Test collections for all

Mark Sanderson
Content

• How to build test collections easily
  • Adapting to context

• What challenges remain with test collections
  • Adapting to a user
Adapting to context

• Need new test collections
• IR matching algorithms vary with context
• Many contexts few test collections
  • <100?
• Why?
  • Perception that test collections are hard to produce
    Implies people will use centrally built ones
    Don’t try to build their own
Thing is...

- test collections aren’t a lot of work to build

- Sheffield
  - Built many
  - 1 week effort per person max.
Series of formation methods

- ISJ
  - Cormack, Palmer, Clarke SIGIR, 1998
  - Other studies in later SIGIRs
- Relevance feedback
  - Soboroff SIGIR, 2001
  - Also studied SIGIR 2004
Cormack, Palmer & Clarke

- Don’t have assessors, have Interactive Searchers and Judgers (ISJ)
  - Try multiple queries,
  - Set aside relevant documents…
  - …for a final set of relevant documents
It does work

- Sheffield used ISJ for geoCLEF 2005
  - 25 topics
    1-2 hours per topic
    Paid a few students to find relevant documents
    Judgements done in a week

- Didn’t have to wait for submissions
By comparison

- CLEF ran pooling in parallel
  - Few contributors to the pool
    First run of the track
  - Sheffield ISJ found many relevant documents
    than (in this case) standard pooling missed
Soboroff

- Assessors not great searchers?
- Get them to do relevance feedback
  - Mark documents in initial system pool as rel
  - Form a new query
  - Mark documents in new system pool as rel.
Suite of methods for you

- Just as good a TREC/CLEF-style pooled collections
- Methods build future proofed collections
  - Should effectively assess IR systems to come
There is the question

• Does your collection need this feature?

• If not
  • You can cut corners
Simple approach

• TREC/CLEF: judge from pool
  • from top 100 (sometimes 50)
• Use pool from top 10
  • A much shallower pool
  • 11%-14% relevance assessor effort
  Compared to top 100
• Small increase in error.
More recent work

- SIGIR 2006
  - Targeted assessment
    Carterette, Allan, Sitaraman
  - Pool sampling
    Aslam, Pavlu, Yilmaz
    4% of pool
Noise tolerant measures

- SIGIR 2003
  - Voorhees, Buckley
  - Bpref-10
- CIKM 2006
  - Yilmaz, Aslam
  - Apparently more stable measure

- Can deal better with patchy qrels
Adapting to users

• Almost all of our collections assume that users have the same definition of relevance
  • Not true!
• Different users view different documents as relevant to the same query
  • Duh!
  • Existing test collections don’t deal with this
Often we have no context

• We could try to find context
• When talking about adaptive IR
  • Do we talk about personalisation with no information?
  • Build collections with many user judgements for the same query
Less is more

• SIGIR 2006
  • Find most relevant doc
    Rank position 1
  • Find next most relevant and different doc
    Rank position 2
• Tested on TREC multi-assessment collection
  • Satisfied more users more of the time
So

• The tools are there
  • What are we waiting for?

• Another reason for these methods being around…
Test collections need to improve

- Generic problems
  - Important to Adaptive IR

- Remember
  - Test collection based evaluation encourages us to…
...build IR systems that...

- Match on sub-sets of query words
  - Often unwanted by users
- Use measures
  - Don’t focus enough on poor queries
- Don’t consider interaction
  - Within topic
  - Across topics
Not enough queries

• Academia
  • 25, 50, 100 topics per year?
• Search engines
  • Hundreds of topics per week
Think about

• The methods described here
• Build new conventional collections
  • Specific to your context

• Think how the methods can help build
  • Collections we’ll need soon
Finally

- NTCIR Evaluation workshop
Still hard work

- Have to create topics
  - Non-trivial
    TREC/CLEF skilled at topic creation

- What’s better (speculation)
  - Your (OK) topics right for your context
  - Their (good) topics right for a different context