# IR Evaluation

Lecture 9

# Evaluation in IR

"How well does the system work?"

- Can be investigated at several levels
  1. Processing: Time and space efficiency
  2. Search: Effectiveness of results
  3. System: Satisfaction of the user
- Here we focus on evaluating retrieval effectiveness

# Retrieval Effectiveness

- In response to a ***query***, an IR ***system*** searches its document ***collection*** and returns a ***ordered list of responses***.
  - called the *retrieved set* or *ranked list*
  - The system employs a search strategy or algorithm
- Measure the quality of a ranked list
  - a better search strategy yields a better ranked list
  - Better ranked lists help the user fill their information need

# Ranking Example

- How good are these search results?
- Could they be better? How?

| Rank | Doc# | Rel? |
|---|---|---|
| 1 | 5 | |
| 2 | 3 | YES |
| 3 | 10 | |
| 4 | 35 | YES |
| 5 | 4 | |
| 6 | 270 | |
| 7 | 14 | YES |
| 8 | 15 | YES |
| 9 | 11 | YES |
| 10 | 1 | |

# Relevant and Retrieved Sets

| | Relevant | Irrelevant |
|---|---|---|
| Retrieved | A | B |
| Not Retrieved | C | D |

- With respect to a given query, the documents can be partitioned into four sets
  - Relevant or not, retrieved or not
  - User says Yes/No, system says Yes/No

# Precision and Recall

| | Relevant | Irrelevant |
|---|---|---|
| Retrieved | A | B |
| Not Retrieved | C | D |

- Precision = fraction of retrieved documents that are relevant
- Recall = fraction of relevant documents retrieved.

$$Precision = \frac{A}{(A \cup B)}$$

$$Recall = \frac{A}{(A \cup C)}$$

# P/R Example

- Precision
  - fraction of retrieved documents that are relevant
- Recall
  - fraction of all relevant documents retrieved.
- What are the precision and recall of this retrieved set?
- What are we missing to answer this?

| Rank | Doc# | Rel? |
|---|---|---|
| 1 | 5 | |
| 2 | 3 | YES |
| 3 | 10 | |
| 4 | 35 | YES |
| 5 | 4 | |
| 6 | 270 | |
| 7 | 14 | YES |
| 8 | 15 | YES |
| 9 | 11 | YES |
| 10 | 1 | |

# From Sets to Rankings

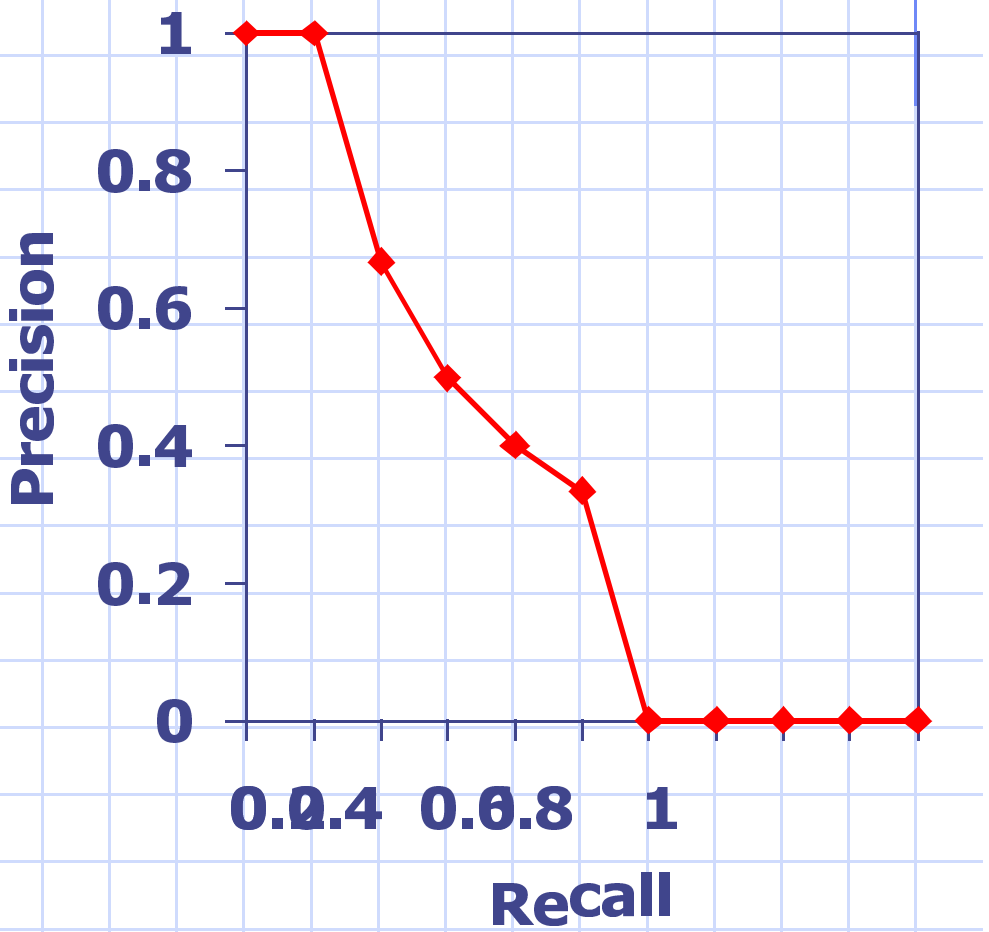
- Precision and Recall are measures of sets
- In a ranked list, we can measure the precision at each *recall point*
  - recall increases when a relevant document is retrieved
  - compute precision at each relevant retrieved document, over that fraction of the retrieved set

# Recall vs. Precision

- There is a tradeoff between recall and precision
  - Can increase recall by retrieving more
  - This can decrease precision
- What does the tradeoff mean?
  - Consider different kinds of *user models*
  - A Web searcher is only going to look at the top 20 documents
  - A patent attorney wants all the relevant documents

# Recall-Precision Graphs

- Shows the retrieval performance at each point in the ranking
- Graph precision at standard recall points
  - 10%, 20%,…,100%
  - Interpolate between points
  - Prec(Rec=r) = max(Prec(Rec >= r))



# Recall and Precision

- The recall-precision graph illustrates the tradeoff made by a search algorithm
  - shows system performance at multiple *operating points*
  - Each user may be interested in a different point on the graph

# Single-number measures

- Often it is useful to have a single number to summarize performance

- Precision at *n* documents retrieved
  - shows set precision at fixed points in the ranking
- R-precision
  - Precision when (# relevant documents) retrieved
- Average precision
  - Average of precision at each relevant document retrieved
  - Precision of an unretrieved relevant document = 0

# van Rijsbergen's F-measure

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{recall(j)} + \frac{1}{precision(j)}}$$

- F(j) = 1 - E(j)
  - so lower values mean lower performance
- If b=1, F(j) is average of precision and recall
- If b=0, F(j) is precision
- If b=Inf, F(j) is recall
- b=2 is a common choice

# Averaging

- We can now measure the effectiveness of a ranked list for a single query
- Want to find the expected performance for an arbitrary query
  - information needs, query lengths, topic coverage, term usage all vary
  - average the measure over many different queries to find the average effectiveness

# Mean Average Precision

- Find the average precision for each query
- Compute the mean AP over all queries
  - This is *macroaverage* - all queries are considered equal
  - A *microaverage* would take the mean of the precision at each recall point for all queries together
- For average recall-precision curves, take average at standard recall points

# Other measures

- Reciprocal rank
- Error
  - Mean absolute error, mean squared error
  - Misses vs. false alarms
- Utility
  - a(rel, retrieved) + b(rel, missed) + c(irrel, retrieved) + d(irrel, missed)

# What is the Right Measure?

- Correspond to a user model
  - Precision: "I'm feeling lucky"
  - Recall: maximizing coverage of topic
  - AP: area under the recall-precision graph
    - effectiveness at each point in the ranking
  - F: explore P/R tradeoff in average precision
  - Reciprocal rank/P@1: known-item searching

# Measure Stability

- Some measures are unstable
  - Short result lists
  - Few relevant documents
- Averaging across queries helps stability
- Need more queries for less stable measures
  - Average precision: 25 ok, 50 good
  - Precision at 10: 150-200 is a good start