Evaluation of IR systems
Introduction

- How to measure user happiness?

- Depends on many factors:
  - Relevance of results
  - User interface design layout
  - Speed of response
  - Target application
    - Web engine: user finds what they want and return to the engine
      - Can measure rate of return users
    - e-commerce site: user finds what they want and make a purchase
      - Is it the end-user, or the e-commerce site, whose happiness we measure?
      - Measure time to purchase, or fraction of searchers who become buyers?
    - Enterprise (company/govt/academic): Care about “user productivity”
      - How much time do my users save when looking for information?
      - Many other criteria having to do with breadth of access, secure access …

Introduction

- System quality
  - How fast does the system index?
    - How many documents/hour for a certain distribution of document sizes?
  - How fast does it search?
    - latency as function of index size
  - How large is the document collection?
  - How expressive is its query language? How fast is it on complex queries?
- all but the last criteria are measurable

Introduction

- To measure ad hoc information retrieval effectiveness in the standard way, we need:
  - A test document collection
  - A test suite of information needs, expressible as queries
  - A set of relevance judgements
    - which documents are relevant/non-relevant for each query a.k.a. Ground Truth, Gold Standard
- Test collection must be of a reasonable size
  - Need to average performance since results are very variable over different documents and information needs

Introduction

- Relevance is assessed relative to an information need, not to a query.

  - For example, the information need:

    "I'm looking for information on whether drinking red wine is more effective at reducing your risk of heart attack than white wine"

    might be translated into the query:

    white AND red AND wine AND heart AND attack AND effective

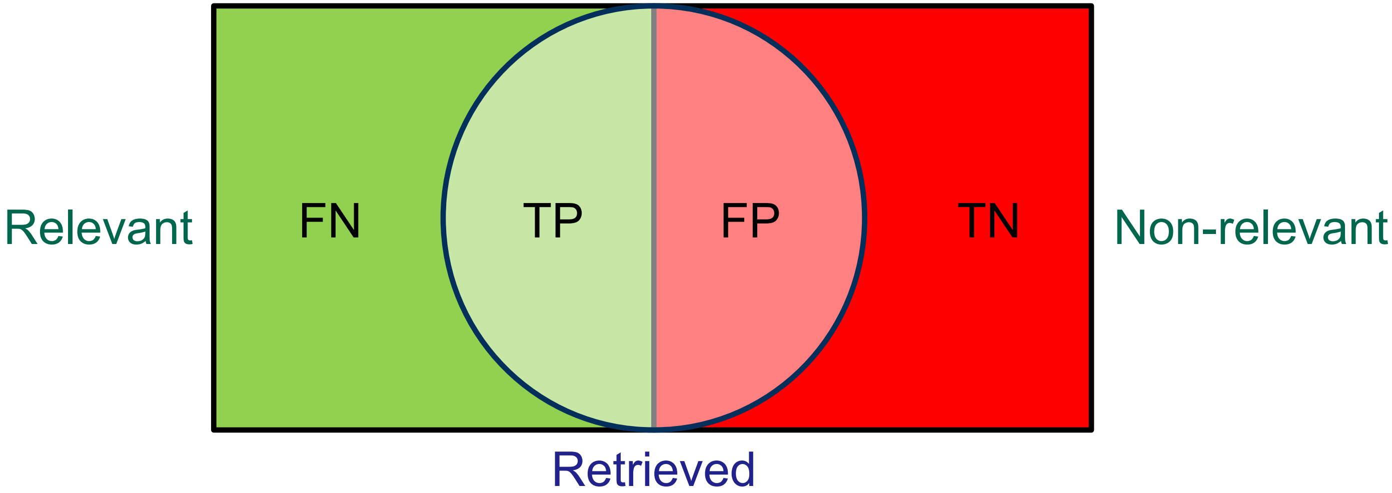
- A document is relevant if it addresses the stated information need, *not just because it contains all the word in the query*

Unranked retrieval: TP, FP, FN, TN

	Relevant	Non-relevant
Retrieved	true positive (TP)	false positive (FP)
Not-retrieved	false negative(FN)	true negative(TN)

Retrieved/Not-retrieved: from IR system
Relevant/Non-relevant: from Ground Truth

True: Retrieved/Not-retrieved corresponds to Relevent/Non-relevant
False: Retrieved/Not-retrieved doesn’t correspond to Relevent/Non-relevant



Unranked retrieval: Precision and Recall

- Precision (P): fraction of retrieved documents that are relevant

$$P = \frac{relevant\ retrieved}{retrieved} = \frac{TP}{TP + FP}$$

  - Measures the “degree of soundness” of the system

- Recall (R): fraction of relevant documents that are retrieved

$$R = \frac{relevant\ retrieved}{relevant} = \frac{TP}{TP + FN}$$

  - Measures the “degree of completeness” of the system

Unranked retrieval: Precision and Recall

- An IR system can get high recall (but low precision) by retrieving all documents for all queries
  - Recall is a non-decreasing function of the number of retrieved documents
  - Precision in good IR systems is a decreasing function of the number of retrieved documents

- Precision can be computed at different levels of recall

- Precision-oriented users
  - Web surfers
- Recall-oriented users
  - Professional searchers, paralegals, intelligence analysts

Unranked retrieval: F-measure

- F-measure(F): weighted harmonic mean of Precision and Recall

$$F = \frac{1}{\alpha \frac{1}{P} + (1 - \alpha) \frac{1}{R}}$$

- $\alpha \in [0,1]$
  - $\alpha = 1 \rightarrow F = P$
  - $\alpha = 0 \rightarrow F = R$
  - Usually $\alpha = 0.5 \rightarrow F = \frac{2 \cdot PR}{P+R}$ (balanced F-measure)
- Trade-off between the “degree of soundness” and the “degree of completeness” of a system

- Weighted harmonic mean: $H = \frac{\sum_{i=1}^{n} \alpha_i}{\sum_{i=1}^{n} \alpha_i \frac{1}{x_i}}$

Evaluation of IR systems
Unranked retrieval: F-measure

- Harmonic mean is a conservative average

    - e.g. 1 document out of 10000 is relevant
    - Retrieving all documents
        - Recall = 100%
        - Precision = 0.01%
        - Arithmetic mean = $\frac{1}{2}(P+R) = 50\%$
        - Harmonic mean (Balanced F-measure) = $\frac{2 \cdot PR}{P+R} = 0.02\%$

- When the value of two number differs, harmonic mean is closer to their minimum than arithmetic or geometric mean

Evaluation of IR systems

Unranked retrieval: Accuracy

- Accuracy(A): fraction of correctly classified documents

$$A = \frac{TP + TN}{TP + TN + FP + FN}$$

- Accuracy is not suitable in the context of IR
  - In many cases data are extremely skewed
    - e.g. 99.99% of documents are Non-relevant
  - In these cases a system tuned to maximize the accuracy will almost always retrieve nothing!
    - Accuracy is 99.99%
    - Recall is $\frac{0}{1}$
    - Precision is $\frac{0}{0}$

Unranked retrieval: Precision and Recall drawbacks

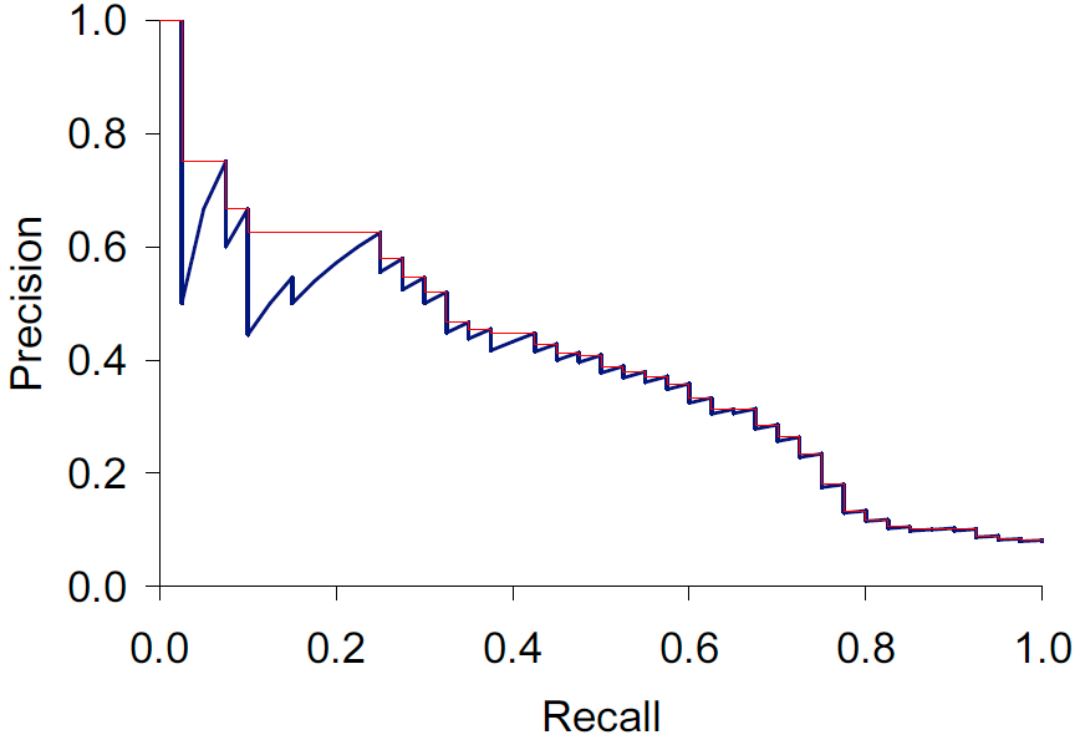
Difficulties in using Precision/Recall

- Average over large corpus/query…
  - Need human relevance assessments
    - People aren't reliable assessors
  - Assessments have to be binary
    - Nuanced assessments?
  - Heavily skewed by corpus/authorship
    - Results may not translate from one domain to another

- *The relevance of one document is treated as independent of the relevance of other documents in the collection*
  - This is also an assumption in most retrieval systems

Ranked retrieval: Precision and Recall

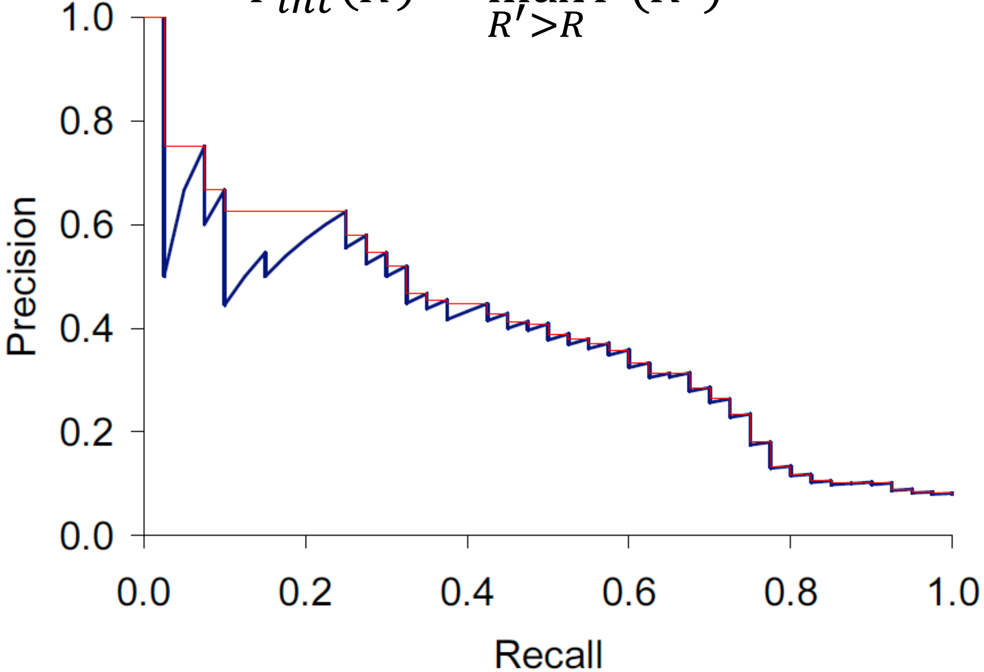
- Precision/Recall/F-measure are set-based measures
  - Unordered sets of documents
- In ranked retrieval systems, P and R are values relative to a rank position
  - Evaluation performed by computing Precision as a function of Recall



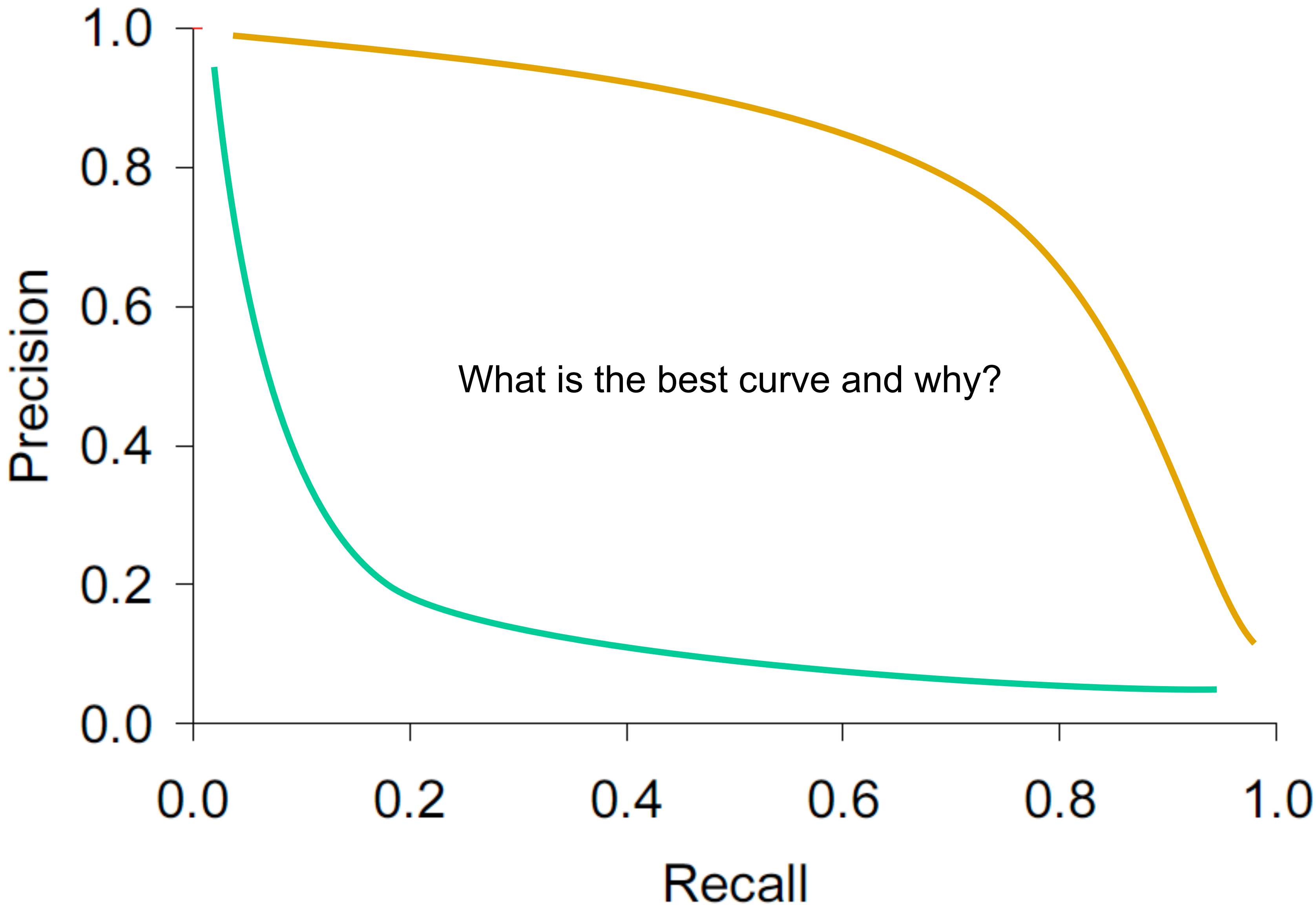
Ranked retrieval: Precision and Recall

- Precision/Recall function
  - If the $(k+1)th$ retrieved document is relevant, then $R(k+1) > R(k)$ and $P(k+1) \geq P(k)$
  - If the $(k+1)th$ retrieved document is non-relevant, then $R(k+1) = R(k)$, but $P(k+1) \leq P(k)$
- To remove the jiggles, use interpolated precision

$$P_{int}(R) = \max_{R'>R} P(R')$$



Ranked retrieval: Precision and Recall



Ranked retrieval: Average Precision

- 11-point interpolated average precision
  - measure precision at 11 recall levels $\{0.0, 0.1, 0.2, \ldots, 1.0\}$
  - compute the arithmetic mean of the precision levels

- mean average precision (MAP)
  - Given a set of queries Q, whose cardinality is $|Q|$

1. Compute the average precision (AP) for each query
   - Average the precision values obtained for the top set of k documents *after each relevant document is retrieved*
   - For a single query, AP is *related* to the area under the un-interpolated Precision/Recall curve
2. Compute the mean AP over the set of queries

Ranked retrieval: Average Precision

- MAP = mean AP over the set of queries

$$MAP(Q) = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \left(\frac{1}{m_i} \sum_{k=1}^{m_i} P(\Re_k) \right)$$

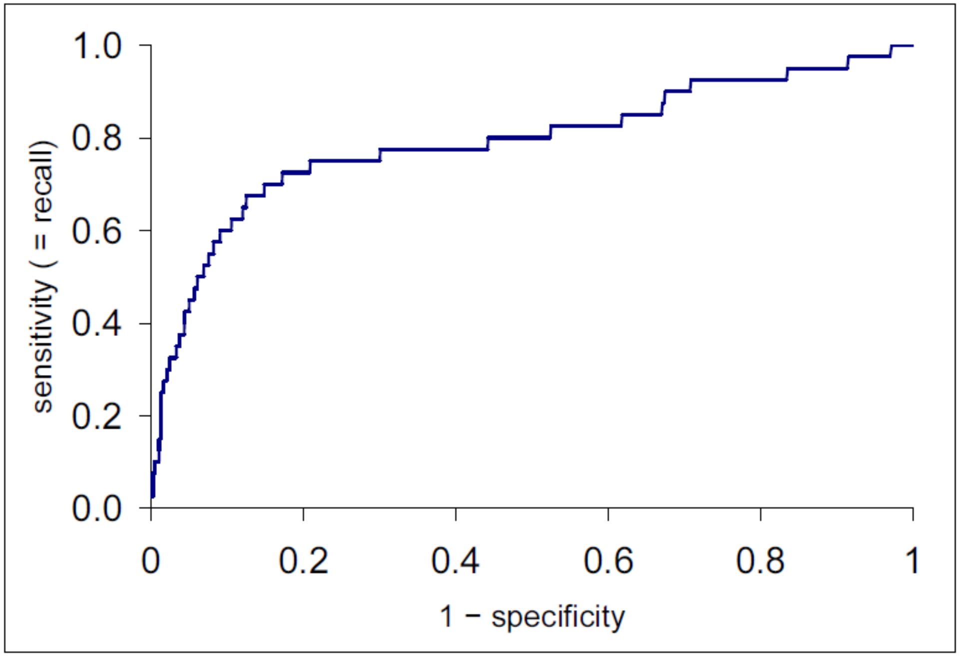
- $\{d_1, \ldots d_{m_i}\}$ documents relevant to query q_i
- $\Re_k$ top-k ranked set of retrieval results

Ranked retrieval: Precision at k, R-precision

- Precision at k
  - Set a fixed value of retrieved results k
  - Compute precision among top-k items
    - pro: does not require any estimate of the size of the set of relevant documents (useful in Web search)
    - con: total number of relevant documents has strong influence on Precision at k
      - e.g. with 8 relevant docs precision at 20 can be at most 0.4
- R-precision
  - Given a relevant set of size Rel
  - Calculate number of relevant documents r in the top-Rel set
    - pros:
      - a perfect system achieves R-precision = 1.0
      - Intuitive meaning: $\frac{r}{Rel}$ = precision at Rel = recall at Rel
    - con: considers only one point on the Precision/Recall curve

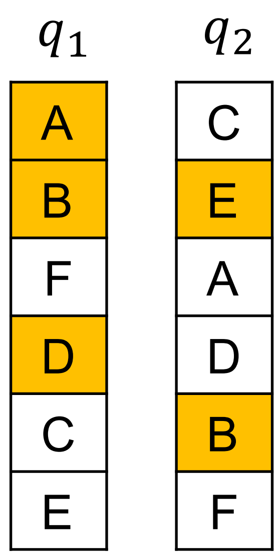
Ranked retrieval: Receiver-Operating-Characteristic (ROC)

- True positive rate (*sensitivity*) vs. false positive rate (1 – *specificity*)
- TP rate = *sensitivity* = Recall = $\frac{TP}{TP+FN} = \frac{retrieved\ relevant}{relevant}$
  - fraction of relevant documents that are retrieved
- FP rate = 1 – *specificity* = $\frac{FP}{FP+TN} = \frac{retrieved\ non\text{-}relevant}{non\text{-}relevant}$
  - fraction of non-relevant documents that are retrieved



Ranked retrieval: example

- An IR system gives the following rankings in response to two queries q_1 and q_2
- The highlighted documents are the ones relevant to the user for a specific query
- Suppose that the whole document collection is shown for each query
  - The total number of relevant and non-relevant documents is known



Ranked retrieval: example

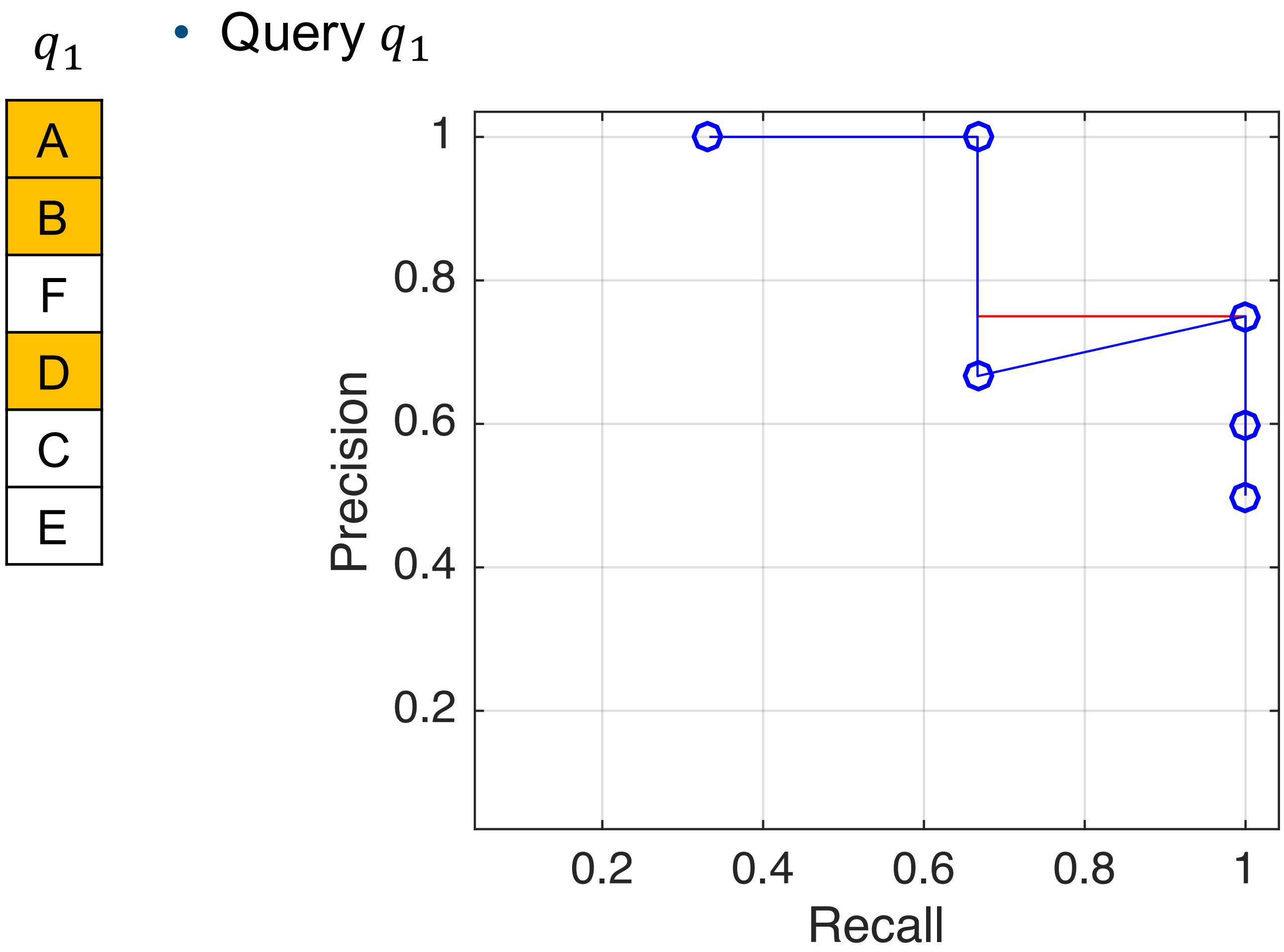
- Draw the Precision-Recall curve for each query

q_1
A
B
F
D
C
E

- Query q_1
  - Precision and Recall at 1 $P(1) = \frac{1}{1}$ $R(1) = \frac{1}{3}$
  - Precision and Recall at 2 $P(2) = \frac{2}{2}$ $R(2) = \frac{2}{3}$
  - Precision and Recall at 3 $P(3) = \frac{2}{3}$ $R(3) = \frac{2}{3}$
  - Precision and Recall at 4 $P(4) = \frac{3}{4}$ $R(4) = \frac{3}{3}$
  - Precision and Recall at 5 $P(5) = \frac{3}{5}$ $R(5) = \frac{3}{3}$
  - Precision and Recall at 6 $P(6) = \frac{3}{6}$ $R(6) = \frac{3}{3}$

Ranked retrieval: example

- Draw the Precision-Recall curve for each query (continue…)



Ranked retrieval: example

- Draw the Precision-Recall curve for each query (continue…)

q_2
C
E
A
D
B
F

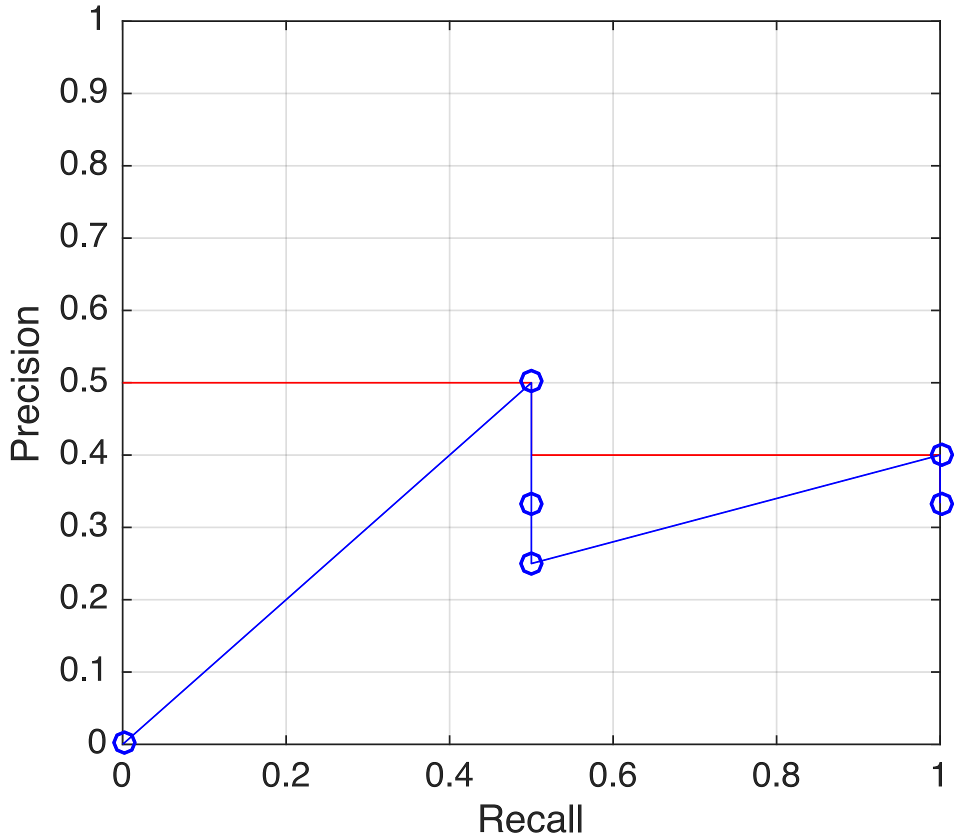
- Query q_2
  - Precision and Recall at 1 $P(1) = \frac{0}{1}$ $\mathrm{R}(1) = \frac{0}{2}$
  - Precision and Recall at 2 $P(2) = \frac{1}{2}$ $\mathrm{R}(2) = \frac{1}{2}$
  - Precision and Recall at 3 $P(3) = \frac{1}{3}$ $\mathrm{R}(3) = \frac{1}{2}$
  - Precision and Recall at 4 $P(4) = \frac{1}{4}$ $\mathrm{R}(4) = \frac{1}{2}$
  - Precision and Recall at 5 $P(5) = \frac{2}{5}$ $\mathrm{R}(5) = \frac{2}{2}$
  - Precision and Recall at 6 $P(6) = \frac{2}{6}$ $\mathrm{R}(6) = \frac{2}{2}$

Ranked retrieval: example

- Draw the Precision-Recall curve for each query (continue…)

q_2
C
E
A
D
B
F

- Query q_2



Ranked retrieval: example

- Determine the R-precision for each query
  - Query q_1
    - $Rel = 3 \rightarrow$ R-precision = $P(3) = \frac{2}{3}$
  - Query q_2
    - $Rel = 2 \rightarrow$ R-precision = $P(2) = \frac{1}{2}$

- Calculate the Mean Average Precision
  - $AP_1 = \frac{1}{3}\big(P(1) + P(2) + P(4)\big) = \frac{11}{12}$
  - $AP_2 = \frac{1}{2}\big(P(2) + P(5)\big) = \frac{9}{20}$
  - $MAP = \frac{1}{2}(AP_1 + AP_2) = \frac{41}{60}$

Evaluation of IR systems
Ranked retrieval: example

- Draw the Receiver-Operating-Characteristic for each query

q_1

A
B
F
D
C
E

- Query q_1
  - $TP_{rate}(1) = R(1) = \frac{1}{3}$ $\quad FP_{rate}(1) = \frac{0}{3}$
  - $TP_{rate}(2) = R(2) = \frac{2}{3}$ $\quad FP_{rate}(2) = \frac{0}{3}$
  - $TP_{rate}(3) = R(3) = \frac{2}{3}$ $\quad FP_{rate}(3) = \frac{1}{3}$
  - $TP_{rate}(4) = R(4) = \frac{3}{3}$ $\quad FP_{rate}(4) = \frac{1}{3}$
  - $TP_{rate}(5) = R(5) = \frac{3}{3}$ $\quad FP_{rate}(5) = \frac{2}{3}$
  - $TP_{rate}(6) = R(6) = \frac{3}{3}$ $\quad FP_{rate}(6) = \frac{3}{3}$

Evaluation of IR systems

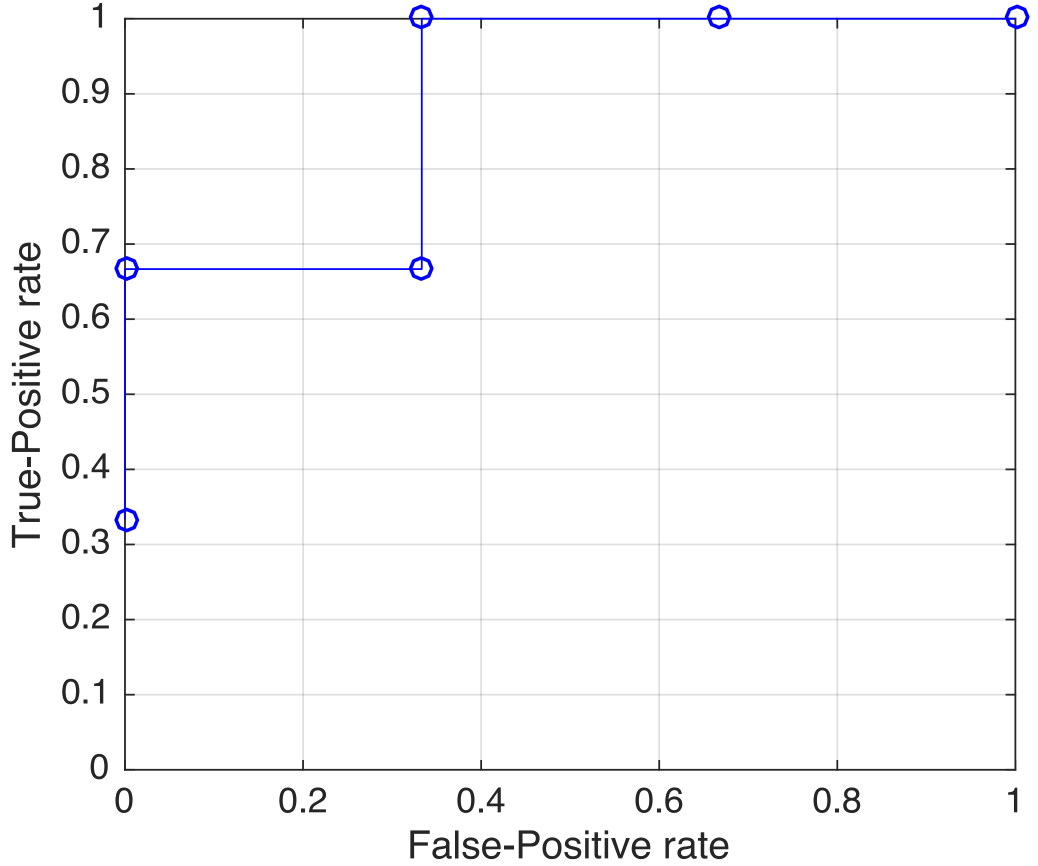
Ranked retrieval: example

- Draw the Receiver-Operating-Characteristic for each query

q_1

A
B
F
D
C
E

- Query q_1



Ranked retrieval: example

- Draw the Receiver-Operating-Characteristic for each query

q_2
C
E
A
D
B
F

- Query q_2
  - $TP_{rate}(1) = R(1) = \frac{0}{2}$ $\quad$ $FP_{rate}(1) = \frac{1}{4}$
  - $TP_{rate}(2) = R(2) = \frac{1}{2}$ $\quad$ $FP_{rate}(2) = \frac{1}{4}$
  - $TP_{rate}(3) = R(3) = \frac{1}{2}$ $\quad$ $FP_{rate}(3) = \frac{2}{4}$
  - $TP_{rate}(4) = R(4) = \frac{1}{2}$ $\quad$ $FP_{rate}(4) = \frac{3}{4}$
  - $TP_{rate}(5) = R(5) = \frac{2}{2}$ $\quad$ $FP_{rate}(5) = \frac{3}{4}$
  - $TP_{rate}(6) = R(6) = \frac{2}{2}$ $\quad$ $FP_{rate}(6) = \frac{4}{4}$

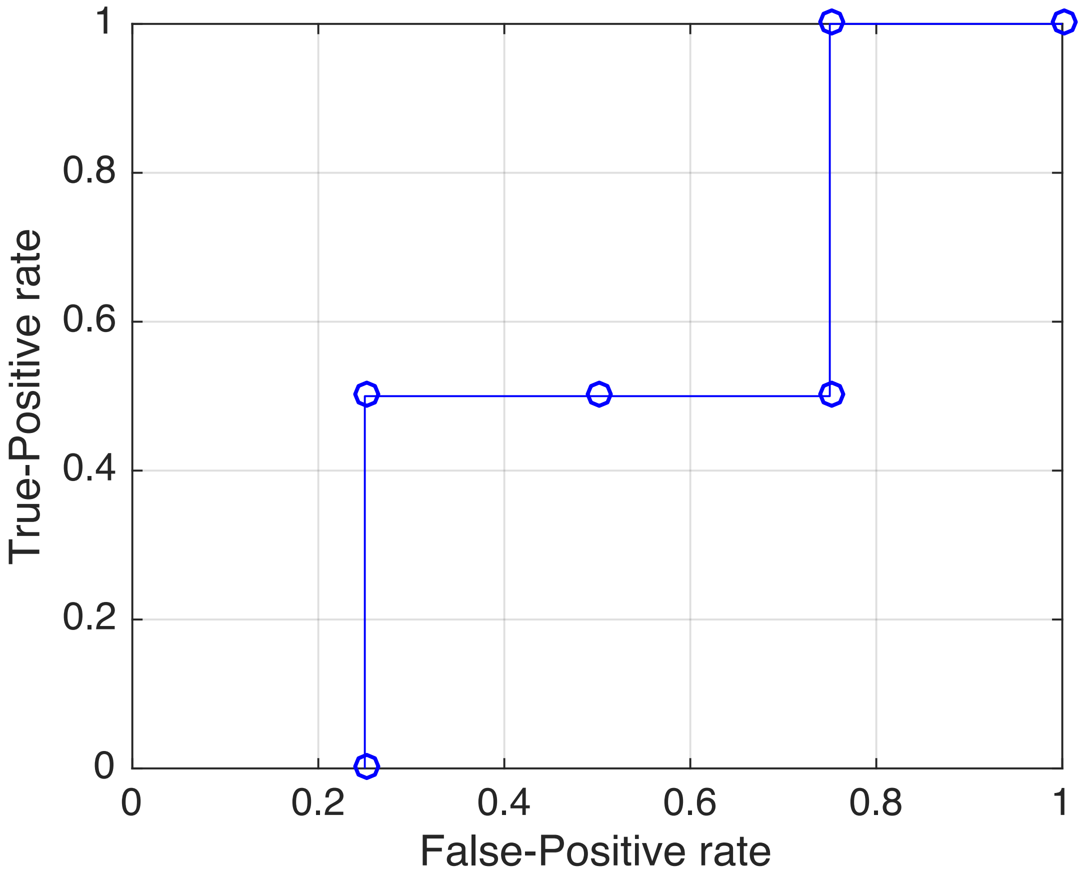
Ranked retrieval: example

- Draw the Receiver-Operating-Characteristic for each query

q_2

C
E
A
D
B
F

- Query q_2



References

- [Baeza-Yates and Ribeiro-Nieto, 1999] R. Baeza-Yates and B. Ribeiro-Nieto, “Modern Information Retrieval”, 1999 (http://www.mir2ed.org/)
- [Manning et al., 2008] C.D. Manning, P. Raghavan and H. Schütze, “Introduction to Information Retrieval”, Cambridge University Press, 2008 (http://nlp.stanford.edu/IR-book/)