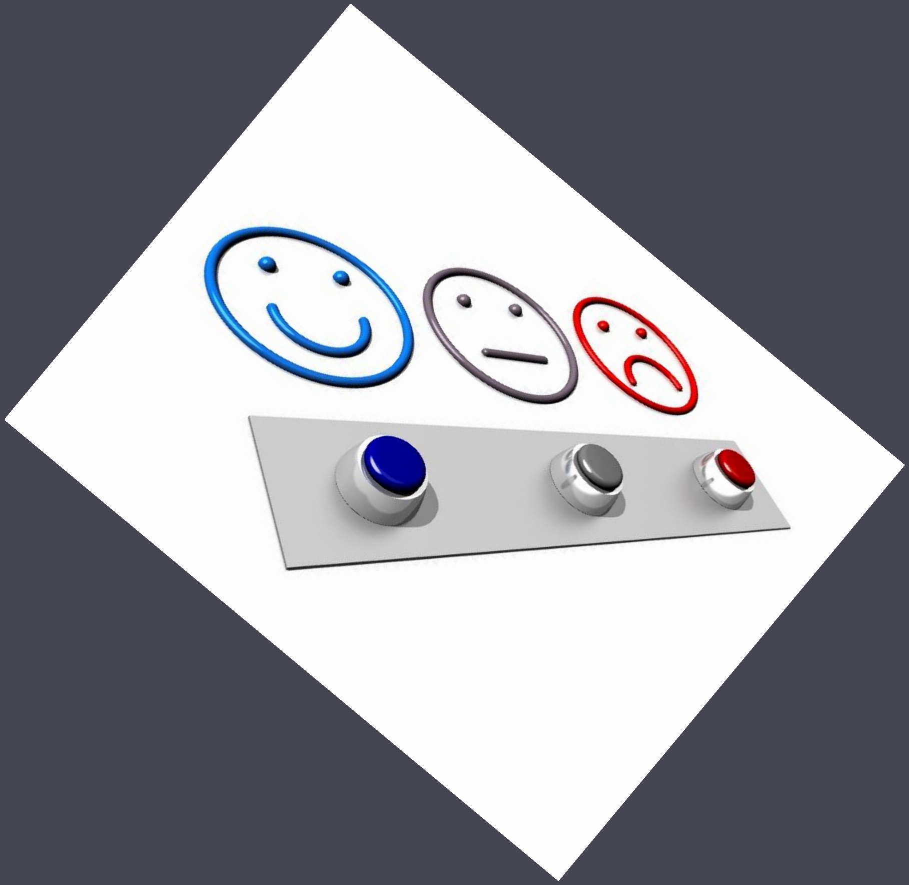


CS 597: SPECIAL TOPICS INFORMATION RETRIEVAL

Evaluation Strategies

Why Evaluate?

- Evaluation is key to building effective and efficient retrieval systems
  - Informally, effectiveness measures the ability of a system to find the *right information*, while efficiency measures how *quickly* things get done
- Effectiveness, efficiency, and cost are related
  - Efficiency and cost targets may impact effectiveness & vice versa
- Data for evaluation
  - Online versus benchmarks

Online Experiments

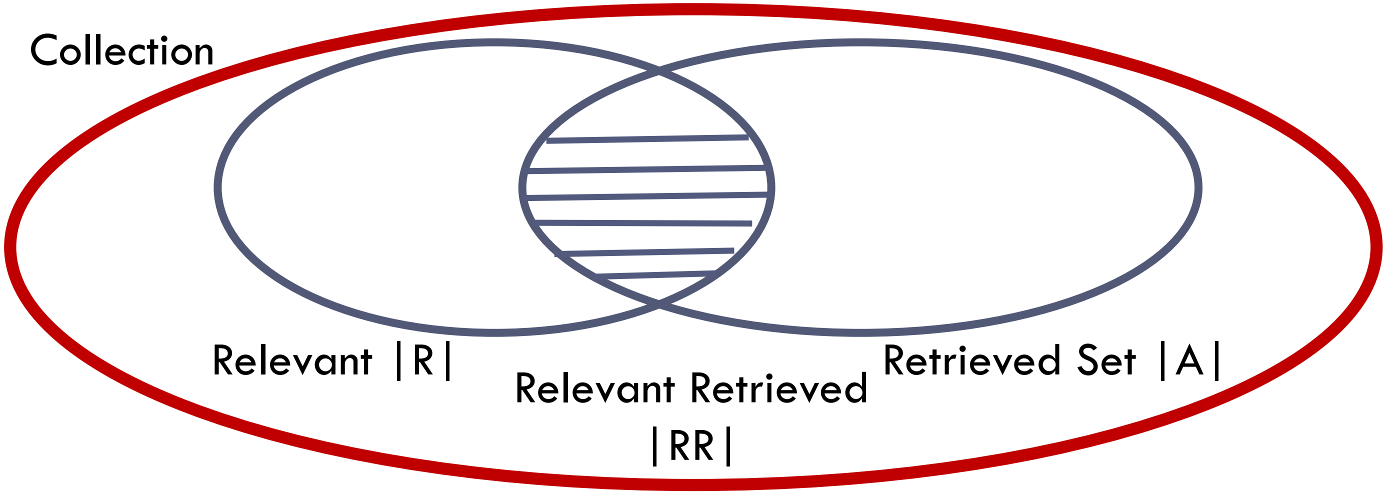
- Actively involve users in gathering information about their uses and preferences, related to an IR system, to evaluate it
  - Need to be representative of the population under evaluation so that any conclusions based on interactions with these users are considered valid
  - Sites such as Mechanical Turk can help, but… do they really?

Test Collections

- AP: Associated Press newswire documents from 1988-1990
  - Queries, topics, and relevance judgments
- Yahoo! Answers Dataset
  - Questions, answers, and metadata for answers and users
- BookCrossing
  - Rated books by users, demographic information on users
- LETOR: Learning to Rank
  - Large query-url pairs, ranking information
- Yahoo Search
  - Query logs to entity search

Effectiveness Measures

- Precision & Recall



$$Precision = \frac{|RR|}{|A|}$$

$$Recall = \frac{|RR|}{|R|}$$

Precision at k (P@k)

Effectiveness Measures

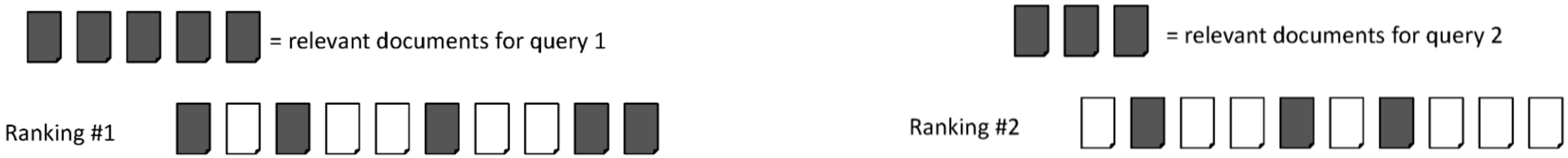
- F-Measure
  - P is Precision and R is recall
  - Weighted variations are also often considered

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

- False positives vs false negatives
  - FP: error that indicates that a non-relevant document is retrieved
  - FN: error that indicates that a relevant document is not retrieved

Effectiveness Measures

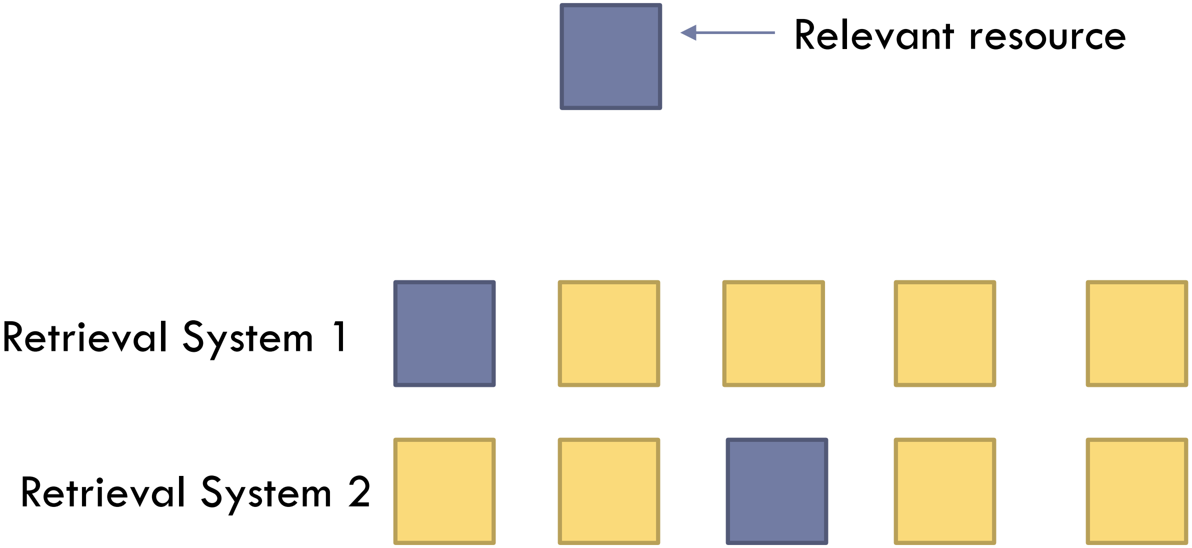
- Mean Average Precision
  - Summarize rankings from multiple tasks by averaging average precision
  - Most commonly used measure in research papers
  - Assumes user is interested in finding many relevant resources for each task
  - Requires many relevance judgments in a collection


= relevant documents for query 1
Ranking #1
= relevant documents for query 2
Ranking #2

$$Precision\ query\ 1 = \frac{5}{10} = 0.50 \qquad Precision\ query\ 2 = \frac{3}{10} = 0.30$$

$$Mean\ Average\ Precision = \frac{\frac{5}{10} + \frac{3}{10}}{2} = 0.40$$

Does Precision Always Work?



Relevant resource

Retrieval System 1

Retrieval System 2

What is the precisions of System 1? And System 2?
Are both system equivalents in terms of performance?

Effectiveness Measures

- Normalized Discounted Cumulative Gain
  - Assumes that
    - Highly relevant resources are more useful than marginally relevant resources
      - Common ranges are (0..1) and (1 ..5)
    - The lower the ranked position of a relevant resource, the less useful it is for the user, since it is less likely to be examined

$$NDCG = \frac{DCG}{IDCG} = \frac{rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2 i}}{\underbrace{rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2 i}}_{\text{Normalization factor}}}$$

Graded relevance of the document at rank *i*

Penalization/reduction/discount factors

Computed for the perfect ranking

Example

	Retrieved Resources									
Given Rank	3	2	3	0	0	1	2	2	3	0
Discounted Gain	3	2	1.89	0	0	0.39	0.71	0.67	0.95	0
DCG	3	5	6.89	6.89	6.89	7.28	7.99	8.66	9.61	**9.61**
Ideal Rank	3	3	3	2	2	2	1	0	0	0
Ideal DCG	3	6	7.89	8.89	9.75	10.52	10.88	10.88	10.88	**10.88**

$$NDCG = \frac{DCG}{IDCG} = \frac{rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2 i}}{rel_1 + \sum_{i=2}^{p} \frac{rel_i}{log_2 i}} = \frac{9.61}{10.88}$$

Effectiveness Measures

- Mean Reciprocal Rank
  - Aims to identify the average number of resources a user has to scan through before identifying a relevant one

Tasks

$$MRR = \underbrace{\frac{1}{|T|}}_{\text{Normalization factor}} \sum_{i=1}^{T} \frac{1}{rank_i}$$

Ranking position of the first relevant (i.e., correct) resource

User-Oriented Measures

- Coverage
  - In RecSys, number of items in a collection that can ever be recommended
- Diversity
  - Degree to which the result set is homogeneous
- Novelty
  - Fraction of relevant documents retrieved that were unknown to the user
- Serendipity
  - Degree to which results are “surprising”

Efficiency Metrics

- Scalability
  - With a growing dataset, how will the system behave?
- Overall Response Performance
  - Real time vs offline tasks
- Query throughput
  - Number of queries processed per unit of time

Significance Tests

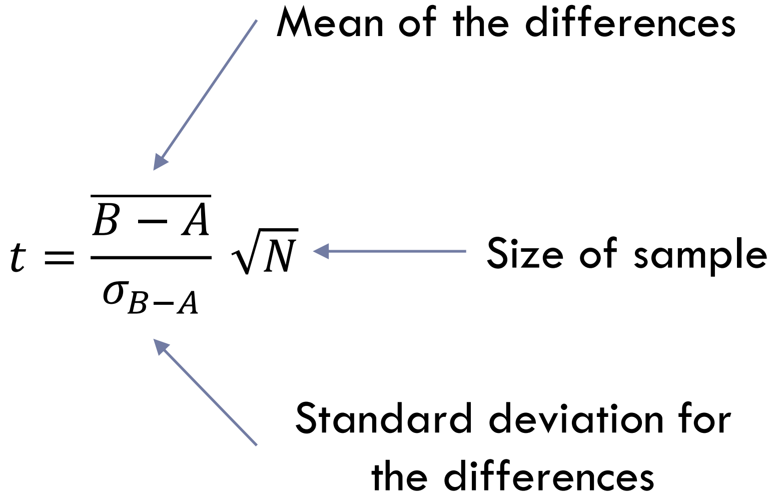
- Given the results from a number of queries, how can we conclude that strategy A is better than strategy B?
  - A significance test enables us to reject the null hypothesis (no difference) in favor of the alternative hypothesis (B is better than A)
  - The power of a test is the probability that the test will reject the null hypothesis correctly
  - Increasing the number of “trials” in the experiment also increases power of test
  - Common significance tests
    - T-test, Wilcoxon signed-ranked test, sign test

Significance Tests

- Procedure for comparing 2 retrieval systems
  1. Compute the effectiveness measure for every task for both systems
  2. Compute a *test statistic* based on a comparison for the effectiveness measure for each task
     - Test statistic depends on the significance test, and is simply a quantity calculated from the sample data that is used to decide whether or not the null hypothesis should be rejected
  3. Use test statistic to compute a P-value, which is the probability that a test statistic value that extreme could be observed is the null hypothesis were true.
     - *Small* P-value suggest that the null hypothesis may be false
  4. The null hypothesis (no difference) is rejected in favor of the alternative hypothesis (e.g., B more effective than A) if P-value is $\leq \alpha$, the significance level
     - Typical values for α are 0.05 and 0.1

Example: t-Test

Task	A	B	B-A
1	25	35	10
2	43	84	41
3	39	15	-24
4	75	75	0
5	43	68	25
6	15	85	70
7	20	80	60
8	52	50	-2
9	49	58	9
10	50	75	25



Mean of the differences

$$t = \frac{\overline{B-A}}{\sigma_{B-A}}\sqrt{N}$$

Size of sample

Standard deviation for the differences

$$t = \frac{21.4}{29.1}\sqrt{10} = 2.33$$

P-value= 0.02

$\leq \alpha = 0.05 \longrightarrow$ Reject Null Hypothesis

$\leq \alpha = 0.01 \longrightarrow$ Accept Null Hypothesis