Evaluating search results

- Classic measures

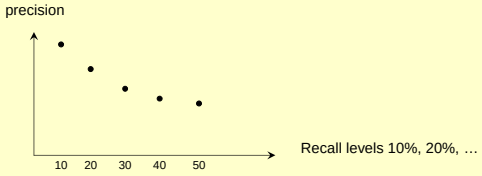
$$\text{Precision} = \frac{(\#\text{ relevant items retrieved})}{(\#\text{ retrieved items})}$$

$$\text{Recall} = \frac{(\#\text{ relevant items retrieved})}{(\#\text{ relevant items})}$$

What is "relevance" – human judge! Yes/no decision

Ranked items

- More can do:
  - Look at precision and recall at any point in ranking
  - Plot precision v.s. recall



Single numbers

- Choose specific position in ranking at which to measure precision
  - E.g. among top 10
- Average the precision after each new relevant item as go down ranking

Relevance by TREC method

Text Retrieval Conference 1992 to present

- Fixed collection per "track"
  - E.g. "*.gov", CACM articles
- Each competing search engine for a track asked to retrieve documents on several "topics"
  - Search engine turns topic into query
  - Topic description has clear statement of what is to be considered *relevant* by *human judge*

Pooling

- Human judges can't look at all docs in collection: thousands to millions
- Pooling chooses subset of docs of collection for human judges to rate relevance of
- Assume docs not in pool not relevant

How construct pool for a topic?
Let competing search engines decide:

- Choose a parameter k (typically 100)
- Choose the top k docs as ranked by each search engine
- Pool = union of these sets of docs
  Between k and (# search engines) * k docs in pool
- Give pool to judges for relevance scoring

Pooling cont.

- $(k+1)^{st}$ doc returned by one search engine either irrelevant or ranked higher by another search engine in competition
- In competition, each search engine is judged on results for top r > k docs returned

Web search evaluation

- Are different kinds of queries – identified inTREC Web Track – some are:
  - Ad hoc
  - Topic distillation: set of key resources small, 100% recall?
  - Home page: # relevant pages = 1 (except mirrors)
- Distinguish for competitors or just judges?

More web/online issues

- Are browser-dependent and presentation dependent issues:
  - On first page of results?
  - See result without scrolling?

Other issues in evaluation

- Does retrieving highly relevant documents really satisfy users?
  - Subjectivity?
- Are there dependences not accounted for?
- Many searches are interactive