A Simple and Efficient Sampling Method for Estimating AP and nDCG

Emine Yilmaz
Microsoft Research, Cambridge, UK

Evangelos Kanoulas Javed Aslam
Northeastern University, Boston, USA

Introduction

- Obtaining relevance judgments
  - Relevance judgments are expensive
- TREC: Depth-k pooling
- Document collections can be very large
  - Depth pooling is still expensive (85600 judgments for TREC8)
    - *3 min/doc, 40 hrs/wk, 50 wks/year ==> 2.14 man-years!*
- Evaluation with incomplete judgments
  - Bpref (Buckley and Voorhees, SIGIR'06)
  - Evaluation using condensed lists (Sakai SIGIR'07)
  - Methods for *ranking* systems with less judgments (Carterette et al. SIGIR'06, Moffat et al. SIGIR'07)
  - Methods *directly estimating measures* with less judgments (Aslam et al. SIGIR'06, Yilmaz and Aslam CIKM'06)

Motivation

- Inferred AP (Yilmaz and Aslam CIKM'06)
  - No confidence intervals associated with the estimates
  - Incomplete relevance judgments random subset of complete judgments
- Importance Sampling (Aslam et al. SIGIR'06)
  - Difficult to compute confidence intervals
  - Overly complicated
- Combine the advantages of the two approaches
  - Confidence intervals for inferred AP
  - Extend inferred AP to incorporate nonrandom judgments

Inferred AP

[Yilmaz and Aslam CIKM06]

- Average precision as a random experiment

  1. Select a relevant document at random
     - Rank of the document : *k*
  2. Select a rank at random from the set {1,….,*k*}
  3. Output the binary relevance of document at this rank.

- Average (step 1) of precisions at relevant documents (steps 2 and 3).

Inferred AP

[Yilmaz and Aslam CIKM06]

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

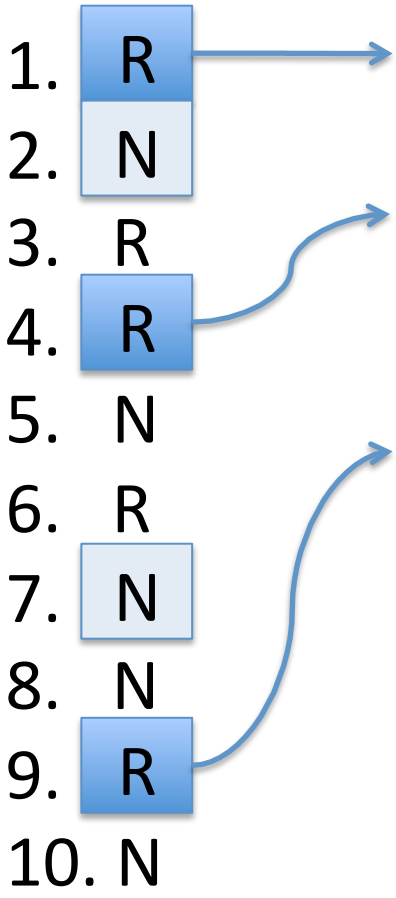
Inferred AP

[Yilmaz and Aslam CIKM06]

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

Inferred AP

[Yilmaz and Aslam CIKM07]



$$\widehat{PC}_1 = \frac{1}{1} \cdot 1 = 1$$

$$\widehat{PC}_3 = \frac{1}{4} \cdot 1 + \frac{3}{4} \cdot \frac{1}{2} = 0.625$$

$$\widehat{PC}_9 = \frac{1}{9} \cdot 1 + \frac{8}{9} \cdot \frac{2}{4} = 0.5555$$

$$\text{infAP} = \frac{1 + 0.625 + 0.5555}{3} = 0.7268$$

Variance in Inferred AP

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

- Inferred AP is unbiased in expectation
- Varies in practice
  - Variance and Confidence Intervals

- Random Experiment can be realized as two stage sampling

Variance in Inferred AP

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

- Two stages sampling
- Stage 1 : sample of *cut-off levels* (relevant documents) and average precisions
  - 1st variance component

Variance in Inferred AP

1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

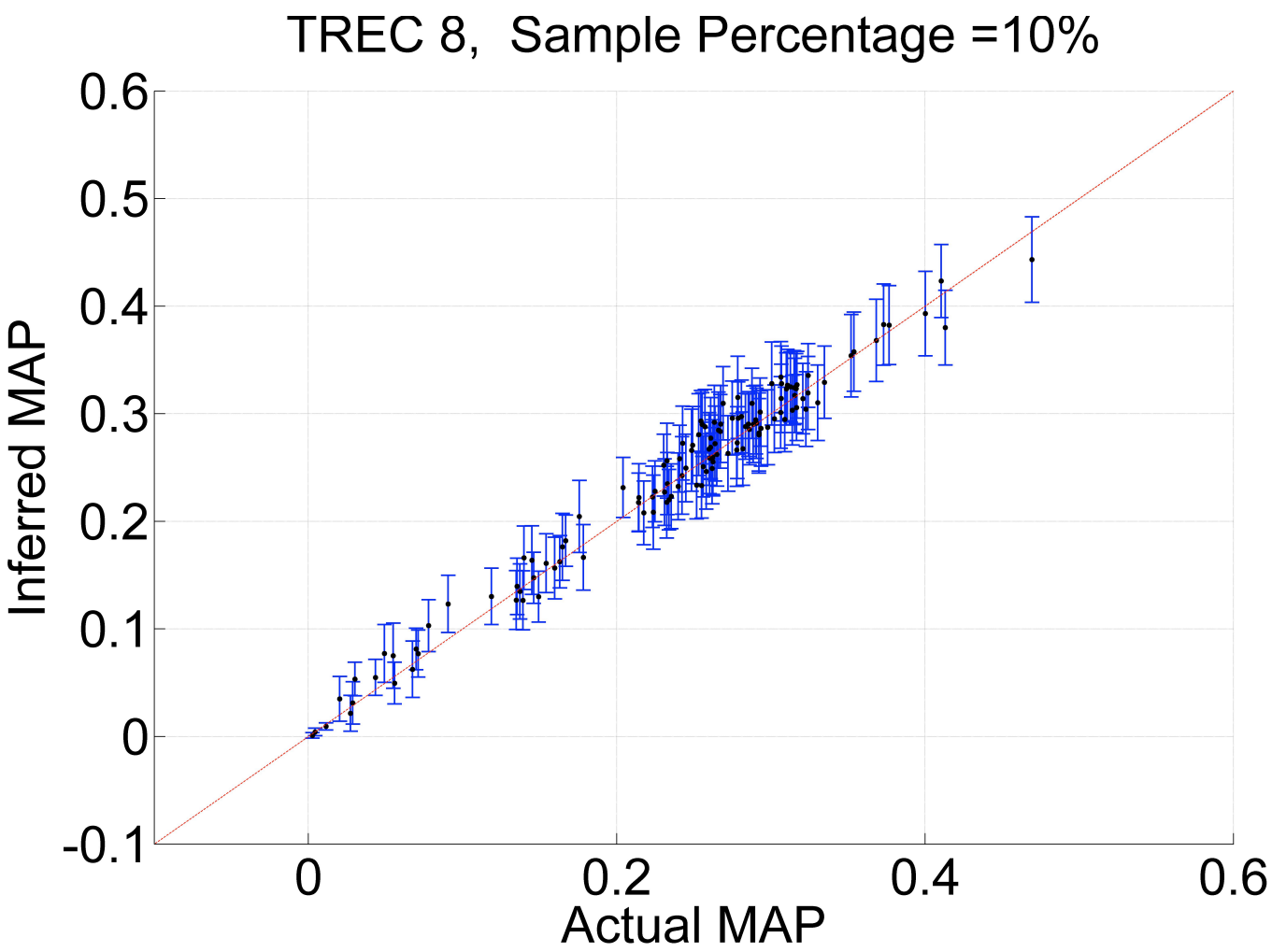
- Two stages sampling
- Stage 2 : sample of *documents* above each selected cut-of level to compute precisions
  - 2^{nd} variance component

Variance in Inferred AP

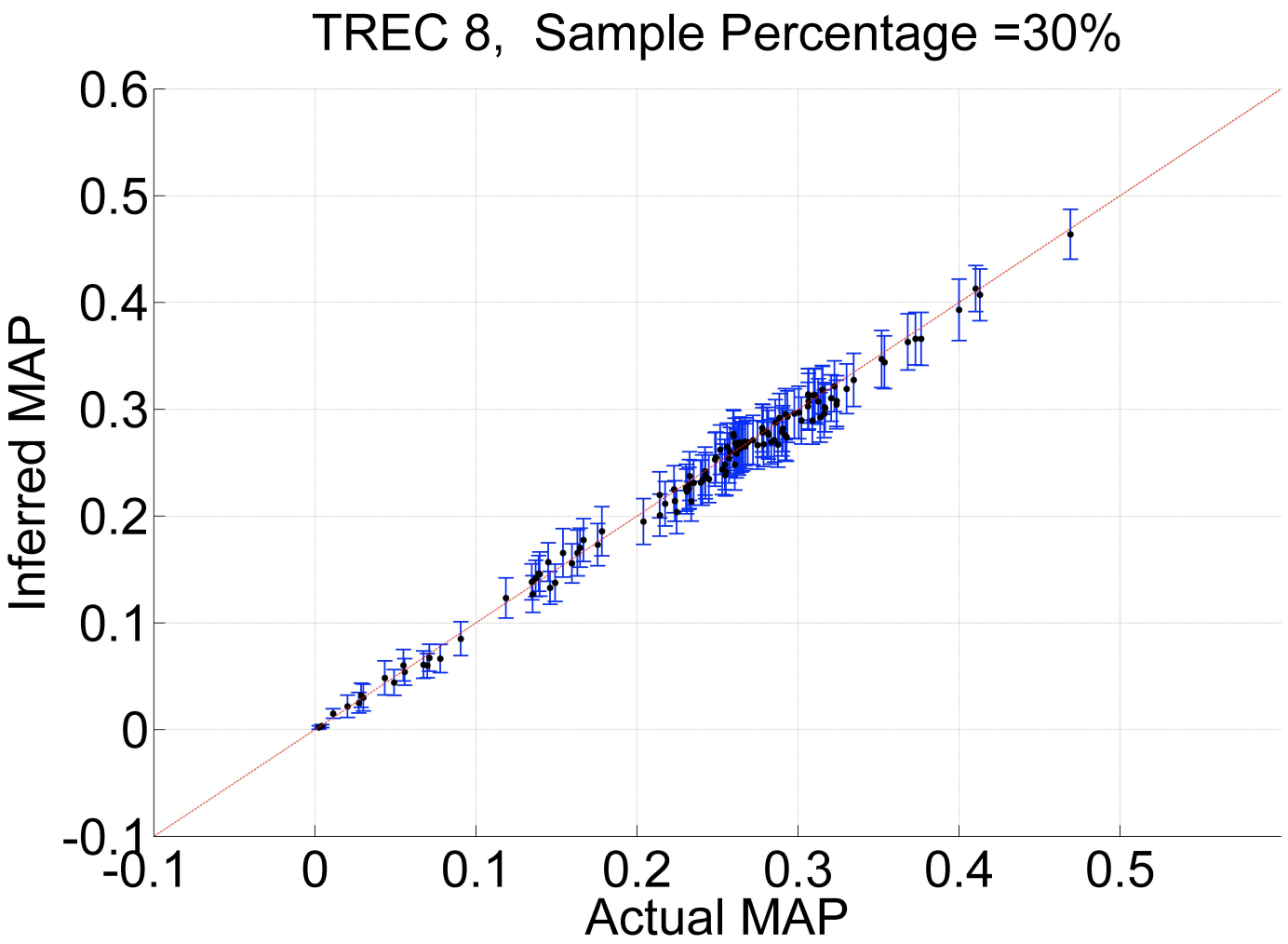
1. R
2. N
3. R
4. R
5. N
6. R
7. N
8. N
9. R
10. N

- Law of Total Variance
  - Total Variance in inferred AP = stage 1 variance + stage 2 variance
- Variance of Mean InfAP = Total Variance in InfAP / (# of Queries)^2
- Assign confidence intervals to Mean InfAP according to Central Limit Theorem

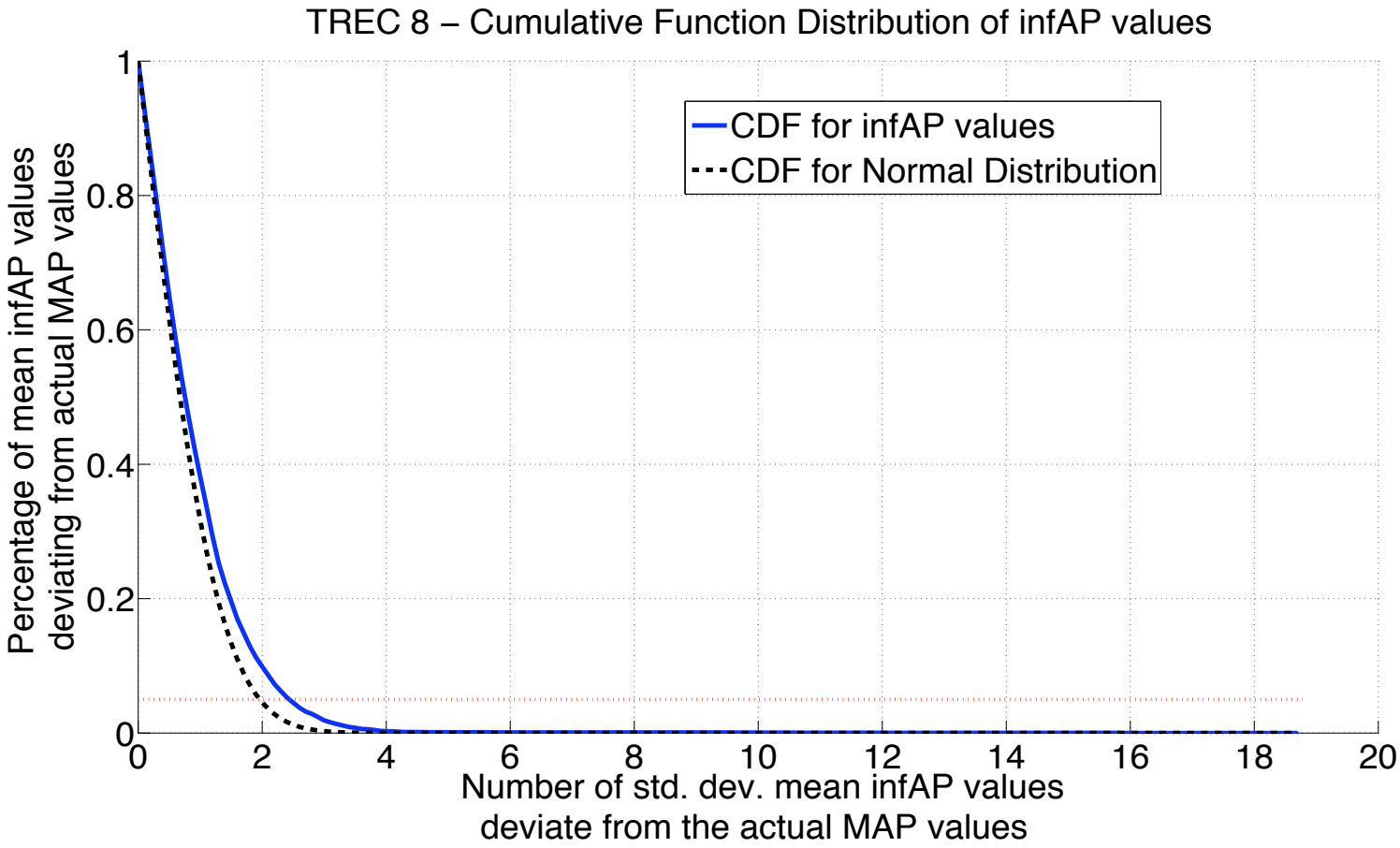
Confidence Intervals for Mean InfAP



Confidence Intervals for Mean InfAP

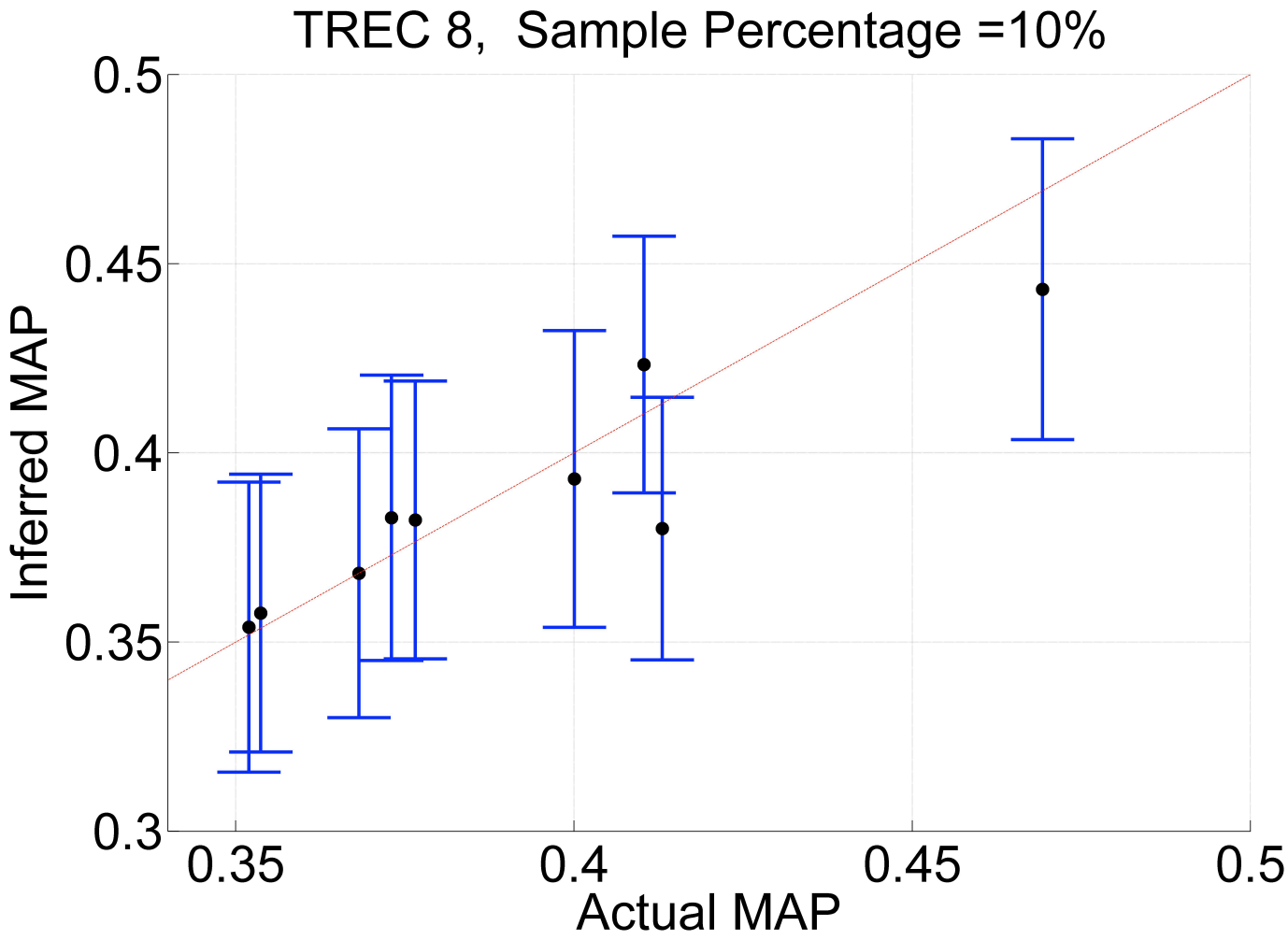


Confidence Intervals for Mean InfAP



- K-S test : for 90% of systems the hypothesis cannot be rejected ($\alpha = 0.05$)

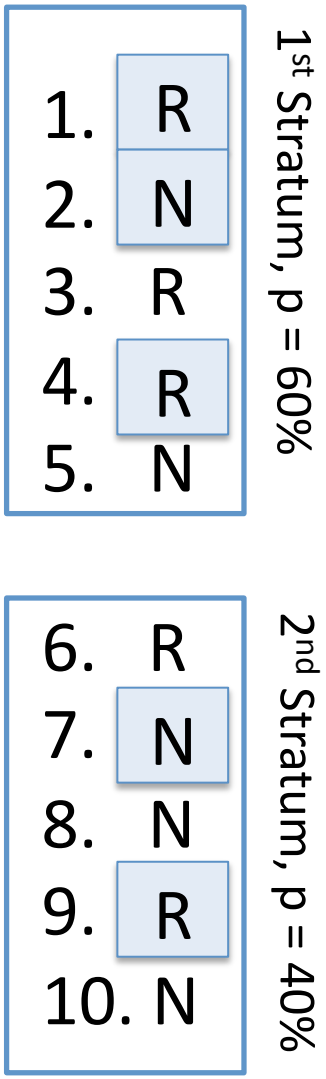
Confidence Intervals for Mean InfAP



Stratified Random Sampling

- Goal :
  - Unbiased estimator of AP
  - Decrease variance in the estimator
- Evaluation measures give more weight to documents towards the top of the list
- “Top-heavy” sampling strategy can reduce variance in Mean InfAP

Stratified Random Sampling



- Divide complete pool of judgments into strata (disjoint contiguous subsets)
- Randomly sample some documents from each stratum to be judged
- Sampling percentage within each stratum can be different
- Evaluate search engines with sampled documents

Extended infAP (xinfAP)

- Select a relevant document at random (1st step)
  - Selected relevant document can fall in any of the strata
  - By the definition of conditional expectation

$$\text{xinfAP} = E[AP] = \sum_{\forall s \in Strata} P_s \cdot E[AP_s]$$

P_s : Probability that a randomly picked rel docs falls into strata s

Extended infAP (xinfAP)

- Select a relevant document at random (1st step)
  - Probability of picking relevant document from stratum s

$$P_s = \frac{R_s}{R_Q}$$

R_s : Num rels within stratum s

R_Q : Num rels in query Q

Extended infAP (xinfAP)

- Select a relevant document at random (1st step)
  - Probability of picking relevant document from stratum s

$$P_s = \frac{R_s}{R_Q} \qquad \begin{array}{l} R_s : \text{Num rels within stratum } s \\ R_Q : \text{Num rels in query } Q \end{array}$$

$$\hat{P}_s \sim \frac{E[R_s]}{E[R_Q]}$$

$$E[R_s] = \frac{|\text{ rel docs sampled from } s\ |}{|\text{ docs sampled from } s\ |} \cdot |\text{ docs in } s\ | \qquad E[R_Q] = \sum_{\forall s} E[R_s]$$

Extended infAP (xinfAP)

1st Stratum, p = 60%

1. R
2. N
3. R
4. R
5. N

$$E[R_{s_1}] = \frac{2}{3} \cdot 5$$

$$E[R_{s_2}] = \frac{1}{2} \cdot 5$$

2nd Stratum, p = 40%

6. R
7. N
8. N
9. R
10. N

$$\hat{P}_{s_1} = \left(\frac{2}{3} \cdot 5\right) / \left(\frac{2}{3} \cdot 5 + \frac{1}{2} \cdot 5\right) = 0.57$$

Extended infAP (xinfAP)

$$\text{xinfAP} = E[AP] = \sum_{\forall s \in Strata} P_s \cdot E[AP_s]$$

- Select a relevant document at random (1st step)
  - Within each stratum:
    - Judged documents uniform random subset of all documents
    - Uniform distribution over the relevant documents
    - $E[AP_s]$ computed as average of precisions at judged relevant documents

Extended infAP (xinfAP)

- Precision at a relevant document at rank *k* (2nd and 3rd step)
  - Select a rank at random from the set {1,….,*k*}
  - Output the binary relevance of document at this rank.

  - Probability 1/k pick the current document

$$E[PC_k] = \frac{1}{k} \cdot 1$$

Extended infAP (xinfAP)

- Precision at a relevant document at rank *k* (2nd and 3rd step)
  - Select a rank at random from the set {1,….,*k*}
  - Output the binary relevance of document at this rank.

  - Probability 1/k pick the current document
  - Probability (k-1)/k pick a document above

$$E[PC_k] = \frac{1}{k} \cdot 1 + \frac{k-1}{k} E[PC \text{ above } k]$$

Extended infAP (xinfAP)

- Precision at a relevant document at rank *k* (2nd and 3rd step)
  - Select a rank at random from the set {1,….,*k*}
  - Output the binary relevance of document at this rank.

  - Probability 1/k pick the current document
  - Probability (k-1)/k pick a document above

$$E[PC_k] = \frac{1}{k} \cdot 1 + \frac{k-1}{k} E[PC \text{ above } k]$$

$$E[PC \text{ above } k] = \sum_{\forall s} \frac{N_s^{k-1}}{k-1} \cdot E_s[PC \text{ above } k]$$

Probability of picking a document (above k) from stratum s

Extended infAP (xinfAP)

- Precision at a relevant document at rank *k* (2nd and 3rd step)
  - Select a rank at random from the set {1,….,*k*}
  - Output the binary relevance of document at this rank.

  - Probability 1/k pick the current document
  - Probability (k-1)/k pick a document above

$$E[PC_k] = \frac{1}{k} \cdot 1 + \frac{k-1}{k} E[PC \text{ above } k]$$

$$E[PC \text{ above } k] = \sum_{\forall s} \frac{N_s^{k-1}}{k-1} \cdot E_s[PC \text{ above } k]$$

$$E_s[PC \text{ above } k] = \frac{\#\text{ judged rel above } k \text{ within } s}{\#\text{ judged above } k \text{ within } s}$$

Extended infAP (xinfAP)

- Precision at a relevant document at rank *k* (2nd and 3rd step)
  - Select a rank at random from the set {1,….,*k*}
  - Output the binary relevance of document at this rank.

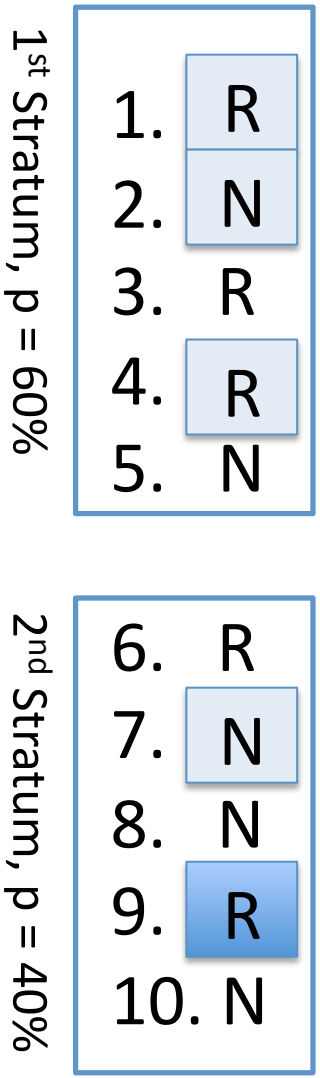
  - Probability 1/k pick the current document
  - Probability (k-1)/k pick a document above

$$E[PC_k] = \frac{1}{k} \cdot 1 + \frac{k-1}{k} E[PC \text{ above } k]$$

$$E[PC \text{ above } k] = \sum_{\forall s} \frac{N_s^{k-1}}{k-1} \cdot E_s[PC \text{ above } k]$$

$$E_s[PC \text{ above } k] = \frac{\#\text{ judged rel above } k \text{ within } s+\varepsilon}{\#\text{ judged above } k \text{ within } s+2\varepsilon}$$

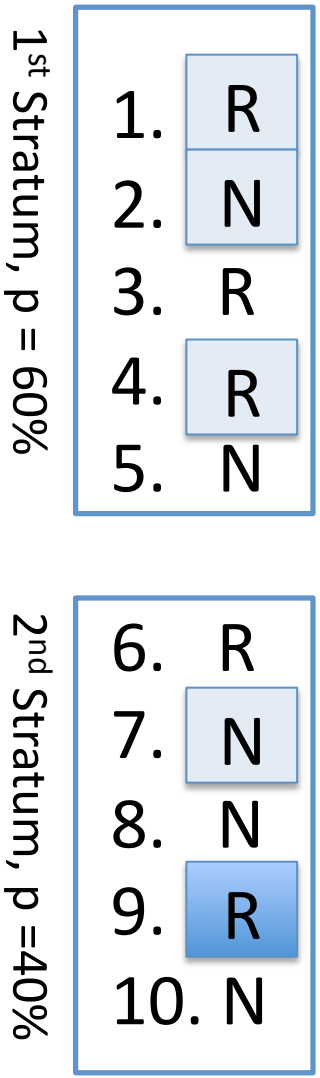
Extended infAP (xinfAP)



$$E[PC_k] = \frac{1}{k} \cdot 1 + \frac{k-1}{k} E[PC \text{ above } k]$$

$$E[PC_9] = \frac{1}{9} \cdot 1 + \frac{8}{9} \cdot \left(\frac{5}{8} \cdot \frac{2}{3} + \frac{3}{8} \cdot \frac{0}{1} \right) = 0.4815$$

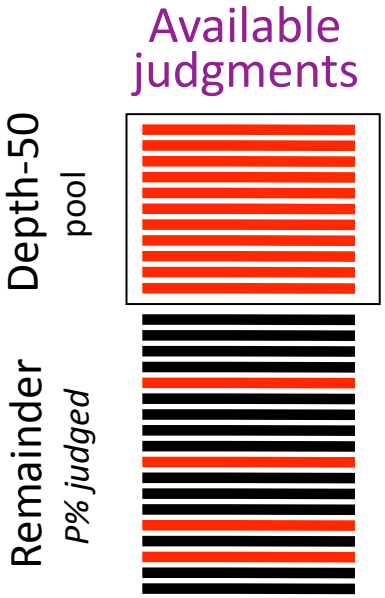
Extended infAP (xinfAP)



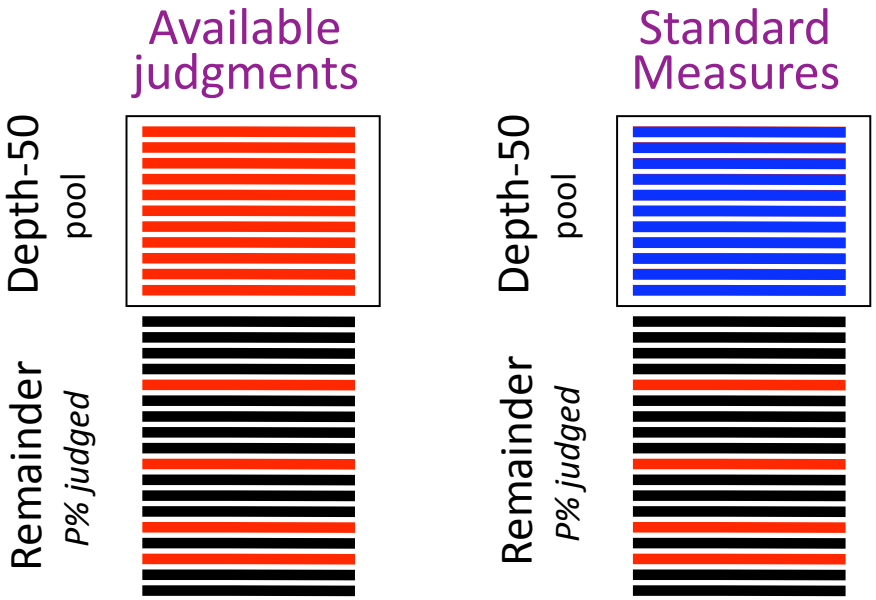
$$E[PC \text{ above } k] = \sum_{\forall s} \frac{N_s^{k-1}}{k-1} \cdot E_s[PC \text{ above } k]$$

$$E[PC_9] = \frac{1}{9} \cdot 1 + \frac{8}{9} \cdot \left(\frac{5}{8} \cdot \frac{2}{3} + \frac{3}{8} \cdot \frac{0}{1} \right) = 0.4815$$

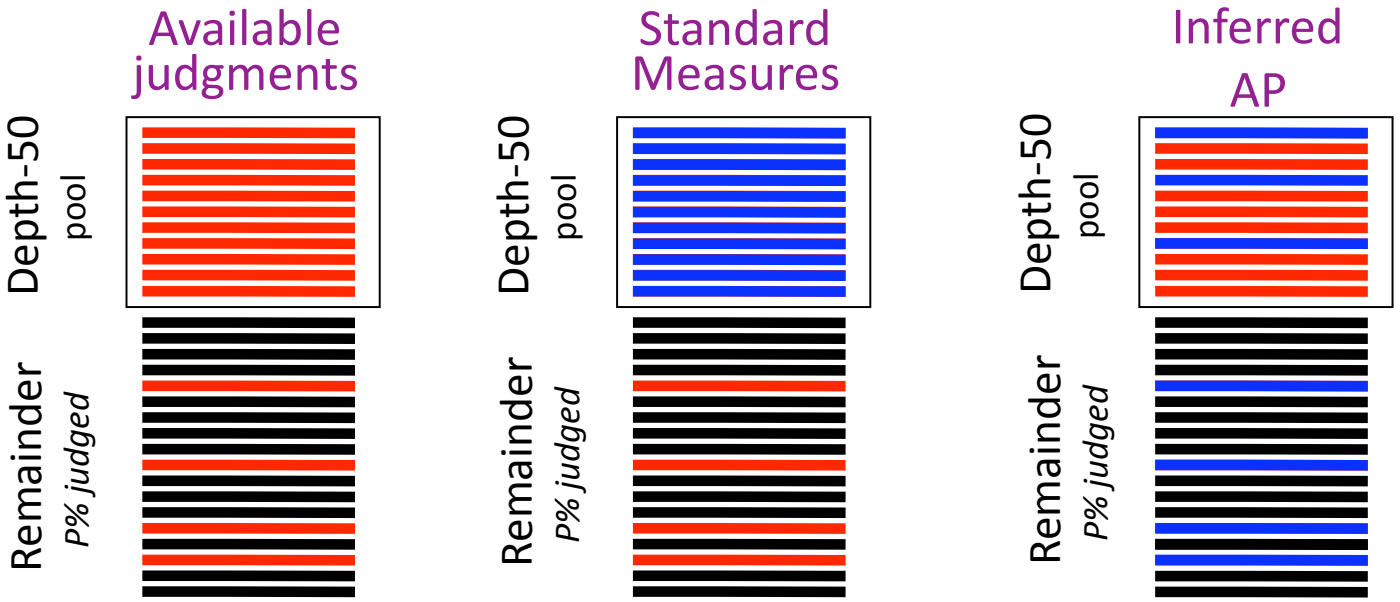
TREC Terabyte ‘06



TREC Terabyte ‘06



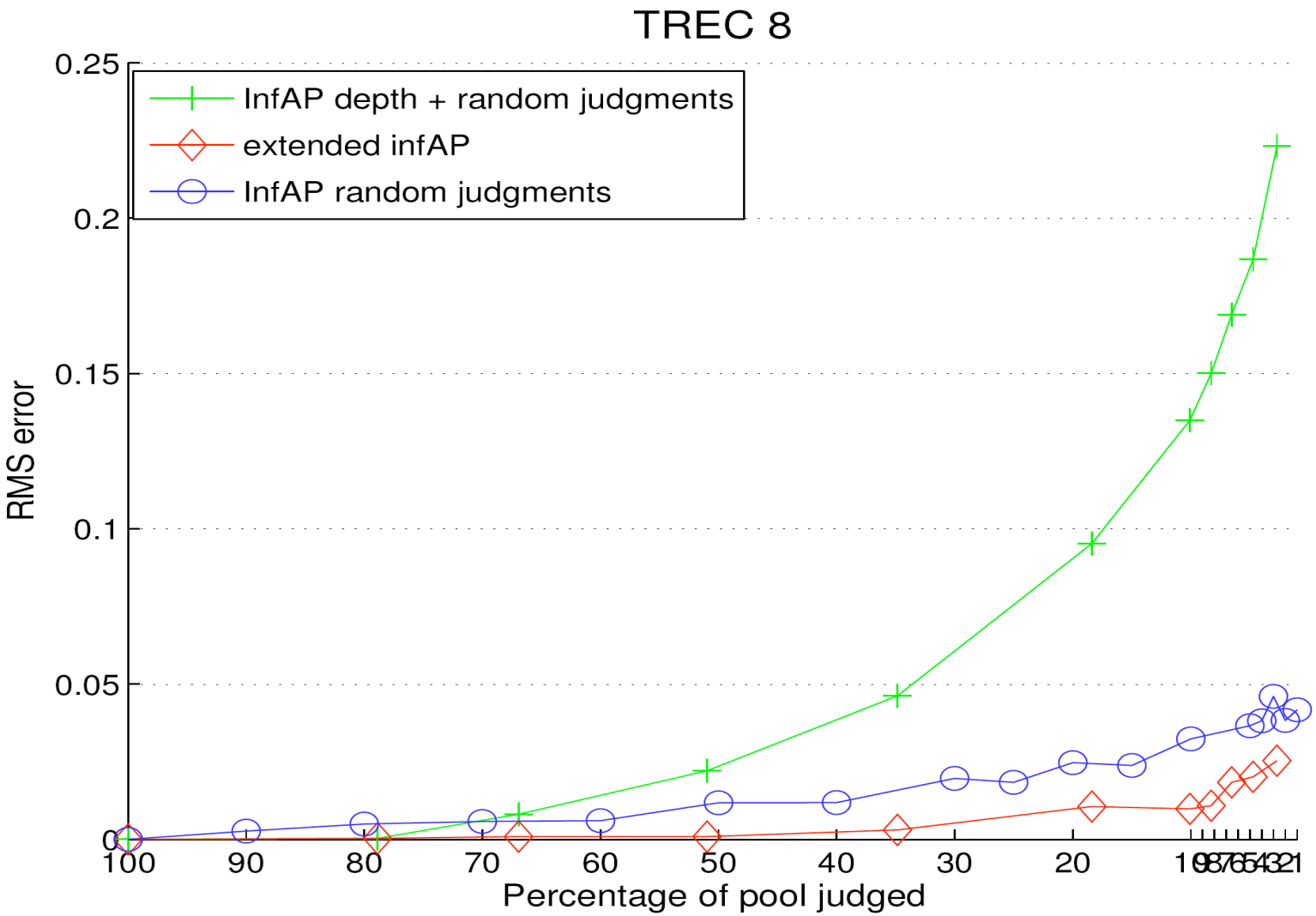
TREC Terabyte ‘06



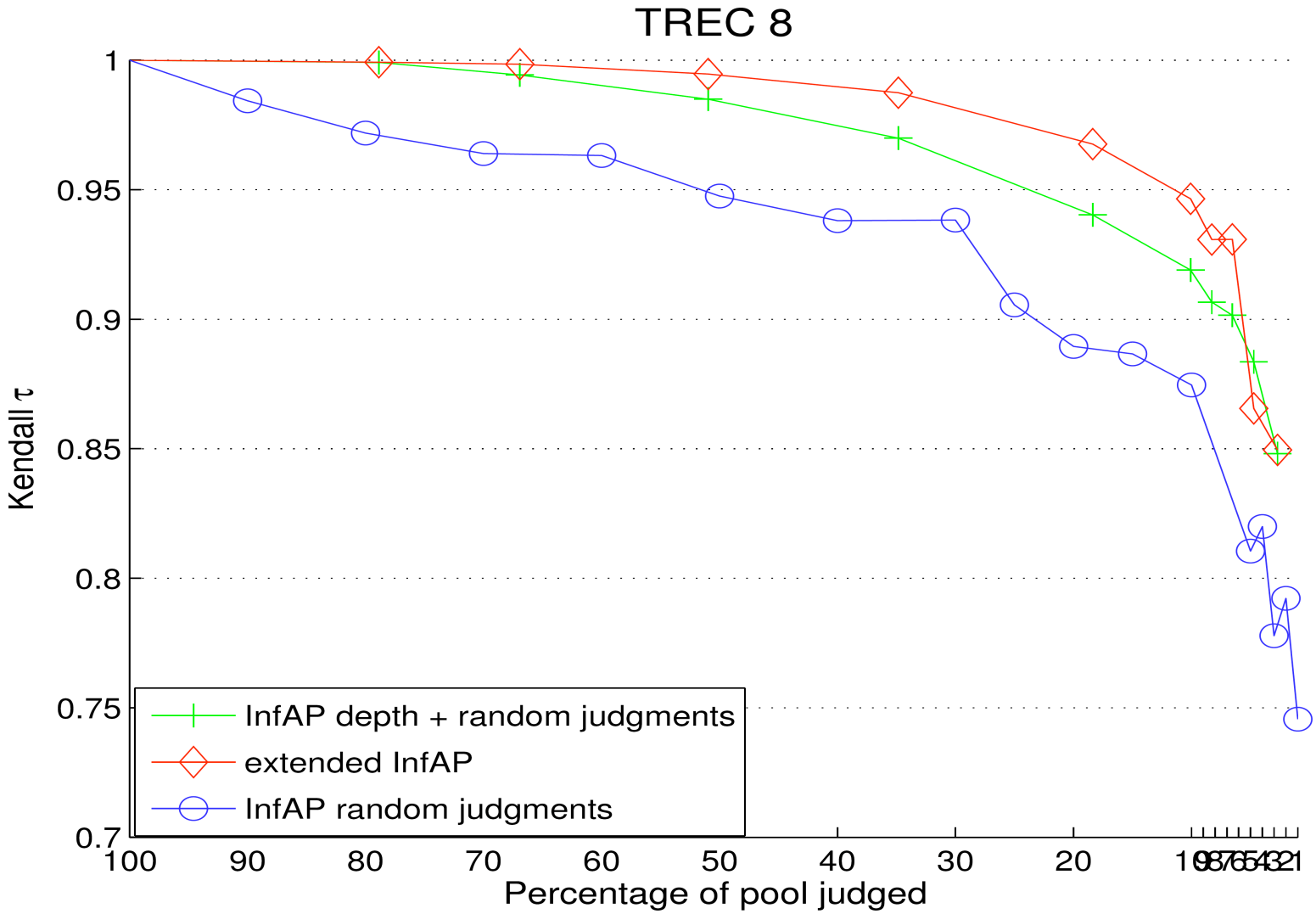
Simulate Terabyte Setup on TREC 8 data

- Assume complete judgments: depth-100 pool
- Form different depth-k pools
  - k ϵ {1,2,3,4,5,10,20,30,40,50}
- For each k compute the total number of documents in depth-k pool
- Randomly sample equal number of documents from the complete judgment set (excluding depth-k pool)
- Assume the remaining documents are unjudged
  - Evaluate search engines with sampled documents

Comparison of the measures : RMS error



Comparison of the measures: Kendall's Tau



Inferred nDCG (infNDCG)

- Apply the same methodology to nDCG

$$nDCG = \frac{DCG}{DCG_I}, \text{where} \quad \text{DCG} = \sum_{i=1}^{Z} \frac{g_i}{\lg(i+1)}$$

- Estimate DCG and DCG_I separately
  - $E[DCG_I]$can be computed using the estimated number of relevant documents (for each relevance grade)

$$\inf nDCG = E[nDCG] = \frac{E[DCG]}{E[DCG_I]}$$

DCG as a Random Experiment

$$\mathrm{DCG} = \sum_{i=1}^{Z} \frac{g_i}{\lg(i+1)}$$

- For each rank *i,* associate a variable $x_i = Z \cdot \frac{g_i}{\lg(i+1)}$
- DCG as a random experiment

1. Select a document at random
   - Rank of the document: *i*

2. Output the value of x_i

Estimating DCG with Incomplete Judgments

- DCG as a random experiment
  1. Select a document at random
     - Rank of the document: *i*

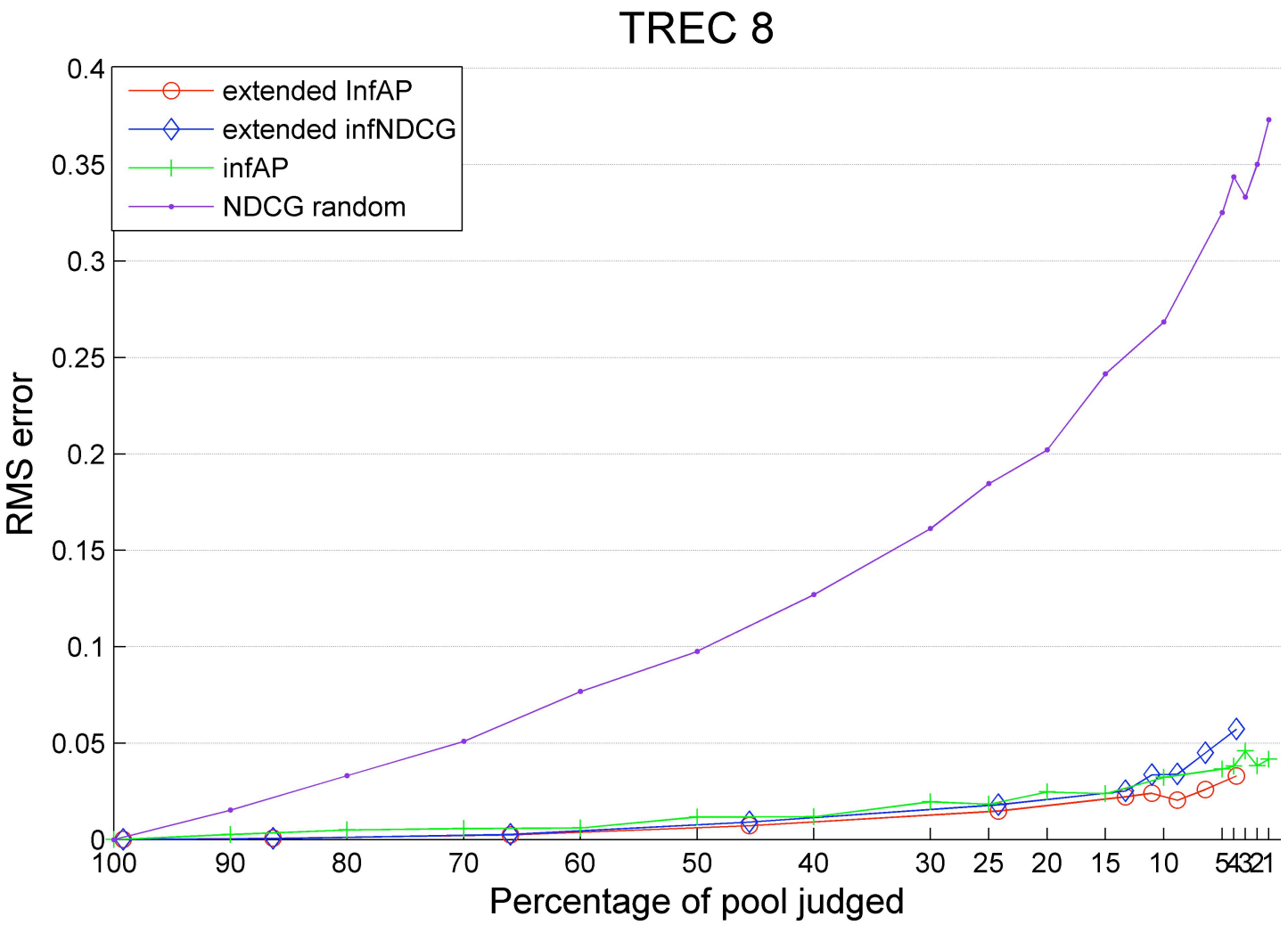
  2. Output the value of $x_i = Z \cdot \frac{g_i}{\lg(i+1)}$

- Due to properties of conditional expectation,

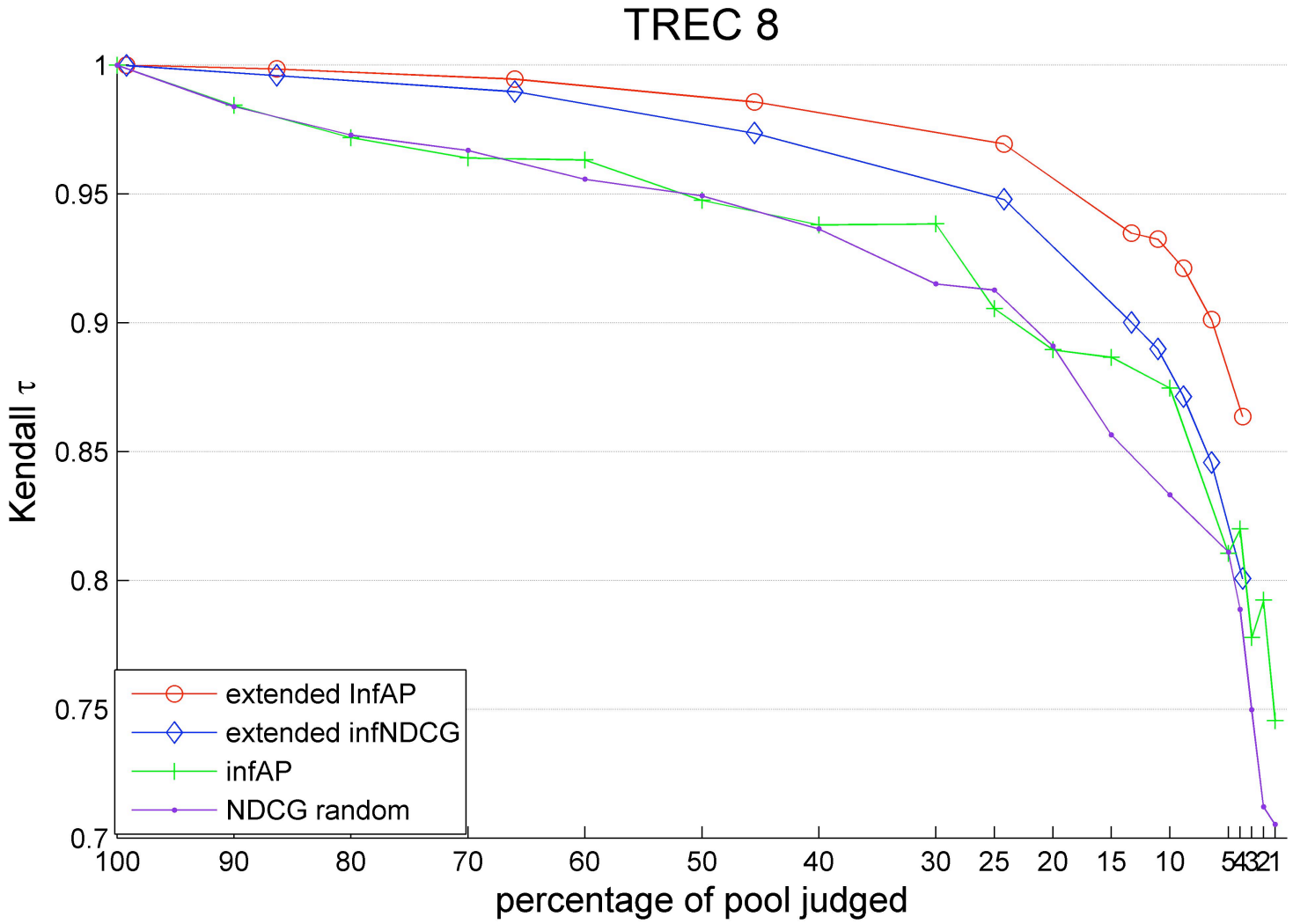
$$E[DCG] = \sum_{\forall s} \frac{Z_s}{Z} E_s[DCG]$$

$$E_s[DCG] = \frac{\sum_{\forall j \in sampled_s} x_j}{|sampled_s|}$$

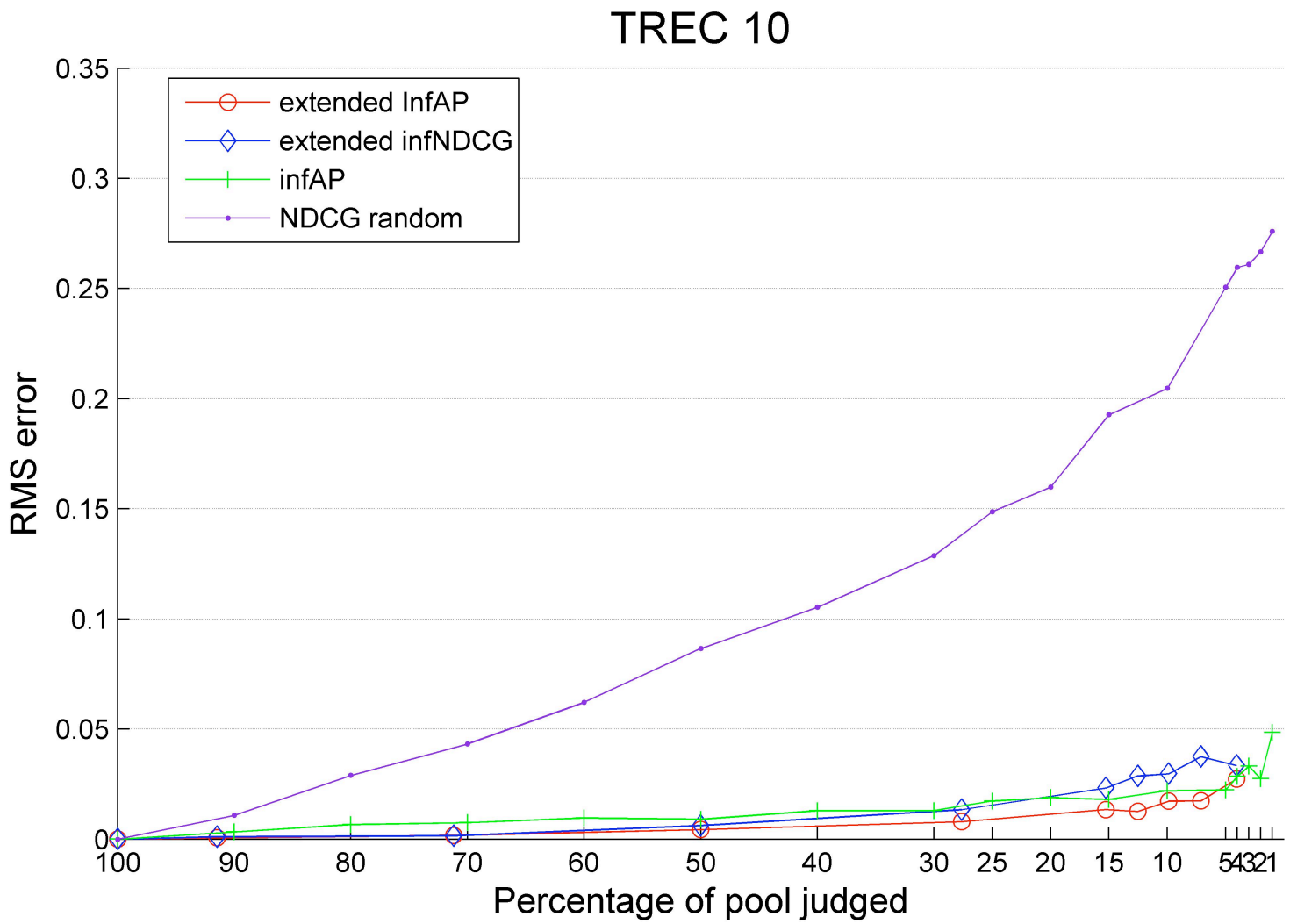
Overall Results: TREC8



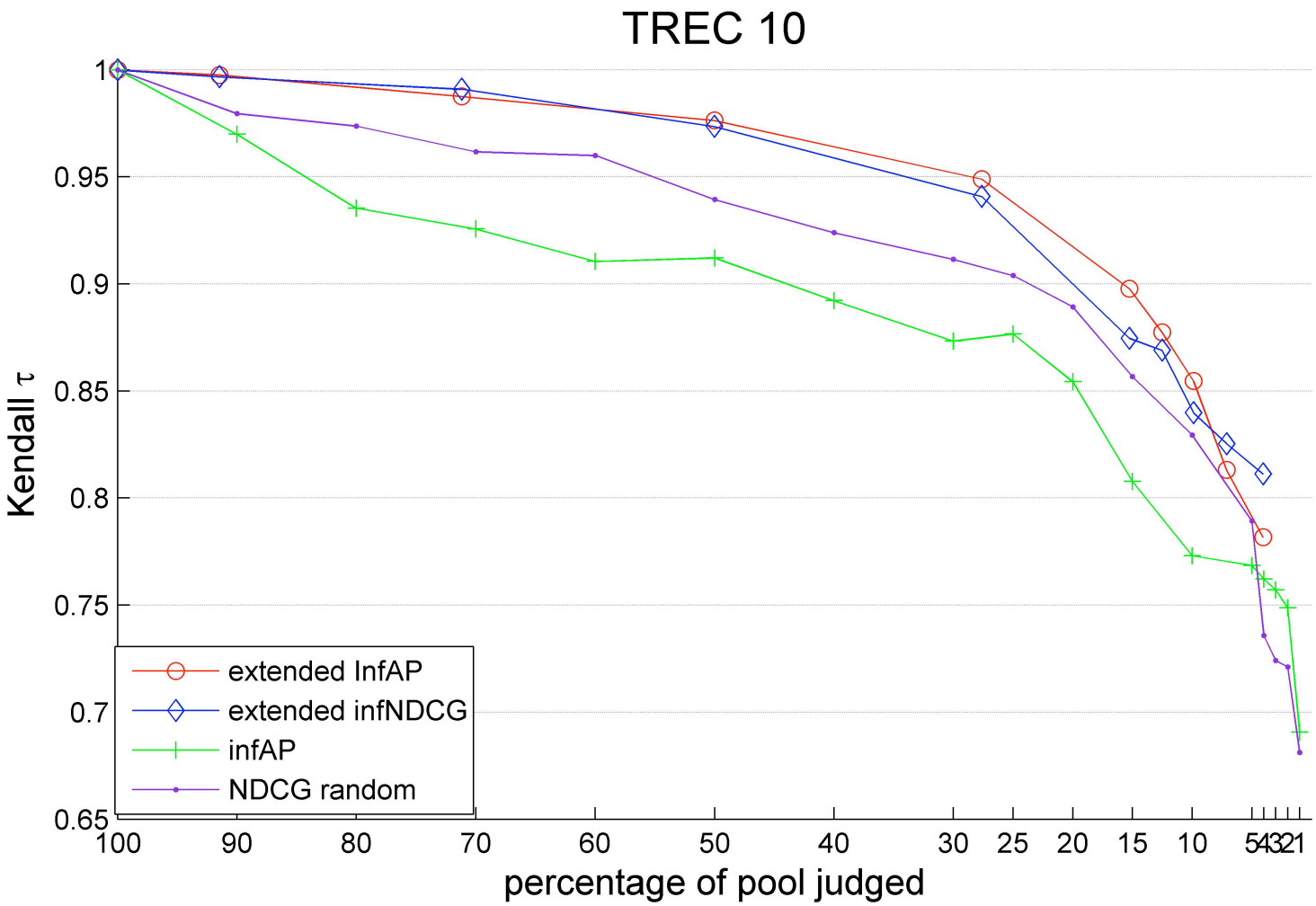
Overall Results: TREC8



Overall Results: TREC10



Overall Results: TREC10



Conclusions