Retrieval Evaluation

- Measures

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

References:

1. *Modern Information Retrieval*, Chapter 4 & Teaching material
2. *Introduction to Information Retrieval*, Chapter 8
3. *Search Engines: Information Retrieval in Practice,* Chapter 8

Evaluation in General

- 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 evaluation
  - E.g.: **information retrieval system**
    - Relevance of retrieved documents is important, besides time and space (**quality of the answer set**)
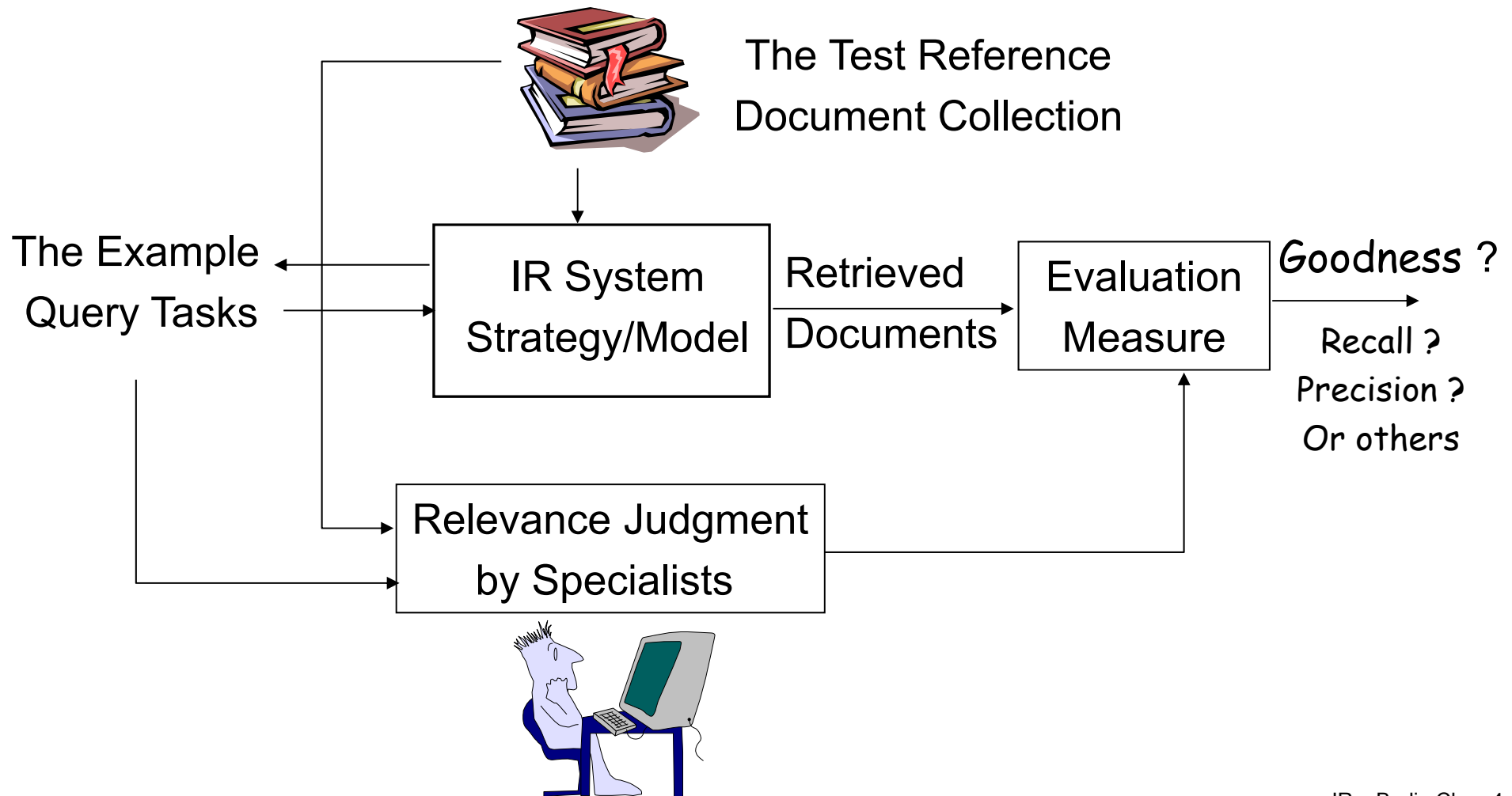 - Discussed here !

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.

Retrieval Evaluation

- To measure how well the system meets the information needs of the users
  - We know that a same result set might be interpreted differently by distinct users
  - However, it is possible to define a quantitative (numeric) **metric** that, on average, have a correlation with the preferences of a population of users

- A common approach to compute such a **metric** is to compared the results produced by the system with the results suggested by humans for the same set of queries

- Retrieval evaluation is a critical and integral component of any modern IR system

Retrieval Evaluation (cont.)

- A pictorial representation

The Test Reference
Document Collection

The Example
Query Tasks

IR System
Strategy/Model

Retrieved
Documents

Evaluation
Measure

Goodness ?

Recall ?
Precision ?
Or others

Relevance Judgment
by Specialists

Evaluation for IR (cont.)

- Test Reference Collection
  - A collection of documents
  - A set of example information requests (queries)
  - A set of relevant documents for each information request
    (or a relevance judgment for each document-query pair)

- 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, or the context in which the query is posed
 - Get a lot more attention since 1990s

Recall and Precision

- **Recall** $\left(\dfrac{|R_a|}{|R|} \right)$
  - The fraction of the relevant documents which has been retrieved

- **Precision** $\left(\dfrac{|R_a|}{|A|} \right)$
  - The fraction of the retrieved documents which is relevant

All Docs

Answer Set $|A|$

Relevant Docs in Answer Set $|R_a|$

Relevant Docs $|R|$
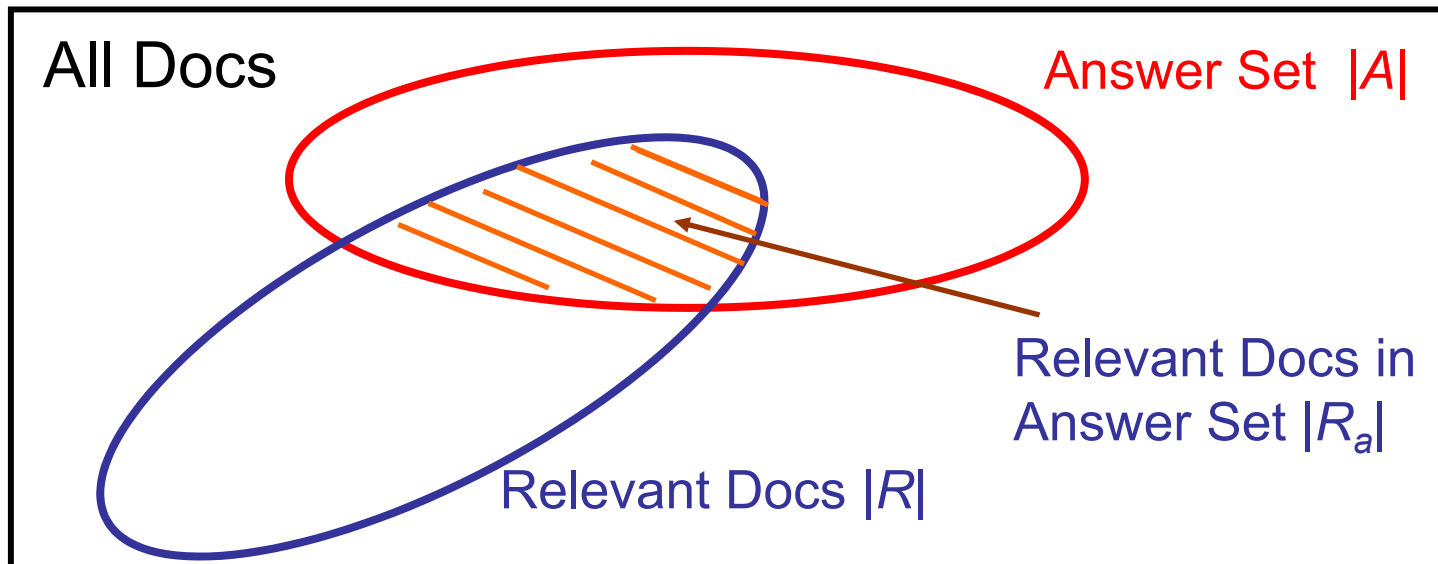
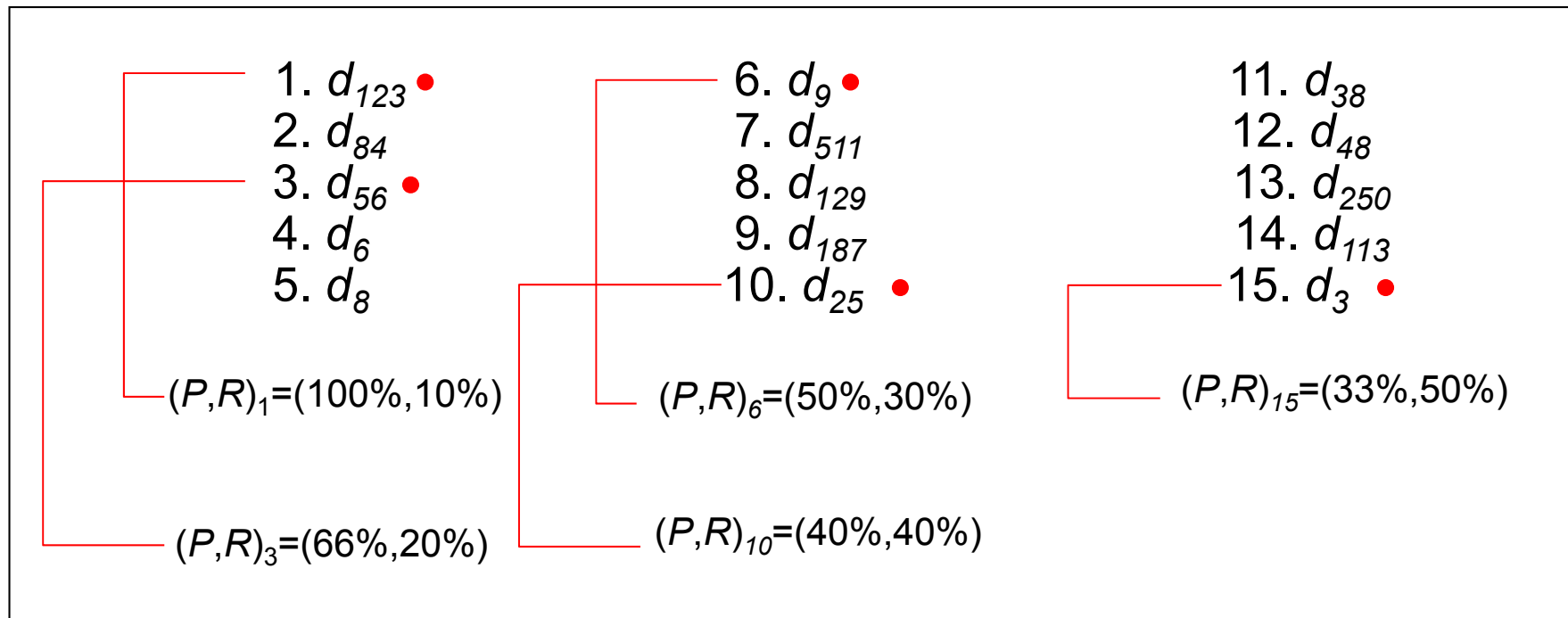# Recall and Precision (cont.)

- 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
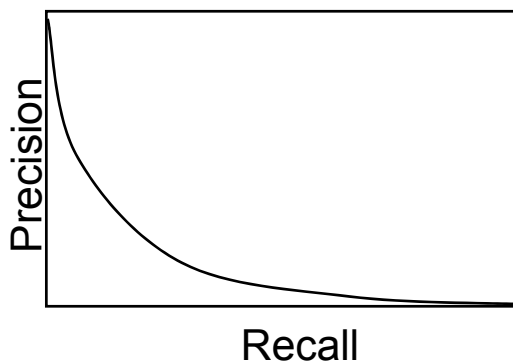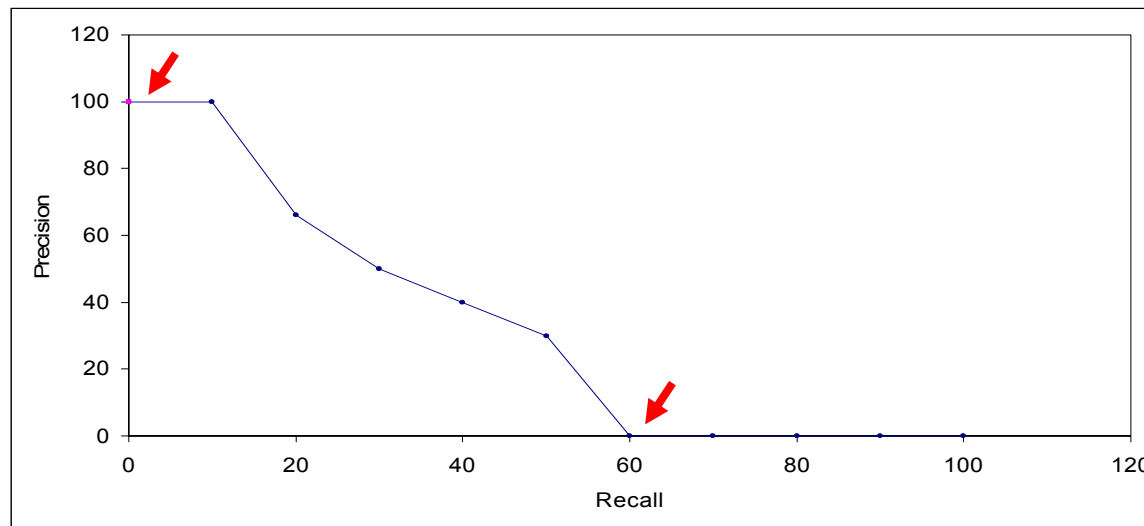
Recall and Precision (cont.)

- 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

1. d_{123} ●
2. d_{84}
3. d_{56} ●
4. d_6
5. d_8

$(P,R)_1 = (100\%, 10\%)$

$(P,R)_3 = (66\%, 20\%)$

6. d_9 ●
7. d_{511}
8. d_{129}
9. d_{187}
10. d_{25} ●

$(P,R)_6 = (50\%, 30\%)$

$(P,R)_{10} = (40\%, 40\%)$

11. d_{38}
12. d_{48}
13. d_{250}
14. d_{113}
15. d_3 ●

$(P,R)_{15} = (33\%, 50\%)$

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%
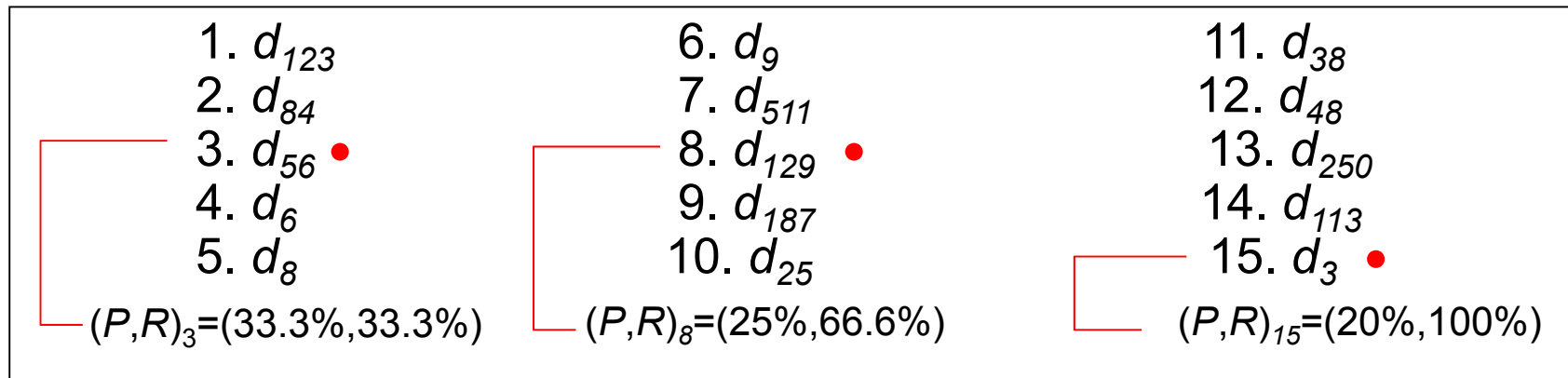
Interpolated Recall-Precision Curve

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

  - Utilization of an interpolation procedure is necessary !

- Example 3.3

  - $R_q=\{d_3, d_{56}, d_{129}\}$

    - Three relevant documents

1. d_{123}	6. d_9	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56} ●	8. d_{129} ●	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d_8	10. d_{25}	15. d_3 ●
$(P,R)_3=(33.3\%, 33.3\%)$	$(P,R)_8=(25\%, 66.6\%)$	$(P,R)_{15}=(20\%, 100\%)$

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

Interpolated Recall-Precision Curve (cont.)

- Interpolated Precisions at standard recall levels

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

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

- Example 3.3 (cont.)

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

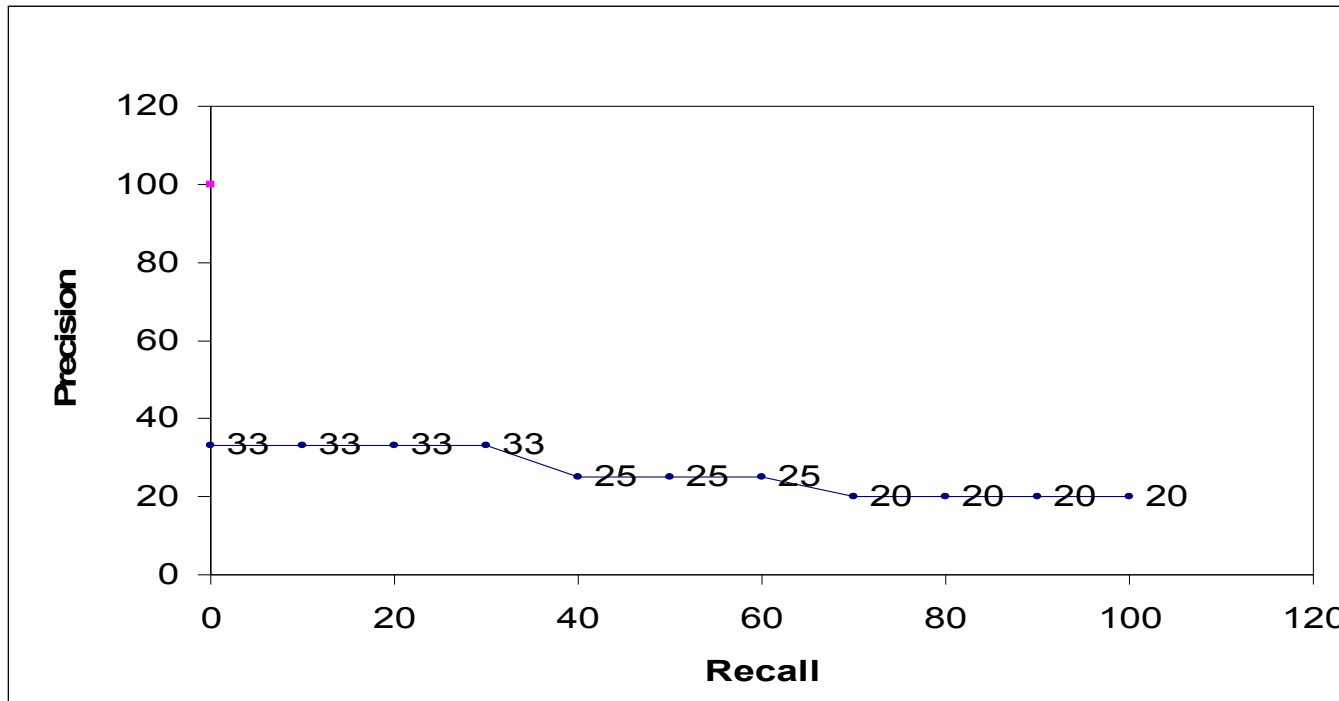$(P,R)_8 = (25\%, 66.6\%)$

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

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

query *i*

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%

Interpolated Recall-Precision Curve (cont.)

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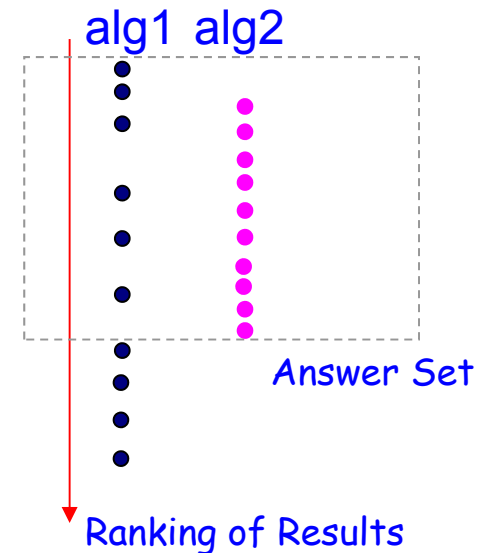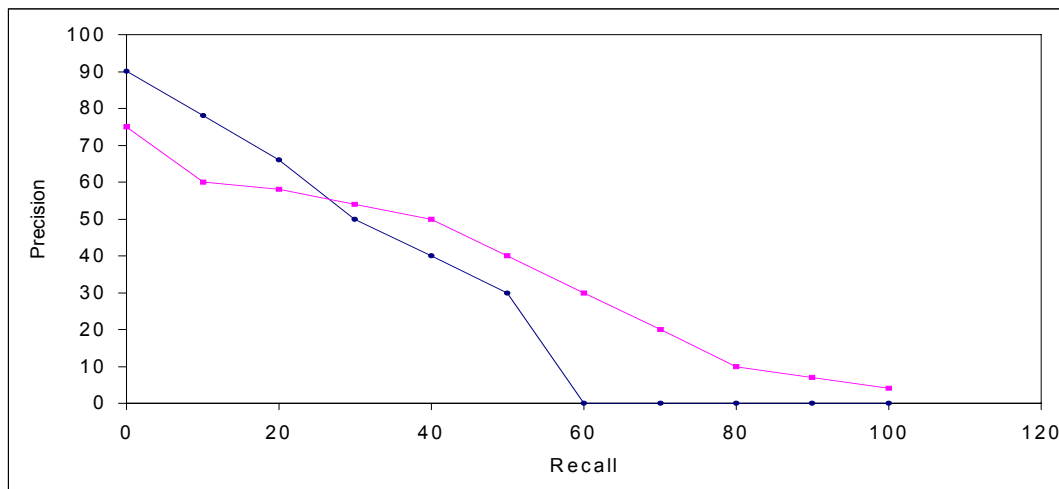 - Interpolated precisions at 11 standard recall levels

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

alg1 alg2
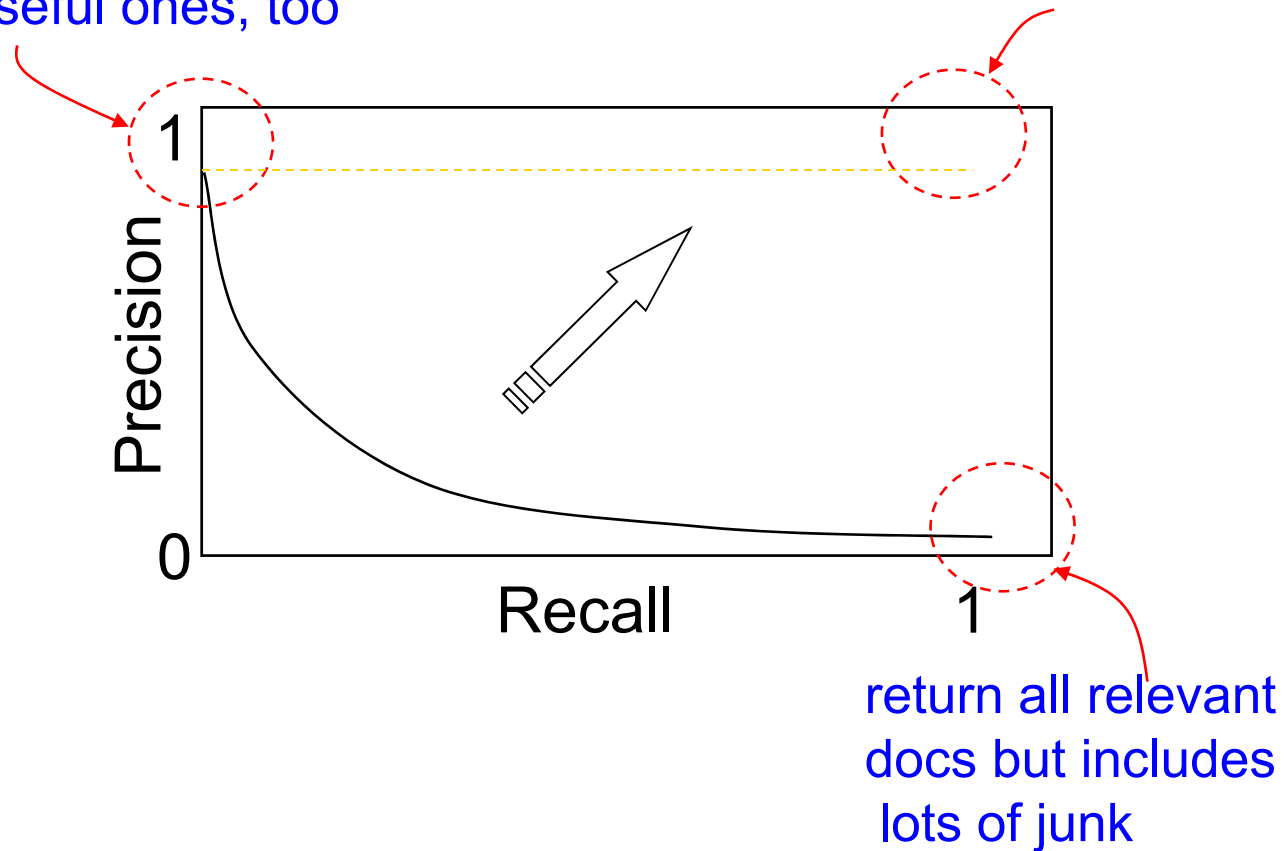

Answer Set

Ranking of Results

- Difficult to determine which of these two results is better

Interpolated Recall-Precision Curve (cont.)

- Trade-off between Recall and Precision

return most relevant docs but miss many useful ones, too

the ideal case

return all relevant docs but includes lots of junk

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
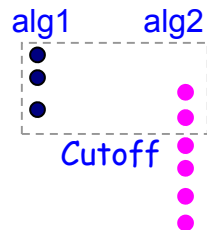
  - But when doc cutoff levels were used

    - An algorithm might present a good average precision at seen relevant docs but have a poor performance in terms of overall recall

alg1   alg2

Cutoff

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

1. d_{123} ●	6. d_9 ●	11. d_{38}
2. d_{84}	7. d_{511}	12. d_{48}
3. d_{56} ● ■	8. d_{129} ■	13. d_{250}
4. d_6	9. d_{187}	14. d_{113}
5. d_8	10. d_{25} ●	15. d_3 ● ■

$R_q = \{d_3, d_5, d_9, d_{25}, d_{39}, d_{44}, d_{56}, d_{71}, d_{89}, d_{123}\}$

•10 relevant documents (●)

=> R-precision = 4/10=0.4

$R_q = \{d_3, d_{56}, d_{129}\}$

•3 relevant document (■)

=> R-precision=1/3=0.33

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

  - Given two 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.)
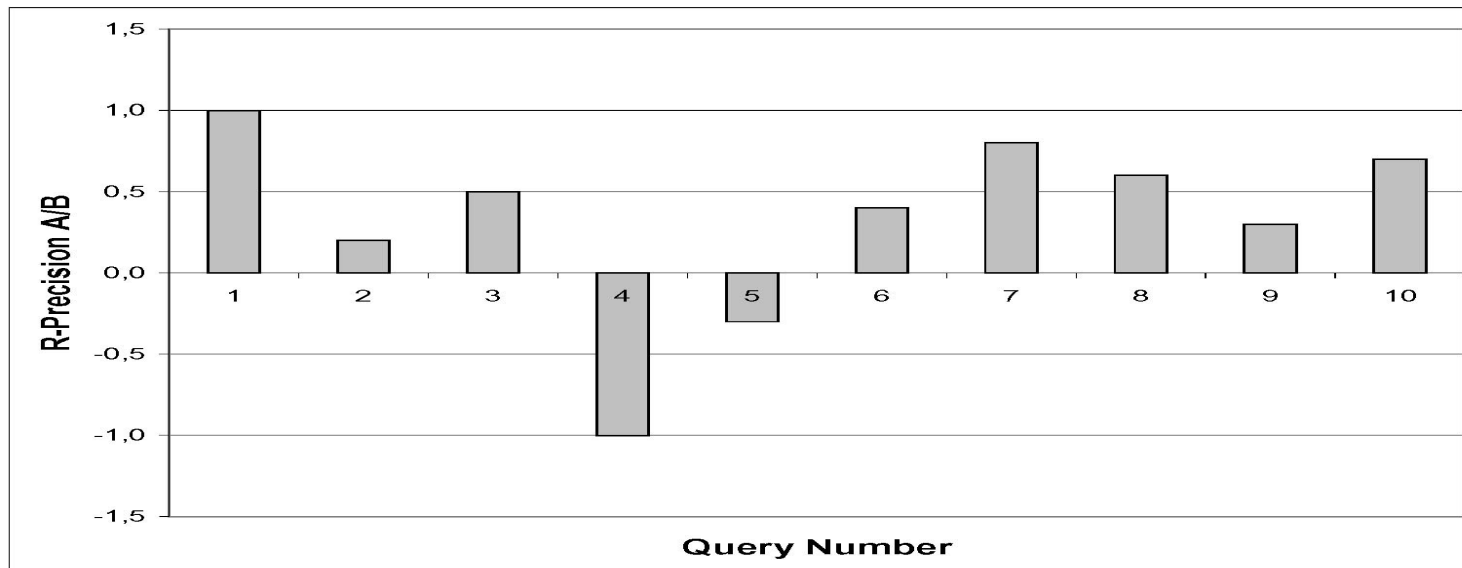
- A positive $RP_{A/B}(i)$ indicates that the algorithm A is better than B for the i-th query and vice versa (0 indicates both algorithms have equivalent retrieval qualities)

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 ?
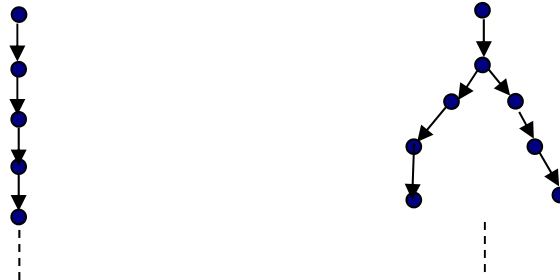
Alternative Measures

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

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

    - $r(j)$: 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

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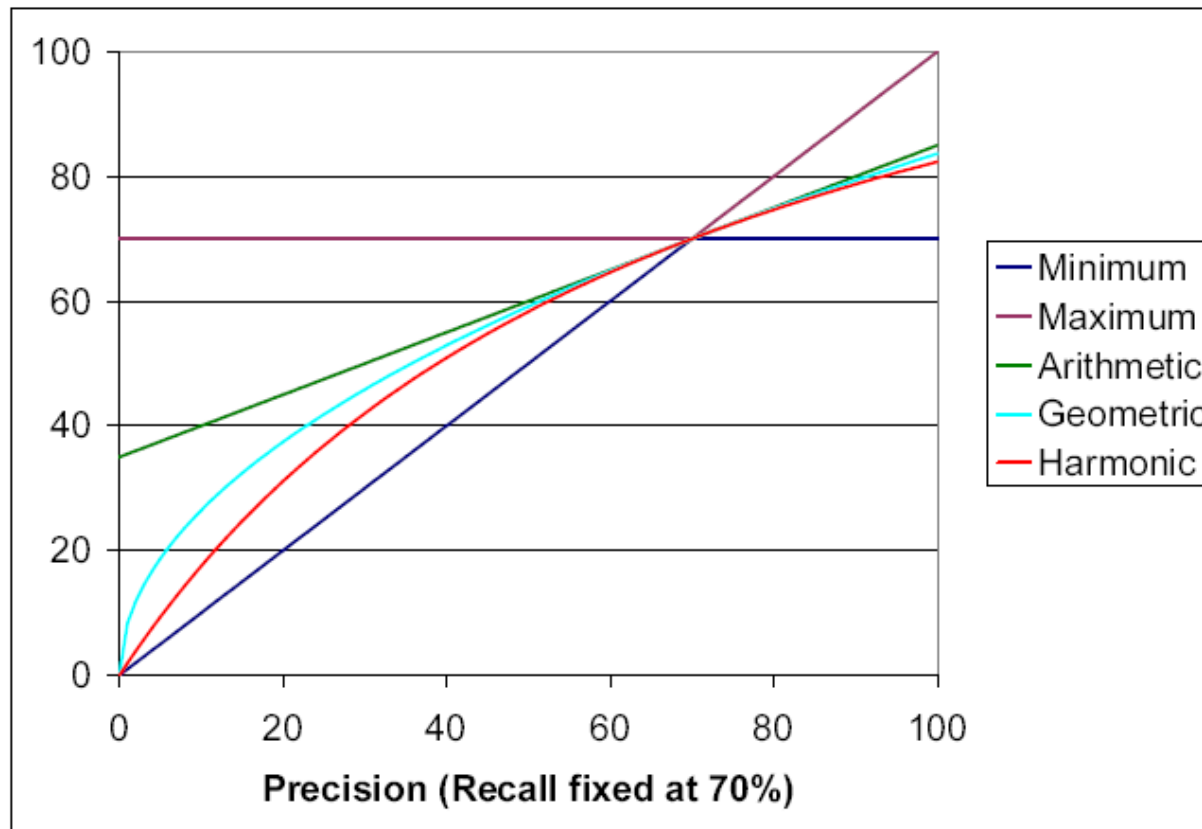 - Allow the user to specify whether he is more interested in recall or precision

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

  - 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!
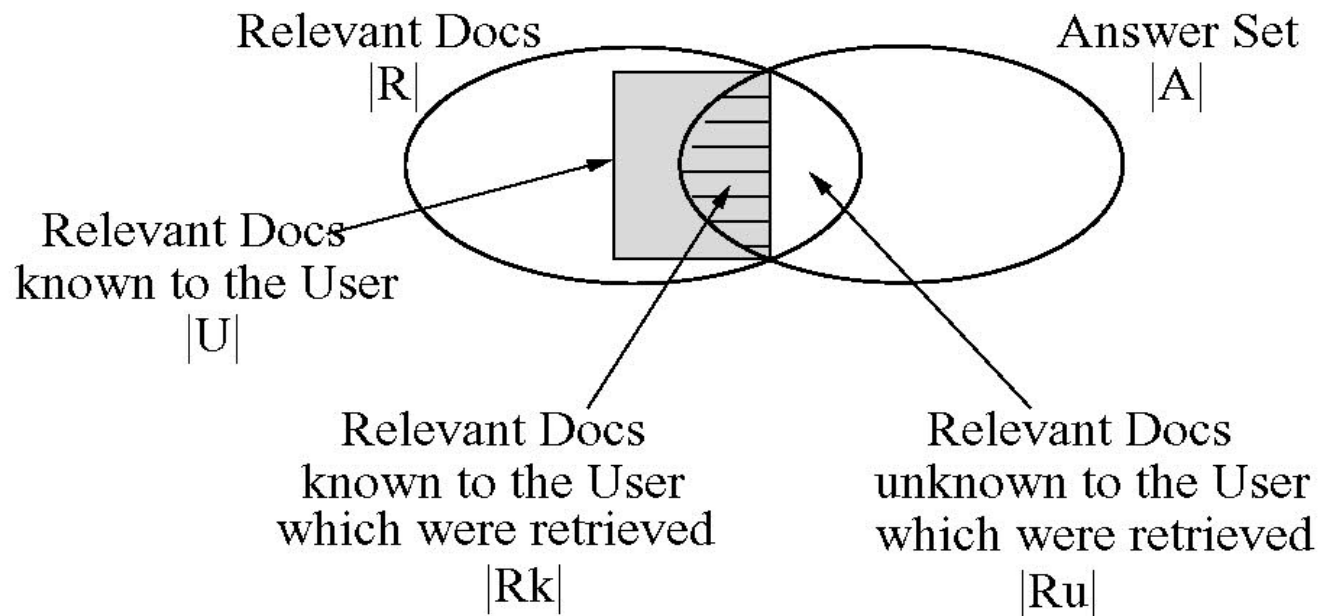
Arithmetic/Geometric/Harmonic Means

► **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.)

Relevant Docs
$|R|$

Answer Set
$|A|$

Relevant Docs
known to the User
$|U|$

Relevant Docs
known to the User
which were retrieved
$|Rk|$

Relevant Docs
unknown to the User
which were retrieved
$|Ru|$

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

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

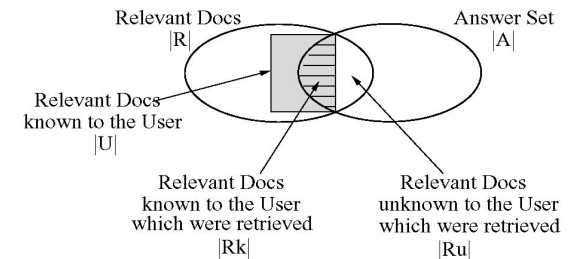- **Novelty ratio** = $\dfrac{|Ru|}{|Ru| + |Rk|}$

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

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

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

Relevant Docs
|R|

Answer Set
|A|

Relevant Docs
known to the User
|U|

Relevant Docs
known to the User
which were retrieved
|Rk|

Relevant Docs
unknown to the User
which were retrieved
|Ru|

- ## 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

$$\frac{|Ru|}{|Ru| + |Rk|}$$
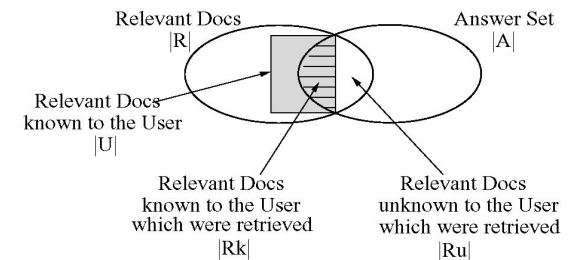
Alternative Measures (cont.)

- ## 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

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

Relevant Docs
|R|

Answer Set
|A|

Relevant Docs
known to the User
|U|

Relevant Docs
known to the User
which were retrieved
|Rk|

Relevant Docs
unknown to the User
which were retrieved
|Ru|

Discounted Cumulated Gain (DCG)

- Precision and recall allow only binary relevance assessments

- As a result, there is no distinction between highly relevant docs and mildly relevant docs

- These limitations can be overcome by adopting graded relevance assessments and metrics that combine them

- The **discounted cumulated gain** (DCG) is a metric that combines graded relevance assessments effectively

Discounted Cumulated Gain (cont.)

- When examining the results of a query, two key observations can be made:

  - Highly relevant documents are preferable at the top of the ranking than mildly relevant ones

  - Relevant documents that appear at the end of the ranking are less valuable

Discounted Cumulated Gain (cont.)

- Consider that the results of the queries are graded on a scale 0–3 (0 for non-relevant, 3 for strong relevant docs)

- For instance, for queries q_1 and q_2, consider that the graded relevance scores are as follows:

$$
\begin{aligned}
R_{q_1} &= \{\, [d_3, 3], [d_5, 3], [d_9, 3], [d_{25}, 2], [d_{39}, 2], \\
&\qquad [d_{44}, 2], [d_{56}, 1], [d_{71}, 1], [d_{89}, 1], [d_{123}, 1] \,\} \\
R_{q_2} &= \{\, [d_3, 3], [d_{56}, 2], [d_{129}, 1] \,\}
\end{aligned}
$$

- That is, while document d_3 is highly relevant to query q_1, document d_{56} is just mildly relevant

Discounted Cumulated Gain (cont.)

- Given these assessments, the results of a new ranking algorithm can be evaluated as follows

- Specialists associate a graded relevance score to the top 10-20 results produced for a given query q
  - This list of relevance scores is referred to as the **gain vector** G

- Considering the top 15 docs in the ranking produced for queries q_1 and q_2, the gain vectors for these queries are:

$$G_1 = (1, 0, 1, 0, 0, 3, 0, 0, 0, 2, 0, 0, 0, 0, 3)$$
$$G_2 = (0, 0, 2, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 3)$$

Discounted Cumulated Gain (cont.)

- By summing up the graded scores up to any point in the
- ranking, we obtain the **cumulated gain** (*CG*)
- For query q_1, for instance, the cumulated gain at the first position is 1, at the second position is 1+0, and so on
- Thus, the *cumulated gain vectors* for queries q_1 and q_2 are given by

$$CG_1 = (1, 1, 2, 2, 2, 5, 5, 5, 5, 7, 7, 7, 7, 7, 10)$$
$$CG_2 = (0, 0, 2, 2, 2, 2, 2, 3, 3, 3, 3, 3, 3, 3, 6)$$

  – For instance, the cumulated gain at position 8 of *CG*1 is equal to 5

Discounted Cumulated Gain (cont.)

- In formal terms, we define
  - Given the gain vector G_j for a test query q_j , the CG_j associated with it is defined as

$$CG_j[i] = \begin{cases} G_j[1] & \text{if } i = 1; \\ \\ G_j[i] + CG_j[i-1] & \text{otherwise} \end{cases}$$

Where $CG_j[i]$ refers to the cumulated gain at the i-th position of the ranking for query q_j

Discounted Cumulated Gain (cont.)

- We also introduce a discount factor that reduces the impact of the gain as we move upper in the ranking

- A simple discount factor is the logarithm of the ranking position

- If we consider logs in base 2, this discount factor will be $\log_2 2$ at position 2, $\log_2 3$ at position 3, and so on

- By dividing a gain by the corresponding discount factor, we obtain the **discounted cumulated gain** (**DCG**)

Discounted Cumulated Gain (cont.)

- More formally,
  - Given the gain vector G_j for a test query q_j, the vector DCG_j associated with it is defined as

$$DCG_j[i] = \begin{cases} G_j[1] & \text{if } i = 1; \\ \frac{G_j[i]}{\log_2 i} + |DCG_j[i-1] & \text{otherwise} \end{cases}$$

Where $DCG_j[i]$ refers to the **discounted cumulated gain** at the i-th position of the ranking for query q_j

Discounted Cumulated Gain (cont.)

- For the example queries q_1 and q_2, the DCG vectors are given by

$$DCG_1 = (1.0, 1.0, 1.6, 1.6, 1.6, 2.8, 2.8, 2.8, 2.8, 3.4, 3.4, 3.4, 3.4, 3.4, 4.2)$$

$$DCG_2 = (0.0, 0.0, 1.3, 1.3, 1.3, 1.3, 1.3, 1.6, 1.6, 1.6, 1.6, 1.6, 1.6, 1.6, 2.4)$$

- Discounted cumulated gains are much less affected by relevant documents at the end of the ranking
- By adopting logs in higher bases the discount factor can be accentuated

DCG Curves

- To produce CG and DCG curves over a set of test queries, we need to average them over all queries

- Given a set of N_q queries, average $\overline{CG}[i]$ and $\overline{DCG}[i]$ over all queries are computed as follows
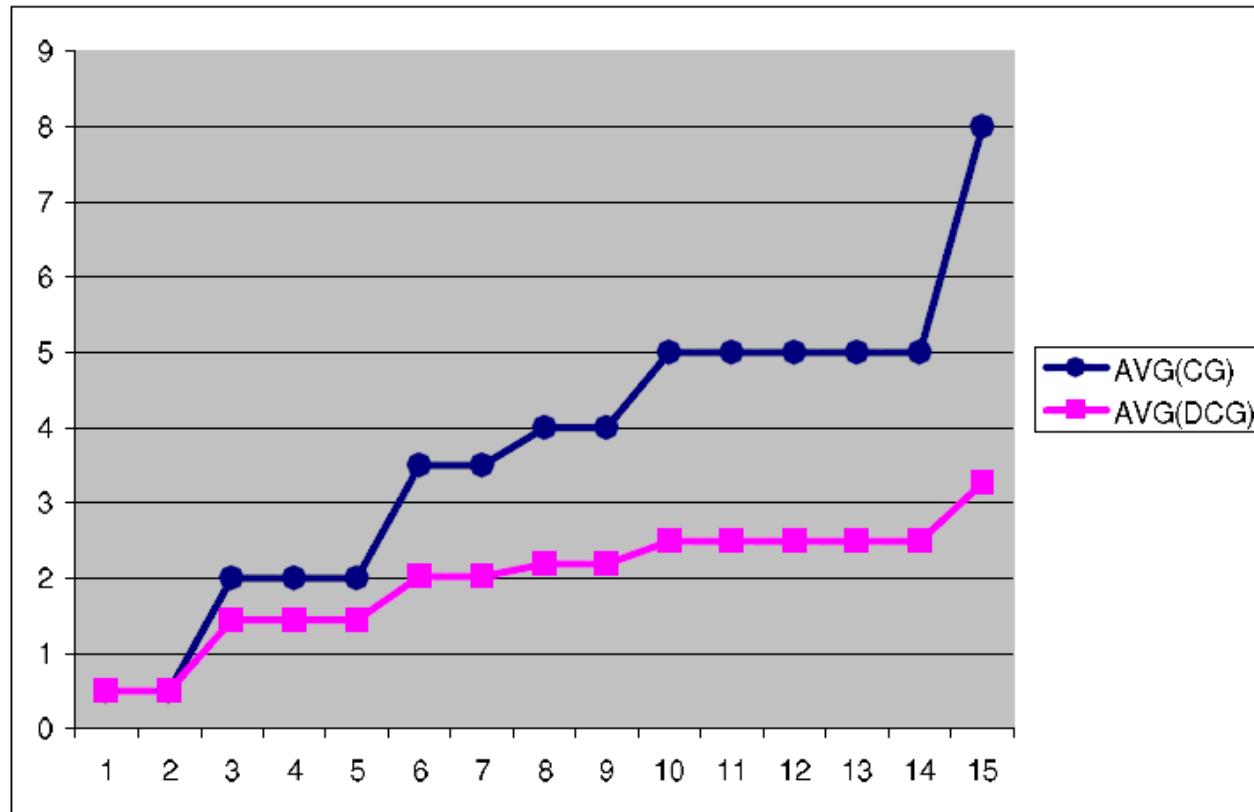
$$\overline{CG}[i] = \sum_{j=1}^{N_q} \frac{CG_j[i]}{N_q}; \qquad \overline{DCG}[i] = \sum_{j=1}^{N_q} \frac{DCG_j[i]}{N_q}$$

- For instance, for the example queries q_1 and q_2, these averages are given by

$$\overline{CG} = (0.5, 0.5, 2.0, 2.0, 2.0, 3.5, 3.5, 4.0, 4.0, 5.0, 5.0, 5.0, 5.0, 5.0, 8.0)$$
$$\overline{DCG} = (0.5, 0.5, 1.5, 1.5, 1.5, 2.1, 2.1, 2.2, 2.2, 2.5, 2.5, 2.5, 2.5, 2.5, 3.3)$$

DCG Curves (cont.)

- Then, average curves can be drawn by varying the rank positions from 1 to a pre-established threshold

Ideal CG and DCG Metrics

- Recall and precision figures are computed relatively to the set of relevant documents

- CG and DCG scores, as defined above, are not computed relatively to any baseline

- This implies that it might be confusing to use them directly to compare two distinct retrieval algorithms

- One solution to this problem is to define a baseline to be used for normalization

- This baseline are the ideal CG and DCG metrics, as we now discuss

Ideal CG and DCG Metrics (cont.)

- For a given test query q, assume that the relevance assessments made by the specialists produced:
  - n_3 documents evaluated with a relevance score of 3
  - n_2 documents evaluated with a relevance score of 2
  - n_1 documents evaluated with a score of 1
  - n_0 documents evaluated with a score of 0

- The ideal gain vector IG is created by sorting all relevance scores in decreasing order, as follows:

$$IG = (3, \ldots, 3, 2, \ldots, 2, 1, \ldots, 1, 0, \ldots, 0)$$

- For instance, for the example queries q_1 and q_2, we have

$$IG_1 = (3, 3, 3, 2, 2, 2, 1, 1, 1, 1, 0, 0, 0, 0, 0)$$

$$IG_2 = (3, 2, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)$$

Ideal CG and DCG Metrics (cont.)

- Ideal CG and ideal DCG vectors can be computed analogously to the computations of CG and DCG

- For the example queries q_1 and q_2, we have

$$ICG_1 = (3, 6, 9, 11, 13, 15, 16, 17, 18, 19, 19, 19, 19, 19, 19)$$

$$ICG_2 = (3, 5, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6, 6)$$

- The ideal DCG vectors are given by

$$IDCG_1 = (3.0, 6.0, 7.9, 8.9, 9.8, 10.5, 10.9, 11.2, 11.5, 11.8, 11.8, 11.8, 11.8, 11.8, 11.8)$$

$$IDCG_2 = (3.0, 5.0, 5.6, 5.6, 5.6, 5.6, 5.6, 5.6, 5.6, 5.6, 5.6, 5.6, 5.6, 5.6, 5.6)$$

Ideal CG and DCG Metrics (cont.)

- Further, average $\overline{ICG}$ and average $\overline{IDCG}$ scores can be computed as follows

$$\overline{ICG}[i] = \sum_{j=1}^{N_q} \frac{ICG_j[i]}{N_q}; \qquad \overline{IDCG}[i] = \sum_{j=1}^{N_q} \frac{IDCG_j[i]}{N_q}$$

- For instance, for the example queries q_1 and q_2, $\overline{ICG}$ and $\overline{IDCG}$ vectors are given by

$$\overline{ICG} = (3.0, 5.5, 7.5, 8.5, 9.5, 10.5, 11.0, 11.5, 12.0, 12.5, 12.5, 12.5, 12.5, 12.5, 12.5)$$

$$\overline{IDCG} = (3.0, 5.5, 6.8, 7.3, 7.7, 8.1, 8.3, 8.4, 8.6, 8.7, 8.7, 8.7, 8.7, 8.7, 8.7)$$

- By comparing the average CG and DCG curves for an algorithm with the average ideal curves, we gain insight on how much room for improvement there is

Normalized DCG   (NDCG)

- Precision and recall figures can be directly compared to the ideal curve of 100% precision at all recall levels

- DCG figures, however, are not build relative to any ideal curve, which makes it difficult to compare directly DCG curves for two distinct ranking algorithms

- This can be corrected by normalizing the DCG metric

- Given a set of N_q test queries, normalized CG and DCG metrics are given by

$$NCG[i] = \frac{\overline{CG}[i]}{ICG[i]}; \qquad NDCG[i] = \frac{\overline{DCG}[i]}{IDCG[i]}$$

Normalized DCG (cont.)

- For instance, for the example queries q_1 and q_2, NCG and NDCG vectors are given by

$$
\begin{aligned}
NCG \ &= \ (0.17, 0.09, 0.27, 0.24, 0.21, 0.33, 0.32, \\
&\qquad 0.35, 0.33, 0.40, 0.40, 0.40, 0.40, 0.40, 0.64) \\
NDCG \ &= \ (0.17, 0.09, 0.21, 0.20, 0.19, 0.25, 0.25, \\
&\qquad 0.26, 0.26, 0.29, 0.29, 0.29, 0.29, 0.29, 0.38)
\end{aligned}
$$

- The area under the NCG and NDCG curves represent the quality of the ranking algorithm

- Higher the area, better the results are considered to be

- Thus, normalized figures can be used to compare two distinct ranking algorithms

Discussion on DCG Metrics

- CG and DCG metrics aim at taking into account multiple level relevance assessments

- This has the advantage of distinguishing highly relevant documents from mildly relevant ones

- The inherent disadvantages are that multiple level relevance assessments are harder and more time consuming to generate

Discussion on DCG Metrics (cont.)

- Despite these inherent difficulties, the CG and DCG metrics present benefits:

    – They allow systematically combining document ranks and relevance scores

    – **Cumulated gain (CG) provides a single metric of retrieval performance at any position in the ranking**

    – It also stresses the gain produced by relevant docs up to a position in the ranking, which makes the metrics more immune to outliers

    – Further, **discounted cumulated gain (DCG)** allows down weighting the impact of relevant documents found late in the ranking

Trends and Research Issues

- A major trend today is research on interactive user interfaces and their evaluation

  - Which evaluation measures are most appropriate?

- Another important trend is **crowdsourcing**

  - Use the population of Web users to conduct well defined evaluation tasks in exchange for small sums of money

- Further, the proposal, the study, and the characterization of alternative measures to precision and recall continue to be of interest

  - Some specific scenarios are not well covered by precision-recall figures