Informatics 225
Computer Science 221

Information Retrieval

Lecture 26

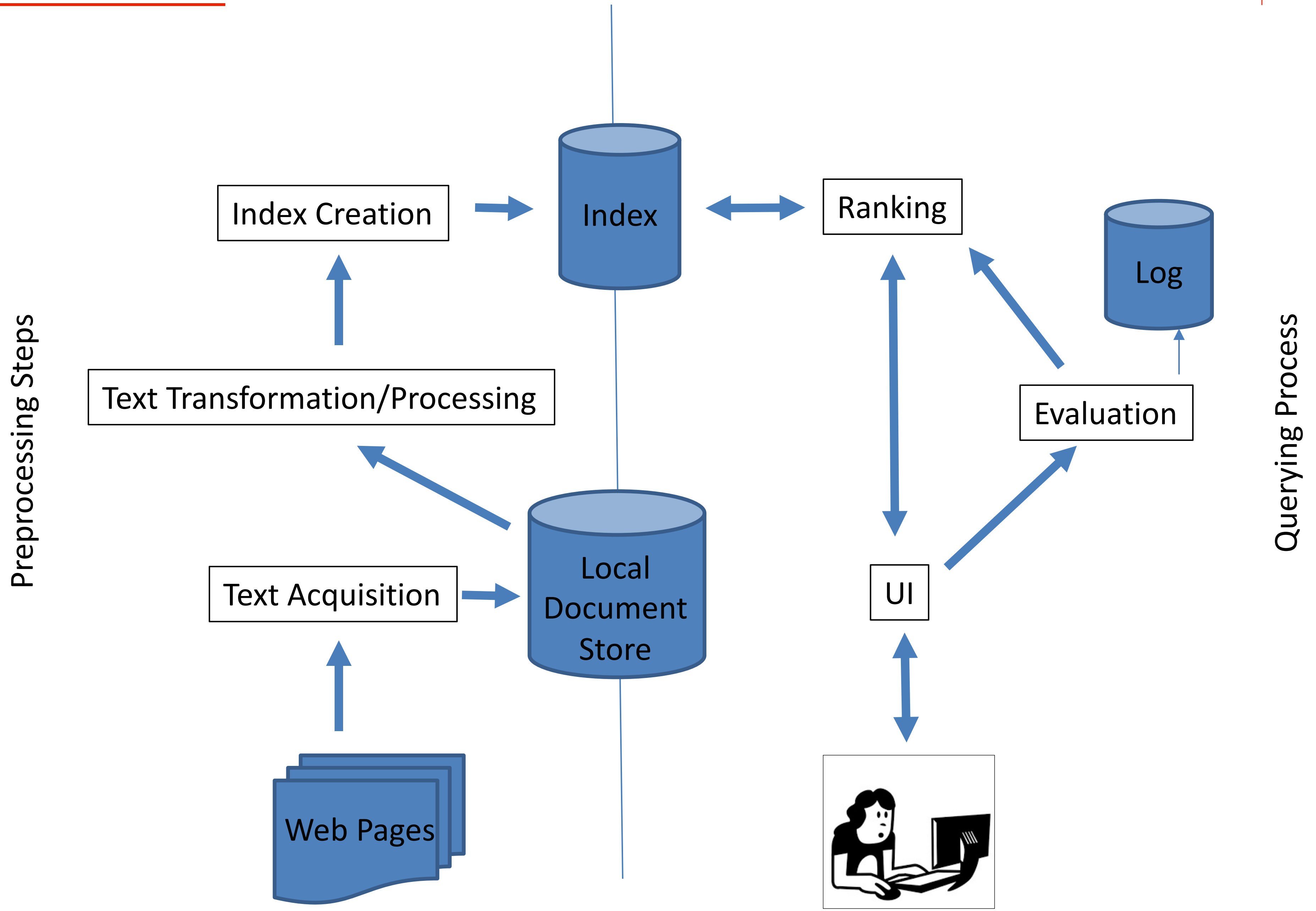
These course materials borrow, with permission, from those of Prof. Cristina Videira Lopes, Addison Wesley 2008, Chris Manning, Pandu Nayak, Hinrich Schütze, Heike Adel, Sascha Rothe, Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie. Powerpoint theme by Prof. André van der Hoek.

Search Engine Evaluation

Information Retrieval

These course materials borrow, with permission, from those of Prof. Cristina Videira Lopes, Addison Wesley 2008, Chris Manning, Pandu Nayak, Hinrich Schütze, Heike Adel, Sascha Rothe, Jerome H. Friedman, Robert Tibshirani, and Trevor Hastie

Architecture



Effectiveness Measures

A is set of relevant documents,
B is set of retrieved documents

	Relevant	Non-Relevant
Retrieved	$A \cap B$	$\overline{A} \cap B$
Not Retrieved	$A \cap \overline{B}$	$\overline{A} \cap \overline{B}$

$$\boxed{\begin{aligned} Recall &= \frac{|A \cap B|}{|A|} \\ Precision &= \frac{|A \cap B|}{|B|} \end{aligned}}$$

Recall : how well the search engine is doing at finding all the relevant documents for a query.
Precision : how well it is doing at rejecting non-relevant documents.

Classification Errors

- *False Positive* (Type I error)
  – a non-relevant document is retrieved

$$Fallout = \frac{|\overline{A} \cap B|}{|\overline{A}|}$$

- *False Negative* (Type II error)
  – Relevant documents that are not retrieved
  – 1- *Recall*

F Measure

- *Is there a single metric that could merge those values?*

F Measure

- *Is there a single metric that could merge those values?*

- *Harmonic mean* of recall and precision

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

  - harmonic mean emphasizes the importance of small values, whereas the arithmetic mean is more affected by outliers that are unusually large

F Measure

- *Is there a single metric that could merge those values?*

- *Harmonic mean* of recall and precision

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

  - harmonic mean emphasizes the importance of small values, whereas the arithmetic mean is more affected by outliers that are unusually large

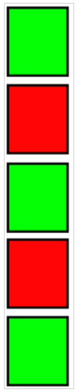
- More general form (weighted)

$$F_\beta = (\beta^2+1)RP/(R+\beta^2 P)$$

  - β is a parameter that determines relative importance of recall and precision

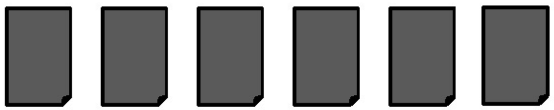
At which point do you compute the precision? Precision@K

- Set a rank threshold K
- Compute % relevant in top K
- Ignore documents ranked lower than K
- Ex:
  - Prec@3 of 2/3
  - Prec@4 of 2/4
  - Prec@5 of 3/5

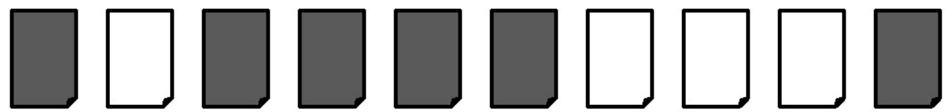


- In similar fashion we have Recall@K

Ranking Effectiveness

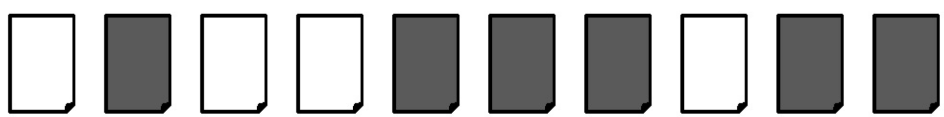
 = the relevant documents

Ranking #1



Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

Ranking #2

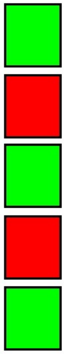


Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

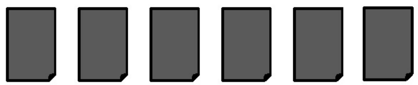
Summarizing a Ranking

- You can:
  - Calculating recall and precision at fixed rank positions
  - Calculating precision at standard recall levels, from 0.0 to 1.0
    - requires *interpolation*
  - **Averaging the precision values from the rank positions where a relevant document was retrieved**

Average Precision

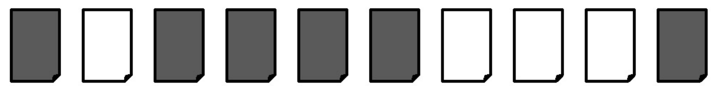
- Consider rank position of each ***relevant*** doc
  - $K_1, K_2, \ldots K_R$
- Compute Precision@K for each $K_1, K_2, \ldots K_R$
- Average precision = average of P@K
- Ex:  has AvgPrec of $$\frac{1}{3}\cdot\left(\frac{1}{1}+\frac{2}{3}+\frac{3}{5}\right)\approx 0.76$$

Average Precision



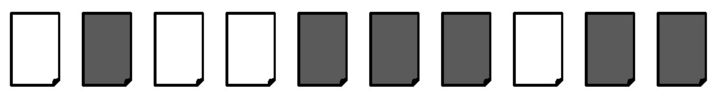
= the relevant documents

Ranking #1



Recall	0.17	0.17	0.33	0.5	0.67	0.83	0.83	0.83	0.83	1.0
Precision	1.0	0.5	0.67	0.75	0.8	0.83	0.71	0.63	0.56	0.6

Ranking #2

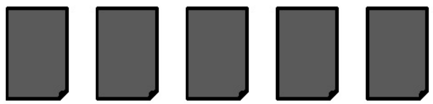


Recall	0.0	0.17	0.17	0.17	0.33	0.5	0.67	0.67	0.83	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.5	0.57	0.5	0.56	0.6

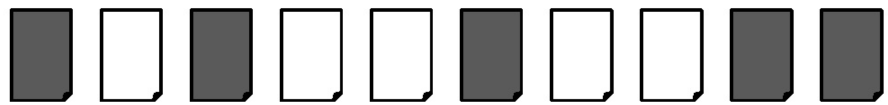
$$\text{Ranking \#1: } (1.0 + 0.67 + 0.75 + 0.8 + 0.83 + 0.6)/6 = 0.78$$

$$\text{Ranking \#2: } (0.5 + 0.4 + 0.5 + 0.57 + 0.56 + 0.6)/6 = 0.52$$

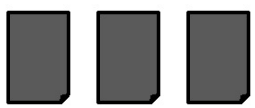
Averaging Across Queries

 = relevant documents for query 1

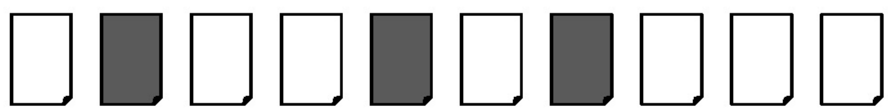
Ranking #1



Recall	0.2	0.2	0.4	0.4	0.4	0.6	0.6	0.6	0.8	1.0
Precision	1.0	0.5	0.67	0.5	0.4	0.5	0.43	0.38	0.44	0.5

 = relevant documents for query 2

Ranking #2



Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

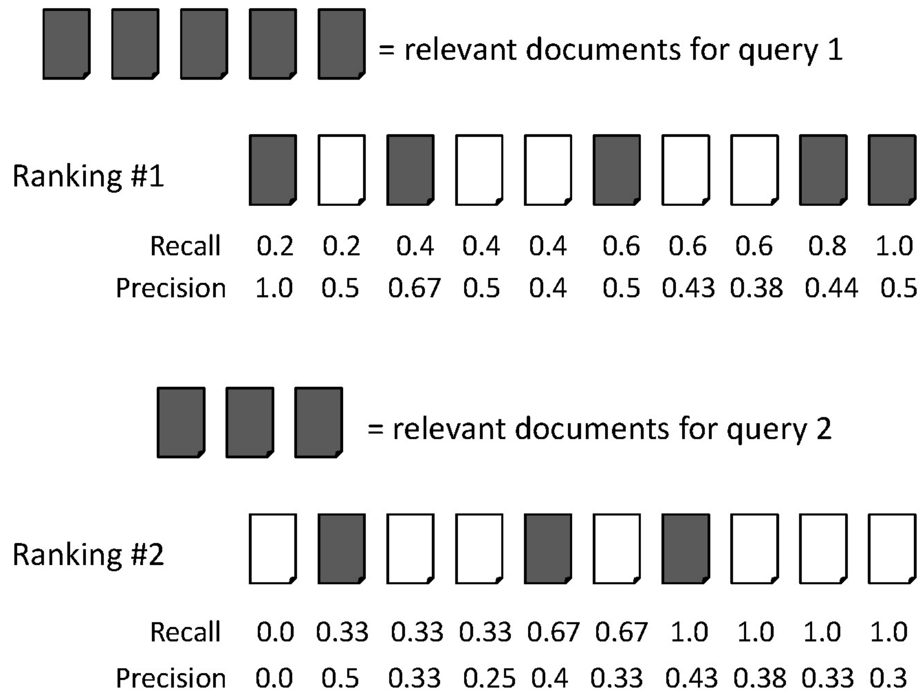
Averaging

- *Mean Average Precision* (MAP)
  - summarize rankings from multiple queries by averaging average precision
  - most commonly used measure in research papers

Averaging

- *Mean Average Precision* (MAP)
  - summarize rankings from multiple queries by averaging average precision
  - most commonly used measure in research papers
  - assumes user is interested in finding many relevant documents for each query
  - requires many relevance judgments in text collection

MAP



$$\textit{average precision query 1} = (1.0 + 0.67 + 0.5 + 0.44 + 0.5)/5 = 0.62$$
$$\textit{average precision query 2} = (0.5 + 0.4 + 0.43)/3 = 0.44$$

$$\textit{mean average precision} = (0.62 + 0.44)/2 = 0.53$$

Averaging

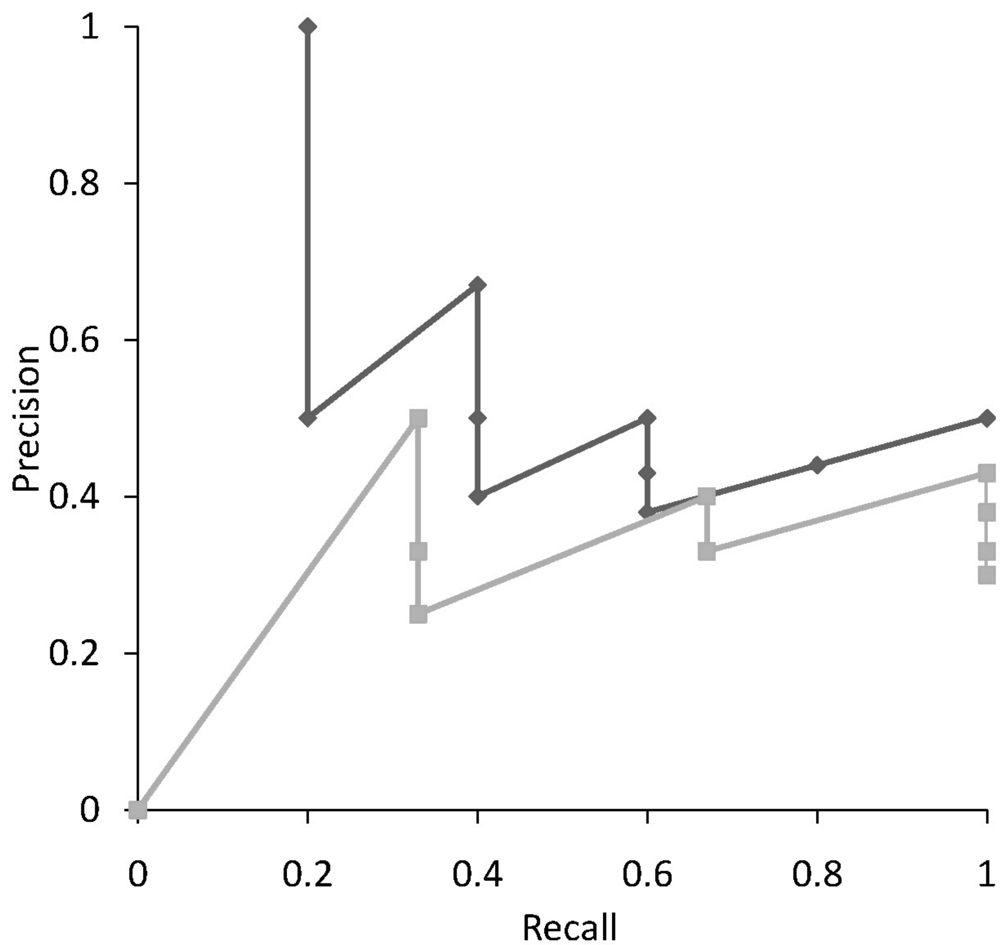
- *Mean Average Precision* (MAP)
  - summarize rankings from multiple queries by averaging average precision
  - most commonly used measure in research papers
  - assumes user is interested in finding many relevant documents for each query
  - requires many relevance judgments in text collection

- Recall-precision graphs are also useful summaries

Summarizing a Ranking

- You can:
  - Calculating recall and precision at fixed rank positions
  - **Calculating precision at standard recall levels, from 0.0 to 1.0**
    - **requires *interpolation***
  - Averaging the precision values from the rank positions where a relevant document was retrieved

Recall-Precision Graph



Interpolation

- Calculate precision at standard recall levels:

$$P(R) = \max\{P' : R' \geq R \wedge (R', P') \in S\}$$

Interpolation

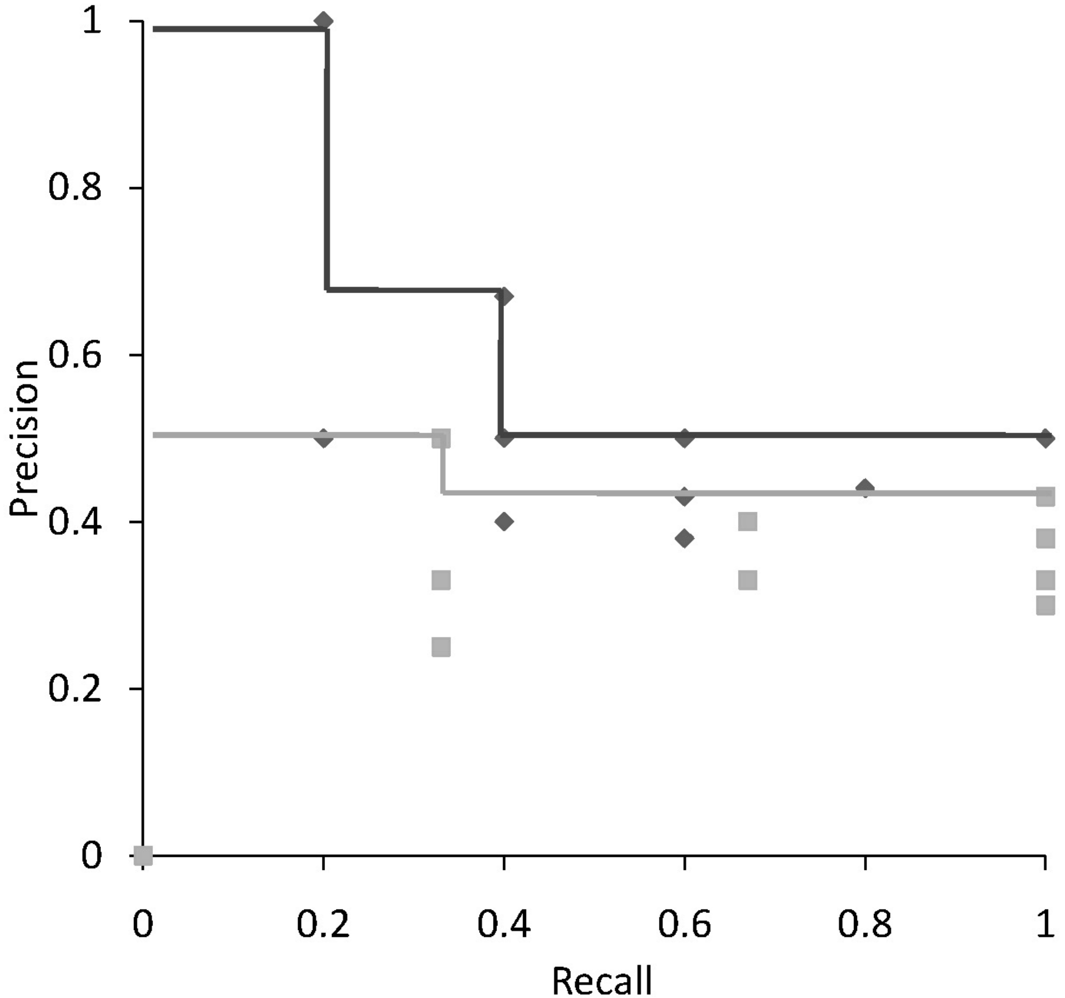
- Calculate precision at standard recall levels:

$$P(R) = \max\{P' : R' \geq R \wedge (R', P') \in S\}$$

  - where *S* is the set of observed (*R*,*P*) points

- Defines precision at any recall level as the *maximum* precision observed in any recall-precision point at a higher recall level
  - produces a step function
  - defines precision at recall 0.0

Interpolation

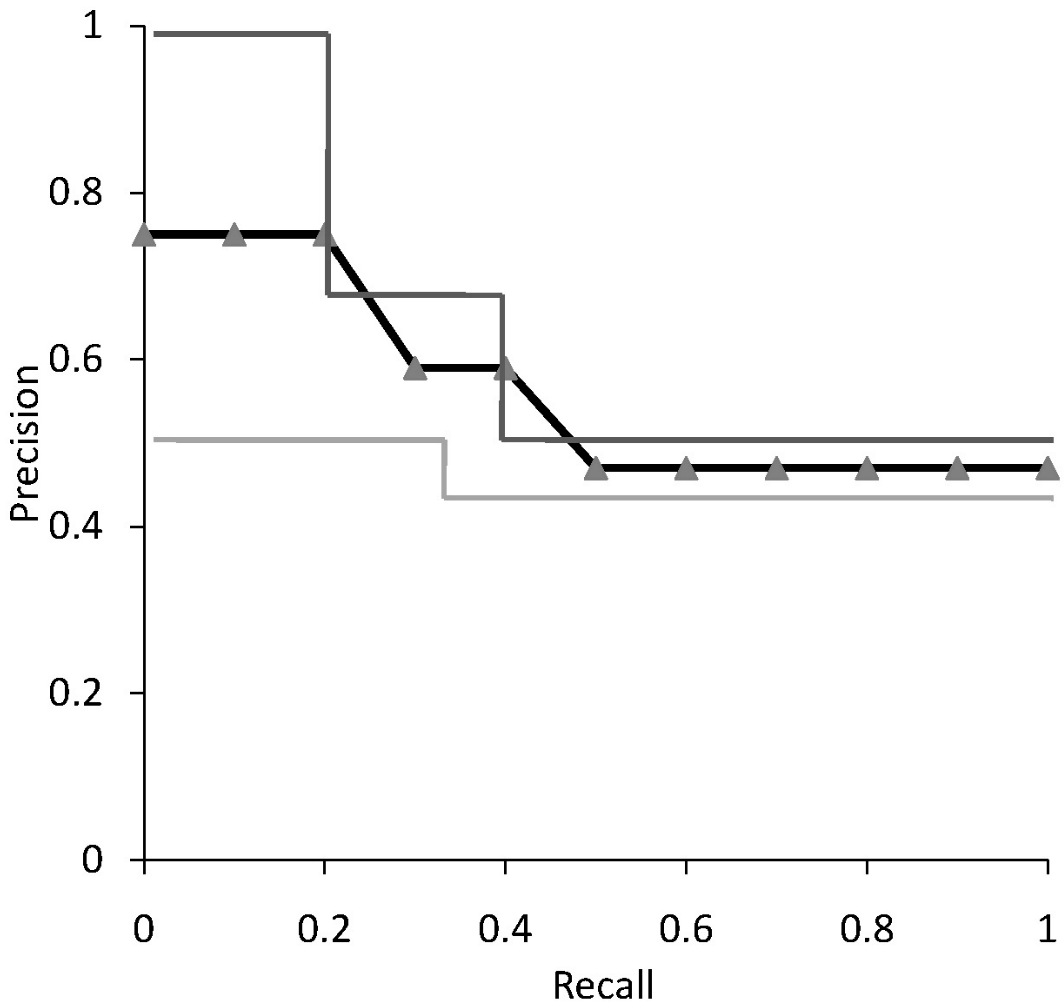


Average Precision at Standard Recall Levels

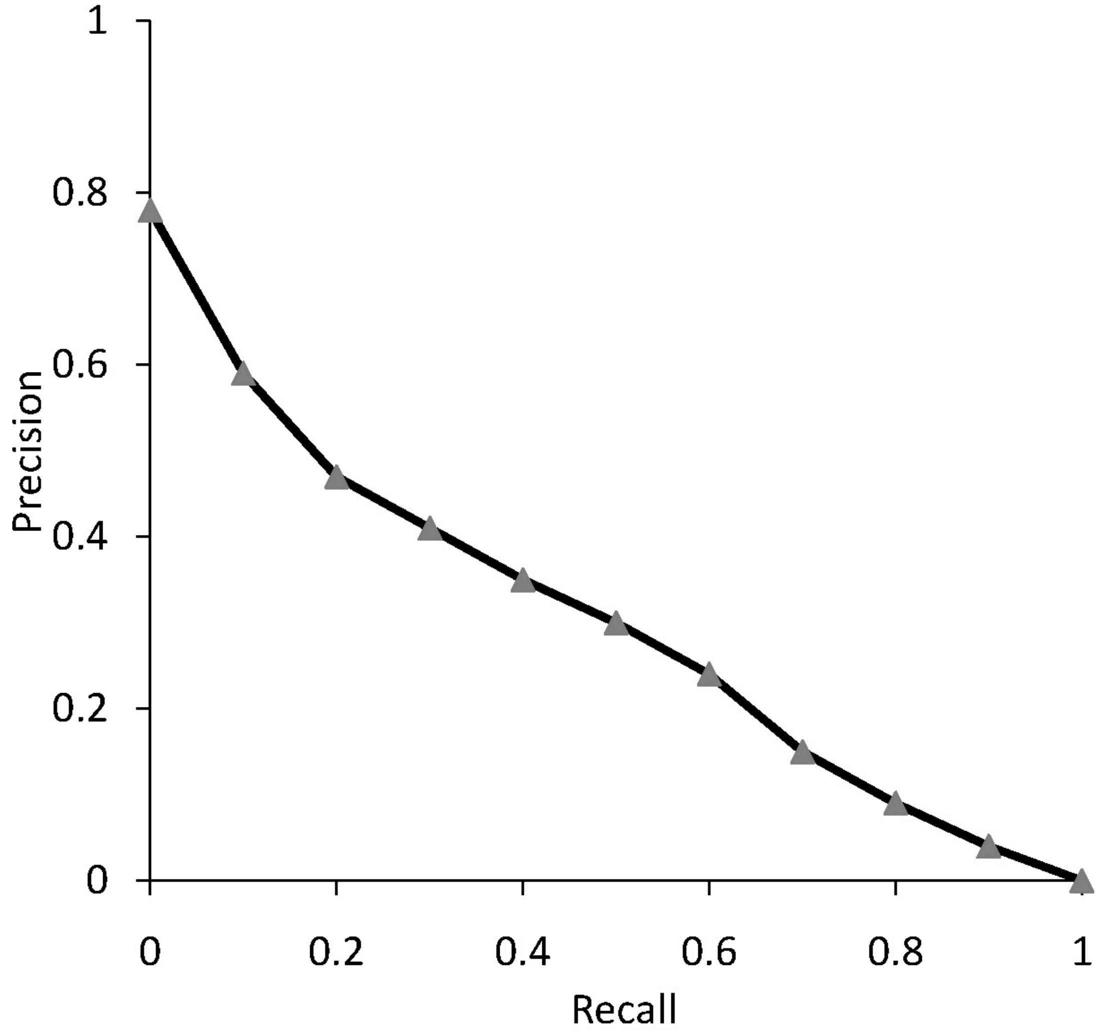
Recall	0.0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
Ranking 1	1.0	1.0	1.0	0.67	0.67	0.5	0.5	0.5	0.5	0.5	0.5
Ranking 2	0.5	0.5	0.5	0.5	0.43	0.43	0.43	0.43	0.43	0.43	0.43
Average	0.75	0.75	0.75	0.59	0.47	0.47	0.47	0.47	0.47	0.47	0.47

- Recall-precision graph plotted by simply joining the average precision points at the standard recall levels

Average Recall-Precision Graph



Graph for 50 Queries



Focusing on Top Documents

- Users tend to look at only the top part of the ranked result list to find relevant documents

- Some search tasks have only one relevant document
  - e.g., navigational search, question answering

- Recall not always appropriate
  - instead need to measure how well the search engine does at retrieving relevant documents at very high ranks

Focusing on Top Documents

- Precision at Rank R
  - R typically 5, 10, 20
  - easy to compute, average, understand
  - not sensitive to rank positions less than R

Focusing on Top Documents

- Precision at Rank R
  - R typically 5, 10, 20
  - easy to compute, average, understand
  - not sensitive to rank positions less than R
- Reciprocal Rank
  - reciprocal of the rank at which the first relevant document is retrieved
  - *Mean Reciprocal Rank (MRR)* is the average of the reciprocal ranks over a set of queries
  - very sensitive to rank position

Discounted Cumulative Gain

- Popular measure for evaluating web search and related tasks

- Two (very reasonable) assumptions:
  - Highly relevant documents are more useful than marginally relevant documents
  - 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

Discounted Cumulative Gain

- Uses *graded relevance* as a measure of the usefulness, or *gain,* from examining a document

Discounted Cumulative Gain

- Uses *graded relevance* as a measure of the usefulness, or *gain,* from examining a document

- Gain is accumulated starting at the top of the ranking and may be reduced, or *discounted*, at lower ranks

Discounted Cumulative Gain

- Uses *graded relevance* as a measure of the usefulness, or *gain,* from examining a document
- Gain is accumulated starting at the top of the ranking and 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

- What if relevance judgments are in a scale of [0,r]? r>2
- Cumulative Gain (CG) at rank n
  - Let the ratings of the n documents be r_1, r_2, …r_n (in ranked order)

$$CG = r_1+r_2+\ldots r_n$$

Discounted Cumulative Gain

- What if relevance judgments are in a scale of [0,r]? r>2

- Cumulative Gain (CG) at rank n
  - Let the ratings of the n documents be r_1, r_2, …r_n (in ranked order)

$$CG = r_1+r_2+\ldots r_n$$

- Discounted Cumulative Gain (DCG) at rank n

$$DCG = r_1 + r_2/\log_2 2 + r_3/\log_2 3 + \ldots\ r_n/\log_2 n$$

  - *We may use any base for the logarithm!*

Discounted Cumulative Gain

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

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

- Alternative formulation:

$$DCG_p = \sum_{i=1}^{p} \frac{2^{rel_i} - 1}{log(1+i)}$$

  - used by some web search companies
  - emphasis on retrieving highly relevant documents

DCG Example

- 10 ranked documents judged on 0-3 relevance scale:

  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

- DCG *(just add the discounted gains cumulatively!)*:

  3, 5, 6.89, 6.89, 6.89, 7.28, 7.99, 8.66, 9.61, 9.61

Normalized DCG

- DCG numbers are averaged across a set of queries at specific rank values
  - e.g., DCG at rank 5 is 6.89 and at rank 10 is 9.61
- 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 Example

- Perfect ranking:

  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

- 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

  – NDCG $\leq$ 1 at any rank position

Summary

- No single measure is the correct one for any application
  - choose measures appropriate for task
  - use a combination
  - shows different aspects of the system effectiveness

Summary

- No single measure is the correct one for any application
  - choose measures appropriate for task
  - use a combination
  - shows different aspects of the system effectiveness

- Analyze performance of individual queries

- Optimize the engine using user behavior (e.g. click logs)

Note: size of click logs

- How large is the click log?
  - bing search logs: 10+ TB/day
  - In existing publications:
    - [Silverstein+99]: 285M sessions
    - [Craswell+08]: 108k sessions
    - [Dupret+08] : 4.5M sessions (21 subsets * 216k sessions)
    - [Guo +09a] : 8.8M sessions from 110k unique queries
    - [Chapelle+09]: 58M sessions from 682k unique queries
    - [Liu+09a]: 0.26PB data from 103M unique queries