Modern Information Retrieval

Retrieval Evaluation

Reference Collections

CFC: The Cystic Fibrosis Collection

Performance Evaluation

- Most common measures of system performance are *time* and *space*
- *Time*: how fast does the system run?
- *Space*: what fraction of the available resources does the system consume?
- Time x Space: good metrics for data retrieval systems and for IR systems
- But, since answers in an IR system are only approximate, we must also evaluate the *quality* of those answers!

Retrieval Performance Evaluation

- To evaluate the quality of the approximate answers, we compare them with a set of *ideal* answers (provided by specialists).
- Clearly, we can only do this for a set of pre-defined example information requests, also referred to as *reference topics.*
- For each reference topic, the *ideal answer set* is provided.
- The documents used for generating the various ideal answer sets form a *reference collection*.

Retrieval Performance Evaluation

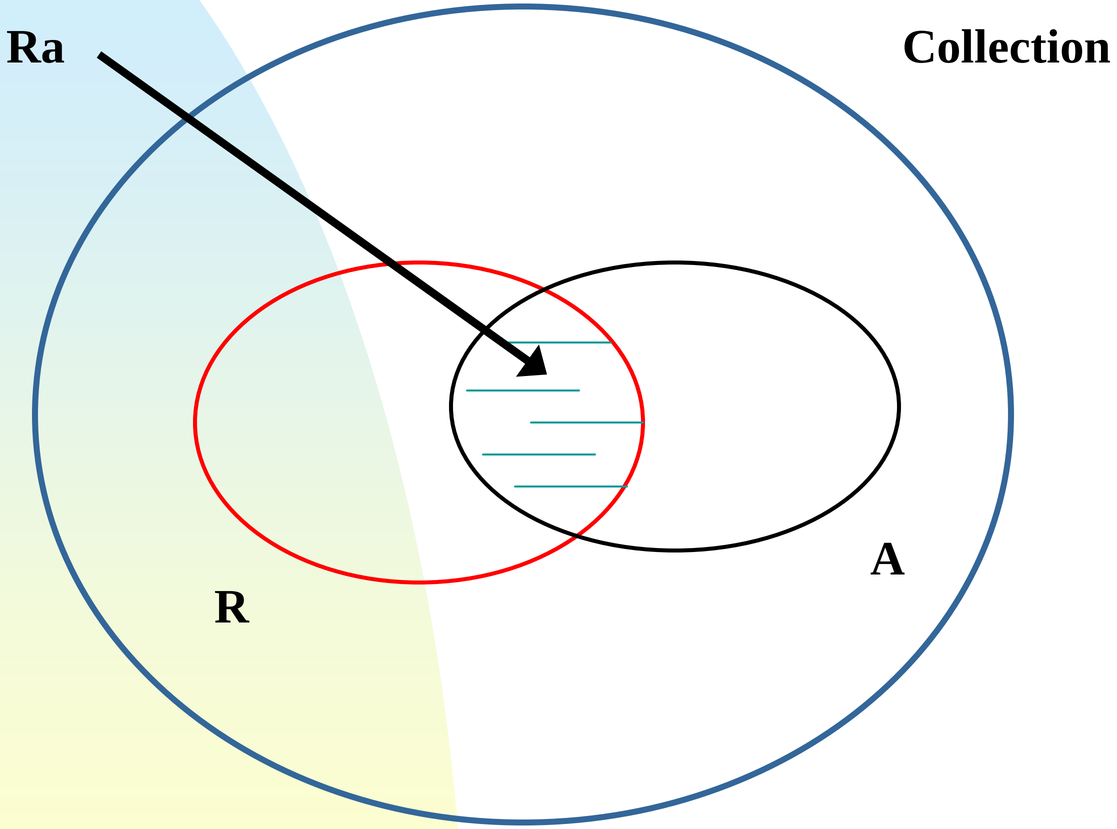
- The evaluation of the quality of a ranking algorithm involves then:
    - ◆ a reference collection
    - ◆ a set of reference topics
    - ◆ an ideal answer set for each reference topic
- The answers generated by a ranking algorithm (such as the vector model) are compared with the ideal answer sets to determine *how good* is the ranking.
- This process of evaluating the quality of a ranking is usually referred to as *retrieval performance evaluation*.

Precision and Recall

- Retrieval performance evaluation is often measured in terms of two metrics: *precision* and *recall*.
- Let,
  - ◆ I : an example information request (topic)
  - ◆ R : the ideal answer set for the topic I
  - ◆ |R| : number of docs in the set R
  - ◆ A : the answer set generated by a ranking strategy we wish to evaluate
  - ◆ |A| : the number of docs in the set A

Precision and Recall

- Relationship between the sets R and A, given I.



|Ra|
Recall = -----
|R|

|Ra|
Precision = -----
|A|

Precision and Recall

- The viewpoint using the sets R, A, and Ra, does not consider that documents presented to the user are ordered (i.e., ranked).
- User sees a ranked set of documents and examines them starting from the top.
- Thus, precision and recall vary as the user proceeds with his examination of the set A.
- Most appropriate then is to plot a curve of precision versus recall.

Precision and Recall

- Let Rq be the set of relevant docs for a query q
  - Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

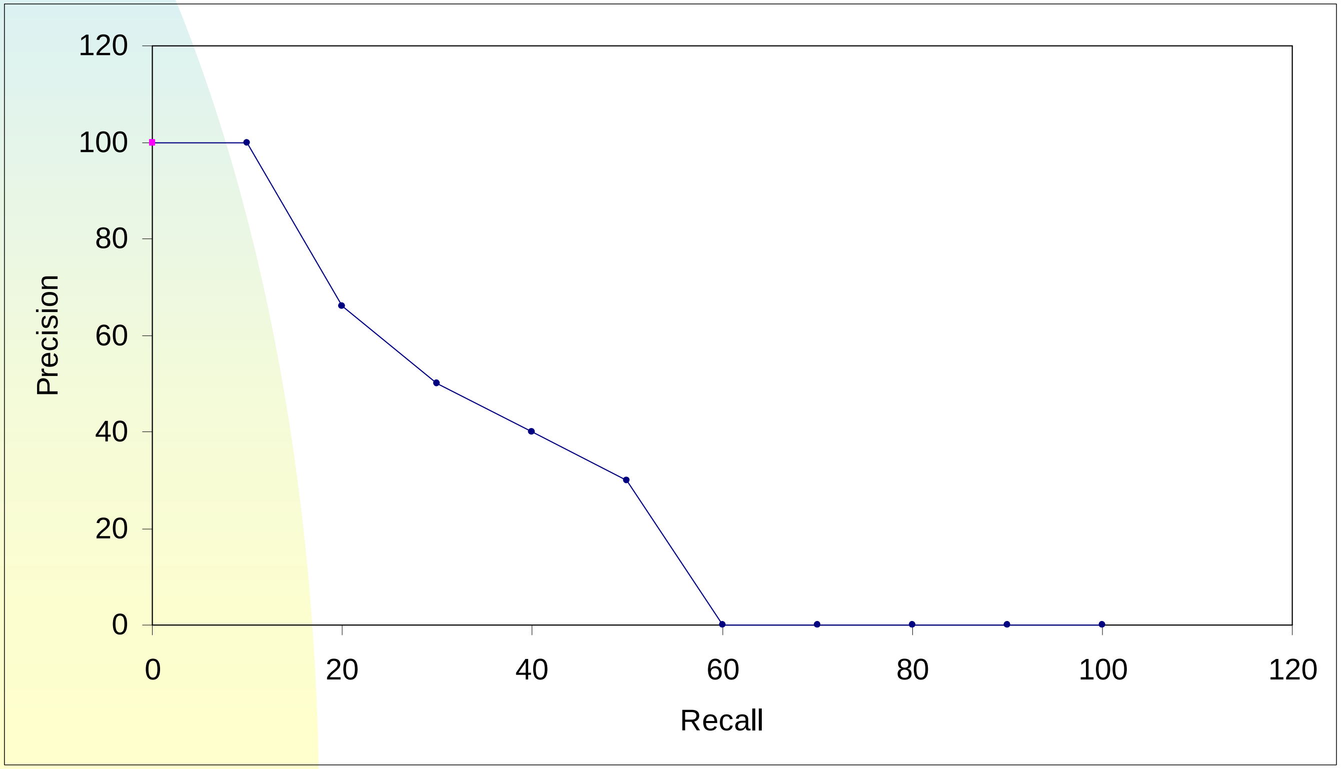
- Consider a new retrieval algorithm that yields the following set of docs as answers to the query q:

1. **d123**
2. d84
3. **d56**
4. d6
5. d8
6. **d9**
7. d511
8. d129
9. d187
10. **d25**
11. d38
12. d48
13. d250
14. d113
15. **d3**

Precision and Recall

- Consider a new retrieval algorithm that yields the following set of docs as answers to the query q:

1. d123
2. d84
3. d56
4. d6
5. d8
6. d9
7. d511
8. d129
9. d187
10. d25
11. d38
12. d48
13. d250
14. d113
15. d3



Precision and Recall

- Precision: a single query. What if multiple queries?
- Let Nq be the number of queries considered. Then,

$$\boxed{\mathrm{Avg}(P(r)) = \mathrm{sum} \; \frac{P_i(r)}{N_q}}$$

where, Pi(r) : precision at recall level r for the ith query.

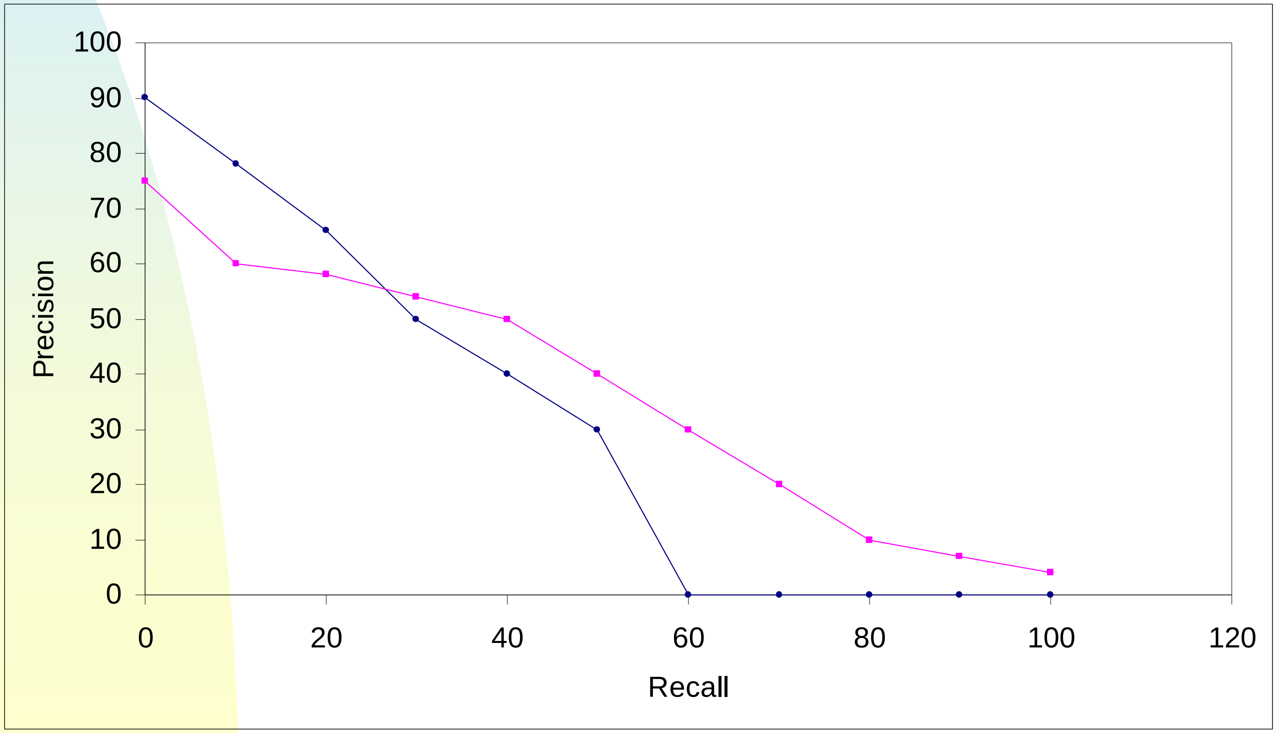
- In case the set Rq of relevant docs includes less than 10 docs, use interpolation:

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

where P(rj) is precision at recall level rj.

Precision and Recall

- Two distinct algorithms can be compared, over a set of Nq queries, by examing their curves of average precision and recall.



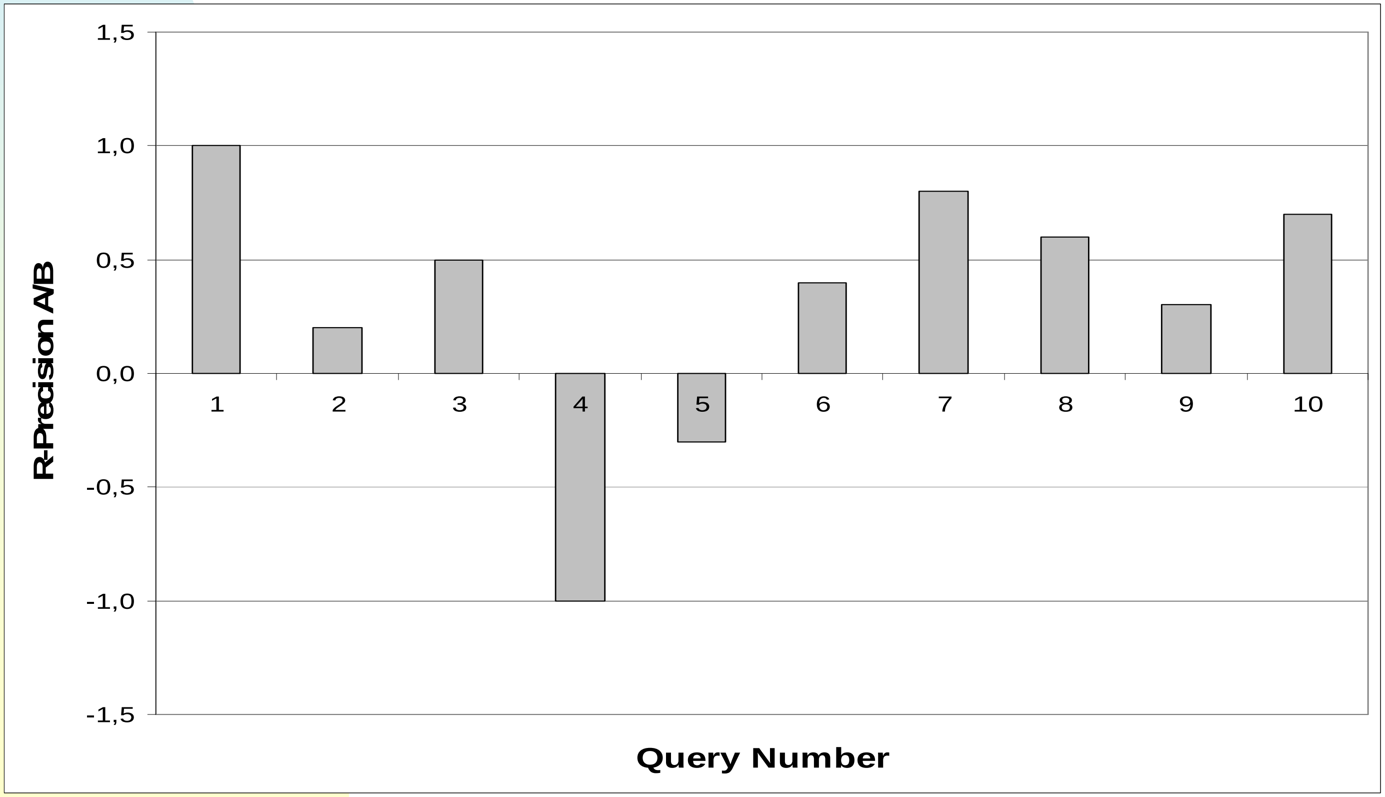
Single Value Summaries

- Precision and recall: average over Nq queries.
- How to evaluate retrieval performance over individual queries?
- Use a single number to summarize retrieval performance for each query.
- Let,
    - R be the total number of relevant docs for a query q.
- Define,
    - R-Precision: precision at the point at which exactly R docs have been examined.

Single Value Summaries

- Consider two retrieval algorithms A and B.
- Let,
  - RPa(i) : R-precision for algorithm A for the ith query
  - RPb(i) : R-precision for algorithm B for the ith query

$$\boxed{RPa/b(i) = RPa(i) - RPb(i)}$$



Trec Collection

- Standard reference collection most referred to nowadays.
- Annual Trec Conference at NIST, Maryland.
- Companies and research groups can then compare their retrieval systems.
- Reference collections are prepared for these comparative experiments:
  - Trec-3 : reference collection with 2 GBytes
  - Trec-6 : reference collection with 5.8 GBytes

Trec Collection

- Trec-6 is composed of docs from:
  - WSJ: Wall Street Journal
  - AP: Associated Press
  - ZIFF: Computer Selects, Ziff-Davis
  - FR: Federal Register
  - DOE: US DOE Publications
  - SJMN: San Jose Mercury News
  - PAT: US Patents
  - FT: Financial Times
  - CR: Congressional Record
  - FBIS: Foreign Broadcast Information Service
  - LAT: LA Times

Trec Collection

- Docs at TREC are represented in SGML:

```
<doc>
<docno> WSJ880406-0090 </docno>
<hl> AT&T Unveils New Services </hl>
<author> Janet Guyon </author>
<text>
American Telephone & Telegraphy Co.
introduced the first of a new generation of
phone services with broad ...
</text>
</doc>
```

Trec Collection

- Topics at TREC are detailed descriptions of information needs:

  <top>

  <num> Number: 168

  <title> Topic: Financina AMTRAK

  <desc> Description:

  A document will address the role of the Federal Government in financing the operation of the National Railroad Transportation Corporation (AMTRAK).

  <narr> Narrative: A relevant document must provide information on the government´s responsability to make AMTRAK an economically viable entity.

  </top>

Benchmark Tasks at Trec-6

- General:
  - Ad hoc
  - Routing
- Specific:
  - Chinese
  - Filtering (new incoming doc relevant?)
  - Interactive (user interacts with system)
  - NLP
  - Cross Languages
  - High precision (retrieve 10 docs in 5 minutes)
  - Spoken document retrieval (broadcast news)
  - Very Large Corpus (7.5 million documents; 20 GBytes)

CFC Collection

- 1,239 documents indexed with the term *cystic fibrosis* in the National Library of Medicine's MEDLINE
- Each doc record is composed of:

MEDLINE accession number
title
major subjects
abstract
citations

author
source
minor subjects
references

CFC Collection

- 100 information requests with extensive relevance judgements:
  - 4 separate relevance scores for each request
  - Scores proviced by human experts and by a medical bibliographer
  - Each score:
    - 0 (not relevant)
    - 1 (marginally relevant)
    - 2 (strongly relevant)

CFC Collection

- Small and nice collection for experimentation
- Number of information requests is large relative to the collection size
- Good relevance judgements
- For online access:
  - http://www.dcc.ufmg.br/irbook
  - http://www.sunsite.dcc.uchile.cl/irbook
  - http://www.sims.berkeley.edu/~hearst/irbook