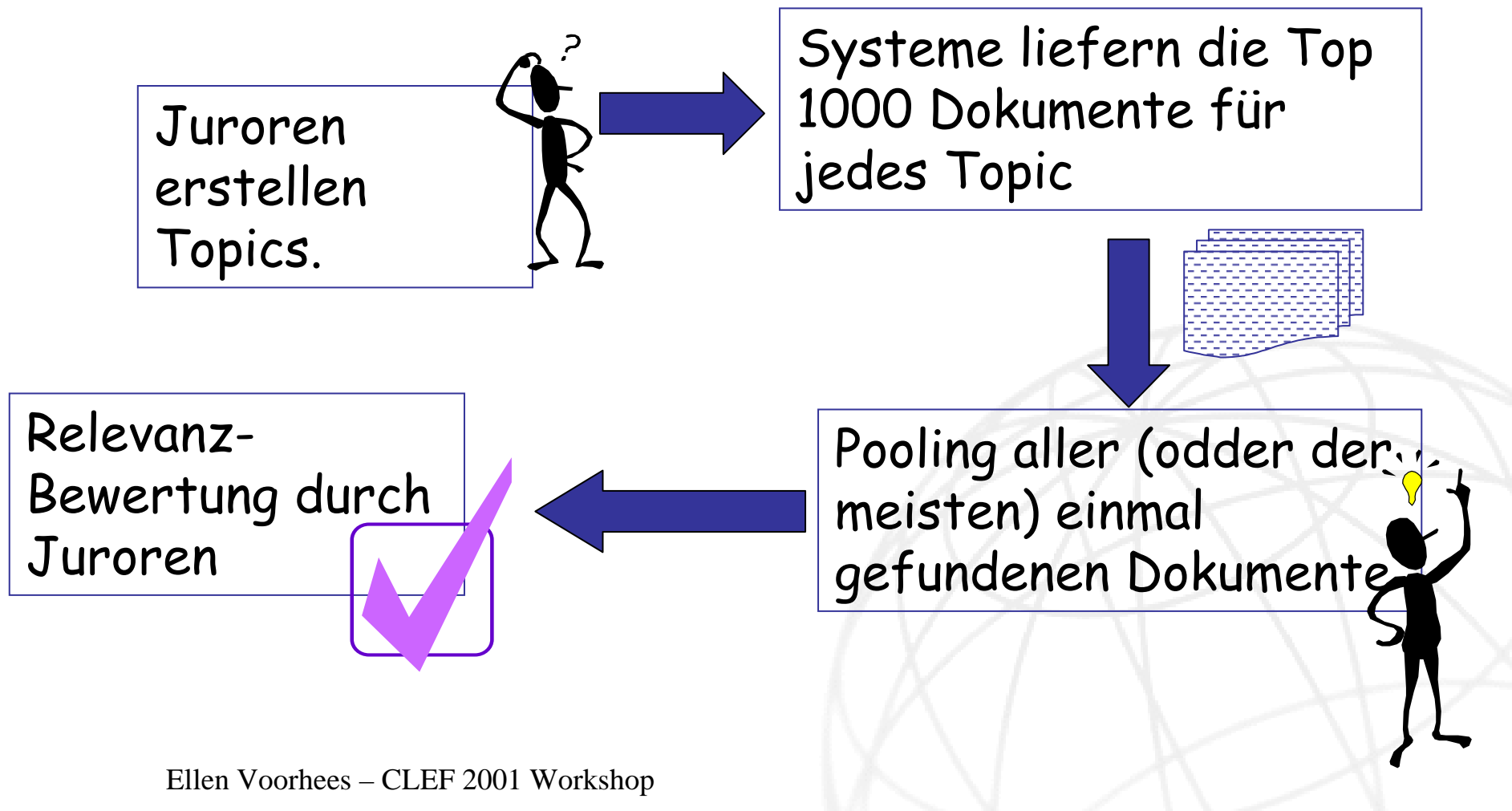
Benchmarks

Evaluierungsforschung

Mandl et al. @ CLEF 2003 - 2006

Robust CLEF
GeoCLEF

Pooling Methode



Ellen Voorhees – CLEF 2001 Workshop



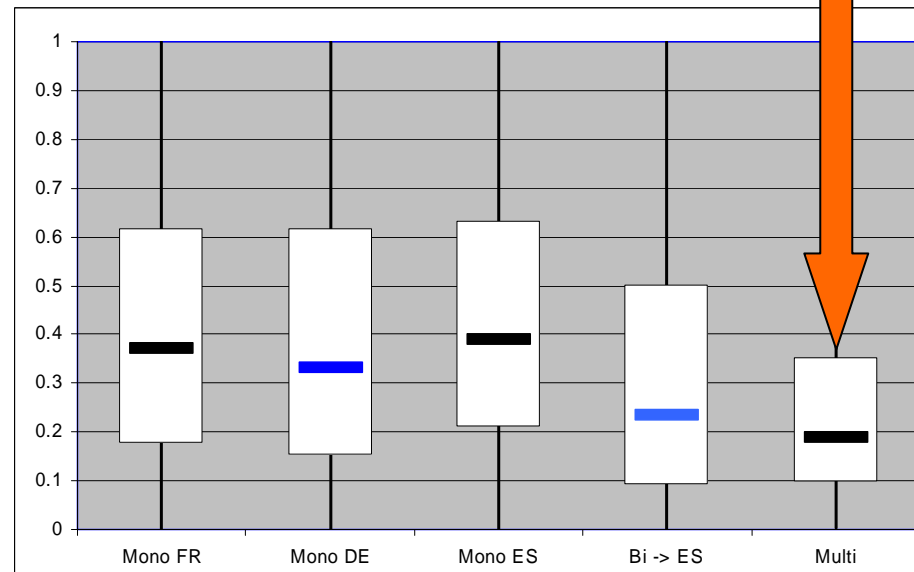
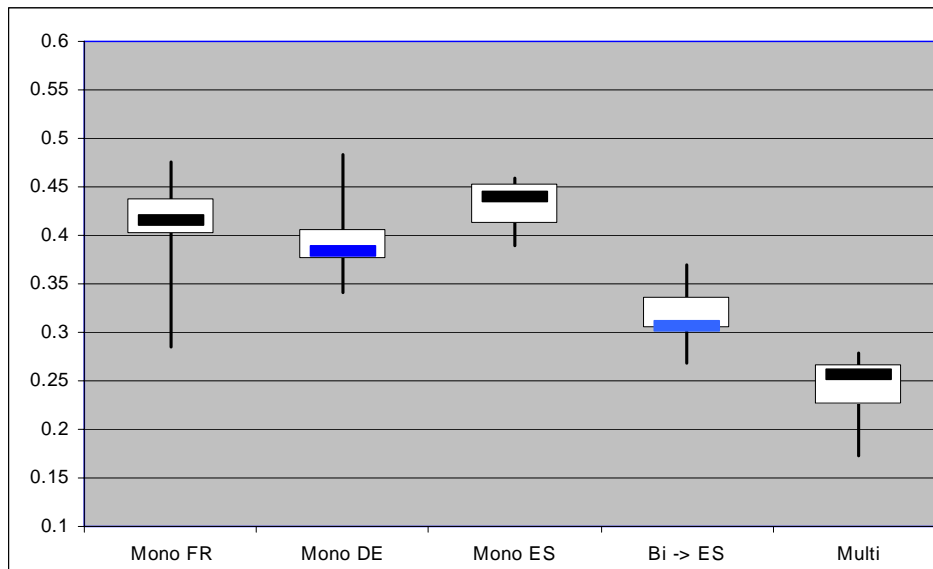
Quelle: TREC presentation slide No. 15

Text REtrieval Conference (TREC)

**These: Der Aufbau des
Benchmarks ist wichtig!**

(Nimm mehr Topics!)

Das lässt uns nicht einschlafen!



- Reichen (die üblichen) 50 Topics aus, um die Systeme zu vergleichen?
 - Zwischen zwei Systemen muss ein gewisser Unterschied bestehen, um statistisch sicher zu sein, dass eines besser ist als das andere
 - Ab 50 Topics liegt der Unterschied unter 5%
 - Teilweise auch deutlich unter 5% (absolut)

Sanderson & Zobel 2005

- „Flache“ Relevanzbewertung (wesentlich weniger Dokumente) bei deutlich mehr Topics führt zu trennschärferen Ergebnissen



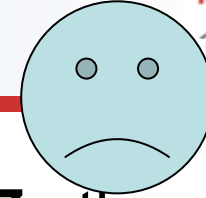
These: Die Wahl des Effektivitäts- maßes ist wichtig!

**(Stell dir den Benutzer
wenigstens vor!)**

- Es gibt mehr als ein Topic
 - Wie fassen wir die Ergebnisse zusammen?
- System Evaluierung
 - Wie gut sind die Ergebnislisten?
- Benutzerzentrierte Evaluierung
 - Wie zufrieden ist der Benutzer?

Der Benutzer sieht die Perspektive der
Evaluierung (=MAP) nie!

Sondern nur die Leistung des Systems für
seine Anfragen

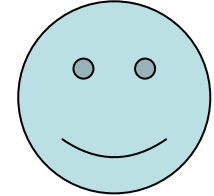


«The **unhappy customer**, on average, **will tell 27** other people about their experience. With the use of the internet, whether web pages or e-mail, that number can increase to the thousands ...»

«**Dissatisfied customers** tell an average of **ten** other people about their bad experience. Twelve percent tell up to twenty people.»

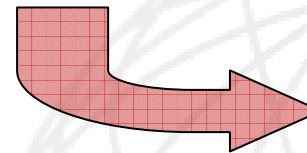
→ ***Bad news travels fast.***

On the other hand, satisfied customers will tell an average of *five* people about their positive experience.



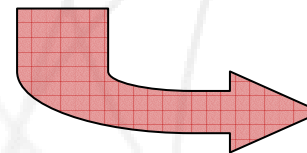
→ ***Good news travels somewhat slower***

Your system should produce less bad news!



improve on
hard topics

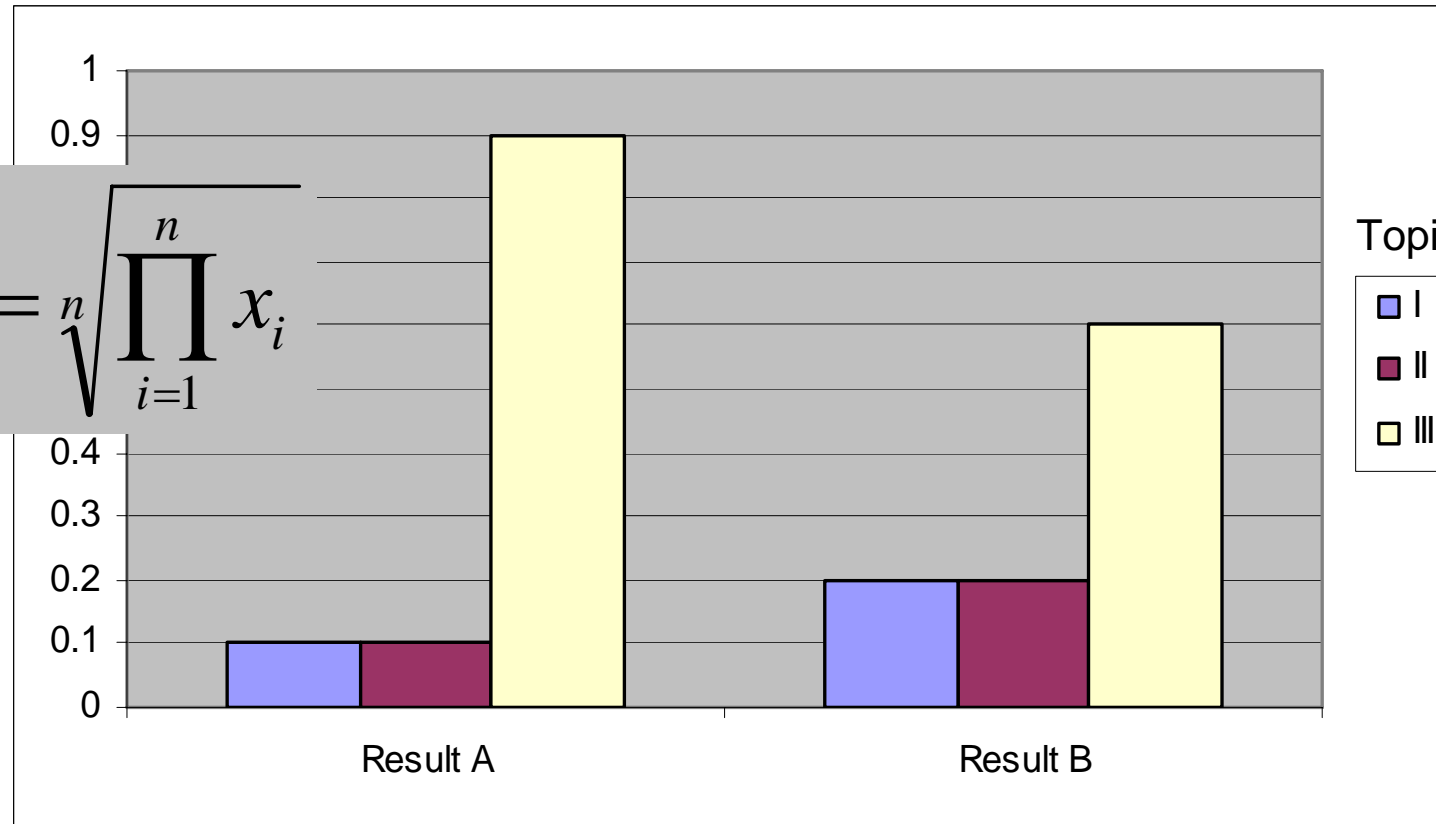
Don't worry too much about the good news



best topics

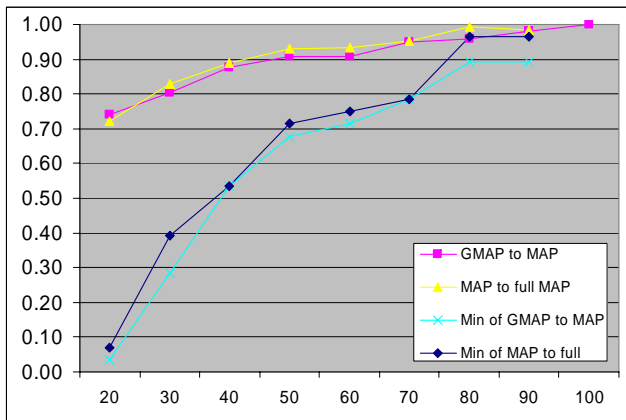
Welches System ist besser?

$$geoAve = \sqrt[n]{\prod_{i=1}^n x_i}$$

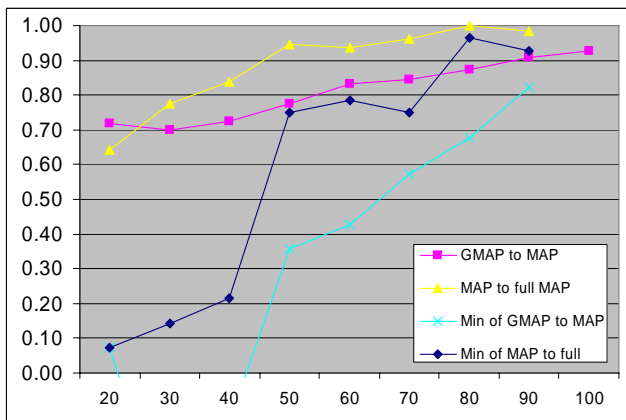


GeoAve	A	0.21	GeoAve	B	0.29
MAP	A	0.37	MAP	B	0.33

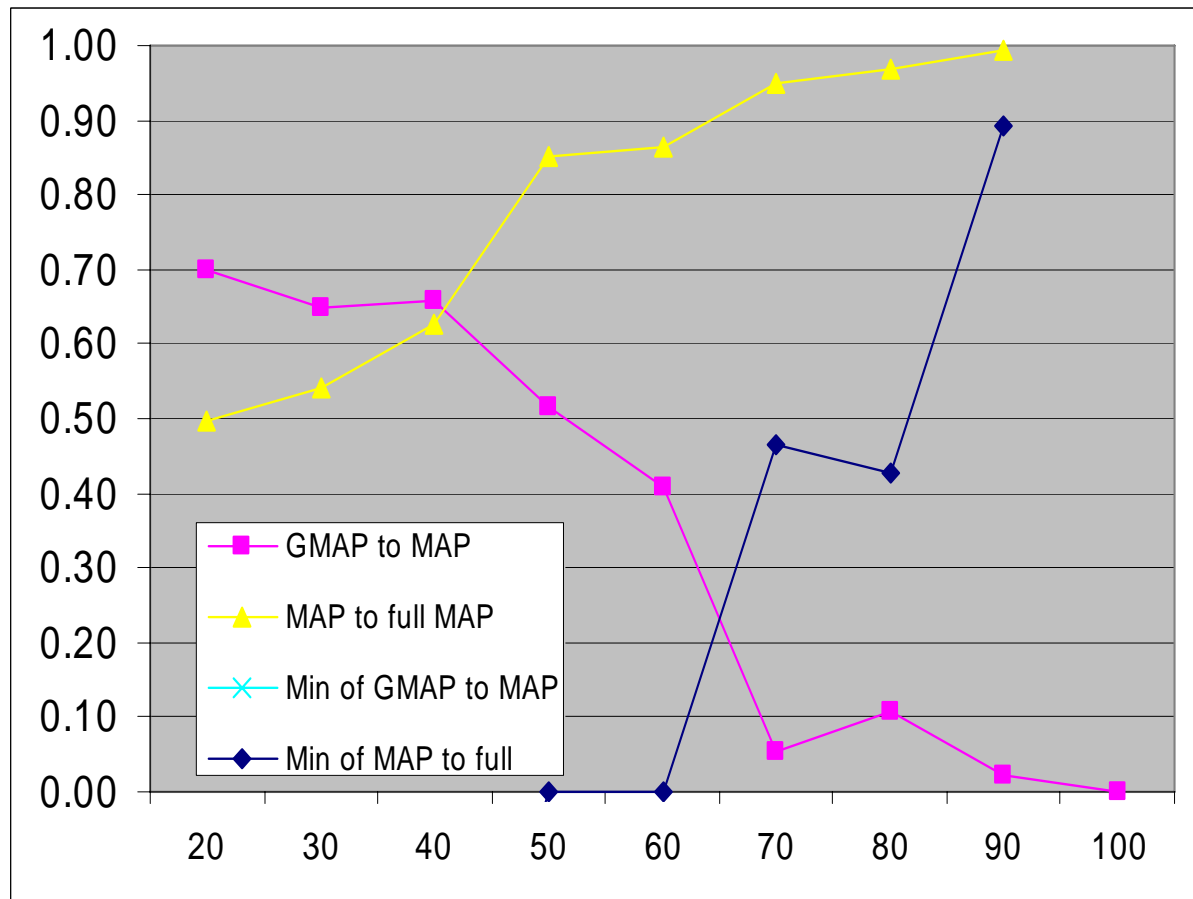
Untermengen von Topics



Monolingual Dutch



Monolingual French



Multilingual task

- Robuste Evaluierung misst etwas anderes
- Beim multilingualen Retrieval weicht dieses andere vom Standards (MAP) ab
- Mehr „schwierige“ Aufgaben -> Robuste Maße ändern viel
- und Robustheit ist nicht nur ein Thema für die Architektur von P2P

These:
**Maße für die Architektur
sind wichtig!**

**Berechne, was die
Robustheit des
Netzes bringt!**

A faint, light gray wireframe globe is visible in the background on the right side of the slide.

- Effizienz-Nachteile gegenüber monolithischen IR (non-distributed IR as upper bound, Liu & Callan 2004)
- Vorteile durch P2P-Architektur
 - Netzwerk-Verkehr (Müller et al. 2005), Robustheit, Balance der Peers, Ressourcen der Peers, etc.
 - Quantifizieren (analog Recall-Precision-Kurve)
- Informationsmanagement-Thema
- Kompromiss suchen zwischen Such-Effektivität und Architektur-Bewertung
 - „trade-off between quality of retrieval and resources consumed“ (Zhou et al. 2004)

These: Das Benutzermodell ist wichtig!

Wer sucht warum?

- Benutzermodell

- TREC: „Dedicated Searcher“

- Recall-orientiert (ein Maß in Zhou et al. 2004)
 - Recall-Precision Kurve als „Upper Bound“ (Liu & Callan 2004)

- Passt das wirklich für P2P ?

- MRR (auch ein Maß in Zhou et al. 2004)
 - *Siehe Robustheit*
 - *Siehe flache Relevanzbewertung*



*Vielen Dank für Ihre
Aufmerksamkeit*