مبانی بازیابی اطلاعات و جستجوی وب

۸- Evaluation & Result Summaries

Original slides by: Hinrich Schütze and Christina Lioma

2

Overview

1. Evaluation benchmarks
2. Unranked evaluation
3. Ranked evaluation
4. Result summaries

What we need for a benchmark

- A collection of documents
  - Documents must be representative of the documents we expect to see in reality.
- A collection of information needs
  - . . .which we will often incorrectly refer to as queries
  - Information needs must be representative of the information needs we expect to see in reality.
- Human relevance assessments
  - We need to hire/pay “judges” or assessors to do this.
  - Expensive, time-consuming
  - Judges must be representative of the users we expect to see in reality.

Standard relevance benchmark

Benchmark	**Description**	**Year**
Cranfield	1398 abstracts of aerodynamics journal articles, 225 queries	1950s
TREC	1.89 million documents (mainly news), 450 information needs (topics)	since 1992
GOV2	25 million web pages	2004
NTCIR and CLEF	cross-language information retrieval	various
20 Newsgroups	18941 articles, 20 topics	~ 1995
Reuters-RCV1	806,791 newswire articles	1996-97

5

Outline

1. Evaluation benchmarks
2. Unranked evaluation
3. Ranked evaluation
4. Result summaries

Measures for a search engine

- How fast does it index
  - e.g., number of bytes per hour
- How fast does it search
  - e.g., latency as a function of queries per second
- What is the cost per query?
  - in dollars

Measures for a search engine

- All of the preceding criteria are measurable: we can quantify speed / size / money
- However, the key measure for a search engine is user happiness.
- What is user happiness?
- Factors include:
  - Speed of response
  - Size of index
  - Uncluttered UI
  - Most important: relevance
- (actually, maybe even more important: it's free)
- Note that none of these is sufficient: blindingly fast, but useless answers won't make a user happy.
- How can we quantify user happiness?

8

Who is the user?

- Who is the user we are trying to make happy?
- Web search engine: searcher. Success: Searcher finds what she was looking for. Measure: rate of return to this search engine
- Web search engine: advertiser. Success: Searcher clicks on ad. Measure: clickthrough rate
- Ecommerce: buyer. Success: Buyer buys something. Measures: time to purchase, fraction of "conversions" of searchers to buyers
- Ecommerce: seller. Success: Seller sells something. Measure: profit per item sold
- Enterprise: CEO. Success: Employees are more productive (because of effective search). Measure: profit of the company

9

Most common definition of user happiness: Relevance

- User happiness is equated with the relevance of search results to the query.
- But how do you measure relevance?
- Standard methodology in information retrieval consists of three elements.
  - A benchmark document collection
  - A benchmark suite of queries
  - An assessment of the relevance of each query-document pair

10

Precision and recall

- Precision (P) is the fraction of retrieved documents that are relevant

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

- Recall (R) is the fraction of relevant documents that are retrieved

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

Precision and recall

	Relevant	Nonrelevant
Retrieved	true positives (TP)	false positives (FP)
Not retrieved	false negatives (FN)	true negatives (TN)

$$P = TP \,/\, (\, TP + FP \,)$$

$$R = TP \,/\, (\, TP + FN \,)$$

12

A combined measure: *F*

- *F* allows us to trade off precision against recall.

$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

- $\alpha \in [0, 1]$ and thus $\beta^2 \in [0,\infty]$
- Most frequently used: balanced *F* with $\beta = 1$ or $\alpha = 0.5$
  - This is the harmonic mean of *P* and *R*: $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$
- What value range of β weights recall higher than precision?

F: Example

	relevant	not relevant	
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

- P = 20/(20 + 40) = 1/3
- R = 20/(20 + 60) = 1/4
- $F_1 = 2\frac{1}{\frac{1}{\frac{1}{3}}+\frac{1}{\frac{1}{4}}} = 2/7$

Optional: Accuracy

- Why do we use complex measures like precision, recall, and *F*?
- Why not something simple like accuracy?
- Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.
- In terms of the contingency table above,

  accuracy = $(TP + TN)/(TP + FP + FN + TN)$.
- Why is accuracy not a useful measure for web information retrieval?

Optional: Exercise

- Compute precision, recall and F_1 for this result set:

	relevant	not relevant
retrieved	18	2
not retrieved	82	1,000,000,000

- The snoogle search engine below always returns 0 results (“0 matching results found”), regardless of the query. Why does snoogle demonstrate that accuracy is not a useful measure in IR?

Optional: Why accuracy is a useless measure in IR

- Simple trick to maximize accuracy in IR: always say no and return nothing
- You then get 99.99% accuracy on most queries.
- Searchers on the web (and in IR in general) want to find something and have a certain tolerance for junk.
- It's better to return some bad hits as long as you return something.
- →We use precision, recall, and *F* for evaluation, not accuracy.

17

Outline

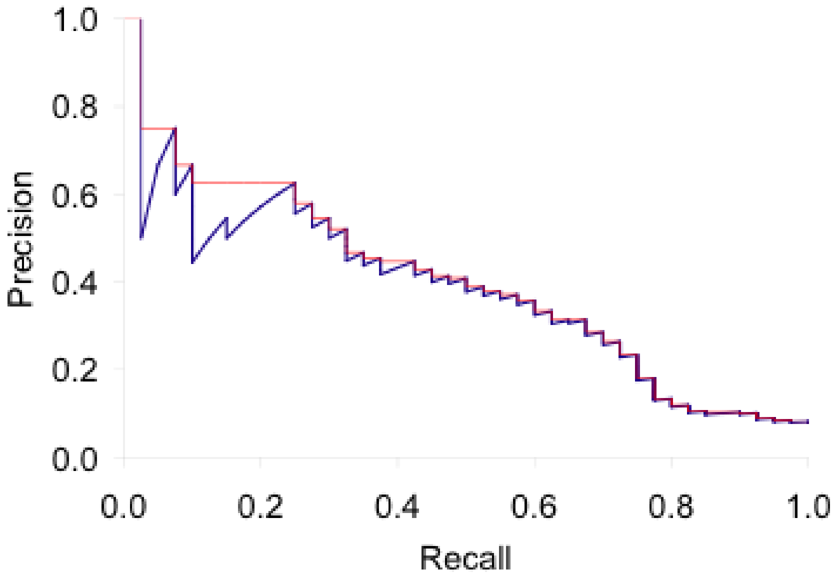
1. Evaluation benchmarks
2. Unranked evaluation
3. Ranked evaluation
4. Result summaries

Precision-recall curve

- Precision/recall/F are measures for unranked sets.
- We can easily turn set measures into measures of ranked lists.
- Just compute the set measure for each "prefix": the top 1, top 2, top 3, top 4 etc results
- Doing this for precision and recall gives you a precision-recall curve.

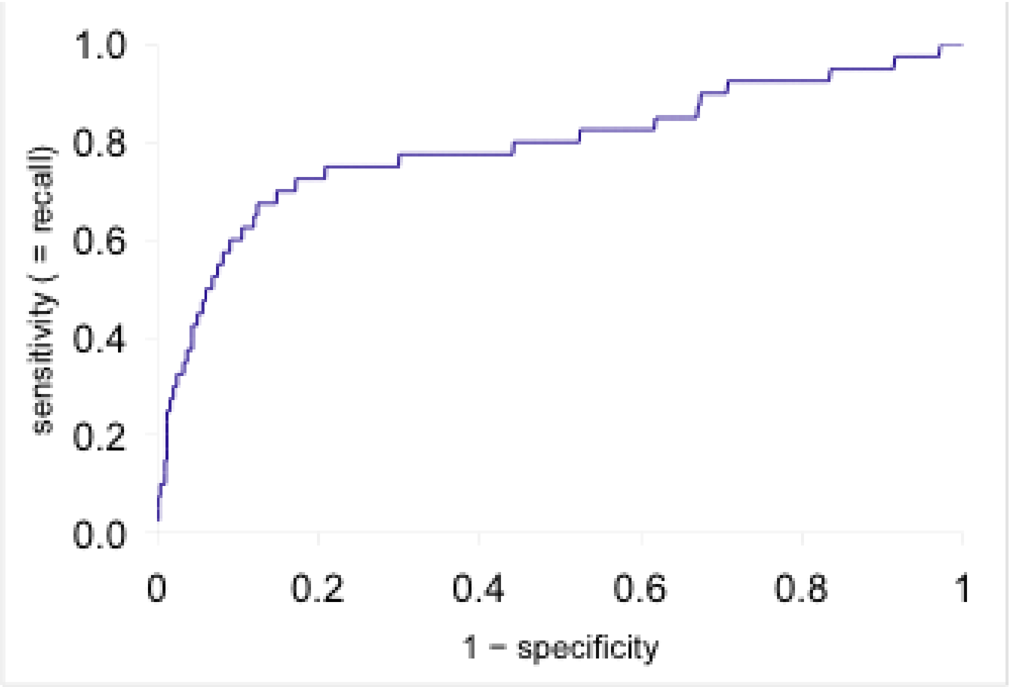
19

A precision-recall curve



- Each point corresponds to a result for the top k ranked hits ($k = 1, 2, 3, 4, . . .$).
- Interpolation (in red): Take maximum of all future points

Optional: ROC curve



- Similar to precision-recall graph
- But we are only interested in the small area in the lower left corner.

21

Validity of relevance assessments

- Relevance assessments are only usable if they are consistent.
- If they are not consistent, then there is no “truth” and experiments are not repeatable.
- How can we measure this consistency or agreement among judges?
- → Kappa measure (optional)

22

Evaluation at large search engines

- Recall is difficult to measure on the web
- Search engines often use precision at top k, e.g., $k = 10$ . . .
- . . . or use measures that reward you more for getting rank 1 right than for getting rank 10 right.
- Search engines also use non-relevance-based measures.
  - Example 1: clickthrough on first result
  - Example 2: Ongoing studies of user behavior in the lab
  - Example 3: A/B testing

A/B testing

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an "automatic" measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness.
- Probably the evaluation methodology that large search engines trust most

24

Optional: Critique of pure relevance

- We've defined relevance for an isolated query-document pair.
- Alternative definition: marginal relevance
- The marginal relevance of a document at position k in the result list is the additional information it contributes over and above the information that was contained in documents $d_1 \ldots d_{k-1}$.
- Exercise
  - Why is marginal relevance a more realistic measure of user happiness?
  - Give an example where a non-marginal measure like precision or recall is a misleading measure of user happiness, but marginal relevance is a good measure.
  - In a practical application, what is the difficulty of using marginal measures instead of non-marginal measures?

25

Outline

1. Evaluation benchmarks
2. Unranked evaluation
3. Ranked evaluation
4. Result summaries

How do we present results to the user?

- Most often: as a list – aka "10 blue links"
- How should each document in the list be described?
- This description is crucial.
- The user often can identify good hits (= relevant hits) based on the description.
- No need to "click" on all documents sequentially

Doc description in result list

- Most commonly: doc title, url, some metadata . . .
- . . . and a summary
- How do we “compute” the summary?

28

Summaries

- Two basic kinds: (i) static (ii) dynamic
- A static summary of a document is always the same, regardless of the query that was issued by the user.
- Dynamic summaries are query-dependent. They attempt to explain why the document was retrieved for the query at hand.

29

Dynamic summaries

- Present one or more “windows” or snippets within the document that contain several of the query terms.
- Prefer snippets in which query terms occurred as a phrase
- Prefer snippets in which query terms occurred jointly in a small window

30

Dynamic summaries

- **We need to cache documents.**
- Snippets should communicate whether and how the document answers the query.
- Ideally: the snippet should answer the query, so we don't have to look at the document.
- Dynamic summaries are a big part of user happiness because ?

31

منابع

- فصل هشتم کتاب An introduction to information retrieval