Search Evaluation

Tao Yang

CS290N

Slides partially based on text book [CMS] [MRS]

Table of Content

- **Search Engine Evaluation**
- **Metrics for relevancy**
  - Precision/recall
  - F-measure
  - MAP
  - NDCG

Difficulties in Evaluating IR Systems

- **Effectiveness is related to the *relevancy* of retrieved items.**
- **Relevancy is not typically binary but continuous. Not easy to judge**
- **Relevancy, from a human standpoint, is:**
  - Subjective/cognitive: Depends upon user's judgment, human perception and behavior
  - Situational and dynamic:
    - Relates to user's current needs. Change over time.
  - E.g.
    - CMU. US Open. Etrade.
    - Red wine or white wine

Measuring user happiness

- **Issue: who is the user we are trying to make happy?**
- **<u>Web engine</u>: user finds what they want and return to the engine**
  - Can measure rate of return users
- **<u>eCommerce site</u>: user finds what they want and make a purchase**
  - Is it the end-user, or the eCommerce site, whose happiness we measure?
  - Measure time to purchase, or fraction of searchers who become buyers?

Aspects of Search Quality

- **Relevancy**
- **Freshness& coverage**
  - Latency from creation of a document to time in the online index. (Speed of discovery and indexing)
  - Size of database in covering data coverage
- **User effort and result presentation**
  - Work required from the user in formulating queries, conducting the search
  - Expressiveness of query language
  - Influence of search output format on the user's ability to utilize the retrieved materials.

System Aspects of Evaluation

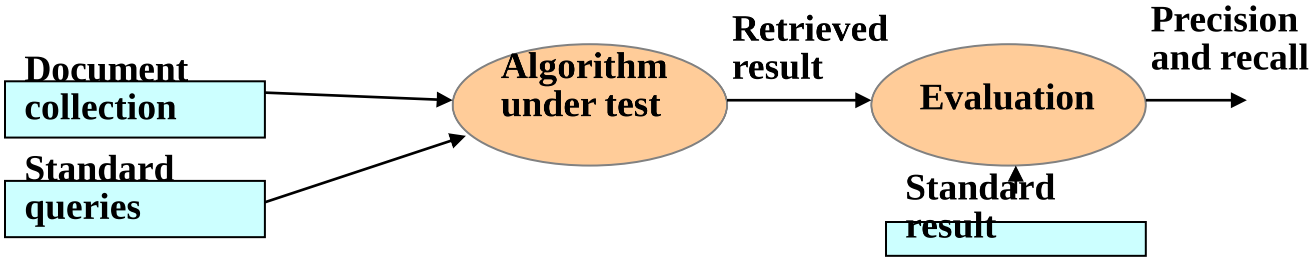
- **Response time:**
  - Time interval between receipt of a user query and the presentation of system responses.
  - Average response time
    - at different traffic levels (queries/second)
    - When # of machines changes
    - When the size of database changes
    - When there is a failure of machines
- **Throughputs**
  - Maximum number of queries/second that can be handled
    - without dropping user queries
    - Or meet Service Level Agreement (SLA)
      - For example, 99% of queries need to be completed within a second.
  - How does it vary when the size of database changes

System Aspects of Evaluation

- **Others**
  - Time from crawling to online serving.
  - Percentage of results served from cache
  - Stability: number of abnormal response spikes per day or per week.
  - Fault tolerance: number of failures that can be handled.
  - Cost: number of machines needed to handle
    - different traffic levels
    - host a DB with different sizes

Relevance benchmarks

- **Relevant measurement requires 3 elements:**
  1. A benchmark document collection
  2. A benchmark suite of queries
  3. Editorial assessment of query-doc pairs
     - Relevant vs. non-relevant
     - Multi-level: Perfect, excellent, good, fair, poor, bad



- **Public benchmarks**
  - TREC: http://trec.nist.gov/
  - Microsoft/Yahoo published learning benchmarks

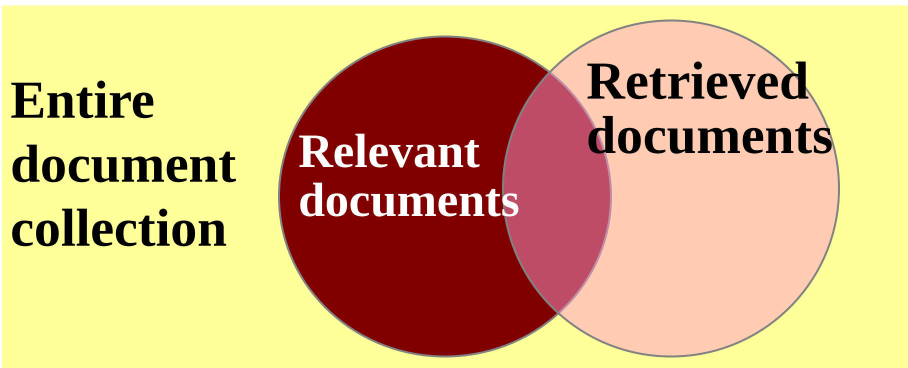
Unranked retrieval evaluation: Precision and Recall

- **Precision: fraction of retrieved docs that are relevant = P(relevant|retrieved)**
- **Recall: fraction of relevant docs that are retrieved = P(retrieved|relevant)**

	Relevant	Not Relevant
Retrieved	tp (True positive)	fp
Not Retrieved	fn	tn

- **Precision P = tp/(tp + fp)**
- **Recall R = tp/(tp + fn)**

Precision and Recall: Another View



	retrieved	not retrieved
irrelevant	retrieved & irrelevant	Not retrieved & irrelevant
relevant	retrieved & relevant	not retrieved but relevant

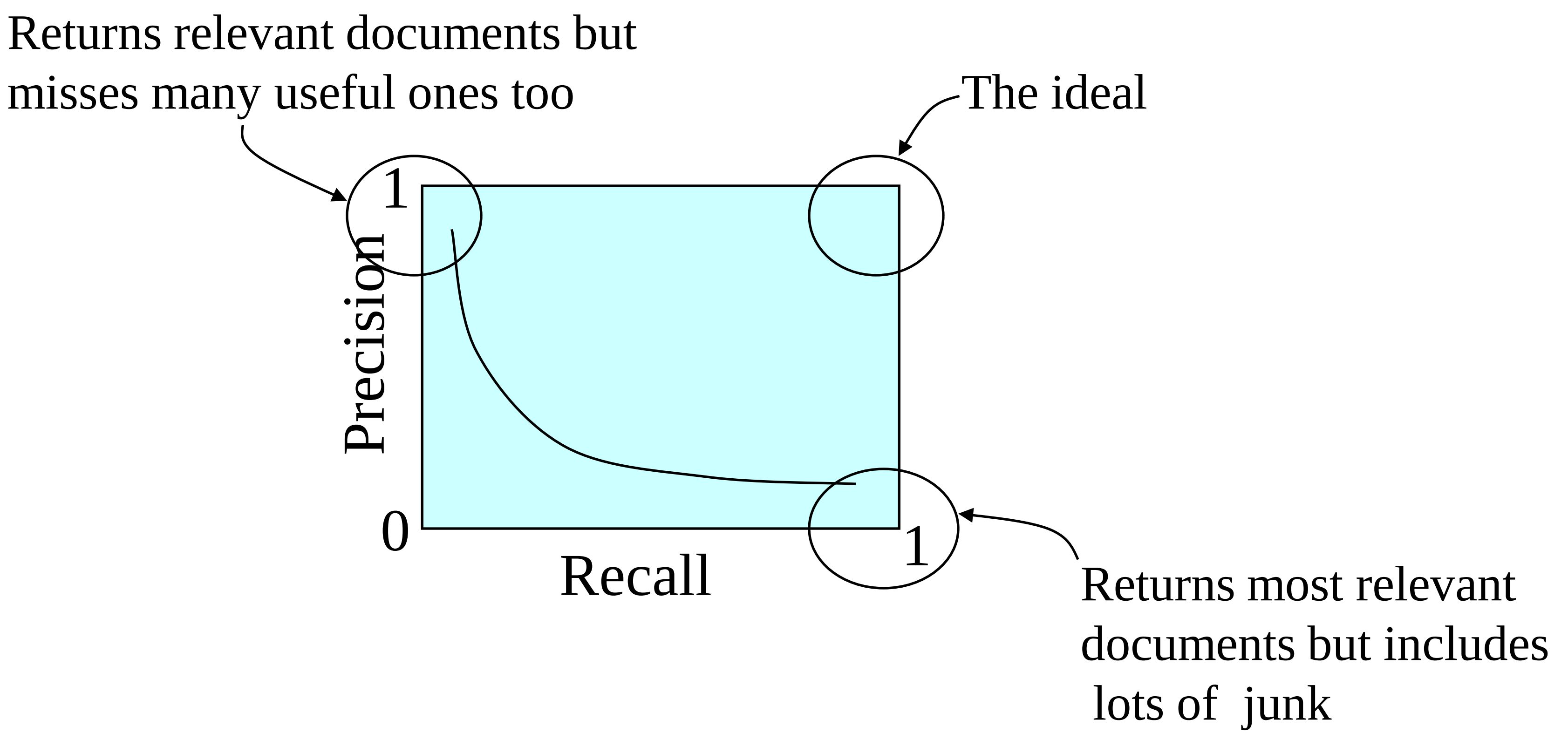
$$recall = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ relevant\ documents}$$

$$precision = \frac{Number\ of\ relevant\ documents\ retrieved}{Total\ number\ of\ documents\ retrieved}$$

Determining Recall is Difficult

- **Total number of relevant items is sometimes not available:**
  - Use queries that only identify few rare documents known to be relevant

Trade-off between Recall and Precision



F-Measure

- One measure of performance that takes into account both recall and precision.
- Harmonic mean of recall and precision:

$$F = \frac{2PR}{P+R} = \frac{2}{\frac{1}{R}+\frac{1}{P}}$$

E Measure (parameterized F Measure)

- **A variant of F measure that allows weighting emphasis on precision over recall:**

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

- **Value of β controls trade-off:**
  - β = 1: Equally weight precision and recall (E=F).
  - β > 1: Weight precision more.
  - β < 1: Weight recall more.

Computing Recall/Precision Points for Ranked Results

- For a given query, produce the ranked list of retrievals.
- Mark each document in the ranked list that is relevant according to the gold standard.
- Compute a recall/precision pair for each position in the ranked list that contains a relevant document.

R- Precision (at Position R)

- **Precision at the R-th position in the ranking of results for a query that has R relevant documents.**

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

R = # of relevant docs = 6

R-Precision = 4/6 = 0.67

Computing Recall/Precision Points: An Example

n	doc #	relevant
1	588	x
2	589	x
3	576	
4	590	x
5	986	
6	592	x
7	984	
8	988	
9	578	
10	985	
11	103	
12	591	
13	772	x
14	990	

Let total # of relevant docs = 6
Check each new recall point:

R=1/6=0.167; P=1/1=1

R=2/6=0.333; P=2/2=1

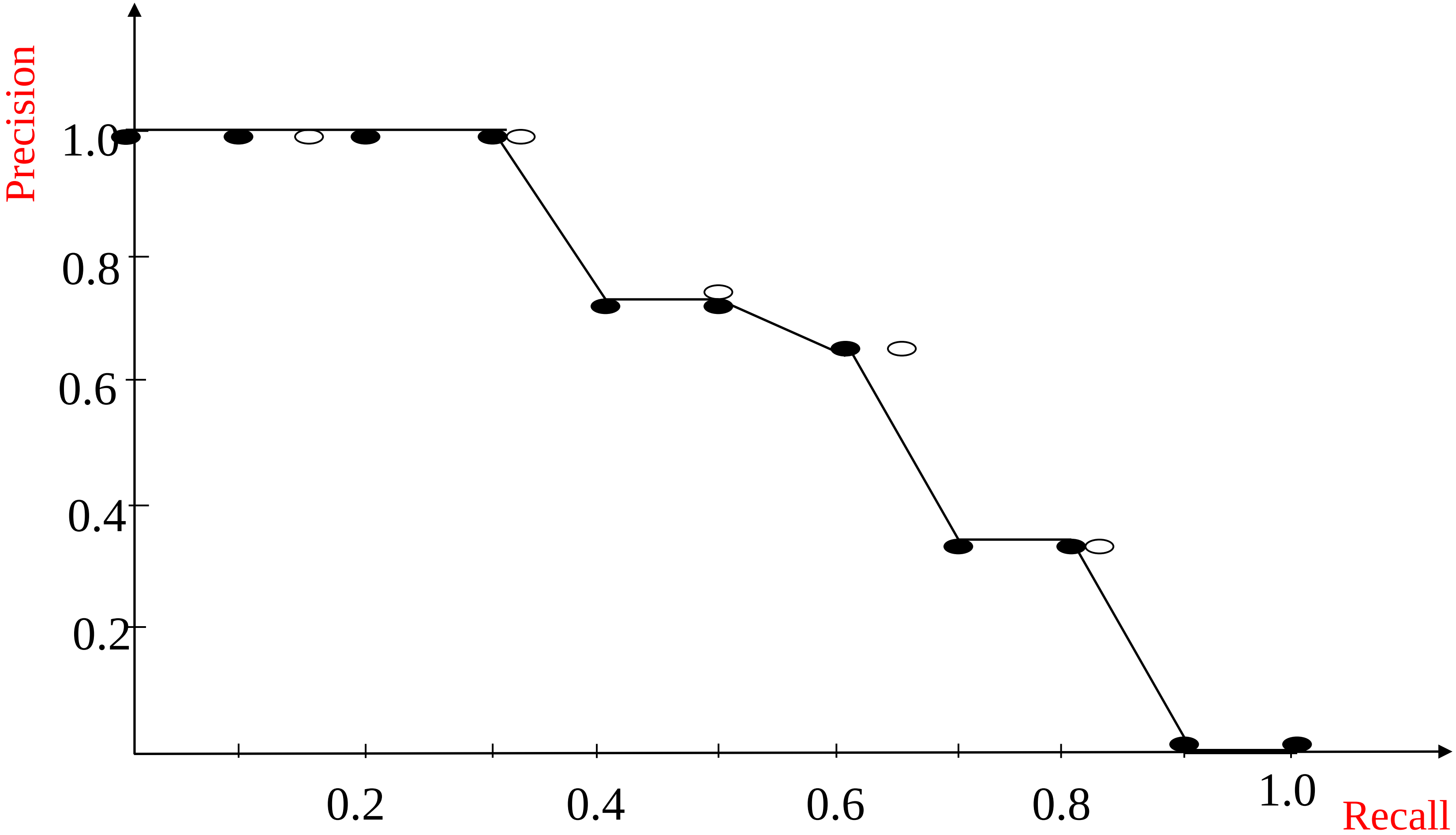
R=3/6=0.5; P=3/4=0.75

R=4/6=0.667; P=4/6=0.667

R=5/6=0.833; p=5/13=0.38

Missing one relevant document.
Never reach 100% recall

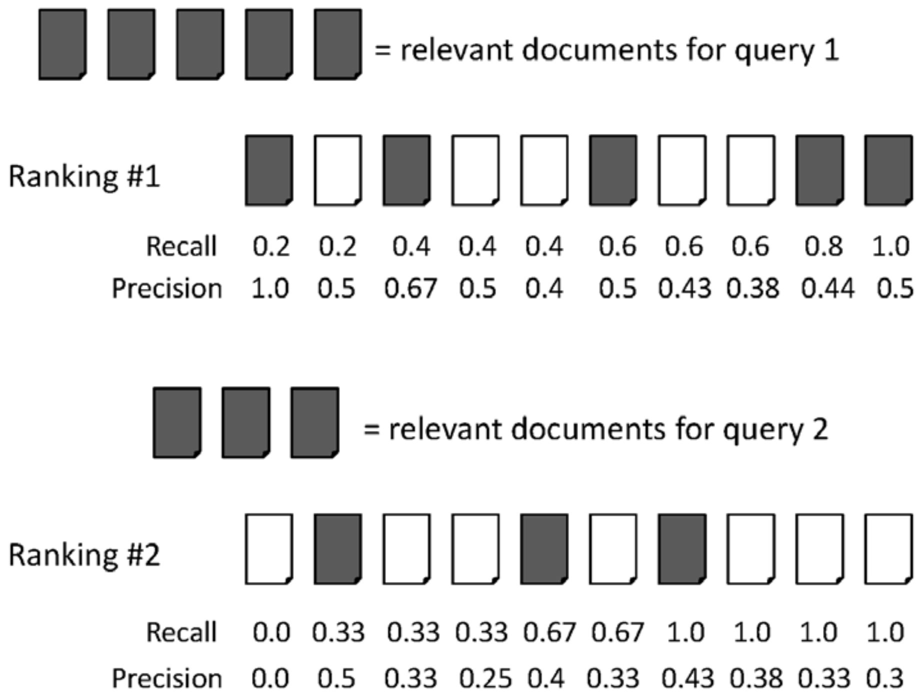
Interpolating a Recall/Precision Curve: An Example



Averaging across Queries: MAP

- ***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 Example:



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

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

Discounted Cumulative Gain

- **Popular measure for evaluating web search and related tasks**
- **Two assumptions:**
  - Highly relevant documents are more useful than marginally relevant document
    - Support relevancy judgment with multiple levels
  - 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 *discounted*, at lower ranks, e.g. 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*** **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@1, @2, etc:**

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

Normalized DCG

- **DCG values are often *normalized* by comparing the DCG at each rank with the DCG value for the *perfect ranking***
  - *Example:*
    - *DCG@5 = 6.89*
    - *Ideal DCG@5=9.75*
    - *NDCG@5=6.89/9.75=0.71*

- **NDCG numbers are averaged across a set of queries at specific rank values**

NDCG Example with Normalization

- **Perfect ranking:**

  3, 3, 3, 2, 2, 2, 1, 0, 0, 0
- **Ideal DCG@1, @2, …:**

  3, 6, 7.89, 8.89, 9.75, 10.52, 10.88, 10.88, 10.88, 10
- **NDCG@1, @2, …**
  - normalized 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