ADVANCED TOPICS IN INFORMATION RETRIEVAL AND WEB SEARCH

Lecture 4:

Information Retrieval Evaluation

Dr. S. Mehdi Vahidipour
Thanks Dr. Momtazi
Vahidipour@kashanu.ac.ir

Based on the text book slides.

Outline

- **Introduction**
- Unranked vs ranked retrieval
  - Evaluation of unranked retrieval
  - Evaluation of ranked retrieval
- Significant tests
- Evaluation data and benchmarks
- Evaluation at large scale data
- Results representation

IR Evaluation

- The key measure for a search engine is user happiness
- Main factors in user happiness
  - Speed of response
  - Uncluttered UI
  - Relevance
  - Free to use
- Note: none of these is sufficient
  - Blindingly fast, but useless answers won’t make a user happy

Relevance

- User happiness is equated with the relevance of search results to the query
- “Relevance to the query” is very problematic
- Example
  - Information need: “I am looking for information on whether drinking milk is effective at reducing your risk of heart attacks.”
  - Query: [milk reduce heart attack effect]
  - Sample document: "At the heart of his speech was an attack in the conference reception for reducing the use of unhealthy elements in producing milk."
  - It is an excellent match for the query but not relevant to the information need.

Relevance

- User happiness can only be measured by relevance to an information need, not by relevance to queries
- Our terminology is sloppy in IR: we talk about query-document relevance judgments even though we mean information-need-document relevance judgments

Evaluation Data

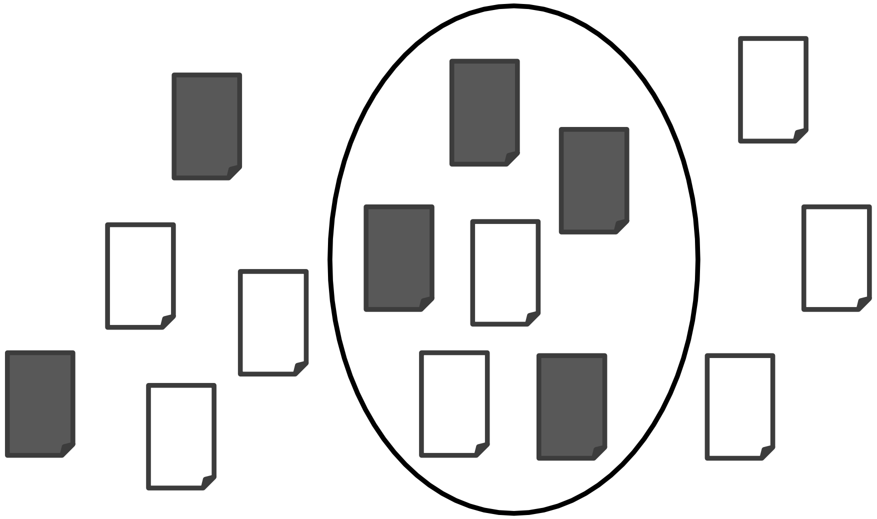
- Standard methodology in information retrieval for measuring relevance consists of three elements:
  - A benchmark document collection
  - A benchmark suite of queries
  - An assessment of the either relevant or nonrelevant for each pair of query and document

Outline

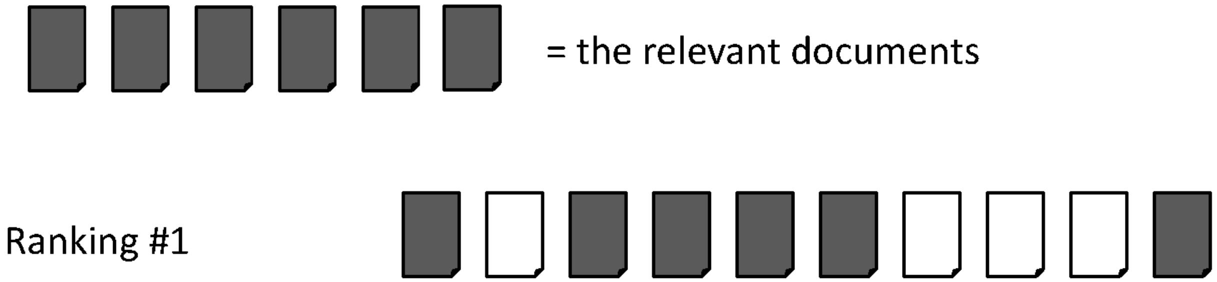
- Introduction
- **Unranked vs ranked retrieval**
  - **Evaluation of unranked retrieval**
  - Evaluation of ranked retrieval
- Significant tests
- Evaluation data and benchmarks
- Evaluation at large scale data
- Results representation

Unranked vs. Ranked Retrieval

- Unranked retrieval
  - Returns a set of documents with no priority
  - A boolean classification as relevance and nonrelevance



- Ranked retrieval
  - Returns a set of ranked documents
  - The position of document in the list is important



Precision and Recall

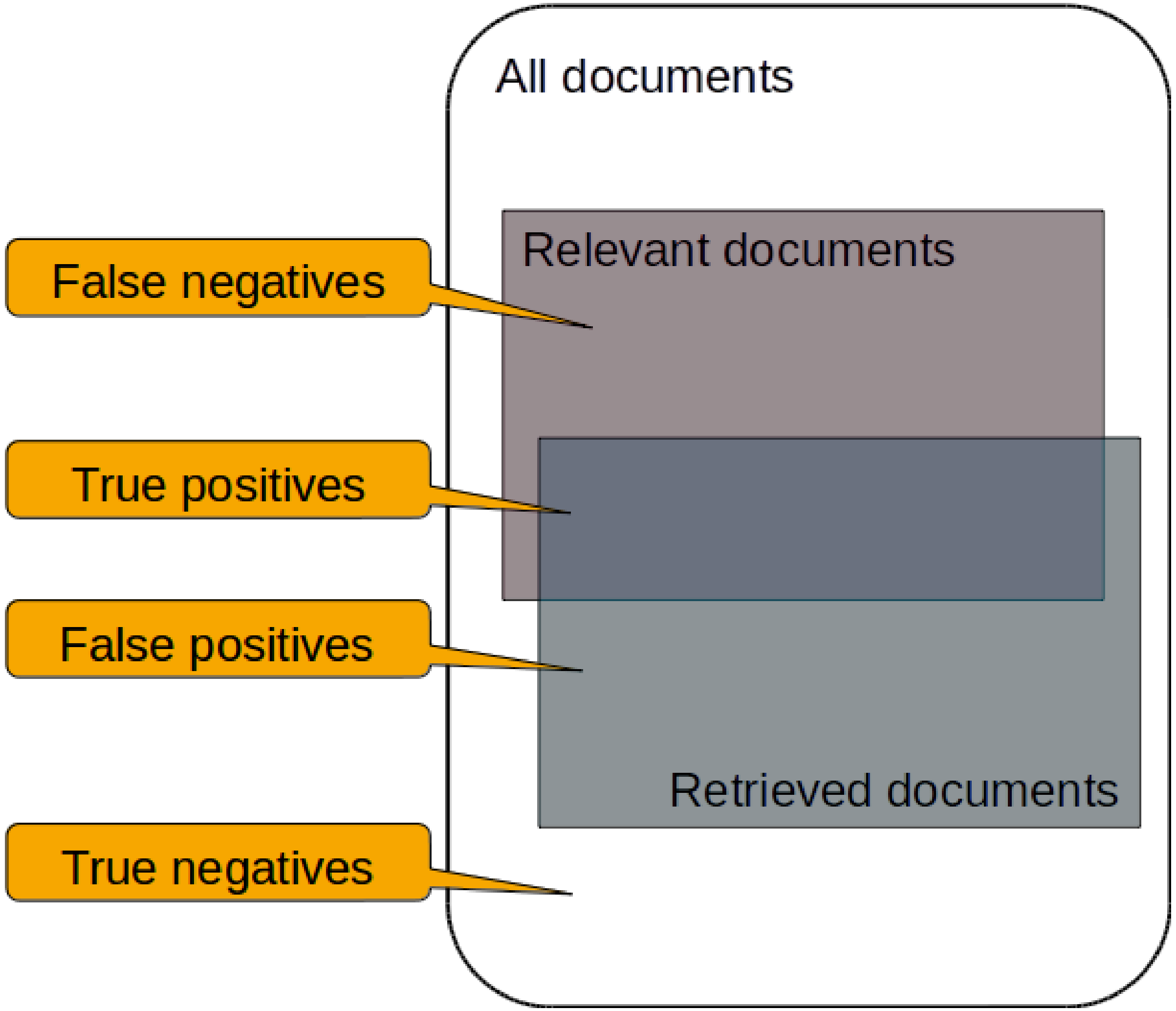
- Required data
  - The set of relevant documents (but we may not find all)
  - The set of retrieved documents (but not all of them are relevant)

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

  Precision = #(relevant items retrieved) / #(retrieved items) = P(relevant|retrieved)

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

  Recall = #(relevant items retrieved) / #(relevant items) = P(retrieved|relevant)



$$\text{Precision} = \frac{\text{True positives}}{\text{Retrieved documents}}$$

$$\text{Recall} = \frac{\text{True positives}}{\text{Relevant documents}}$$

Precision/Recall Tradeoff

- Find algorithm that maximizes precision.
  - Or minimizes classification errors (false positives)
  - Return nothing!

- Find algorithm that maximizes recall.
  - Return everything!

- Solution:
  - Considering both measures at the same time by F-Measure

F-Measure

- General formula

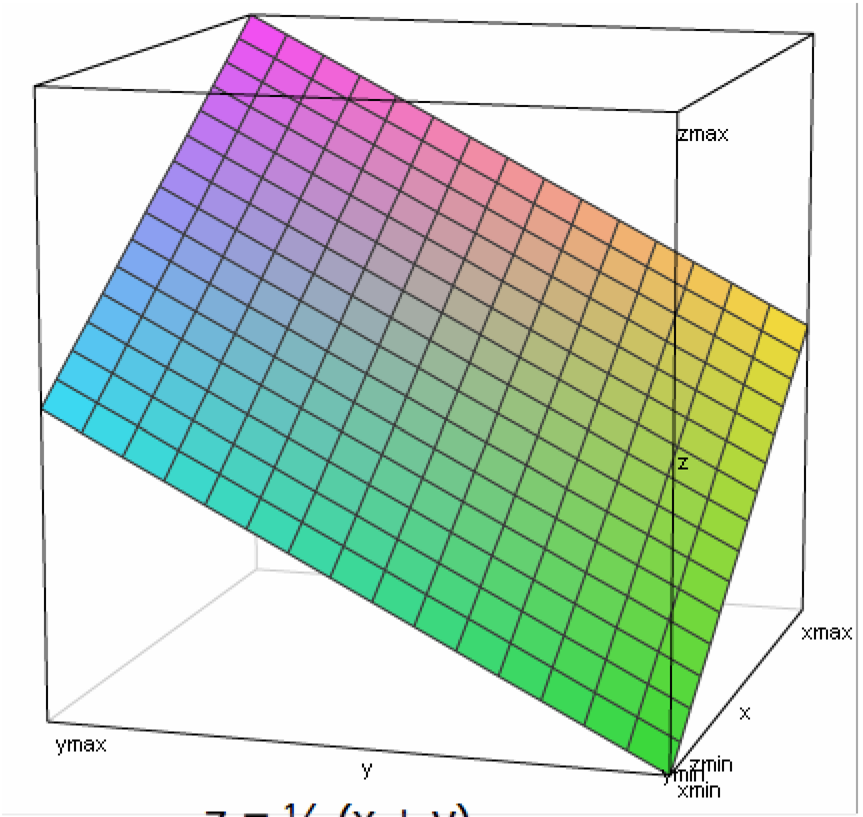
$$F = \frac{1}{\alpha\frac{1}{P} + (1-\alpha)\frac{1}{R}} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \quad \text{where} \quad \beta^2 = \frac{1-\alpha}{\alpha}$$

$\alpha \in [0, 1]$ and thus $\beta^2 \in [0, \infty]$

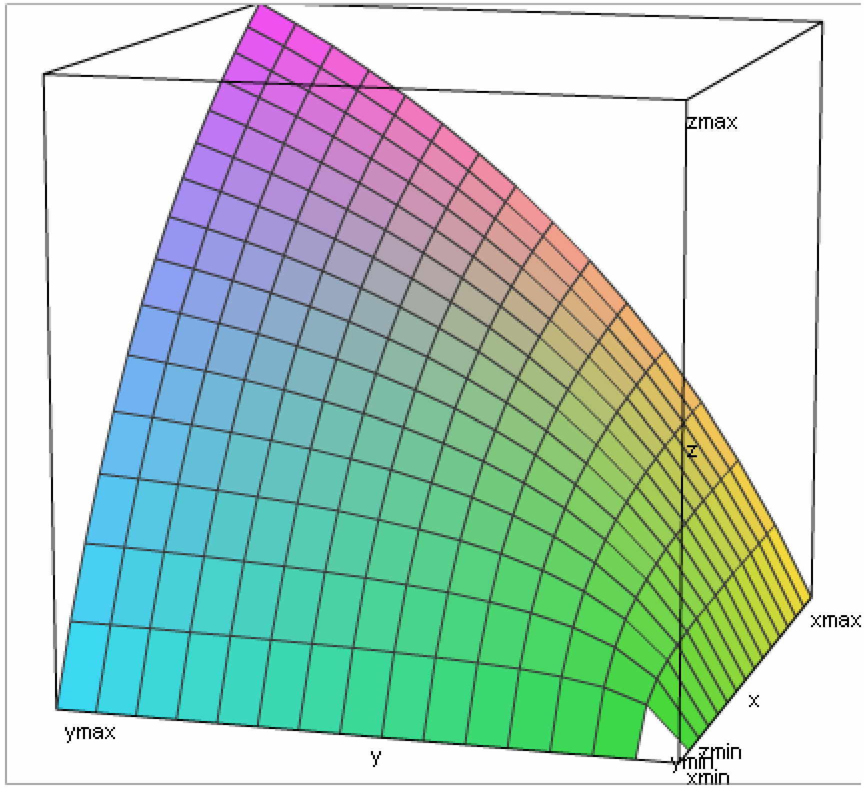
- Harmonic mean of recall and precision ($\beta^2 = 1$)

$$F = \frac{1}{\frac{1}{2}(\frac{1}{R} + \frac{1}{P})} = \frac{2RP}{R+P}$$

Arithmetic Mean (Average) vs Harmonic Mean (F-Measure)



$z = \frac{1}{2}(x + y)$



$z = 2(x \cdot y) / (x + y)$

Example

	relevant	not relevant	
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

Example

	relevant	not relevant	
retrieved	20	40	60
not retrieved	60	1,000,000	1,000,060
	80	1,000,040	1,000,120

$$P = \frac{20}{20+40} = \frac{1}{3}$$

$$R = \frac{20}{20+60} = \frac{1}{4}$$

$$F_1 = \frac{2 \times 1/3 \times 1/4}{1/3 + 1/4}$$

Accuracy

- Why not using a simpler measure like accuracy?
- Accuracy is the fraction of decisions that are correct (relevant/nonrelevant) .

Accuracy = (TP + TN)/(TP + FP + FN + TN).

- True Negative item is big enough to reduce the impact of other items

- Simple trick: always say no and return nothing => get 99.99% accuracy on most queries
- Searchers on the web (and in IR in general) want to find something and have a certain tolerance for junk
  - It's better to return some bad hits as long as you return something

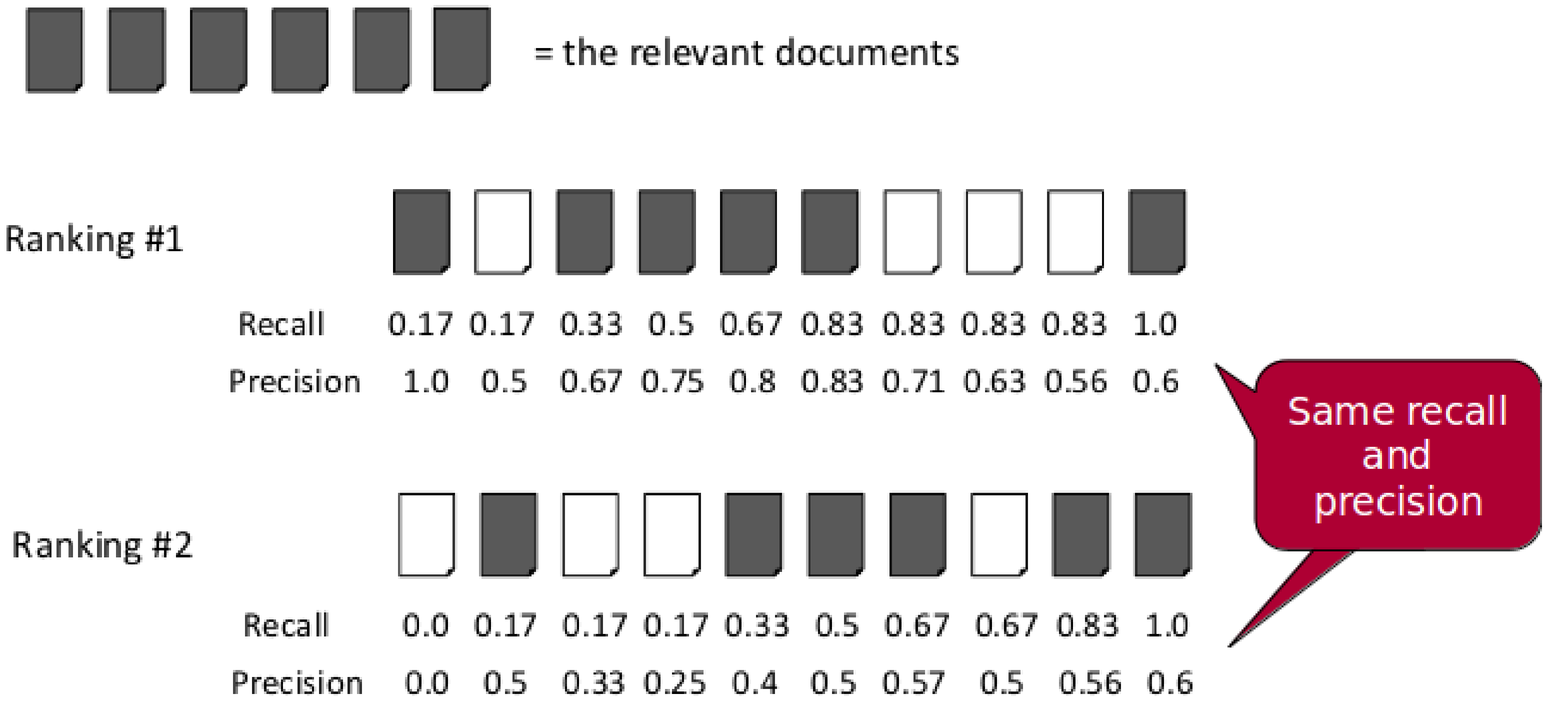
→ We use precision, recall, and F for evaluation, not accuracy.

Outline

- Introduction
- **Unranked vs ranked retrieval**
  - Evaluation of unranked retrieval
  - **Evaluation of ranked retrieval**
- Significant tests
- Evaluation data and benchmarks
- Evaluation at large scale data
- Results representation

Ranking Effectiveness

- Problem: Evaluate ranking, not just Boolean classification
- Idea: Calculate precision and recall at every rank position



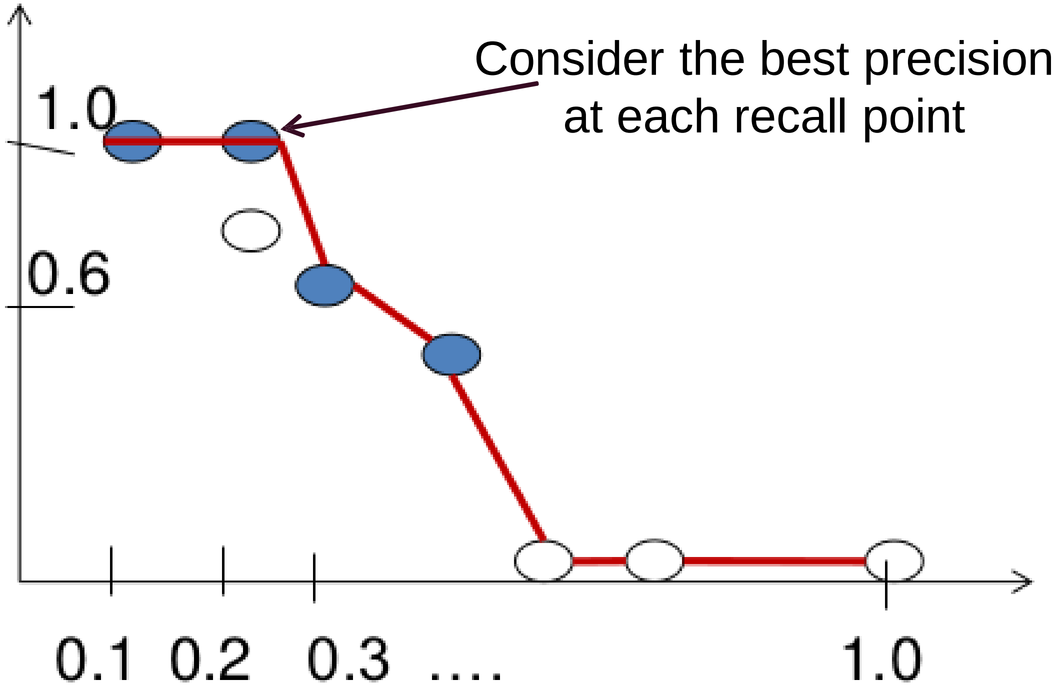
Ranking Effectiveness

- Problem: Long lists are unwieldy and difficult to compare
- Three ideas:
  - Calculating precision at standard recall levels, from 0.0 to 1.0 in increments of 0.1 => "Precision-Recall Curve"
  - Averaging the precision values from the rank positions where a relevant document was retrieved => "Average Precision"
  - Calculating precision at small number of fixed rank positions = > "Precision at rank k"
    - Ignores ranking after p; ignores ranking within 1 to p

Precision-Recall (PR) Curve

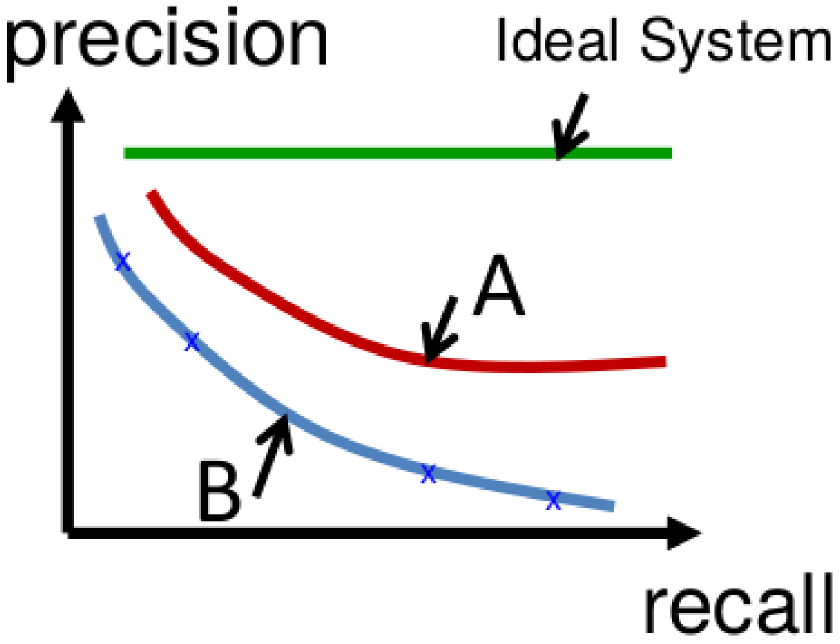
- Assume the total number of relevant documents in collection: 10

	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	?	
	10/10	



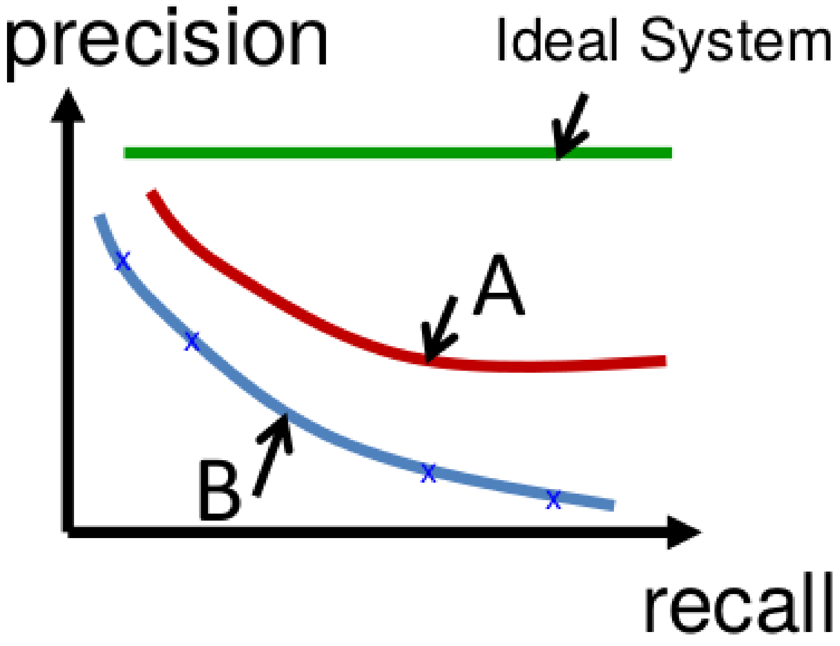
Precision-Recall (PR) Curve

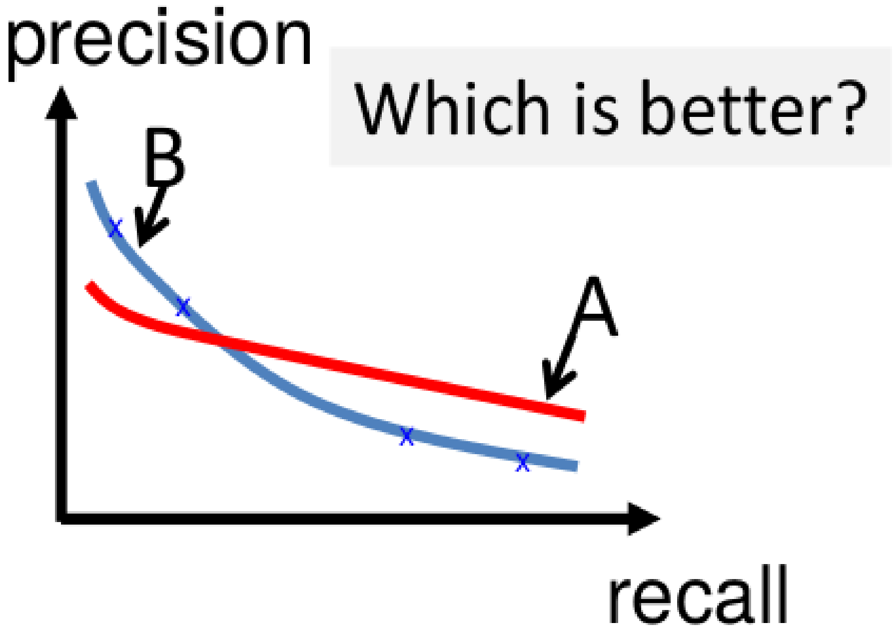
- Comparing PR Curves



Precision-Recall (PR) Curve

- Comparing PR Curves





Average Precision

- The average of precision at every cutoff where a new relevant document is retrieved
  - Normalizer = the total # of relevant docs in the retrieved collection
  - Sensitive to the rank of each relevant document

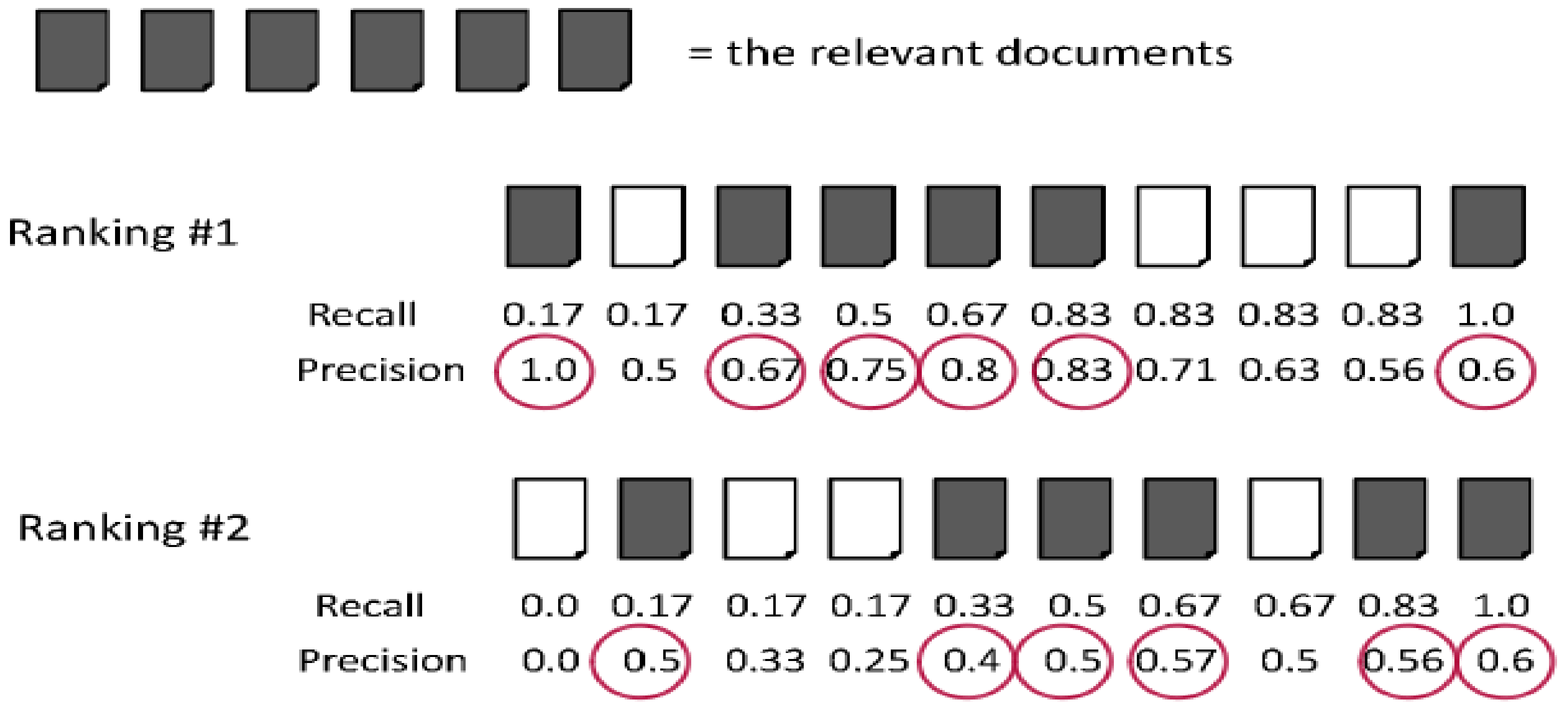
	Precision	Recall
D1 +	1/1	1/10
D2 +	2/2	2/10
D3 –	2/3	2/10
D4 –		
D5 +	3/5	3/10
D6 –		
D7 –		
D8 +	4/8	4/10
D9 –		
D10 –	?	
	10/10	

AP = (1/1 + 2/2 + 3/5 + 4/8+….) / 10

AP = (1/1 + 2/2 + 3/5 + 4/8) / 4

Average Precision

- Ranking #1: (1.0+0.67+0.75+0.8+0.83+0.6) / 6 = 0.78
- Ranking #2: (0.5+0.4+0.5+0.57+0.56+0.6) / 6 = 0.52



Mean Average Precision

- Evaluate ranking algorithm for more than one query
- Each ranking produces average precision
- Take average of those numbers

=> Mean Average Precision (MAP) (= average average precision)

- Most commonly used measure in research papers

Precision@k

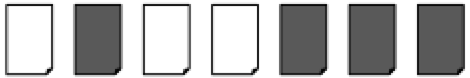
- Users tend to look at only the top part of the ranked result list to find relevant documents; e.g., first 1 or 2 result pages
- Measure how well the search engine does at retrieving relevant documents at very high ranks
- "Precision at Rank k"
  - K is typically 5, 10, 20
  - Easy to compute, easy to average over queries, easy to understand
  - But not sensitive to rank positions less than k
  - Single relevant document can be ranked anywhere

- Alternative: Reciprocal Rank

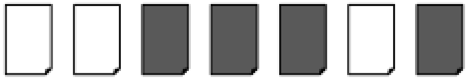
Reciprocal Rank

- Reciprocal of the rank at which the first relevant document is retrieved
- Very sensitive to rank position, regards only first relevant document

Reciprocal rank: 1/2



Reciprocal rank: 1/3



- Mean Reciprocal Rank (MRR) is the average of the reciprocal ranks over a set of queries

Discounted Cumulative Gain (DCG)

- Popular measure for evaluating web search and related tasks
- Uses graded relevance as a measure of the usefulness, or gain, from examining a document

- Two assumptions
  - Highly relevant documents are more useful than marginally relevant document
  - The lower the ranked position of a relevant document, the less useful it is for the user, since it is less likely to be examined

- Gain is accumulated starting at the top of the ranking
  - May be reduced, or discounted, at lower ranks
  - Typical discount is 1/log(rank)
  - With base 2, the discount at rank 4 is 1/2, and at rank 8 it is 1/3

Discounted Cumulative Gain (DCG)

- DCG is the total gain accumulated at a particular rank p:

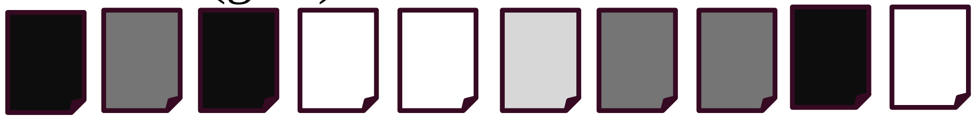
$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

  - rel_i is graded relevance of document at rank i.
  - Can use “Bad” = 0 to “Perfect” = 3 or 5

Discounted Cumulative Gain (DCG)

$$DCG_p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

- Example:
  - 10 ranked documents judged on 0-3 relevance scale (gain):
    - 3, 2, 3, 0, 0, 1, 2, 2, 3, 0



  - Discounted gain:
    - 3, 2/1, 3/1.59, 0, 0, 1/2.59, 2/2.81, 2/3, 3/3.17, 0

    = 3, 2, 1.89, 0, 0, 0.39, 0.71, 0.67, 0.95, 0
  - Discounted Cumulative Gain at each position:
    - 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61
  - DCG@5 = 6.89   DCG@10 = 9.61

Normalized DCG

- DCG values are often normalized by comparing the DCG at each rank with the DCG value for the perfect ranking.
- Makes averaging easier for queries with different numbers of relevant documents

=> NDCG $\leq$ 1 at any rank position

Normalized DCG

- Example
  - Original result:
    - 3, 2, 3, 0, 0, 1, 2, 2, 3, 0
  - Original DCG values
    - 3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61
  - Perfect ranking for the ten results:
    - 3, 3, 3, 2, 2, 2, 1, 0, 0, 0
  - Ideal DCG values:
    - 3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10.88
  - NDCG values (divide actual by ideal):
    - 1, 0.83, 0.87, 0.76, 0.71, 0.69, 0.73, 0.8, 0.88, 0.88