max planck institut informatik

Dealing with Incomplete Judgments in Cascade Measures

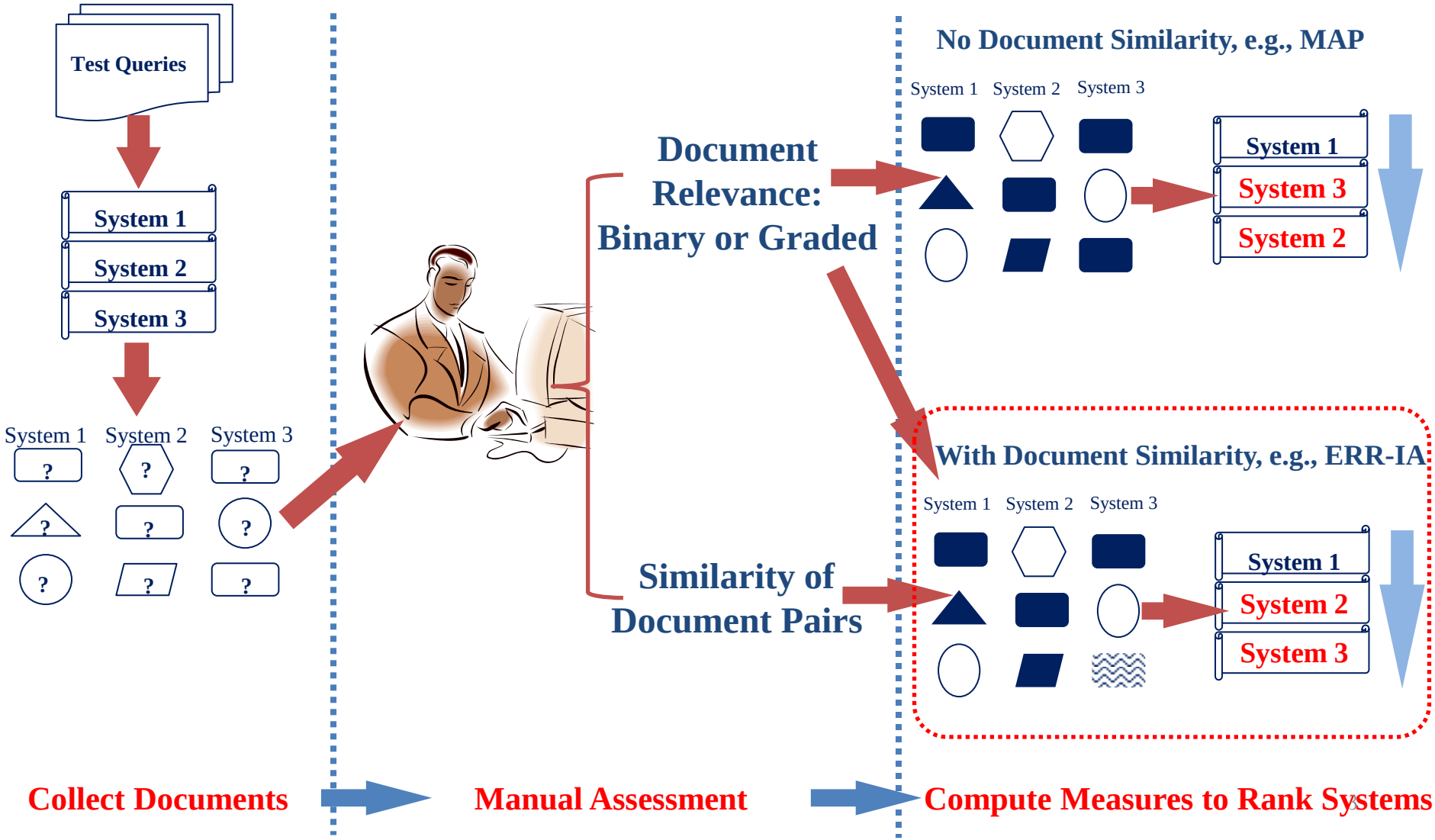
Kai Hui[1], Klaus Berberich[1], Ida Mele[2]
[1]Max Planck Institute for Informatics
{khui, kberberi}@mpi-inf.mpg.de
[2] Università della Svizzera italiana (USI)
Ida.mele@uci.ch

The 3rd ACM International Conference on the Theory of Information Retrieval

Background
Consume Incomplete Judgments

- Manual judgments in the evaluation for diversification are especially expensive, due to the judgments are based on individual subtopics
- Selectively label fewer documents or reuse existing judgments is desirable, but both lead to incomplete judgments
- Established measures require complete judgments for an accurate evaluation, and only consume manual labels (e.g., -1, 0, 1, 2…)

Background
IR Evaluation Pipeline Revisit



Background
Revisit the Evaluation for Diversification

- **Rain Man**

Subtopic 1. Where can I watch the full “Rain Main” movie online?
Subtopic 2. Find information about the real person on which the Rain Man movie is based.
Subtopic 3. Find movie reviews of “Rain Man”.
Subtopic 4. Find quotes from the “Rain Man” movie.
Subtopic 5. Find the lyrics to Eminem’s “Rain Man”.

- Manual assessors provide judgments for individual subtopics, such as document d_1 is judged as relevant (label 1) to subtopic 1.

Motivation
Novel Measures on Sparse Judgments

- Existing works successfully catered for incomplete judgments when few are missing (less than 50%)
- We attempt to develop measures that can accurately evaluate on more sparse judgments, namely, when missing more than 50% judgments
- Fully utilize the judgments by employing the content of documents beyond the labels

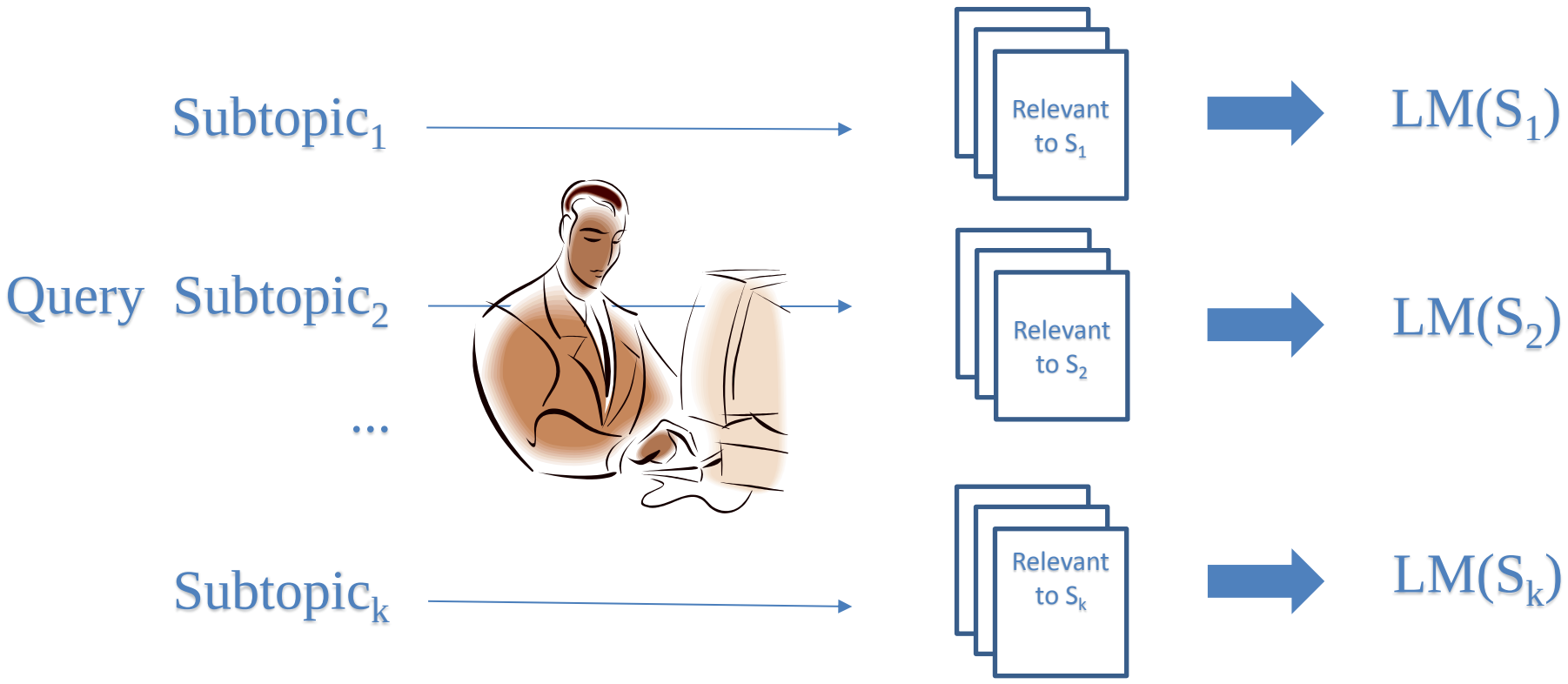
Objective

Develop novel evaluation measures for diversification, consuming sparse judgments.

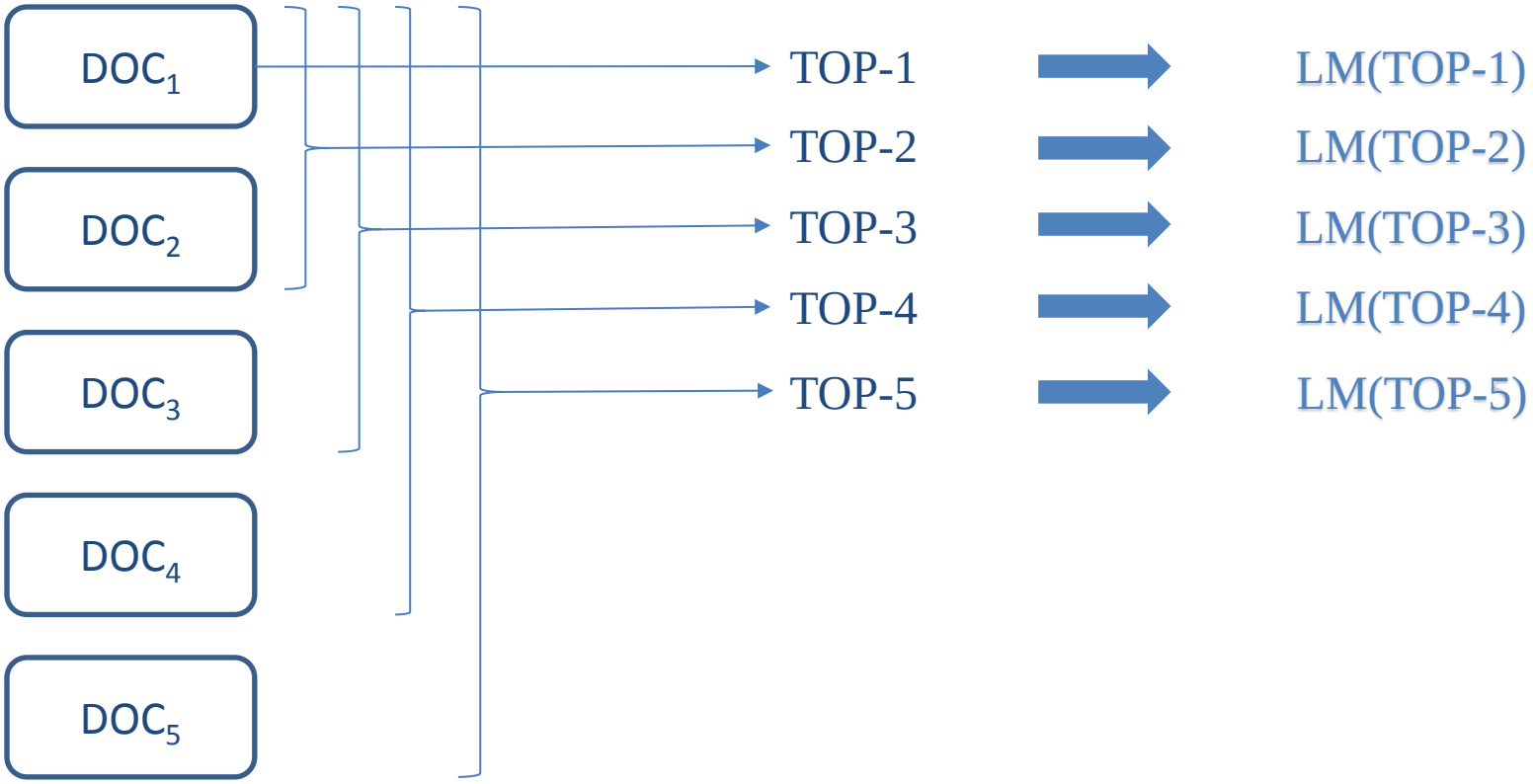
Method Overview

- Represent each subtopic with a language model (LM) based on manual judgments
- Represent a ranking by a series of language models (LM) estimated for top-1, top-2, …, top-k documents
- Compute the divergence between individual language models over each position for individual subtopics
- Convert the matrix of divergence into a scalar

Method
(1) Represent Subtopics with LMs



Method
(2) Represent Ranking with Cascade of LMs



Method
(3) KL-Divergence for Evaluation

	LM(TOP-1)	LM(TOP-2)	LM(TOP-3)	LM(TOP-4)	…	LM(TOP-k)
$LM(S_1)$						
$LM(S_2)$						
…						
$LM(S_i)$						

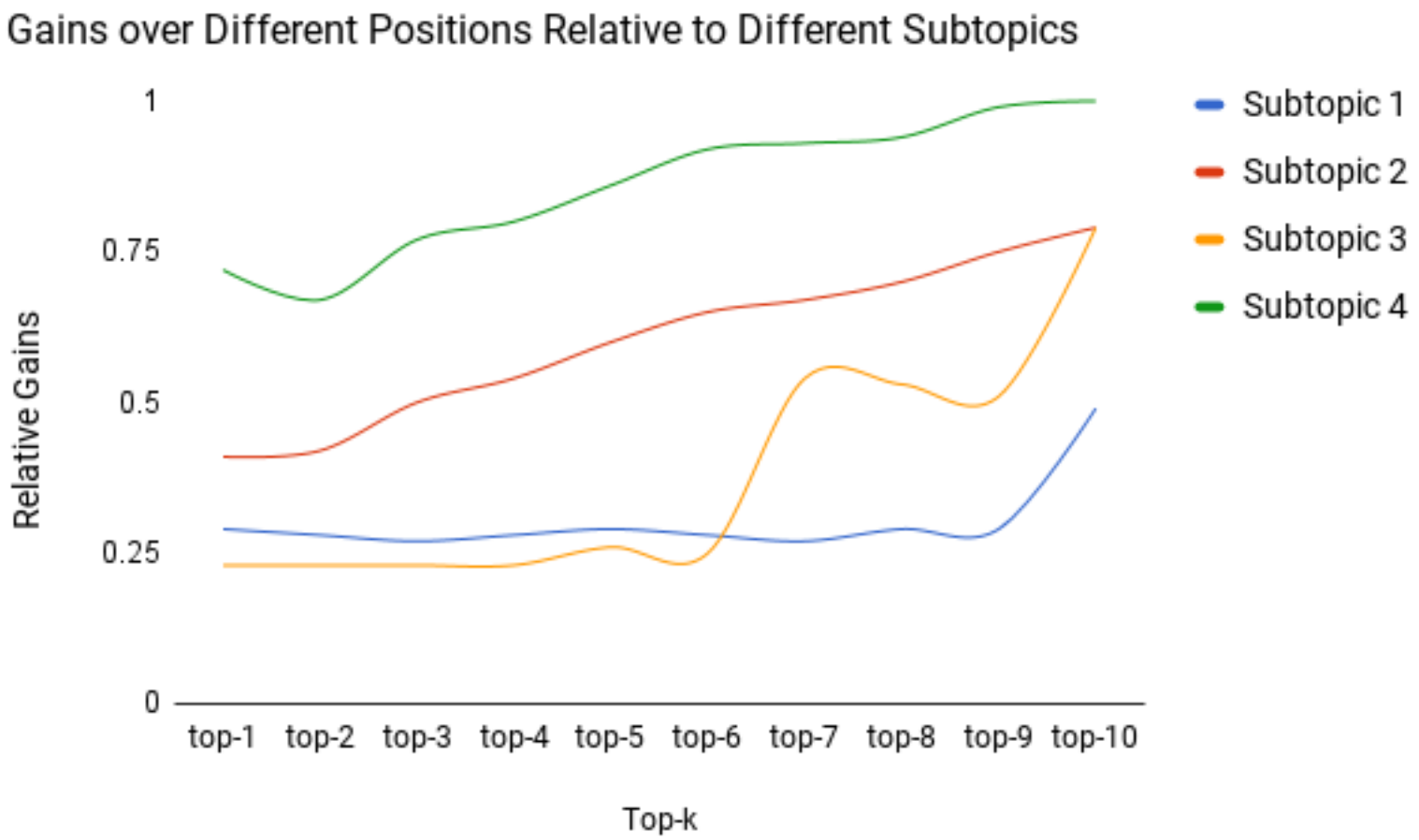
$$g(i,k)=max\left(0,1-\frac{KLD(LM(\mathrm{Si})||LM(\mathrm{TOP-k}))}{KLD(LM(\mathrm{Si})||LM(\mathrm{D}))}\right)$$

KLD between the LM of TOP-k and the LM of individual subtopics

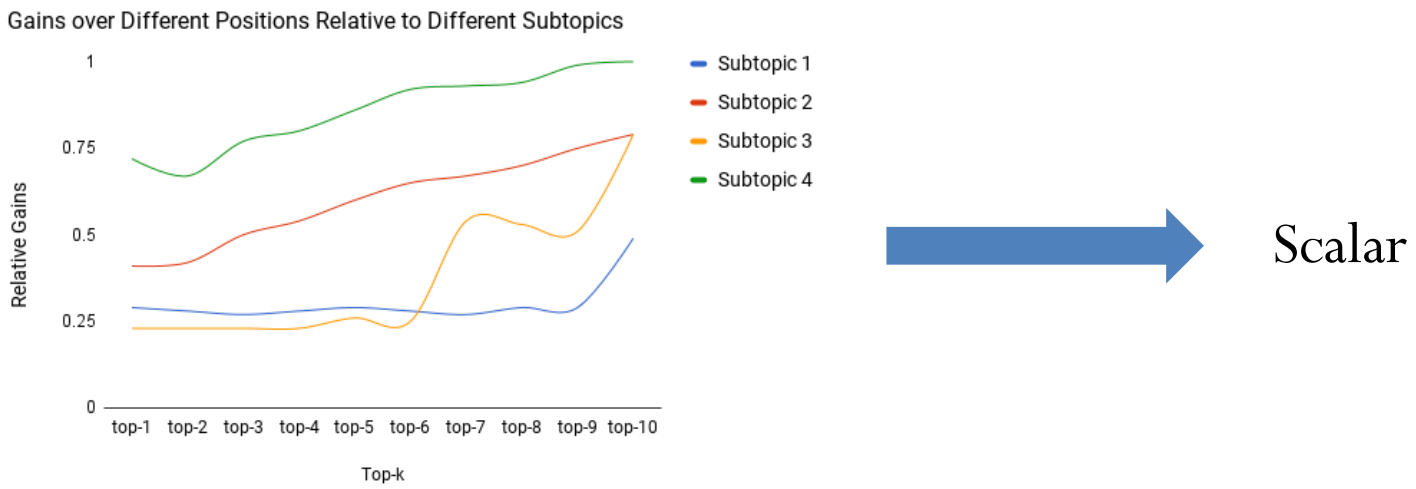
Normalization factor: relative to a background language model, e.g., a collection language model.

While going deeper in a ranking, the gains of relevant information are computed relative to different subtopics at each position.

Method
(3) KL-Divergence for Evaluation



Method
(4) Summarize the KL-Divergence



- At each position, among different subtopics employ the maximum of the absolute gains **(Abs)** or of the gain difference relative to the previous position **(Delta)**
- Sum up the gains over individual positions after rank-biased normalization **(RB)**

Evaluation
Experimental Setting

❑ **Dataset**

TREC Web Track 2011–2014, 64 k labeled documents, 200 queries

❑ **Simulate incomplete judgments**

Randomly sample p% judgments as incomplete judgments

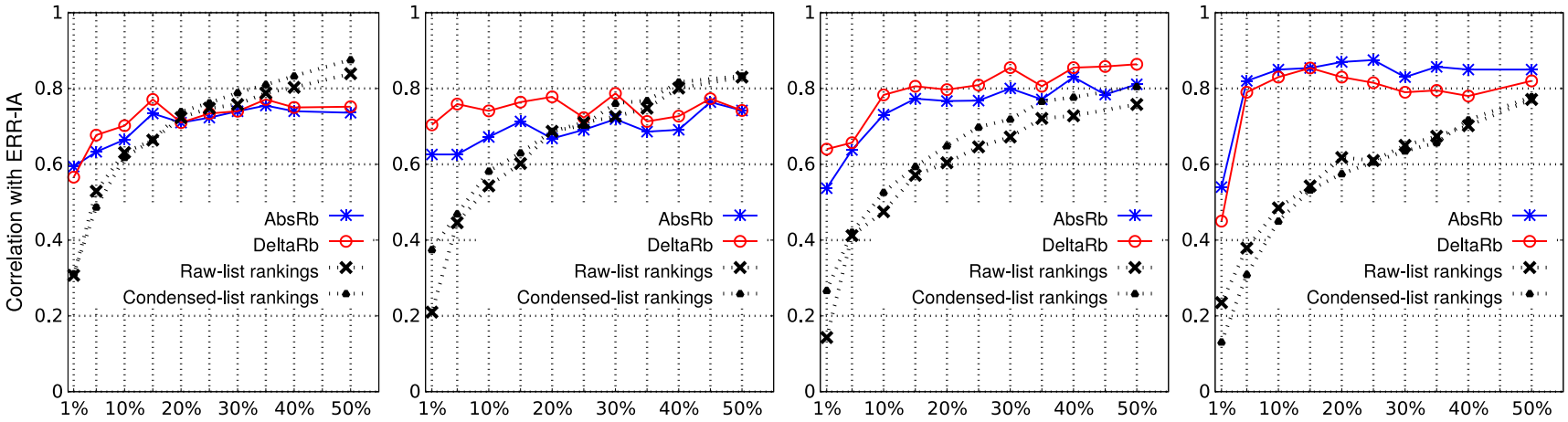
❑ **Baselines**

Treat unlabeled documents as non-relevant or non-existent

❑ **Benchmark**

Kendall's τ correlation: approximation to the system ranking under standard measures with complete measures

Evaluation Results



- Results on four years
- x-axis represents the percentage of the available judgments; y-axis is the correlation
- Dashed black curves are established measures (baselines)
- Red/blue curves represent two variants of the proposed measures
- Proposed measures are robust after more than 10% judgments are available
- However, it is hard for the proposed measures to get beyond 0.9 correlation

max planck institut informatik

Thank You!

contact: khui@mpi-inf.mpg.de