Measuring Test Collection Reusability

Ben Carterette, Evgeniy Gabrilovich, Vanja Josifovski, and Donald Metzler

Dept. of Computer & Information Sciences, University of Delaware
Yahoo! Research

Test Collections

- A *test collection* for evaluating a retrieval task comprises:
  - A corpus
  - A sample of queries
  - A set of relevance judgments
    - Usually drawn from a set of existing rankings
- After initial cost of construction, test collections allow virtually free off-line, system-based evaluation of ranking functions

Reusable Test Collections

- Test collection construction cost is high
- *Reusable* test collections amortize the cost over many off-line evaluation
  - Use the same test collection to evaluate "new" ranking functions
- Reusability may be limited when:
  - Not enough queries in the sample
  - Not enough judgments per query
  - New systems retrieve many unjudged documents

Previous Work

- One of TREC's goals is reusable test collections
  - Encourage diverse set of system submissions
  - Judge them deeply (depth 50 or 100)
  - Produces tens of thousands of judgments that cover most uses [Zobel '98; Voorhees '01]
- One assumption often made:
  - Unjudged documents are not relevant
  - Reasonable when there are many judgments

Previous Work

- How many and what types of errors should we expect when reusing collections?
  - Possibly under different assumptions about unjudged documents
- Buttcher et al. (2007), Sakai (2008), Sanderson & Joho (2004), Buckley et al. (2006)

Our Approach

- Quantify reusability in terms of *uncertainty*
  - Uncertainty when estimating the effectiveness of a new system using an existing set of judgments
  - Greater uncertainty → less reusability
- How to measure uncertainty?
  - Confidence intervals (means, std. errors)
  - Easy to compute measures that correlate well with the std. error

Expectation and Variance

- Relax assumption about unjudged documents:
  - Each unjudged document has a probability distribution over relevance grades
- Use distributions to calculate expectations and variances of evaluation measures
  - Summing over all possible judgments to all unjudged documents

Example

prec@5 = 1/5 under usual assumption

Depending on how ?s resolve:

- RNNNN, prec@5 = 1/5; RNNNR, prec@5 = 2/5
- RNRNN, prec@5 = 2/5; RNRNR, prec@5 = 3/5
- RRNNN, prec@5 = 2/5; RRNNR, prec@5 = 3/5
- RRRNN, prec@5 = 3/5; RRRNR, prec@5 = 4/5

Interval Measures of Reusability

- Step 1: For measure *m*, estimate its expectation and variance for each query in a set of n, e.g.

$$E[prec@5] = \frac{1}{5}\sum_{i=1}^{5} p_i \qquad Var[prec@5] = \frac{1}{5^2}\sum_{i=1}^{5} p_i(1-p_i)$$

- Step 2: Estimate mean and variance of *mean*(*m*)

$$\mathrm{E}[\overline{m}] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[m(Q_i)] \qquad \mathrm{Var}[\overline{m}] = \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}[m(Q_i)]$$

- Step 3: Construct confidence interval

$$\left[\mathrm{E}[\overline{m}] - z_{\frac{\alpha}{2}}\sqrt{\frac{\mathrm{Var}[\overline{m}]}{n}}, \mathrm{E}[\overline{m}] + z_{\frac{\alpha}{2}}\sqrt{\frac{\mathrm{Var}[\overline{m}]}{n}}\right]$$

Estimating p_i

- For judged documents, p_i is 0 or 1
- Estimate p_i for unjudged documents using logistic regression classifier
- Features
  - Similarity of unjudged docs to judged docs
  - Precision, average precision, etc.
  - Number of results judged, relevant, non-relevant, unjudged

Alternatives to CIs

- Calculating confidence intervals is expensive
  - Training relevance classifiers for each query in a large set is time-consuming
  - Time complexity of variance can be high
    - $O(n^3)$ for average precision
  - Complexity also grows with number of judgment grades
- Can we use point measures to predict the width of the confidence interval?

Point Measures of Reusability

- Recall, precision, other evaluation measures
- Number of unjudged documents
- Mean Average Reuse

$$reuse@k(Q) = \frac{judged@k(Q)}{k}$$

$$AR(Q) = \frac{1}{judged(Q)} \sum_i reuse@i(Q)$$

- Recall + Mean Average Reuse

$$\widehat{\sigma} = -0.0023 + 1.1154 \cdot rec + 0.1088 \cdot MAR$$

Experiment

- Investigate whether confidence interval width provides good sense of ability to evaluate system
- Determine accuracy of CI width predictions
- Outline of experiment:
  - Start with a full test collection
  - Reduce its judgments in a principled way
  - Use smaller set to estimate CIs
  - Compare to “true” performance

Data

All systems submitted to two TREC tracks:

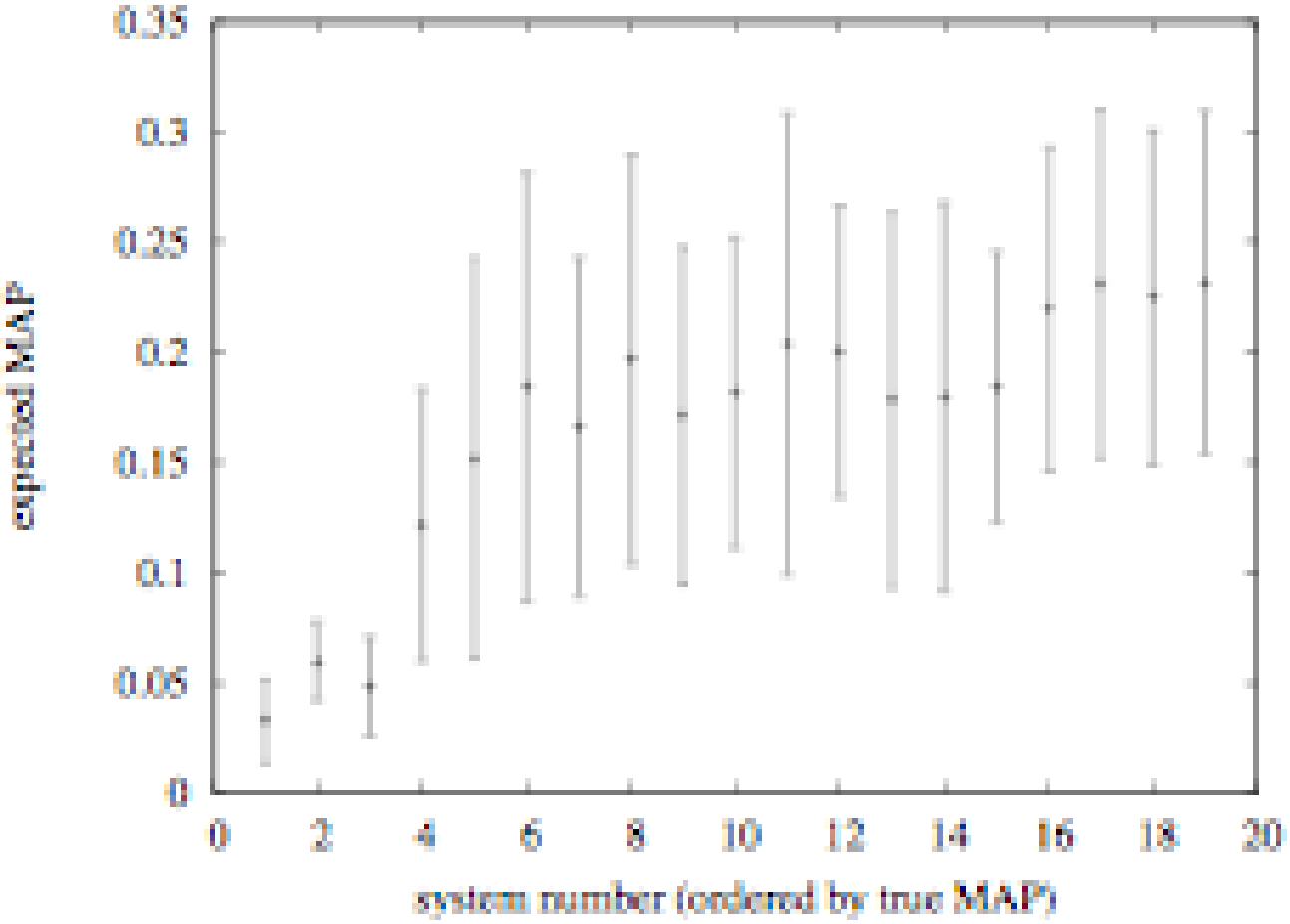
track	sites	systems	topics	judgments	relevant
Web 2004	18	74	225	88,566	1,763
Robust 2005	17	74	50	37,798	6,561

- Judgments were collected using pooling
  - Top k documents from systems from each submitting site were pooled together and judged
- Systems reflect a very diverse set of retrieval methods
- Tracks reflect different retrieval tasks
  - Web includes high-precision tasks
  - Robust is a high-recall task

Basic Experimental Procedure

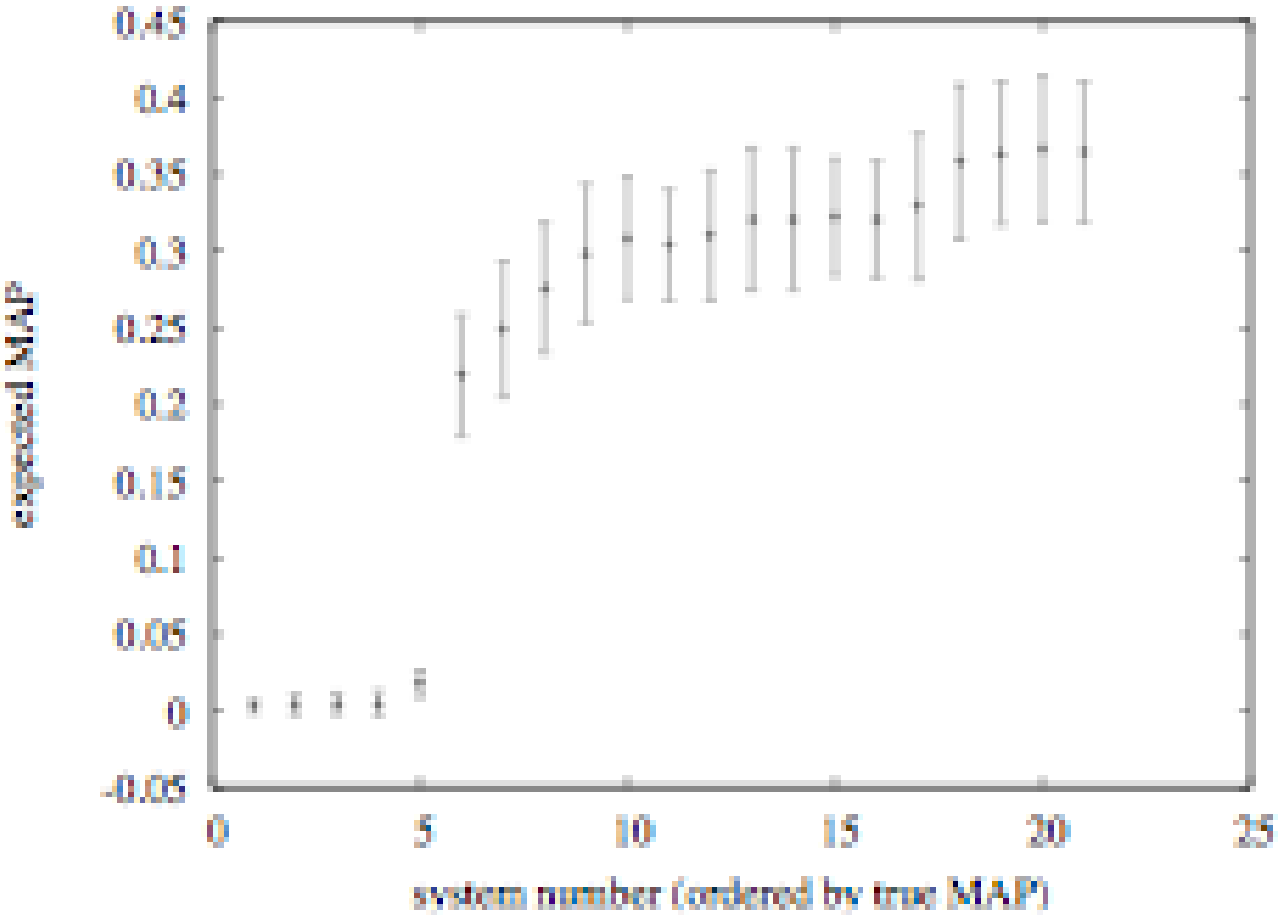
- Simulate pooling and judging to form a smaller set of relevance judgments
  - Select m sites to form the pool
  - Judge their systems to depth k
- Use that pool to evaluate the systems that did not contribute to it
  - i.e. train a relevance classifier, compute means and confidence intervals

Example Confidence Intervals



Means and 95% confidence intervals for 19 systems from 5 sites evaluated using a pool of depth 10 from 5 systems from 1 site

Example Confidence Intervals

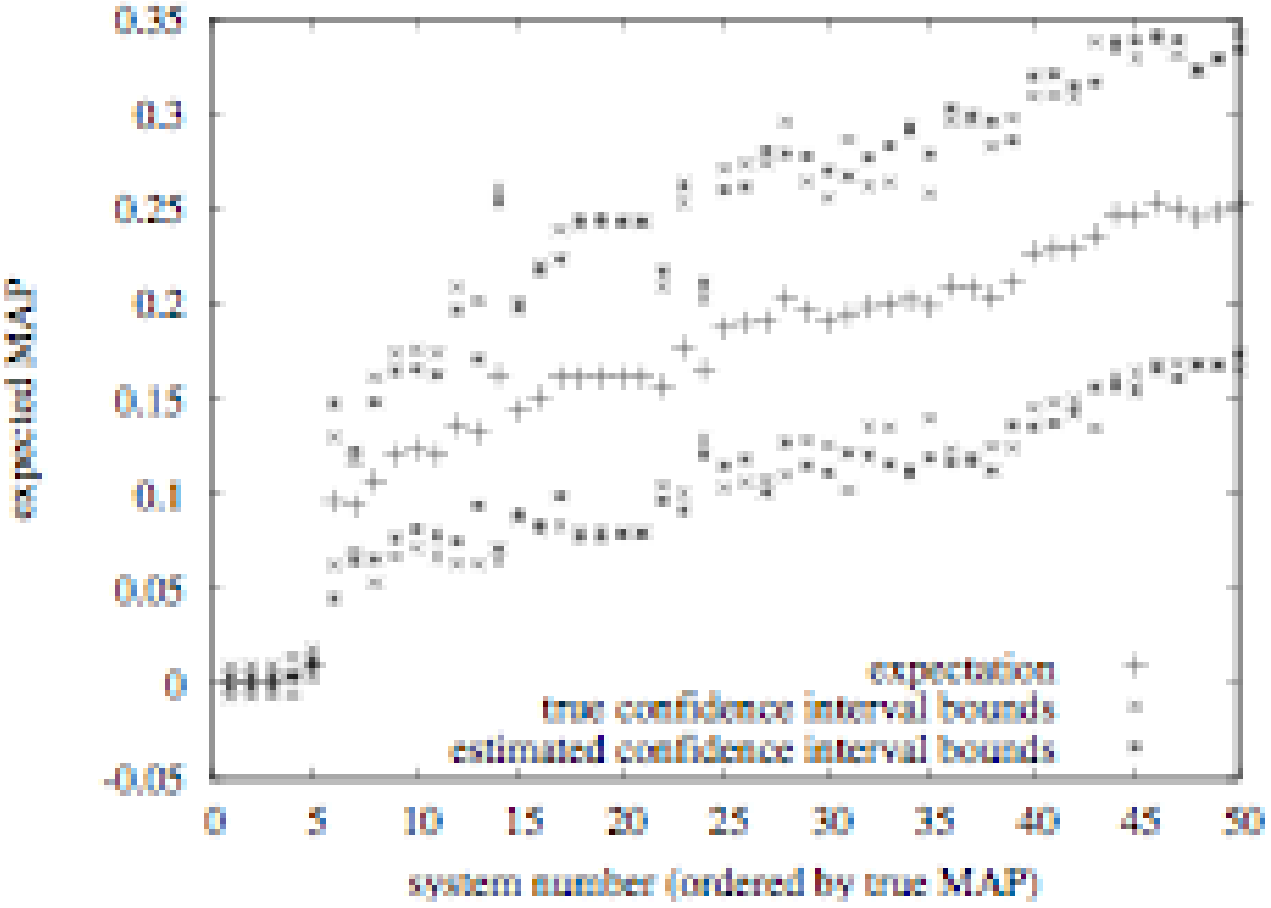


Means and 95% confidence intervals for 21 systems from 5 sites evaluated using a pool of depth 10 from 14 systems from 3 sites

Full Experimental Process

1. Choose m_1 sites to form the pool of depth k
2. Choose m_2 sites to calculate full CI from pooled judgments
3. Learn CI width as a function of one or more point estimates
4. Predict CI width on remaining systems; compare to “true” value

Example CI Prediction



True and predicted confidence intervals based on depth-10 pool of one site

Results

m_1	m_2	k	judged	rel	τ_{test}	σ_{test}	$\rho_{\sigma_{m_2},\widehat{\sigma_{m_2}}}$
1	5	1	644	103	0.536	0.010	0.899
1	5	5	2,247	224	0.855	0.026	0.953
1	5	10	3,885	257	0.870	0.025	0.915
1	5	20	7,367	392	0.901	0.022	0.928
1	10	5	2,197	199	0.829	0.028	0.904
1	10	10	4,175	269	0.851	0.025	0.888
1	10	20	7,200	371	0.875	0.023	0.865
3	5	1	1,028	206	0.812	0.025	0.891
3	5	5	5,267	361	0.926	0.026	0.915
3	5	10	10,372	520	0.960	0.017	0.921
3	5	20	20,573	668	0.968	0.012	0.869
5	5	1	1,724	252	0.872	0.031	0.905
5	5	5	7,287	478	0.960	0.018	0.892
5	5	10	15,941	611	0.970	0.013	0.871
5	5	20	27,000	777	0.975	0.009	0.855

Conclusions

- Reusability measures can be used to determine if offline tests will yield accurate estimates of online effectiveness
- Confidence interval estimates are most flexible
  - Practitioners can determine if estimates are within acceptable bounds
  - Can be complicated to compute
- Points measures strongly correlate with standard error of estimates


Questions?