Information Retrieval (CSD510)

Evaluation

February 14, 2024

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.

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

Relevance: query vs. information need

- Relevance to what?
- First take: relevance to the query
- "Relevance to the query" is very problematic.
- Information need i : "I am looking for information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine."
- This is an information need, not a query.
- Query q: [*red wine white wine heart attack*]
- Consider document d' : *At the heart of his speech was an attack on the wine industry lobby for downplaying the role of red and white wine in drunk driving.*
- d is an excellent match for query q . . .
- d is not relevant to the information need i.

Unranked retrieval - Precision and Recall

- Precision (P) is the fraction of retrieved documents that are relevant
  Precision $= \frac{\#(relevant\ \ items\ \ retrieved)}{\#(retrieved\ \ items)}$
- Recall (R) is the fraction of relevant documents that are retrieved
- Recall $= \frac{\#(relevant\ \ items\ \ retrieved)}{\#(reelevant\ \ items)}$

Unranked retrieval - Precision and Recall

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

$$
\begin{aligned}
P &= TP/(TP+FP) \\
R &= TP/(TP+FN)
\end{aligned}
$$

Unranked retrieval - Precision and Recall

- You can increase recall by returning more docs.
- Recall is a non-decreasing function of the number of docs retrieved.
- A system that returns all docs has 100
- The converse is also true (usually): It's easy to get high precision for very low recall.

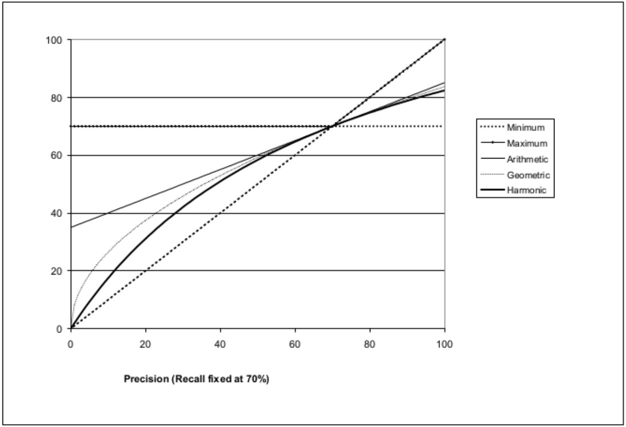
A combined measure: F

- F allows us to trade off precision against recall.
  $F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$, where, $\beta^2 = \frac{1-\alpha}{\alpha}$
- $\alpha\ in[0,1]$ and $\beta^2 = [0, \infty]$
- Most frequently used: balanced F with $\beta = 1$ or $\alpha = 0.5$
  - This is the harmonic mean of P and R: $\frac{1}{F} = \frac{1}{2}(\frac{1}{P} + \frac{1}{R})$
- Accuracy is the fraction of decisions (relevant/nonrelevant) that are correct.
- In terms of the contingency table above, accuracy $= (TP + TN)/(TP + FP + FN + TN)$
- Simple trick to maximize accuracy in IR: always say no and return nothing.
- You then get 99.99% accuracy on most queries.

F: Why harmonic mean?

- Why don't we use a different mean of P and R as a measure?
  - e.g., the arithmetic mean
- **Objective:** Punish really bad performance on either
- Taking the minimum achieves this.
- But minimum is not smooth and hard to weight.
- F (harmonic mean) is a kind of smooth minimum. precision or recall.

$F1$ and other averages



- We can view the harmonic mean as a kind of soft minimum

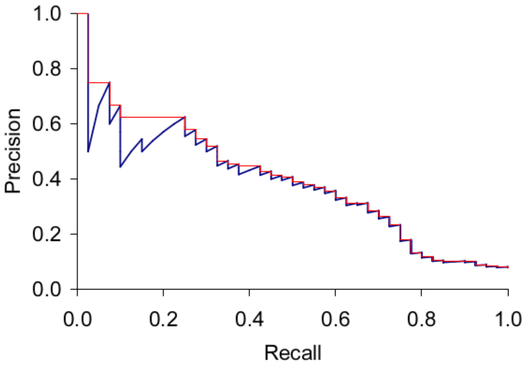
Difficulties in using precision, recall and F

- We need relevance judgments for information-need-document pairs – but they are expensive to produce.
- For alternatives to using precision/recall and having to produce relevance judgments – see end of this lecture.

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.

A precision-recall curve



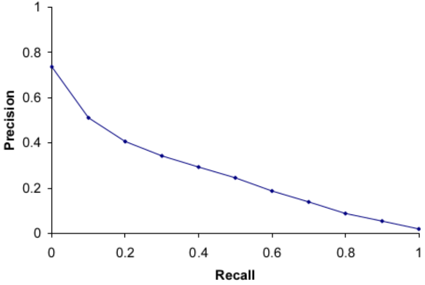
- Each point corresponds to a result for the top k ranked hits ($k = 1, 2, 3, 4, \ldots$)
- Interpolation (in red): Take maximum of all future points
- Rationale for interpolation: The user is willing to look at more stuff if both precision and recall get better.

11-point interpolated average precision

Recall	Interpolated Precision
0.0	1.00
0.1	0.67
0.2	0.63
0.3	0.55
0.4	0.45
0.5	0.41
0.6	0.36
0.7	0.29
0.8	0.13
0.9	0.10
1.0	0.08

11-point average: $\approx$ 0.425

Averaged 11-point precision/recall graph



- Compute interpolated precision at recall levels 0.0, 0.1, 0.2, ...
- Do this for each of the queries in the evaluation benchmark
- Average over queries
- This measure measures performance at all recall levels.
- The curve is typical of performance levels at TREC.
- Note that performance is not very good!

Mean average precision (MAP)

- Most standard among the TREC community is Mean Average Precision (MAP).
- Provides a single-figure measure of quality across recall levels.
- For a single information need, Average Precision is the
  - average of the precision value obtained for the set of top k documents existing after each relevant document is retrieved.
  - this value is then averaged over information needs
- If the set of relevant documents for an information need $q_i \in Q$ is $d_1, \cdots, d_m$ and R_{jk} is the set of ranked retrieval results from the top result until you get to document d_k, then

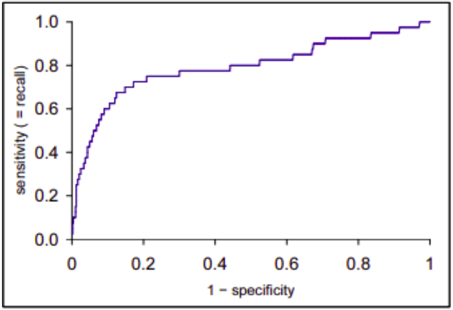
$$MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} PRECISION(R_{jk})$$

- When no relevant document retrieved, the precision value is taken to be 0.
- The MAP is roughly the average area under the precision-recall curve for a set of queries.

PRECISION AT k

- The precision and recall based measures may not be germane to users such as web surfers.
- The number of relevant documents in the first few pages matters the most.
- **Precision at k:** Number of relevant documents in the first k documents.
- It has the advantage of not requiring any estimate of the size of the set of relevant documents.

ROC curve



- Plots the true positive rate (sensitivity) against the false positive rate (1 − specificity).
- sensitivity = recall = true positive rate = $\frac{tp}{tp+fn}$
- false positive rate = (1-specificity)=$\frac{fp}{(fp+tn)}$
- Sensitivity is not a good measure since the number of *true negatives* are very large w.r.t. *true positives*.

NDCG - Normalized Discounted Cumulative Gain

- NDCG is designed for situations of non-binary notions of relevance
- It is evaluated over some number k of top search results
- For a set of queries Q, let $R(j, d)$ be the relevance score assessors gave to document d for query j, then

$$NDCG(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)} - 1}{log_2(1 + m)}$$

- Z_{kj} is a normalization factor calculated to make it so that a perfect ranking's NDCG at k for query j is 1.

Kappa agreement

- Relevance judgments given by humans are quite idiosyncratic and variable.
- The success of an IR system depends on how good it is at satisfying the needs of these idiosyncratic humans.
- To address the variation in relevance judgement by the humans we attempt to measure the agreement between human judges.
- Measure: **Kappa statistic**
- $kappa = \frac{P(A)-P(E)}{1-P(E)}$
  - $P(A)$ is the proportion of the times the judges agreed
  - $P(E)$ is the proportion of the times they would be expected to agree by chance.
- $P(E)$ estimated from *marginal statistics* to calculate expected agreement.