AIW & Information Retrieval (UE18CS322)
Unit 3

Aronya Baksy

March 2021

1 Evaluation Metrics for IR Systems

- Evaluation may be done quantitatively in terms of
  - Index construction time
  - Search Time (queries per second)
  - Cost per query (engg cost vs revenue gained)
  - User satisfaction (UI, speed, relevance)
- In web search, the user is either a **searcher** (rate of return of search engine) or an **advertiser** (click through rate). Both their satisfaction are correlated
- In an enterprise setting, **buyers** (time to purchase, fraction of "conversions" of searchers to buyers), **sellers** (profit per item sold) and **executives** (company profit) are the users. Buyer satisfaction is correlated with seller/executive satisfaction.
- Relevance always measured w.r.t. **information need**, not actual queries.

1.1 Relevance Benchmark

- Consists of a benchmark corpus, information needs expressed as queries, and a binary judgement of relevance for each query-doc pair.

1.1.1 Cranfield experiments

- Premise: Retrieved documents' relevance is a good proxy of a system's utility in satisfying users' information need
- Procedure:
  1. 1398 abstracts from journal articles on aerodynamics, and 225 queries.
  2. Exhaustive relevance judgments of all (query, document) pairs
  3. Compare different indexing system over this collection

1.1.2 TREC Benchmark

- Text Retrieval Conference, organized by the NIST (USA). The most popular benchmark from this set is the TREC Ad Hoc used between 1992 and '99.
- 1.89 million documents, mainly NewsWire articles, 450 information needs
- No exhaustive relevance judgments – too expensive
- Rather, NIST assessors' relevance judgments are available only for the top K documents returned by a system which was entered in the TREC evaluation
- **GOV2** is another TREC/NIST collection consisting of 25 million webpages.
- **NTCIR** for East Asian languages, and cross-language IR

	relevant	nonrelevant
retrieved	true positive (TP)	false positive (FP)
not retrieved	false negative (FN)	true negative (TN)

Figure 1: Confusion Matrix

1.2 Unranked Evaluation

- Precision : fraction of retrieved documents that are relevant. Recall: fraction of relevant documents that were retrieved.
- Precision prefers systems that retrieve less number of highly relevant documents, Recall prefers high number of retrieved documents.
- Precision decreases as the number of retrieved documents increases (unless in perfect ranking), while recall keeps increasing
- The F-score is a combination of the precision and recall:

$$P = \frac{TP}{TP + TN} \tag{1}$$

$$R = \frac{TP}{TP + FN} \tag{2}$$

$$F = \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}} \tag{3}$$

- The HM is used as it is a smooth minimum, and it punishes bad performance on either the P or the R. It is known that $HM \leq GM \leq AM$
- AM for a naive engine that simply returns everything as result is 0.5 which is too high.

1.3 Ranked Evaluation

- Equation (3) can also be written as

$$F = \frac{1}{\frac{\alpha}{P} + \frac{1-\alpha}{R}} = \frac{(\beta + 1)PR}{\beta P + R} \tag{4}$$

$$\beta = \frac{1 - \alpha}{\alpha} \tag{5}$$

- $\beta < 1$ emphasizes precision, $\beta > 1$ emphasizes recall.
- On one query, different systems tend to perform the same. On different queries, there is large variance in performance of a single system
- COmputing P and R for top 1, 2, 3, 4, and so on documents leads to a P-R curve (P on vertical, R on horizontal). A P-R curve is a sawtooth:
  - Next document considered is irrelevant, hence recall constant but precision reduces (vertical decrease)
  - Next document considered is relevant, hence recall and precision increase, a rightward upward movement.
- Interpolate the points by taking maximum of all future points
- The **Mean Average Precision** (MAP) is a summary measure derived from the graph.

- For 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
- MAP is the average of average precisions over all the queries considered.
- For a single query, average precision approximates the area under the precision recall curve
- Set of information needs should be diverse for MAP to be an effective measure across systems