# LBSC 708x: E-Discovery
# Evaluating E-Discovery

William Webber
CIS, University of Maryland

Spring semester, 2012

# Outline

Why and what of evaluation

Measuring retrieval quality

Evaluating retrieval processes

Which is better: automated or manual retrieval?

# Outline

# Why evaluate?

Blair and Maron (1985).

- Gave lawyers a Boolean retrieval system, to work on production for real case.
- Asked them to iteratively reformulate queries until they were confident that they had retrieved 75% of relevant material.
- When lawyers finished, asked lawyers to check sample of unretrieved documents.
- Found they had only retrieved 20% of relevant material.

Moral: don't trust intuition to tell you how good your retrieval process is.

# Measuring quality

Three (at least) ways to go about “measuring quality” of an e-discovery process:

1. Measure how well a process has performed in a given retrieval.
2. Evaluate how well a process performs in general, particularly in comparison with other processes.
3. Assess whether a process is managed in a way that is likely engender quality.

# Certifying best practice

ISO 9001-style quality management systems:

- Certify not that your process is best practice
- . . . but that you have a meta-process that allows you to drive your process towards best practice:
  - measurement of quality
  - organizational learning from experience
  - continuous improvement cycles

We'll not talk about this further here.

# Evaluating quality of process

Goal: measurement of quality of process, independent of particular task or production.

- Comparative: often, absolute quality matters less than “which of these options is best?”
- Predictive: helps answer the question, “Which process should I use for this case?”

Difficulty: how generalize from specific productions to productions in general. How can we “randomly sample” productions?

# Measurement of retrieval effectiveness

Measure the effectiveness (comprehensiveness, accuracy) of a particular production.

- how effectively has process retrieved documents according to the operative conception of relevance?
- how consistent and correct is the operative conception of relevance that was arrived at?

Might be performed as part of certifying to the opposing side and to the court that our production in response to the opposing request is adequate.

# Outline

# Considerations in measuring retrieval effectiveness

- Effort: how to allocate effort between validation of quality of result and improvement of that quality.
- Effectiveness: how do we measure effectiveness? What consitutes an adequate retrieval, and how do we establish it?
- Agreement: how to communicate quality of production, and conception of relevance, to other side, to achieve their agreement.
  - and how to do so without given away evidence or strategy

# Binary independence of relevance

## Definition (Goal of Information Retrieval)

To produce all and only documents responsive (relevant) to a production request.

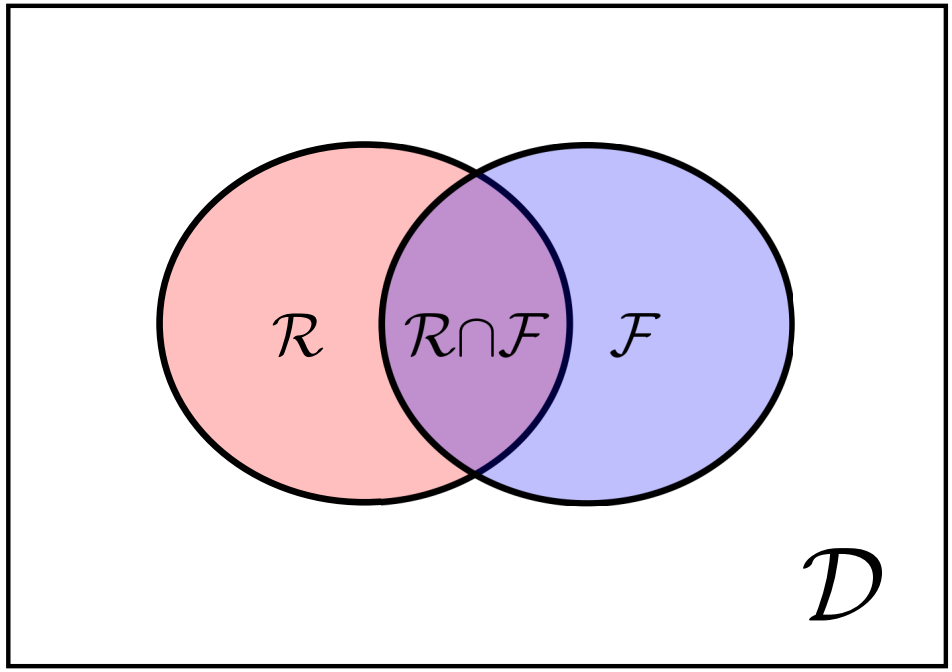
- Suggests a binary ground truth of independently relevant and irrelevant documents.
- Evaluation is about counting proportions of relevant documents.

# Criticisms of the binary model

Problems with the binary independence model:

- Dependence: if document A contains all the information of document B, and we have already produced document A, is document B still (equally) relevant as if we hadn't produced document A?
- Degrees: some documents are more important or more "relevant" than others (so-called "hot documents"). Documents that arrive in later EDRM stages more important than those that don't:
  - Documents that help determine strategy.
  - Documents that are introduced into evidence.

# Recall, precision, F1




**Precision** the proportion of retrieved documents that are relevant.

**Recall** the proportion of relevant documents that are retrieved.

**F score** the harmonic mean of precision and recall.

# Evaluation metric: recall

$$\frac{p + (z^2_{\alpha/2}/2n) \pm z_{\alpha/2}\sqrt{[p(1-p) + z^2_{\alpha/2}/4n]/n}}{1 + z^2_{\alpha/2}/n} \quad (1)$$

- But we don't know all the relevant documents!
- So we have to sample to estimate the number of relevant documents in the unretrieved (and often in the retrieved) set
- . . . and use sample to set probabilistic lower bound on recall

# Confidence intervals

- In sampling from the retrieved set of documents, we want to say something like “we’re 95% confident that 55% $\pm$ 2% of retrieved documents are relevant”
  - Practitioners sometimes summarize this, rather confusingly, as “95% $\pm$ 2% confidence”.
- A sample size of 2399 documents gives a $\pm$2% estimate with 95% confidence.
  - . . . provided that the proportion relevant is not too far from 50%.
- Bound on recall a little bit more complex . . . but same idea

# Stopping rule

- Former method provides a static measure of effectiveness
- However, actual production is iterative
  - we could always spend more effort looking for relevant documents
- What we want is a stopping rule:
  - Is it worth the effort to continue looking, or is the production complete enough that we can stop now?

# Stopping rules

- One stopping rule is, “stop when lower bound on recall is above $X\%$”
- Another stopping rule is, “stop when proportion of novel, hot documents in remainder of corpus is below $X\%$”
  - For example, “sample 2399 unretrieved documents and if none are hot and novel, stop”
  - Sets an upper bound of 0.15% of unretrieved documents being hot and novel
- Note that latter rule provides pragmatic solution to problems of dependence, degree of relevance.

# Considerations in measuring retrieval effectiveness

- Effort: how to allocate effort between validation of quality of result and improvement of that quality.
- Effectiveness: how do we measure effectiveness? What consitutes an adequate retrieval, and how do we establish it?
- Conception: how to we assess the accuracy of the conception of relevance that was used in the production?

# Assessing conception of relevance

- Conception of relevance subjective
- But we only need to satisfy other side's conception, not match all possible conceptions
- For classification-based approaches:
  - describe, agree upon method of choosing seed documents (e.g. by search terms)
  - send other side all assessed documents (training and test)
  - . . . give them opportunity to dispute assessments
- For manual review approaches:
  - . . . ?

# Outline

# System evaluation: which is best process?

- Like to have a general evaluation of (relative) process effectiveness
- . . . to determine which process to choose for a production
- . . . to minimize risk that we get to the end of the production and find our results are bad

# Factors in retrieval quality

Various factors involved:

**Effectiveness** Recall, precision, etc.

**Effort** We can always improve our results with more effort

**Measurability** How reliably can we measure how well we did?

**Agreeability** How well can we persuade the other side and (ultimately) the judge that our process is reasonable?

# Common basis for comparison: test collection

- Retrieval effectiveness, difficulty depends heavily on task, and corpus
- Therefore, need common corpus, common tasks to compare systems
- Such mixture of corpus, tasks, and relevance assessments known as a test collection

# What basis of relevance?

- Trickiest part of test collection formation is, assessing documents for relevance
- As with concrete evaluation, collection to big for exhaustive assessment, must sample documents for assessment
- But what basis for determining relevance of individual documents?

# Assessor disagreement

Voorhees (IPM, 2000)

- If two different assessors assess a set of documents
- . . . and we compare the sets of documents that they find relevant
- . . . only 50% documents that either find relevant will be found relevant by both (50% overlap)

Corollary: one manual review has upper bound effectiveness of 0.66 precision/recall when measured against another manual review. May be ok for relative evalution; problematic for

absolute evaluation.

# Again, conception doesn't need to be objective

- As with concrete evaluation, we don't have to meet everyone's conception of relevance; only a particular person's conception
- In a live evaluation task, can have a single figure (e.g. a lawyer) whose conception of relevance each team is trying to match.
- This is the approach taken in the TREC Interactive Task, with their Topic Authority (TA).

# Reusing subjective conception of relevance

- But how to capture subjective conception of relevance for a reusable collection, when TA is no longer available?
- Need some way of objectively specifying this conception, that is comparable to what live teams received.
  - For machine classification: actual relevance assessments of TA
  - For other approaches: . . . ?

# Outline

# Comparing manual and automated review

- A particularly hot question at the moment: is automated retrieval as good as (better than) manual review?
- How to answer this question?

# Compare with manual review?

- One approach: see which system gets closest to manual review
- This assumes manual review is gold (perfect), or at least silver (near-perfect), standard
- Unfortunately not: assessor disagreement and error is very common.

# Experiment comparing manual and automate review

Roitblat, Kershaw, and Oot (JASIST, 2010).

- Take a production done by Verizon using exhaustive manual review (cost: $14 million)
- Ask two automated retrieval vendors to redo production
- Also have two manual review teams re-review a sample of original production

# Comparing automated to manual review: results

| Assessor pair | kappa | f1 | prec.1 | prec.2 |
|---|---|---|---|---|
| Original vs. Manual A | 0.16 | 0.28 | 0.49 | 0.20 |
| Original vs. Manual B | 0.15 | 0.27 | 0.54 | 0.18 |
| Manual A vs. Manual B | 0.24 | 0.44 | 0.48 | 0.40 |
| Original vs. Auto C | 0.25 | 0.34 | 0.46 | 0.27 |
| Original vs. Auto D | 0.29 | 0.38 | 0.53 | 0.29 |

Table: Inter-assessor agreement reported by RKO, JASIST 2010.

- Automated systems agreed better with original production than manual review
- But everyone disagreed with each other pretty badly
- Conclusion: automated review no less awful than manual

# Compare to gold standard

- Really, what manual and automated methods are trying to do is to meet an individual lawyer's conception of relevance
- Need to compare them in this setup
- That can be (kind of) retrojected onto the TREC interactive task setup

# Assessors versus automated systems

| Topic | Team | Rec | Prec | F1 |
|---|---|---|---|---|
| t201 | System A | 0.78 | 0.91 | 0.84 |
| | TREC (Law Students) | 0.76 | 0.05 | 0.09 |
| t202 | System A | 0.67 | 0.88 | 0.76 |
| | TREC (Law Students) | 0.80 | 0.27 | 0.40 |
| t203 | System A | 0.86 | 0.69 | 0.77 |
| | TREC (Professionals) | 0.25 | 0.12 | 0.17 |
| t204 | System I | 0.76 | 0.84 | 0.80 |
| | TREC (Professionals) | 0.37 | 0.26 | 0.30 |
| t207 | System A | 0.76 | 0.91 | 0.83 |
| | TREC (Professionals) | 0.79 | 0.89 | 0.84 |

Table: Automated and manual effectiveness.

Best automated team better than manual review teams for all but one topic. (Grossman and Cormack, JOLT 2011)

# END