RETRIEVAL EVALUATION

Outline

- Intro
- Basics of probability and information theory
- Retrieval models
- Retrieval evaluation
  - Basic measures: precision, recall
  - Combined measures
  - Measures for integrating user ratings
  - Ranking measures
- Link analysis
- From queries to top-k results
- Social search

Different types of evaluation

- **Efficiency evaluation**

  Objective measurements
  - Answer time analysis
  - Space consumption analysis

- **Effectiveness evaluation**

  Subjective measurements (user satisfaction, surprise, etc.)
  - Quality of returned results in terms of relevance
  - Online testing with human evaluators

Relevance judgements

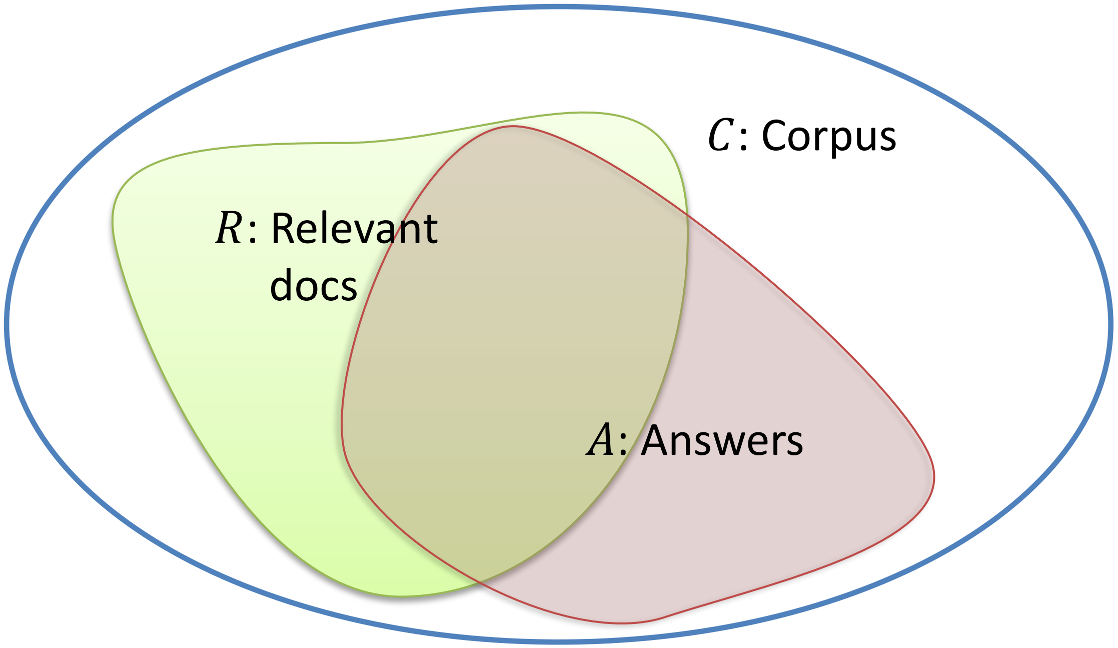
- Require knowledge (or even expertise) about good and poor results with regard to search need
- Are time-consuming and expensive
- Can not be done for every document in the corpus

- **Pooling** for **explicit relevance feedback**
  - Top-k documents returned by one (or multiple) search engine(s) are merged into a pool
  - Duplicates are removed
  - Human evaluators give binary relevance feedback

- **Query logs** for **implicit relevance feedback**
  - Contain tuples of the form (UserIP, query, URL, click, time, …)
  - Can be used to infer preferences

$$\left.\begin{array}{l} d_1 \text{ - no click} \\ d_2 \text{ - click} \\ d_3 \text{ - no click} \end{array}\right\} \Rightarrow \begin{array}{l} d_2 > d_1 \\ d_2 > d_3 \end{array}$$

Precision & recall



$$Precision = \frac{|R \cap A|}{|A|} \qquad Recall = \frac{|R \cap A|}{|R|}$$

- **False positive**: non-relevant document in the answer set (analogous for **true negative**)
- **False negative**: relevant document not in the answer set (analogous for **true positive**)
- Optimizing for precision

  $\Leftrightarrow$ increasing the probability of a result (in the answer set) being relevant
- Optimizing for recall

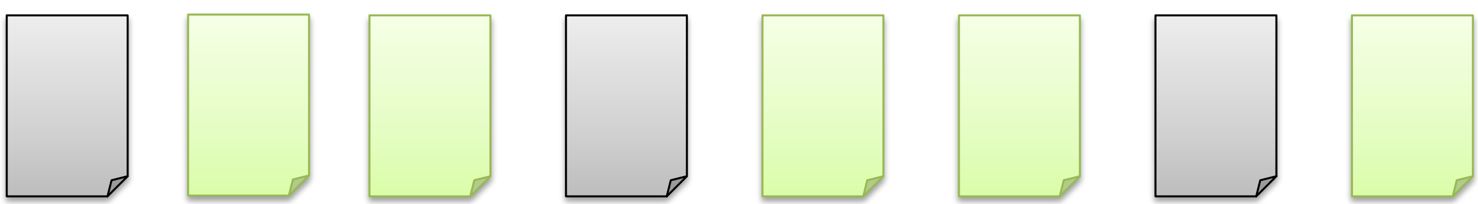
  $\Leftrightarrow$ increasing the probability of a relevant doc being in the answer set

Precision@k

- Consider top-k retrieved documents as the answer set
- Compute $precision@i$ ($P@i$) for all i on this answer set

- Toy example

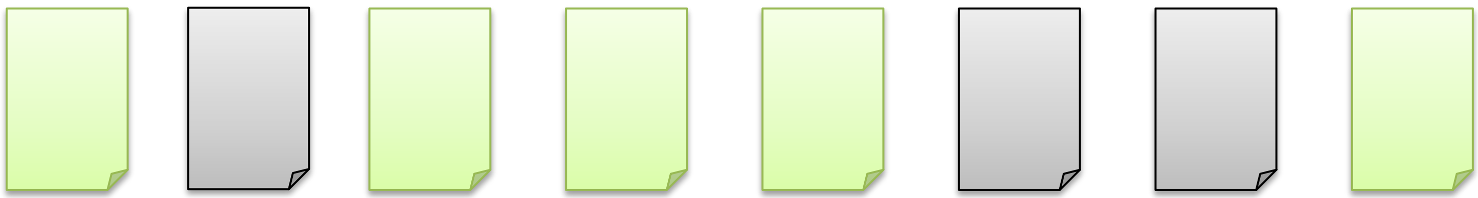
$P@i$ 1 0.5 0.66 0.75 0.8 0.66 0.57 0.625



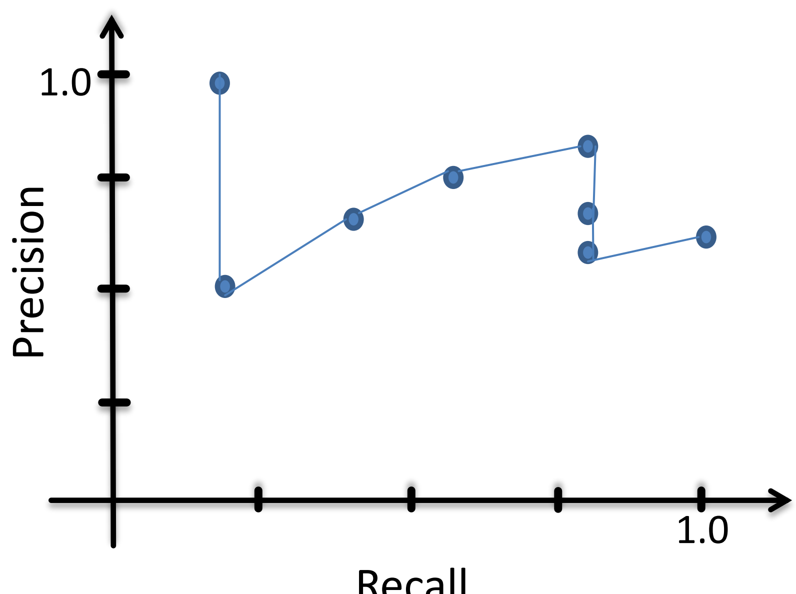
$P@i$ 0 0.5 0.66 0.5 0.6 0.66 0.57 0.625

Precision-recall curves (1)

- Toy example: suppose we have found all relevant documents

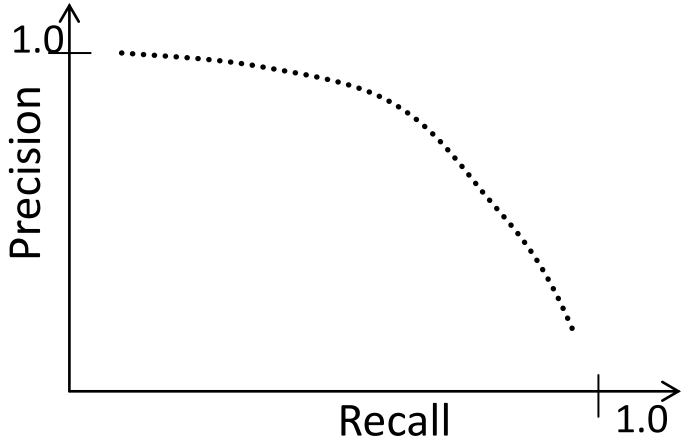


P@i	1	0.5	0.66	0.75	0.8	0.66	0.57	0.625
Recall	0.2	0.2	0.4	0.6	0.8	0.8	0.8	1

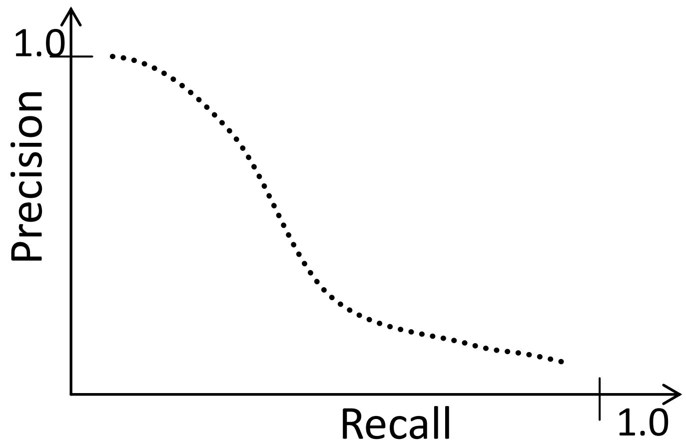


Precision-recall curves (2)

- Ideal performance



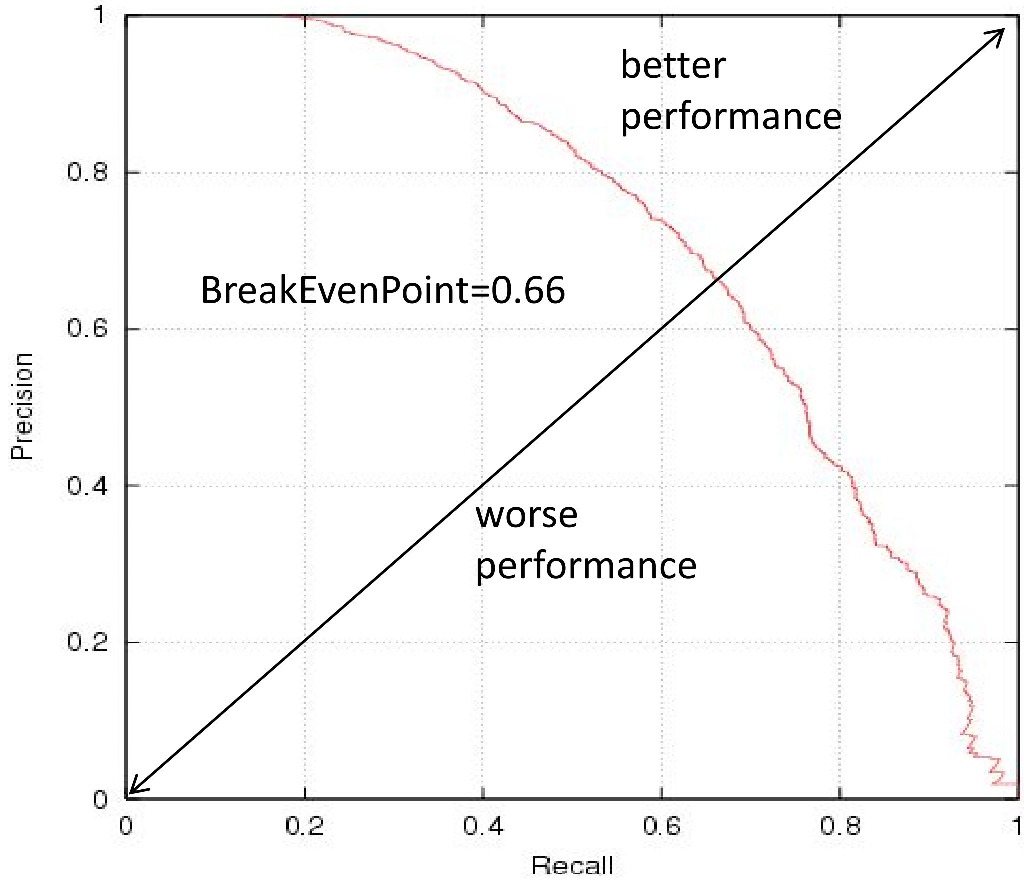
- Typical perfomance



Often reported as effectiveness measure: area under the curve (AUC)

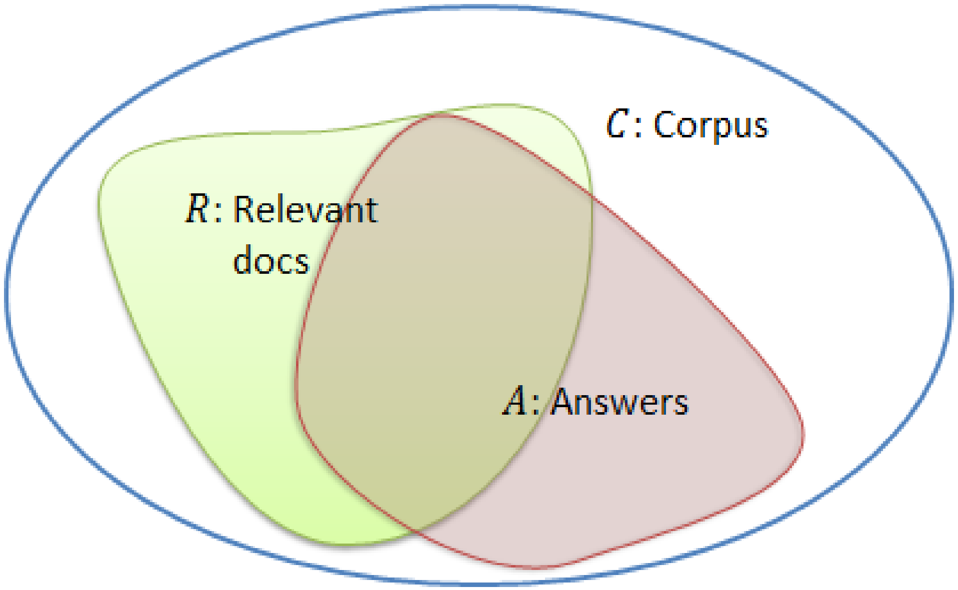
Break-even-point of precision and recall

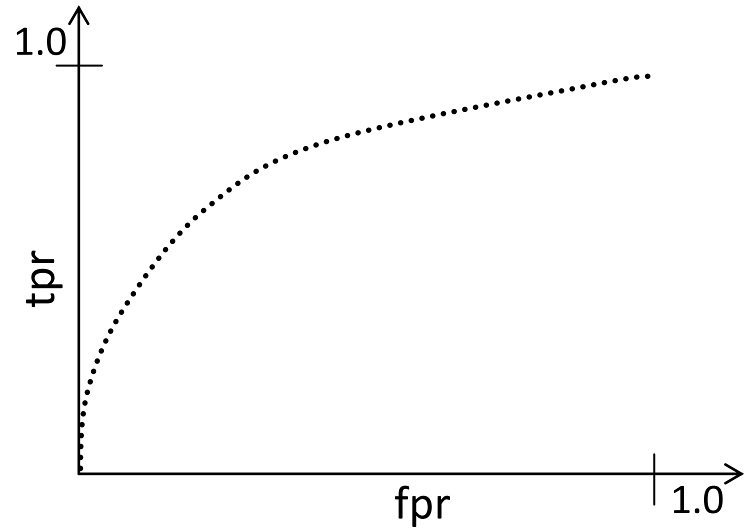
- $Precision = v = Recall$



ROC (receiver-operating characteristics) curves

- True-positive rate vs. false-positive rate





$$True\ Positive\ Rate(TPR) = Recall = \frac{|A \cap R|}{|R|}$$

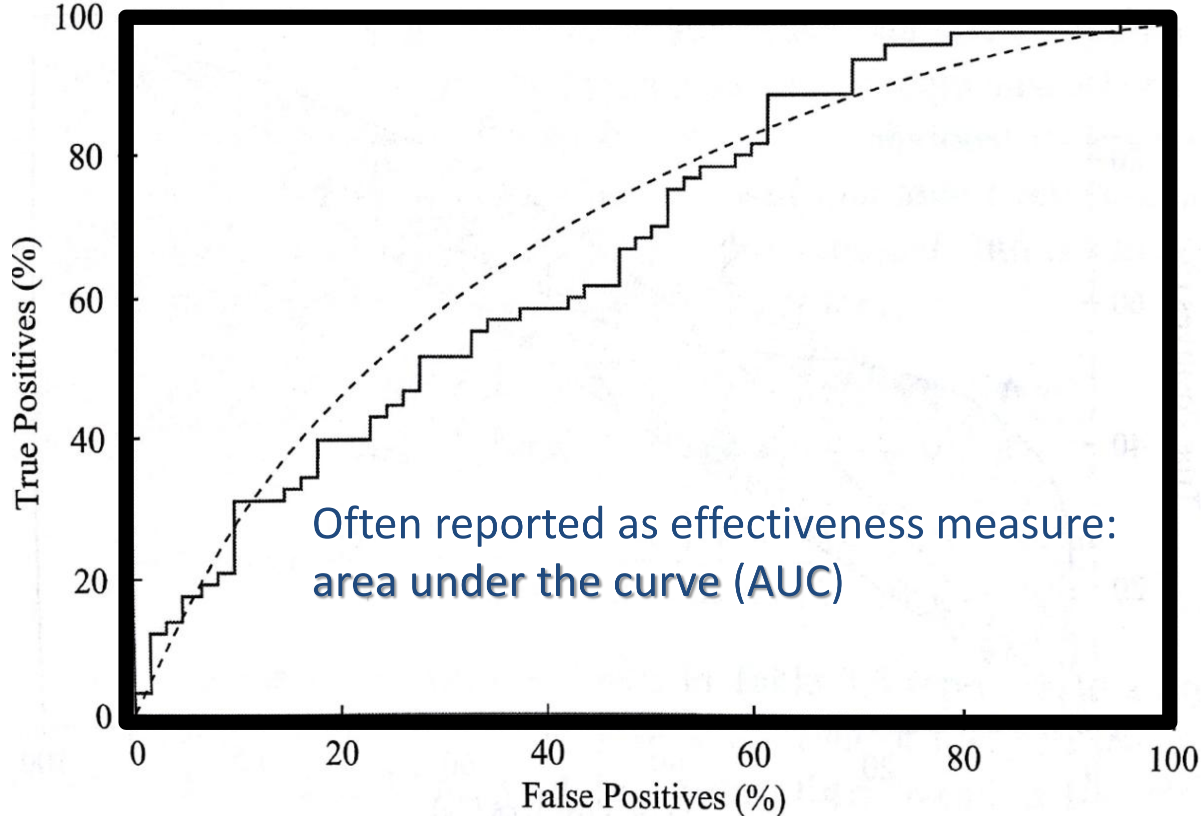
$$False\ Postitive\ Rate(FPR) = Fallout = \frac{|A \cap \bar{R}|}{|\bar{R}|}$$

ROC (receiver-operating characteristics) curves

- Plotting **true-positive rate** vs. **false-positive rate**

Ranking	Relevant
1	yes
2	yes
3	no
4	yes
5	yes
6	yes
7	yes
8	yes
9	no
10	yes
11	no
12	yes
13	yes
14	no
15	yes
16	no
17	yes
18	no
19	yes
...	...

Source: I.Witten, E. Frank, M. Hall: Data Mining – Practical Machine Learning Tools and Techniques



Often reported as effectiveness measure: **area under the curve (AUC)**

F-measure

- Combining precision and recall

$$F = \frac{1}{\beta \frac{1}{Precision} + (1-\beta)\frac{1}{Recall}}$$

- For $\beta = 0.5$ we get the harmonic mean:

$$F = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

- Mitigates the influence of large precision or recall values (to prevent bias towards large outlying values)

- Example: for $Recall = 0.2$ and $Precision = 0.9$, harmonic mean is $F \approx 0.33$

Macro evaluation

- Consider benchmark of n queries $q_1, \ldots, q_n$ and corresponding results
- For a user-oriented evaluation, **average precision, average recall**, and **average F-measure over all queries** are suitable measures
- **Macro precision**

$$Precision_{macro} = \frac{1}{n}\sum_{i=1}^{n} Precision(q_i)$$

- **Macro recall**

$$Recall_{macro} = \frac{1}{n}\sum_{i=1}^{n} Recall(q_i)$$

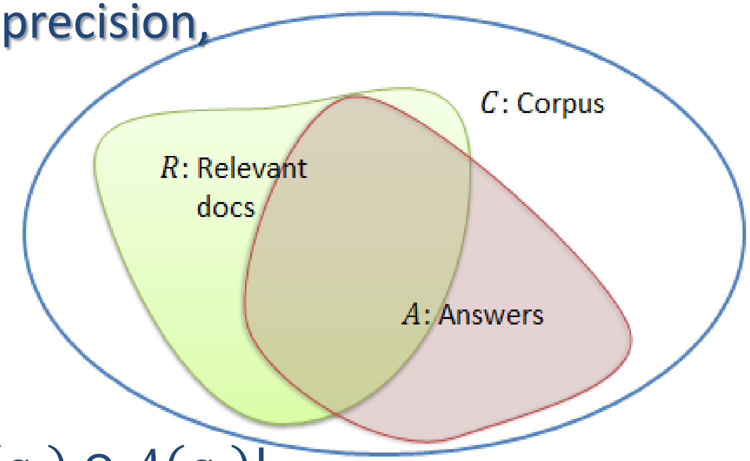
- **Macro F-measure**

$$F_{macro} = \frac{1}{n}\sum_{i=1}^{n} F(q_i)$$

Micro evaluation

- Consider benchmark of n queries $q_1, \ldots, q_n$ and corresponding results

- For a system-oriented evaluation, "**overall" precision**, "**overall" recall**, and "**overall" F-measure** are better suited.



- **Micro precision**

$$Precision_{micro} = \frac{\sum_{i=1}^{n} |R(q_i) \cap A(q_i)|}{\sum_{i=1}^{n} |A(q_i)|}$$

- **Micro recall:** analogously
- **Micro F-measure**

$$F_{micro} = \frac{2 \cdot Precision_{micro} \cdot Recall_{micro}}{Precision_{micro} + Recall_{micro}}$$

Interpolated average precision

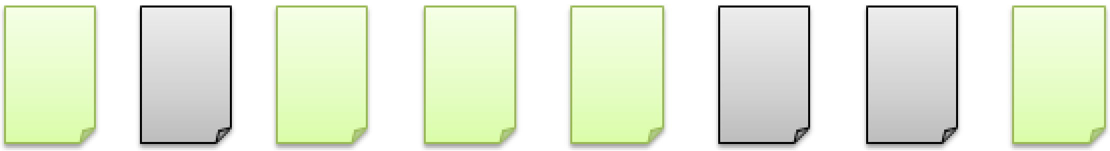
- Average precision for query q computed over different recall levels (for a given step width, e.g., 0.2)
- Let $Prec(Rl) = \max\{Prec' : Rl' \geq Rl \wedge (Prec', Rl')\ is\ observed\}$ (maximum precision observed in any recall-precision point at a higher or equal recall level)
- The interpolated average precision is defined as

$$IAP = \frac{1}{1/\Delta Rl} \sum_{i=1}^{1/\Delta Rl} Prec(i \cdot \Delta Rl)$$

- Upper bound of the area under the precision-recall curve

Interpolated average precision: example

- For recall levels at step width 0.2, compute the interpolated average precision

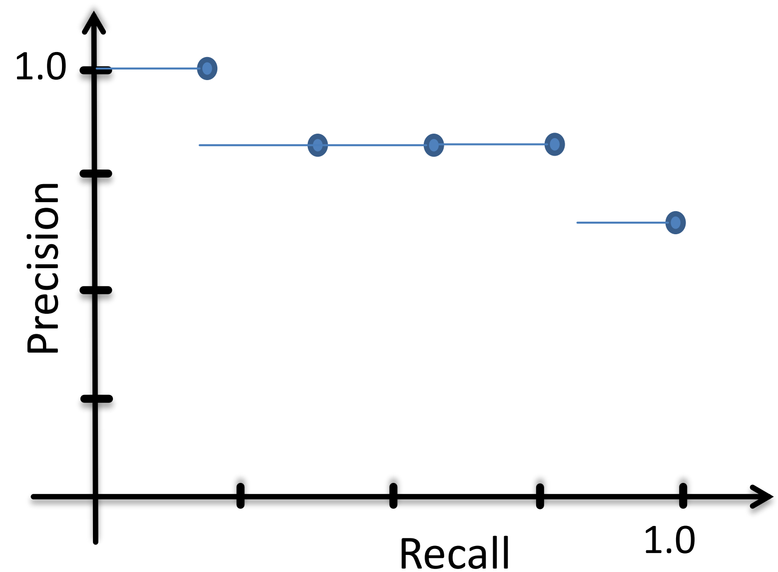


$P@i$	1	0.5	0.66	0.75	0.8	0.66	0.57	0.625
$Recall$	0.2	0.2	0.4	0.6	0.8	0.8	0.8	1

- Remember: $Prec(Rl) = \max\{Prec' : Rl' \geq Rl \wedge (Prec', Rl')\ is\ observed\}$

$$IAP = \frac{1}{1/\Delta Rl} \sum_{i=1}^{1/\Delta Rl} Prec(i \cdot \Delta Rl)$$

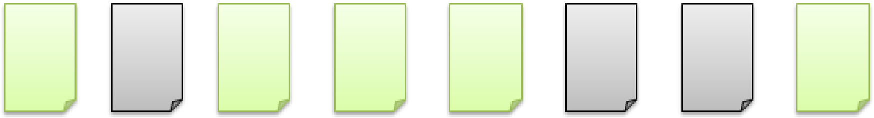
$$= 0.2 \cdot (1 + 3 \cdot 0.8 + 0.625)$$
$$= 0{,}805$$



Mean average precision (MAP)

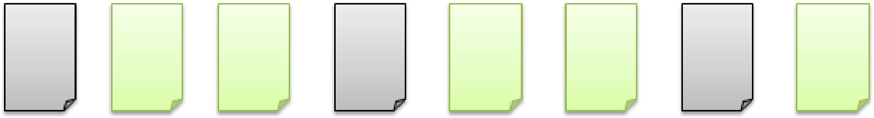
- Consider benchmark of n queries $q_1, \ldots, q_n$ and corresponding results

Answers for q_1:



$P@i$ 1 0.5 0.66 0.75 0.8 0.66 0.57 0.625

Answers for q_2:



$P@i$ 0 0.5 0.66 0.5 0.6 0.66 0.57 0.625

- Generally:

$$MAP = \frac{1}{n}\sum_{q_i} AvePrecision(q_i)$$

$$= \frac{1}{2}\left(\frac{(1 + 0.66 + 0.75 + 0.8 + 0.625)}{5} + \frac{(0.5 + 0.66 + 0.6 + 0.66 + 0.625)}{5}\right)$$

- Other possibility:

$$MAPI = \frac{1}{n}\sum_{q_i} IAP(q_i)$$

- Note: $MAPI$ corresponds to the **macro-average of per-query interpolated average precision** (with standard step width between recall levels 0.01)

Mean reciprocal rank (MRR)

- How effectively does a search system retrieve the first relevant result?
- Consider queries $q_1, \dots, q_n$ and corresponding ranked result lists
- $frr(q_i)$ denotes the rank of the **first relevant result** for any q_i
- The **mean reciprocal rank** is defined as

$$MRR = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{frr(q_i)}$$

- Variations are possible (e.g., summand is 0 if $frr(q_i) > threshold$)

Discounted cumulative gain (DCG)

- Previous evaluation measures were based on binary relevance feedback (i.e., result is **either relevant or non-relevant**)

- Is it possible to **integrate ratings for degree of relevance into evaluation** of effectiveness?

- Consider query q with ranked results, where $res(i)$ stands for the result at rank i

$$DCG = \sum_i \frac{2^{rating(res(i))} - 1}{\log(1+i)}$$

where, for example

$$rating(res(i)) = \begin{cases} 0, & if\ res(i)\ is\ irrelevant \\ 1, & if\ res(i)\ is\ ok \\ 2, & if\ res(i)\ is\ relevant \end{cases}$$

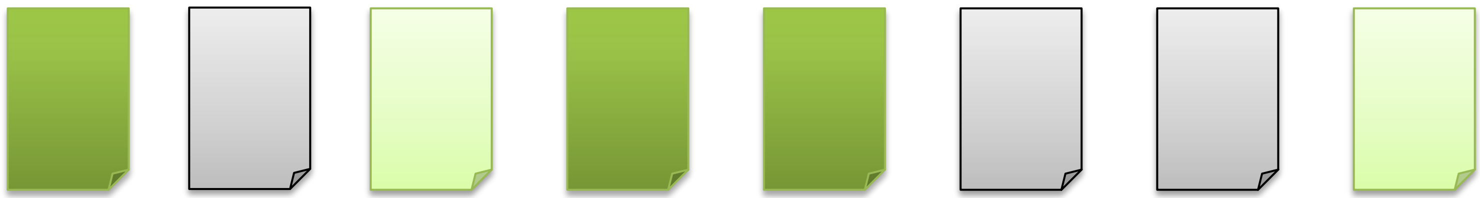
- Punishes result lists with many relevant results ranked lower than less relevant ones

Normalized discounted cumulative gain (NDCG)

- Normalize DCG by the DCG of the optimal ranking (of the query results)

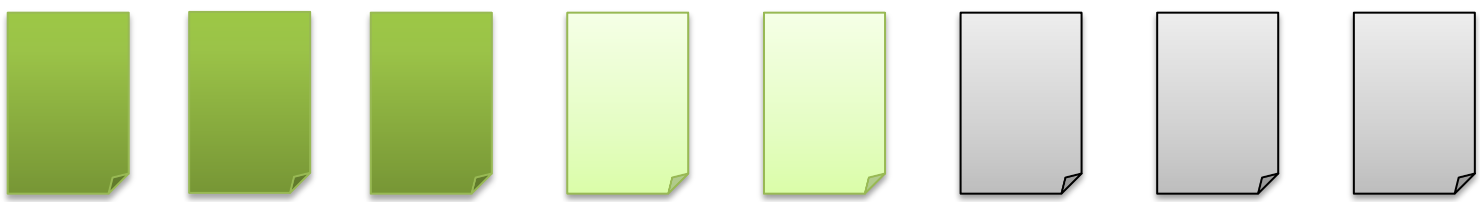
- Example

Actual ranking



$$DCG = \frac{2^2-1}{\log(2)} + \frac{2^0-1}{\log(3)} + \frac{2^1-1}{\log(4)} + \frac{2^2-1}{\log(5)} + \frac{2^2-1}{\log(6)} + \frac{2^0-1}{\log(7)} + \frac{2^0-1}{\log(8)} + \frac{2^1-1}{\log(9)}$$

Optimal ranking



$$oDCG = \frac{2^2-1}{\log(2)} + \frac{2^2-1}{\log(3)} + \frac{2^2-1}{\log(4)} + \frac{2^1-1}{\log(5)} + \frac{2^1-1}{\log(6)} + \frac{2^0-1}{\log(7)} + \frac{2^0-1}{\log(8)} + \frac{2^0-1}{\log(9)}$$

$$NDCG = \frac{DCG}{oDCG}$$

Result diversity

- Results should cover different aspects of the user's search need
- For ambiguous query (e.g., *Paris*), diversify results and hope that top-k results will satisfy the user's search need
- Important in sponsored search, e.g., *giant* could be a good term for "Giant Company Software", the movie "Giant", or Giant bikes
- General measure for result diversity

$$Div@k = \lambda \sum_{d \in top\_k} relevance(d) + (1-\lambda) \sum_{d,d' \in top\_k} dissimilarity(d,d')$$

Comparing rankings

- For two rankings π_1, π_2 of results to the same query
- $\mathit{Overlap@k}$ (similarity measure)

$$Overlap@k(\pi_1, \pi_2) = \frac{|top\_k(\pi_1) \cap top\_k(\pi_2)|}{k}$$

- **Footrule distance**
  - Let $S := top\_k(\pi_1) \cup top\_k(\pi_2)$

$$FRDist(\pi_1, \pi_2) = \frac{1}{|S|} \sum\nolimits_{e \in S} |\pi_1(e) - \pi_2(e)|$$

- **Kendall's τ measure** (distance measure)

$$K_\tau(\pi_1, \pi_2) = \frac{\left| \left\{ \begin{matrix} (a,b) \in S \times S \ | \ (a \neq b) \wedge \\ (\pi_1(a) > \pi_1(b) \wedge \pi_2(a) < \pi_2(b) \vee \pi_1(a) < \pi_1(b) \wedge \pi_2(a) > \pi_2(b)) \end{matrix} \right\} \right|}{|S|(|S|-1)}$$

- **Note:** $FRDist(\pi_1, \pi_2) \geq K_\tau(\pi_1, \pi_2) \geq \frac{1}{2} FRDist(\pi_1, \pi_2)$

Summary

- Basic measures
  - Precision (@k), recall
  - Precision-recall curves, break-even-point
  - ROC curves
  - Area under the curve (AUC)
- Combined measures
  - F-Measure
  - Micro, macro average (of precision, recall, F-measure)
  - Interpolated precision
  - Mean average precision (MAP)
- Measures for integrating user ratings
  - (Normalized) discounted cumulative gain ((N)DCG)
- Diversification
- Ranking measures
  - Overlap@k, Footrule distance, Kendall's τ