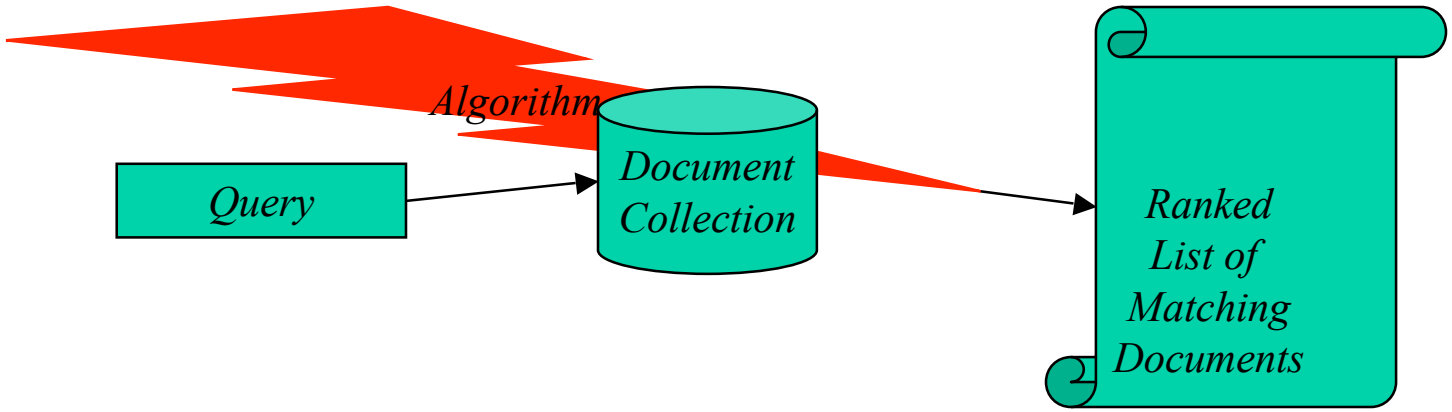
Why System Evaluation?

- There are many retrieval models/ algorithms/ systems, which one is the best?
- What is the best component for:
  - Ranking function (dot-product, cosine, …)
  - Term selection (stopword removal, stemming…)
  - Term weighting (TF, TF-IDF,…)
- How far down the ranked list will a user need to look to find some/all relevant documents?

IR Systems



or, determine, given a query q, P(d | Q)
for all documents d in a collection and then
sort the documents according to P(D | Q)

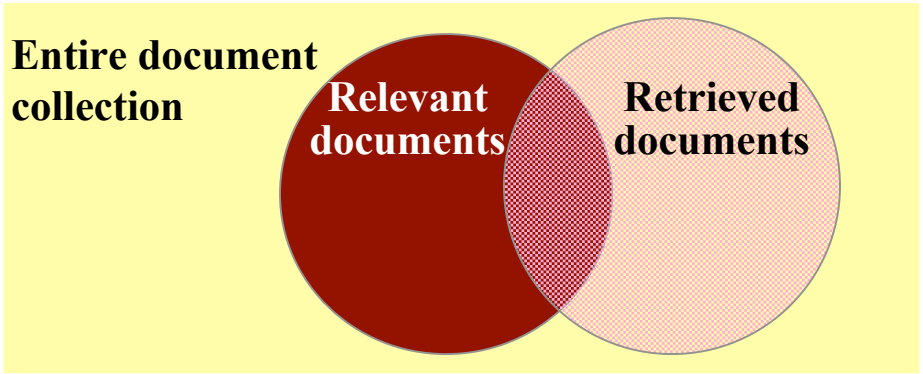
Difficulties in Evaluating IR Systems

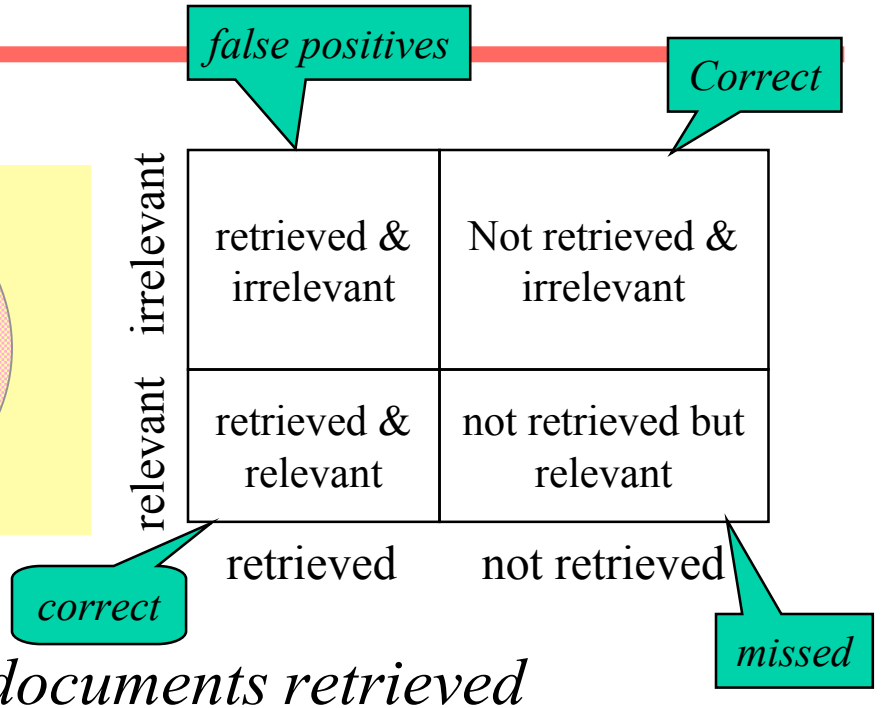
- Effectiveness is related to the ***relevancy*** of retrieved items.
- Relevancy is not typically binary but continuous.
- Even if relevancy is binary, it can be a difficult judgment to make.
- Relevancy, from a human standpoint, is:
  - Subjective: Depends upon a specific user's judgment.
  - Situational: Relates to user's current needs.
  - Cognitive: Depends on human perception and behavior.
  - Dynamic: Changes over time.

Human Labeled Corpora (Gold Standard)

- Start with a corpus of documents.
- Collect a set of queries for this corpus.
- Have one or more human experts exhaustively label the relevant documents for each query.
- Typically assumes binary relevance judgments.
- Requires considerable human effort for large document/query corpora.

Precision and Recall





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

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

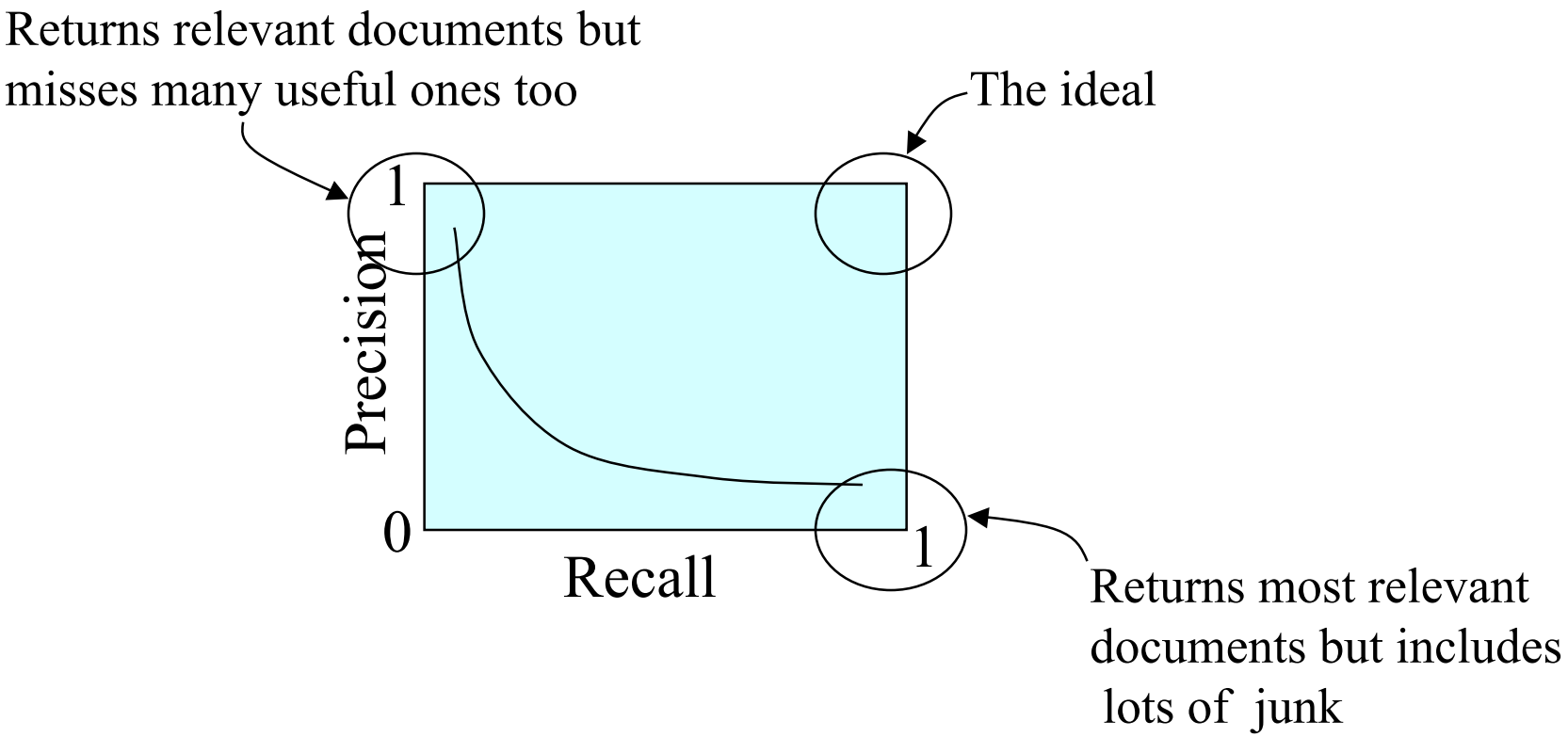
Precision and Recall

- Precision
  - The ability to retrieve top-ranked documents that are mostly relevant.
- Recall
  - The ability of the search to find ***all*** of the relevant items in the corpus.
  - Recall is not well defined for Search Engines

Determining Recall is Difficult

- Total number of relevant items is sometimes not available:
  - Sample across the database and perform relevance judgment on these items.
  - Apply different retrieval algorithms to the same database for the same query. The aggregate of relevant items is taken as the total relevant set. (This is the definition of the relevant set used by TREC rather than a hand-labeled corpus)

Trade-off between Recall and Precision



Computing Recall/Precision Points

- 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.

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

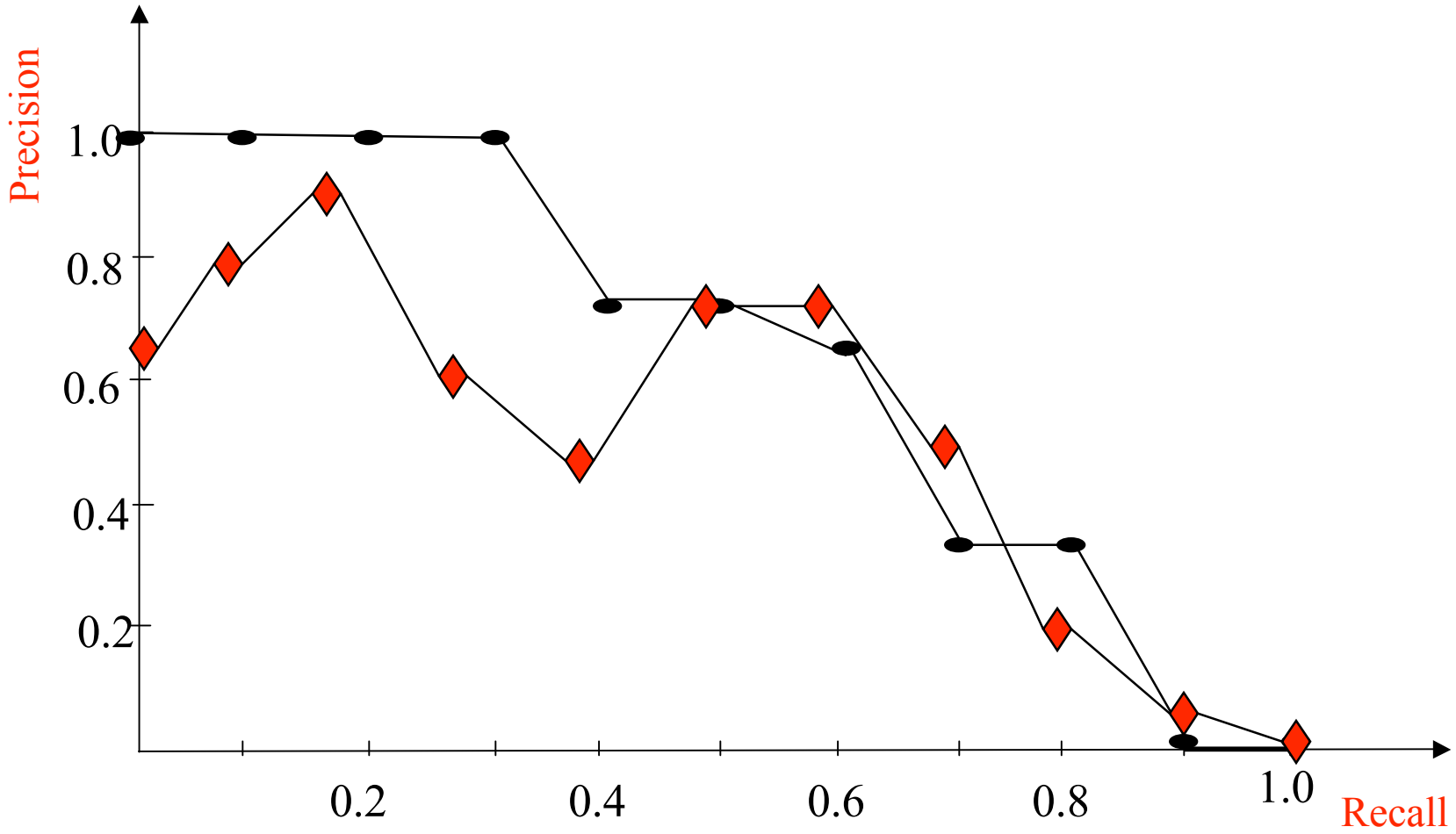
R=3/6=0.5; P=3/4=0.75

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

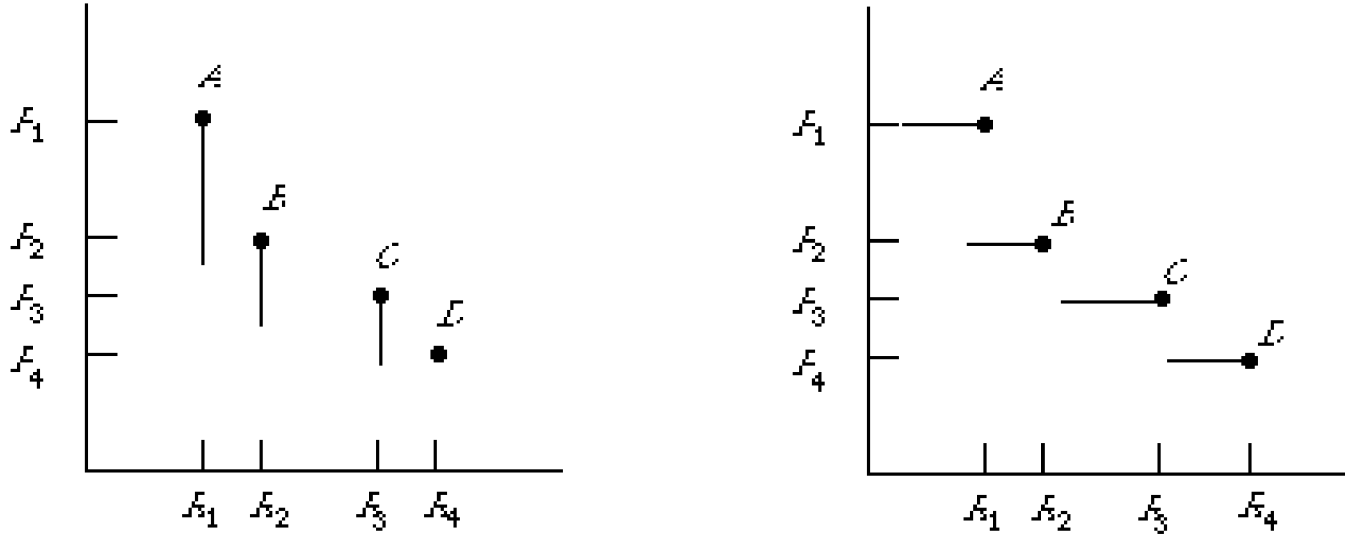
R=5/6=0.833; p=5/13=0.38

Missing one relevant document. Never reach 100% recall

Recall/Precision Curve: Examples



Interpolating Recall & Precision



The left figure above shows the ranges of precisions obtained for several tests at specified levels of recall. The right figure shows the range of recalls obtained at specified levels of precision. Problem: should you fix recall then get precision? fix precision then get recall? fix num retrieved and get precision and recall?

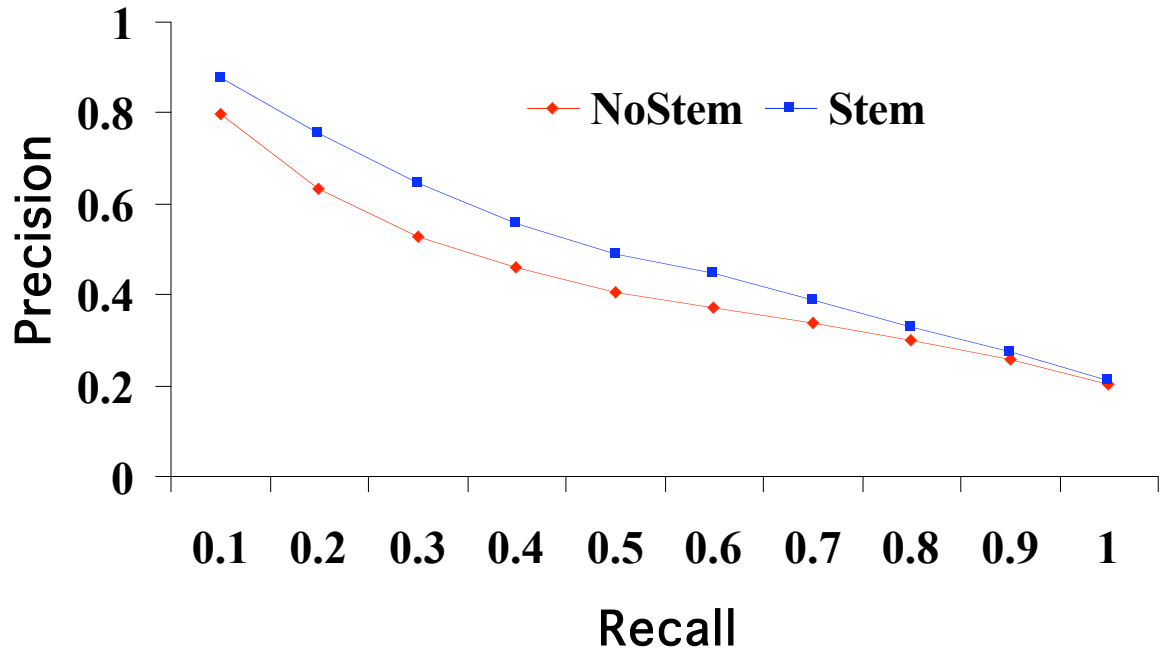
For search engines, how should you measure?

Average Recall/Precision Curve

- Typically average performance over a large ***set*** of queries.
- Compute average precision at each standard recall level across all queries.
- Plot average precision/recall curves to evaluate overall system performance on a document/query corpus.

Compare Two or More Systems

- The curve closest to the upper right-hand corner of the graph indicates the best performance



Problem, Which curve is better?

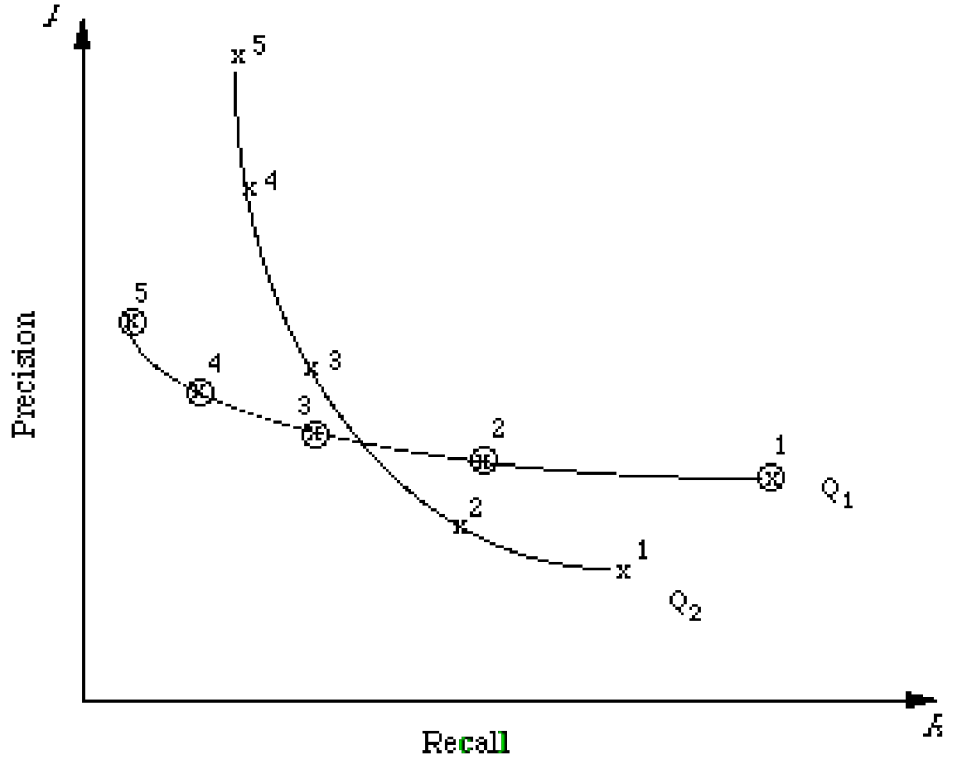
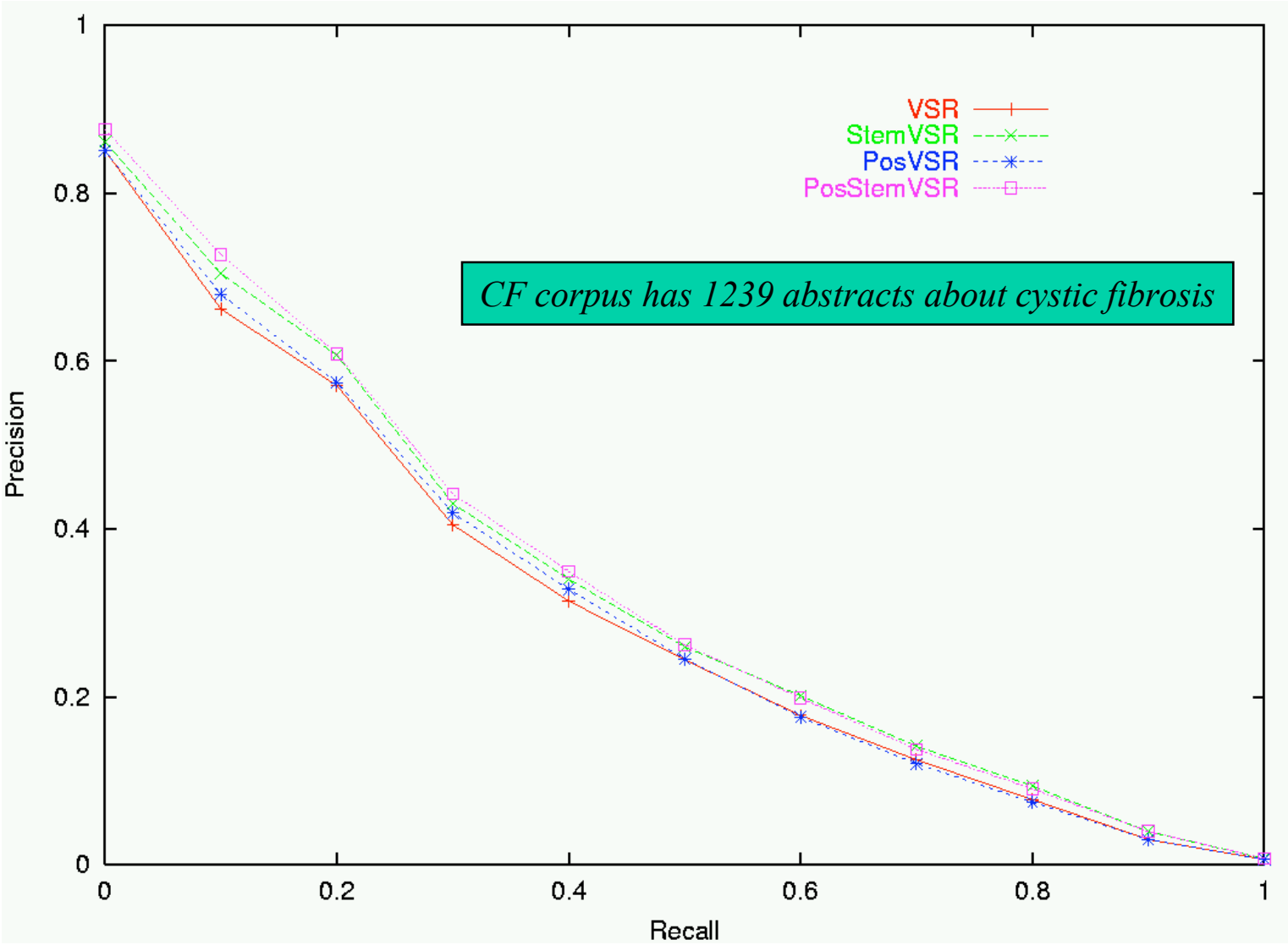


Figure 7.1. The precision-recall curves for two queries. The coordinates indicate the settings of the control parameter λ

Sample RP Curve for CF Corpus



R- Precision

- 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

F-Measure

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

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

- Compared to arithmetic mean, both need to be high for harmonic mean to be high.
- Assumes a threshold has been set

Modeling Performance -- Swets

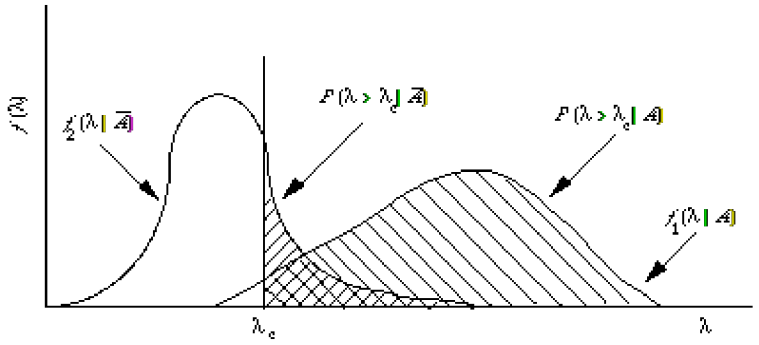


Figure 7.5. Two normal distributions for λ, one, $N(\mu_1, \sigma_1)$, on the set of relevant documents A with density of $f_1'(\lambda \mid A)$, the other, $N(\mu_2, \sigma_2)$, on the set of non-relevant documents $\bar{A}$ with density $f_2'(\lambda \mid \bar{A})$. The size of the areas shaded in a N-W and N-E direction represents recall and fallout respectively

- Idea is, if you know the two distributions you can manipulate lambda to give you the precision/recall you want.
- Problems?

Modeling Performance -- SMART

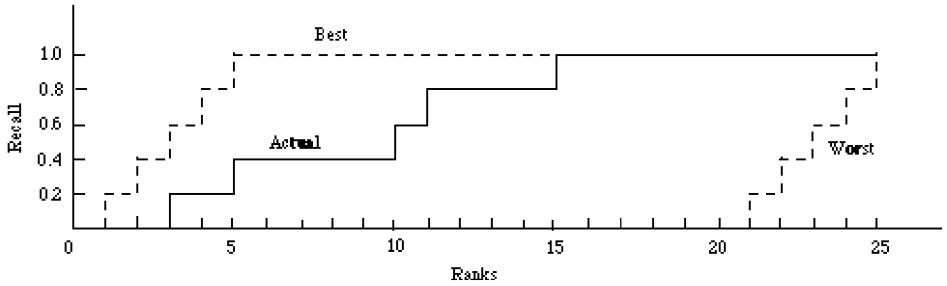


Figure 7.9 An illustration of how the normalized recall curve is bounded by the best and worst cases. (Adapted from Robertson[15], page 99)

- Idea is to show recall as a function of number of documents retrieved
- Based of assumption that each retrieval has a non-zero cost. ?
- Problems with applying this model to search engines? How can be model be revised to overcome those problems?

Fallout Rate

- Problems with both precision and recall:
  - Number of irrelevant documents in the collection is not taken into account.
  - Recall is undefined when there is no relevant document in the collection.
  - Precision is undefined when no document is retrieved.

$$Fallout = \frac{no.\,of\ \ nonrelevant\ items\ retrieved}{total\ no.\,of\ \ nonrelevant\ items\ in\ the\ collection}$$

Subjective Relevance Measure

- *Novelty Ratio*: The proportion of items retrieved and judged relevant by the user and of which they were previously unaware.
  - Ability to find ***new*** information on a topic.
- *Coverage Ratio*: The proportion of relevant items retrieved out of the total relevant documents ***known*** to a user prior to the search.
  - Relevant when the user wants to locate documents which they have seen before (e.g., the budget report for Year 2000).

Other Factors to Consider

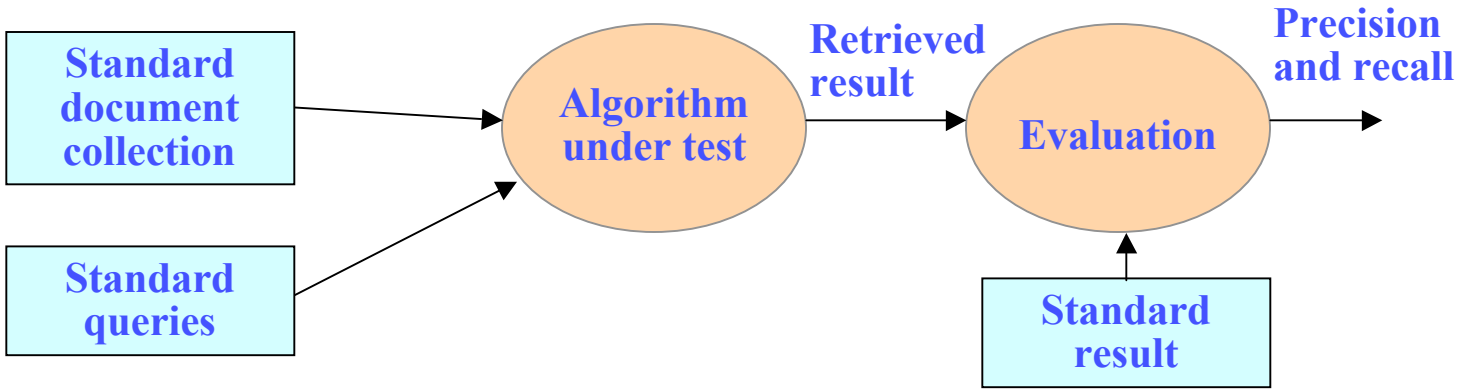
- *User effort*: Work required from the user in formulating queries, conducting the search, and screening the output.
- *Response time*: Time interval between receipt of a user query and the presentation of system responses.
- *Form of presentation*: Influence of search output format on the user's ability to utilize the retrieved materials.
- *Collection coverage*: Extent to which any/all relevant items are included in the document corpus.
- *Nth Location* -- How far down the ranked list will a user need to look to find some/all relevant documents

Experimental Setup for Benchmarking

- ***Analytical*** performance evaluation is difficult for document retrieval systems because many characteristics such as relevance, distribution of words, etc., are difficult to describe with mathematical precision.
- Performance is measured by ***benchmarking***. That is, the retrieval effectiveness of a system is evaluated on a *given set of documents*, *queries*, and *relevance judgments*.
- Performance data is valid only for the environment under which the system is evaluated.

Benchmarks

- A benchmark collection contains:
  - A set of standard documents and queries/topics.
  - A list of relevant documents for each query.
- Standard collections for traditional IR:
  - Smart collection: ftp://ftp.cs.cornell.edu/pub/smart
  - TREC: http://trec.nist.gov/



Benchmarking – The Problems

- Performance data is valid only for a particular benchmark.
- Building a benchmark corpus is a difficult task.
- Benchmark web corpora are just starting to be developed.
- Benchmark foreign-language corpora are just starting to be developed.

The TREC Benchmark

- TREC: Text REtrieval Conference (http://trec.nist.gov/)
  Originated from the TIPSTER program sponsored by Defense Advanced Research Projects Agency (DARPA).

- Became an annual conference in 1992, co-sponsored by the National Institute of Standards and Technology (NIST) and DARPA.

- Participants are given parts of a standard set of documents and TOPICS (from which queries have to be derived) in different stages for training and testing.

- Participants submit the P/R values for the final document and query corpus and present their results at the conference.

TREC Advantages

- Large scale (compared to a few MB in the SMART Collection).
- Relevance judgments provided.
- Under continuous development with support from the U.S. Government.
- Wide participation:
  - TREC 1: 28 papers 360 pages.
  - TREC 4: 37 papers 560 pages.
  - TREC 7: 61 papers 600 pages.
  - TREC 8: 74 papers.

Evaluation

- Summary table statistics: Number of topics, number of documents retrieved, number of relevant documents.
- Recall-precision average: Average precision at 11 recall levels (0 to 1 at 0.1 increments).
- Document level average: Average precision when 5, 10, .., 100, … 1000 documents are retrieved.
- Average precision histogram: Difference of the R-precision for each topic and the average R-precision of all systems for that topic.

Seeking Better Web Searches

Parts of a Search Engine

- Crawler
- Indexer
  - Part of IR
- Retrieval Engine
  - Part of IR
  - The focus of this article

Crawler

- The data source
  - old problems
    - how to get the data
    - how to keep it up to date
  - New Solutions
    - froogle
    - google scholar
    - google book search
    - google catalogs
    - None of this does anything about the crawling problem

Indexing

- Only new stuff here is massive parallelism
  - Google & Beowulf

Retrieval

- Clustering
  - mooter.com, clusty.com, kartoo.com
- Collaborative Filtering
  - tivo, amazon.com
- Augmented search
  - use stuff on your disk to bias search
  - use past queries to bias search
  - Use location (if mobile) to direct search
- Specialized search
  - the rise of “niche” engines
    - some supported by major engines (eg. scholar)

More on Retrieval

- Hum a few bars
- Draw something
- Free text?

Evaluating Search Engines

- How do you judge the effectiveness of a search engine
  - when is some new thing useful?
    - How does a search engine provider evaluate?
    - How does / should a person evaluate?