

Towards Good Evaluation of Individual Topics

Chris Buckley – Sabir Research

Current Test Collection Situation

- Cranfield Methodology
 - Goal is to fairly compare systems
 - Fixed static document collection
 - “Large” number of fixed topics
 - Fixed relevance judgments, from single user per topic
 - Binary, or slightly better, levels of relevance
 - Various evaluation measures, depending on goals

Need For Many Topics

- Several papers have shown we want 50+ topics
 - Buckley, Voorhees Sigir 2004
- Caused by single topic uncertainty
 - System-topic interactions
 - Unknown topic difficulty
 - Uncertainty due to choice of measure
 - Uncertainty in actual measurement
 - Uncertainty due to relevance judgments
- We accept first two causes, for the most part we ignore the last three

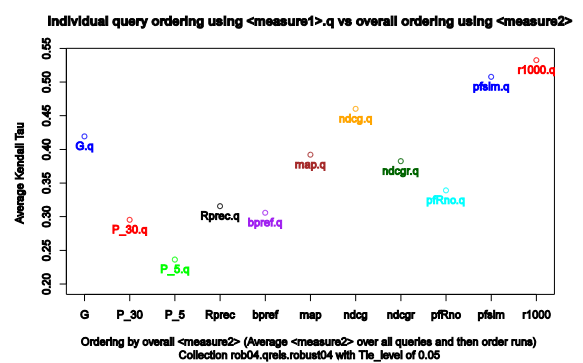
Costs of Poor Single Topic Evaluation

- Requires more topics
- Forces a focus on averages of measures
- Failure analysis is very difficult
 - Individual topic measure numbers can’t be trusted
 - Is improvement due to solving system-topic interaction problem, or just random chance due to uncertainty
- No bounds on measurement error
 - Needed for some environments (legal eDiscovery)

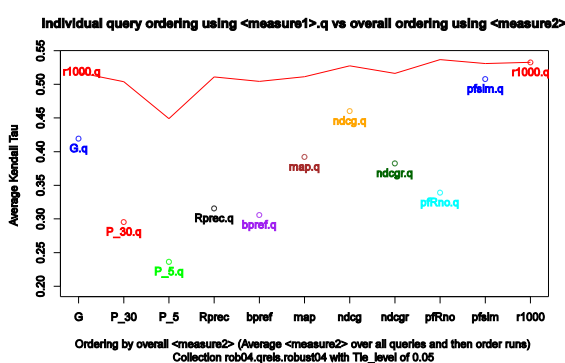
Current Individual Topic Measure Values

- How good are they?
 - Compare ranking of systems on individual topics with the overall ranking of systems. (Kendall Tau)
- Look at what makes a measure better on individual topics
- Initial plots are the Robust04 Track
 - 249 topics
 - All runs are automatic
 - Large number relevance judgments, “Complete”

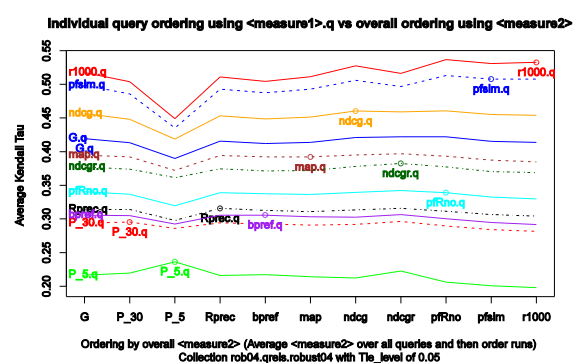
Topics Predicting Overall Rankings (Same Measure)



Topics Predicting Overall Rankings (Recall 1000)



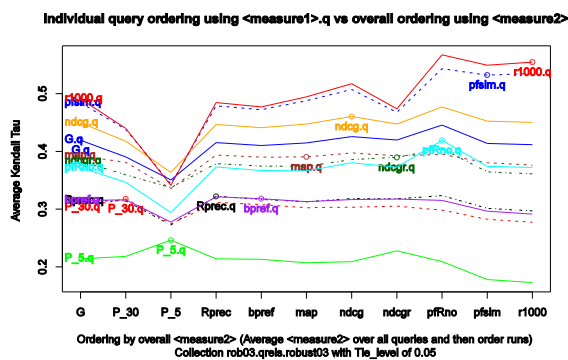
Topics Predicting Overall Rankings (Robust04)



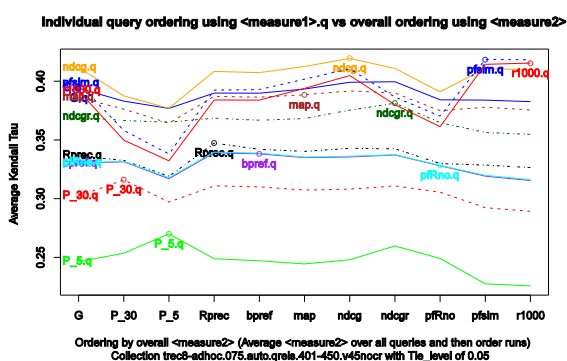
Implications

- Narrow ranges indicates measures are basically the same here, with the exception of P_5
 - Measures do not agree with their own overall average much more than they agree with the other overall measures
- Measures have large differences in predictive power of individual topics
- Measures are ordered by the amount of information used in them
 - Suggests differences show measurement error

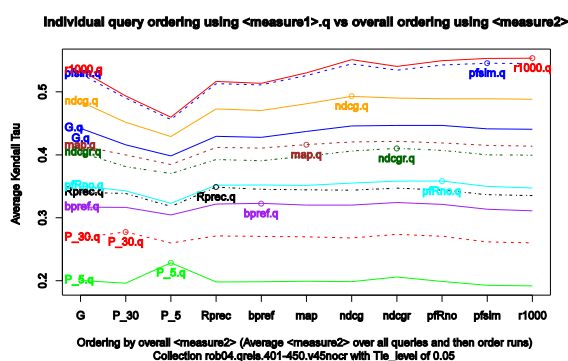
Topics Predicting Overall Rankings (Robust03)



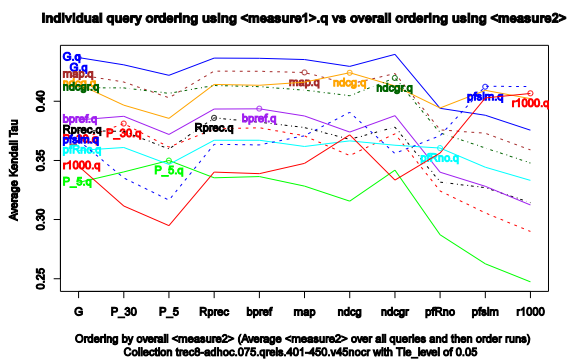
Topics Predicting Overall Rankings (TREC8 adhoc auto)



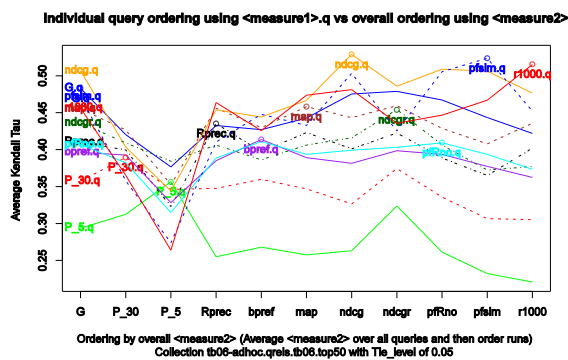
Topics Predicting Overall Rankings Robust04 runs using TREC8 qrels



Topics Predicting Overall Rankings (TREC8 auto+manual)



Topics Predicting Overall Rankings (tb06 auto+man)



Lessons Learned So Far...

- Individual measures don't do a terrible job of ranking systems
 - Future work: can we categorize topics that rank systems well?
- Quality of ranking strongly influenced by the amount of information used and measurement error

Evaluation Failure Analysis

- MAP: heavily used and heavily studied.
- Number of papers examining the failure of MAP to fairly evaluate user's needs
- Turpin, Scholer - Sigir 2006
 - Claim: Users can't distinguish between systems which have MAP ranging between .55 and .95
 - Methodology may have some problems, but...
 - I completely agree with the results
 - I can't distinguish between such systems!

(cont):Relevance Disagreements

- My experience: for most system/topics with high MAP, top non-relevant docs are all marginally non-relevant at worst
 - RIA failure analysis (one topic) agrees with this.
- Users do NOT reliably agree on relevance
 - ~40% overlap in different users relevant docs
 - Harman, TREC 4
 - Cormack TREC 6,7
 - Buckley TREC 2008

(cont) Relevance Disagreements

- All standard measures have strong measurement error due to relevance disagreements
 - Is MAP more affected than others? Unknown.
- How much is this measurement error reflected in earlier plots?
- How do we use reduce this measurement error?

Multi-level Relevance Judgments?

- Binary judgments an artifact of IR history
 - Fine for small collections
- Multi-level judgements increases information available to measures
 - That reduces measurement error
- But
 - Introduces parameters of value of multiple levels
 - Introduces inconsistencies between topics
 - Doesn't reduce relevance disagreements

Preference Relationships

- Establish preferences among docs for user.
 - Much more direct reflection of user's need (in many cases) than absolute threshold of binary or multi-level relevance judgment.
 - No parameters.
- But
 - Impossible to get full coverage of a topic from a single user while maintaining consistency.
 - Doesn't solve relevance disagreement problem

Multi-user Preferences!

- Establish preference relationships on possibly overlapping small subsets for a topic, one subset per user.
- Represents disagreements between users
 - Adds information to reduce measurement error.
 - Computationally feasible to cover needed judgments (no consistency requirement)
- But
 - Need new evaluation measures

TREC_EVAL 9.0

- http://trec.nist.gov/trec_eval
 - Been floating around for over a year
 - Complete rewrite
- Implements several preference measures
- Implements several multiple user approaches
 - All measures can be averaged over multiple users
 - Some measures can be micro-averaged
- Need practical experience
 - TREC relevance feedback track next year?

Single Topic Evaluation

- Field has neglected, since we want multiple topics to completely compare systems
- Needed for several purposes including failure analysis, error bounds, and understanding
- Current measurement error is high
- Need to use more information in our measures, and more accurate information
 - Must include different user opinions
- Multiple user preference relations a solution

Questions?