COMP4650/6490 Document Analysis

Evaluation of IR Systems

ANU School of Computing

Administrative Matters

- Class Representatives
  - Contact details are available on Wattle
  - Please send them your feedback
  - Thank you to all candidates
- Labs
  - Lab1 solution released
  - Lab2 will be made available later today
- Quiz 1
  - Open: 11am Monday 7 August
  - Due: 5pm Friday 11 August
  - First 3 lectures in IR section are assessed
  - Marks and answers will be released after due date

So far…

We looked at:

- Boolean retrieval (unranked results)
- Ranked retrieval

Outline

- Purpose of evaluation
  - Why do we need evaluation
  - What do we want to evaluate
- Test collection
  - Three components of test collections
  - Standard test collections
  - Build large test collection
- Evaluation of unranked retrieval sets
  - Precision, recall, accuracy and F-measure
- Evaluation of ranked retrieval results
  - Precision-recall curve, interpolated precision
  - Single number metrics: MAP and MRR

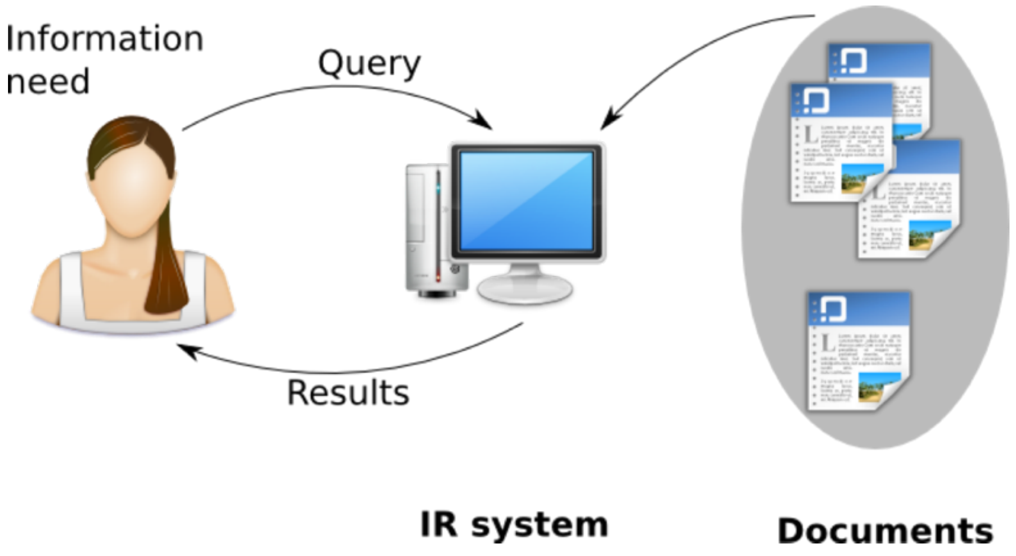
Outline

- Purpose of evaluation
  - Why do we need evaluation
  - What do we want to evaluate
- Test collection
  - Three components of test collections
  - Standard test collections
  - Build large test collection
- Evaluation of unranked retrieval sets
  - Precision, recall, accuracy and F-measure
- Evaluation of ranked retrieval results
  - Precision-recall curve, interpolated precision
  - Single number metrics: MAP and MRR

Purpose of Eval.

Why do we need evaluation?

- To build IR systems that satisfy user's information needs



- Given multiple candidate systems, which one is the best?

Purpose of Eval.

What do we want to evaluate?

- System efficiency
  - Speed, storage, memory, cost
- System effectiveness
  - Quality of search results
  - Does it find what I'm looking for?
  - Does it return lots of junk?

We will focus on evaluating system **effectiveness**.

Purpose of Eval.

Improve System Effectiveness

IR system design choices

- Which tokeniser? Which stemmer? Lemmatisation? Remove stop words?
- Which scoring method?
- tf-idf or wf-idf?
- Length normalisation or not?

What are the best choices?

Outline

- Purpose of evaluation
  - Why do we need evaluation
  - What do we want to evaluate
- Test collection
  - Three components of test collections
  - Standard test collections
  - Build large test collection
- Evaluation of unranked retrieval sets
  - Precision, recall, accuracy and F-measure
- Evaluation of ranked retrieval results
  - Precision-recall curve, interpolated precision
  - Single number metrics: MAP and MRR

Test Collection

- A *test collection* is a collection of relevance judgment on (query, document) pairs
- Example

Query 1

- Doc 1: **relevant**
- Doc 2: irrelevant
- Doc 3: irrelevant
- Doc 4: **relevant**
- Doc 5: irrelevant

Query 2

- Doc 1: irrelevant
- Doc 2: irrelevant
- Doc 3: **relevant**
- Doc 4: irrelevant
- Doc 5: **relevant**

- This relevancy information is known as the *ground truth*
- It is typically constructed by trained human annotators

Test Collection

Three Components of Test Collections

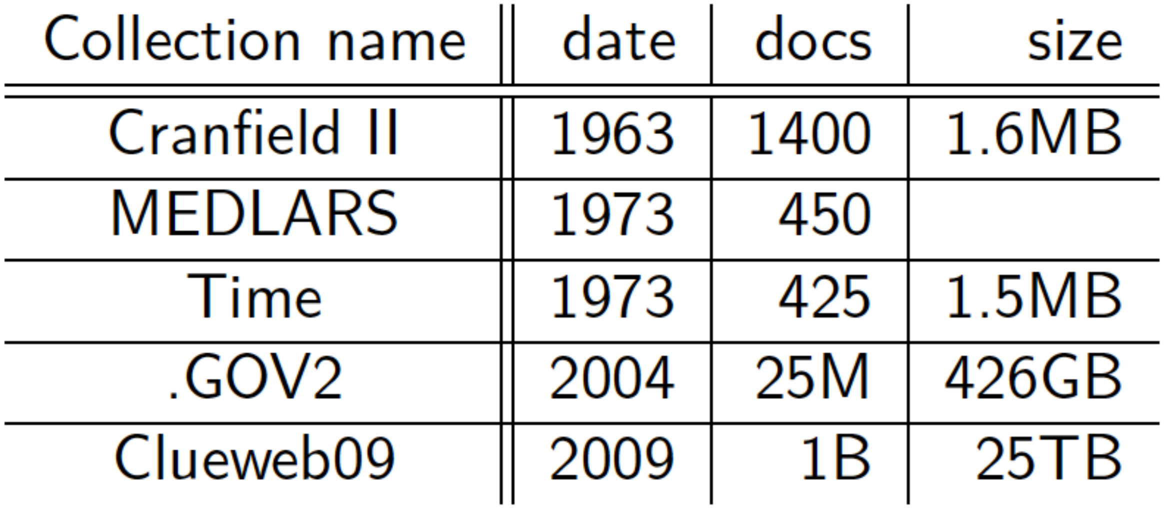
1. A collection of documents
2. A test suite of information needs
   - Expressible as queries
3. A set of relevance judgments
   - Usually binary assessment of either *relevant* or *irrelevant* for each (query, document) pair

Test Collection

Relevance Judgment

- Relevance is assessed relative to an *information need*, not a query
- Example
  - If our information need is:
    Information on whether drinking red wine is more effective at reducing your risk of heart attacks than white wine
  - Candidate query: *wine red white heart attack effective*
- A document is relevant if it addresses the information need
- The document does not need to contain all/any of the query terms

Standard Test Collections



Collection name	date	docs	size
Cranfield II	1963	1400	1.6MB
MEDLARS	1973	450	
Time	1973	425	1.5MB
.GOV2	2004	25M	426GB
Clueweb09	2009	1B	25TB

Table: Examples of test collections

And now also blogs, Twitter, legal, patents, chemical, genomic, ...

Test Collection

Build Large Test Collection

- Time-consuming and expensive process
- Recent collections do not have relevance judgments on all possible (query, document) pairs
- Assess relevance only for a subset of documents for each query
  - The pooling approach: given a query, try multiple IR systems and obtain a set of candidate documents
- Multiple judges assess the relevance of a candidate document given information need

Outline

- Purpose of evaluation
  - Why do we need evaluation
  - What do we want to evaluate
- Test collection
  - Three components of test collections
  - Standard test collections
  - Build large test collection
- Evaluation of unranked retrieval sets
  - Precision, recall, accuracy and F-measure
- Evaluation of ranked retrieval results
  - Precision-recall curve, interpolated precision
  - Single number metrics: MAP and MRR

Evaluation of Retrieval Results

Two Evaluation Settings

- Evaluation of *unranked* retrieval sets (Boolean retrieval)
  - Ranks of retrieved documents are not important
  - Retrieved documents vs. Not retrieved documents
- Evaluation of *ranked* retrieval results
  - Rank of retrieved documents are important
  - Relevant documents should be ranked above irrelevant documents

Unranked Eval.

Evaluation of Unranked Retrieval Sets

Example: Suppose we have 10 documents, and the system returns 4 documents for the query

Retrieved (Returned) docs:

- Doc 2: **relevant** (ground truth)
- Doc 4: irrelevant
- Doc 5: irrelevant
- Doc 7: **relevant**

Not retrieved docs:

- Doc 1: irrelevant
- Doc 3: irrelevant
- Doc 6: irrelevant
- Doc 8: **relevant**
- Doc 9: irrelevant
- Doc 10: irrelevant

How can we evaluate the performance of this system?

Contingency Table

- Contingency table is a summary table of retrieval results

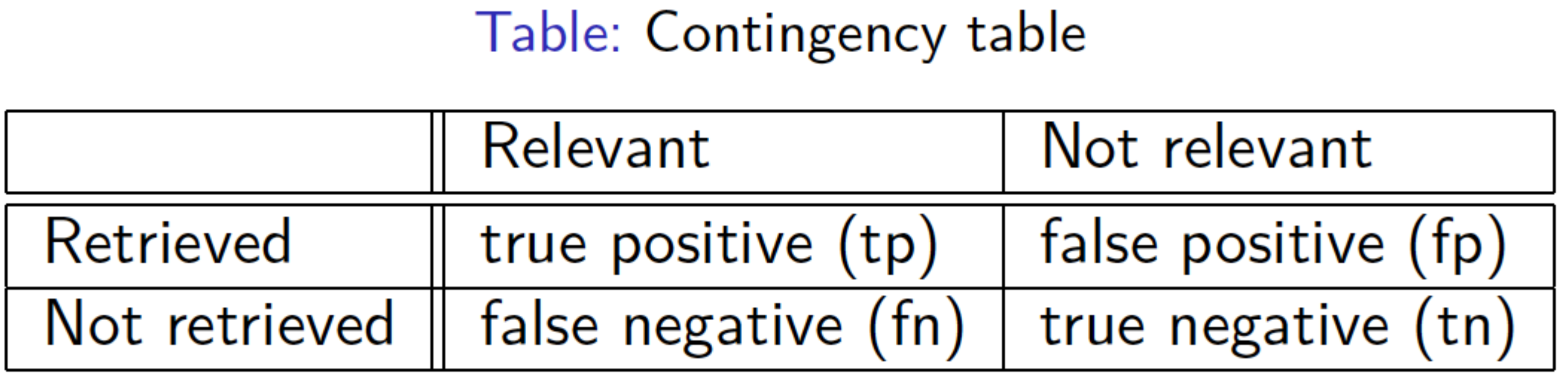


Table: Contingency table

	Relevant	Not relevant
Retrieved	true positive (tp)	false positive (fp)
Not retrieved	false negative (fn)	true negative (tn)

- tp: Number of **relevant** documents returned by system
- fp: Number of irrelevant documents returned by system
- fn: Number of **relevant** documents NOT returned by system
- tn: Number of irrelevant documents NOT returned by system

Unranked Eval.

Precision and Recall

Precision: fraction of retrieved documents that are relevant

$$Precision = \frac{\#\text{of relevant docs retrieved}}{\#\text{of retrieved docs}} = \frac{tp}{tp + fp}$$

Recall: fraction of relevant documents that are retrieved

$$Recall = \frac{\#\text{of relevant docs retrieved}}{\#\text{of relevant docs}} = \frac{tp}{tp + fn}$$

Related measures
https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Precision and Recall

Example

Table: 10 document example

	Rel.	Not rel.
Retrieved	2	2
Not retrieved	1	5

- Precision $= \frac{2}{2+2} = 0.50$
- Recall $= \frac{2}{2+1} = 0.66$
- System with high precision and recall is always preferable.

Accuracy

- Accuracy: fraction of relevant documents that are correct

$$Accuracy = \frac{\#\text{of correctly classified docs}}{\#\text{of total docs}} = \frac{tp + tn}{tp + tn + fp + fn}$$

- Accuracy is NOT appropriate for evaluating IR systems
  - Assume we have 100 documents, and only 1 document is relevant given a certain query

Table: System 1

	Rel.	Not rel.
Retrieved	0	0
Not retrieved	1	99

Table: System 2

	Rel.	Not rel.
Retrieved	1	4
Not retrieved	0	95

  - Accuracy: 0.99 (System 1), 0.96 (System 2)
  - System 1 performs better in terms of accuracy but retrieved no relevant documents

Unranked Eval.

F-Measure

F-Measure is the weighted harmonic mean of precision (P) and recall (R)

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}, \qquad \alpha \in [0, 1]$$

- A single measure that trades off precision and recall
- $\alpha > 0.5$, emphasises precision, e.g. $F = P$ if $\alpha = 1$
- $\alpha < 0.5$, emphasises recall, e.g. $F = R$ if $\alpha = 0$
- $\alpha = 0.5, F_1 = \frac{2PR}{P + R}$

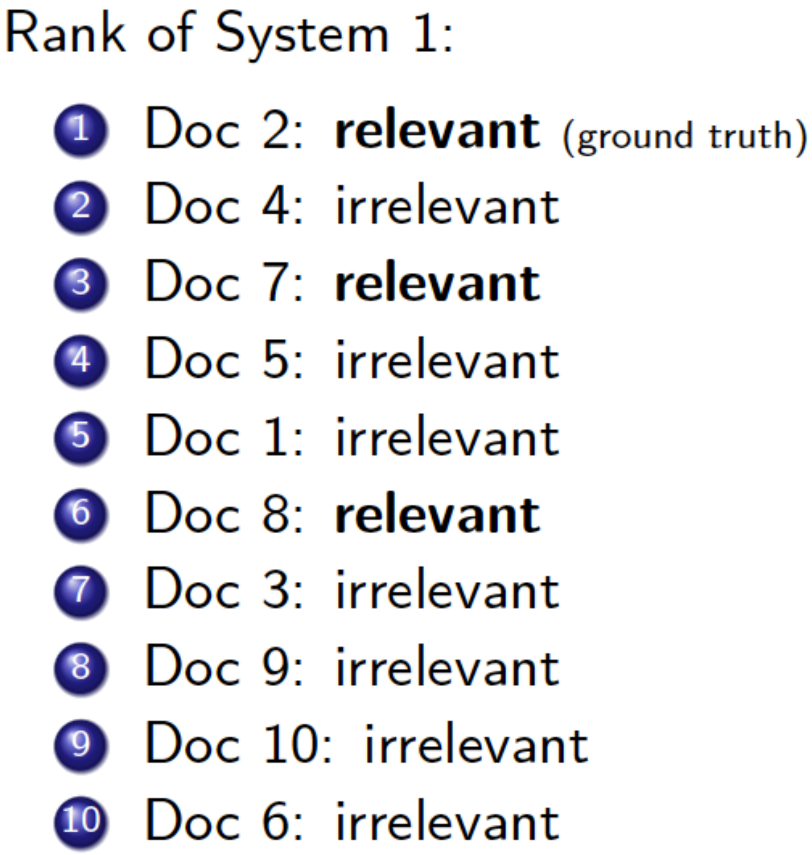
Outline

- Purpose of evaluation
  - Why do we need evaluation
  - What do we want to evaluate
- Test collection
  - Three components of test collections
  - Standard test collections
  - Build large test collection
- Evaluation of unranked retrieval sets
  - Precision, recall, accuracy and F-measure
- Evaluation of ranked retrieval results
  - Precision-recall curve, interpolated precision
  - Single number metrics: MAP and MRR

Evaluation of Ranked Retrieval Results

Example

- Given a query, an IR system retrieved all 10 documents in our collection, and generates ranked results
- Precision & Recall cannot be directly applied in this case
- Need a metric to measure the performance of ranked list!



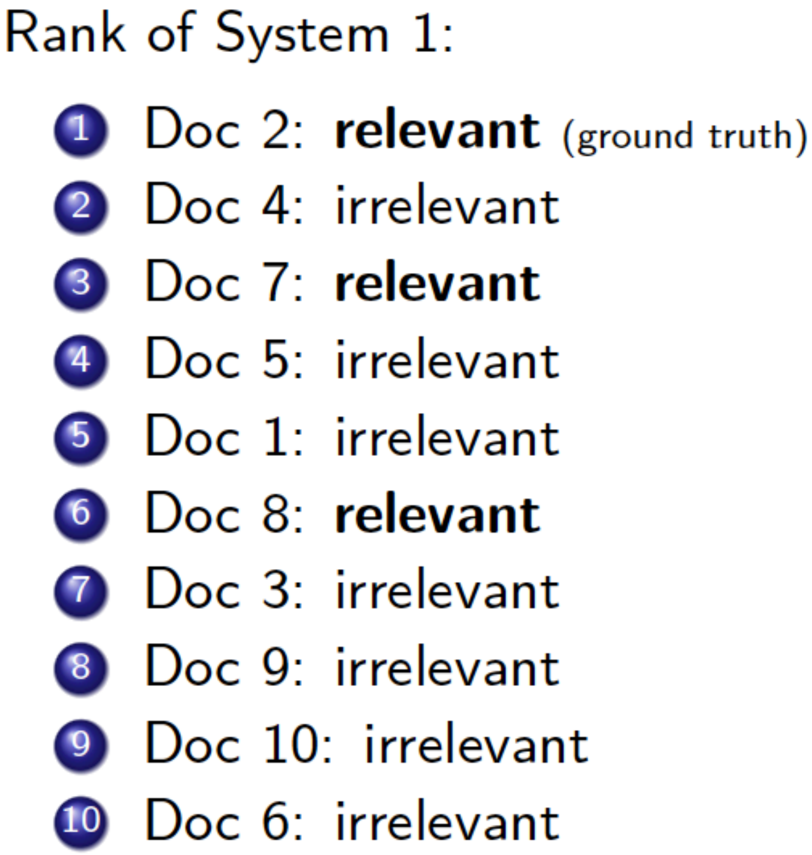
Rank of System 1:

1. Doc 2: **relevant** (ground truth)
2. Doc 4: irrelevant
3. Doc 7: **relevant**
4. Doc 5: irrelevant
5. Doc 1: irrelevant
6. Doc 8: **relevant**
7. Doc 3: irrelevant
8. Doc 9: irrelevant
9. Doc 10: irrelevant
10. Doc 6: irrelevant

Evaluation of Ranked Retrieval Results

Example

- Given a query, an IR system retrieved all 10 documents in our collection, and generates ranked results
- Precision & Recall cannot be directly applied in this case
- Need a metric to measure the performance of ranked list!



Rank of System 1:

1. Doc 2: **relevant** (ground truth)
2. Doc 4: irrelevant
3. Doc 7: **relevant**
4. Doc 5: irrelevant
5. Doc 1: irrelevant
6. Doc 8: **relevant**
7. Doc 3: irrelevant
8. Doc 9: irrelevant
9. Doc 10: irrelevant
10. Doc 6: irrelevant

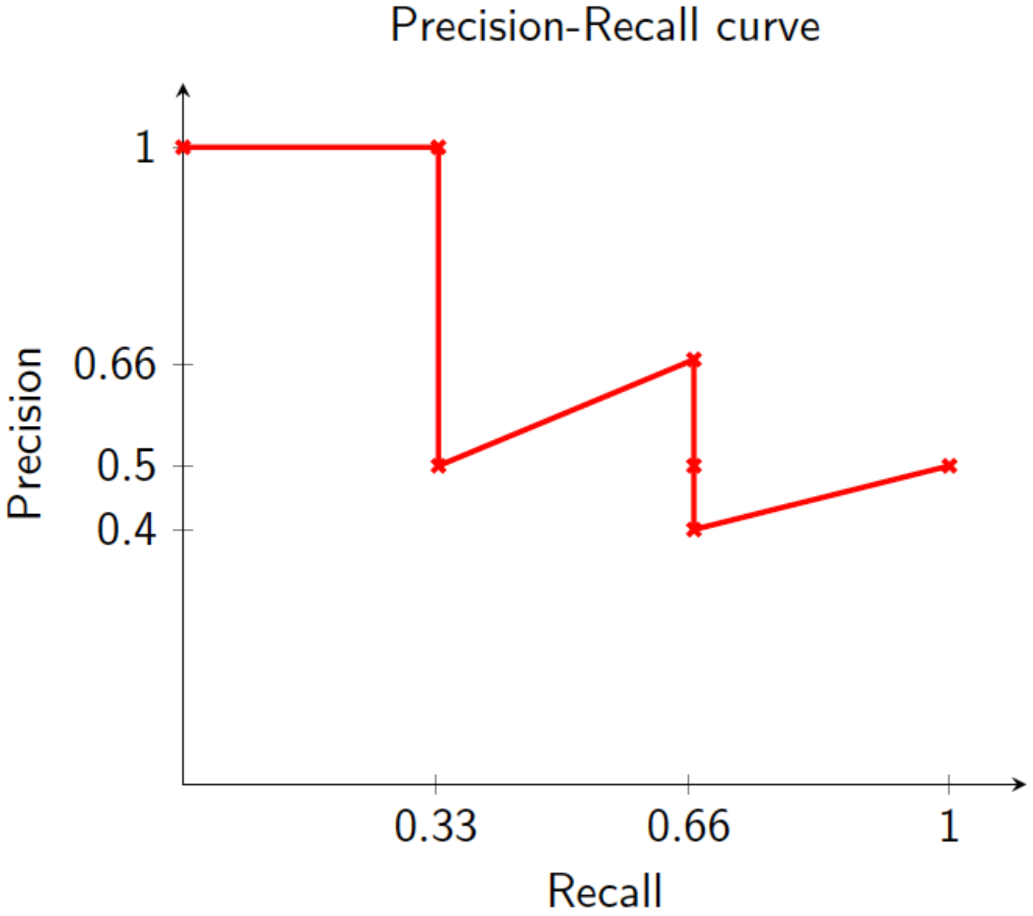
How can we quantify the performance of this result?

Ranked Eval.

Precision-Recall Curve

Precision and recall of the top-*k* retrieved documents

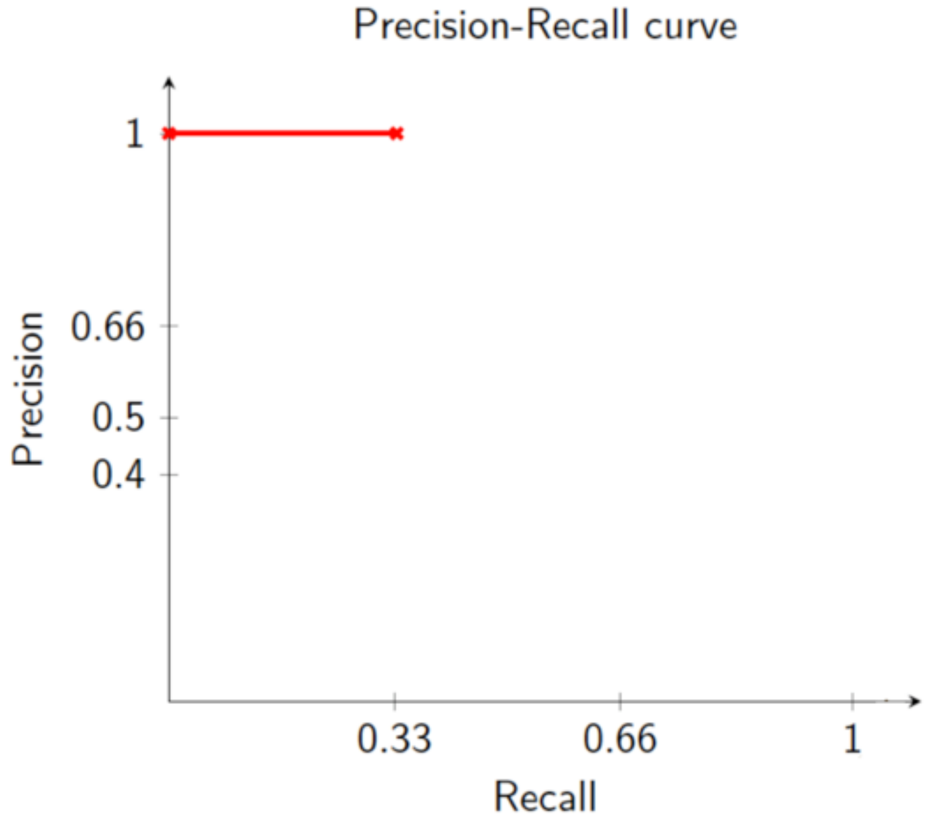
1. Doc 2: **relevant**
2. Doc 4: irrelevant
3. Doc 7: **relevant**
4. Doc 5: irrelevant
5. Doc 1: irrelevant
6. Doc 8: **relevant**
7. Doc 3: irrelevant
8. Doc 9: irrelevant
9. Doc 10: irrelevant
10. Doc 6: irrelevant



Precision-Recall Curve

Compute recall and precision at each rank *k* using the top-*k* retrieved documents, and plot the (recall, precision) points until recall is 1

1. Doc 2: **relevant** (1/3, 1)

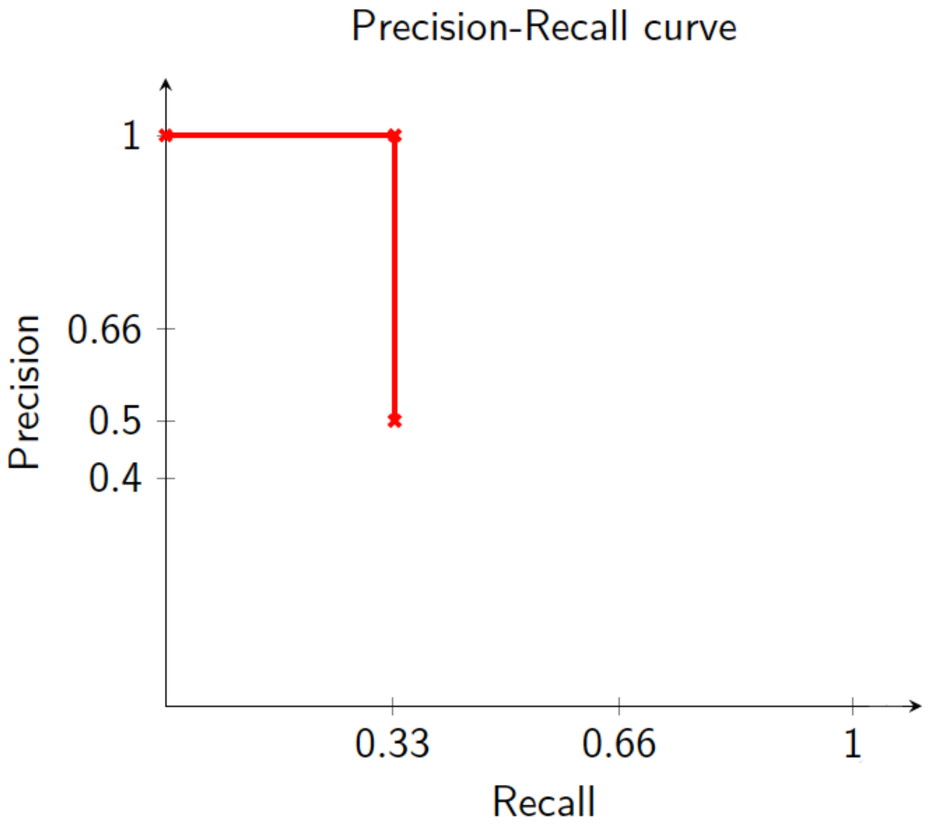


Ranked Eval.

Precision-Recall Curve

Compute recall and precision at each rank *k* using the top-*k* retrieved documents, and plot the (recall, precision) points until recall is 1

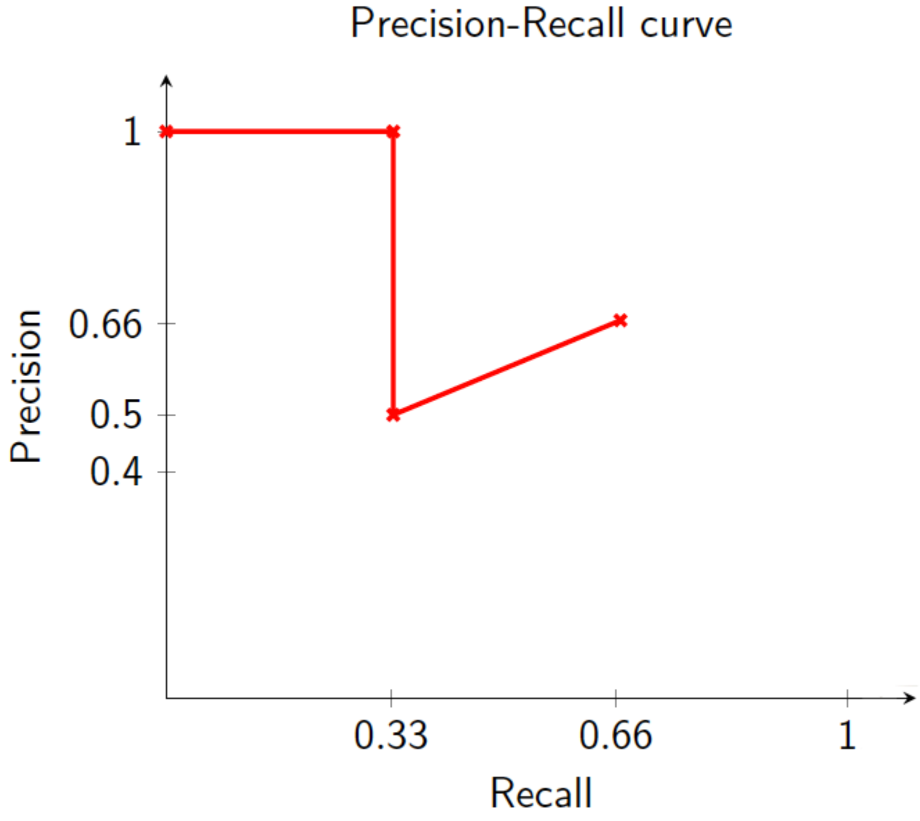
1. Doc 2: **relevant** (1/3, 1)
2. Doc 4: irrelevant (1/3, 1/2)



Precision-Recall Curve

Compute recall and precision at each rank *k* using the top-*k* retrieved documents, and plot the (recall, precision) points until recall is 1

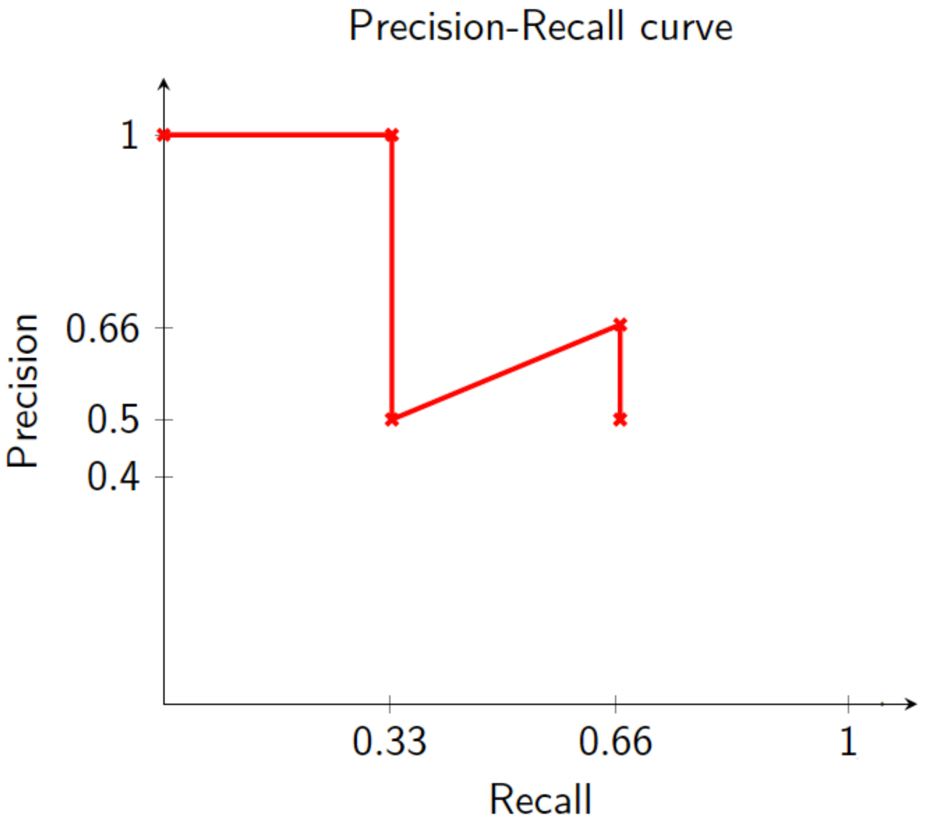
1. Doc 2: **relevant** (1/3, 1)
2. Doc 4: irrelevant (1/3, 1/2)
3. Doc 7: **relevant** (2/3, 2/3)



Precision-Recall Curve

Compute recall and precision at each rank *k* using the top-*k* retrieved documents, and plot the (recall, precision) points until recall is 1

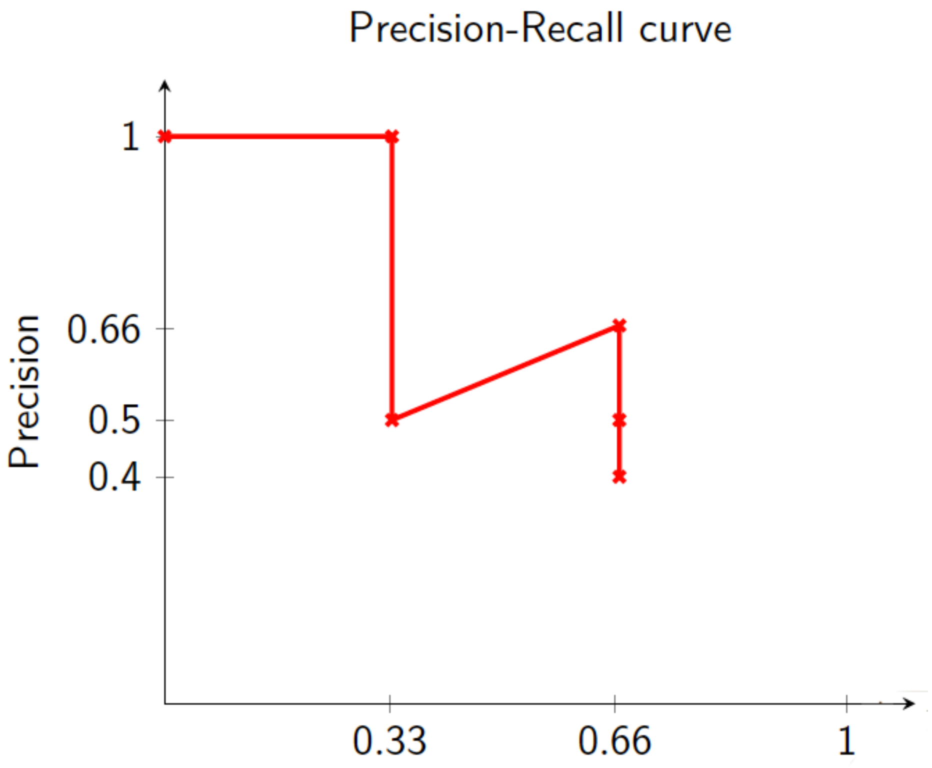
1. Doc 2: **relevant** (1/3, 1)
2. Doc 4: irrelevant (1/3, 1/2)
3. Doc 7: **relevant** (2/3, 2/3)
4. Doc 5: irrelevant (2/3, 2/4)



Precision-Recall Curve

Compute recall and precision at each rank *k* using the top-*k* retrieved documents, and plot the (recall, precision) points until recall is 1

1. Doc 2: **relevant** (1/3, 1)
2. Doc 4: irrelevant (1/3, 1/2)
3. Doc 7: **relevant** (2/3, 2/3)
4. Doc 5: irrelevant (2/3, 2/4)
5. Doc 1: irrelevant (2/3, 2/5)

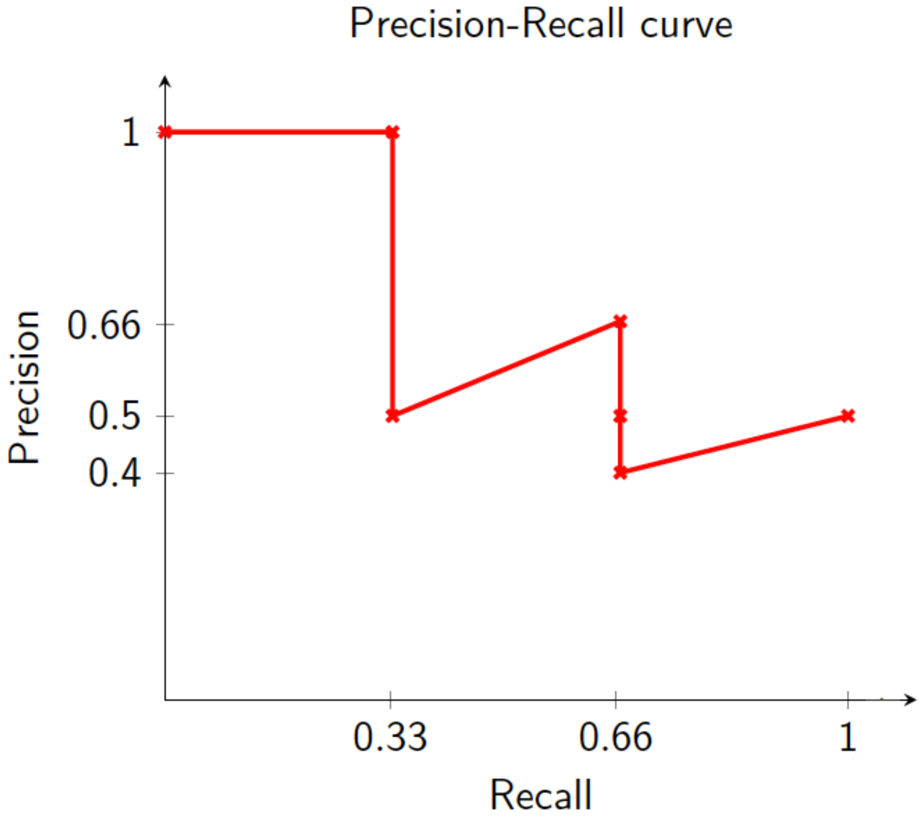


Ranked Eval.

Precision-Recall Curve

Compute recall and precision at each rank *k* using the top-*k* retrieved documents, and plot the (recall, precision) points until recall is 1

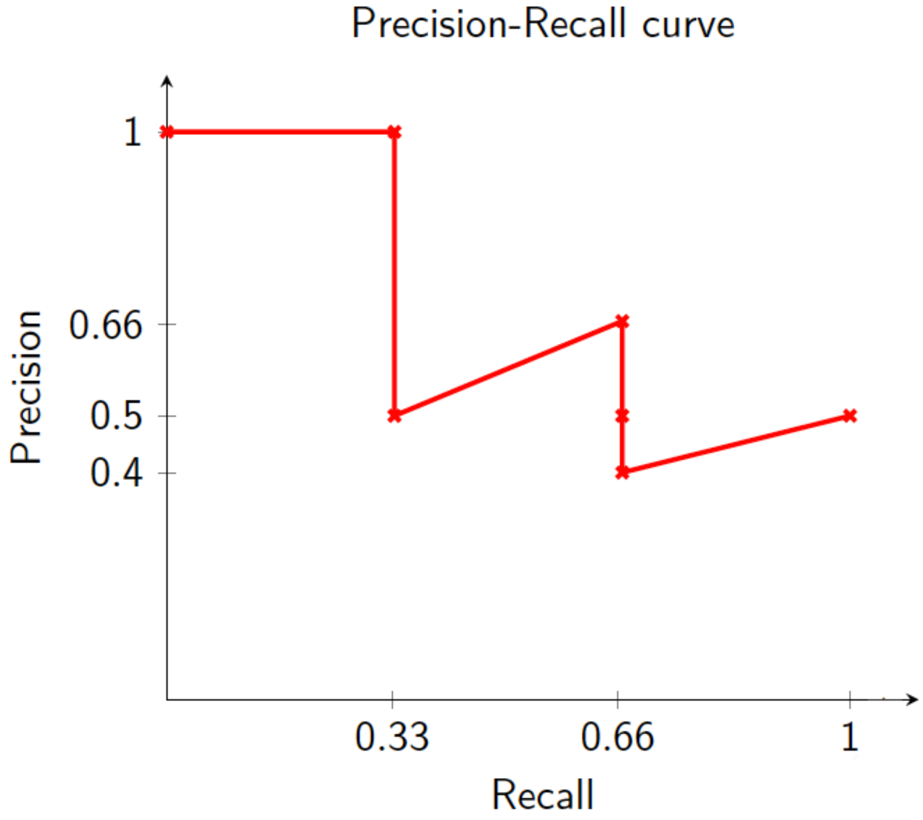
1. Doc 2: **relevant** (1/3, 1)
2. Doc 4: irrelevant (1/3, 1/2)
3. Doc 7: **relevant** (2/3, 2/3)
4. Doc 5: irrelevant (2/3, 2/4)
5. Doc 1: irrelevant (2/3, 2/5)
6. Doc 8: **relevant** (1, 3/6)



Precision-Recall Curve

Compute recall and precision at each rank *k* using the top-*k* retrieved documents, and plot the (recall, precision) points until recall is 1

1. Doc 2: **relevant** (1/3, 1)
2. Doc 4: irrelevant (1/3, 1/2)
3. Doc 7: **relevant** (2/3, 2/3)
4. Doc 5: irrelevant (2/3, 2/4)
5. Doc 1: irrelevant (2/3, 2/5)
6. Doc 8: **relevant** (1, 3/6)
7. Doc 3: irrelevant
8. Doc 9: irrelevant
9. Doc 10: irrelevant
10. Doc 6: irrelevant



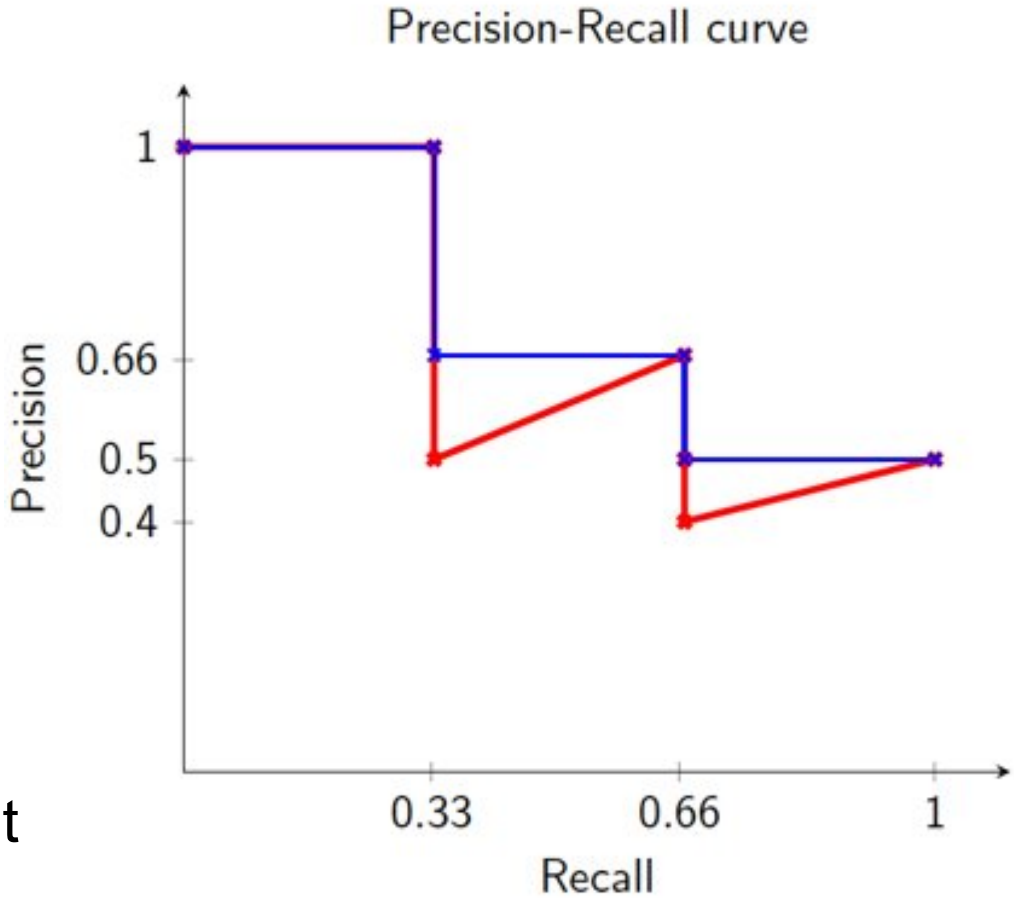
Ranked Eval.

Interpolated Precision

- At a given recall level use the maximum precision at all higher recall levels

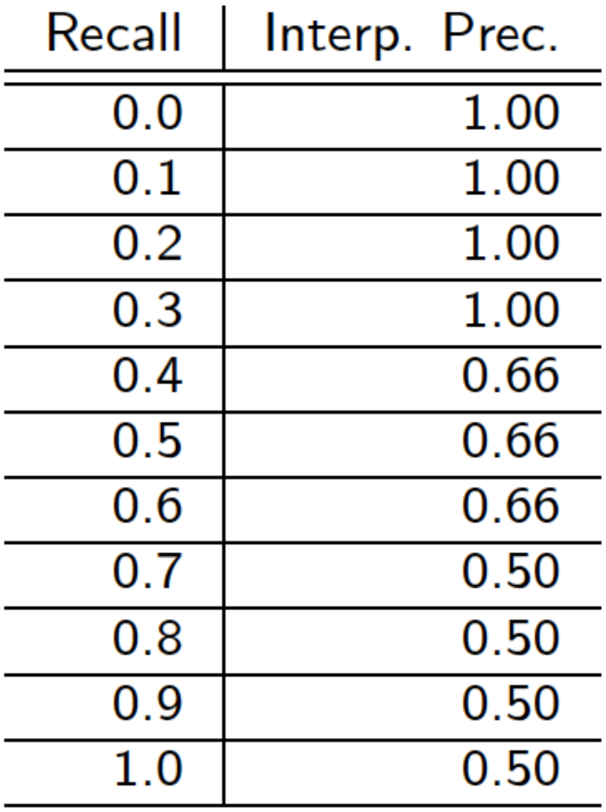
$$p_{\text{interp}}(r) = \max_{r' \geq r} p(r')$$

- Intuition:
  there’s no disadvantage to retrieving more documents if both precision and recall improve
- Makes it easier to interpret



Interpolated Precision

- For system evaluation we need to *average across many queries*
- It is not easy to average a PR curve in its current form
- Solution: 11-point interpolated average precision
- Each point in the 11-point interpolated precision is *averaged across all queries* in the test collection
- A perfect system will have a straight line from (0,1) to (1,1)

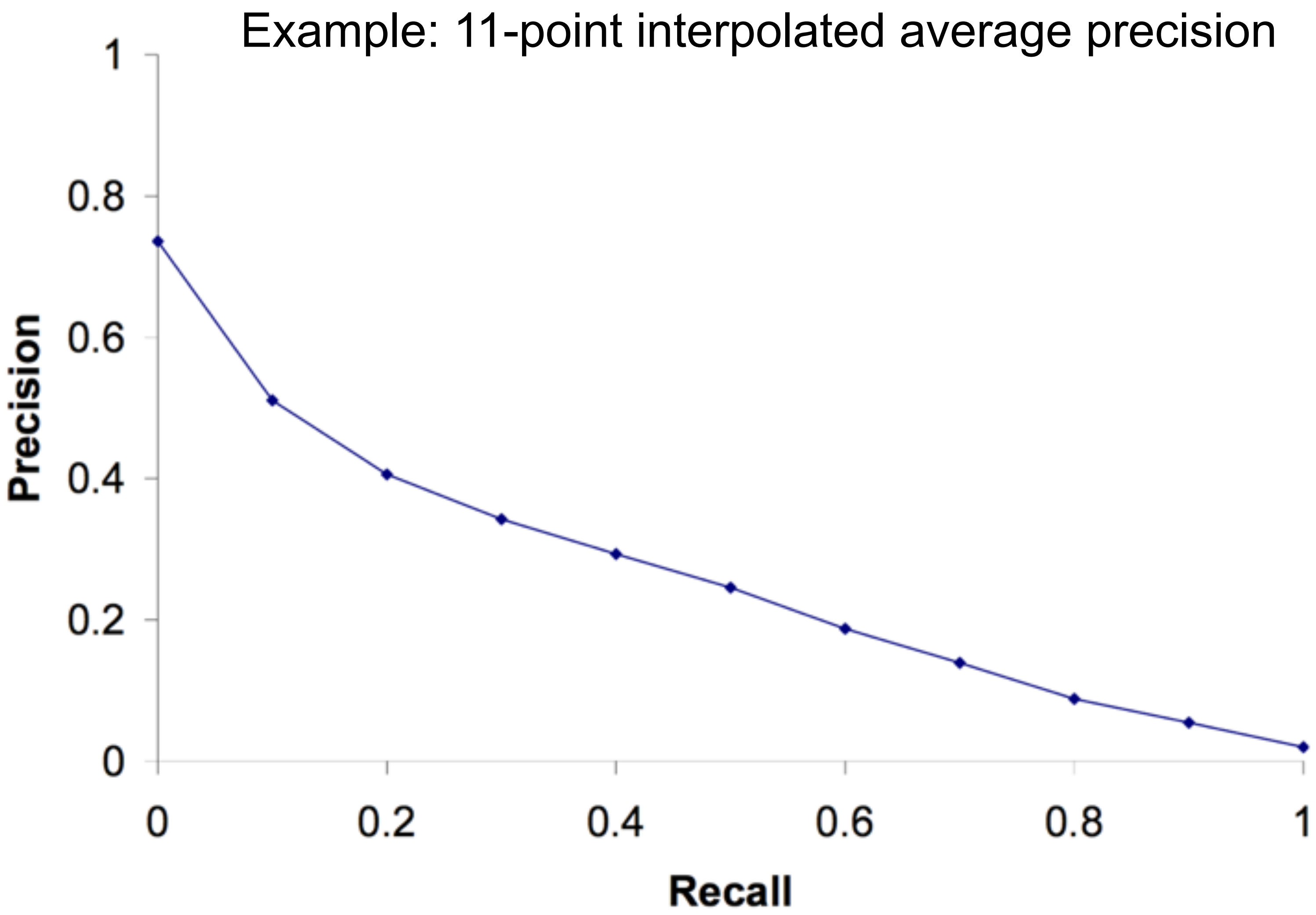


Recall	Interp. Prec.
0.0	1.00
0.1	1.00
0.2	1.00
0.3	1.00
0.4	0.66
0.5	0.66
0.6	0.66
0.7	0.50
0.8	0.50
0.9	0.50
1.0	0.50

Table: 11-point interpolated precision.

Ranked Eval.

Interpolated Precision



Ranked Eval.

Single Number Metrics

- Precision-Recall curves can be useful but sometimes we would like to use a single number to compare systems
- Average Precision and MAP
- Mean Reciprocal Rank (MRR)
- etc.

Average Precision and MAP

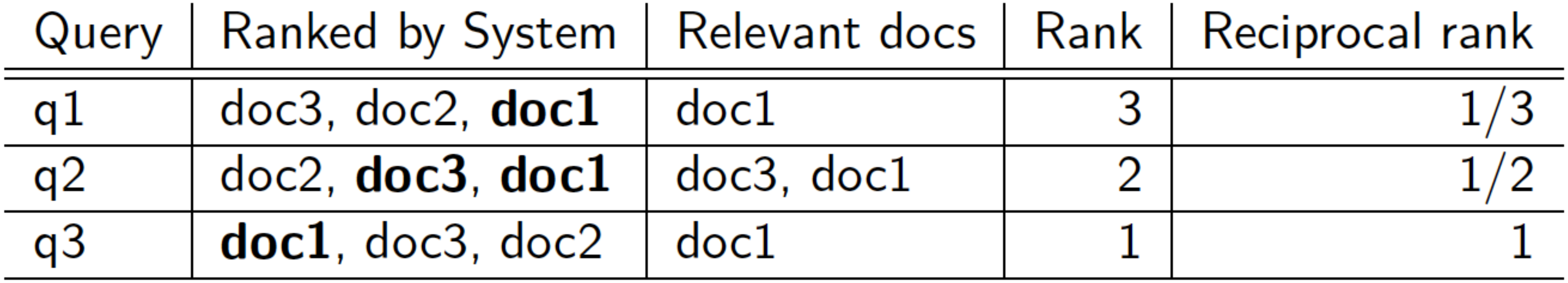
- Average precision (AP) for a single information need q

$$\text{AP}(q) = \frac{1}{m} \sum_{k=1}^{m} \text{Precision}(\mathscr{R}_k)$$

  - m: the number of relevant documents for q
  - $\mathscr{R}_k$: the set of ranked retrieval results from the top document down to the k-th relevant document
- Mean average precision (MAP)
  - The mean of the average precision (AP) for many information needs
  - MAP is a single-figure measure of quality across recall levels

Mean Reciprocal Rank (MRR)

- MRR is the averaged inverse rank of the first relevant document
- For if we only care about how high in the ranking the first relevant document is.



Query	Ranked by System	Relevant docs	Rank	Reciprocal rank
q1	doc3, doc2, **doc1**	doc1	3	1/3
q2	doc2, **doc3**, **doc1**	doc3, doc1	2	1/2
q3	**doc1**, doc3, doc2	doc1	1	1

Table: MRR example (from Wikipedia)

$$\text{Mean Reciprocal Rank} = \frac{1}{3}\left(\frac{1}{3} + \frac{1}{2} + 1\right) = 0.61$$

Ranked Eval.

Other Ranking Measures

- Precision at k: The proportion of relevant document in the top k retrieved documents
- R-precision: The precision at $|Rel|$ documents returned (Rel is a set of known relevant document for an information need), i.e. the same as precision at $|Rel|$
- Receiver Operating Characteristics (ROC) curve and ROC-AUC (i.e. area under the ROC curve)
- Normalised Discounted Cumulative Gain (NDCG) which requires graded relevance judgements (i.e. scores)

Summary

- Purpose of evaluation
  - Why do we need evaluation
  - What do we want to evaluate
- Test collection
  - Three components of test collections
  - Standard test collections
  - Build large test collection
- Evaluation of unranked retrieval sets
  - Precision, recall, accuracy and F-measure
- Evaluation of ranked retrieval results
  - Precision-recall curve, interpolated precision
  - Single number metrics: MAP and MRR

References

- Chapter 8, Introduction to Information Retrieval
- Some lecture slides are from
  - Pandu Nayak and Prabhakar Raghavan, CS276
  - Information Retrieval and Web Search, Stanford University