



**Lecture 8:**  
**I: IR evaluation**  
**II: Adaptive Hypertext**

**Johan Bollen**  
Old Dominion University  
Department of Computer Science

*[jbollen@cs.odu.edu](mailto:jbollen@cs.odu.edu)*

*<http://www.cs.odu.edu/~jbollen>*



## IR evaluation

### Why we must evaluate?

1. Myriad of possible IR systems
2. Theoretical and logical appeal
3. User preferences and communities
4. Need for metric and benchmarks to determine relative effectiveness

### Golden standard: relevancy

1. IR: query  $\rightarrow$  results  $\rightarrow$  relevancy?
2. Who determines relevancy?

(a) User?

(b) Expert?

3. What is relevancy?

### Need for formalization of metrics based on relevancy

1. Able to deal with different queries, collections
2. Objective and quantitative metric
3. Simple and intuitive



## Concept of Precision and Recall

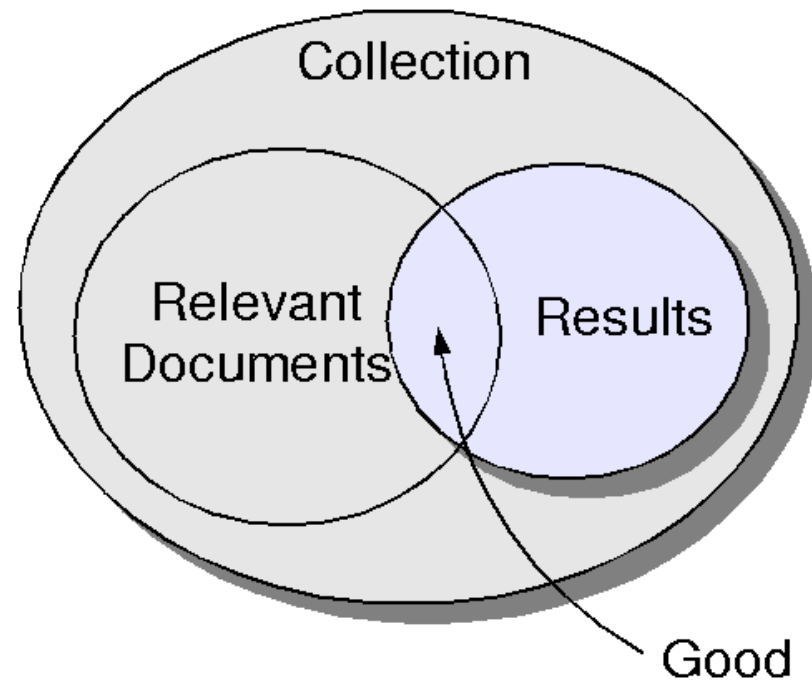
### Two accepted meanings

1. Your favorite DJ's
2. Two metrics of IR performance

### Collection v. query

1. Given we have a query
2. Given we have a collection
3. We have a set of documents that are all relevant to query

- (a) Relevancy remains to be defined
- (b) Let's assume we have an adequate understanding
4. Our IR system returns its results
5. Two questions
  - (a) How many irrelevant results?
  - (b) How well do results cover set of all relevant results in collection?



Relevancy formalized



## Precision and Recall

Let us denote a user information request as  $I$   
and the set of document relevant to that request as  $R$ .

For any request  $I$ , a set of results is returned:  $A$

The cardinality of  $A$  and  $R$  is denoted:  $|A|$  and  $|R|$

We denote the set of relevant results in  $A$  as  $R_a$ , so  $R_a \subseteq A$ .

Recall

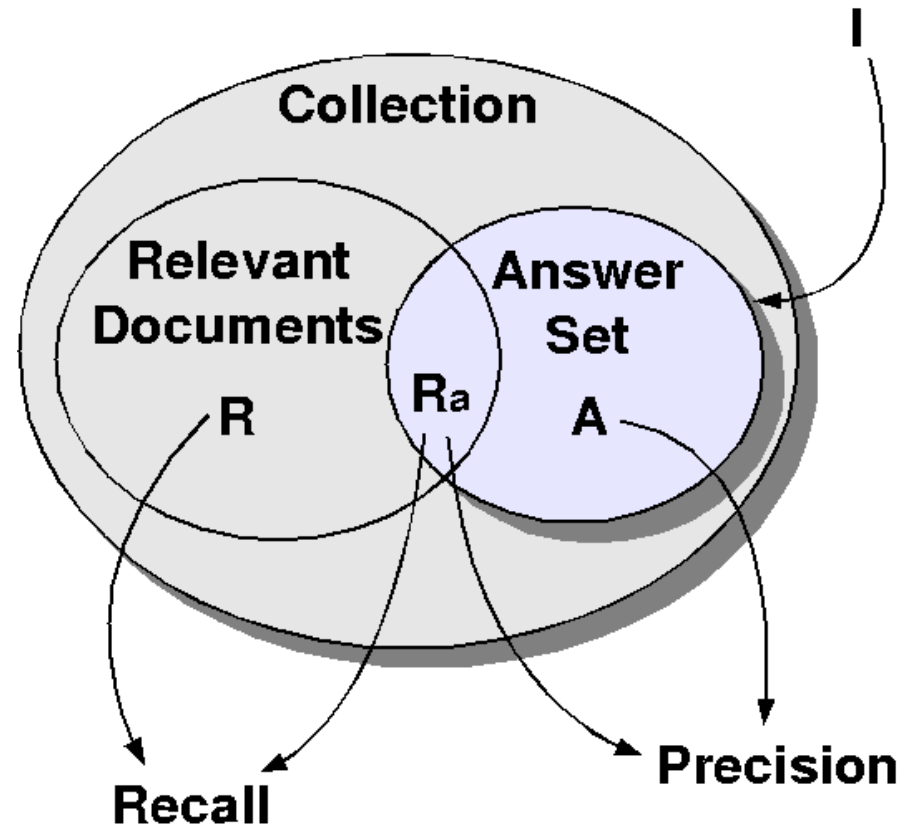
$$\text{recall} = \frac{|R_a|}{|R|}$$

or the ratio of relevant results in the answer set over the total number of possibly relevant results for  $I$  in the collection

Precision

$$\text{precision} = \frac{|R_a|}{|A|}$$

or the ratio of relevant results over all results in the answer set



## Precision and Recall



## Precision and Recall

### Precision

1. Represents how many relevant results returned by IR system
2. Indicates **quality** of results
3. corresponds to false positives

### Recall

1. Represents how many relevant items were captured in toto by IR system
2. Indicates **coverage** of results

3. Corresponds to false negatives

### Precision v. Recall

1. IR system is a little like fishing
2. Fine net catches a lot of small and large fish: many of them not good catch, but you catch most fish
3. Coarse net catches only large fish: but many fish not caught
4. You can not have both fine and coarse net!



## Determining Precision and Recall figures

Assume set of queries: for each query we can determine precision and recall.

Assume a set of queries  $Q = \{q_1, q_2, \dots, q_t\}$ . The ideal and complete answer set of all relevant documents for query  $q_1$  is:

$$R_{q_1} = \{d_7, d_5, d_6, d_{20}, d_{25}\}$$

We invent an IR model, and apply it to the collection  $R$ ,  $R_q \subset R$  using

It produces a ranking of results:

- 1\*  $d_7$
- 2  $d_3$
- 3\*  $d_{20}$
- 4  $d_1$
- 5  $d_{662}$
- 6  $d_{429}$
- 7\*  $d_{25}$
- 8  $d_{99}$
- 9  $d_{723}$
- 10\*  $d_6$

We take into account the ranking produced by the IR system because the user examines results in this order.

Result 1:  $d_7 \in R_{q_1}$ , precision=100%, recall=20%

Result 3:  $d_{20} \in R_{q_1}$ , precision=66%, recall=40%

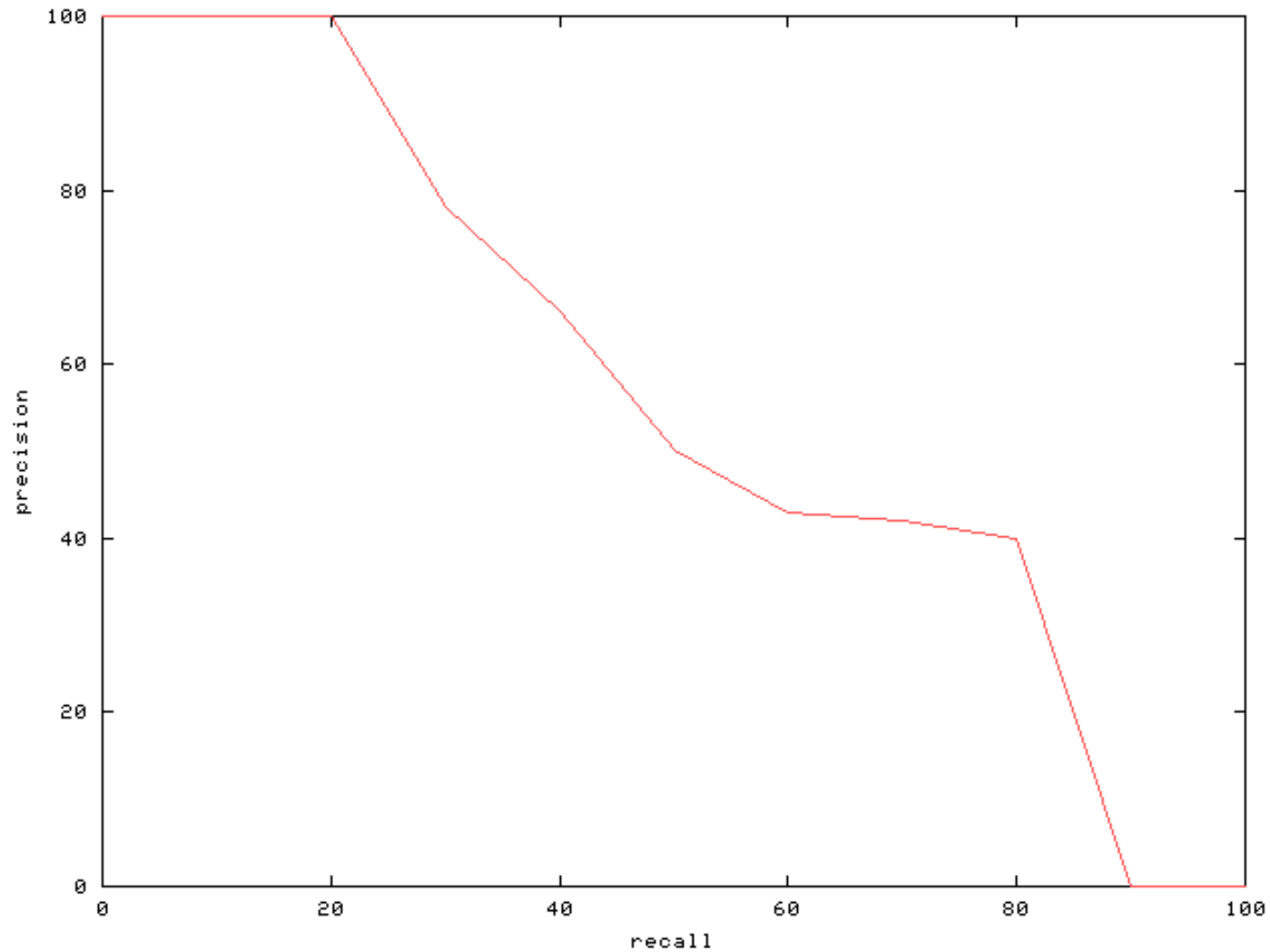
Result 5:  $d_{25} \in R_{q_1}$ , precision=43%, recall=60%

Result 9:  $d_6 \in R_{q_1}$ , precision=40%, recall=80%

Above 80% recall, precision=0



## Precision and Recall curve





## Precision at standard recall levels

1. Notion of precision at recall levels
  - (a) Given that we accept a certain recall level
  - (b) What will the precision of a given IR system be?
  - (c) Standard levels: 0, 10, 20, ..., 100 (11 levels)
2. PR calculated over range of queries
  - (a) Requirement to extrapolate values
  - (b) Aggregate different precision values for level of recall
3. Averaging procedures and aggregate PR numbers



## Average precision

Assume  $\overline{P}(r)$  indicates average precision at recall level  $r$

$P_i(r)$  represents precision at recall level  $r$  for query  $i$ .

then:

$$\overline{P}(r) = \sum_{i=1}^{N_q} \frac{P_i(r)}{N_q}$$

Need to extrapolate recall levels

For any given query, recall level may not be precisely within standard level.

Use maximum of neighboring recall levels.



## Single Value Summaries

1. We often need to know how an IR system performs for single queries
  - (a) Reveal exceptions and specific problems
  - (b) Compare different IR systems over set of given queries
2. Averaging at relevant results for given query
  - (a) Sample precision at relevant results in ranking
  - (b) Average values
  - (c) Encourages systems that rank relevant results highly
3. R-precision
  - (a) Look at total number of relevant documents,  $|R_q|$  for query  $q$
  - (b) Sample precision at ranking in R-th position
4. Precision histograms
  - (a) Compare R-precision for different queries across different IR systems
  - (b) Subtract R-precision system 1 from R-precision system 2
  - (c) Allows comparison of two system for same set for queries



## Notes on Precision and Recall

1. excellent DJs
2. validity problems
  - (a) Relevance:
    - i. Subjective
    - ii. Depends on more than term matches
    - iii. Different to consistently evaluate across queries
  - (b) All or nothing
    - i. Set of relevant documents is not so crisp
    - ii. Precision and Recall do not take into account degree of

- relevancy
- iii. Actual objective should be to compare degree of relevancy for query or IR system and for external relevancy estimator
- iv. Novelty and serendipity: butterfly effect on user satisfaction
- (c) Scale
  - i. Large collections: difficult to determine **all** relevant documents
  - ii. Difficult to construct IR evaluation framework



## Standardized evaluation framework: Reference collections

1. Lack of consistent frameworks for PR evaluation

2. Different systems, collections, queries

TREC - Text Retrieval Conference

1. Started in 1992, Conference format
2. Test-collection sampled from variety of sources, 5.8gb and retrieval tasks
3. Set of information requests
4. Set of relevant results determined by “pooling method”
5. Set of statistics collected for each

system

CACM and ISI collections

1. CACM

- (a) Communications of the ACM (1979)
- (b) Data includes categories, references, etc
- (c) 52 test information requests

2. ISI

- (a) Sample of Institute of Scientific Information collection
- (b) 25 test information requests