Lecture 5: IR Evaluation

William Webber (`william@williamwebber.com`)

COMP90042, 2014, Semester 1, Lecture 5

What we'll learn today

- How (in principle) to build a reusable test collection for evaluating IR systems
- How to evaluate and compare IR systems against such a test collection, using effectiveness metrics

Meeting human information needs

[SELECT * FROM customers WHERE city='Sydney' AND age > 45]
[(jaguar OR irvine OR webber) AND (race OR competition OR "grand prix") w/10 (statistics OR results OR scores)]
[jaguar race statistics]

- Free text queries are not formal representations of information sought (unlike SQL or Boolean queries)
- Rather, they are informal, suggestive approximations of what user wants (which user themselves may not exactly know)
- Does not place onus on user to
- "Correct" answers are not formally definable
- "Models" only guides, don't determine theoretical correctness

All models are wrong, but some are useful – George E. P. Box, 1987

- System must do "best it can" to:
  - Infer user's intent
  - Predict result responsiveness to this intent

Correct answers not formally definable

[(jaguar OR irvine OR webber) AND (race OR competition OR "grand prix") w/10 (statistics OR results OR scores)]
[jaguar race statistics]

- How well the system's results (for a query, for all queries) meet a user's need is referred to as the system's *effectiveness*
- And the process of determining this effectiveness (for a given query, a given set of queries, or in general) is known as *effectiveness evaluation*
- Cannot use evaluation regimes such as "a correct system is one in which all documents returned contain all query keywords"
- Ultimately, effectiveness defined by user's satisfaction with or utility from results.

Direct human evaluation

- Obvious evaluation method: direct evaluation with human users, effectiveness measure from:
  - reported satisfaction
  - completion of tasks
- But method too expensive, slow for comparing, tuning many different formulae or parameters:
  - TF $= f_{d,t}$ *OR* $\log(f_{d,t} + 1)$ *OR* ...
  - Pivoted DLN slope $s = 1.0$ *OR* 0.9 *OR* ...
  - PRF with 1 or 3 or 5 or ... top documents
  - Rocchio parameter $\alpha = 0.4$ *OR* 0.5 *OR* 0.6 *OR* ...
  - Across 200 different queries
- Complexities of experimental setup (user to evaluate 20 results for one query, without learning or fatiguing)

Automated testing

- We want evaluation setup that can be run automatically
- While still being based upon human perceptions of effectiveness
- To achieve this, we will have to make some simplifications!
- Begin with "maximal" set of simplifications applied
- ... to create (traditional, TREC, Cranfield) *test collection model*

Framework

- User has information need
- Express this need as a query
- System runs query against corpus
- Returns ranked list of documents
- Effectiveness is how well this ranked list satisfies information need

Simplifying assumption 1: Ad-hoc

Retrieval is *Ad-Hoc*

- Query is made once
  - No opportunity for refinement, feedback
- We have no prior knowledge of the user (their interests, preferences)
- We have no prior knowledge of behaviour of other users for this query

Simplifying assumption 2: Relevance

Effectiveness based upon *relevance*

- Each document is either relevant or irrelevant to information need
  - Note: more exact to speak of "relevance to information need" than "relevance to query"
- Relevance is binary (document is either wholly relevant or wholly irrelevant)
- Relevance of one document in result independent of relevance of other documents in result (no redundancy, diversity)
- Effectiveness of result is function of relevance of documents in result

Test collection

With these assumptions, automated effectiveness evaluation performable with a reusable *test collection*, consisting of three (main) components:

Corpus
: set of documents

Queries
: set of queries to run against corpus
  - Sometimes supplemented by fuller descriptions of underlying information need
  - In which case we speak of "topics"

Qrels
: for each document and query, a (human) judgment of whether that document is relevant to (the information need underlying) that query

Converting document ranking into relevance vector

Retrieval run

Docid	Score
CR93H-9548	0.5436
CR93H-12789	0.4958
CR93H-10580	0.4633
CR93H-14389	0.4616
AP880828-0030	0.4523
CR93H-10986	0.4383
...	

Qrels

Docid	Rel
AP880828-0030	0
AP881226-0140	1
AP881227-0083	0
CR93H-14389	0
CR93H-9548	1
CR93H-10580	0
CR93H-10986	1
CR93H-12789	0
...	
...	

Relevance vector

$$\langle 1, 0, 0, 0, 0, 1, \ldots \rangle \tag{1}$$

- Take retrieval run as a ranking of document ids (already a very abstracted representation!)
- Look up relevance of document ids in qrels dictionary
- Convert run into relevance vector

Quantifying effectiveness

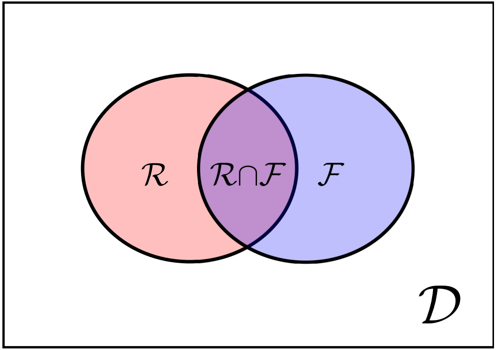
Effectiveness of result is function of relevance of documents in result

- Need function to express effectiveness of relevance vector as a single number

$$m(\langle < 1, 0, 1, 0, 0, 1, 1, \ldots \rangle >) \rightarrow 0.8 \tag{2}$$

- This function an *effectiveness metric*
- And the number it reports an *effectiveness score*

Recall and precision



$\mathcal{R}$ relevant documents

$\mathcal{F}$ retrieved documents

Two fundamental (set-based) measures:

Recall Proportion of relevant documents retrieved, $\frac{|\mathcal{R} \cap \mathcal{F}|}{|\mathcal{R}|}$

Precision Proportion of retrieved documents relevant, $\frac{|\mathcal{R} \cap \mathcal{F}|}{|\mathcal{F}|}$

Precision @ k

Simple measure, *prec@k*:

- Truncate ranking to depth k
- Calculate precision of prefix

$$\text{p@k}(\vec{f}) = \frac{1}{k} \sum_{i=1}^{k} f_i \quad (3)$$

$$\begin{aligned} \text{p@5}(\langle 1,1,0,1,0,1,0,0,1,0\ldots\rangle) &= \text{p@5}(\langle 1,1,0,1,0\rangle) \\ &= \tfrac{1}{5}\cdot 3 \\ &= 0.6 \end{aligned} \quad (4)$$

$\text{rec@k}(\vec{f}) = c_q \cdot \text{p@k}(\vec{f})$, where c_q is a query-dependent constant:

- Why?
- What is c_q?

Precision @ k

Two objections to Precision @ k:

Not rank-sensitive

- Doesn't reward better rankings up to k:

$$\text{p@5}(\langle 1,0,0,0,0\rangle) = \text{p@5}(\langle 0,0,0,0,1\rangle) \tag{5}$$

- More exactly: rank-sensitivity is very coarse; ranks up to k get same weight of $1/k$; ranks beyond k get weight of 0

Not recall-sensitive

- Ignores number of relevant documents for query, $R_q = |\mathcal{R}_q|$
- Maximum p@100 when $R_q = 1$ is 0.01.
- p@5 $= 1.0$ easier where $R_q = 1000$ than $R_q = 5$

This important where aggregating scores over multiple queries

Mean average precision (MAP)

Mean average precision:

> *The average precision at each point in the ranking a relevant document occurs:*

In practice

- ranking generally truncated at some depth k (e.g. $k = 1000$)
- relevant documents not in ranking given precision 0

$$\mathrm{AP}(\vec{f}; k, q) = \frac{1}{R_q} \sum_{i=1}^{k} f_i \cdot \mathrm{p@}i(\vec{f}) \tag{6}$$

A model for MAP

Simple model of user behaviour and resulting utility:

- User views $\vec{f}$ from top, stops when r_u seen relevant docs
- r_u is a random variable:

$$\Pr(r_u = i) = \begin{cases} 1/R_q, & \text{if } 0 < i \leq R_q. \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

  - (Unrealistically assumes user "knows" R_q)
- Let d_{r_u} be the rank of the r_u'th relevant document. Then mean average precision definable as:[1]

$$\mathrm{AP}(\vec{f}) = \mathbb{E}[\mathrm{p@}d_{r_u}(\vec{f})] \tag{8}$$

[1] Corrected, 2014-03-20. Previous version incorrectly had $\mathbb{E}[\mathrm{p@}r_u(\vec{f})]$

TREC

- Since early 1990s, academic IR evaluation focused around collaborative evaluation "competitions", that:
  - share effort of creating collection (particularly, evaluating documents for relevance to queries)
  - provide common benchmark for performance
- First and most famous of these is TREC (Text REtrieval Conference), run annually, based at NIST in US.
- Typical ad-hoc TREC collection contains:
  - 50 topics (queries, with more extended relevance statements), authored by experience independent searchers
  - Qrels for top 100 results returned by each participant to each query (pooling) (remaining documents assumed irrelevant), judged by topic authors
  - (Externally) results submitted by participants

Example TREC datasets

TREC 5, 1996

Topic

⟨num⟩ Number: 252
⟨title⟩ Topic: Combating Alien Smuggling
⟨desc⟩ Description: What steps are being taken by governmental or even private entities world-wide to stop the smuggling of aliens.
⟨narr⟩ Narrative: To be relevant, a document must describe an effort being made (other than routine border patrols) in any country of the world to prevent the illegal penetration of aliens across borders.

Qrels

Topic	Docid	Rel
252	AP881226-0140	1
252	AP881227-0083	0
252	CR93E-10038	0
252	CR93E-1004	0
252	CR93E-10211	0
252	CR93E-10529	1
...		

Runfile

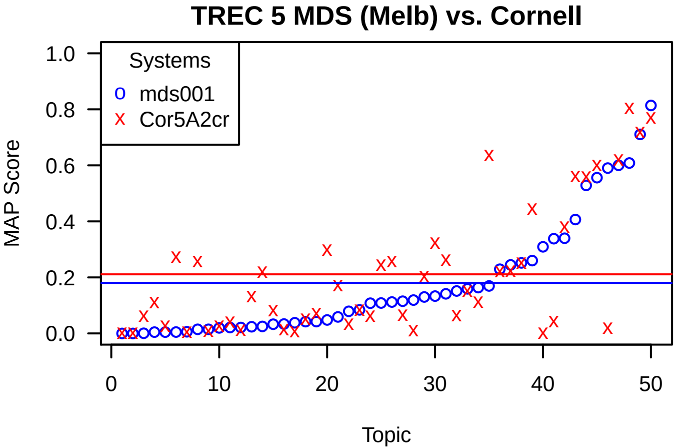
Topic	Docid	Score
252	CR93H-9548	0.5436
252	CR93H-12789	0.4958
252	CR93H-10580	0.4633
252	CR93H-14389	0.4616
252	AP880828-0030	0.4523
252	CR93H-10986	0.4383
...		

Comparing systems on a test collection

Compare two systems on test collection:

- Run each system against each topic
- Calculate per-topic effectiveness score under selected metric (e.g. AP)
- Calculate systems score on collection as mean of topic scores
- Compare systems by mean score
- Test mean score differences for statistical significance

MDS (Melbourne) vs. Cornell



- Mean AP scores: MDS 0.180, Cornell 0.211
- Not statistically significant ($p > 0.05$ in 2-tailed, paired t test)

Extending test collection model: multi-grade relevance

- Go beyond binary relevance to allow multiple relevance levels
- E.g. "irrelevant", "marginal", "relevant", "highly", "essential"
- Requires metric support (e.g. nDCG, RBP)

Pros

- Allows finer-grade relevance assessment
- Widely used by search engines, because:
  - Many "relevant" results
  - Short result list (10 results)
  - Emphasis on getting top results

Cons

- May place more load on assessor
- Unclear if gives better (deep-rank) assessment than binary

Extending test collection model: diversity

- Similar documents make each other redundant in results list
- Query may have many intents or aspects

  Intents different topics underlying same query

  Aspects different parts of information about the one topic
- Want to avoid redundancy, reward diversity in results list

Pros

- Very important aspect of practical retrieval satisfaction, utility

Cons

- Places a much heavier load on assessor / organizers

(IR'ers recognized this issue for decades, but only in past decade did they "bite the bullet")

Extending test collection model: multi-session

- In practice, a user can refine their query, search interactively
- System should respond to a query differently if it is a refinement
- Recent attempts to do this in a test collection
- ... but very difficult!
- May have to be approached through interaction studies (see next)

Automatic user feedback methods

Automatic user feedback methods available on working, heavily-used system (e.g. web search engine):

- Click-through statistics (if a user clicks on a result, treat that result as “correct”)
- Try different result lists on users, and observe click and other behaviour:
  - A/B testing (show different result lists to different users)
  - Result interleaving (interleave results from two algorithms in the one list)

Looking back and forward

Back

- Retrieval effectiveness must be measured against human perception
- Human-in-loop too expensive for regular experiments
- Test collection "cans" human as qrels
- Metric calculates score from relevance vector
- Compare two systems by scores on set of topics from one collection

Looking back and forward

Forward

- With almost all text analytic techniques, human judgment is ultimately required, and "how do we evaluate this?" becomes a crucial question
- Next lecture looks at the (difficult-to-evaluate!) text analytical technique of (document) clustering
- Text classification is "evaluation-based tuning on steroids": take human relevance assessments and use them to automatically develop your model

Further reading

- Overview of one of the TREC conferences, for instance TREC 5[2]
- Chapter 3, "Technical Background", of William Webber, *Measurement in Information Retrieval Evaluation*[3] (PhD Thesis; Melbourne, 2010)

[2] `http://trec.nist.gov/pubs/trec5/papers/overview.ps.gz` (note: gzipped postscript)

[3] `http://www.williamwebber.com/research/wew-thesis-PhD.pdf`