Test Collections

Lecture 10

IR Evaluation

- We can use precision and recall to measure the performance of IR systems
- These systems might be
  - operational systems, running in the field
  - experimental systems in the laboratory
  - prototype retrieval algorithms
- Creating good test queries and useful document collections is hard
- So we often build standard *test collections*

What is a test collection?

1. A collection of documents
2. A set of information needs or queries
3. Relevance judgments

- Examples: CRAN, CACM, TREC

Document Collection

- Language
- Genre
- Origin
- Time period, era
- Quality, style
- Authorship

- Structure
- Media
- Labels, tags
- Categories
- Formats, encoding
- Availability

Information needs, search topics

- Information needs are diverse
  - Users are interested in different things
  - Usually not what you expect...
  - Search performance varies across topics and queries
- Test topics should reflect this!
- Variety is crucial for reliable experiments

A Sample TREC topic

<top>
<num> Number: 351
<title> Falkland petroleum exploration

<desc> Description:
What information is available on petroleum exploration in the South Atlantic near the Falkland Islands?

<narr> Narrative:
Any document discussing petroleum exploration in the South Atlantic near the Falkland Islands is considered relevant. Documents discussing petroleum exploration in continental South America are not relevant.
</top>

Relevance Judgments

- Relevance is complicated
  - Users "know it when they see it"
  - Users disagree about what's relevant
- Assessors need well-defined guidelines
  - "A document is relevant if it contains *any* information you would use in compiling a report on the topic." -- TREC relevance
- Should reflect experimental task

Why standard test collections?

- A test collection is an experimental tool
- It allows other experimenters to
  - understand your results
  - compare their results to yours
  - reproduce your results
- Often built for a specific purpose
  - retrieval, filtering, classification, clustering

Cranfield II Experiments

- Goal: measure effect of two different index languages on search effectiveness
- The "Cranfield Collection"
  - 1400 aeronautical engineering abstracts
  - 225 one- or two-sentence topics
- Experimental assumptions
  - Relevance = topical similarity
    - Static information need
    - All documents equally desirable
  - Relevance judgments are complete and representative of the user population

Cranfield topics

.I 001
.W
what similarity laws must be obeyed when constructing aeroelastic models of heated high speed aircraft .
.I 008
.W
can a criterion be developed to show empirically the validity of flow solutions for chemically reacting gas mixtures based on the simplifying assumption of instantaneous local chemical equilibrium .
.I 009
.W
what chemical kinetic system is applicable to hypersonic aerodynamic problems .

The TREC Workshops

- Text REtrieval Conferences
- Started in 1992
- Framework for evaluating retrieval tasks using large test collections
- Anyone can participate
  - get the data, run your system, submit the results
- Results and experiences are shared at the workshop every November

TREC Tracks

- TREC began with two tasks
  - *ad hoc retrieval*
  - *routing*
- and added several tracks over the years
- some tracks use different collections
- Not all tracks run in all years

Tracks

- Filtering
- Question answering
- Web
- VLC (100GB)
- Interactive
- Cross-Language
- Chinese
- Video
- Query
- Spoken Document Retrieval

The TREC Collections

- The "classic" TREC collections
  - 5 CDs (~5GB) of text
  - newswire: AP, WSJ, SJMN, FBIS, FT, LAT
  - gov’t documents: patents, CR, FR
  - 450 search topics, with relevance judgments covering different subsets of the collection
- The TREC Web collections
  - 100GB from the Internet Archive (1997)
  - 2GB and 10GB subsets
  - 18GB .GOV collection (2002)
- Different tracks use different collections

The TREC Main Track ("ad hoc")

- At NIST...
  - Assessors create 50 new search topics
  - Guidelines and topics are released to participants
- Participants (universities, labs, companies...)
  - Use their systems to search the collections for relevant documents for each topic
  - Submit their top 1000 for each topic to NIST
- Back at NIST...
  - Assessors make relevance judgments
  - Runs are evaluated using the judgments and results sent back to participants

TREC Relevance Judgments

- Collections are too large for complete judgments
- Pooling
  - Top 100/topic/run placed in a pool (no duplicates)
  - Assessor judges only documents in the pool
- Studies have shown that
  - Yes, some relevant documents are missed
  - But it doesn't change the rankings of systems
  - Judgments usable by non-participating systems
  - Disagreements by assessors don't affect system rankings

The TREC collections

- The topics are written and released without judgments
- Judgments are set after all results are in
- Therefore, each year a new collection is produced
  - document set (e.g., CDs 4 and 5)
  - that year’s topics (e.g. 351-400)
  - relevance judgments for those topics on those documents

Lessons from TREC

- Larger collections
  - Can provide much better research results
  - Complete judgments are impossible
  - We can use pooling to overcome this
- Methodology usable for lots of tasks
  - retrieval was just the start
  - filtering, web search, speech, video retrieval
  - CLEF and NTCIR evaluations