Chapter 8: Evaluation in information retrieval

- How do we determine whether certain IR system is effective or not?
- IR is a highly *empirical discipline*
- User utility? User happiness?

8.1 Information retrieval system evaluation

- *Relevant* and *nonrelevant* documents
- Gold standard or ground truth judgment of relevance
- A test suite of information needs to be of a reasonable size: around 50 information needs has usually been found to be a sufficient minimum.
- Relevance is assessed relative to an information need, *not a query.*
- To evaluate a system, we require an *overt expression* of an information need.
- We will assume a binary decision of relevance.
- Wrong to report results on a test collection after tuning the parameters against it: Need to use one or more *development test collections* instead for tuning

8.2 Standard test collections

- Focus particularly on test collections for *ad hoc information retrieval system*
- The Cranfield collection
- TREC
- GOV2
- NTCIR: East Asian language and cross-language information retrieval
- CLEF
- Reuters-21578 and Reuters-RCV1
- 20 Newsgroups

8.3 Evaluation of unranked retrieval sets

- Most frequent and basic measures: *Precision* and *Recall*
- Accuracy: Not an appropriate measure for IR problems
  - The data is extremely skewed: Normally over 99.9% of the documents are in the nonrelevant category
  - Trying to label some documents as relevant will almost always lead to a high rate of false positives
- But users are assumed to have a certain tolerance for seeing some false positives, as long as they get some useful information.
- The advantage of using precision and recall: one is more important than the other in many circumstances.
- You can always get a recall of `1` by retrieving *all* documents for all queries.
- Precision usually decreases as the number of documents retrieved is increased.
- In general, we want to get some amount of recall while tolerating only a certain percentage of false positives.
- *F measure*: The weighted harmonic mean of precision and recall
  - $F = \frac{1}{\alpha \frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R}$ where $\beta^2 = \frac{1-\alpha}{\alpha}$
  - Balanced F measure F_1: Equal weights; $\alpha = \frac{1}{2}$ or $\beta = 1$.
  - $\beta < 1$ emphasize precision while $\beta > 1$ emphasize recall.
- Why harmonic mean?: We can reach the arithmetic mean of 50% when we get 100% recall by retrieving all documents for all queries
  - However, assuming 1 document in 10,000 is actually relevant, the harmonic mean would be 0.02%
  - The harmonic mean is always less than or equal to the arithmetic mean and geometric mean.
  - *When the values of two numbers differ greatly, the harmonic mean is closer to their minimum than to their arithmetic mean.*

8.4 Evaluation of ranked retrieval results

- Previously mentioned evaluation measures are computed using unordered set of documents
- We need to extend these to evaluate the *ranked* retrieval results.
- *Precision-Recall Curve*: Distinctive saw-tooth shape; If the $(k+1)$-th document retrieved is nonrelevant then recall is the same as for the top `k` documents, but precision has dropped.
  - If it is relevant, then both precision and recall increase, and the curve jags up and to the right.
  - **Interpolated precision** p_{interp} at a certain recall level r: The *highest* precision found for any recall level $r' \geq r$; $\max_{r' \geq r} p(r')$

- **11-point interpolated average precision**
  1. For each information need in the *test collection*, measure interpolated precisions at the 11 recall levels: $0.0, 0.1, 0.2, \cdots, 1.0$.
  2. At each recall level, calculate the arithmetic mean of all the interpolated precisions.
- **Mean Average Precision (MAP)**: Have especially good discrimination and stability
  - 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.
  - Then MAP is the average of average precisions over information needs.
  - $\text{MAP}(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk})$
  - When a relevant document is not retrieved at all, the precision value would be 0.
  - For a single information need, the average precision approximate the area under the uninterpolated precision-recall curve
  - SO the MAP is roughly the *average area* under the precision-recall curve, for a set of queries.
  - Since the MAP for a test collection is the arithmetic mean of average precision values of individual information needs, this has the effect of weighting each information need *equally.*
  - There is normally more agreement in MAP for a particular information need across systems, than for MAP scores for different information needs for the same system.
    - A set of test information needs must be large and diverse enough to be representative of system effectiveness across diffferent queries.
- **Precision at** k: Measuring precision at fixed low levels of retrieved results
  - Least stable and does not average well, as the total number of relevant documents for a query has a strong influence on precision at k.
- **R-precision**: Kinda like Precision at k
  - Have a some known set of relevant documents called rel: This set may not be exactly complete.
    - This allows us to adjust for the actual size of relevant documents, in contrast to precision at k with a fixed size
  - Obtain top $|rel|$ results from the system.
  - If the total of `r` results are found to be relevant, then
    - R-precision is $\frac{r}{|rel|}$.
    - Note that recall is also $\frac{r}{|rel|}$, hence R-precision is identical to the *break-even point.*
  - Somewhat unclear why you should be interested in break-even point rather than the point which maximizes F measure or other relevant points.
  - Empirically, R-precision is highly correlated with MAP.
- **ROC Curve**: Plots the true positive rate against *false positive rate*
  - For a good system, the graph climbs steeply on the left side.
  - Specificity (True negative rate?) was not seen as very useful notion as the value would be always almost 1 for all information needs.
    - The 'interesting' part of precision-recall curve is $0 < \text{recall} < 0.4$.
  - Report the area under the ROC curve
- **Normalized Discounted Cumulative Gain (NDCG)**: Often employed with ML approaches to ranking
  - Designed for ***non-binary*** notions of relevance
  - Evaluated over some number k of top results
  - For a set of queries Q, let $R(j, d)$ be the relevance score given to document d for query j.
  - $NDCG(Q, k) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} Z_{kj} \sum_{m=1}^{k} \frac{2^{R(j,m)} - 1}{\log_2(1+m)}$
    - Z_{kj} is a normalization factor to make the perfect NDCG at k for a single query j to be 1
    - For queries for which $k' < k$ documents are retrieved, the last summation is done up to k'.

8.5 Assessing relevance

- Test information needs should be appropriate for predicted usage of the system.
  - Using random combination of query terms as an information need is not a good idea: They will not resemble the actual distribution of information needs
- We also need to collect *relevance assessments*
  - For large collections, usual for relevance to be assessed only for a subset of the documents for each query
  - *Pooling*: Relevance assessed over a subset of the collection, which is formed from the top k documents returned by different IR systems and perhaps other sources
- Humans are not good at reliably reporting gold standard judgements
  - Their relevance judgements are quite idiosyncratic and variable. But we need to satisfy these people's needs anyway
  - *Kappa statistic*: A common measure for agreement between judges; designed for categorical judgements
    - Corrects a simple agreement rate for the rate of *chance* agreement

        - 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
            - If they are making a two-class decision and we have no more assumptions, then this would be 0.5
            - However, normally the class distribution assigned is *skewed*, so we use *marginal* statistics to calculate the expected agreement.
            - Marginals across judges or separate marginals for each judges?: Across judges is more conservative in the presence of systematic differences in assessments across judges
        - Rule of Thumb:
            - Kappa value above 0.8: Good agreement
            - Between 0.67 and 0.8: Fair agreement
            - Below 0.67: Data providing a *dubious* basis for an evaluation
    - In the TREC evaluations and medical IR, the level of agreement normally falls in the range of *fair*.
        - The fact that we see modest human agreement on a binary judgement is one reason for not requiring more fine-grained relevance labeling.
    - Making one or the other of two judges' opinions as the gold standard: Litle impact on the *relative* effectiveness ranking

8.5.1 Critiques and justifications of the concept of relevance

- The standard model of relevant/non-relevant documents makes it possible to carry out comparative experiments
- In practice, despite the simplification, the standard formal measures are quite good enough, and recent work for optimizing formal evaluation measures in IR has succeeded brilliantly.
- But There are still some problems latent within the abstractions used:
    - The relevance of one document is treated as independent of the relevance of other documents in the collection.
    - Assessments are binary; there aren't any nuanced assessments of relevance. They are simply treated as absolute and objective decisions.
- *Marginal relevance*: Whether a document still has distinctive usefulness *after* the user has looked at certain other documents
    - Maximizing marginal relevance requires returning documents that exhibit diversity and novelty.
        - Using distinct facts or entities as evaluation units? - directly measures true utility to the user but harder to create a test collection

8.6 A broader perspective: System quality and user utility

- Despite formal evaluation measures, we are ultimately interested in how *satisfied* each users are with the results the system gives?
- Could do user studies
- Other system aspects that allow quantitative evaluation and the issue of user utility

8.6.1 System issues

- How fast does it index/search?
- Query language expressiveness
- How large is its document collection?

8.6.2 User utility

- User happiness is elusive to measure

8.6.3 Refining a deployed system

- A/B testing
    - Change exactly one thing and evaluate measures on randomly selected users

8.7 Results snippets

- Wish to present a results list *informative enough* for the users to do final ranking of the documents
- *Snippet*: A short summary of the document, designed to allow the user to decide its relevance
    - Title
    - Short summary: How to design automatically extract this?
- *Static* summaries and *dynamic* (query-dependent) summaries
    - Dynamic summaries attempt to explain *why* a particular document was retrieved

- Better ways to do text summarization: How to choose good sentences?
  - Combine positional factors (first and last sentences, etc.) and content factors (emphasizing sentences with key terms that have low DF but high frequency and good distribution with the document)
  - Or doing full text generation
- Keyword-in-context (KWIC) snippets: Summaries that contain one or several of the query terms
- Dynamic summaries are generated in conjunction with scoring: users prefer snippets that read well because they contain complete phrases.
- Dynamic summaries make IR system designs complicated