max planck institut
informatik

Selective Labeling and Incomplete Label Mitigation for Low-Cost Evaluation

Kai Hui, Klaus Berberich
Max Planck Institute for Informatics
khui@mpi-inf.mpg.de
kberberi@mpi-inf.mpg.de

Sep. 01, 2015

MAX-PLANCK-GESELLSCHAFT

max planck institut
informatik

Overview

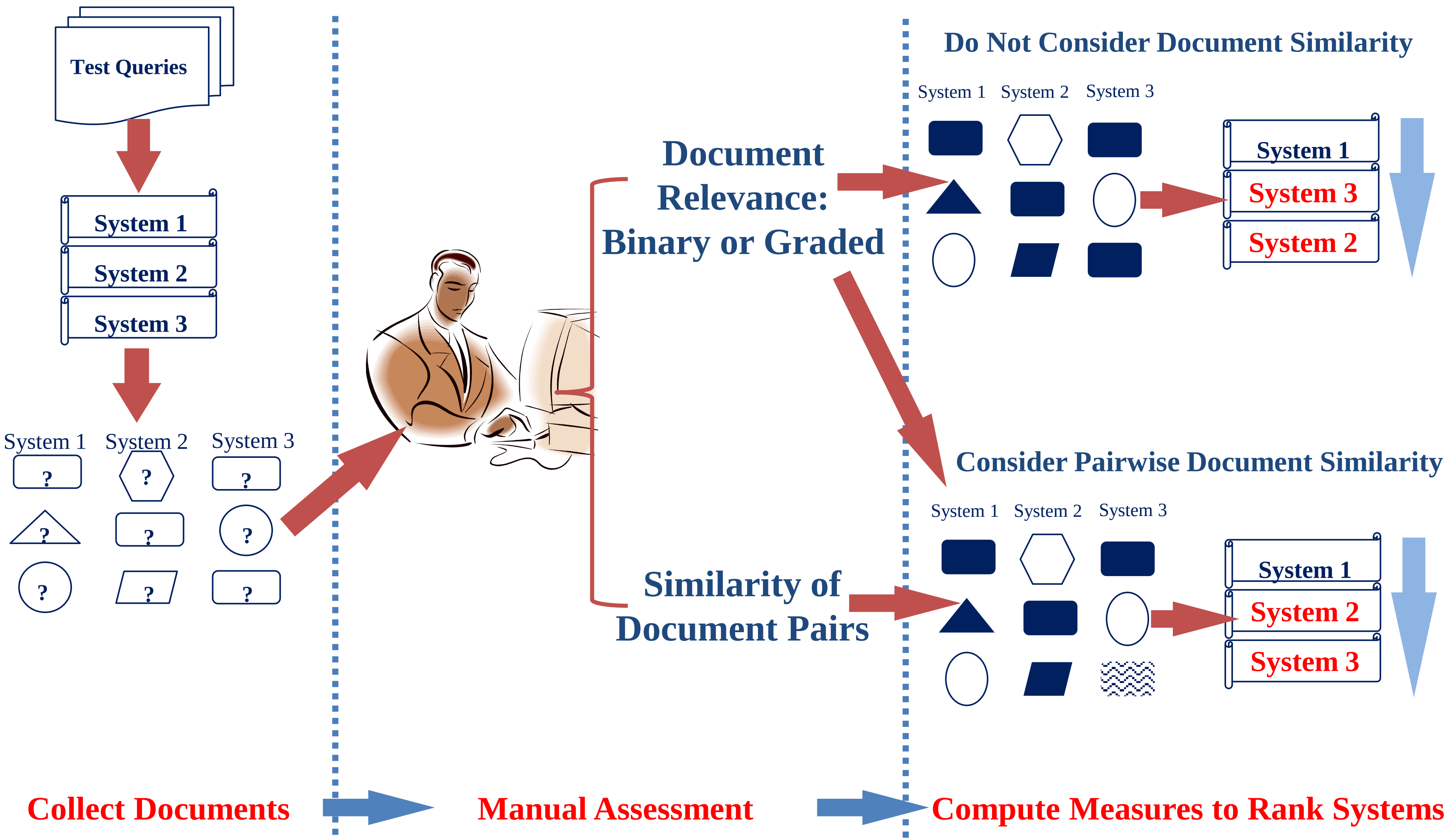
- ❑ Background: IR Evaluation
- ❑ Objectives & Related Work
- ❑ MaxRep Selective Labeling
- ❑ Experimental Results
- ❑ Conclusion

Overview

- ❑ Background: IR Evaluation
- ❑ Objectives & Related Work
- ❑ MaxRep Selective Labeling
- ❑ Experimental Results
- ❑ Conclusion

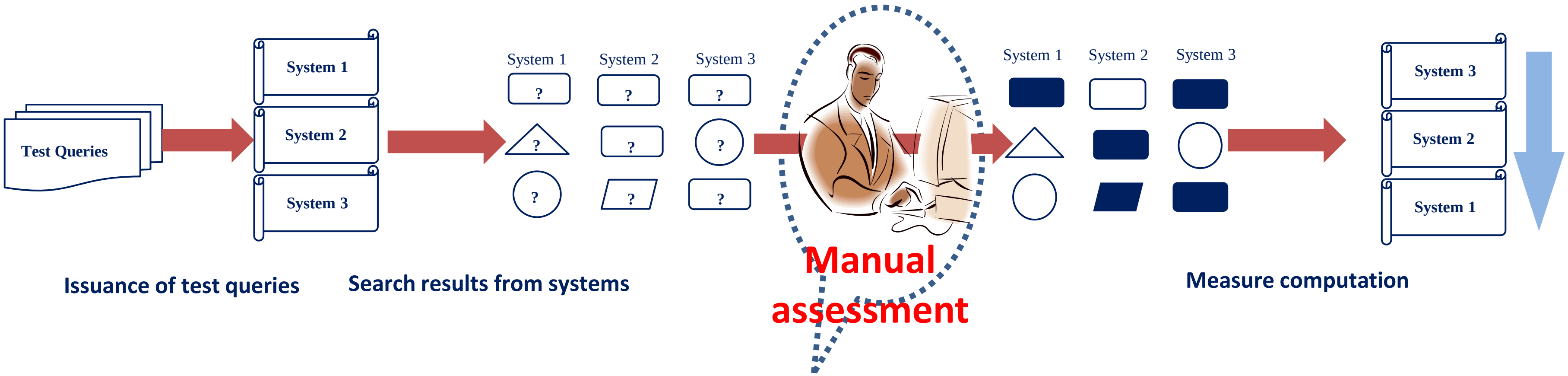
max planck institut
informatik

Background: IR Evaluation Pipeline



max planck institut
informatik

Expensive Cost of Evaluation



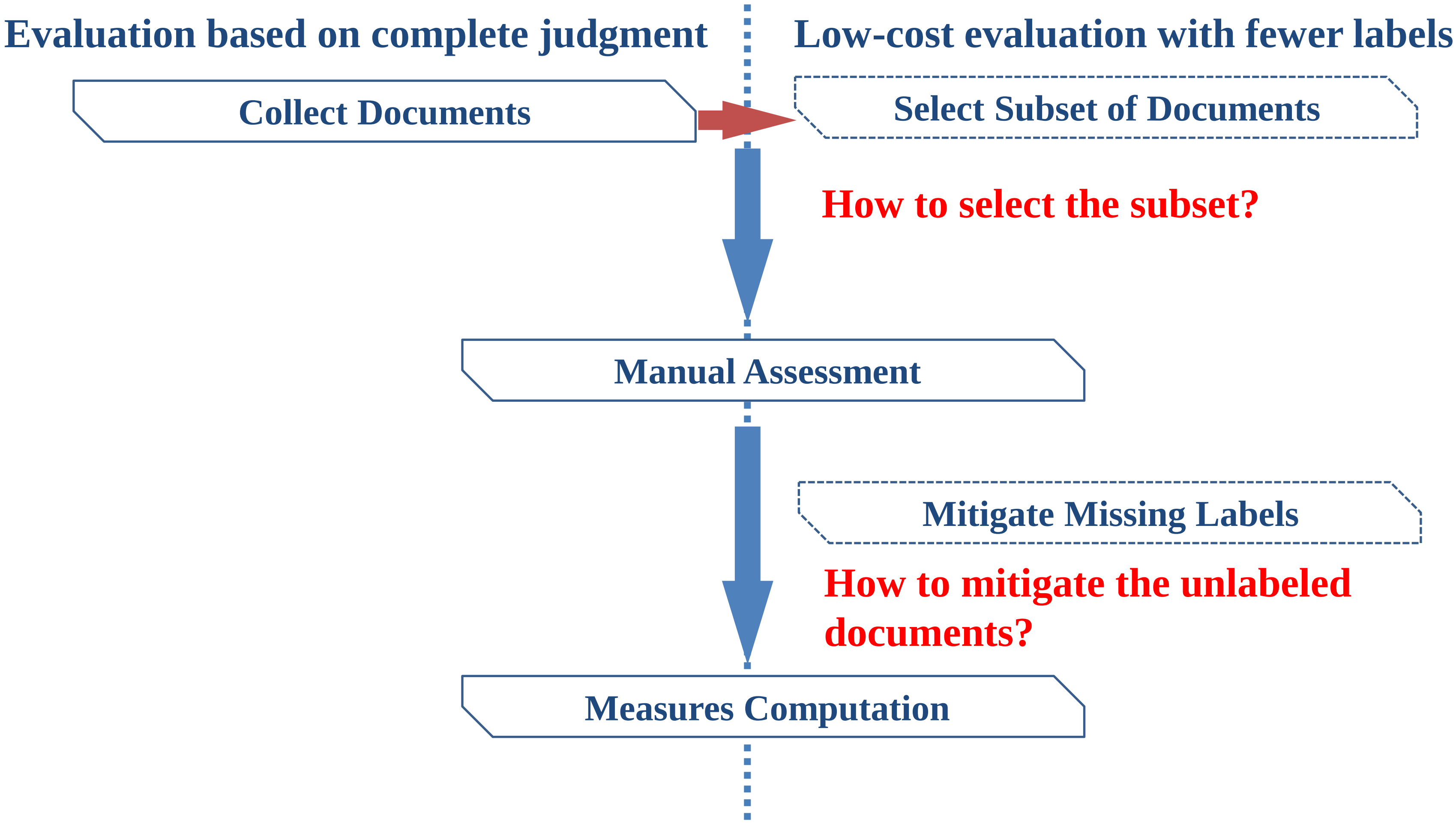
Statistics of Labels from TREC Web Track			
Year	**#Systems**	**Pooling depth**	**#Total labeled doc**
2011	62	25	19,344
2012	48	20/30	16,036
2013	61	10/20	14,336
2014	42	25	14,429

max planck institut informatik

Overview

- Background: IR Evaluation
- Objectives & Related Work
- MaxRep Selective Labeling
- Experimental Results
- Conclusion

Objective: Evaluation with Fewer Labels



Related Work

Evaluation based on complete judgment

Collect Documents

Manual Assessment

Measures Computation

Low-cost evaluation with fewer labels

Select Subset of Documents

Different selection strategies:

- Sampling: uniformly sampling, statAP
- Pooling: incremental pooling
- Active selection: MTC, RTC

Mitigate Missing Labels

Different mitigation methods:

- Regard missing labels as non-relevant or non-existing: default in TREC, indAP & condensed list
- Distinct labeled non-relevant documents: bpref
- As random variables: infAP, eMAP (MTC)
- Predict missing labels

max planck institut informatik

Overview

- ❑ Background: IR Evaluation
- ❑ Objectives & Related Work
- ❑ **MaxRep Selective Labeling**
- ❑ Experimental Results
- ❑ Conclusion

Framework: Selective Labeling & Label Prediction

Observations

- ❑ **Cluster Hypothesis:** relevant documents are clustered
- ❑ **Label Bias**: there exist more non-relevant documents than relevant documents

Selective Labeling Strategies

Cluster Hypothesis → representative documents

Label Bias → documents are more likely to be relevant

General Framework

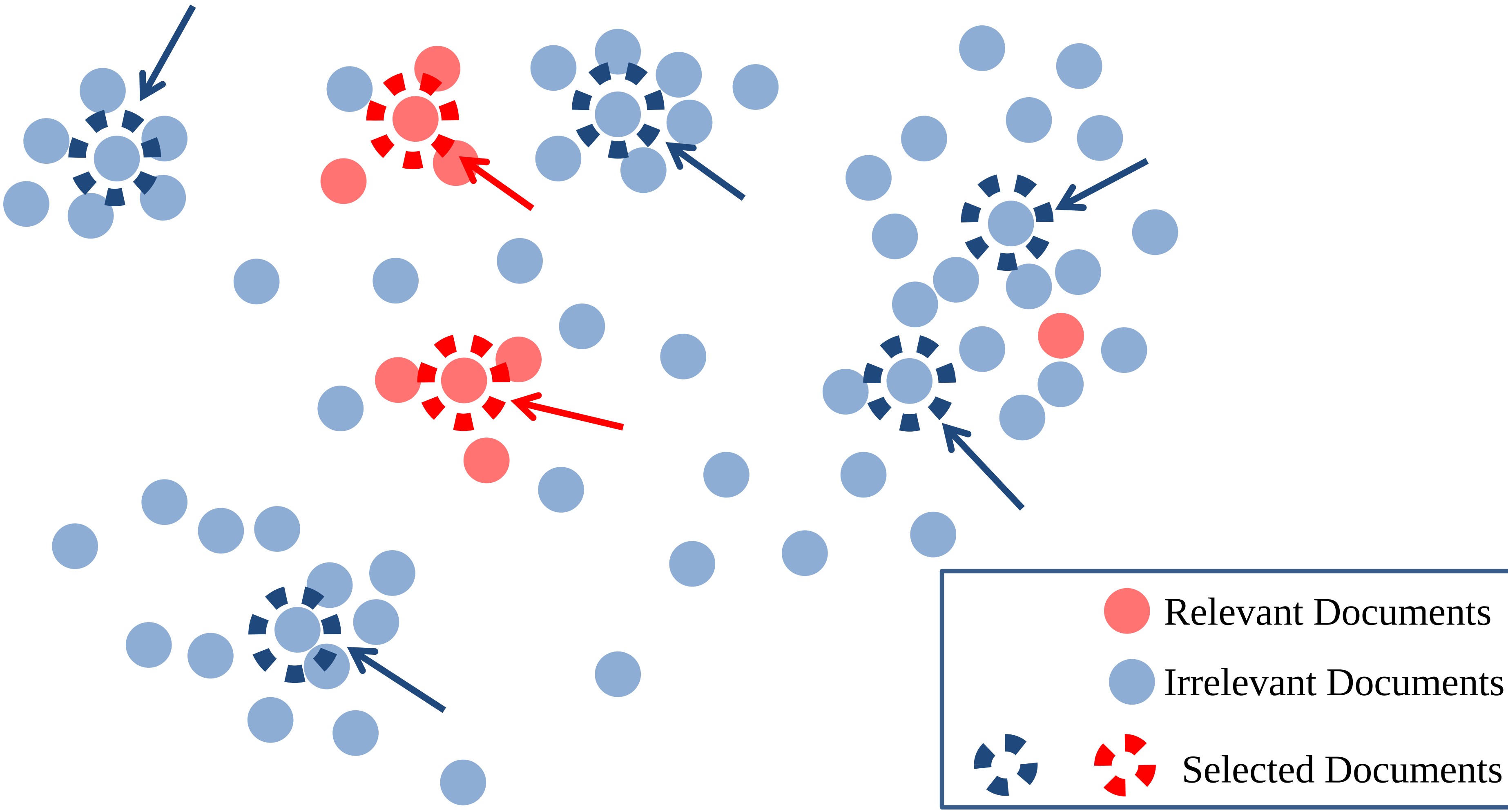
Selective Labeling + Label Prediction

Label Prediction

Standard text classification method: SVM with linear kernel

max planck institut
informatik

MaxRep Example:
Document Vector Space for A Given Query



MaxRep Method: Representative Documents

Representativeness of Document Subset

- Document subset L with t documents from document collection D
- Representativeness of L is the aggregating maximum coverage of the remaining documents D

$$f(L) = \sum_{d_i \in D_q} \max_{d_j \in L} sim(d_i, d_j)$$

MaxRep Method: Encode Document Relevance

Encode Document Relevance in Selection

- AP-Prior: documents ranked higher by rivaling systems are more likely to be relevant. n denotes length of ranking, r is the rank

$$P[r] \approx \frac{1}{2n} log \frac{n}{r}$$

- Allocate aggregated AP-Prior weight w to document

$$f(L) = \sum_{d_i \in D_q} \max_{d_j \in L} \mathbf{w_i} sim(d_i, d_j)$$

Optimization Target: $L_k^* = \underset{L_k}{\operatorname{argmax}} \; f(L) \quad \text{s.t.} \quad |L| = k$

Overview

- ❑ Background: IR Evaluation
- ❑ Objectives & Related Work
- ❑ MaxRep Selective Labeling
- ❑ Experimental Results
- ❑ Conclusion

Experimental Setting

❑ **Dataset**

TREC Web Track 2011–2014 on ClueWeb 09 & 12, 64 k labeled documents, 200 queries

❑ **Ground truth measure**

Mean Average Precision (MAP)

❑ **Benchmark**

➢ Kendall's τ correlation:

Approximation to the system ranking

➢ Root Mean Square Error (RMSE):

Approximation to the MAP values

Methods in Comparison

Different alternatives in two building blocks:

Select Subset of Documents

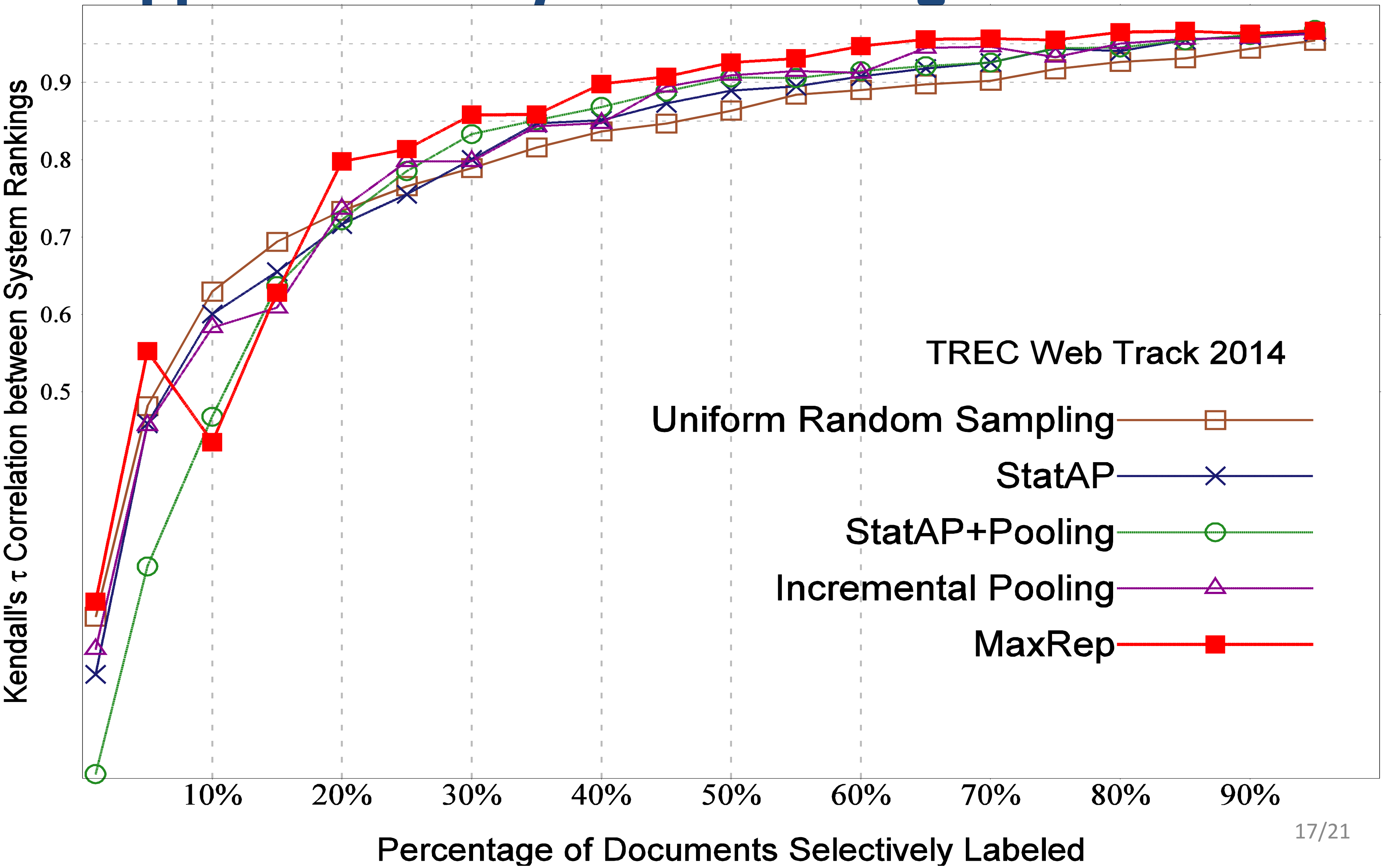
- ❑ Uniform random sampling
- ❑ Incremental pooling
- ❑ Non-uniform random sampling: statAP
- ❑ MaxRep: maximum representative

Mitigate Missing Labels

- ❑ trec-map: missing labels as non-relevance
- ❑ bpref: separates labeled non-relevance
- ❑ indAP: missing labels as non-exist
- ❑ infAP: estimator of precision at rank r
- ❑ statAP: adjusts inclusion probability
- ❑ Predict-map: SVM based label prediction

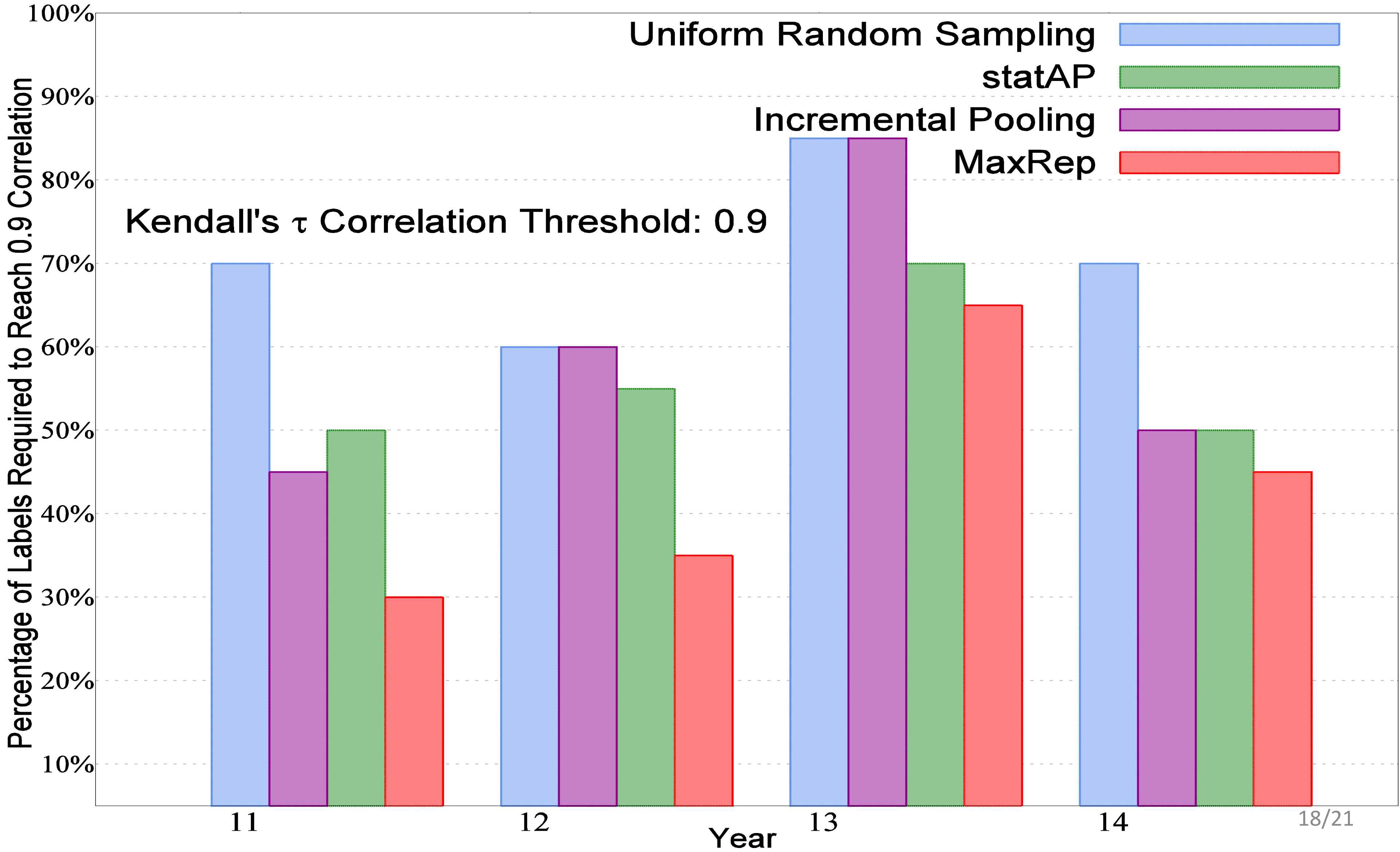
max planck institut
informatik

Approximate System Ranking: Kendall's τ

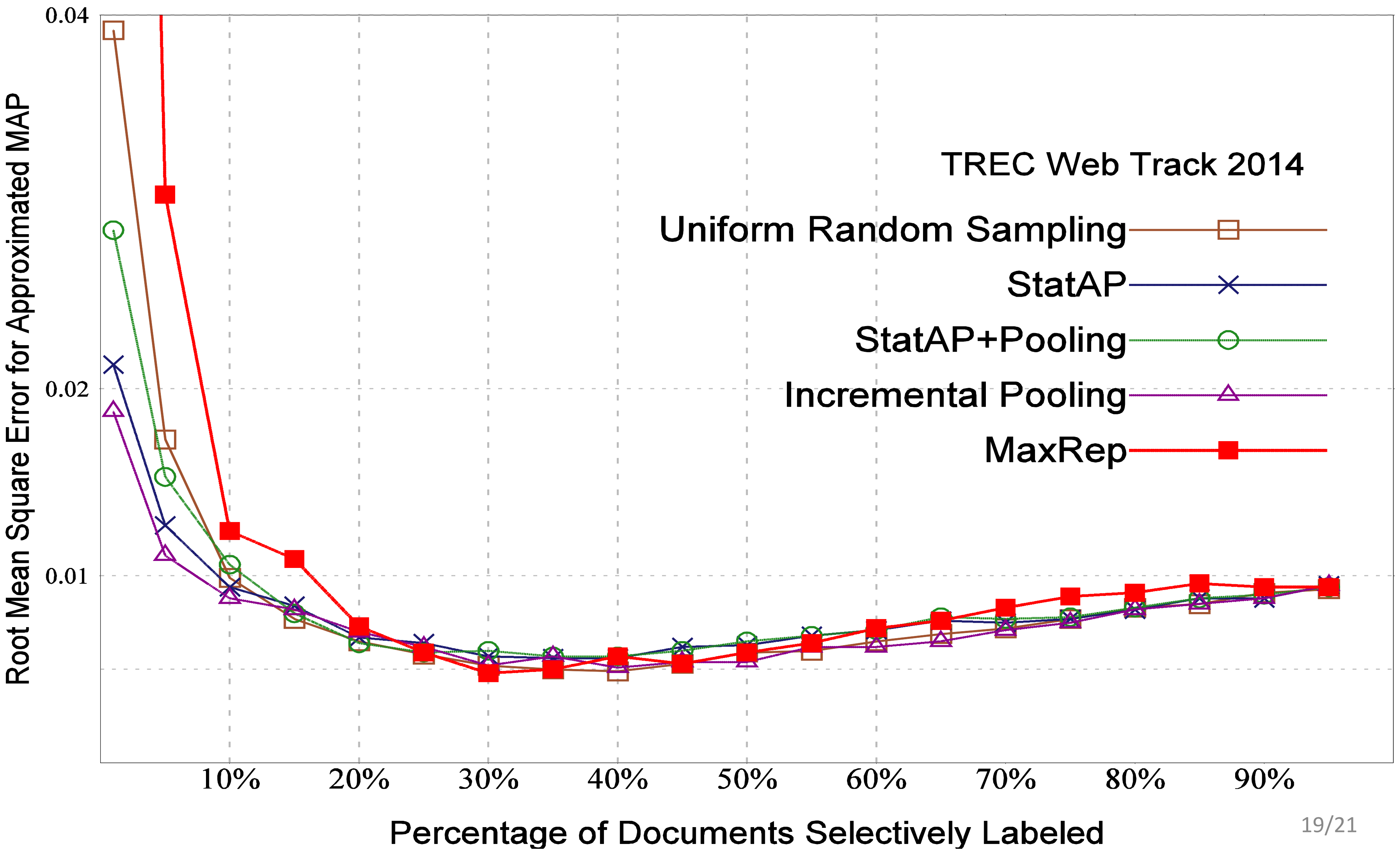


max planck institut informatik

Summarization of Kendall's τ



Approximate MAP Score: RMSE



Overview

- Background: IR Evaluation
- Objectives & Related Work
- MaxRep Selective Labeling
- Experimental Results
- Conclusion

Conclusion

- **Label prediction is a robust and viable** strategy to mitigate incomplete labels, with at least 20% of documents as training data
- **A novel strategy MaxRep is proposed**, considering both ranking information and document contents and seeking to select a representative subset of documents to label
- **Large scale experiments on TREC Web Track** data confirmed MapRep outperforms other strategies when label prediction is used
- For future works, **novelty & diversity will be considered**, and corresponding measures, like ERR-IA, will be approximated

Thanks!