Information Access and Retrieval (GBX9MO23)

Evaluation of Information Retrieval Systems

M2R – MOSIG

2021-2022

Philippe Mulhem

Philippe.Mulhem@imag.fr

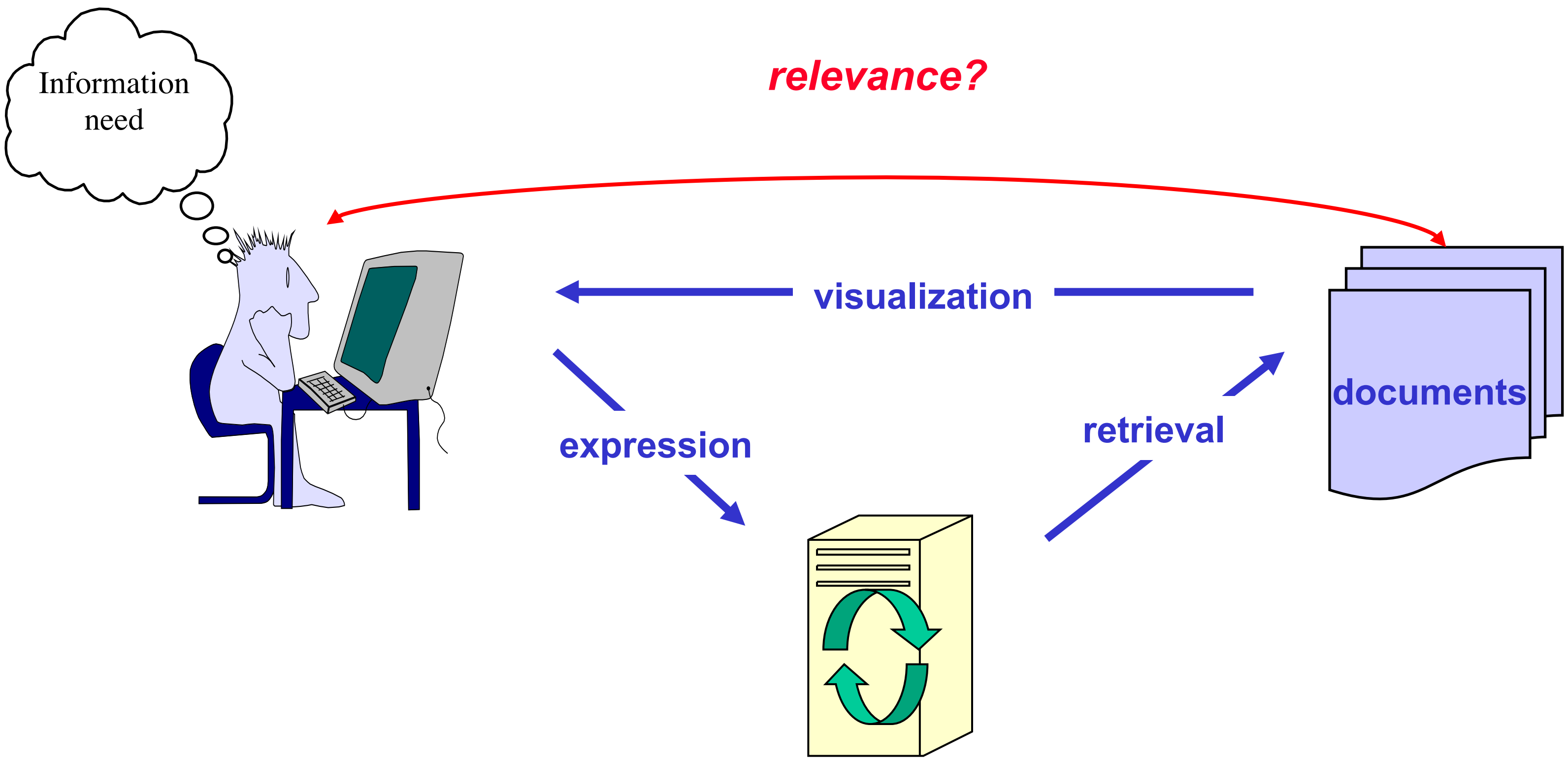
http://lig-membres.imag.fr/mulhem/

Outline

1. Introduction
2. Recall/Precision measures
3. Recall/Precision curves
4. Mean Average Precision
5. F-measure
6. Precision@x documents
7. Discounted Cumulated Gain
8. Test Collection
9. trec_eval software
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 or physical, needed to users to express queries
  - 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 (retrieval of relevant docs), comparing IRSs in a theoretical way (using their model) is a unsolved problem

⇒ so: use black box tests

We compare the results of a system with *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
    - Expert users assess the relevance of each doc of the corpus according to each query

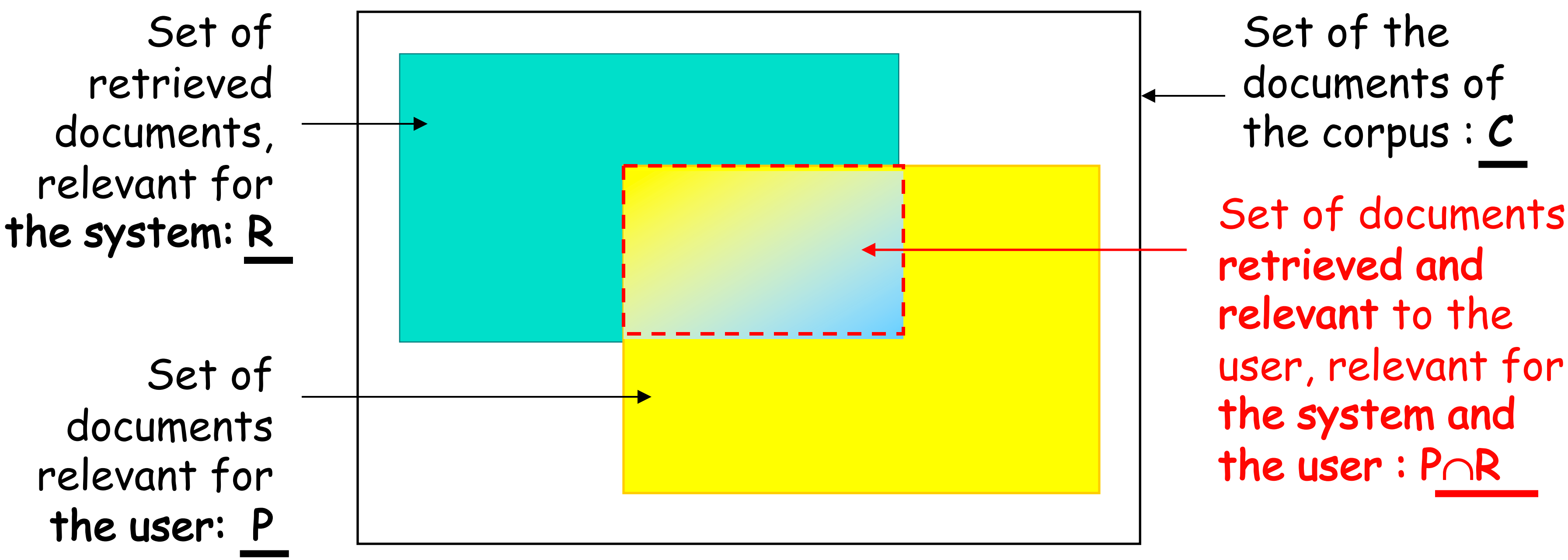
    → These are the ideal answers

      → classically binary: relevant/non-relevant

      → may be numbers (4 → highly relevant, …, 0 → non-relevant)
  – one (or several) evaluation measure (s)
    - Well defined
    - That analyse one aspect of the quality of systems
      – Ex. quality of the system for the top-documents in the result, …

2. Recall/precision measures

- To compare user (ideal) and system relevances:
  - Using binary relevance assessments



2. Recall/precision measures

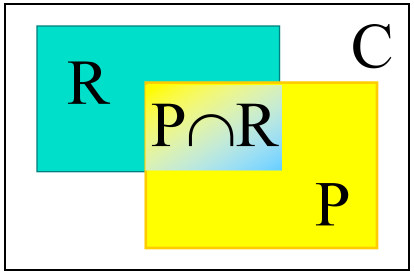
- The essential criteria are:
  - *recall*: ability of the system to give in the answer all the relevant documents according to the user
  - *precision*: ability of the system to give in the answer only relevant documents according to the user

These two criteria are antagonistic:

- Most of the time, when we improve one we degrade the other…

2. Recall/precision measures

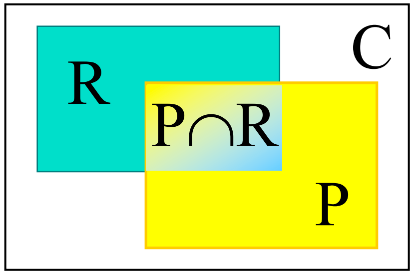
- The recall is the ratio of
  - The number of retrieved documents by the system and relevant to the user
  - Divided by the number of all the documents of the corpus that are relevant to the user



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

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

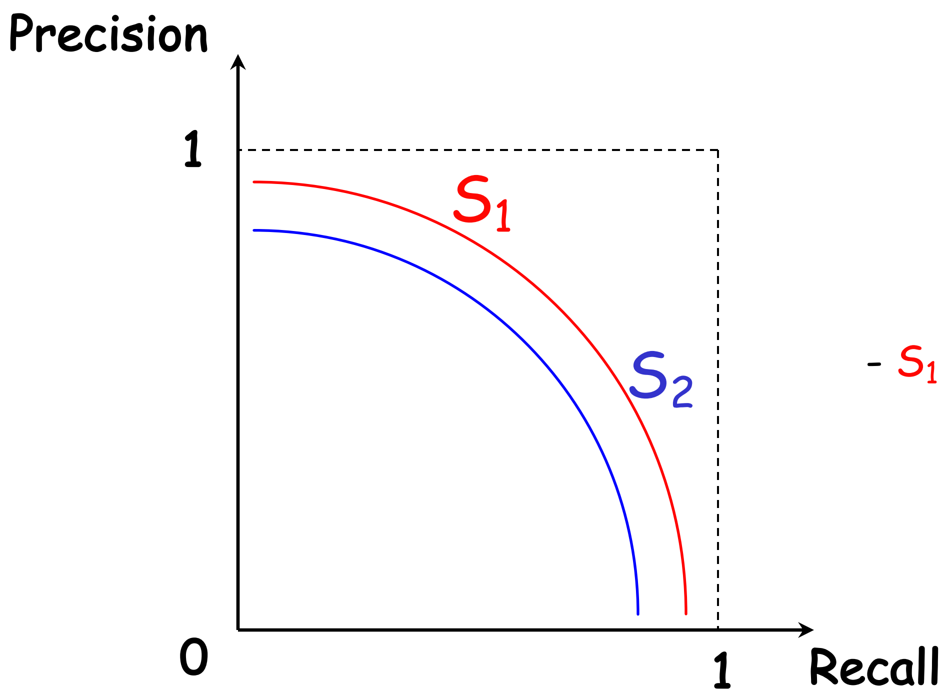
- For one query and one system : 2 real values
  - Example: a system gives 5 documents, among them 3 are relevant, knowing that there are 10 relevant documents in the corpus:
    - Recall = 3 / 10 = 0.3
    - Precision = 3 / 5 = 0.6
- BUT, no use of rankings:

pos	S1 result	S2 result
1	d235	d5
2	d56	d12
3	d786	d235
4	d451	d976
5	d67	d376

  - same recall (0.3) and same precision (0.6) values for S1 and S2, but S1 is "better"
- We need more detailed evaluations
  - Recall/precision diagrams

3. Recall/precision diagrams

- Comparison of 2 systems S1 et S2



- S_1 is better than S_2

3. Recall/precision diagrams

- 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 result as answer, then we do the same for the two first results of the system, the first three results, and so on, until each retrieved document is processed.

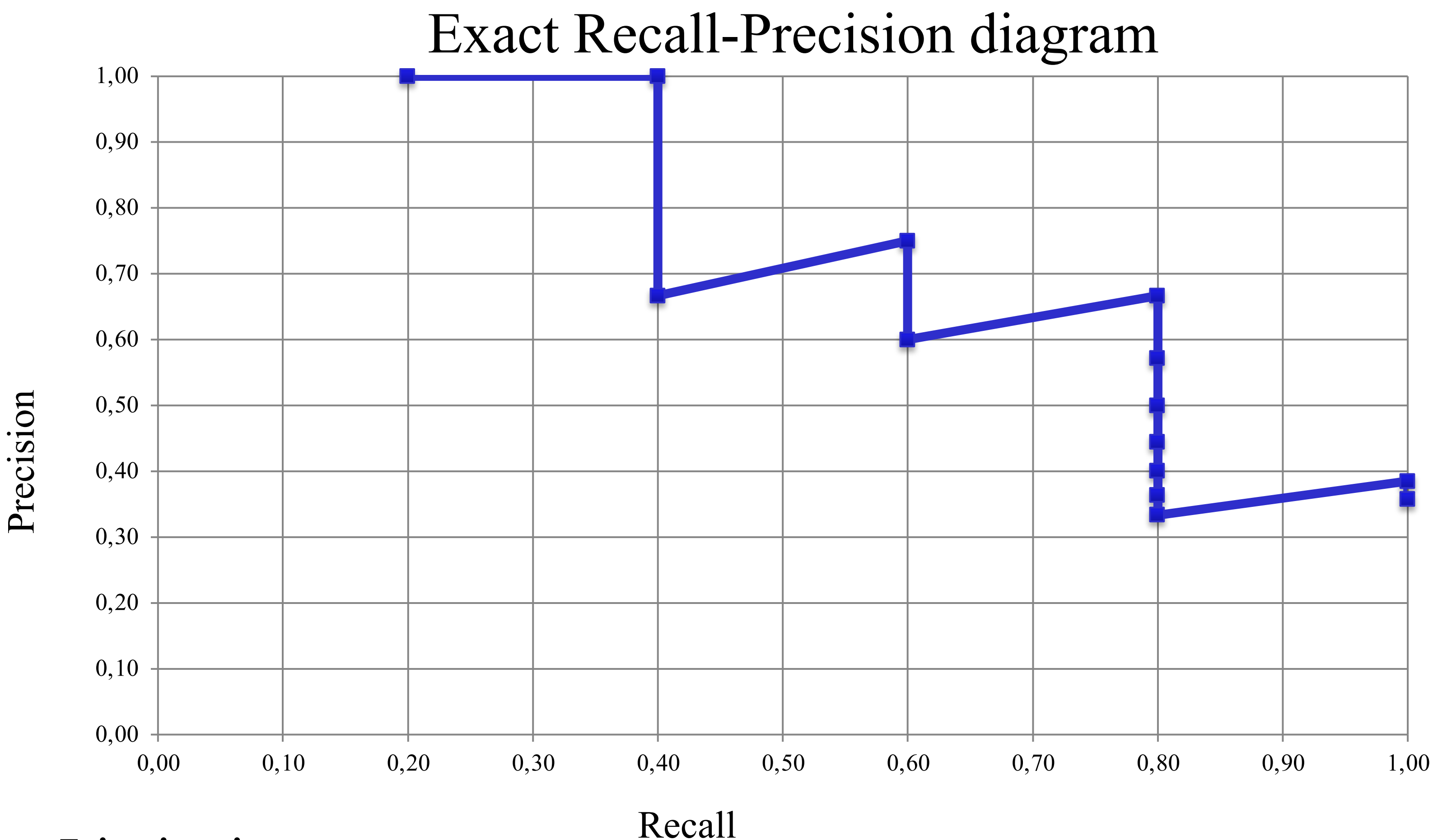
3. Recall/precision diagrams

- Corpus of 200 documents, one query Q that have 5 relevant docs {572, 588, 589, 590, 592}

Exact Recall-Precision table

			recall	precision
position	document	is relevant	p and r / p	p and r / r
1	588	X	0,20	1,00
2	589	X	0,40	1,00
3	576		0,40	0,67
4	590	X	0,60	0,75
5	986		0,60	0,60
6	592	X	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	X	1,00	0,38
14	990		1,00	0,36

3. Recall/precision diagrams



- Limitation
  Difficult to fuse exact R-P curves for several queries,
  => problem of merging the recall values

3. Recall/precision diagrams

- Solution: Interpolated Recall/Precision diagrams
  - Fix 11 recall points R={0, 0.1, 0.2, …, 0.9, 1}
  - Rule of the maximum

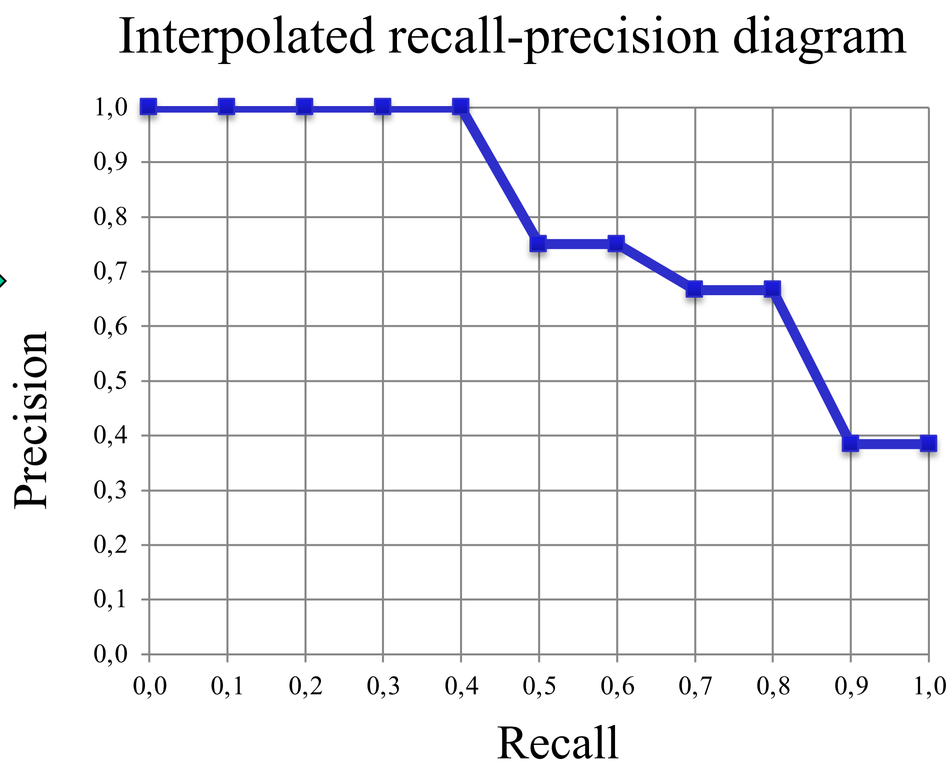
    for each recall point v_r in R, select in the exact table the lines with recall greater or equal than v_r and pick the max of precision in these lines

    -> classically, begin with v_r=0, then 0.1, then 0.2, …, then 1.0
  - Example from the exact table of slide 14
    - With v_r= 0.6, the max precision = 0.75 (from 4th result)
  - When, for a recall point, there is no precision value in the exact table according to the rule of the maximum, then we force its interpolated precision to 0 (i.e., the theoretical minimal precision value).
    - In the table of slide 14 ends in line 12, the precision is 0 for v_r=1

3. Recall/precision diagrams

Interpolated recall/precision table

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

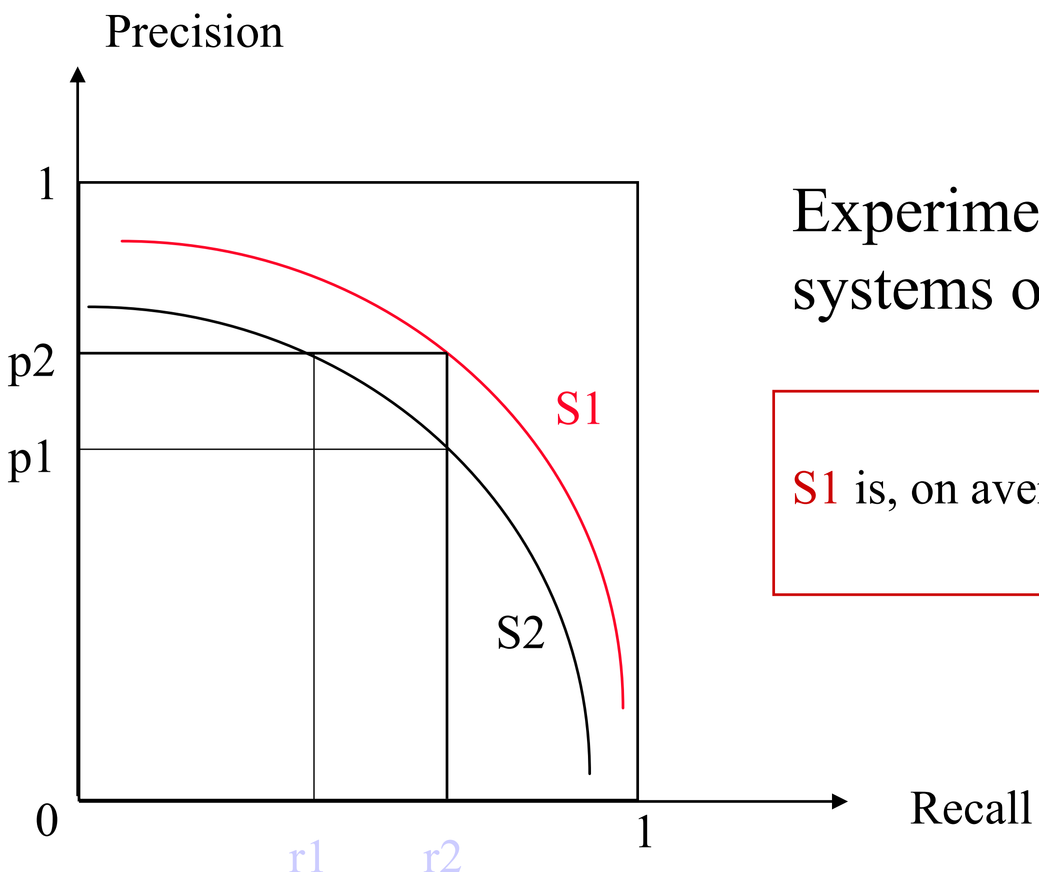


3. Recall/precision diagrams

- A full evaluation considers many queries
- For nbQ queries > 1:
    1. Generate interpolated table for each query
    2. Average on each of the 11 recall points for all the nbQ queries
    3. Generate the overall recall/precision table + diagram of a system.

3. Recall/precision diagrams

- Comparing systems

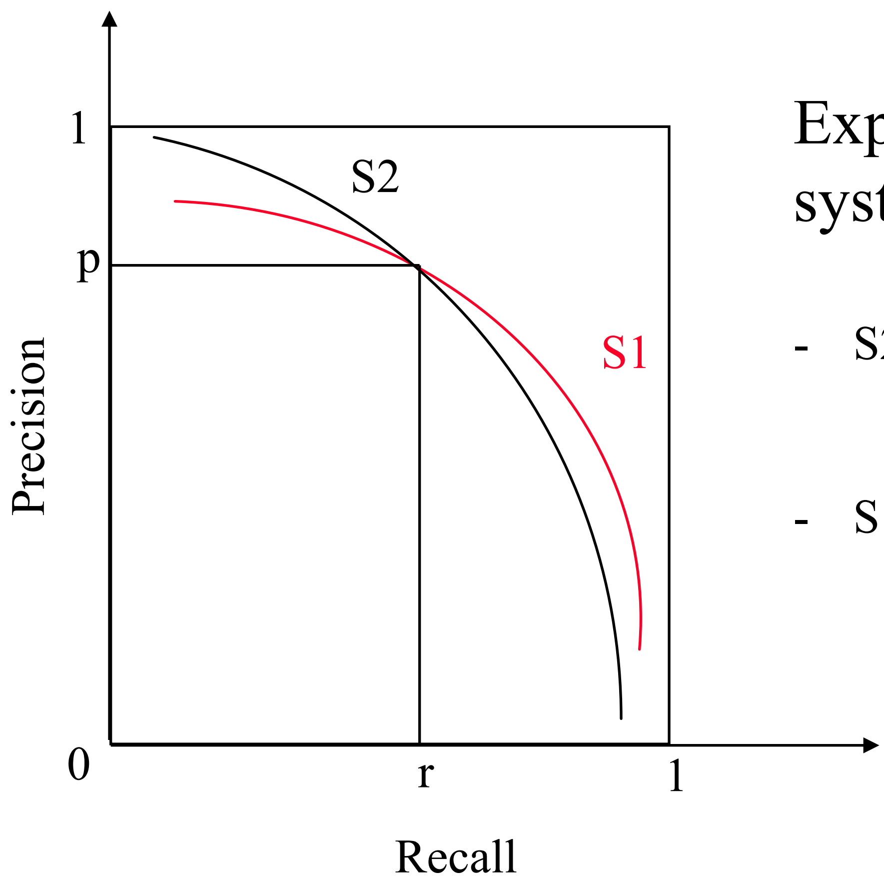


Experimental comparison of systems on a test collection :

S1 is, on average, always better than S2

3. Recall/precision diagrams

- Comparing systems

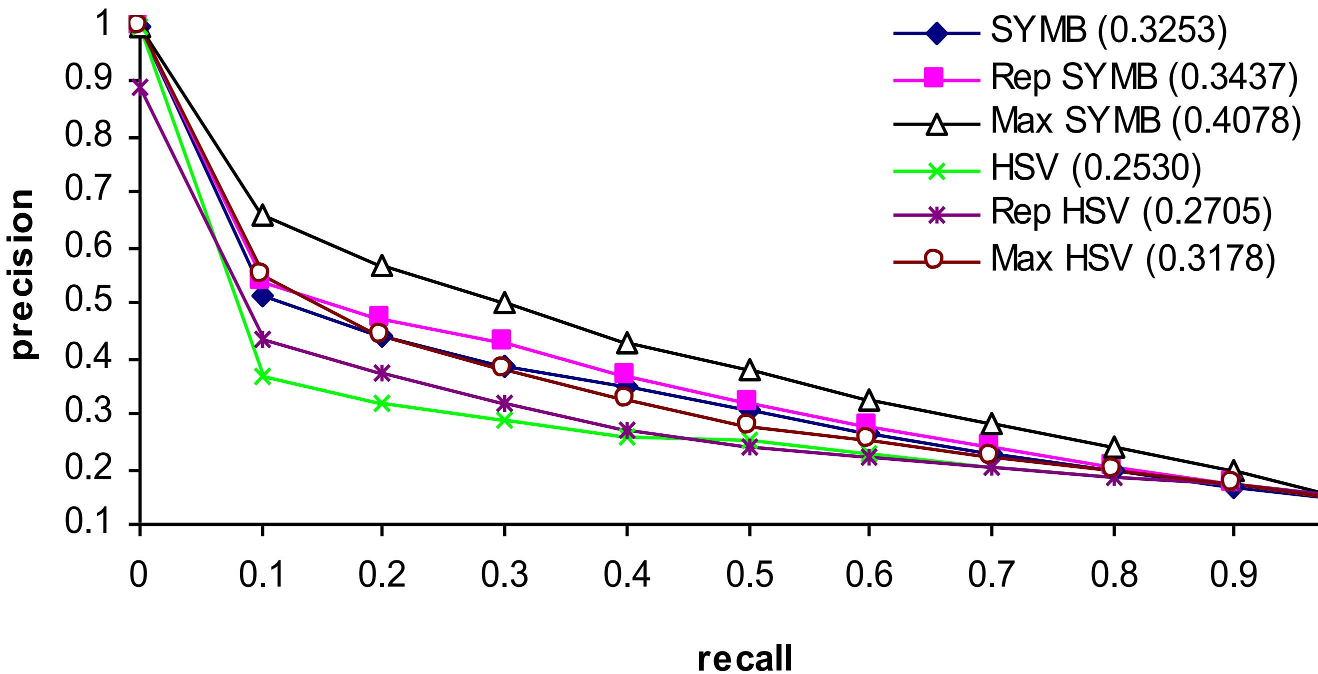


Experimental comparison of systems on a test collection :

- S2 is better than S1 for precision
  - Web search
- S1 is better than S2 for recall
  - Side effects of medicine drugs

3. Recall/precision diagrams

– A real diagram



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 <u>for one query</u>
    - precision computed after each relevant document (from the exact table), averaged

$$AP = \frac{\sum_{k=1}^{n} \text{Prec(k) . rel(k)}}{|P|}$$

      - P : set of relevant documents, Prec(k) precision value at result k,

$$rel(k) = \begin{cases} \text{1 if document at position k is relevant} \\ \text{0 otherwise} \end{cases}$$

      - on the previous example: AP=0.76 (from table slide 14)
  - MAP mean of the average precision over all query

5. F-measure

- Integrates recall and precision in one value (harmonic mean)
- General form : $F_\beta = \frac{(1+\beta^2).precision.recall}{\beta^2.precision + recall}$

- In IR: $\beta = 1$

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

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 (table slide 14):
    - P@5 = 0.60
    - P@10=0.40
    - P@15=0.33

7. Normalized Discounted 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>

      1 2 3 4 5 6
  - Obtain the gain value for each document:

    G[j]=gain(R[j])

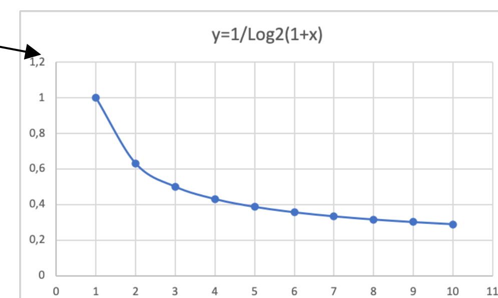
    - Ex : G= <1, 2, 0, 0, 2, 1>

Simple user model: more chances to look at the first doc, then less chances to look at the second results, …)

  - Compute the discounted gain for each document:

    DG[j]=gain(R[j])/$\log_2$(j+1))

    - Ex : DG = < 1, 1.26, 0, 0, 0.77, 0.36 >



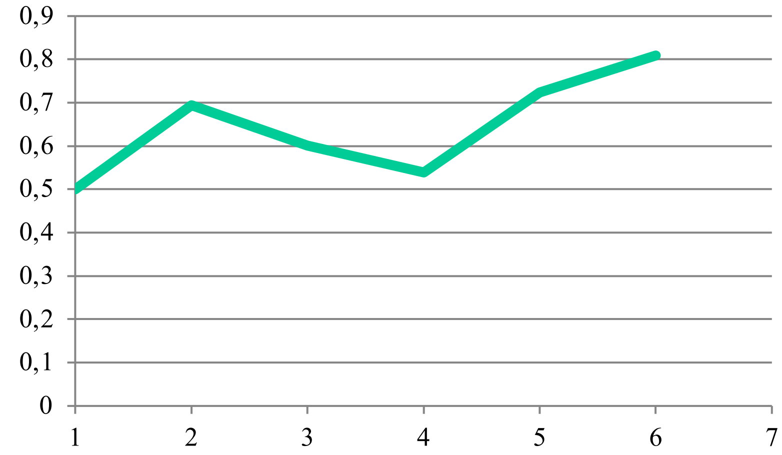
  - Compute cumulated gain at rank i: $$DCG[i]=\sum_{j=1}^{i} DG[j]$$
    - Ex : DCG=<1, 2.26, 2.26, 2.26, 3.04, 3.39>

7. Normalized Discounted Cumulated Gain

- Normalization by using an ideal list I, list of the gains of the relevant documents of R sorted by decreasing gain value (ex. 4 docs with relevance of 2, 2, 1, 1)
  - Ex : I=<2, 2, 1, 1, 0, 0>
- Discounted gain for the ideal list between the position 1 and i :
  - Ex : DCI=<2 3,26 3,76 4,19 4,19 4,19>
- Normalized Cumulated Gain : $nDCG[i] = \frac{DCG[i]}{DCI[i]}$
  - Ex : nDCG=<0,5 0,69 0,60 0,54 0,72 0,81>

7. Normalized Discounted Cumulated Gain

- Curve (on our example):



  – Difficult to read…

- More readable values using nDCG
  – nDCG@x: value AT x documents retrieved
    - On our example: nDCG@5=0.72

7. Normalized Discounted Cumulated Gain

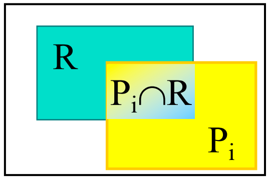
- Cumulated gain compares an ideal result list to the result obtained
- Uses importance in rank (top more important):
  - Simulate user behaviour
- Takes into account non binary values of relevance
  + this is good !
  – difficult to interpret curves results
  + use nDCG@x

8. Test collections

- Recall/precision/nDCG need <u>test collections</u>
- A test collection = a set of resolved queries q_i on a corpus
  - A large fixed corpus C of documents (> 100K)
  - queries representative of real user interests
    - diverse queries (subject, style, vocabulary)
    - large number (> 30)
  - Relevance assessments (which doc of C is relevant/not relevant for which query)
- Hard to assess manually the queries on the full corpus

→ Assess only on *pooled* results [Voorhes 2001]

- we run the queries q_i on several state of the art systems S_j, each system gets a result list per query $P_{i,j}$
- we make a union of each results *sets* per query : $P_i = \bigcup_j P_{i,j}$
- we evaluate user relevance on the P_i (=> not all the collection)



=> the non-assessed documents are considered non-relevant

8. Test collections

- Use of pooling
  - 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

  - If you test a system S *a posteriori* (i.e., it is not used in the pool) it <u>may be underevaluated</u>
    - S may retrieve relevant results that were not considered in the pool, so marked as non-relevant…

8. Test collections

- Evaluation measures adapted to pooling
  - Use condensed lists: evaluation of results considering ONLY the judged documents (different from slide 29)
    - Example, if a system gives $R = <d_{23}, d_{56}, d_9, d_{135}, d_{87}, d_4>$ and if d_{135} is not in the pool, then it's like $R' = <d_{23}, d_{56}, d_9, \cancel{d_{135}}, d_{87}, d_4>$, and we compute evaluations measures on R'.

  - P'@5 : P@x using only *judged* documents
  - nDCG'@5 : nDCG@5 using only *judged* documents
  - P' (resp. nDCG') has the same role than P (resp. nDCG)

9. Trec-eval

- Software that generates the tables for the recall/precision diagrams and AP, P@5, 10, 20, 50 and 100 documents, and nDCG, 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 using overlapping of characters).
  - Discounted Cumulated Gain used in eval
  - On large collections, difficult to make evaluations
    - One solution (TREC) pool the results for several systems
    - Evaluation measures take into account the pooling
  - Transfert such evaluation using log-files (Web search) ?
  - Transfert such evaluation from one test collection to another one ?

10. Conclusion

- To do
  - Understand classical IR evaluation (Cranfield Paradigm)
  - Understand recall/precision measures and diagrams (redo the example, and make others removing one relevant 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.