

Low-Cost Evaluation for Novelty and Diversity

Kai Hui

Max-Planck-Institut für Informatik, D5

khui@mpi-inf.mpg.de

Outline

- Background
- Motivation
- Related Works
- Preliminary Results
- Our Method
- Next Steps
- Discussions

Outline

- **Background**

- Novelty and Diversity Evaluation

- Motivation
- Related Works
- Preliminary Results
- Our Method
- Next Steps
- Discussions

TREC Novelty and Diversity Example

Query: No. 196 from WT2012

Query: sore throat

Subtopic 1: What causes a sore throat?

Subtopic 2: Find home remedies for a sore throat.

Subtopic 3: Find information on throat cancer.

Subtopic 4: What does it mean when my throat is sore on only one side?

Manual Label Example

| Query-id | Subtopic | Docid | Label |
|----------|----------|---------------------------|-------|
| 196 | 1 | clueweb09-enwp02-06-01125 | 1 |
| 196 | 2 | clueweb09-enwp02-06-01125 | 0 |
| 196 | 3 | clueweb09-enwp02-06-01125 | 1 |
| 196 | 4 | clueweb09-enwp02-06-01125 | 0 |

.....

Novelty and Diversity Measure Example: ERR-IA

$$ERR - IA = \sum_{top-k} \frac{1}{k} \sum_{subtopic\ i} g_{k,i} (1 - \alpha)^{c(k,i)}$$

$g_{k,i}$: relevance labels for the k -th document on subtopic i

$c(k, i)$: the count of relevant documents for subtopic i before the k -th document

α : the parameter for penalizing the repeating subtopics, normally set as 0.5

Their definition is directly based on the relevance labels, thus the evaluation quality highly depends on the labels.

Datasets

Document Collections

- ClueWeb09 Category A (CwA): **500 M English web pages**
- ClueWeb09 Category B (CwB): **50 M English web pages**
- Constructed Dataset (CwC): **450 M web pages from CwA but not in CwB**

Query Sets and Labels

- TREC Web Track (WT) 2009-2012, 200 queries with their labeled documents
- Runs for evaluation: **48 for 2009, 32 for 2010, 62 for 2011, and 48 for 2012**

Outline

- Background

- **Motivation**

Towards Lost-Cost Evaluation

- Related Works

- Preliminary Results

- Our Method

- Next Steps

- Discussions

Cost of Evaluation

- Top-k pooling: collect top-k from all candidate runs to generate a pool
- Manually label every subtopic document pair in the pool
- Limit to shallow pooling depth, e.g., top-25

| Year | #Query | #Systems | Pooling size | #Total labeled doc | #Labeled Relevant doc |
|------|--------|----------|--------------|--------------------|-----------------------|
| 2009 | 50 | 48 | 20 | 24,817 | 4,942 |
| 2010 | 50 | 32 | 20 | 15,352 | 6,553 |
| 2011 | 50 | 62 | 25 | 19,344 | 5,030 |
| 2012 | 50 | 48 | 20/30 | 16,036 | 5,559 |

Low Cost Evaluation Framework

Three Components in Evaluation:

Test Query Set



Document Collections



Manual Labeling

Select fewer candidate documents to label

- **Better discriminativeness**, e.g., only few systems can return a certain relevant document
- **Based on document content**, e.g., centroid documents in the similarity space are more representative
- **Combine the above two methods**

Reuses existing labels.

- **The existence of unlabeled documents** make existing measures not reusable.
- Our preliminary experiments show that Pool@20 only covers **25%-30% relevant documents**.

Objective

Objective

- Reconstruct the ranking of runs according to ERR-IA with incomplete judgments
- Compare the ranking of runs determined by ERR-IA with complete judgments
- Incompleteness comes from either the selection of initial pool or the reuse of the existing judgments

Measures

- Kendall's τ is used to measure the correlation among rankings
- Kendall's τ : ranging between -1 and 1

$$\tau = \frac{(\text{Number of Concordant Pairs}) - (\text{Number of Discordant Pairs})}{\text{Total Number of Pair Combinations}}$$

Outline

- Background
- Motivation
- **Related Works**
- Preliminary Results
- Our Method
- Next Steps
- Discussions

Related Work for Low Cost Evaluation

- **Sample documents to be labeled:**

Identify the crucial documents to label by selecting documents with best discriminating ability, like MTC. (Carterette et. al, 06)

- **Learning to predict the missing labels**

Mitigate the missing labels by predicting labels. Only has been tested for mitigating small portion of missing labels on adhoc task.

(Büttcher, 07) (Carterette & Allan, 07)

- **Condensed list of relevance judgment**

Remove all the unjudged documents instead of regarding them as irrelevant, tend to over-estimate the unlabeled systems. (Sakai, 13)

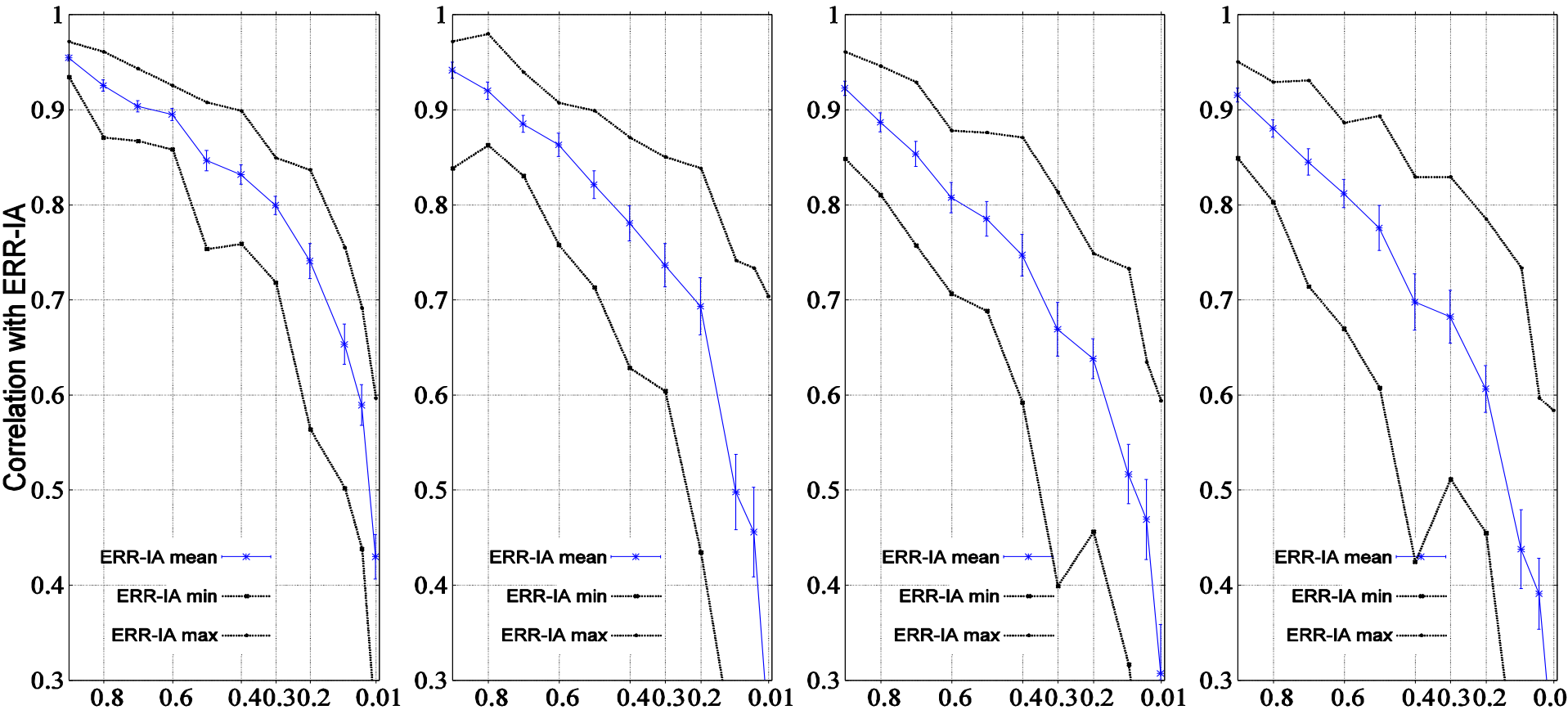
- **Reduces query numbers:**

Use fewer queries for testing and conclude statistical significant result. Most are retrospective method, the performance is unclear. (Robertson, 11)

Outline

- Background
- Motivation
- Related Works
- **Preliminary Results**
 - (1) Label Fewer Documents
 - Our Method
 - Next Steps
 - Discussions

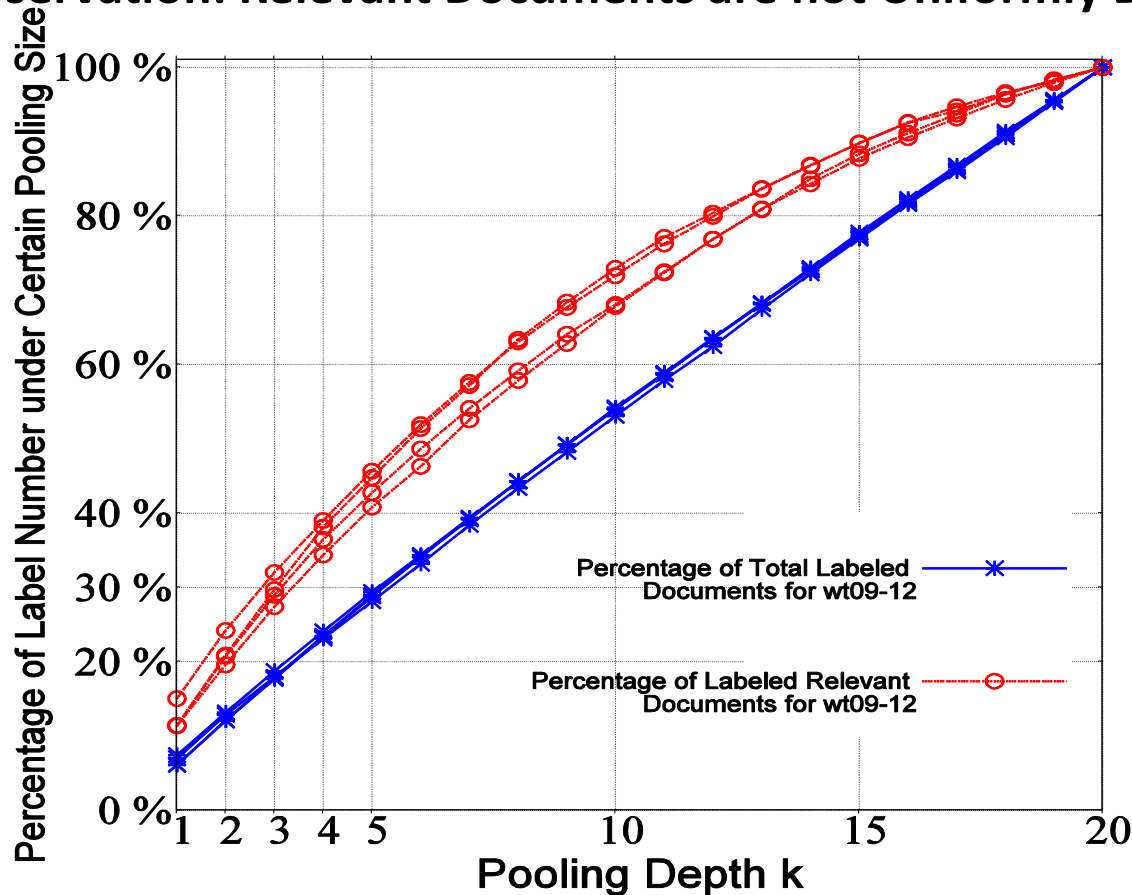
Randomly Select Fewer Documents to Label



- Percentage of available labels with random sampling versus the Kendall's tau correlation, repeated 30 times: **evaluation with 60% of labels is not reliable**
- Low-cost evaluation can't be implemented by simply random sampling, and the evaluation is sensitive to the completeness of the labels

How to Select?

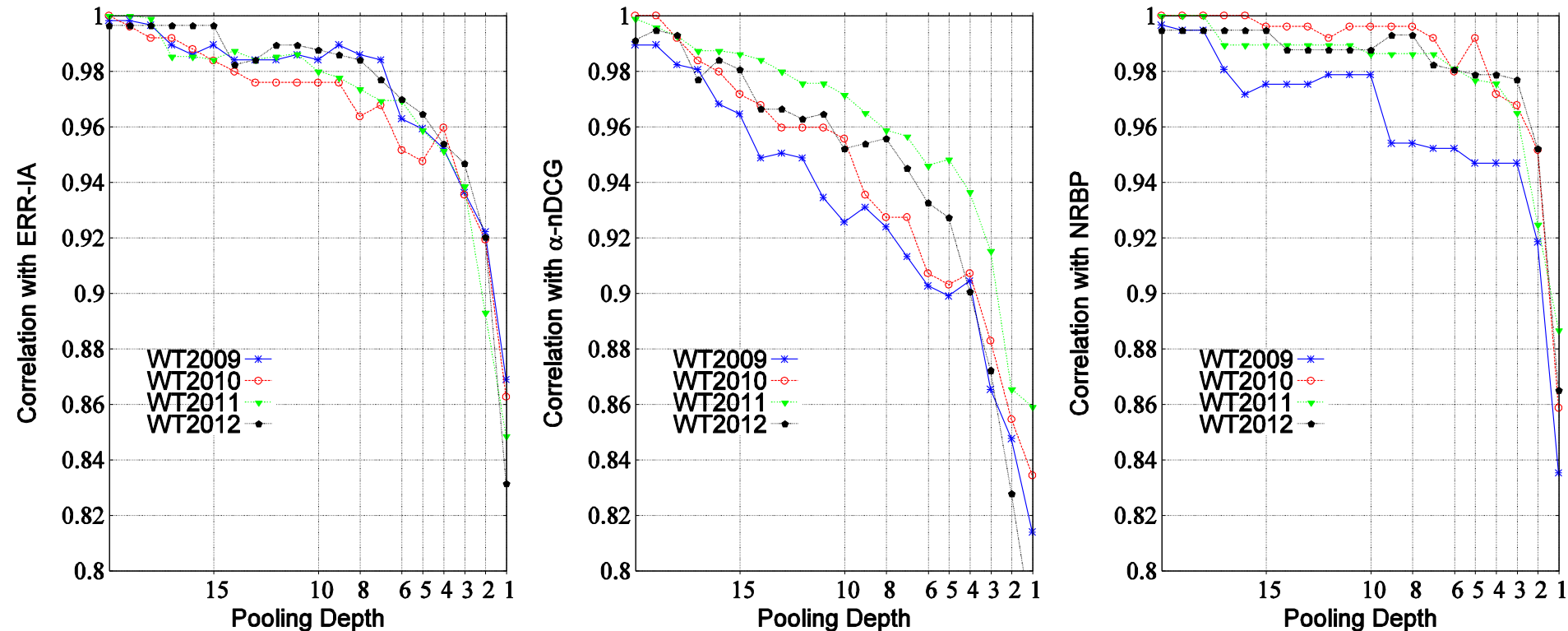
Observation: Relevant Documents are not Uniformly Distribute



- Left: percentage of **relevant (red)** / **total (blue)** w.r.t. the pooling depth
- Right: percentage of **total labels** w.r.t. **the relevant labels**
- **Relevant documents distribute not uniformly on the pooling depth, i.e., shallow pool contain larger portion of the relevant documents**

Select According to the Rank from Rival Systems

Preliminary Results: Labels on Shallow Pool Simulate Complete Evaluation Perfectly



- Kendall's tau correlation between the ranking determined by full judgment w.r.t. the judgment on different pooling depth, for measures ERR-IA, α -nDCG and the NRBP
- **Label on Top-6 pool (40% of total labels) is enough to get over 0.9 correlation w.r.t. complete judgment**
- We can use document content or discriminativeness to further reduce the label numbers¹⁶

Outline

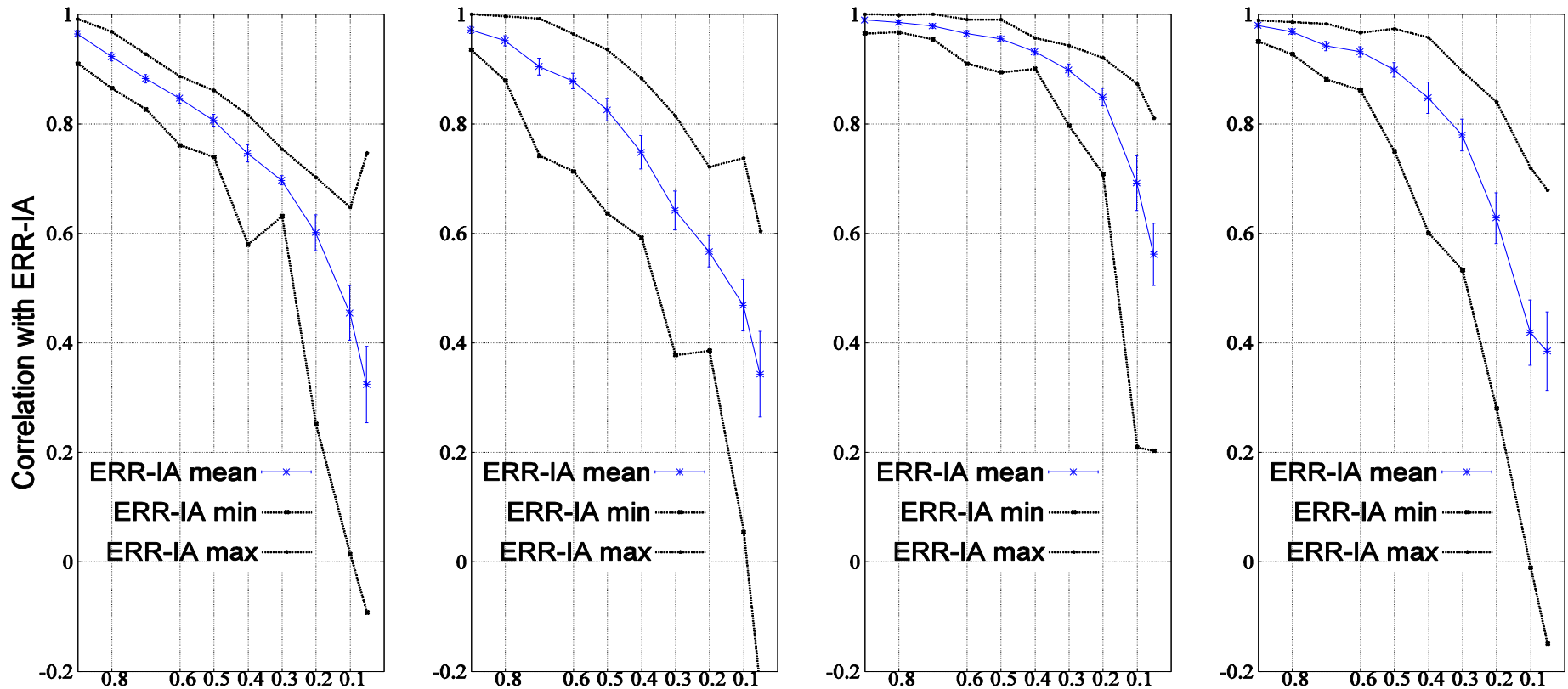
- Background
- Motivation
- Related Works
- **Preliminary Results**

(2) Reuse the Labels

- Our Method
- Next Steps
- Discussions

Reuse the Existing Labels

Leave N Out: Evaluate Systems without Contribution to the Pool



- Percentage of certain percentage of systems versus the Kendall's tau correlation, repeating 30 times: evaluation with less than 50%-60% systems contributing labels is not reliable
- **The existing measurement is not reusable, being biased towards systems without contributions to the pool**

Existence of the Unlabeled Documents: Predict the Labels

When might they exist?

- When the evaluated system was not included in the pooling
- On a new document collection
- When going deeper than the pooling depth, e.g., ERR-IA@30

Why are they problematic?

- Direct dependence on the labels, e.g., ERR-IA
- Missing labels have to be mitigated before evaluation

How to deal with them?

- Regard unlabeled as irrelevant: underestimates (Sakai et.al, 12)
- Remove unlabeled documents (condensed list): overestimate (Sakai et.al, 12)
- Predict the labels for unseen documents (Büttcher et.al, 07)(Carterette & Allan, 07)

Outline

- Background
- Motivation
- Related Works
- Preliminary Results
- **Our Method**

Proposed Reusable Evaluation Measures

- Next Steps
- Discussions

Mitigate the Unlabeled Documents: Predict the Missing Labels



Given query, subtopics, and top-k documents to evaluate.

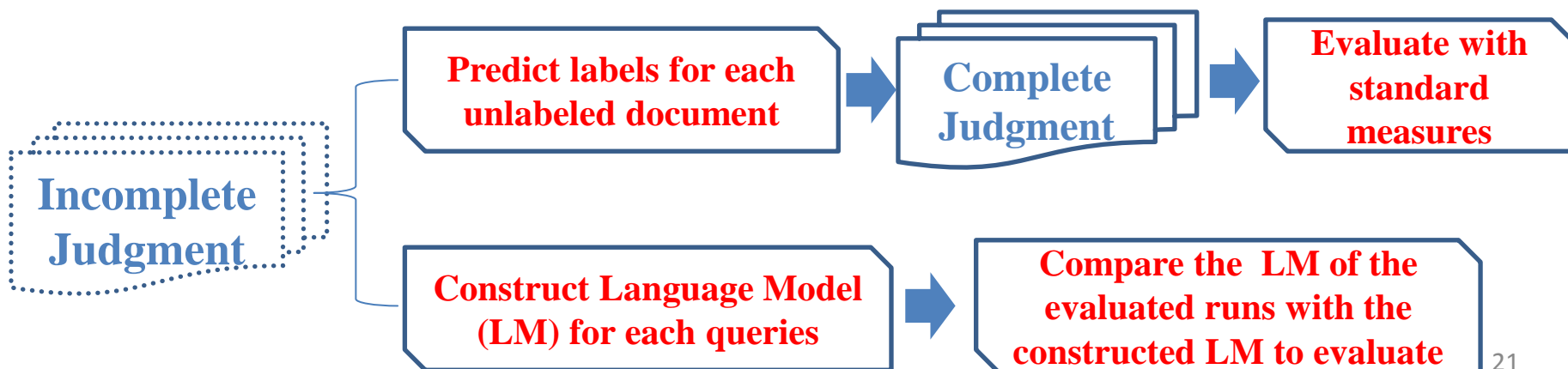
Options for prediction:

- **Pointwise prediction** (Büttcher et.al 07) (Carterette & Allan, 07)

<query, subtopic, doc>: <relevant, irrelevant>

- **Listwise prediction**

<query, subtopics, top-k docs>: the diversity and relevance of top-k docs



Pointwise Prediction

Predict Labels for Document Subtopic Pair

Prediction Setting

- Given information of query, subtopic and the tf-idf for each document
- Select terms with the decreasing order of the collection frequency
- Similar setting with (Büttcher et.al 07)

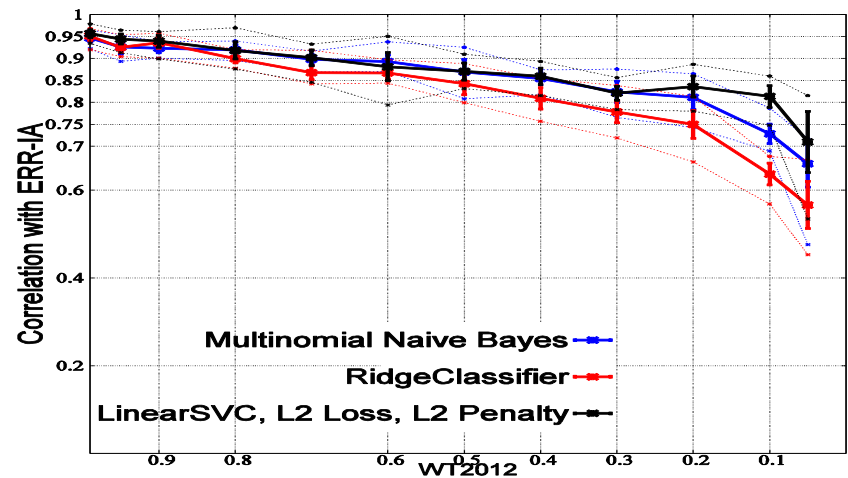
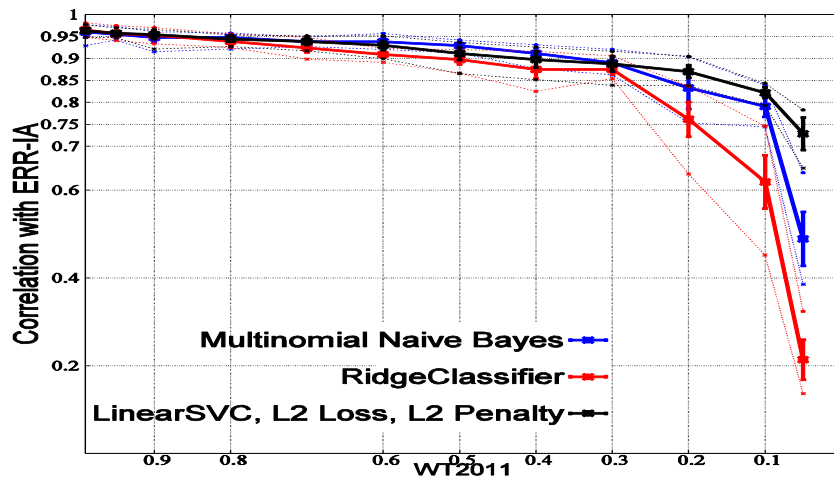
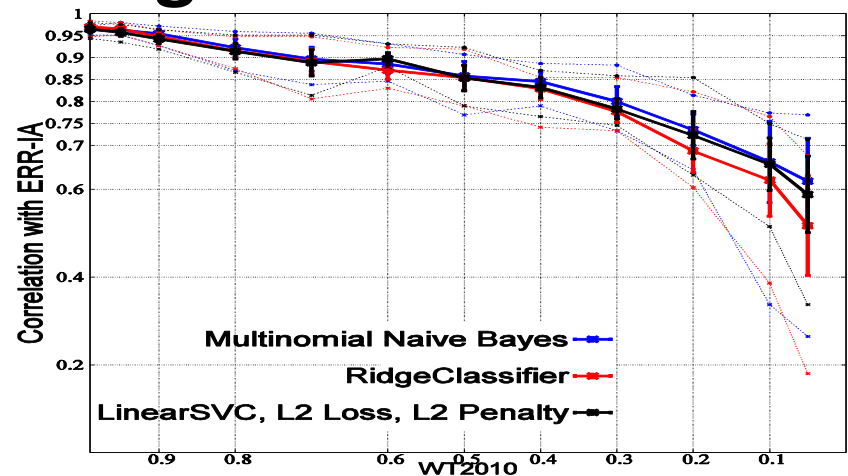
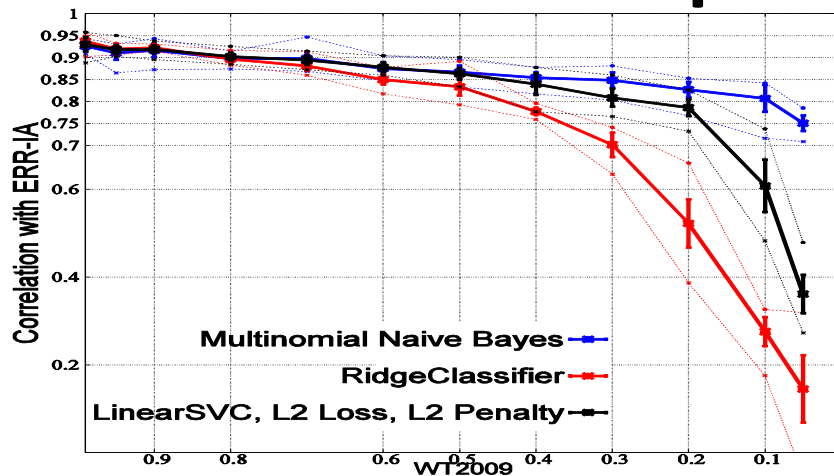
Learning Tools

- Scikit package based on python (<http://scikit-learn.org/>)
- Use Naïve Bayes, Linear Regression and the SVM to train
- With default setting of the parameters from the toolkit

Performance

- Partially mitigate the missing labels, can reconstruct the ranking of systems with as less as 40%-50% available labels

Pointwise Prediction on Randomly Sampled Incomplete Judgment



- Percentage of available labels with random sampling versus the Kendall's tau correlation, repeated 6 times: **evaluation with less than 30%-50% of labels is not reliable**
- Pointwise prediction can partially mitigate the missing labels

Outline

- Background
- Motivation
- Related Works
- Preliminary Results
- Our Method
- **Next Steps**
- Discussions

Future Work in Low-Cost Evaluation

Reusable Evaluation:

- Generation of subtopic distribution: text summarization, snippet generation etc..
- Regression: from distance matrix to the value of effectiveness measure
$$\text{Measure Score} = f(\text{Abs}, \text{Delta})$$
- Pairwise/Pointwise prediction to predict the missing labels

Select Documents to Label:

- Select documents taking into account their discriminativeness and content.

Outline

- Background
- Motivation
- Related Works
- Preliminary Results
- Our Method
- Next Steps
- **Discussions**

Discussions

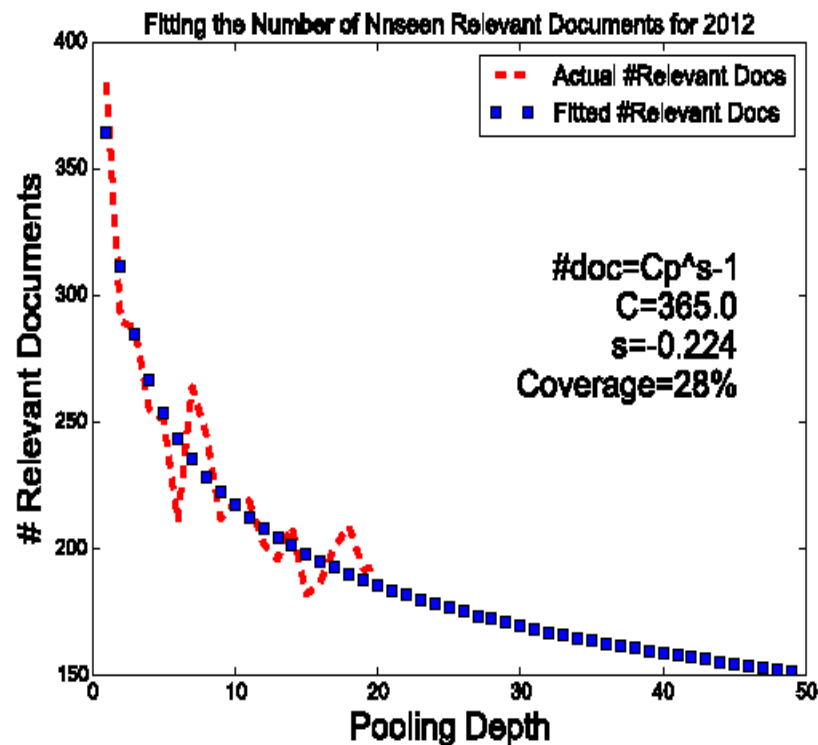
- Is novelty & diversity evaluation at deeper depth, e.g., at 40, meaningful?
- Whether we want to reuse the judgment on new document collections?
- How to confirm the existence of the relevant unlabeled documents?
- Which features, other than document discriminativeness and content are promising for selective labeling?

Thank You!
Q & A

References

- T. Sakai, Z. Dou, R. Song, and N. Kando: The reusability of a diversified search test collection. In Proceedings of AIRS 2012, pages 26–38, 2012
- Justin Zobel: How reliable are the results of large-scale information retrieval experiments?, Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, p.307-314, August 24-28, 1998, Melbourne, Australia .
- Stefan Büttcher , Charles L. A. Clarke , Peter C. K. Yeung , Ian Soboroff: Reliable information retrieval evaluation with incomplete and biased judgments, Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval, July 23-27, 2007, Amsterdam, The Netherlands
- Ben Carterette , James Allan: Semiautomatic evaluation of retrieval systems using document similarities, Proceedings of the sixteenth ACM conference on Conference on information and knowledge management, November 06-10, 2007, Lisbon, Portugal
- Ben Carterette , James Allan , Ramesh Sitaraman: Minimal test collections for retrieval evaluation, Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval, August 06-11, 2006, Seattle, Washington, USA
- Robertson, S: On the contributions of topics to system evaluation. In *European conference on information retrieval* (pp. 129–140).
- T. Sakai: The unreusability of diversified search test collections. In Proceedings of EVIA 2013, 2013.

Difficulties in Reusing the Labels: Unlabeled Documents



X-axis: Pooling Depth

Y-axis: Number of new relevant documents

How many do we miss? (Zobel, 98)

- Count the number of new relevant documents in pool@i given labeled pool@i-1
- Fit the curve for existing pool and predict for deeper pool, e.g., pool@100
- Pool@20 covers 25% - 30% relevant documents
- Unlabeled relevant documents are due to their low rank, e.g., rank at 50

Why do the unlabeled documents exist?

- Incomplete coverage of the relevant documents in existing judged pool.