MSCC-IRWS
Information Retrieval and Web Search
Evaluation 2

Jason Farina

Faculty of Computing Science
Griffith College Dublin

jason.farina@gcd.ie

Where are we?

- Done:
  - Basic precision / recall
  - Interpolated precision
- To Do:
  - Precision at n documents
  - R-precision
  - Mean Average Precision
  - bPref
  - Which metric to choose??

Example Reminder

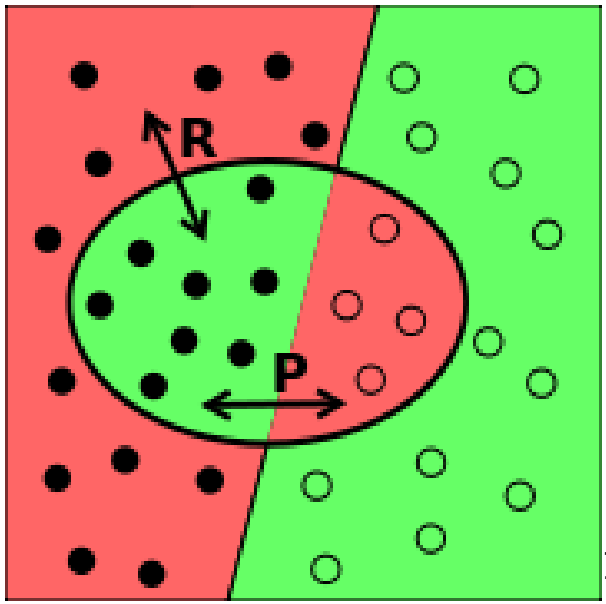
- Consider a query, q, on document collection C where $|C| = 800$
- $Rel = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$
- The ranked list of retrieved documents, RET, is given by:

1.	d_{123} (r)	6.	d_9 (r)	11.	d_{38}
2.	d_{84}	7.	d_{511}	12.	d_{48}
3.	d_{56} (r)	8.	d_{129}	13.	d_{250}
4.	d_6	9.	d_{187}	14.	d_{113}
5.	d_8	10.	d_{25} (r)	15.	d_3 (r)

F1-Score

- Single value metrics are helpful to have
- One of the most common metrics produces a harmonic combination of precision and recall
- This is the F1 score (also known as the E-measure) proposed by van Rijsbergen.
- This is given as: $F1 = \frac{1+\beta^2}{\frac{\beta^2}{R}+\frac{1}{P}}$

F1-Score



[1]

[1] image taken from wikipedia

F1-Score

- looking at the formula in more detail:
- We know that
- $RECALL = \frac{\#RetrievedRelevant}{\#Relevant}$
- $\#RetrievedRelevant = \#TruePositives$
- $\#Relevant = \#TruePositives + FalseNegatives$
- So $Recall(R) = \frac{TP}{TP+FN}$

F1-Score

- $PRECISION = \frac{\#RetrievedRelevant}{\#Retrieved}$
- $\#RetrievedRelevant = \#TruePositives$
- $\#Retrievedt = \#TruePositives + FalsePositives$
- So $Precision(P) = \frac{TP}{TP+FP}$
- We allow α to be our weighting so if $\alpha(P)$ is the weight of precision then$(1 - \alpha)R$ will be the weight of Recall

F1-Score

- What is β?
  - β allows us to decide the relative importance of precision and recall in evaluating performance
  - When $\beta = 1$ we refer to the metric as the balanced metric.
  - This puts equal weight on precision and recall
  - As β drops more weight is placed on precision
  - As it rises more weight is placed on recall

F1-Score

- This measure allows us to represent the effectiveness of our model for a user that places β times as much importance on recall over precision. The higher β, the more relevant Recall is to the user
- $F = \frac{(1+\beta^2)PR}{\beta^2 P+R}$
- This formula is derived from CJs "*EffectivenessMeasure*"(chapter 7 p 120 onwards) $F = 1 - E$

Precision at n / R-Precision

- Sometimes we are interested in the precision amongst the top n results
- This is particularly suited to web search systems as people look no further than the first few results.
- For instance we may be interested in the precision after 3 documents have been retrieved (P@3).
- In the case of our example system this value is 0.66 as two of the first three documents were relevant to the query.
- Similarly we might be interested in the precision at a particular recall level.
- For instance the R-precision at $R = 0.3$ in our original example is 0.5 (i.e. at the point where we reached a recall level of 30%, the precision was 0.5).

Mean Average Precision

- Mean Average Precision (MAP) is the most commonly used metric in IR literature to compare the performance of systems.
- It involves three steps:
  1. Firstly, we must calculate the precision for at each relevant document. This is very similar to how we calculated the precision at the various recall points when we were drawing interpolated precision/recall graphs. Relevant documents that are not included in the result set are given a precision of 0.

Mean Average Precision

2. The *Average Precision* for this query is found by taking the average of these precision scores (i.e. *Average Precision* is for one query).
3. The same procedure must be performed for all the queries being used for evaluation. Taking the Average Precision for each query, we get the average of these, which gives us *Mean Average Precision* (*Mean Average Precision* (MAP) is for multiple queries).

Example

- We calculate the precision at each relevant document in the result set. Documents not returned are given a precision score of 0.

Document	Precision	Document	Precision
d_{123}	1.0	d_5	0.0
d_{56}	0.67	d_{39}	0.0
d_9	0.5	d_{44}	0.0
d_{25}	0.4	d_{71}	0.0
d_3	0.33	d_{89}	0.0
		Average	0.29

- MAP is calculated over all of the queries as the average of averages!

bPref: Binary Preference

- All of the evaluation techniques that we have mentioned to this point are based upon the *Cranfield Paradigm*.
- In this, *test collections* and queries are created that have a known set of relevant documents associated with them.
- The key point is that for each query, *every* document in the collection is judged to be relevant or non-relevant.
- The metrics that we have studied require these *complete relevance judgments*.

bPref: Binary Preference

- With small collections this Cranfield Paradigm is perfect (i.e. there are complete relevance judgments).
- However, as document collection have become larger, complete judgments have become less common and we have *incomplete judgments*. This means that some documents have not been judged so they may or may not be relevant to the test queries.
- With large-scale IR collections (such as those based on the web) this complete judgment is impossible to achieve, with potentially millions of documents to be judged for relevance against hundreds of queries.

bPref: Binary Preference

- For the evaluation metrics we have seen so far, they are simplified by assuming that unjudged documents are non-relevant.
- It was noticed that many unjudged documents had an adverse effect on the evaluation scores being achieved.
  - **NB:** This is not to say the retrieval was worse: after all, the fact of a document being relevant or not is not affected by whether or not somebody has judged it.
- This is problematic as evaluation scores are no longer accurately reflecting the effectiveness of the retrieval.
- For an example, what would the effect on the average precision score on our example be if nobody had judged document d_9?

bPref: Binary Preference

- The idea behind *bpref*[2] is that these unjudged documents should not impact so largely on the evaluation score
- *bPref* functions by ignoring those documents that have not been judged.
- The only ones considered are the judged relevant documents and the judged non-relevant
- This is the accepted metric for large-scale IR systems where complete relevance judgments are impossible to achieve

[2] Buckley & Voorhees, Retrieval Evaluation with Incomplete Information, SIGIR 2004

bPref: Binary Preference

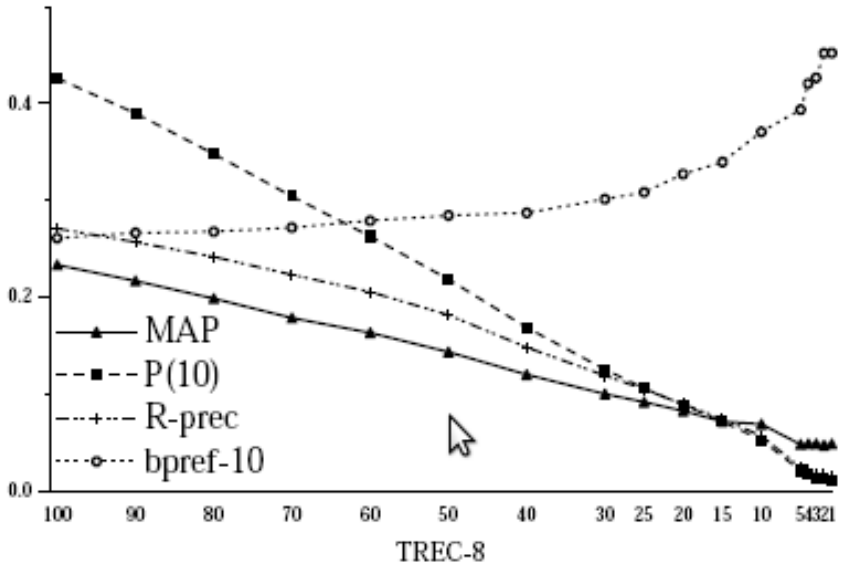
- bPref, for a query with R relevant documents is calculated as:
  - $B = \frac{1}{R} \sum_r 1 - \frac{|n \text{ ranked higher than } r|}{R}$
- where $r \in R$ and n is a member of the first R judged non-relevant documents
- In other words:
  - For each relevant document in the result set:
    - Count the number of non-relevant documents above it in the result set (this is $|n$ ranked higher than $r|$). This cannot be greater than the total number of relevant documents (R)
    - The score for that document is $1 - \frac{|n \text{ ranked higher than } r|}{R}$
  - Average this score over all documents

bPref: Binary Preference

- This works well in most situations.
- However, When R is very small (i.e. there are only one or two relevant documents) it fails
- To get over this we use bPref-10 instead given by:
  - $B10 = \frac{1}{R}\sum_r 1 - \frac{|n \text{ ranked higher than } r|}{10+R}$
- where n is a member of the first $10 + R$ judged non-relevant documents

Effect of bPref

- From Buckley and Voorhees 2004 it can be seen that when complete judgments are available there is no noticable difference between MAP and bPref-10
- As the judgments become less complete bpref is more stable than the others (graph from Buckley & Voorhees)



Which metric should I use?

- We have looked at numerous different metrics for IR evaluation
- All have their advantages and disadvantages
- We need to use an appropriate metric in order to evaluate an IR system's performance
- How this is defined is dependent on the final use of the system

Which metric should I use?

Uses of IR systems

- IR systems are used in many places
  - web search;
  - intranet search;
  - research environments;
  - desktop search;
  - ...

Which metric should I use?

- A general rule is that if complete judgments are available any metric can be used
- MAP gives a good indication of performance within a single metric as it is averaging the results over multiple queries
- For collections that do not have complete judgments, bPref is a more suitable metric
- Real-world IR problems normally do not follow this and hence bPref is becoming more standard
- For tasks such as web search (due to user behaviour), metrics like P@10 might be used.