Retrieval Performance Evaluation

- Measures

Berlin Chen
Department of Computer Science & Information Engineering
National Taiwan Normal University

References:

1. *Modern Information Retrieval*, Chapter 3
2. *Introduction to Information Retrieval*, Chapter 8
3. *Search Engines: Information Retrieval in Practice,* Chapter 8

Introduction

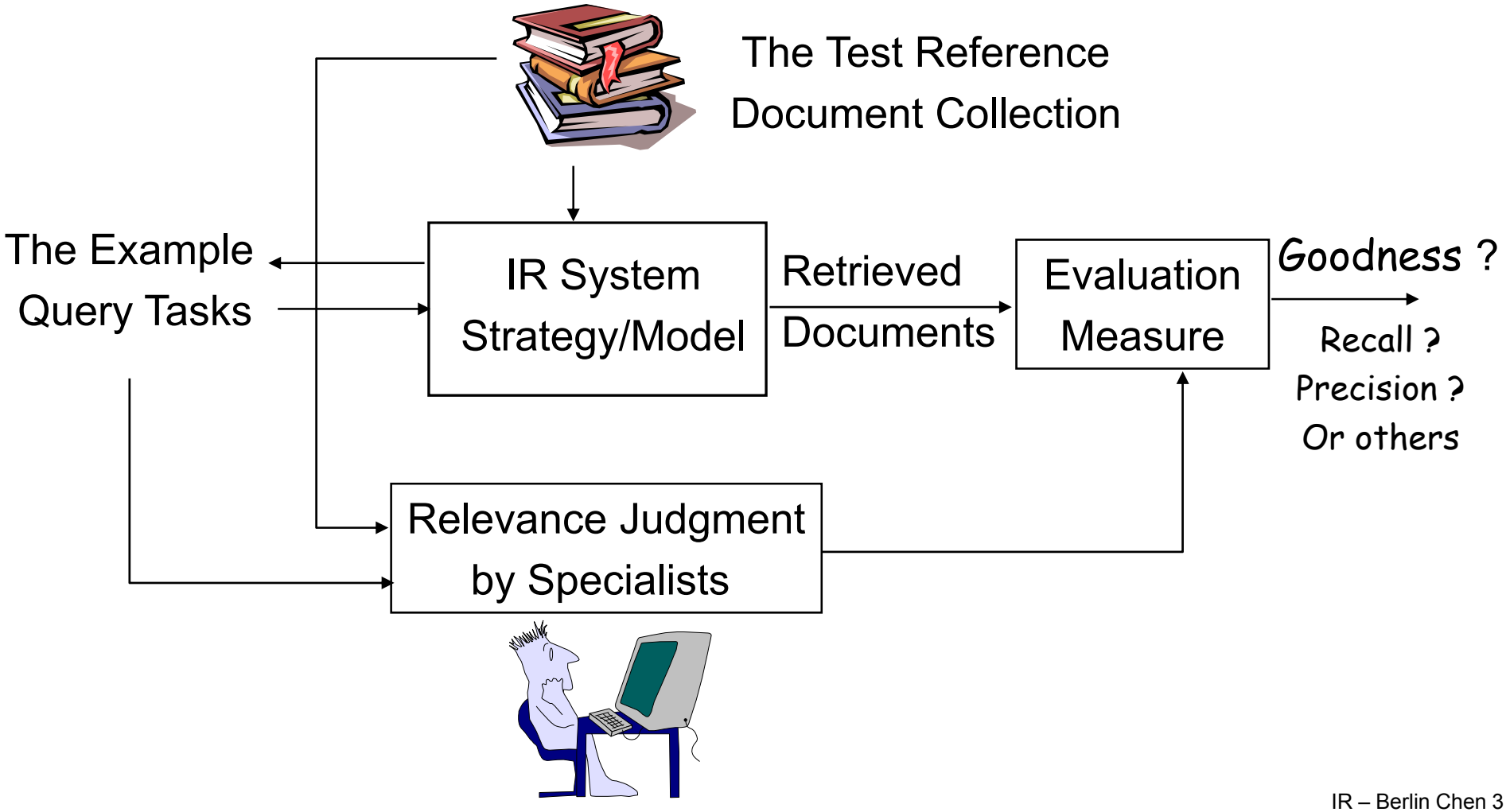
- Functional analysis
  - Functionality test or error analysis instead
- Performance evaluation
  - E.g.: **Data retrieval system**
    - The shorter the response time, the smaller the space used, the better the system is
    - Tradeoff between time and space
- Retrieval performance evaluation
  - E.g.: **information retrieval system**
    - Relevance of retrieved documents is important, besides time and space (quality of the answer set)
  - Discussed here !

Different objectives

In IR, since the user's query is inherently vague, the retrieved documents are not exact answers and have to be ranked according to their relevance to the query.

Introduction (cont.)

- Retrieval performance evaluation (cont.)



Introduction (cont.)

- Test Reference Collection
  - A collection of documents
  - A set of example information requests (queries)
  - A set of relevant documents for each information request

- Evaluation Measure
  - Qualify the similarity between the set of documents retrieved and the set of relevant documents provided by the specialists (assessors)
  - Provide an estimation of the **goodness** of the retrieval strategy

Batch and Interactive Mode

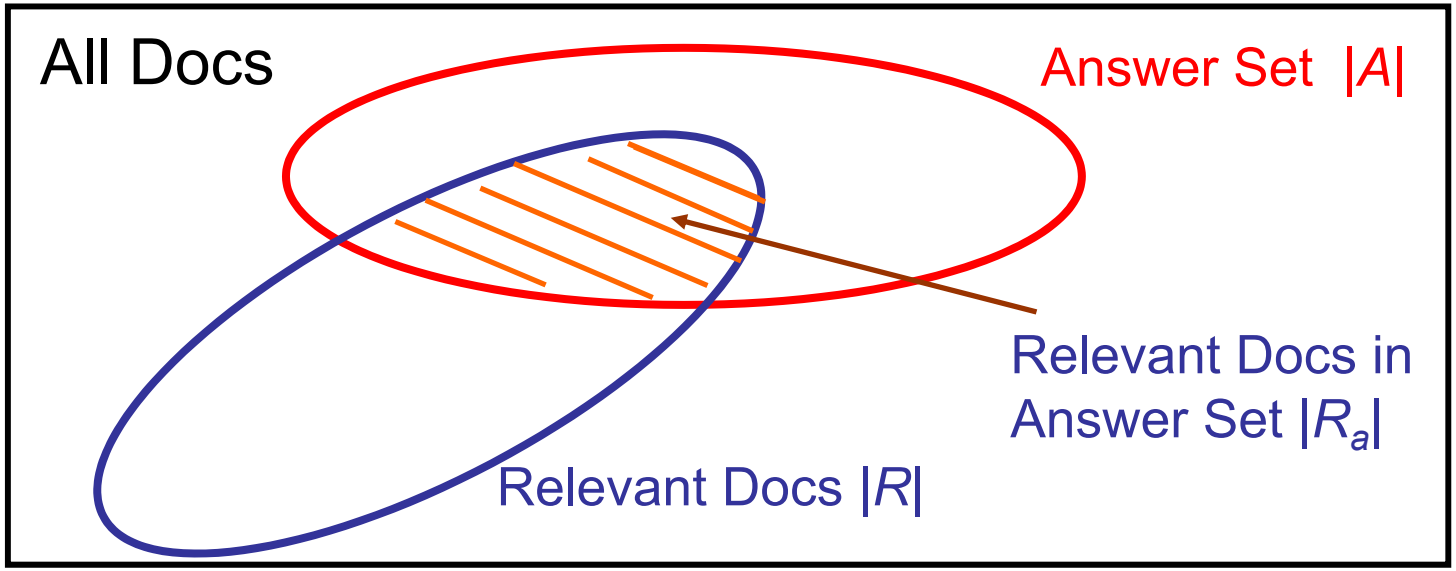
Consider retrieval performance evaluation

- Batch mode (laboratory experiments)
  - The user submits a query and receives an answer back
  - **Measure**: the quality of the generated answer set
  - Still the dominant evaluation (Discussed here !)
    - Main reasons: repeatability and scalability

- Interactive mode (real life situations)
  - The user specifies his information need through a series of interactive steps with the system
  - **Measure**: user effort, interface design, system's guidance, session duration
  - Get a lot more attention since 1990s

Recall and Precision

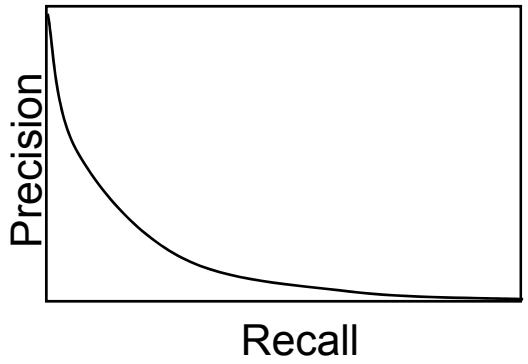
- Recall ($\frac{|R_a|}{|R|}$)
  - The fraction of the relevant documents which has been retrieved
- Precision ($\frac{|R_a|}{|A|}$)
  - The fraction of the retrieved documents which is relevant



Recall and Precision (cont.)

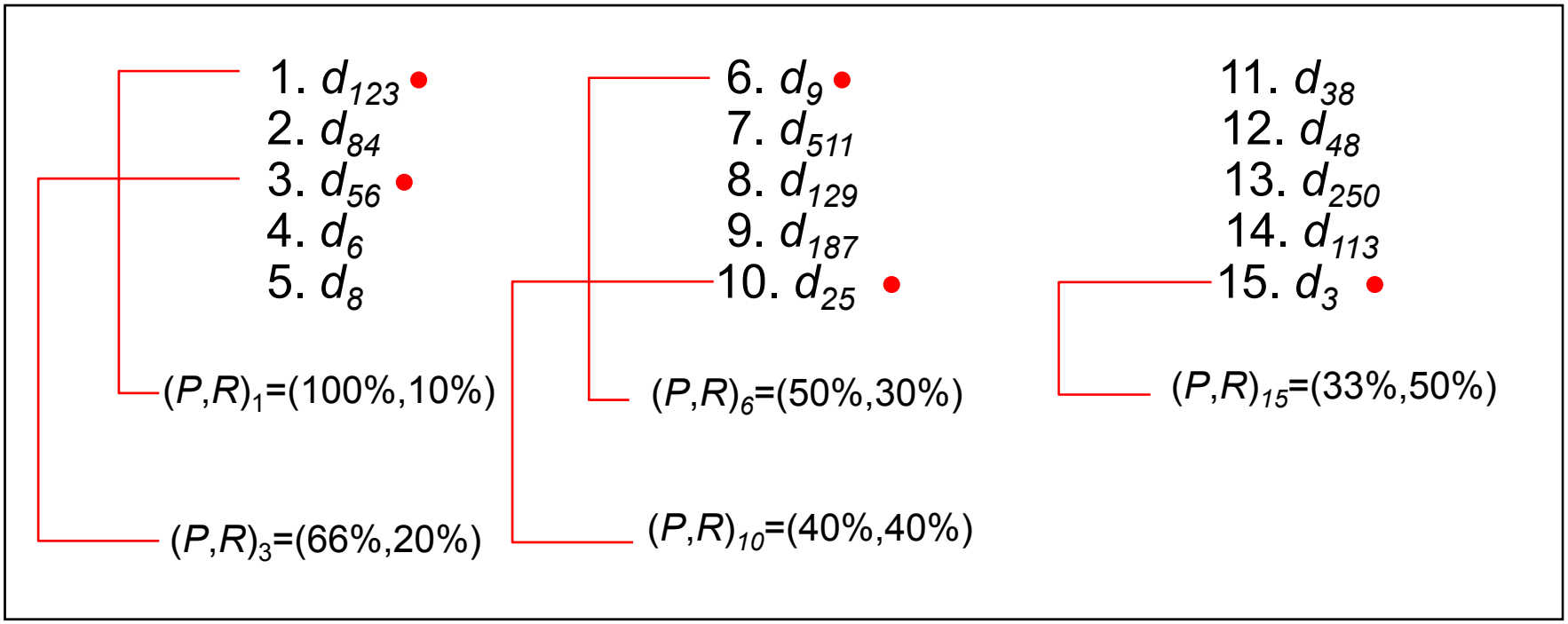
- Recall and precision assume that all the documents in the answer set have been examined (or seen)

- However, the user is not usually presented with all the documents in the answer set A at once
  - Sort the document in A according to a degree of relevance
  - Examine the ranked list starting from the top document (*increasing in recall, but decreasing in precision*)
    - Varying of recall and precision measures
    - A precision versus recall curve can be plotted



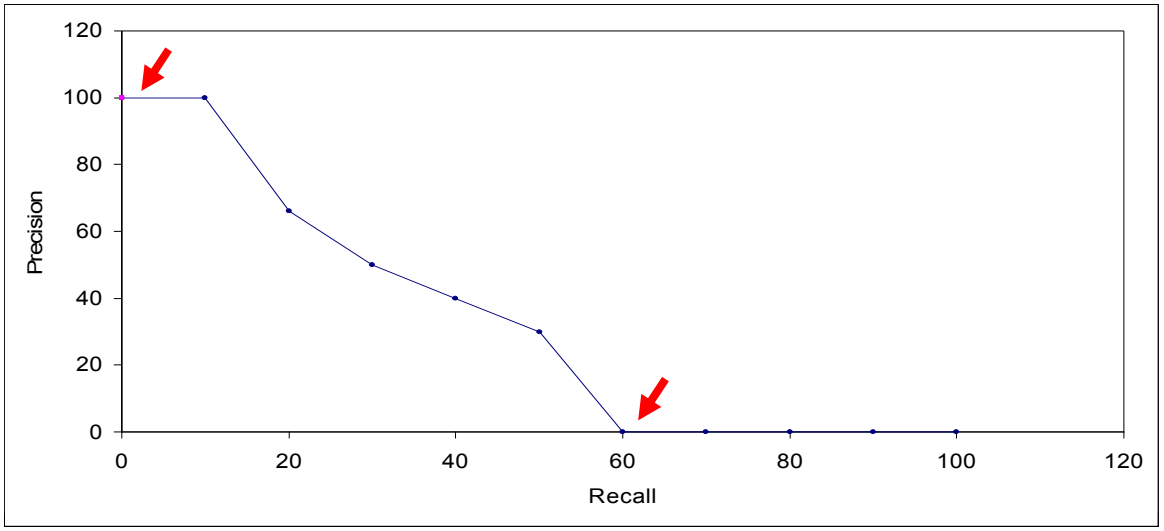
Recall and Precision (cont.)

- Example 3.2
  - $R_q=\{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$
    - Ten relevant documents, five included in Top 15
  - A ranking of the documents for the given query *q*



Recall and Precision (cont.)

- Example 3.2 (count.)



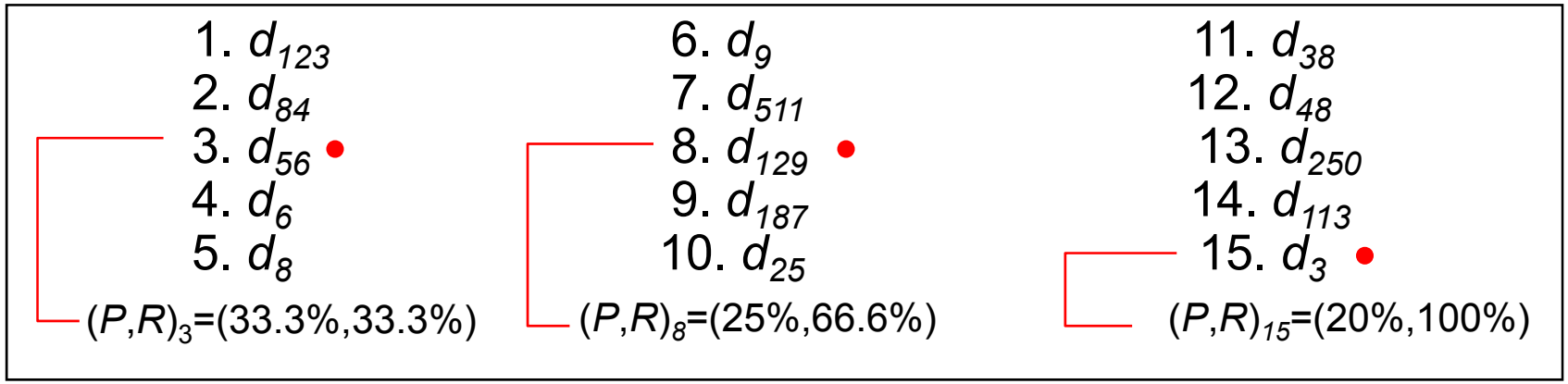
  - The precision versus recall curve is usually plotted based on 11 standard recall levels: 0%,10%,….,100%
  - In this example
    - The precisions for recall levels higher than 50% drop to 0 because no relevant documents were retrieved
    - There was an interpolation for the recall level 0%

Interpolated Recall-Precision Curve

- Since the recall levels for each query might be distinct from the 11 standard recall levels

  Salton, 1983

  – Utilization of an interpolation procedure is necessary !
- Example 3.3
  – $R_q=\{d_3, d_{56}, d_{129}\}$
    - Three relevant documents



- How about the precisions at recall levels 0%, 10%,… ,90%

Interpolated Recall-Precision Curve (cont.)

- Interpolated Precisions at standard recall levels

$$\overline{P}(r_j) = \max_{r_j \le r \le r_{j+1}} P(r)$$

  - the *j*-th standard recall level (e.g., r_5 is recall level 50%)
- Example 3.3 (cont.)

$(P,R)_3$=(33.3%,33.3%)

$(P,R)_8$=(25%,66.6%)

$(P,R)_{15}$=(20%,100%)

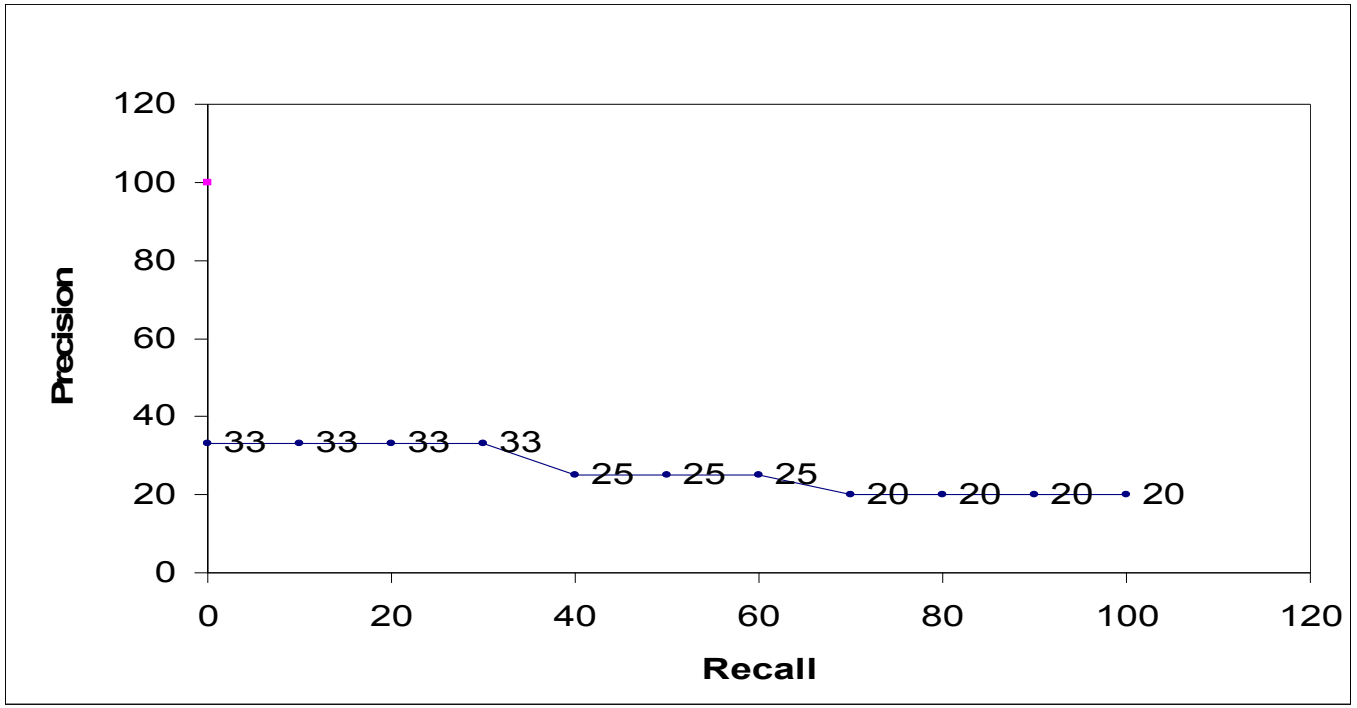
Precision	Recall
33.3%	0%
33.3%	10%
33.3%	20%
33.3%	30%
25%	40%
25%	50%
25%	60%
20%	70%
20%	80%
20%	90%
20%	100%

$$\overline{P}_i(r_j) = \max_{r_j \le r \le r_{j+1}} P_i(r)$$

query *i*

Interpolated Recall-Precision Curve (cont.)

- Example 3.3 (cont.)
  - Interpolated precisions at 11 standard recall levels



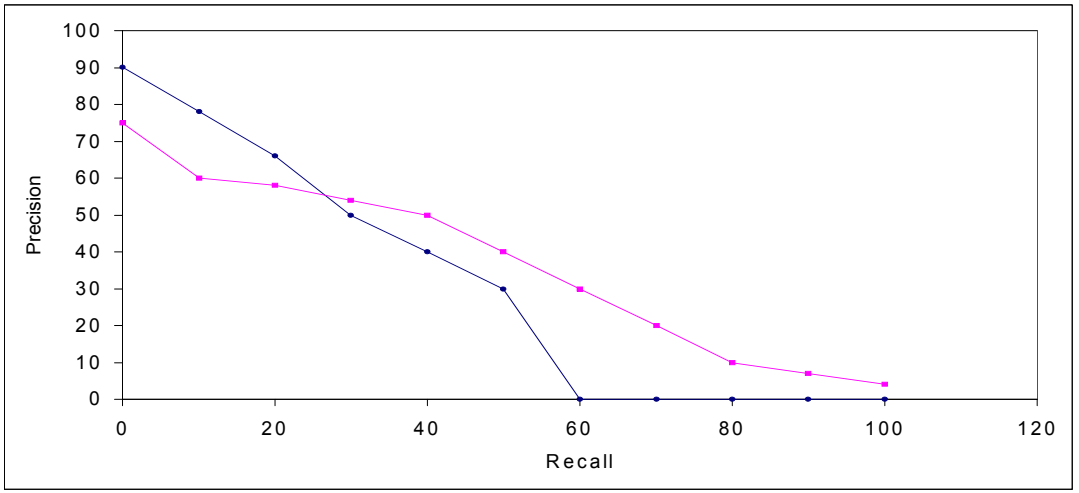
Interpolated Recall-Precision Curve (cont.)

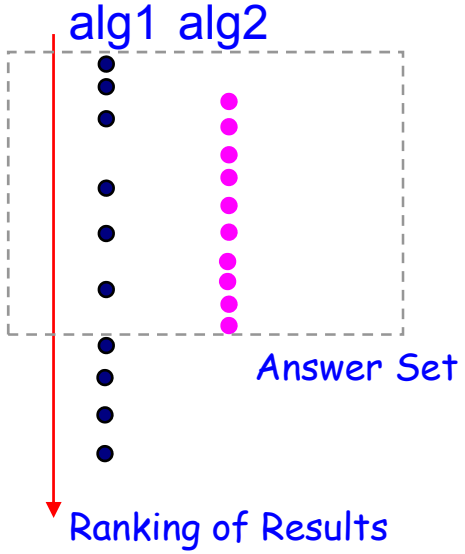
- Evaluate (average) the retrieval performance over all queries

$$\overline{P}_{all}(r_j) = \frac{1}{N_q}\sum_{i=1}^{N_q}\overline{P}_i(r_j)$$

On different recall levels

- Example 3.4: average interpolated recall-precision curves for two distinct retrieval algorithms

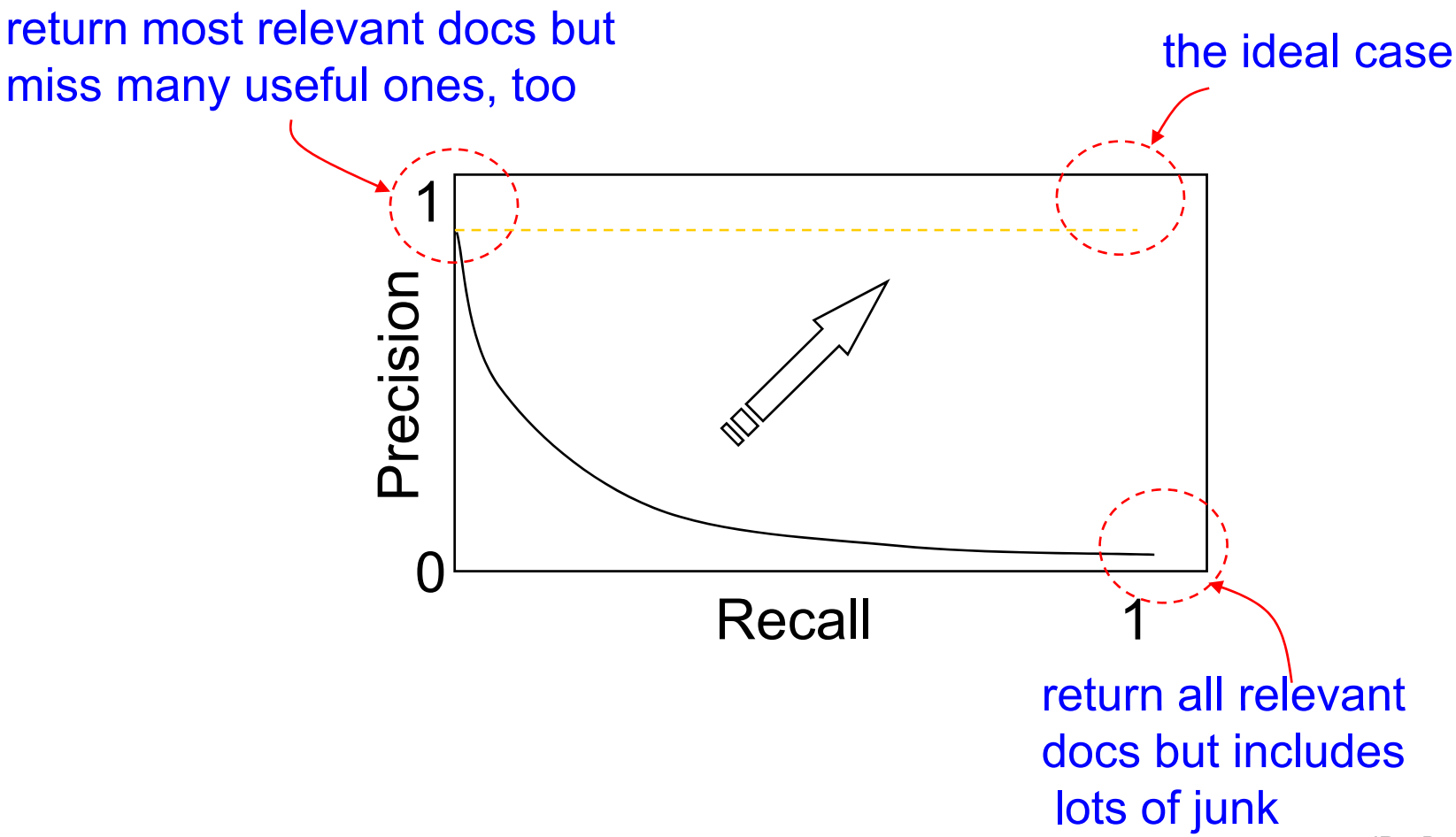




  - Difficult to determine which of these two results is better

Interpolated Recall-Precision Curve (cont.)

- Trade-off between Recall and Precision



Interpolated Recall-Precision Curve (cont.)

- Alternative: average precision at a given document cutoff values (levels)
  - E.g.: compute the average precision when top 5, 10, 15, 20, 30, 50 or 100 relevant documents have been seen
  - Focus on how well the system ranks the top *k* documents
    - Provide additional information on the retrieval performance of the ranking algorithm
  - We can take (weighted) average over results

Interpolated Recall-Precision Curve (cont.)

- Advantages
  - Simple, intuitive, and combined in single curve
  - Provide quantitative evaluation of the answer set and comparison among retrieval algorithms
  - *A standard evaluation strategy* for IR systems

- Disadvantages
  - Can't know *true recall value* except in small document collections (*document cutoff levels are needed*!)
  - Assume a strict document rank ordering

Single Value Summaries

- Interpolated recall-precision curve
  - Compare the performance of retrieval algorithms over a set of example queries
    - Might disguise the important anomalies
  - How is the performance for each individual query ?

- A single precision value (for each query) is used instead
  - Interpreted as a summary of the corresponding precision versus recall curve
    - Just evaluate the precision based on the top 1 relevant document ?
    - Or averaged over all relevant documents

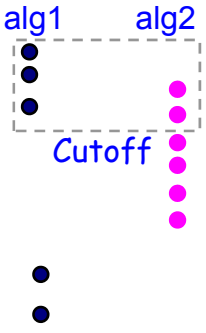
Single Value Summaries (cont.)

- Method 1: Average Precision at Seen Relevant Documents
  - A single value summary of the ranking by averaging the precision figures obtained after each new relevant doc is observed

1. d_{123} ● (P=1.0)	6. d_9 ● (P=0.5)	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56} ● (P=0.66)	8. d_{129}	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d_8	10. d_{25} ● (P=0.4)	15. d_3 ● (P=0.3)

(1.0+0.66+0.5+0.4+0.3)/5=0.57

  - It favors systems which retrieve relevant docs quickly (early in the ranking), i.e., this measure depends heavily on highly ranked relevant documents
  - But when doc cutoff levels were used
    - An algorithm might present a good average precision at seen relevant docs but have a poor performance in terms of overall recall



Mean Average Precision (*m*AP)

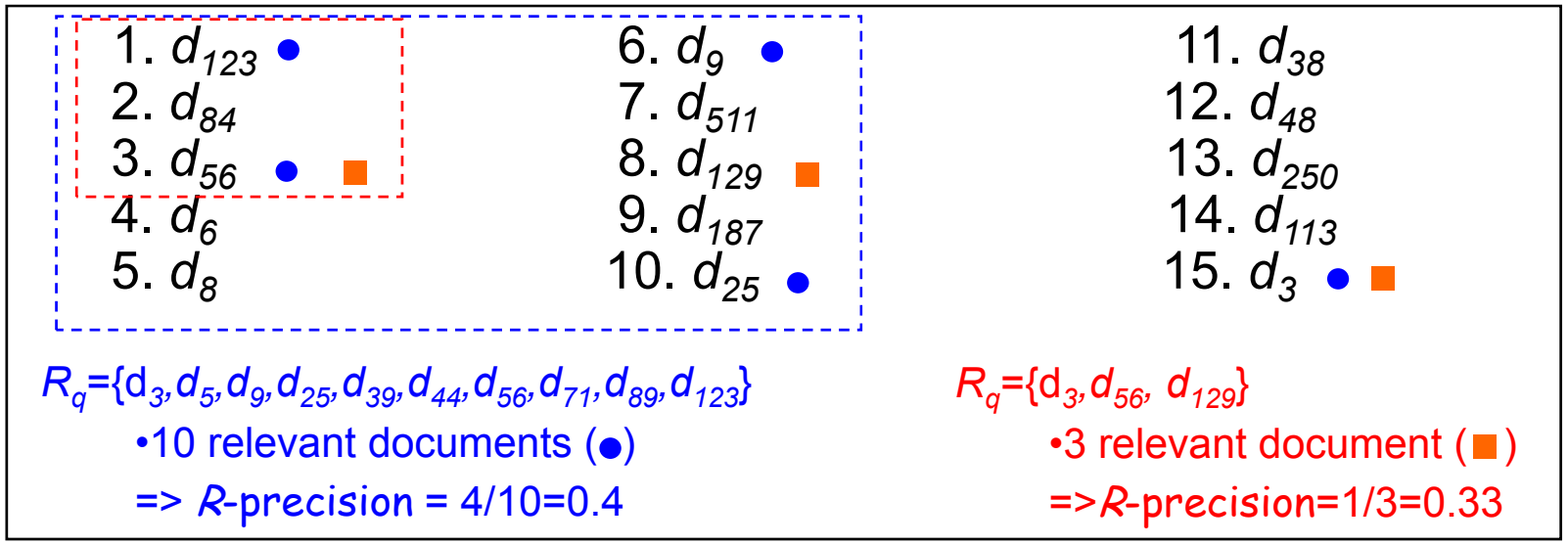
- Averaged at relevant docs and across queries
  - E.g. relevant docs ranked at 1, 5, 10, precisions are 1/1, 2/5, 3/10,
    - non-interpolated average precision (or called *Average Precision at Seen Relevant Documents* in textbook) =(1/1+2/5+3/10)/3
  - Mean average Precision (denoted as *m*AP or MAP)

$$\frac{1}{|Q|}\sum_{q=1}^{|Q|}(\text{non}-\text{interpolated average precision})_q$$

- Widely used in IR performance evaluation

Single Value Summaries (cont.)

- Method 2: R-Precision
  - Generate a single value summary of ranking by computing the precision at the *R*-th position in the ranking
    - Where *R* is the total number of relevant docs for the current query



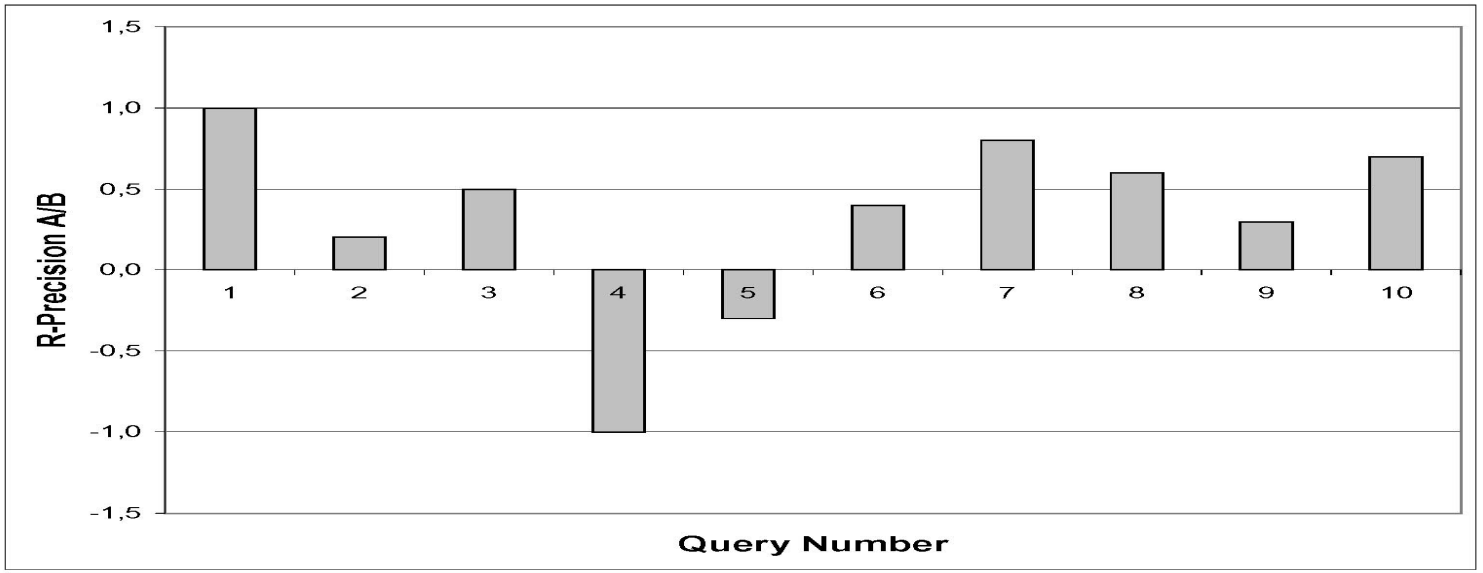
Single Value Summaries (cont.)

- Method 3: Precision Histograms
  - Compare the retrieval history of two algorithms using the R-precision graph for several queries
    - A visual inspection
  - Example 3.5
    - Algorithms *A, B*
    - The difference of R-precision for the *i*-th query:

$$RP_{A/B}(i) = RP_A(i) - RP_B(i)$$

Single Value Summaries (cont.)

- Method 3: Precision Histograms (cont.)
  - Example 3.5 (cont.)



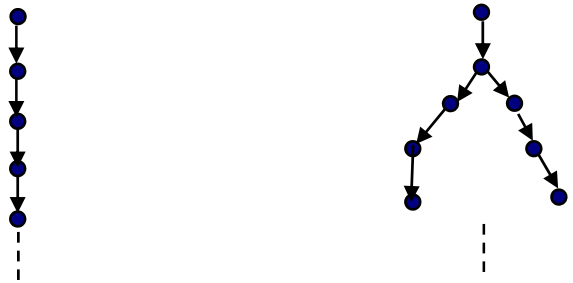
- A positive $RP_{A/B}(i)$ indicates that the algorithm A is better than B for the i-th query and vice versa

Single Value Summaries (cont.)

- Method 4: Summary Table Statistics
  - A statistical summary regarding the set of all the queries in a retrieval task
    - The number of queries used in the task
    - The total number of documents retrieved by all queries
    - The total number of relevant documents which were effectively retrieved when all queries are considered
    - The total number of relevant documents which could have been retrieved by all queries
    - …

Precision and Recall Appropriateness

- The proper estimation of maximal recall requires knowledge of all the documents in the collection
- Recall and precision are related measures which capture different aspects of the set of retrieved documents
- Recall and precision measure the effectiveness over queries in batch mode
- Recall and precision are defined under the enforcement of linear ordering of the retrieved documents
  - Partial Ordering ?



Alternative Measures

- Method 1: The Harmonic Mean (F Measure)
  - The harmonic mean *F* of recall and precision

$$F(j) = \frac{2}{\frac{1}{r(j)} + \frac{1}{P(j)}} = \frac{2 \cdot P(j) \cdot r(j)}{P(j) + r(j)}$$

    - *r*(*j*): the recall for the *j*-th document in the ranking
    - *P*(*j*): the precision for the *j*-th document in the ranking
  - Characteristics
    - F = 0: no relevant documents were retrieved
    - F = 1: all ranked documents are relevant
    - A high F achieved only when both recall and precision are high
    - Determination of the maximal F can be interpreted as an attempt to find the best possible compromise between recall and precision

Harmonic mean emphasizes the importance of small values, whereas arithmetic mean is affected by large values.

Alternative Measures (cont.)

- Method 2: The E Measure

van Rijsbergen 1979

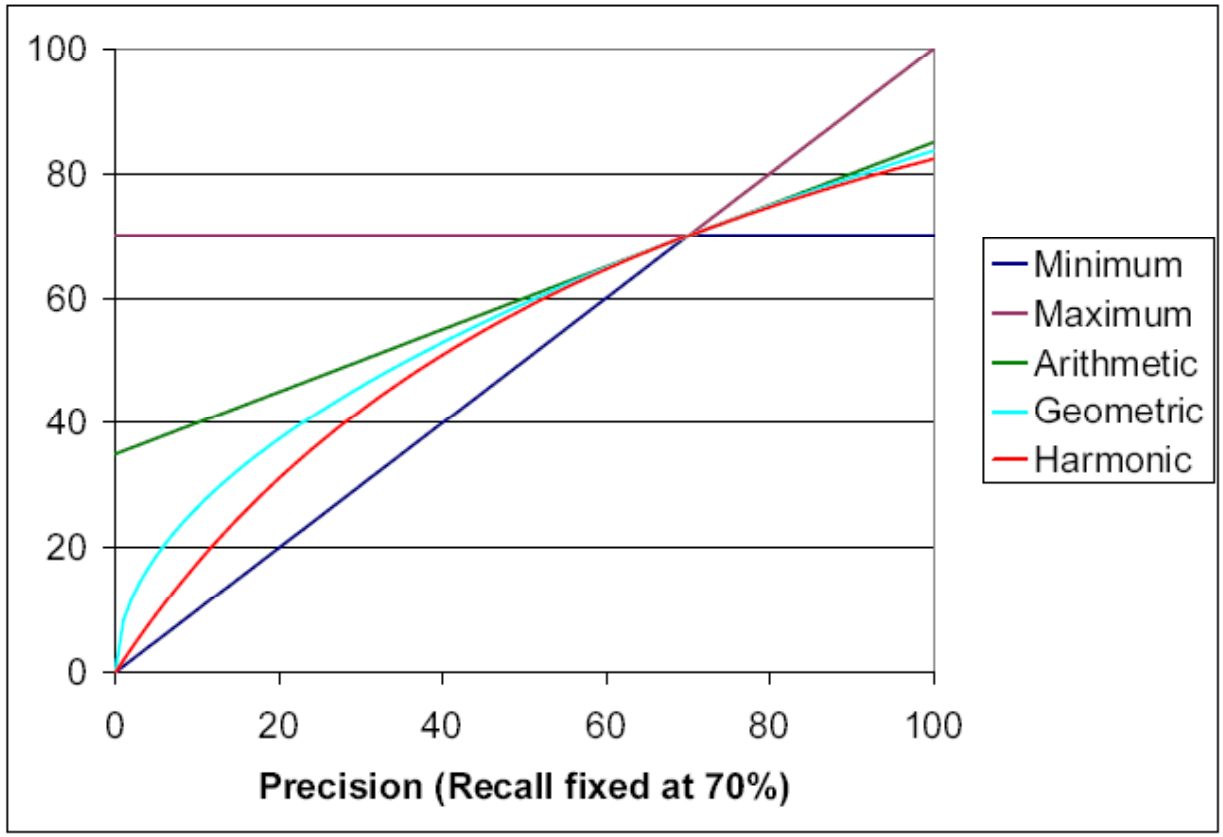
  - Another measure which combines recall and precision
  - Allow the user to specify whether he is more interested in recall or precision

$$E(j) = 1 - \frac{1 + b^2}{\frac{b^2}{r(j)} + \frac{1}{P(j)}} = 1 - \frac{(1 + b^2) \cdot P(j) \cdot r(j)}{b^2 \cdot P(j) + r(j)}$$

  - Characteristics
    - $b = 1$: act as the complement of F Measure
    - $b > 1$: more interested in recall
    - $b < 1$: more interested in precision

Wrong statements in the Textbook!

Arithmetic/Geometric/Harmonic Means



► **Figure 8.1** Graph comparing the harmonic mean to other means. The graph shows a slice through the calculation of various means of precision and recall for the fixed recall value of 70%. The harmonic means is always less than either the arithmetic or geometric mean, and often quite close to the minimum of the two numbers. When the precision is also 70%, all the measures coincide.

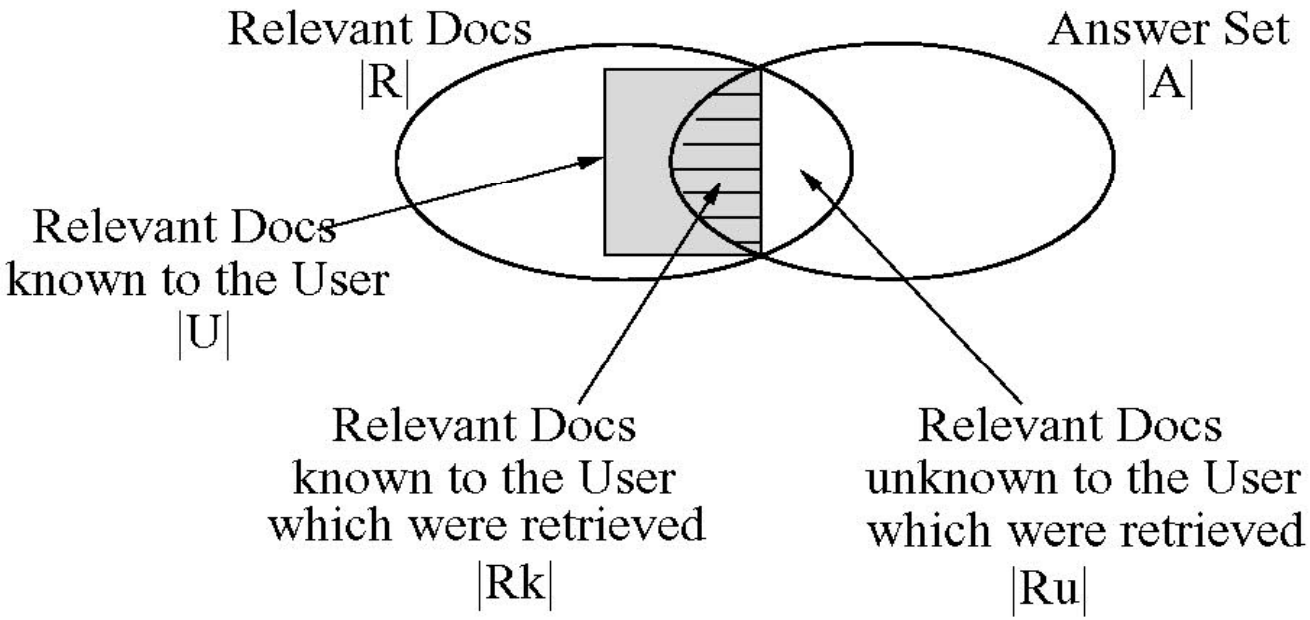
Alternative Measures (cont.)

- Method 3: User-Oriented Measures
  - Problematic assumption of recall and precision
    - The set of relevant documents for a query is the same, independent of the user
  - However, different users have a different interpretation of document relevance

  - User-oriented measures are therefore proposed
    - Coverage ratio
    - Novelty ratio
    - Relative recall
    - Recall effect

Alternative Measures (cont.)

- Method 3: *User-Oriented Measures* (cont.)



- **Coverage ratio** = $\frac{|Rk|}{|U|}$

- Relative recall = $\frac{|R_k|+|Ru|}{|U|}$

- **Novelty ratio** = $\frac{|Ru|}{|Ru|+|Rk|}$

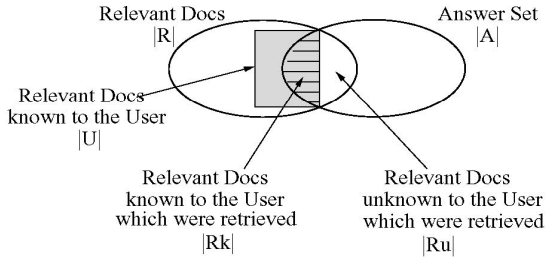
- Recall effect = $\frac{|U|}{|A|}$

Measure the ability to reveal new relevant docs

Alternative Measures (cont.)

- Coverage ratio
  - The fraction of relevant docs **known** to the user which has been retrieved
  - High →find most of the relevant docs user expected to see

$$\frac{|Rk|}{|U|}$$



- Novelty ratio
  - The fraction of relevant docs retrieved which is **unknown** to the user
  - High →find (reveal) many new relevant docs (information) the user previously unknown

$$\frac{|Ru|}{|Ru|+|Rk|}$$

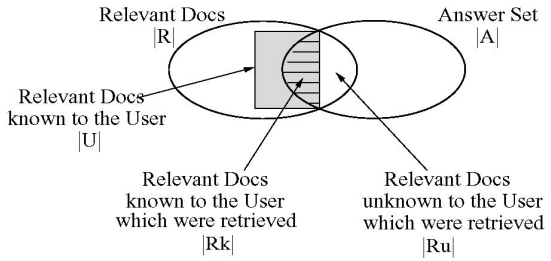
Alternative Measures (cont.)

- Relative recall
  - The ratio between the number of relevant docs found by the system and the number of relevant docs the user expects to find

$$\frac{|R_k|+|Ru|}{|U|}$$

- Recall effect
  - The ratio between the number of relevant docs the user expects to find and the number of docs found by the system

$$\frac{|U|}{|A|}$$



Homework - 1

Homework #1 :Evaluation Measures

The the query-document relevance information (AssessmentTrainSet.txt) for a set of queries (16 queries) and a collection of 2,265 documents is provided. An IR model is then tested on this query set and save the corresponding ranking results in a file (ResultsTrainSet.txt) . Please evaluate the overall model performance using the following two measures.

1. Interpolated Recall-Precision Curve:

$$\overline{P}_i(r_j) = \max_{r_j \le r \le r_{j-1}} P_i(r)$$ (for each query)

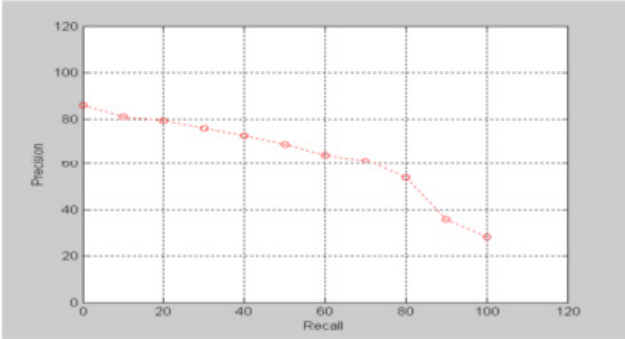
$$\overline{P}_{all}(r_j) = \frac{1}{N_q}\sum_{i=1}^{N_q} \overline{P}_i(r_j)$$ (overall performance)

2. (Non-interpolated) Mean Average Precision:

$$\frac{1}{|Q|}\sum_{q=1}^{|Q|}(\text{non}-\text{interpolated average precision})_q$$

, where "non-interpolated average precision" is "average precision at seen relevant documents" introduced in the textbook.

Example 1: Interpolated Recall-Precision Curve



Example 2: (Non-interpolated) Mean Average Precision

mAP=0.63787418