Introduction to Information Retrieval

Chapter 2: Evaluation of IRS

Dr. SAID KADRI

Associate Professor

Department of Computer Science, Faculty of Mathematics and Informatics,

University Mohamed Boudiaf of M'sila

E-mail: kadri.said28@gmail.com

Website: https://kadrisaid28.wixsite.com/sgadri

Table of content

Evaluation of Information Retrieval Systems IRS

Evaluation criteria for an IR system:

- Fast indexing
- Fast searching
- Expressivity of the query language
- Size of the supported collection.
- User interface (clearness of the input form and of the output list, e.g. snippets, etc).

Evaluation of IRS: Used Measures

Precision and Recall

- Given a query:
  - ❑ ***Are all retrieved documents relevant?***
  - ❑ ***Have all the relevant documents been retrieved?***

- Measures for system performance:
  - ❑ The first question is about the precision of the search

$$Precision = \frac{\#relevant\ \ retrieved}{\#retrieved}$$

⇨ The proportion of relevant retrieved documents to all retrieved documents:

■ The second is about the completeness (recall) of the search.

$$Recall = \frac{\#relevant\ retrieved}{\#relevant}$$

⇨ The proportion of relevant documents that are retrieved, out of all relevant documents available (retrieved + not retrieved)

In other term:

	Relevant	**Not relevant**
Retrieved	TP (True Positive)	FP (False Positive)
Not retrieved	FP (False Positive)	TN (True Negative)

Table of contingency

$$Precision = \frac{tp}{tp + fp}$$

$$Recall = \frac{tp}{tp + fn}$$

F-Measure

- Measure relating precision and recall (combine the two previous measures: precision, recall)
- Uses The weighted harmonic average of precision and recall, the traditional F-measure or balanced F-score is:

$$F = \frac{2 * Precision * Recall}{Precision + Recall}$$

Accuracy

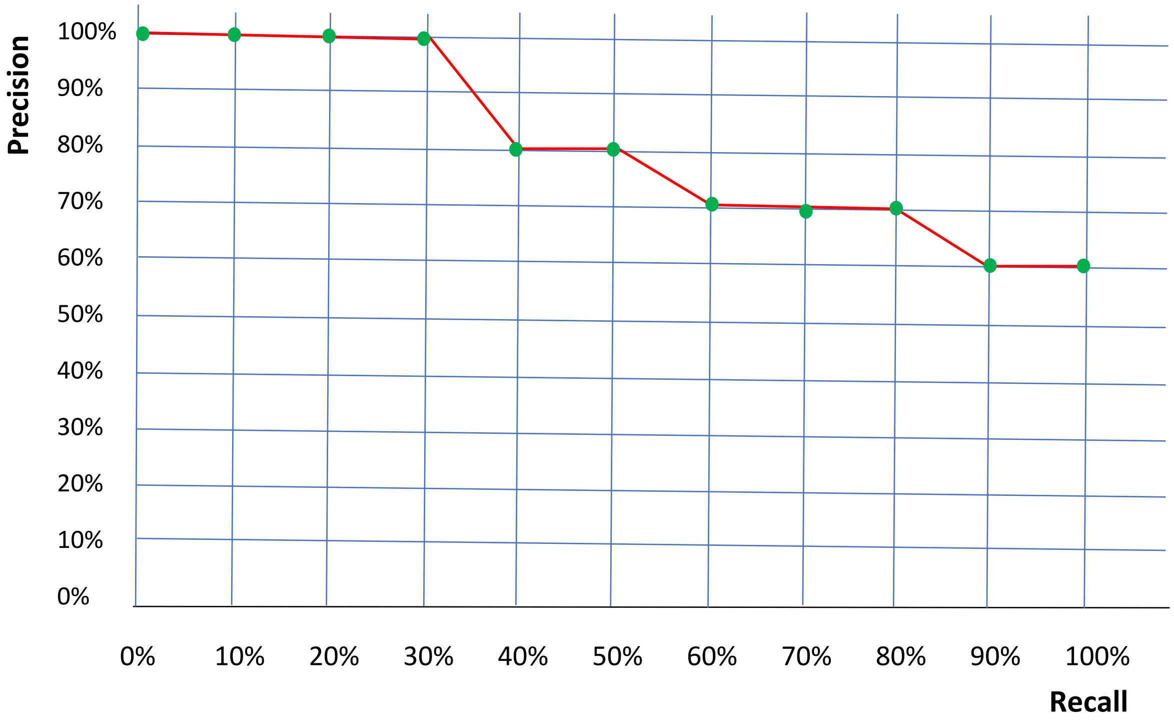
Proportion of the classification relevant/not relevant that is correct

$$accuracy = \frac{tp + tn}{tp + fp + tn + fn}$$

Precision-recall curve

Example: in the following example, we obtain the interpolated at all 11 recall levels in the table. The precision-recall curve is shown below.

i	$P(r_i)$	r_i
0	100%	0%
1	100%	10%
2	100%	20%
3	100%	30%
4	80%	40%
5	80%	50%
6	71%	60%
7	70%	70%
8	70%	80%
9	62%	90%
10	62%	100%



The Precision-Recall Curve

Compare different retrieval algorithms

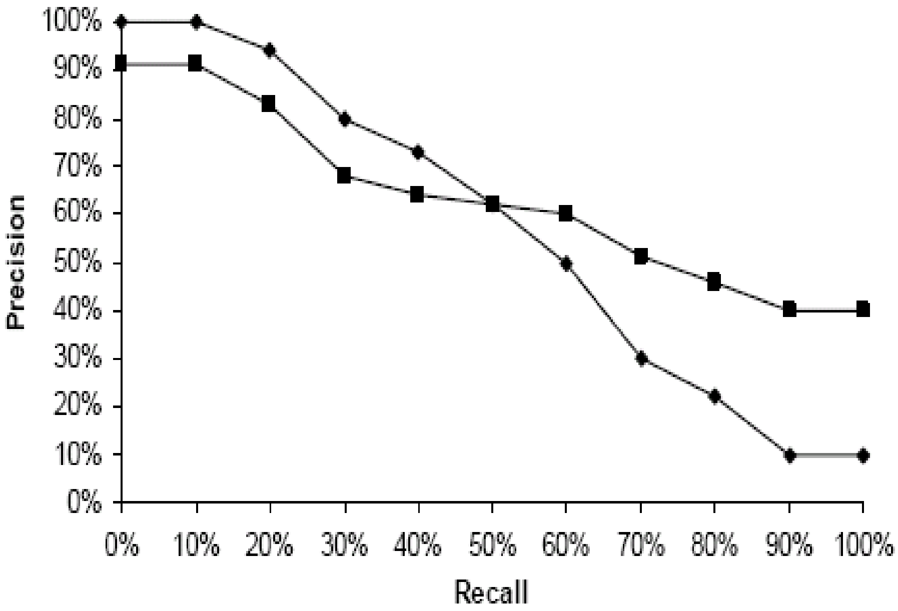


Fig. 6.5. Comparison of two retrieval algorithms based on their precision-recall curves

Compare with multiple queries

- Compute the average precision at each recall level.

$$\overline{p}(r_i) = \frac{1}{|Q|}\sum_{j=1}^{|Q|} p_j(r_i), \tag{22}$$

where Q is the set of all queries and $p_j(r_i)$ is the precision of query j at the recall level r_i. Using the average precision at each recall level, we can also draw a precision-recall curve.

- Draw precision recall curves
- Do not forget the F-score evaluation measure.

Rank precision

- Compute the precision values at some selected rank positions.
- Mainly used in Web search evaluation.
- For a Web search engine, we can compute precisions for the top 5, 10, 15, 20, 25 and 30 returned pages
  - ❑ As the user often looks at more 30 pages.
- Recall is not very meaningful in Web search.
  - ❑ Why?
  - ❑ Because of the large collection of documents and its variation)

Standard train and test collections

Name	Size	Description
Cranfield collection	1398 abstracts of journal articles, plus 255 queries.	about aerodynamics, gathered in UK in the 1950s
TREC (Text REtrieval Conference)		collection maintained by the US National Institute of Standards and Technology since 1992
TREC Ad Hoc Track	1.89 million documents for 450 topics	test collection used for 8 evaluation compagnies led from 1992 to 1999,
TREC 6-8	over 528000 newswires	
GOV2	25 millions of webpages	collection maintained by the NIST, larger than other test collections, but smaller than current collection supported by WWW search engines)
NTCIR (Nii Test Collection for IR systems)	//	various test collections focusing on East Asian languages, mainly used for cross-language IR
CLEF (Cross Language evaluation Forum)	//	Collection focusing on European languages
REUTERS : Reuters 21578 and REUTERS RCV1	21 578 newswires articles and 806791documents,	mainly used for text classification