Test Collections for Adaptive IR

Ellen Voorhees

NIST

National Institute of Standards and Technology
Technology Administration, U.S. Department of Commerce

"Many forms of Government have been tried, and will be tried in this world of sin and woe. No one pretends that democracy is perfect or all-wise. Indeed, it has been said that democracy is the worst form of government except all those other forms that have been tried from time to time."

Sir Winston Churchill
November 11, 1947

No one pretends that test collections are perfect or all-wise. Indeed, it has been said that test collections are terrible for IR research except that they're better than current alternatives.

Me

Traditional Cranfield

- Test collection abstraction
  - originally created specifically to avoid user effect
  - rationale is that abstracted task is a necessary but not sufficient proxy for real user task
    - Sparck Jones calls such an abstract task a "core competence"
- Consequences of abstraction
  - gain control over variables that enable more experimental power at lower cost
  - lose level of realism not accounted for in abstracted task

So it's simple…

- A test collection for adaptive/interactive IR is a new abstraction that
  - includes a minimal amount of information necessary for representing the salient aspects of adaptation
  - while retaining as many of the benefits of Cranfield test collections as possible
    - benefits include experimental power, relatively low cost, generalizability of results, broadly useful

Except that…

- All evidence to date suggests that even tiny extensions to Cranfield collections imply either huge increase in cost or huge loss of power for IR experiments
- If we accept that cost and decide to go ahead, there is no consensus/data to suggest what constitutes a core competence for adaptive IR

What Evidence?

- Even for Cranfield, "user effect" is single biggest variable
  - here, "user effect" is topic + relevance judgments
  - strong negative correlation between variability and experimental power
  - strong positive correlation between power and cost
  - all additions to Cranfield will increase variability

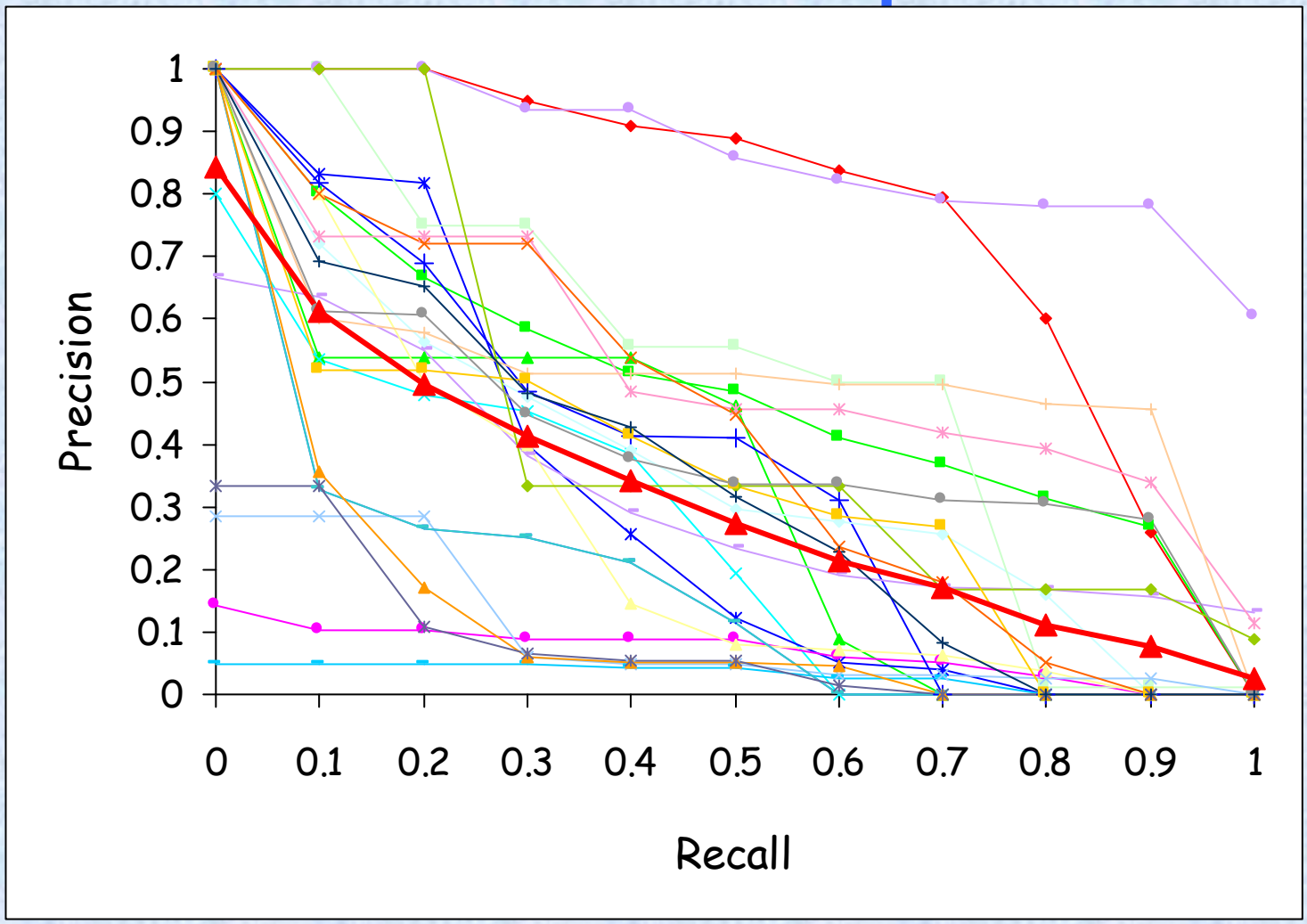
What Evidence?

- Long history of interactive experiments illustrates difficulties
  - Belkin's triumph at SIGIR'94
  - TREC-6 experiments showed common system comparison not worth its expense
  - Hersh and Turpin experiments show difficulty of demonstrating improvement in interactive setting [SIGIR 2000, 2001, 2006]
  - TREC 2004 HARD track

User Effect in Cranfield

- Anova test of TREC-3 results showed:
  - system, topic, interactions all significant
  - topic effect generally much larger than system effect
  - [Banks, Over, and Zhang, 1999]
- More informally, it's easy to find some one query for which your system is best

Interpolated R-P Curves for Individual Topics



Sensitivity Analysis

- With archive of TREC results, have empirically determined the relationship between number of topics, Δ of scores, & error rate [Voorhees & Buckley, 2002]
  - error rate decreases as topic set increases
  - error rate decreases with larger Δ, but then power is reduced
- Sakai [2006] reaches same conclusions
  - more defensible mathematical underpinnings
  - no need for extrapolation
  - (used NTCIR archive rather than TREC results)

Basic Procedure

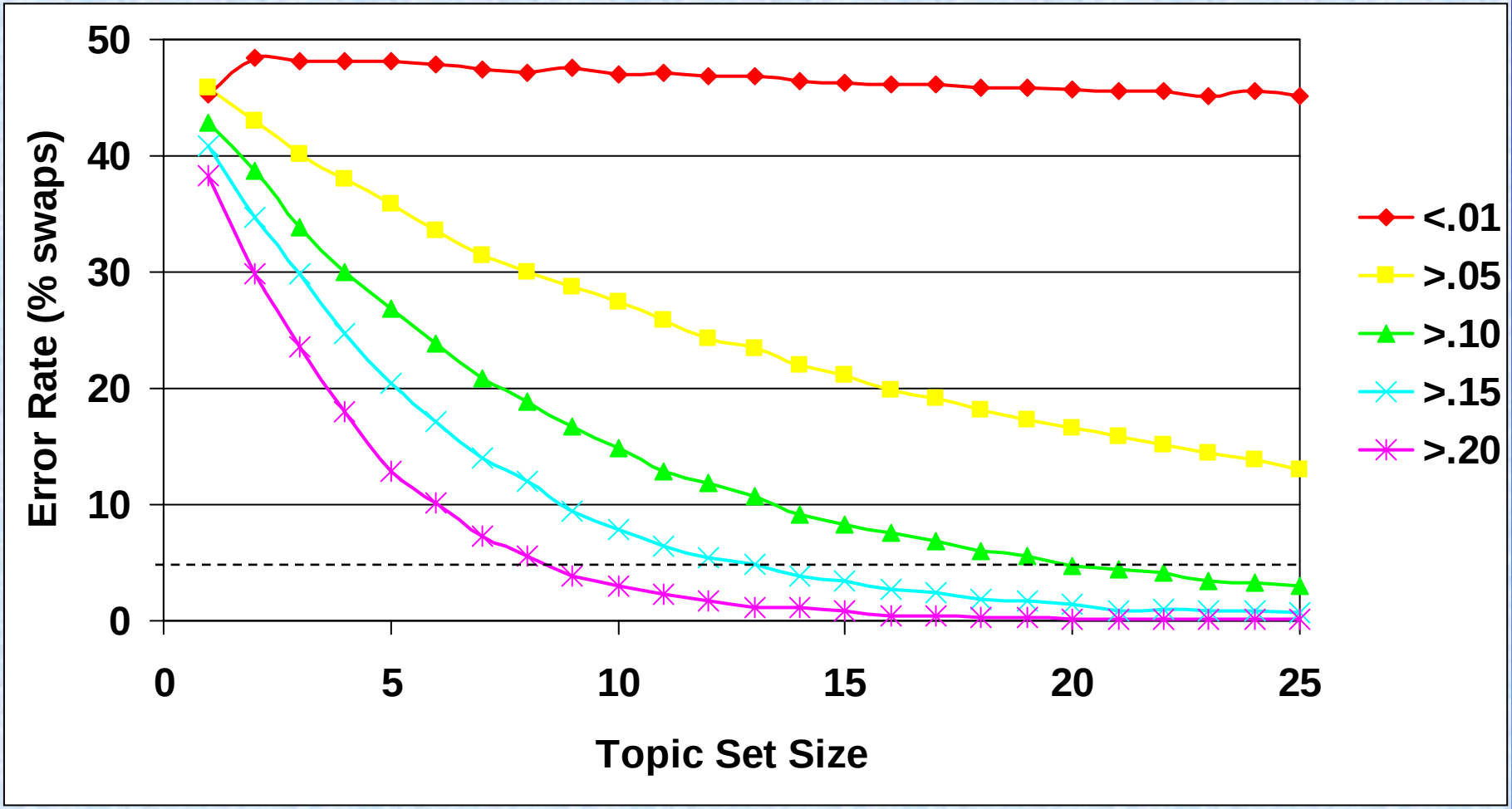
- Estimate likelihood of a changed decision by comparing two runs many times using different topic sets, counting the number of times A>B and B>A
- Probability of a changed decision for one run pair is

$$P(\text{swap}) = \frac{\text{MIN}(|A>B|, |B>A|)}{\text{number of comparisons}}$$

Basic Procedure

- Average P(swap) over many different pairs of runs
- Compute probability as a function of topic set size and observed difference between scores by conditioning counts accordingly
- P(swap) defines an error rate in that it specifies how likely it is that the outcome of one experiment leads to the wrong conclusion

Error Rate by Topic Set Size



Grand averages computed over all TRECs, 50 trials per run pair

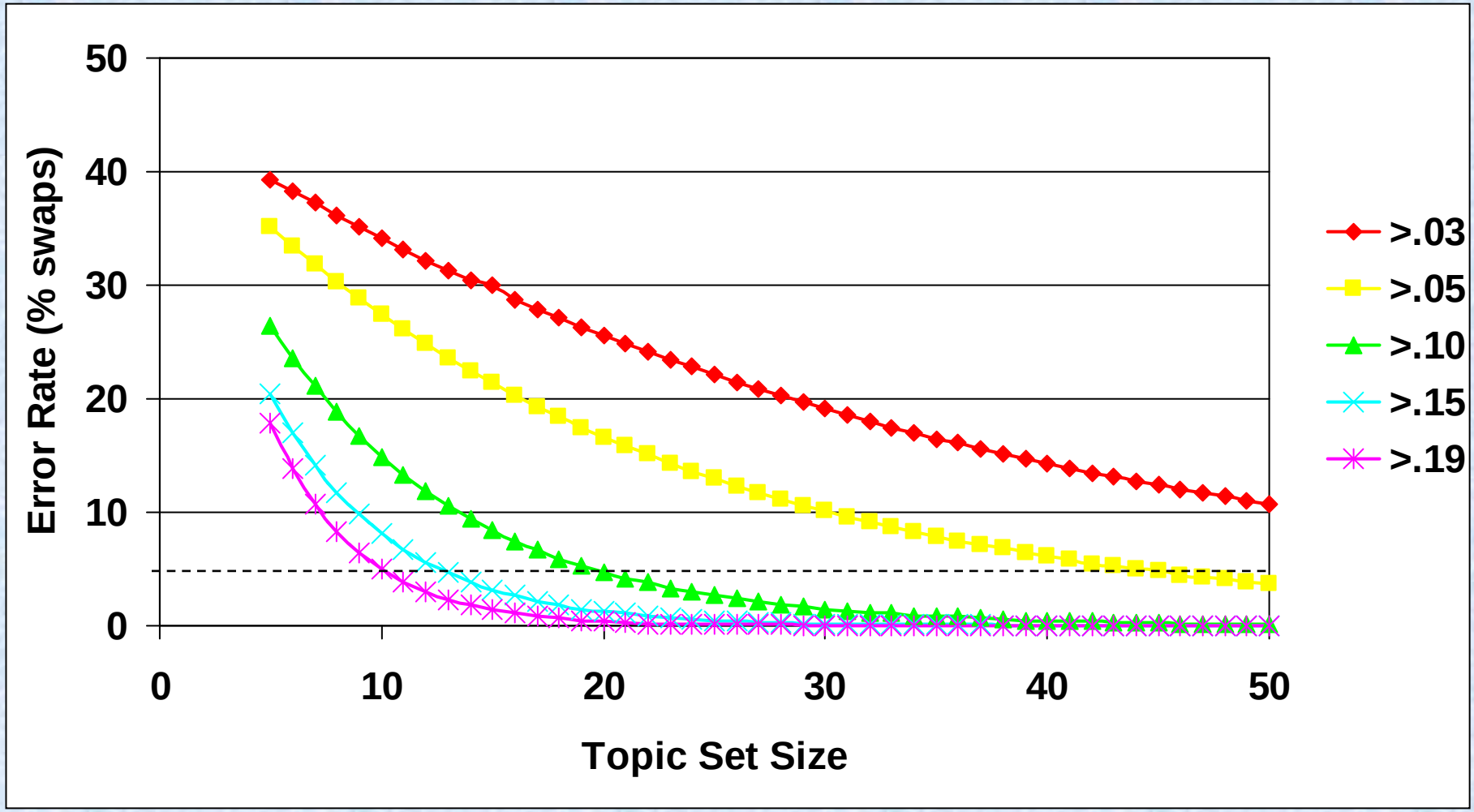
25 is Small!

- For topic set size of 25,
  - P(swap) = 13% when $0.05 \leq \delta < 0.06$ ("noticable" difference)
  - P(swap) = 3% when $.010 \leq \delta < 0.11$ ("material" difference)
  - and these are **absolute** differences, not **relative** differences
    - best runs on TREC collections have ~ 0.3 MAP
    - 0.1 absolute difference at least 33% relative

☛ Topic set sizes of 25 are too small for reliable comparisons with common Δ's

- P(swap) = 23% when $.03 \leq \delta < .04$

Extrapolated Error Rates



50 is Adequate?

- For topic set size of 50,
  - $P(swap)_{extrapolated}$ = 4% when $.05 \leq \delta < .06$
  - $P(swap)_{extrapolated}$ = 1.5% when $.10 \leq \delta < .11$
- But a 10% relative difference for quality TREC runs (absolute difference of about .03) still has error rate > 10%
  - $P(swap)_{extrapolated}$ = 11% when $.03 \leq \delta < .04$

Number of Topics

- The critical factor in reliability of results
  - in most stable case, 25 topics is probably too few; anything less untenable
    - system comparison using standard test collection
    - MAP as evaluation measure
  - even more topics needed in other cases
    - system comparisons using less stable measures
    - user-in-the-loop studies need vastly more topics
      - introduce even larger amount of variability
      - Robertson [1990] calculates hundreds of topics per user to obtain significance in non-matched-pair tests

Interactive Studies

- Costs very high:
  - start-up: need complete systems for all alternatives
  - need even more topics than for system comparisons, but more topics imply more subjects
  - continued tension between "reality" and generalizability

TREC-6 Interactive Track

- Explicit design to perform cross-site comparisons by comparing to common baseline system
  - attempt to get effect of n^2 comparisons of systems by only performing n comparisons
  - minimum experimental design used six topics in Latin square; some sites used multiple blocks
  - assumption that common system would control for inter-site variance not supported in subsequent analysis and incurs its own costs

Hersh and Turpin

- Series of experiments to validate that results from system comparisons are actually meaningful to users
  - concluded that better relevance ranking did not translate into better performance on a specific task… (e.g., 67% improvement in MAP led to negligible improvement in number instances found)
  - …so concluded that system comparisons not trustworthy
  - I contend results are actually an indictment that their interactive experiments had little power & as such they are a good case study of why interactive experiments are difficult

TREC 2004 HARD Track

- High Accuracy Retrieval from Documents
- Goal: improve ad hoc retrieval by customizing the search to the user
  - current systems return results for "average" user
  - necessarily limits effectiveness of system for particular user
- Ad hoc task with additional information
  - metadata supplied in topic statement
  - information collected from *clarifying form*
  - varying unit of retrieval (passage vs. full doc)

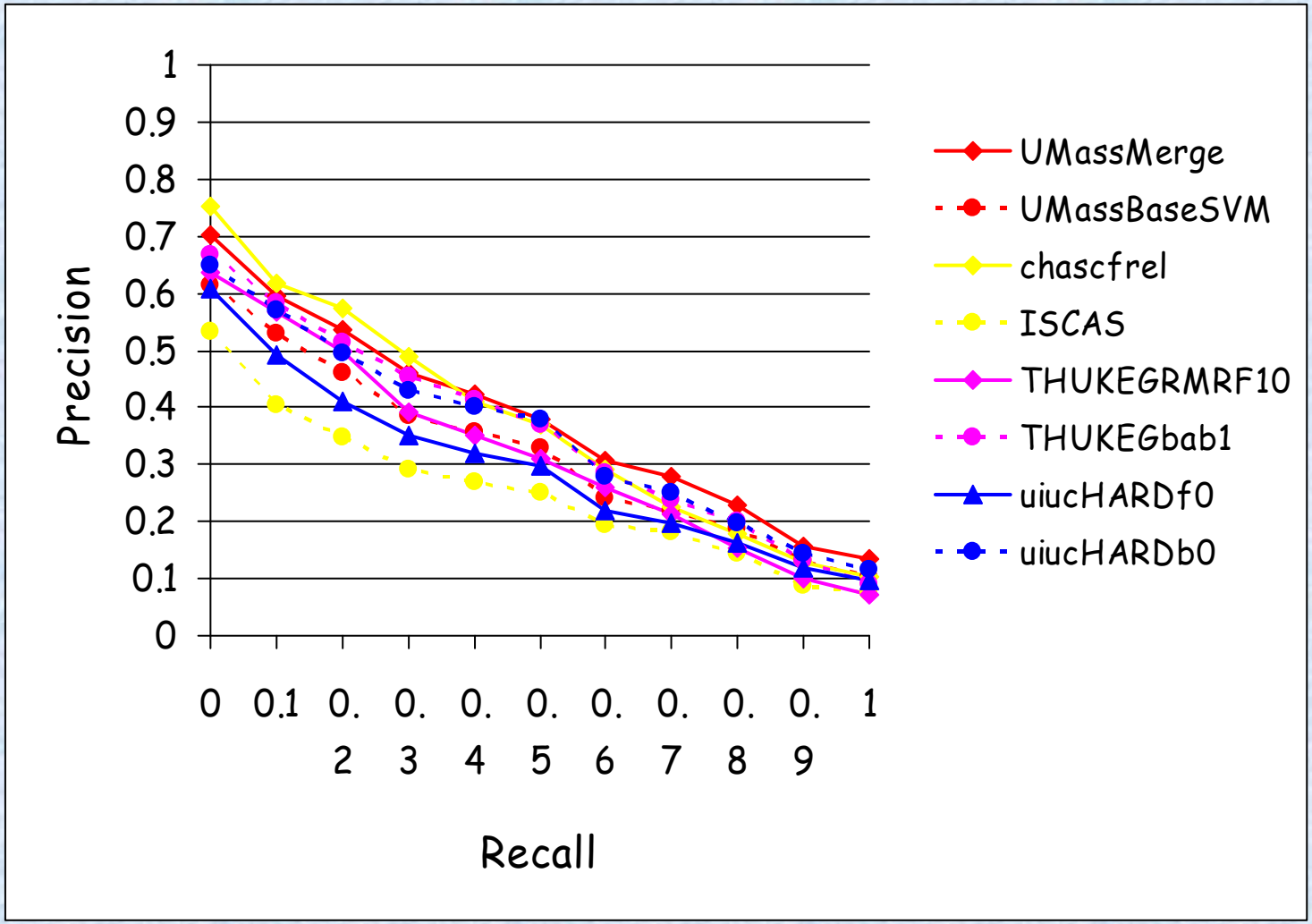
Additional Information

- Metadata from topic statements
  - familiarity [little, much]
  - genre [news-report, opinion-editorial, other, any]
  - geography [US, non-US, any]
  - subject domain [free text]
  - related text (either on-topic or relevant)
- Clarifying forms
  - assessor (surrogate user) spends at most 3 minutes/topic responding to topic-specific form
  - example uses:
    - sense resolution
    - relevance judgments

HARD Protocol

- Perform baseline runs using standard topics
- Receive extended topics and/or clarification form responses
- Perform additional (non-baseline) runs exploiting additional info

Top HARD runs vs. Baseline



Sorted by MAP of higher run using HARD-rel judgments

HARD Track

- Explicit (and reasonable) attempt at defining "interactive lite" task
  - still have problems with insufficient topics per category…
  - …so power still lower than traditional ad hoc task
  - …still have increased costs in development of categories, topics in categories, clarification forms, etc.

Adaptive IR Test Collections

- Cranfield is successful because of its carefully calibrated level of abstraction
  - contains sufficient fidelity to real user tasks to be informative
  - is sufficiently abstract to be broadly applicable, feasible, relatively inexpensive
- Adaptive IR test collections need a new, but also carefully selected abstract task
  - no agreement in AIR literature on what characterizes an adaptive task
  - vital that characteristics to be modeled be minimal set possible

Ways Forward

- Definition of the truly distinguishing features of adaptive IR
- Development of a protocol that captures just those features
  - don't want a characterization of an entire user task
  - instead, focus on just the intrinsic features that define the core competency

A Final Cautionary Note

- For test collections, bias is much worse than incompleteness
  - smaller, fair judgment sets always preferable to larger, potentially-biased sets
  - need to carefully evaluate effects of new pool building paradigms with respect to bias introduced
- Operational tests subject to biases that are difficult to assess or control
  - such tests difficult to generalize as a result