Philosophy of IR Evaluation

Ellen Voorhees

NIST

Evaluation: How well does system meet information need?

- System evaluation: how good are document rankings?
- User-based evaluation: how satisfied is user?

NIST

Why do system evaluation?

- Allows sufficient control of variables to increase power of comparative experiments
  - laboratory tests less expensive
  - laboratory tests more diagnostic
  - laboratory tests necessarily an abstraction
- It works!
  - numerous examples of techniques developed in the laboratory that improve performance in operational settings

NIST

Cranfield Tradition

- Laboratory testing of retrieval systems first done in Cranfield II experiment (1963)
  - fixed document and query sets
  - evaluation based on relevance judgments
  - relevance abstracted to topical similarity
- Test collections
  - set of documents
  - set of questions
  - relevance judgments

NIST

Cranfield Tradition Assumptions

- Relevance can be approximated by topical similarity
  - relevance of one doc is independent of others
  - all relevant documents equally desirable
  - user information need doesn't change
- Single set of judgments is representative of user population
- Complete judgments (i.e., recall is knowable)
- [Binary judgments]

NIST

The Case Against the Cranfield Tradition

- Relevance judgments
  - vary too much to be the basis of evaluation
  - topical similarity is not utility
  - static set of judgments cannot reflect user's changing information need
- Recall is unknowable
- Results on test collections are not representative of operational retrieval systems

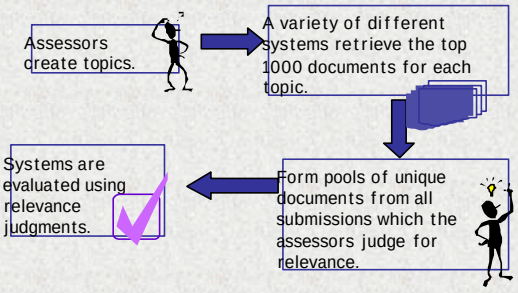
NIST

Response to Criticism

- Goal in Cranfield tradition is to compare systems
  - gives *relative* scores of evaluation measures, not absolute
  - differences in relevance judgments matter only if relative measures based on those judgments change
- Realism is a concern
  - historically concern has been collection size
  - for TREC and similar collections, bigger concern is realism of topic statement

NIST

Using Pooling to Create Large Test Collections



NIST

Documents

- Must be representative of real task of interest
  - genre
  - diversity (subjects, style, vocabulary)
  - amount
  - full text vs. abstract

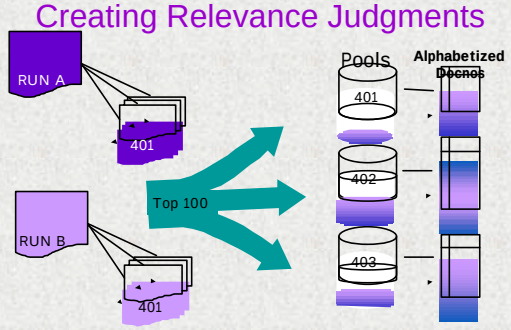
NIST

Topics

- Distinguish between statement of user need (topic) & system data structure (query)
  - topic gives criteria for relevance
  - allows for different query construction techniques

NIST

Creating Relevance Judgments



NIST

Test Collection Reliability

- Recap
  - test collections are abstractions of operational retrieval settings used to explore the relative merits of different retrieval strategies
  - test collections are reliable if they predict the relative worth of different approaches
- Two dimensions to explore
  - inconsistency: differences in relevance judgments caused by using different assessors
  - incompleteness: violation of assumption that all documents are judged for all test queries

NIST

Inconsistency

- Most frequently cited "problem" of test collections
  - undeniably true that relevance is highly subjective; judgments vary by assessor and for same assessor over time ...
  - … but no evidence that these differences affect comparative evaluation of systems

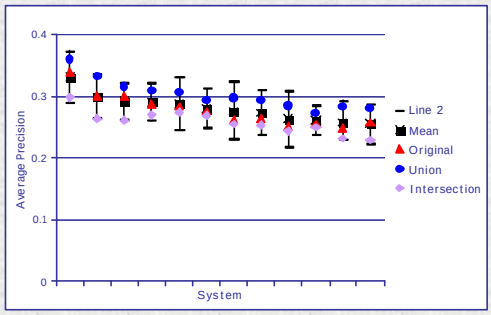
NIST

Experiment:

- Given three independent sets of judgments for each of 48 TREC-4 topics
- Rank the TREC-4 runs by mean average precision as evaluated using different combinations of judgments
- Compute correlation among run rankings

NIST

Average Precision by Qrel



NIST

Effect of Different Judgments

- Similar highly-correlated results found using
  - different query sets
  - different evaluation measures
  - different groups of assessors
  - single opinion vs. group opinion judgments
- Conclusion: comparative results are stable despite the idiosyncratic nature of relevance judgments

NIST

Incompleteness

- Relatively new concern regarding test collection quality
  - early test collections were small enough to have complete judgments
  - current collections can have only a small portion examined for relevance for each query; portion judged is usually selected by pooling

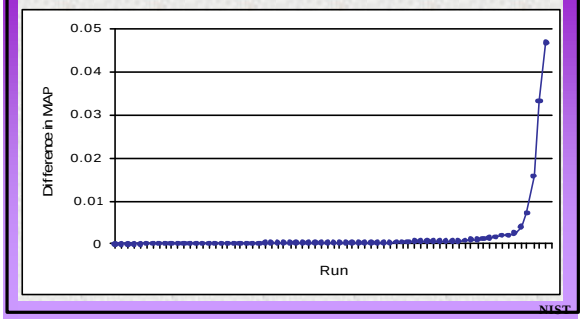
NIST

Incompleteness

- Study by Zobel [SIGIR-98]:
  - Quality of relevance judgments does depend on pool depth and diversity
  - TREC judgments not complete
    - additional relevant documents distributed roughly uniformly across systems but highly skewed across topics
  - TREC ad hoc collections not biased against systems that do not contribute to the pools

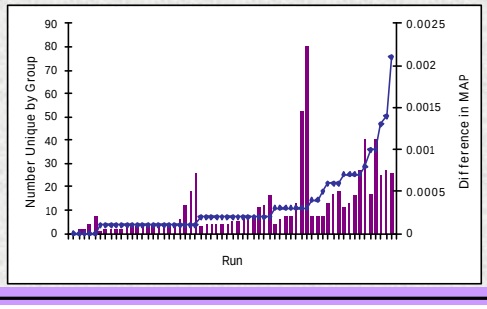
NIST

Uniques Effect on Evaluation



NIST

Uniques Effect on Evaluation: Automatic Only



NIST

Incompleteness

- Adequate pool depth (and diversity) is important to building reliable test collections
- With such controls, large test collections are viable laboratory tools
- 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

NIST

Cross-language Collections

- More difficult to build a cross-language collection than a monolingual collection
  - consistency harder to obtain
    - multiple assessors per topic (one per language)
    - must take care when comparing different language evaluations (e.g., cross run to mono baseline)
  - pooling harder to coordinate
    - need to have large, diverse pools for all languages
    - retrieval results are not balanced across languages
    - haven't tended to get recall-oriented manual runs in cross-language tasks

NIST

Cranfield Tradition

- Test collections are abstractions, but laboratory tests are useful nonetheless
  - evaluation technology is predictive (i.e., results transfer to operational settings)
  - relevance judgments by different assessors almost always produce the same comparative results
  - adequate pools allow unbiased evaluation of unjudged runs

NIST

Cranfield Tradition

- Note the emphasis on ***comparative*** !!
  - absolute score of some effectiveness measure not meaningful
    - absolute score changes when assessor changes
    - query variability not accounted for
    - impact of collection size, generality not accounted for
    - theoretical maximum of 1.0 for both recall & precision not obtainable by humans
  - evaluation results are only comparable when they are from the same collection
    - a subset of a collection is a different collection
    - direct comparison of scores from two different TREC collections (e.g., scores from TRECs 7&8) is invalid

NIST