UNIVERSITY of DELAWARE UNIVERSITY of DE

Evaluation

(Search and Data Mining)

Hui Fang
Department of Electrical and Computer Engineering
University of Delaware

UNIVERSITY of DELAWARE UNIVERSITY of DE

The importance of Evaluation

- The ability to measure differences
  - How well does a search engine work?
  - Is search engine A better than search engine B?
- Evaluation drives what to research
  - Identify techniques that work and fail

UNIVERSITY of DELAWARE UNIVERSITY of DE

Evaluation Criteria

- Effectiveness
  - How accurate are the search results?
- Efficiency
  - How quickly can a user get the results?
- Usability
  - How useful is the system for real user tasks?

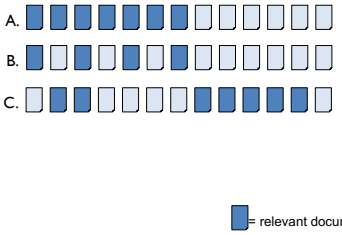
UNIVERSITY of DELAWARE UNIVERSITY of DE

Evaluation Methodology

- A test collection
  - documents
  - queries
  - relevance judgments
- Measure of effectiveness
  - A numeric score used to quantify the quality of the search results
  - Most common measures are based on *precision* and *recall*
- Common Practices
  - Use multiple measures to get different views of performance
  - Test with multiple collections

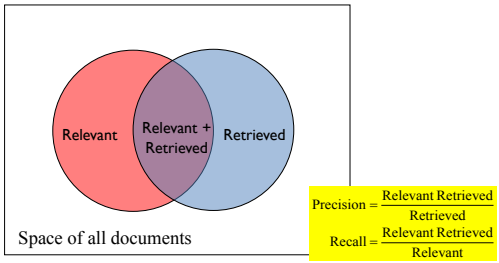
UNIVERSITY of DELAWARE UNIVERSITY of DE

Which search engine is the best?



UNIVERSITY of DELAWARE UNIVERSITY of DE

Set-based Measures



$$\text{Precision} = \frac{\text{Relevant Retrieved}}{\text{Retrieved}}$$

$$\text{Recall} = \frac{\text{Relevant Retrieved}}{\text{Relevant}}$$

Combining Precision and Recall: F-Measure

$$F = \frac{(\beta^2 + 1)PR}{\beta^2 P + R}$$

$$F_1 = \frac{1}{\frac{1}{2}(\frac{1}{R} + \frac{1}{P})} = \frac{2PR}{P+R}$$

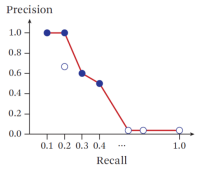
- F1 measure is popular (when beta = 1)
  - Harmonic mean of P and R
    - Inverse of average of their inverses
  - Heavily penalizes low values of P or R
    - Compared to standard average

Evaluating Ranking: Precision-Recall (PR) Curve

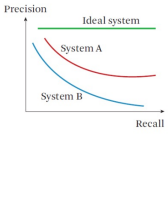
Evaluating Ranking: Precision–Recall (PR) Curve

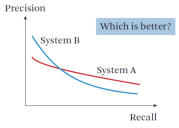
Total number of relevant documents in collection: 10

	Precision	Recall
D_1+	1/1	1/10
D_2+	2/2	2/10
D_3-	2/3	2/10
D_4-		
D_5+	3/5	3/10
D_6-		
D_7-		
D_8+	4/8	4/10
D_9-		
$D_{10}-$	?	10/10



Comparison of two PR curves





Summarize a Ranking: MAP

- For each query, given a ranked list of documents
  - Starting from the top of the ranking list, whenever we see a relevant document, we compute the precision up to that point.
    - For example, if the first relevant document is at the 2nd rank, the precision is ½.
  - If a relevant document never gets retrieved, we assume the precision corresponding to that document is zero.
  - We compute the average of the precision over all the relevant documents.
- Mean Average Precisions (MAP)
  - arithmetic mean average precision over a set of queries

Summarizing a ranking: Mean Average Precision (MAP)

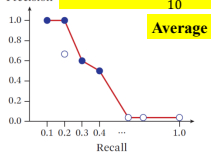
Evaluating Ranking: Precision–Recall (PR) Curve

Total number of relevant documents in collection: 10

	Precision	Recall
D_1+	1/1	1/10
D_2+	2/2	2/10
D_3-	2/3	2/10
D_4-		
D_5+	3/5	3/10
D_6-		
D_7-		
D_8+	4/8	4/10
D_9-		
$D_{10}-$	?	10/10

$$\frac{\frac{1}{1} + \frac{2}{2} + \frac{3}{5} + \frac{4}{8} + 0 + 0 + 0 + 0 + 0 + 0}{10}$$

Average precision



- Quiz 1
  - There are 10 relevant documents in the collection. A search engine return 5 documents, and 2 of them are relevant.
  - What is precision?
  - What is recall?

Normalized Discounted Cumulative Gain (nDCG)

- **Measure the total utility of the top K documents to a user**
  - Cumulative Gain (CG) at rank n
    - Let the ratings of the n documents be r1, r2, …rn (in ranked order), CG = r1+r2+…rn
- **Utility of a lowly ranked document is discounted**
  - Discounted Cumulative Gain (DCG) at rank n
    - DCG = r1 + r2/$\log_2$2 + r3/$\log_2$3 + … rn/$\log_2$n
- **Normalized to ensure comparability across queries**
  - Normalize DCG at rank n by the DCG value at rank n of the ideal ranking
  - The ideal ranking would first return the documents with the highest relevance level, then the next highest relevance level, etc

Evaluation with Multi-level judgements

	Gain	Cumulative gain	Discounted cumulative gain
D_1	3	3	3
D_2	2	3 + 2	3 + 2/log 2
D_3	1	3 + 2 + 1	3 + 2/log 2 + 1/log 3
D_4	1	3 + 2 + 1 + 1	…
D_5	3	…	
D_6	1		
D_7	1		
D_8	2		
D_9	1		
D_{10}	1		

$$\text{Normalized DCG} = \frac{\text{DCG@10}}{\text{IdealDCG@10}}$$

$$\text{DCG@10} = 3 + 2/\log 2 + 1/\log 3 + \cdots + 1/\log 10$$

$$\text{IdealDCG@10} = 3 + 3/\log 2 + 3/\log 3 + \cdots + 3/\log 9 + 2/\log 10$$

Relevance level: r = 1 (non-relevant), 2 (marginally relevant), 3 (very relevant)

Assume: there are 9 documents rated "3" in total in the collection