Test Collections and IR Experimentation

CISC489/689-010, Lecture #11

Wednesday, March 18th

Ben Carterette

Last Time

- Search engine evaluation, esp. system-based evaluation
- Measures:
  - Precision
  - Recall
  - Average precision
  - R-precision
  - DCG
- All measures require *relevance judgments*
- Average over a set of queries

IR Experimentation

- Comparing different engines requires a controlled experimental setting
- Many different factors influence engine effectiveness:
  - Corpus, queries, relevance judgments
  - Parsing decisions, stop list, stemming algorithm
  - Indexing method, compression algorithm, query processing method
  - Retrieval model, model features, model parameters
- These should be controlled to the greatest extent possible to answer the relevant questions

Evaluation Corpus

- A *test collection* consists of:
  - a corpus of documents or other things to search
  - a set of queries with underlying information needs
  - relevance judgments on documents
- The use of test collections and system-based effectiveness evaluation is called the *Cranfield methodology.*
  - Named after Cranfield Aeronautics, where it was invented in the 60s.

Constructing a Test Collection

- Where do the documents come from?
  - Depends on task, domain, availability, …
- Where do the queries come from?
  - Query logs, users who can describe their information need, …
- What do the queries and information needs look like?
  - Depends on task, domain, …
- Where do the relevance judgments come from?
  - Assessors judge documents w.r.t. information needs

Corpus Examples

- News corpora
  - AP: Associated Press articles from 1988-1992.
  - TDT: News articles from AP, Wall Street Journal, NY Times, LA Times, plus audio transcripts
- Web corpora
  - GOV2: Web pages downloaded from .gov domain in 2004
  - Wikipedia: Top 10% of Wikipedia pages
- Genomics corpora
  - Medical journals, clinical reports