GBGI9U07: multimedia document: description and automatic retrieval

2. Evaluation of indexing and retrieval methods

Georges Quénot

Multimedia Information Indexing and Retrieval Group

Laboratory of Informatics of Grenoble

March 2022

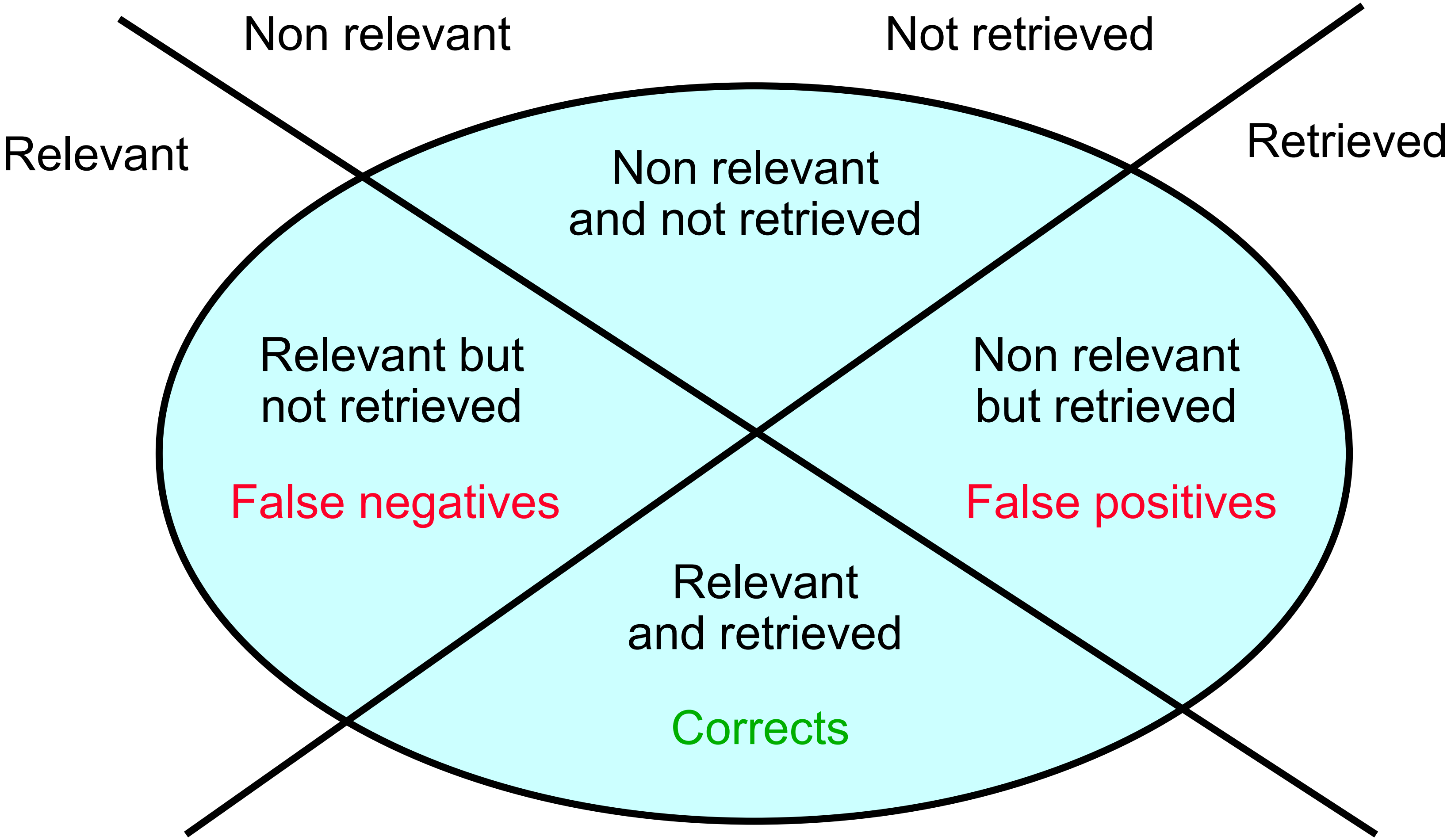
Evaluation : general principles

- A well posed problem or “task”:
  - A corpus,
  - A “ground truth”,
  - A metric,
  - A protocol.
- Annotation / assessment.
- Periodical workshops.
- Organizers and participants.
- Collaborative work.
- Results and presentation of methods.

Tasks : classification or search

- Classification:
  - Split a set into positives and negatives,
  - Predefined classes to recognize,
  - Classical learning from examples,
- Search:
  - Find documents relevant for a query,
  - No predefines classes,
  - The query may be seen as an example (or a set of examples),
  - Higher level learning (the system learns its optimal parameters from development collections).

Metrics: precision and recall
From relevant and non relevant sets



Metrics: precision and recall
From relevant and non relevant sets

$$\text{Recall} = \frac{\text{Retrieved and Relevant}}{\text{Relevant}} = \frac{\text{Corrects}}{\text{Relevant}}$$

$$\text{Precision} = \frac{\text{Retrieved and Relevant}}{\text{Retrieved}} = \frac{\text{Corrects}}{\text{Retrieved}}$$

$$\text{F-measure} = \frac{2 \text{ x Corrects}}{\text{Retrieved + Relevant}}$$

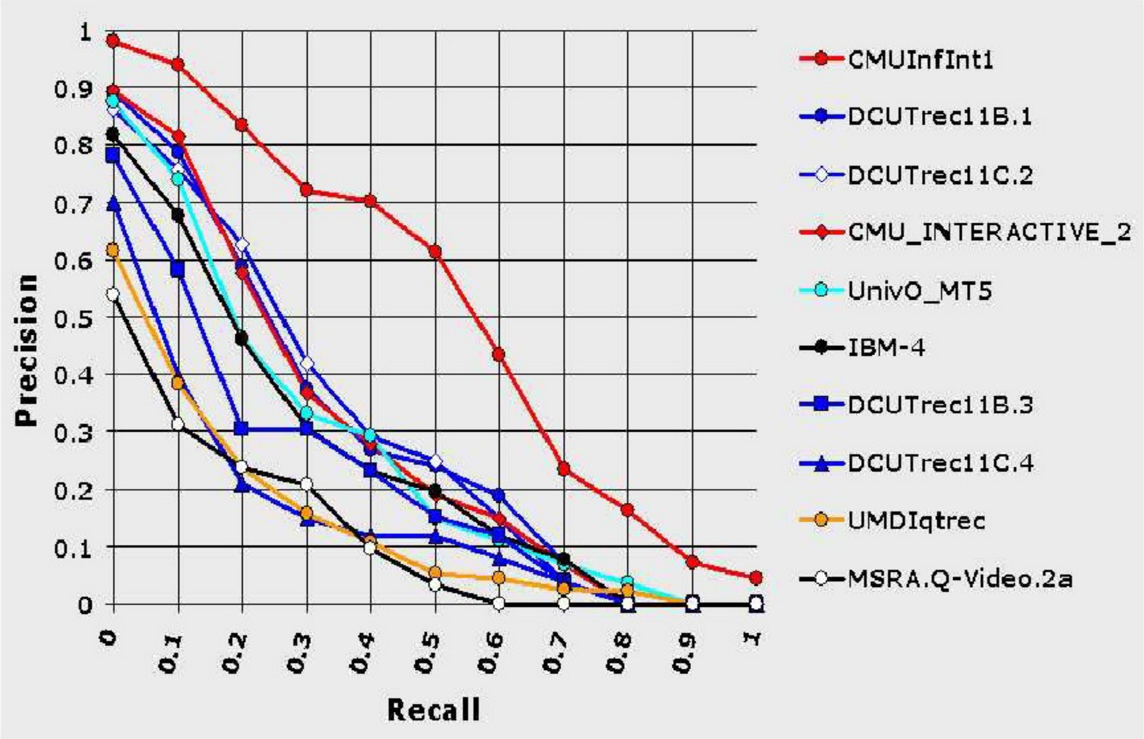
$$\text{Error rate} = \frac{\text{False positives + False negatives}}{\text{Relevant}}$$

Metrics: Recall × Precision curves
From ranked lists

- Results ranked from most probable to least probable: more informative that just “relevant / non relevant”.
- For each *k*: set Ret_k of the *k* first retrieved items
- Fixed set *Rel* of the relevant items
- For each *k*: Recall(Ret_k, *Rel*), Precision(Ret_k, *Rel*)
- Curve joining the (Recall, Precision) points with *k* varying from 1 to *N* = total number of documents.
- Interpolation: Precision = f(Recall) → Continuous curve

- “Standard” program: **`trec_eval`**
  (ranked lists, relevant sets) → RP curve, MAP, ...

Metrics: Recall × Precision curves From ranked lists



- Mean Average Precision (MAP): area under the Recall × Precision curve (`trec_eval`)

Global measures

MAP: Mean Average Precision

$$\text{F-measure} = \frac{2 \text{ x Corrects}}{\text{Retrieved + relevant}}$$

P@10: precision on the10 first documents

P@100: precision on the100 first documents

$$\text{Error rate} = \frac{\text{False positives + False negatives}}{\text{Relevant}}$$

Pooling

- Practical impossibility to judge all documents for all queries,
- A posteriori judgment on a small part of the corpus only,
- Fusion of the N first elements of the list from the set of tested systems (N = from 100 to 1000 typically),
- Judgment of these elements only,
- Documents not judged are considered as non relevant,
- The computation is done as if everything was judged.

Pooling

- Bias : relevant documents are ignored:
  - Recall is (generally) over-estimated,
  - Precision is (generally) under-estimated.
- Bias is small if:
  - There are enough queries,
  - There are enough systems,
  - Pooling is deep enough.
- Similar effect for the whole set of systems
  - Comparison between systems are significant,
  - The ranking between systems is stable.