Information Retrieval and Web Search

Cornelia Caragea

Computer Science
University of Illinois at Chicago

Credits for slides: Mooney

Evaluation of Retrieval Models

Required Reading

- "Information Retrieval" textbook
  - Chapter 8: Evaluation in IR

Announcements

- Second programming assignment will be out today.

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?

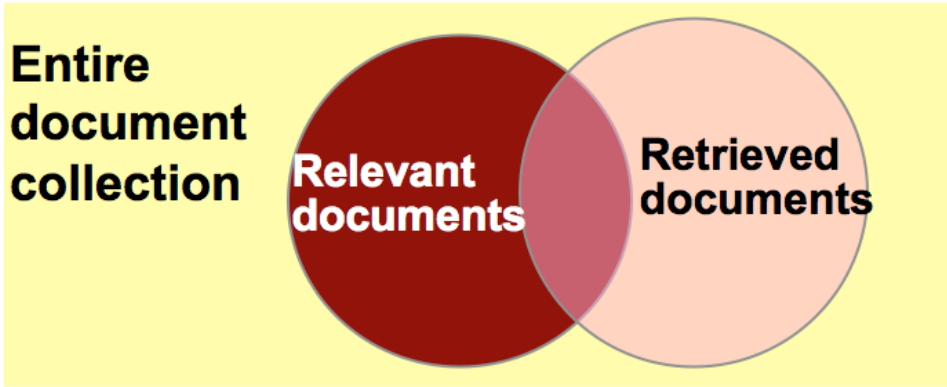
Difficulties in Evaluating IR Systems

- Effectiveness is related to the relevancy of retrieved items.
- Typically, relevancy is not binary but continuous.
- Even if relevancy is binary, it can be a difficult judgment to make.
- From a human standpoint, relevancy is:
  - Subjective: Depends upon a specific user's judgment.
  - Situational: Relates to a 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



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

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

Contingency Table

	Relevant	Non-relevant
Retrieved	tp	fp
Not retrieved	fn	tn

Precision and Recall

From all the documents that are relevant out there, how many did the IR system retrieve?

$$\text{Recall} = \frac{tp}{tp + fn}$$

From all the documents that are retrieved by the IR system, how many are relevant?

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + fp + fn + tn}$$

Is Accuracy an appropriate measure to evaluate IR systems?

Precision and Recall

From all the documents that are relevant out there, how many did the IR system retrieve?

$$\text{Recall} = \frac{tp}{tp + fn}$$

From all the documents that are retrieved by the IR system, how many are relevant?

$$\text{Precision} = \frac{tp}{tp + fp}$$

$$\text{Accuracy} = \frac{tp + tn}{tp + fp + fn + tn}$$

Is Accuracy an appropriate measure to evaluate IR systems?
No. The data is extremely skewed ($\approx 99.9\%$ of documents are non-relevant).

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.

Determining Recall is Difficult

- Total number of relevant items is not always 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.

Computing Recall/Precision Points

- For a given query, produce the ranked list of retrievals.
- Adjusting a threshold on this ranked list produces different sets of retrieved documents, and therefore different recall/precision measures.
- 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: Example 1

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

Computing Recall/Precision Points: Example 2

n	doc #	relevant
1	588	x
2	576	
3	589	x
4	342	
5	590	x
6	717	
7	984	
8	772	x
9	321	x
10	498	
11	113	
12	628	
13	772	
14	592	x

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

Computing Recall/Precision Points: Example 2

n	doc #	relevant
1	588	x
2	576	
3	589	x
4	342	
5	590	x
6	717	
7	984	
8	772	x
9	321	x
10	498	
11	113	
12	628	
13	772	
14	592	x

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=2/3=0.667

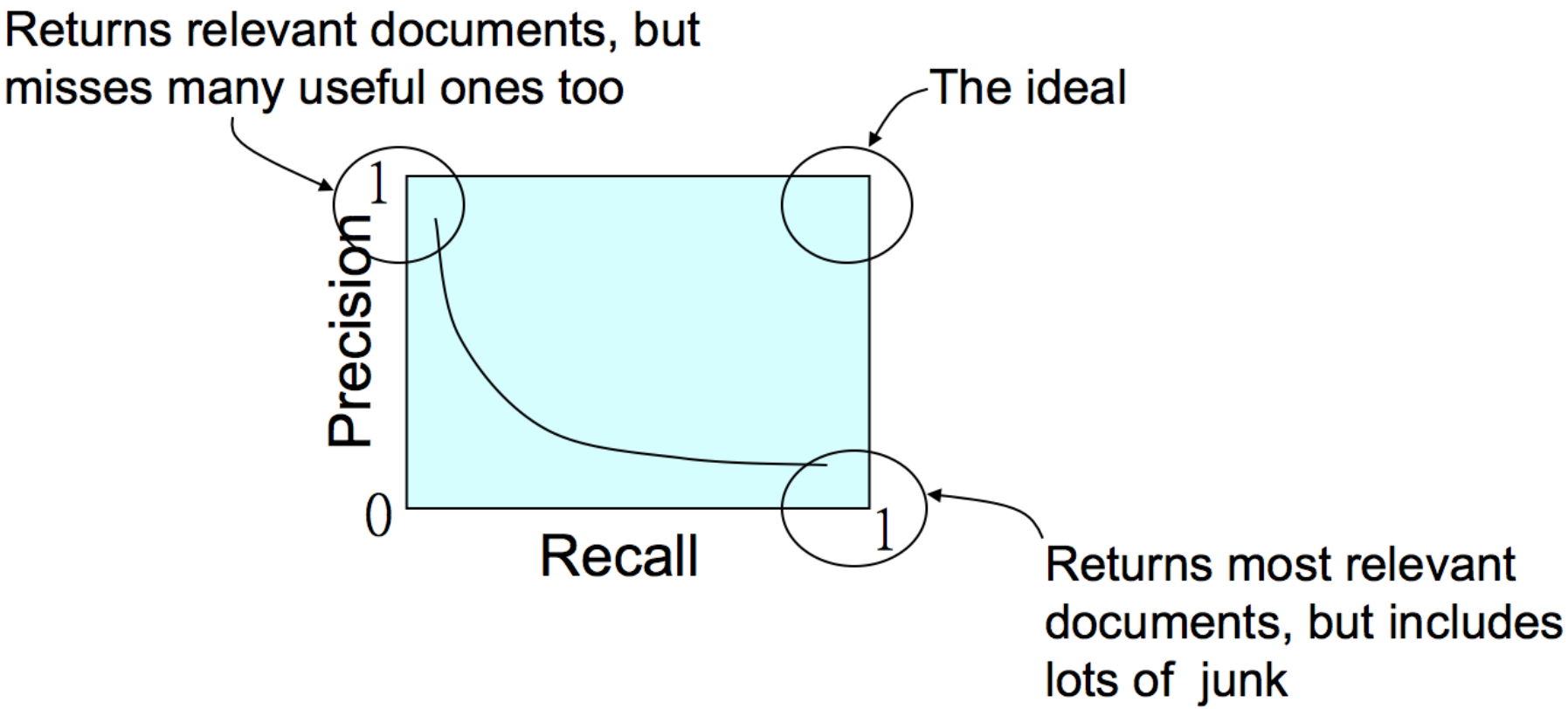
R=3/6=0.5; P=3/5=0.6

R=4/6=0.667; P=4/8=0.5

R=5/6=0.833; P=5/9=0.556

R=6/6=1.0; p=6/14=0.429

Trade-off between Recall and Precision

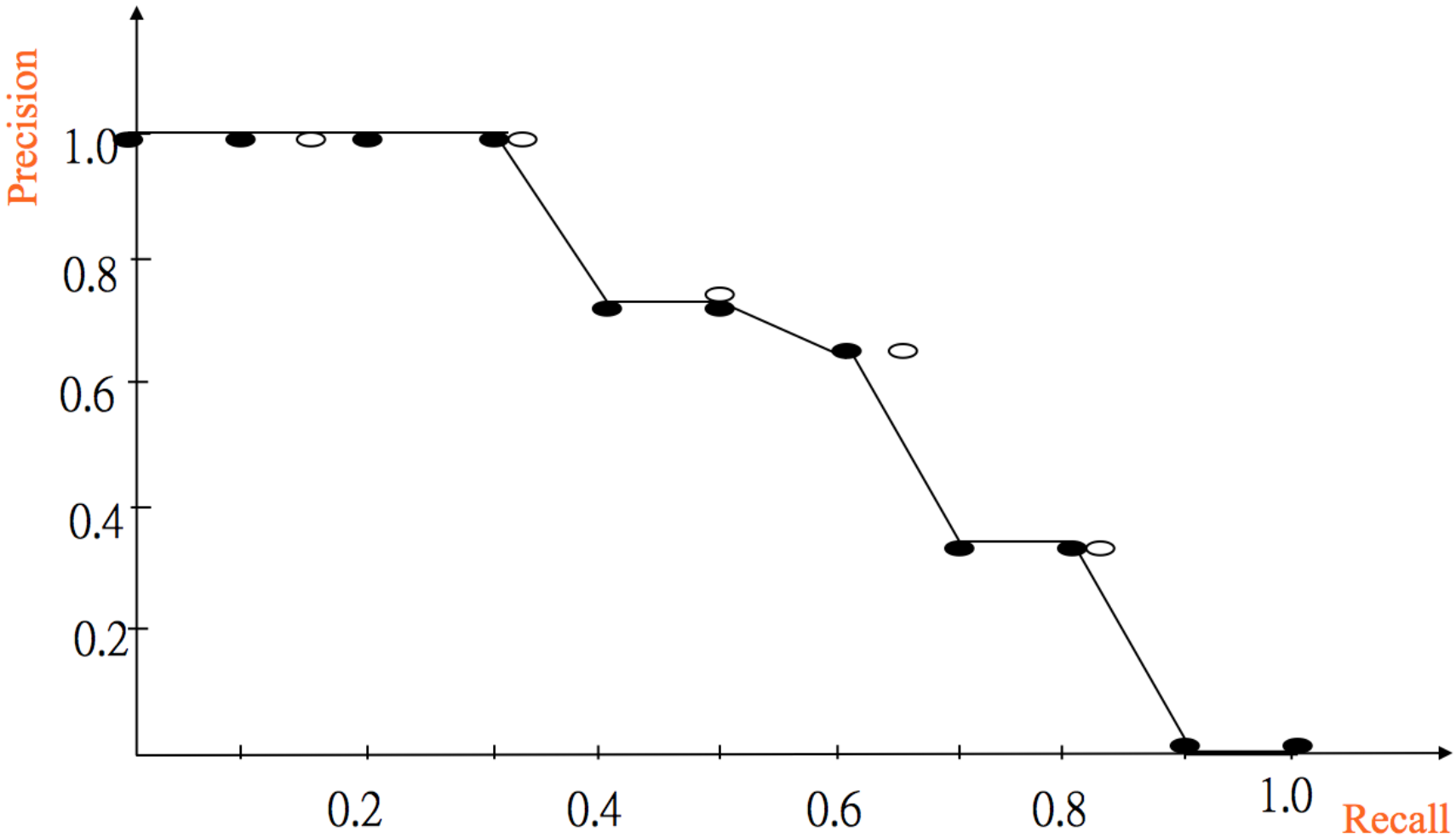


Interpolating a Recall/Precision Curve

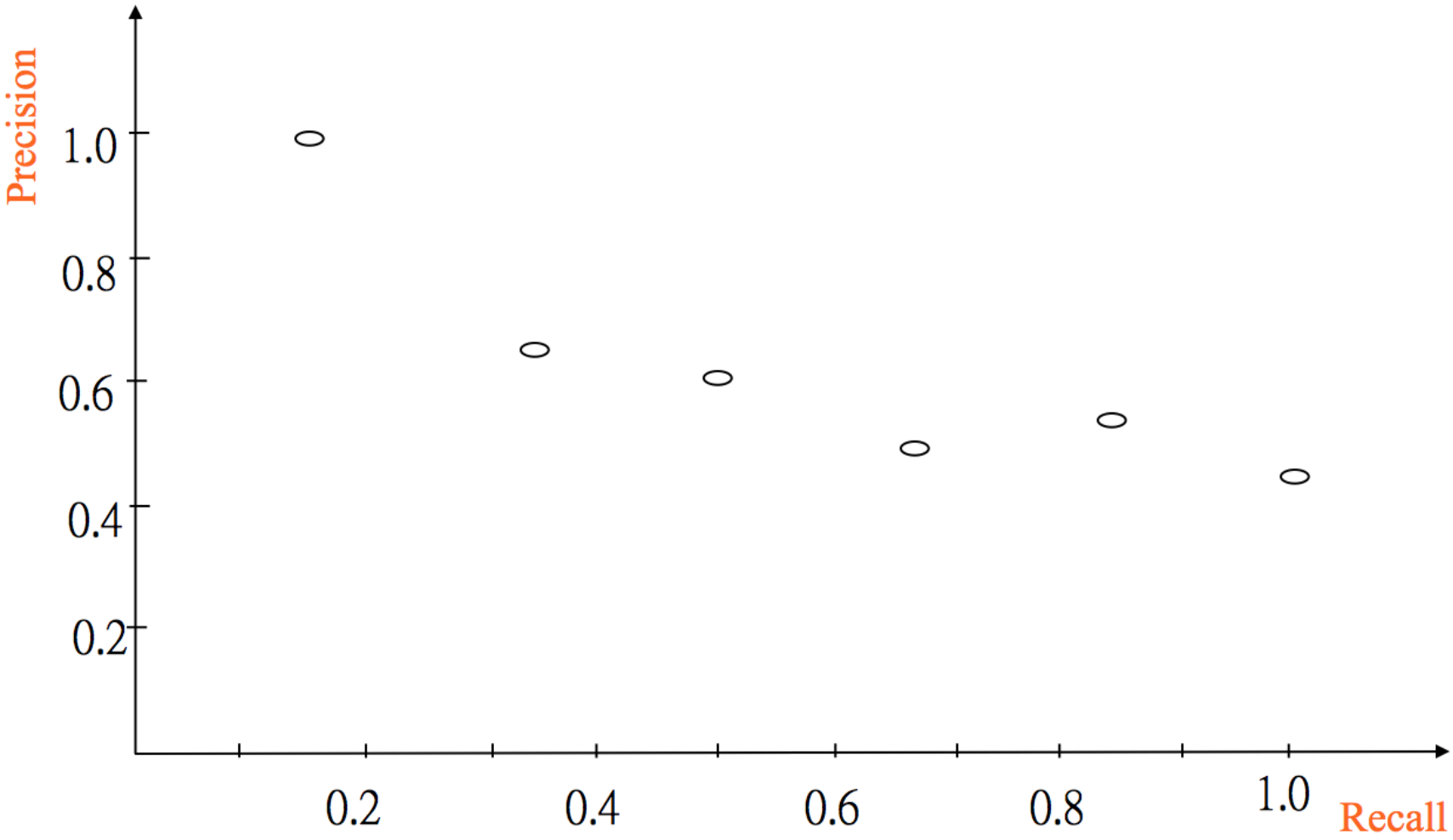
- Interpolate a precision value for each standard recall level:
  - $r_j \in \{0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$
  - $r_0 = 0.0, r_1 = 0.1, \cdots, r_{10} = 1.0$
- The interpolated precision at a certain recall level is defined as the highest precision found for any recall level $r' \geq r$:

$$Prec(r) = \max_{r' \geq r} Prec(r')$$

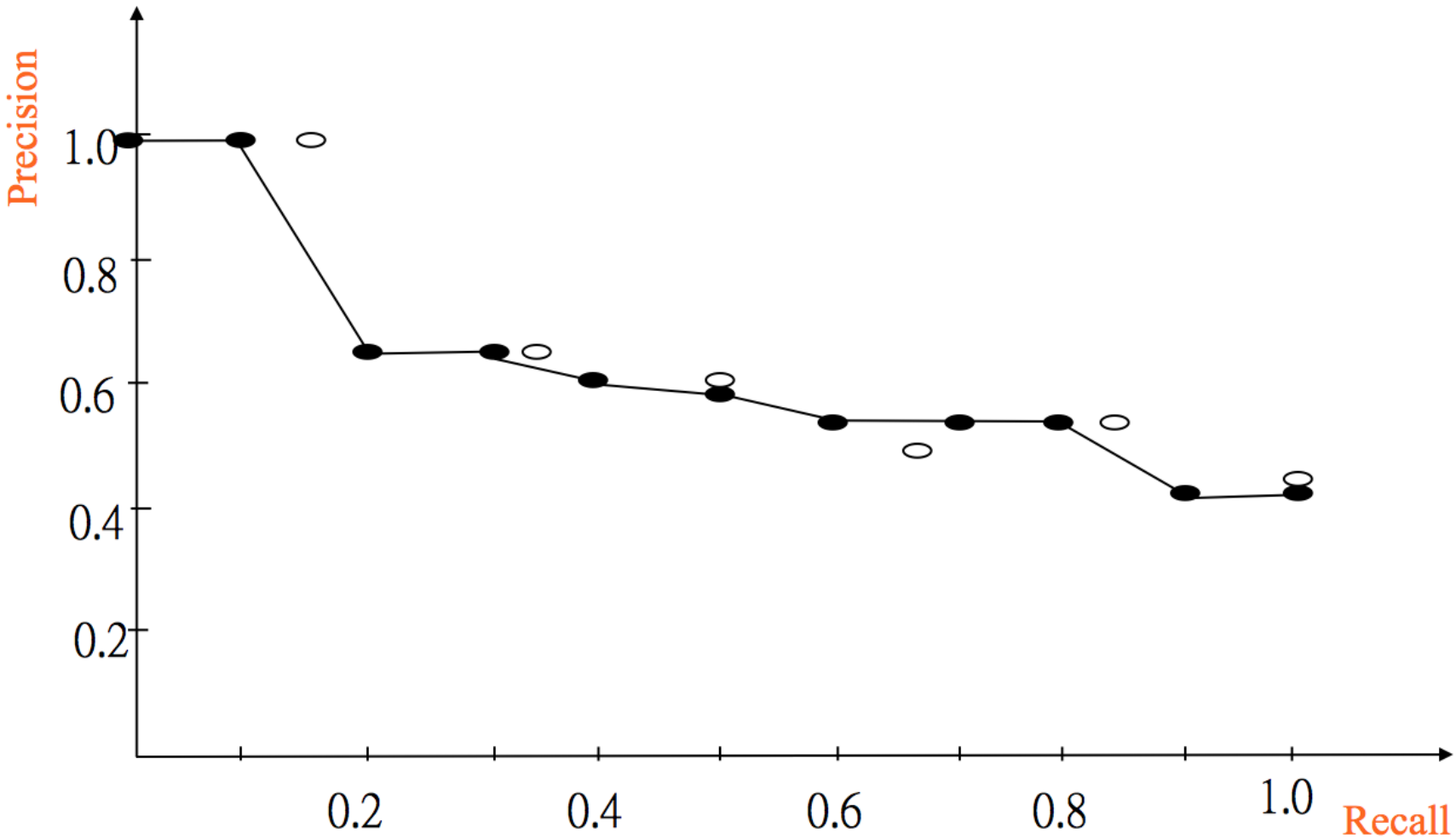
Interpolating a Recall/Precision Curve: Example 1



Interpolating a Recall/Precision Curve: Example 2



Interpolating a Recall/Precision Curve: Example 2

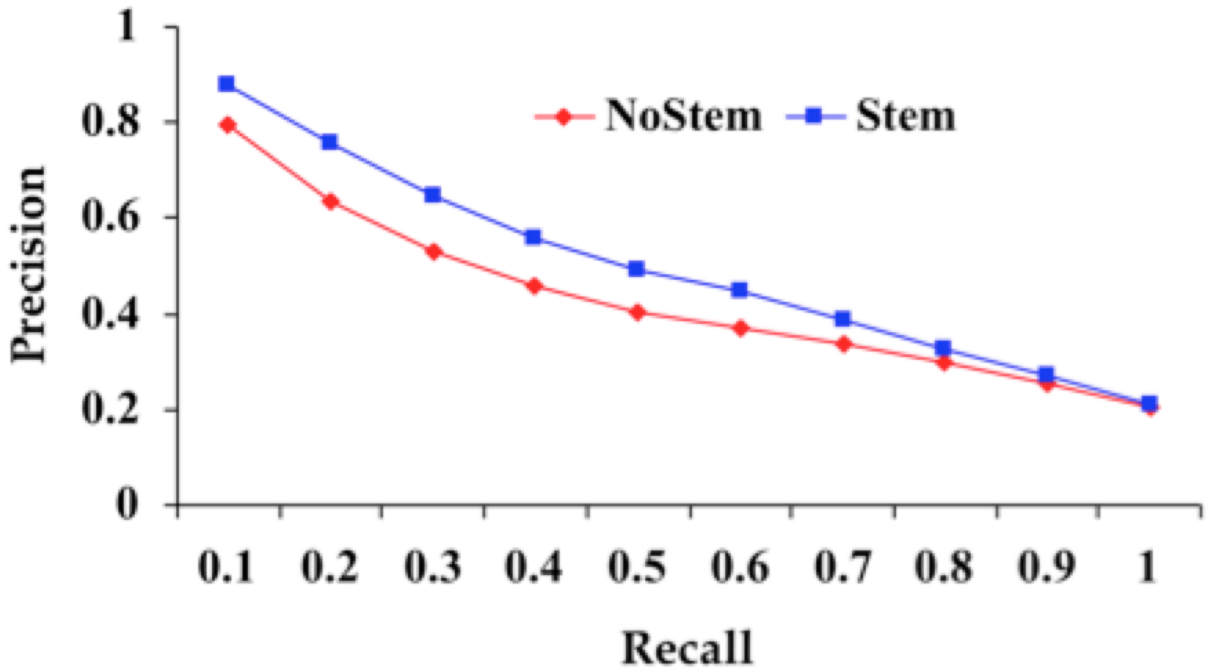


Average Recall/Precision Curve

- Typically, we calculate 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



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:

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

- Compared to arithmetic mean, both need to be high for harmonic mean to be high.

E-Measure (parameterized F Measure)

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

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

  - $\beta = 1$: Equally weight precision and recall (E=F).

Mean Average Precision (MAP)

- Average Precision: Average of the precision values at the points at which each relevant document is retrieved.
  - Ex1: $(1 + 1 + 0.75 + 0.667 + 0.38 + 0)/6 = 0.633$
  - Ex2: $(1 + 0.667 + 0.6 + 0.5 + 0.556 + 0.429)/6 = 0.625$
- Mean Average Precision: Average of the average precision values for a set of queries.

Mean Reciprocal Rank (MRR)

- Mean Reciprocal Rank:

$$MRR = \frac{1}{|Q|} \sum_{q=1}^{|Q|} \frac{1}{r_q}$$

where r_q is the rank at which the first correct prediction was found for query q in Q.

Issues with Relevance

- Marginal Relevance: Do later documents in the ranking add new information beyond what is already given in higher documents.
  - Choice of retrieved set should encourage **diversity** and **novelty**.
- 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 2014).

Other Factors to Consider

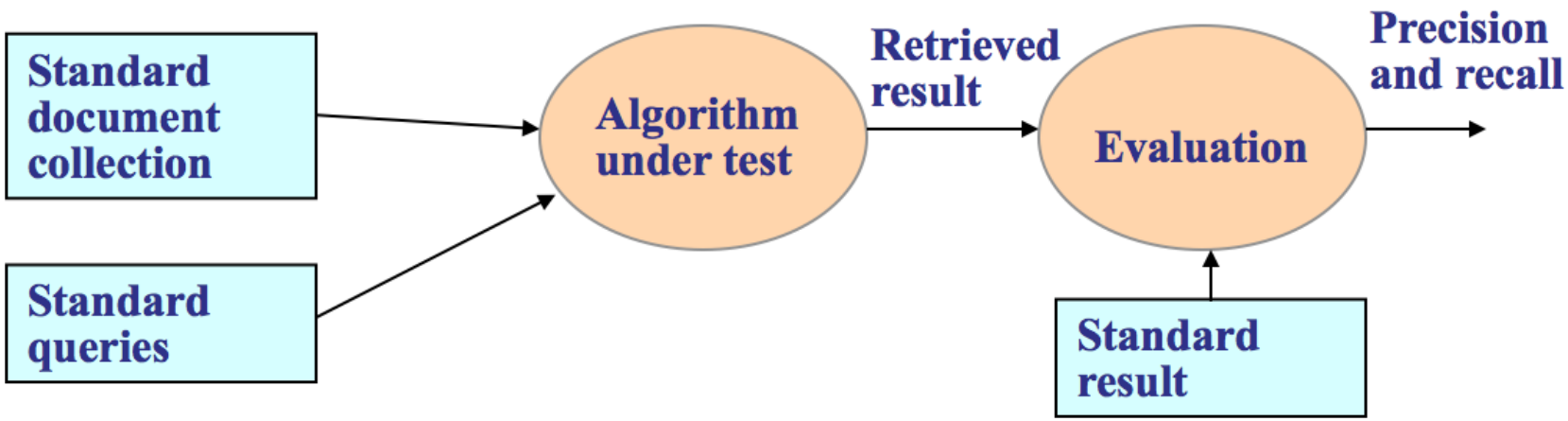
- **User effort:** Work required from the user in formulating queries, conducting the search, and screening the output.
- **Response time:** Time interval between receipt of a user query and the presentation of system responses.
- **Form of presentation:** Influence of search output format on the user's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

Experimental Setup for Benchmarking

- 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 Collections

- Contain:
  - A set of standard documents and queries/topics.
  - A list of relevant documents for each query.



Benchmarking - The Problems

- Performance data is valid only for a particular benchmark.
- Building a benchmark corpus is a difficult task.
- Benchmark foreign-language corpora are more and more developed.

- Interesting paper on building a large dataset for emotion detection: http://www.aclweb.org/anthology/P17-1067