Evaluation of Information Retrieval Systems
M2R - MOSIG

Philippe Mulhem

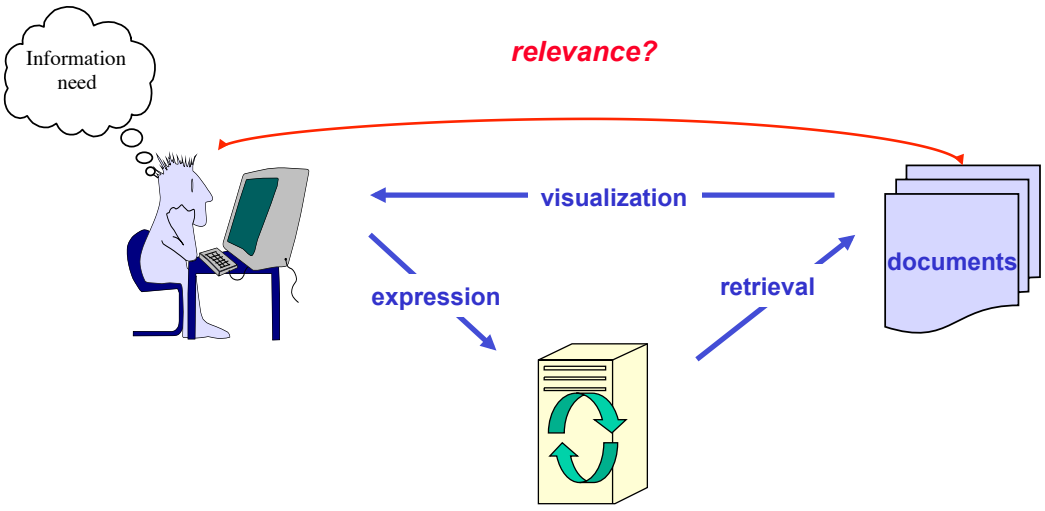
Philippe.Mulhem@imag.fr

Outline

1. Introduction
2. Recall/Precision measures
3. Recall/Precision curves
4. Mean Average Precision
5. F-measure
6. Precision@x documents
7. Cumulated Gain
8. Test Collection
9. trec_eval
10. Conclusion

1. Introduction

- Challenge of Information Retrieval:
  - Content base access to documents that satisfy an user's information



1. Introduction

- Parameters
  - the effort, intellectual ou physical, needed to users to express queries, and see result documents
  - response time
  - display of results (user's capability to use the retrieved documents)
  - corpus quality according to the user's needs
  - capability of the system to retrieve all the relevant documents and to avoid retrieving unrelevant ones.

1. Introduction

- For the last point, comparing IRSs in a theoretical way (using their model) is a unsolved problem

$\Longrightarrow$ use black box tests

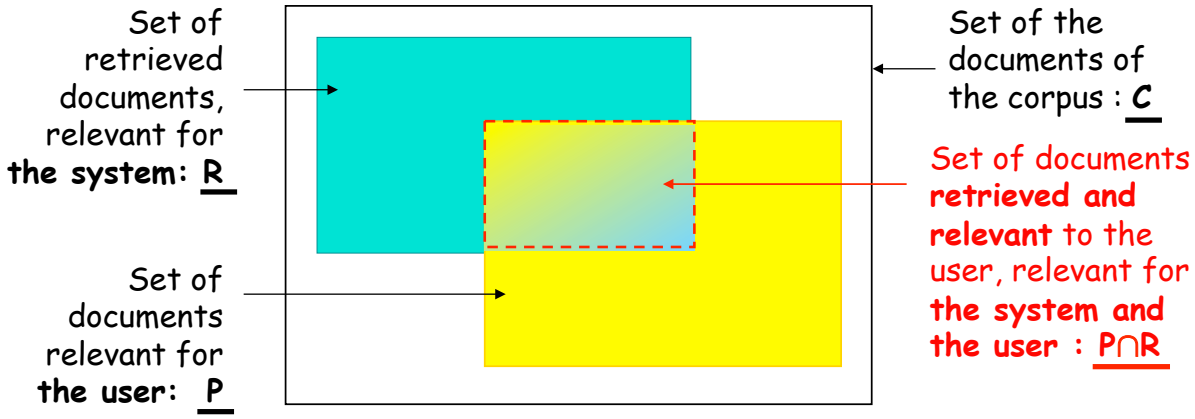
we match up the results of a system in comparison to others, when considering ideal answers to given queries.

1. Introduction

- Test collection (Cranfield Paradigm)
  - a set of documents (corpus) C
  - a set of queries on C
  - a set of relevant documents for each query
  - one (or several) evaluation measure (s)

2. Recall/precision measures

- Objective :
  - Compare user and system relevances

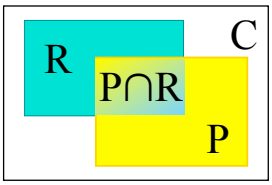


2. Recall/precision measures

- The essential criteria are:
  - Recall: ability of the system to give in the answer all the relevant documents according to the user
  - Precision : ability of the system to give in the answer only relevant documents according to the user
  - These two criteria are antagonistic…

2. Recall/precision measures

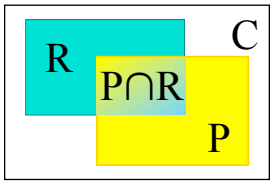
- The recall is the ratio of
  - The number of retrieved documents by the system and relevant to the user
  - Divided by the number of all the documents of the corpus that are relevant to the user



$$recall = \frac{|P \cap R|}{|P|} \qquad \in [0,1]$$

2. Recall/precision measures

- The precision is the ratio of
  - The number of retrieved documents by the system and relevant to the user
  - Divided by the number of the documents retrieved by the system



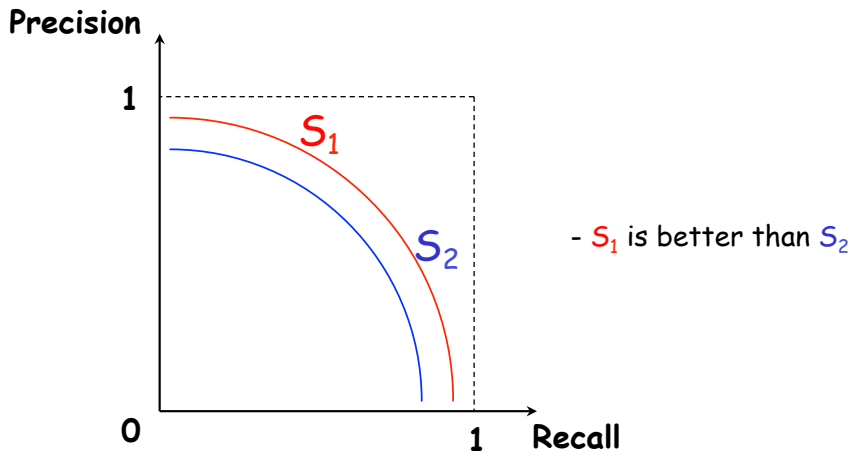
$$precision = \frac{|P \cap R|}{|R|} \qquad \in [0,1]$$

2. Recall/precision measures

- For one query and one system : 2 real values
  - Example: a system gives 5 documents, among them 3 are relevant, knowing that there are 10 relevant documents in the corpus:
    - Recall = 3 / 10
    - Precision = 3 / 5
- We need more detailed evaluations
  - Recall/precision diagrams

3. Recall/precision diagrams

- Recall/Precision diagrams
  - Comparison of 2 systems S1 et S2



3. Recall/precision diagrams

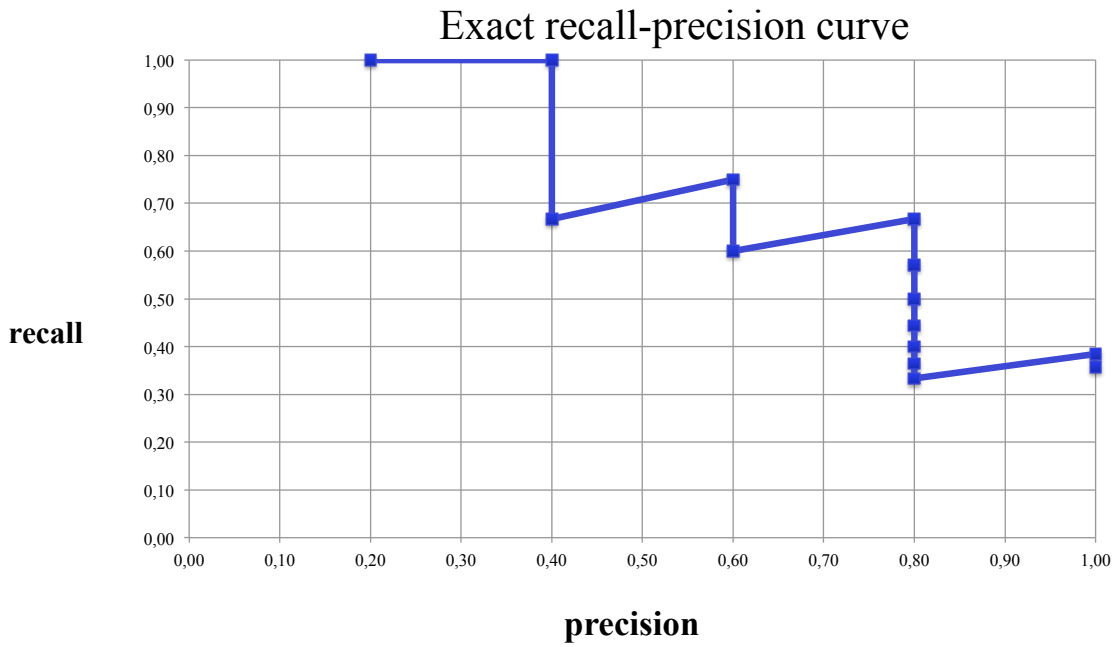
- Recall/Precision diagrams (2)
  - Show the evolution of the precision and the recall with sorted results
  - Method:
    - We compute the precision and the recall when considering only the first document as answer, then we do the same for the two first results of the system, and so on, until each retrieved document is processed.

3. Recall/precision diagrams

- corpus of 200 documents, Q has 5 relevant docs {572, 588, 589, 590, 592}

			recall	precision
		is relevant	p and r / p	p and r / r
1	588	1	0,20	1,00
2	589	1	0,40	1,00
3	576		0,40	0,67
4	590	1	0,60	0,75
5	986		0,60	0,60
6	592	1	0,80	0,67
7	884		0,80	0,57
8	988		0,80	0,50
9	578		0,80	0,44
10	985		0,80	0,40
11	103		0,80	0,36
12	591		0,80	0,33
13	572	1	1,00	0,38
14	990		1,00	0,36

3. Recall/precision diagrams



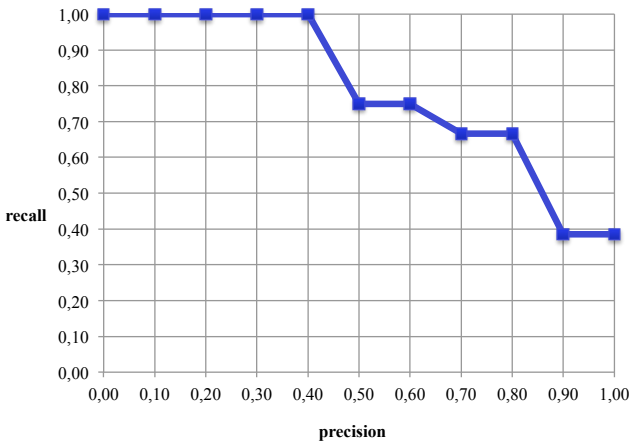
3. Recall/precision diagrams

- Recall/Precision diagrams
  - fix the 11 recall points 0, 0.1, 0.2, …, 0.9, 1
  - Rule of the maximum

    for each recall point v_r, keep the max of precision from recall greater or equal than v_r
  - For instance, in the figure :
    - at recall 0.1, we have the maximum precision at 1, obtained at recall 0.2
  - If a recall point does not have a precision value according to the rule of maximum, then force the precision to 0 (i.e. the min precision value).

3. Recall/precision diagrams

Recall	Precision
0	1
0.1	1
0.2	1
0.3	1
0.4	1
0.5	0.75
0.6	0.75
0.7	0.6667
0.8	0.6667
0.9	0.3846
1	0.3846

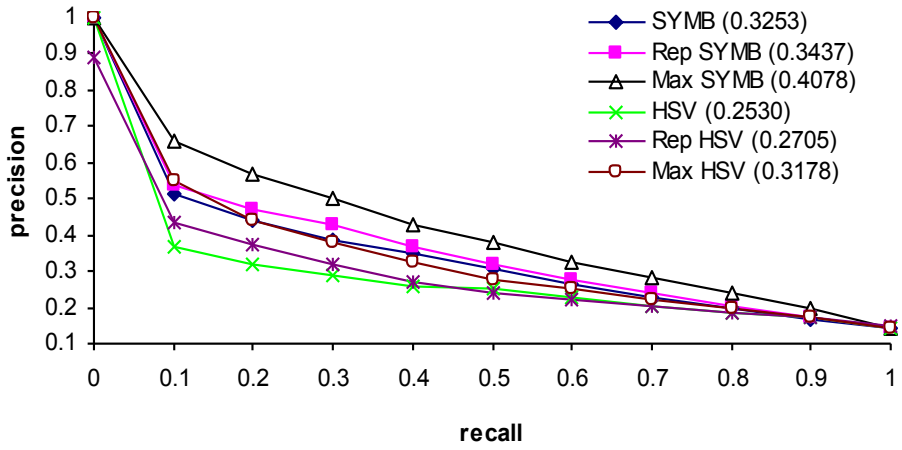


3. Recall/precision diagrams

- For nbQ queries, average on each of the 11 recall points for all the nbQ queries, to obtain the overall recall/precision diagram of a system.

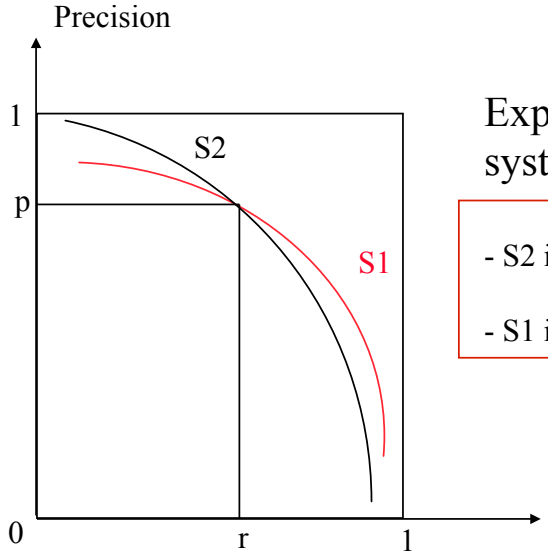
3. Recall/precision diagrams

- Recall/Precision diagrams (7)
  - A real diagram (from [Mulhem & al. 2003])



3. Recall/precision diagrams

- Comparing systems

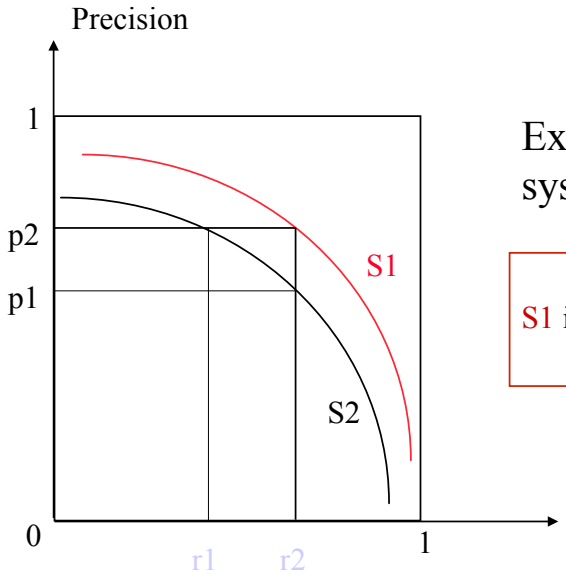


Experimental comparison of systems on a test collection :

- S2 is better than S1 for precision
- S1 is better than S2 for recall

3. Recall/precision diagrams

- Comparing systems



Experimental comparison of systems on a test collection :

S1 is, on average, always better than S2

4. Mean Average Precision

- AP and MAP
  - The idea here is to get a general view of the quality of a system, using only one value.
  - AP : average precision for one query
    - precision computed after each relevant document, averaged
      - on the previous example, aP=0.7603
  - MAP mean of the average precision over each query

5. F-measure

- Integrates recall and precision in one value (harmonic mean)
- General form :

$$F_\beta = \frac{(1+\beta^2).precision.recall}{\beta^2.precision + recall}$$

- General use in IR,

$$F_1 = \frac{2.precision.recall}{precision + recall}$$

6 Precision @x documents

- We evaluate the precision after x documents retrieved, and average over queries
- Useful when evaluating system for first results (10 or 20 for instance)
  - for instance in our example :
    - P@5 = 0.60
    - P@10=0.40
    - P@15=0.33

7. Cumulated Gain

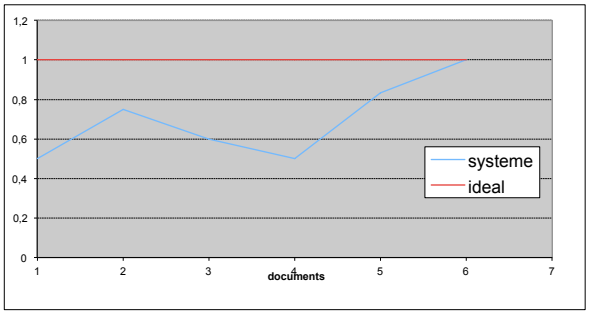
- Cumulated Gain
  - Use of the result list from a system for a query : R
    - Ex: R = <d_{23}, d_{56}, d_9, d_{135}, d_{87}, d_4>
  - Obtain the gain value for each of the documents: G[j]=gain(R[j])
    - Ex : G= <1, 2, 0, 0, 2, 1>
  - Definition of cumulated gain at rank i : $CG[i] = \sum_{j=1}^{i} G[j]$
    - Ex list : CG=<1, 3, 3, 3, 5, 6>

7. Cumulated Gain

  - Normalization by using an ideal list I, list of the gains of the relevant documents for R sorted by decreasing gain value
    - Ex : I=<2, 2, 1, 1, 0, 0>
  - Cumulated Gain for the ideal list between the position 1and i : $CI[i] = \sum_{j=1}^{k} I[k]$
    - Ex : CI=<2, 4, 5, 6, 6, 6>
  - Normalized Cumulated Gain : $nxCG[i] = \frac{xCG[i]}{xCI[i]}$
    - Ex : nxCG=<0.5, 0.75, 0.6, 0.5, 0.83, 1>

7. Cumulated Gain

- curve obtained on the example



- Cumulated gain compares an ideal result list to the result obtained
- Takes into account non binary values of relevance, which is good, but difficult to interpret results

8. Test collections

- As explained before, a test collection is ideally composed of a set of resolved queries.
  - queries representative of real user interests
  - diverse queries (subject, style, vocabulary)
  - large number (> 30)
- For a large corpus (100K or more), it is difficult to evaluate queries on the full corpus

→ use of *pooled* results [Voorhes 2001]

- we run the queries on several systems, each system get a list of result per query
- we make a union of each results per query
- we evaluate user relevance on the sets generated (so, not all the collection)

8. Test collections

- Impact on "global" recall/precision values
  - potential decrease of precision
  - potential increase of recall
- BUT
  - For the MAP, it has been shown that the ranking of systems are kept.

- Note: it impacts if your system is not used in the pool, because results that may be relevant are marked non-relevant…

9. Trec-eval

- Software downloaded on internet. It generates the tables for the recall/precision diagrams and avg. prec. @ 5, 10, 20, 50 and 100 documents, and other measures

  - http://trec.nist.gov/trec_eval/trec_eval.8.1.tar.gz

10. Conclusion

- Limitations
  - Binary relevance assessments for precision/recall based measures (unrealistic but widely used). INEX tried to extend this on structured documents (interpolated recall/precision).
  - Not many evaluations with Cumulated Gain
  - On large collections, difficult to make evaluations
    - One solution (TREC) pool the results for several systems.
  - Hypothesis that relevance is independent of the ranking
    - In reality : If D_1 is presented before D_5, then may be D_5 is not relevant any more, because D_1 contains similar information that D_5 for a user need.

10. Conclusion

- To do
  - Understand classical IR evaluation (Cranfield Paradifgm)
  - Understand recall/precision measures and curves (redo the example, and make others removing on relevat document found, etc.)
  - Understand the nDGC computation.

Bibliography

- R. Baeza-Yates and B. Ribeiro-Neto, Retrieval Evaluation, Chapter 3 of Modern Information Retrieval, Addison Wesley 1999.
- J. Tague-Stucliffe, The pragmatics of Information Retrieval Experimentation, Revisited, Information Processing and Management, 28(4), 467-490, Elsevier, 1992.
- D. Harmann, The TREC Conferences, Proceedings of HIM'95, Konstanz, pp. 9-28.
- K. Järvelin and J. Kekäläinen, Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems 20(4), 422-446, 2002.
- E. Voorhees, The Philosophy of Information Retrieval Evaluation, Proceedings of the second Workshop CLEF on Evaluation of Cross_Language Information Retrieval Systems, pp. 355-370, LNCS 2406, Springer Verlag, 2001.