The last half-century:

a perspective on experimentation in information retrieval

Stephen Robertson

Microsoft Research and City University

On the tradition

- Cranfield
  - Two Cranfield experiments
  - Relevance, recall, precision
- Medlars
  - Evaluation of the Medlars Demand Search Service
- Portable test collections
  - The ‘ideal’ test collection

On the tradition

- The Book: *Information Retrieval Experiment*
- Trying to get at user interaction
- TREC
  - The ideal test collection writ large
- Search engine training
  - Multiple parameters
  - Machine learning

On the tradition

- Some TREC contributions
  - Competition and the ‘unseen’ principle
  - The Pool method
  - Measures and experimental methods
  - The collections
- The TREC influence
  - Spin-offs (CLEF, NTCIR, INEX &c)
  - Search engine training

On having a tradition

- Traditions are good and bad
  - Guidelines or straightjacket
- Pros of the Cranfield-TREC tradition
  - Understanding of evaluation
  - Methods
  - Data
  - High standards

On having a tradition

Cons of the Cranfield-TREC tradition

- Much to learn
  - hard for a PhD student 😐
- Hard to do experiments well
- Bias towards laboratory experiments
  - … and against user experiments
- Hard to get papers accepted
- Only one kind of experiment

On its current status

Dominant!

1. Experiments at TREC / CLEF / DUC &c
2. Experiments using data / methods from the above
3. Theory and models
4. Publication (of all of the above)
5. Operational systems

On the things that stay the same

Documents

(real ones)

Requests

representing information needs

Relevance judgements

on individual documents

Measures

relevance-based – recall, precision &c.

On the things that change

Documents

Requests

Relevance judgements

Measures

… and then some

Documents

Books on shelves
Bibliographic units
Webpages
Virtual webpages
Locations
Passages
Answer fragments
Logical units
Inferred answers

Requests

Note: cannot divorce from relevance

Ideal: catch a user with a need

(archetype: the 1968 Medlars experiment)

Source documents

Constructed from lists

Actors

Logged queries

Relevance judgements

1. Who?
   - User / need
   - Substitute judge with description of need
   - Authoritative source of information
   - Judge interpreting different needs

Relevance judgements

2. What scale?
   - Grades
   - Binary
   - Grades
   - Grades representing user groups
     - Types of need
     - Statistical prevalence

Relevance judgements

3. Which docs?
   - All
   - External source(s)
   - Pools
   - Small pools
   - Sampled

The problem of collection building

Measures

- Sets
  - Recall and precision
  - Fbeta
  - Utility functions
- Ranked output
  - Recall-Precision curve
  - Recall-Fallout (ROC) curve

Measures

- A point on the curve
  - P@n
  - RPrec
  - MRR
- Measures on the whole curve
  - Area under ROC = pairwise error probability
  - Normalized recall / precision
  - Average precision
  - Non-interpolated AP

Measures

- Some criteria for measures
  - Set- or rank-based
  - Top-heavy
- User-oriented
  - Realistic
  - Transparent
- Experiment-oriented
  - Reliable
  - Valid

Measures

- Some criteria for measures
  - Decision-orented
    - Reliable system ranking
    - Valid system ranking
  - Optimization-oriented
    - Smooth, differentiable?
    - Convex?
    - Reliable optimum identification
    - Valid optimum identification

… and then some

- Experimental design
  - Overfitting and all that
  - ‘Does it work’
  - ‘Does x work better than y’
  - Optimization
  - Train-test-holdout; cross-validation
  - Size of model

Final remarks

The tradition that began with Cranfield is alive and kicking, half a century later.

It carries both opportunities and constraints.

It presents two main problems:

- How and when to push its boundaries
- How and when to transcend it

Neither of these should involve throwing it away!