Search Evaluation

Tao Yang

CS293S Fall 2022

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

Table of Content

- **Search Engine Evaluation**
- **Metrics for relevance**
  - Precision/recall
  - F-measure
  - MAP
  - NDCG, MRR
- **Creating Test Collections for IR Evaluation**
- **A/B testing**

Aspects of Search Quality

- **Relevance**
- **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, the size of database changes, and 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

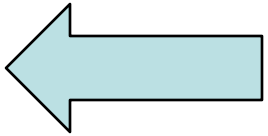
- **Others**
  - Time from crawling to online serving.
  - Percentage of results served from cache
  - Stability: number of abnormal response spikes per day or per week.
  - Fault tolerance: number of failures that can be handled.
  - Cost: number of machines needed to handle
    - different traffic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

Difficulties in Evaluating Relevance of IR Systems

- **Effectiveness is related to the *relevance* of matched items.**
  - Relevance is not typically binary but continuous.
  - Relevance, 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.
    - CMU. US Open
- **Measure happiness of users**
  - Web engine: A user finds what they want and uses again
    - Measure rate of return users
  - eCommerce site: user finds what they want and make a purchase
    - Measure time to purchase, or fraction of searchers who become buyers?

Table of Content

- **Search Engine Evaluation**
- **Metrics for relevance**
  - Precision/recall
  - F-measure
  - MAP
  - NDCG, MRR
- **Creating Test Collections for IR Evaluation**
  - A/B testing



Unranked retrieval evaluation: Precision and Recall

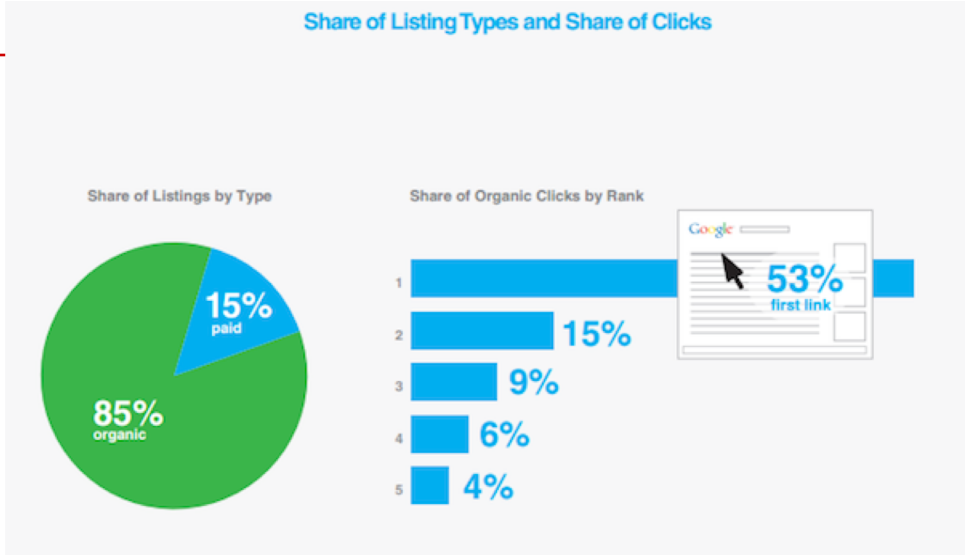
- **Precision: fraction of retrieved docs that are relevant = P(relevant|retrieved)**
- **Recall: fraction of relevant docs that are retrieved = P(retrieved|relevant)**

	Relevant	Not Relevant
Retrieved	tp (True positive)	fp (false positive)
Not Retrieved	fn (false negative)	tn (true negative)

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

Precision at Position R

- For a given query, a user only reviews top results. Ranking order in the results affects relevance.



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

Precision@6= 4/6 = 0.67

Precision at the R-th position in the ranking of results for a query measures % of relevant results by position R

Recall/Precision at a Position: 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

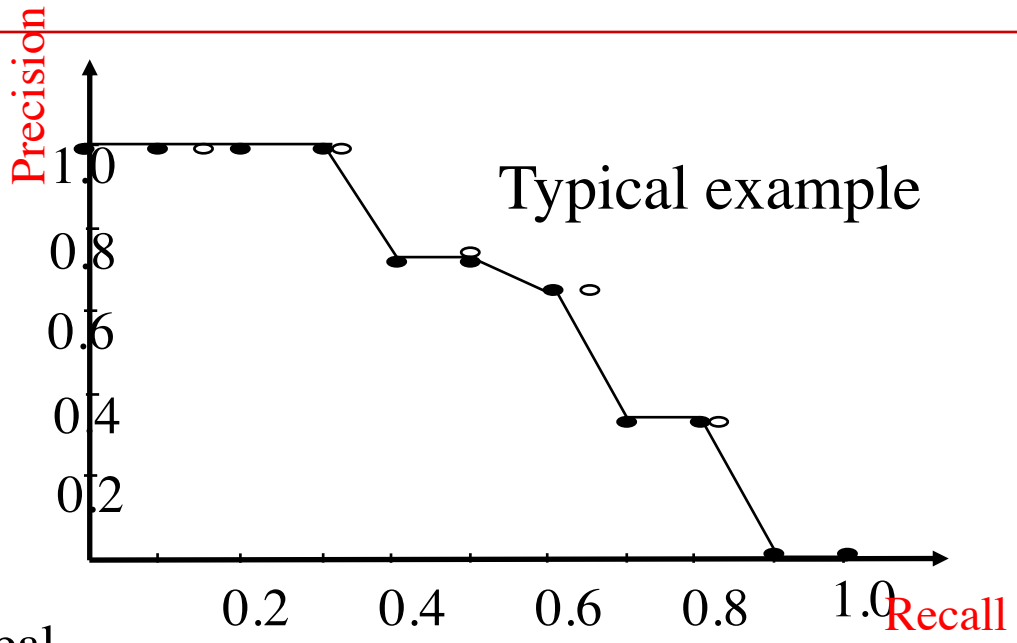
R=3/6=0.5; P=3/4=0.75

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

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

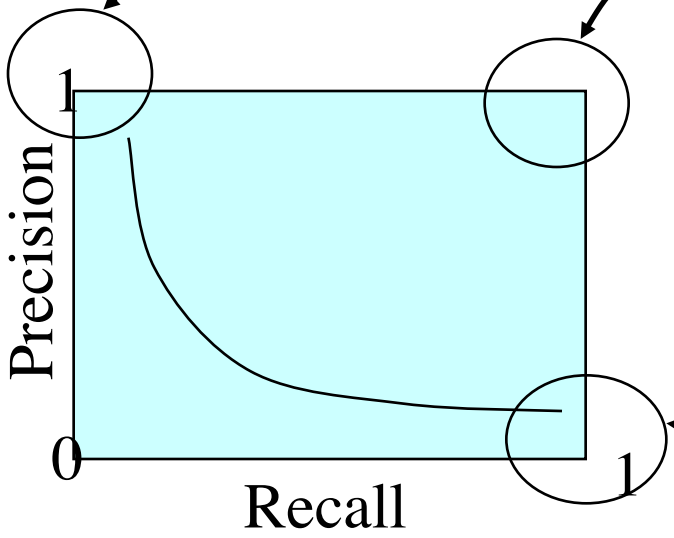
Missing one relevant document. Never reach 100% recall

Trade-off between Recall and Precision



Returns relevant documents but misses many useful ones too

The ideal



Returns most relevant documents but includes lots of junk

F-Measure

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

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

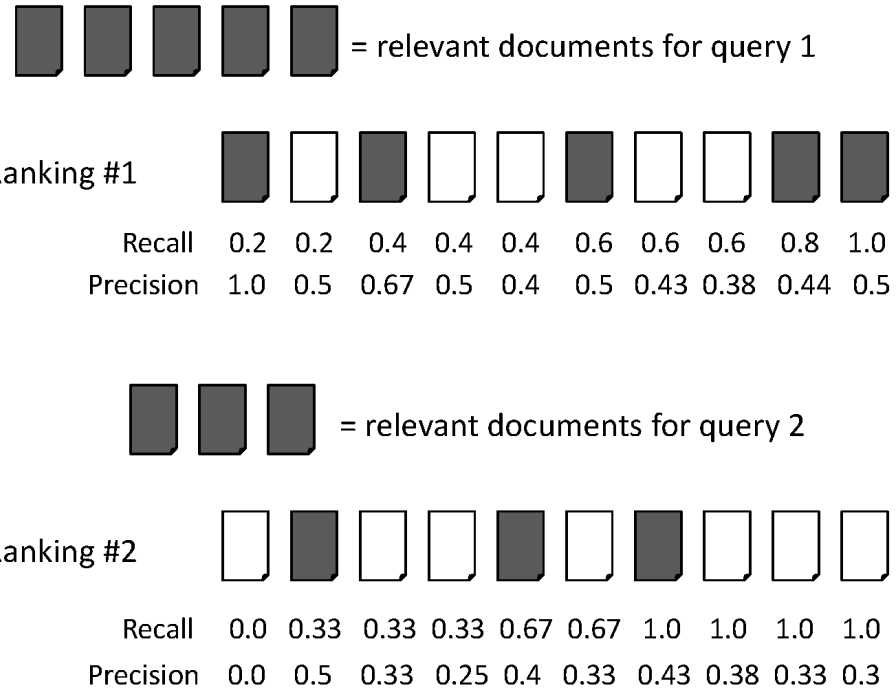
- A variant of F measure that allows weighting emphasis on precision over recall:

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

- 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.

Averaging across Queries: MAP

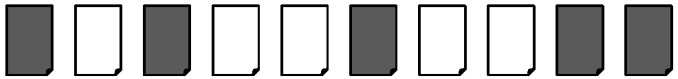
- ***How to evaluate when there are many queries***
- ***Mean Average Precision* (MAP)**
  - summarize rankings from multiple queries by averaging average precision
  - assumes user is interested in finding many relevant documents for each query



MAP Example:


= relevant documents for query 1

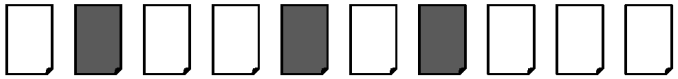
Ranking #1



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



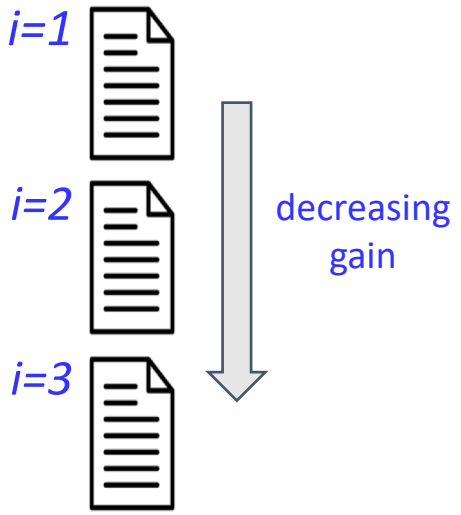
Recall	0.0	0.33	0.33	0.33	0.67	0.67	1.0	1.0	1.0	1.0
Precision	0.0	0.5	0.33	0.25	0.4	0.33	0.43	0.38	0.33	0.3

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

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

Evaluation Metrics: 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@p* is the total gain accumulated at a particular rank *p:***

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

  - Used in the rest of slides and our exercises
  - Alternative formulation:

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

  - 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, @3 etc:**

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

  DCG@3= sum of DCG at top 3 = 3+2+1.89 = 6.89

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*

$$\mathrm{nDCG}(R,q) = \frac{\mathrm{DCG}(R,q)}{\mathrm{IDCG}(R,q)}$$

- **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
- **My ranking:**

  3, 2, 3, 0, 0, 1, 2, 2, 3, 0
  - DCG@1, @2, etc:

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

Evaluation Metrics: Mean Reciprocal Rank

Simplified metric for relevant/irrelevant judgement labels while considering ranking positions

Mean reciprocal rank (MRR) for a set of queries Q:

$$\mathrm{MRR} = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{\mathrm{rank}_i}.$$

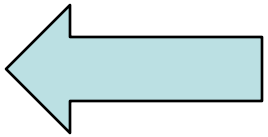
Rank of first relevant doc for the i-th query

Reciprocal rank (RR)

$$\mathrm{RR}(R, q) = \frac{1}{\mathrm{rank}_i}$$

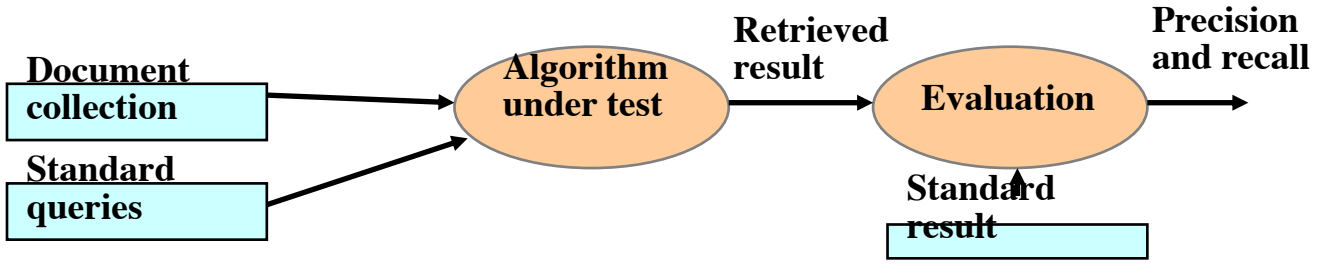
Table of Content

- **Search Engine Evaluation**
- **Metrics for relevance**
  - Precision/recall
  - F-measure
  - MAP
  - NDCG, MRR
- **Creating Test Collections for IR Evaluation**
- **A/B testing**



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

From document collections to test collections

- **Still need**
  - Test queries
  - Relevance assessments
- **Test queries**
  - Must be germane to docs available
  - Best designed by domain experts
  - Random query terms generally not a good idea
- **Relevance assessments**
  - Human judges, time-consuming
  - Are human panels perfect?

Popular Benchmarks for Relevance Evaluation

ClueWeb 09
TREC Robust04
MS MARCO Dev
MS MARCO Passage/Document Ranking
TREC Deep Learning 2019-2021 based on MS MARCO

Dataset	Domain	# Query	# Doc	Query Length	Doc Length	# judgements per query	Graded relevance
ClueWeb09	Web	150	50M	1-5	857	90	yes
Robust04	News	250	0.5M	1-4	479	70	yes
MS MARCO passage - dev	Q&A, Web	6980	8.8M	2-15	57	1	no
TREC DL 19		43				95	yes
TREC DL 20		54				66	yes
MSMARCO Doc - dev		5193	3.2M	2-15	1131	1	no

Datasets from The Text REtrieval Conference (TREC)

- TREC is co-sponsored by the National Institute of Standards and Technology (NIST) and U.S. Department of Defense since 1992
- **TREC Ad Hoc task from first 8 TRECs is standard IR task**
  - Human evaluation of pooled results returned
  - More recently other related things: Web track, HARD
- **A TREC query (TREC 5)**

<top>

<num> Number: 225

<desc> Description:

What is the main function of the Federal Emergency Management Agency (FEMA) and the funding level provided to meet emergencies? Also, what resources are available to FEMA such as people, equipment, facilities?

</top>

Test Collections: TREC Robust 04

- **0.5 million news articles**
  - from the Financial Times, the Federal Register 94, the LA Times, and FBIS (i.e. TREC disks 4&5, minus the Congressional Record). The Robust test set contains 250 topics:
- **250 topics selected in 2004 with query answer judgement labels**

ID: 336

Title: black bear attacks

Description: A relevant document would discuss the frequency of vicious black bear attacks worldwide and the possible causes for this savage behavior.

Narrative: It has been reported that food or cosmetics sometimes attract hungry black bears, causing them to viciously attack humans. Relevant documents would include the aforementioned causes as well as speculation preferably from the scientific community as to other possible causes of vicious attacks by black bears. A relevant document would also detail steps taken or new methods devised by wildlife officials to control and/or modify the savageness of the black bear.

Test Collections: MS MARCO

- Created in 2016, featuring 100,000 real Bing questions and a human generated answer, and gradually expanded to 1M questions
  - Collected for answering web queries with many question queries
- Web documents with many passages per each document
  - Passage collection
  - Document collection
- Sparse judgments
  - About 1 relevant passage per question

ID: 130510

Text: definition declaratory judgment

ID: 1131069

Text: how many sons robert kraft has

ID: 1131069

Text: when did rock n roll begin?

ID: 1103153

Text: who is thomas m cooley

TREC Deep Learning datasets

- Based on MS MARCO passages / docs
- Judgments of many answers per query for a limited number of queries

MS MARCO and Robust04: Stats

Documents and Mean / Median Lengths

Corpus	$\|\mathcal{C}\|$	$\overline{L}(\mathcal{C})$	$\widetilde{L}(\mathcal{C})$
MS MARCO passage corpus	8,841,823	56.3	50
MS MARCO document corpus	3,213,835	1131.3	584
Robust04 corpus (TREC disks 4&5)	528,155	548.6	348

Queries, Query Lengths, and Judgments

Dataset	$\|q\|$	$\overline{L}(q)$	$\|J\|$	$\|J\|/q$	$\|\text{Rel}\|/q$
MS MARCO passage retrieval (train)	502,939	6.06	532,761	1.06	1.06
MS MARCO passage retrieval (development)	6,980	5.92	7,437	1.07	1.07
MS MARCO passage retrieval (test)	6,837	5.85	-	-	-
MS MARCO document retrieval (train)	367,013	5.95	367,013	1.0	1.0
MS MARCO document retrieval (development	5,193	5.89	5,193	1.0	1.0
MS MARCO document retrieval (test)	5,793	5.85	-	-	-
TREC 2019 DL passage	43	5.39	9,260	215.4	95.4
TREC 2019 DL document	43	5.51	16,258	378.1	153.4
Robust04	249	(title) 2.7 (narr.) 15.3 (desc.) 40.2	311,410	1250.6	69.9

Standard relevance benchmarks: Others

- **GOV2**
  - Another TREC/NIST collection
  - 25 million web pages
  - Largest collection that is easily available
  - But still 3 orders of magnitude smaller than what Google/Yahoo/MSN index
- **ClueWeb**
  - Upto 1 billion web pages.
- **NTCIR**
  - East Asian language and cross-language information retrieval
- **Cross Language Evaluation Forum (CLEF)**
  - This evaluation series has concentrated on European languages and cross-language information retrieval.
- **Many others**

Can we avoid human judgment?

- **No**
  - But once we have test collections, we can reuse them (so long as we don't overtrain too badly)
  - Makes experimental work hard
    - Especially on a large scale
- **In some specific settings, can use proxies**
- **Search engines also use non-relevance-based measures.**
  - Clickthrough on first result
    - Not very reliable if you look at a single clickthrough … but pretty reliable in the aggregate.
  - Studies of user behavior in the lab
  - A/B testing

A/B testing

- Purpose: Test a single innovation (variation)
- Prerequisite: Website with large traffic
- Have most users use old system
  - Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an “automatic” measure
  - Clickthrough.
  - Now we can directly see if the innovation (variation) does improve user happiness.