MLRF Lecture 04

J. Chazalon, LRDE/EPITA, 2019

IR evaluation

Lecture 04 part 03

How to evaluate a retrieval system?

We need a set of queries for which we know the expected results
“Ground truth”, aka “targets”, “gold standard”…

To compare 2 methods, we need to use the same database and the same queries.

Many measures / indicators.

Core criterion: is a result relevant (binary classification)?

Precision and Recall

Used to measure the balance between

- Returning many results, hence a lot of the relevant results present in the database, but also a lot of noise
- Returning very few results, leading to less noise, but also less relevant results

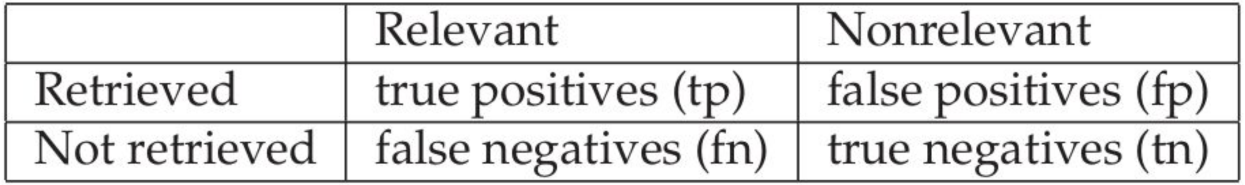
Precision and Recall

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

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



	Relevant	Nonrelevant
Retrieved	true positives (tp)	false positives (fp)
Not retrieved	false negatives (fn)	true negatives (tn)

$$\begin{aligned} P &= tp/(tp+fp) \\ R &= tp/(tp+fn) \end{aligned}$$

F-measure

F measure is the weighted harmonic mean of precision and recall

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

where α ∈ [0, 1] and thus β² ∈ [0, ∞]

The default value is β = 1, leading to:

$$\boxed{F_{\beta=1} = \frac{2PR}{P+R}}$$

How to evaluate a ranked retrieval system?

When results are ordered, more measures are available.

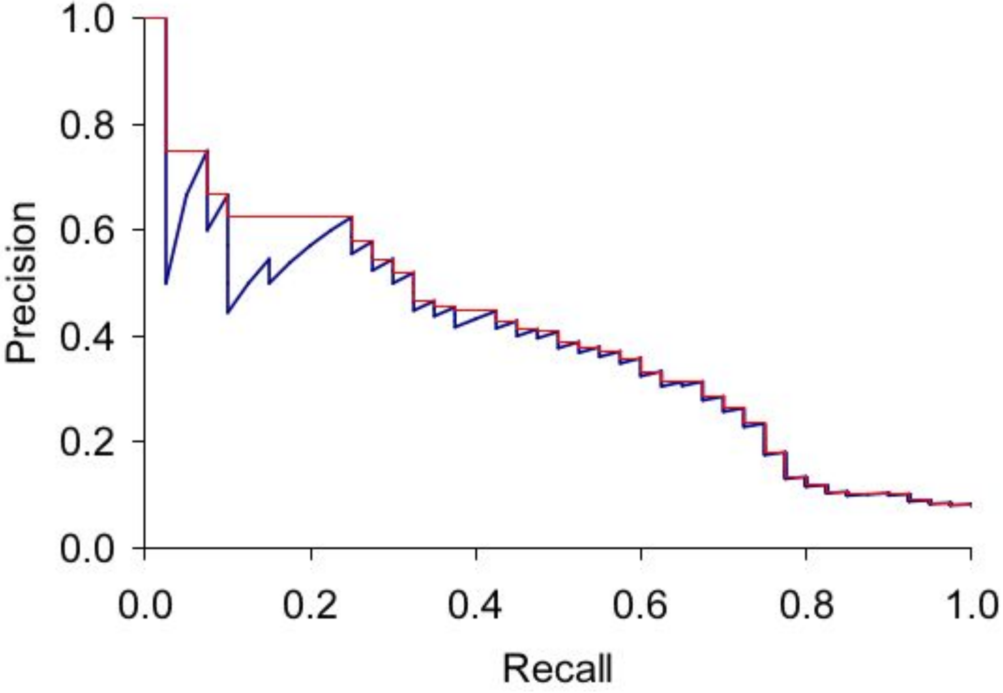
Common useful measures are:

- The precision-recall graph and the mean average precision
- The ROC graph and the area under it (AUC)

Precision-recall graph

Plotting the points

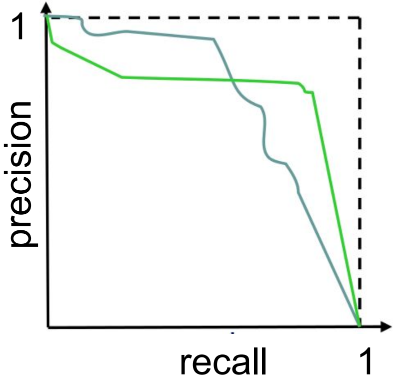
```
For a given query
For each result
    if the result is relevant
    set x = #tp / #expected
    set y = #tp / #returned
```



The recall always increases while we scan the result list.

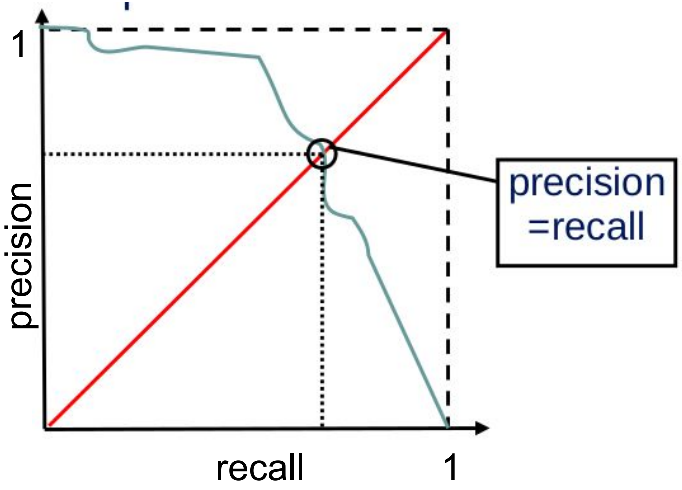
Equal Error Rate and Average Precision

Which one is the best?

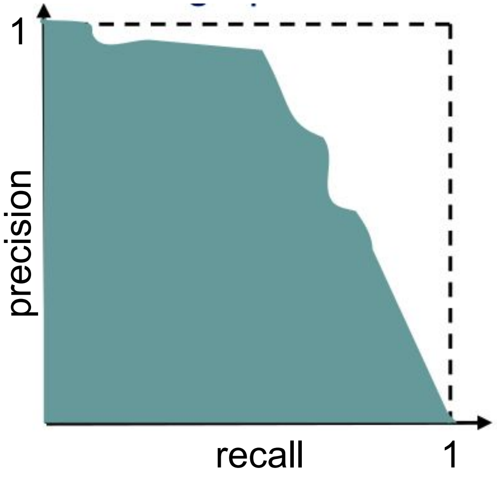


Note: the PR graph does not provide a total order
⇒ need more indicators

Equal Error Rate



Average precision



Mean average precision at k — mAP (@k)

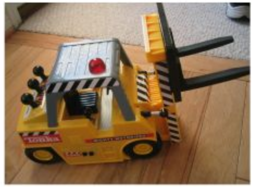
Mean of the average precision of several queries,
when considering k results for each query

⇒ makes evaluation tractable with very large databases

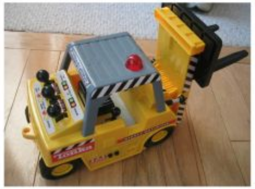
Computed using the trapezoid technique [on board]

Exercise

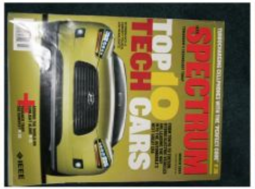
For this query and the following results, plot the precision/recall graph and compute the average precision.



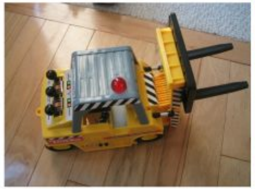
1



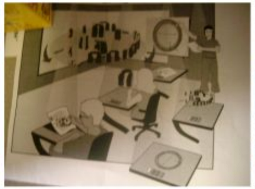
2



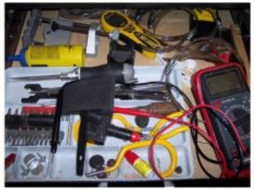
3



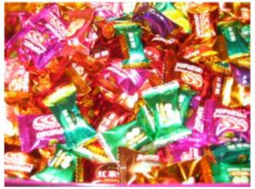
4



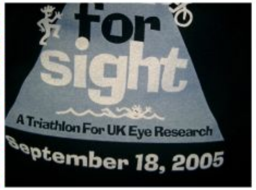
5



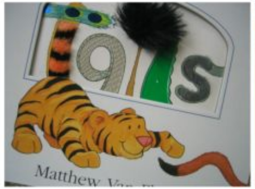
6



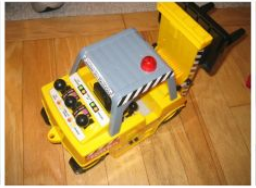
7



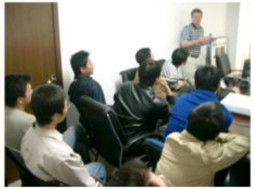
8



9



10



ROC & others

[next time, more useful for classification]