情報検索システム特論

Advanced Information Retrieval Systems
第4回 Lecture #4

2023-05-01

湯川 高志 Takashi Yukawa

Retrieval Evaluation

Evaluation criteria is very important for comparing information retrieval systems

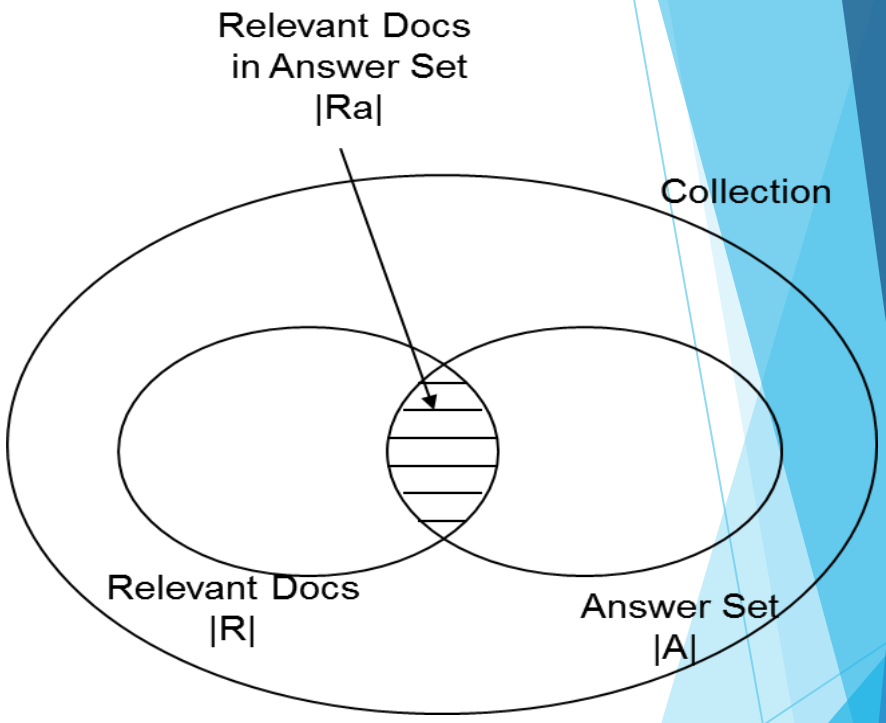
Precision and Recall (1)

- Retrieval performance evaluation is often measured in terms of two metrics
  - Precision
  - 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 (2)

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

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



Precision and Recall (3)

- 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

- 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
  - **0.1 (10%) step of recall vs precision at each recall point**

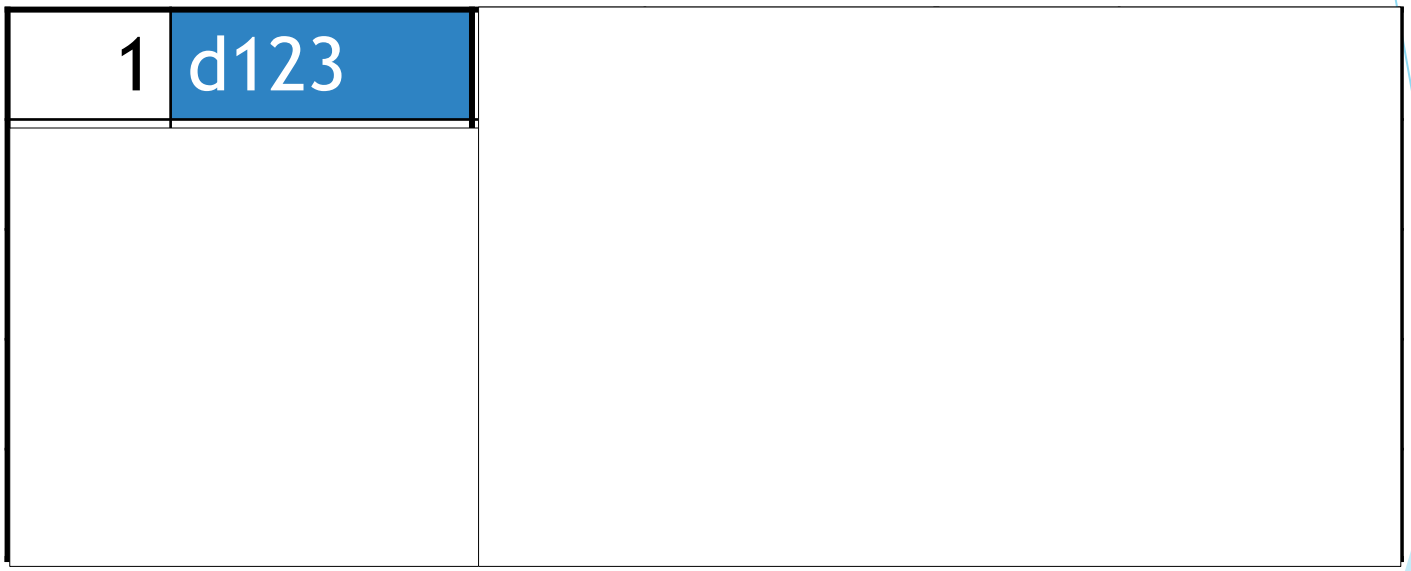
Precision and Recall: An Example

- Rq: the set of relevant docs for a query q
  - Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}
- A retrieval algorithm that yields the following set of docs as answers to the query q

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

Precision and Recall: An Example (cont'd)

- Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}



$$P = \frac{1}{1} = 1\ (100\%) \qquad R = \frac{1}{10} = 0.1\ (10\%)$$

Precision and Recall: An Example (cont'd)

- Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

1	d123
2	d84
3	d56

$$P = \frac{2}{3} = 0.67\ (67\%) \qquad R = \frac{2}{10} = 0.2\ (20\%)$$

Precision and Recall: An Example (cont'd)

- Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

1	d123
2	d84
3	d56
4	d6
5	d8

$$P = \frac{2}{3} = 0.67\ (67\%) \qquad R = \frac{2}{10} = 0.2\ (20\%)$$

Precision and Recall: An Example (cont'd)

- Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

1	d123	6	d9
2	d84		
3	d56		
4	d6		
5	d8		

$$P = \frac{3}{6} = 0.5\ (50\%) \qquad R = \frac{3}{10} = 0.3\ (30\%)$$

Precision and Recall: An Example (cont'd)

- Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

1	d123	6	d9
2	d84	7	d511
3	d56	8	d129
4	d6	9	d187
5	d8	10	d25

$$P = \frac{4}{10} = 0.4\ (40\%) \qquad R = \frac{4}{10} = 0.4\ (40\%)$$

Precision and Recall: An Example (cont'd)

- Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

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

$$P = \frac{5}{15} = 0.33\ (33\%) \qquad R = \frac{5}{10} = 0.5\ (50\%)$$

Precision and Recall: An Example (cont'd)

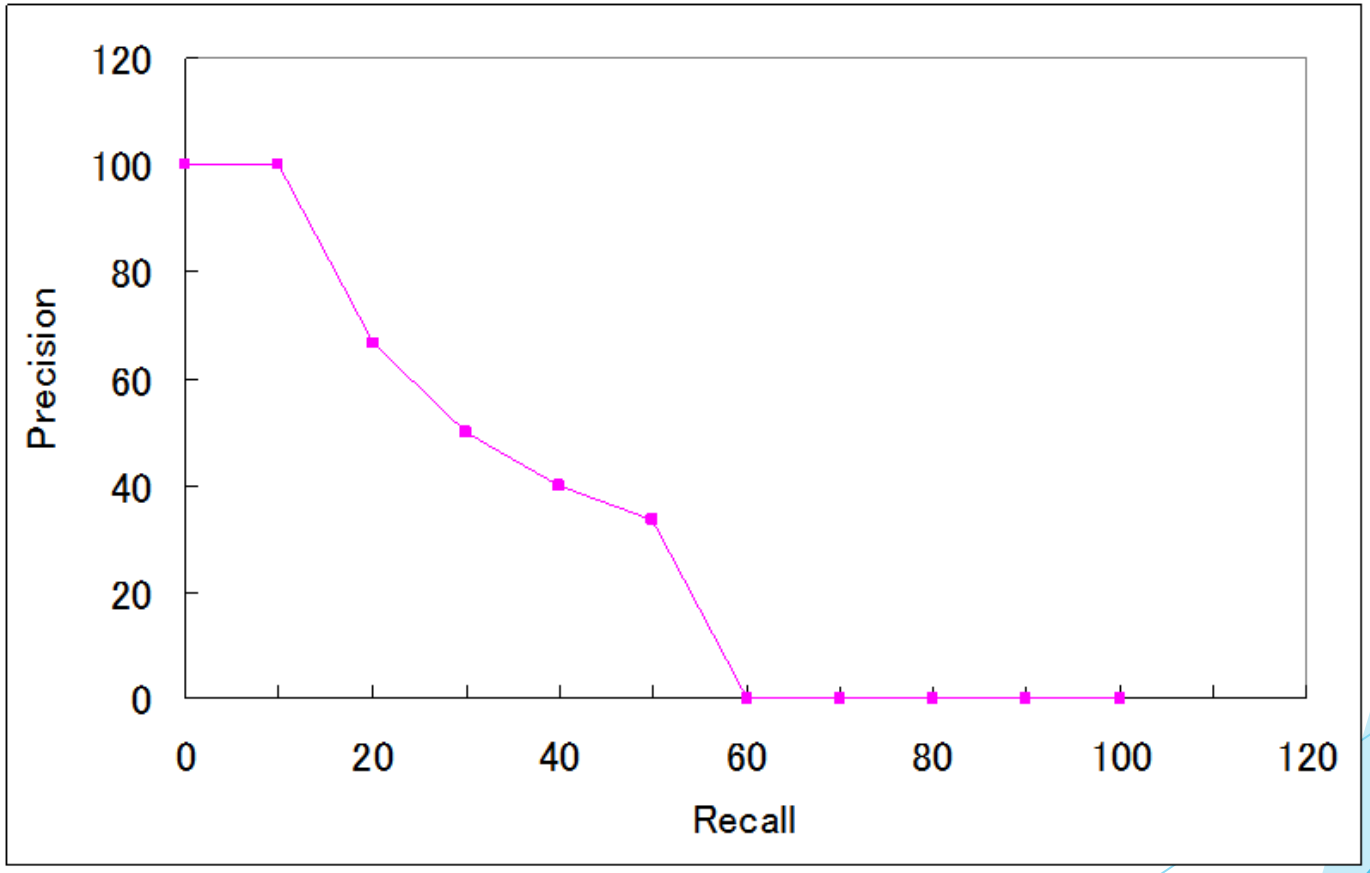
- Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

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

$P = 0$

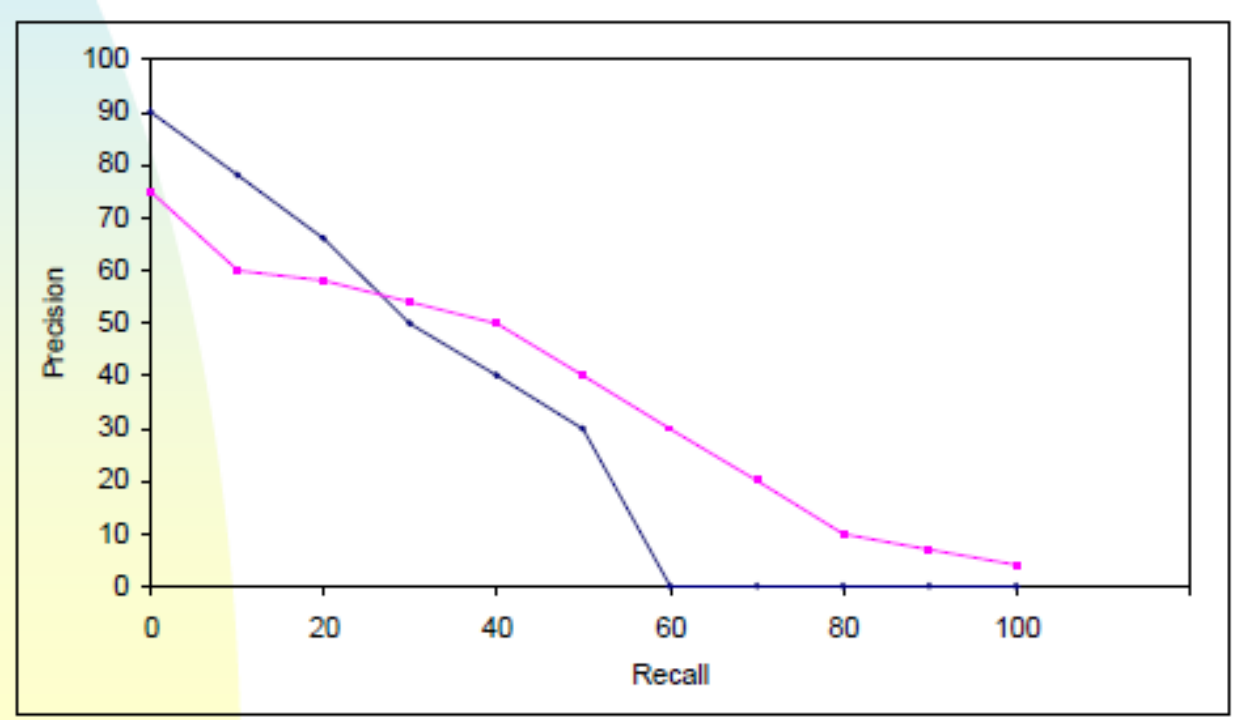
$R \geq 0.6\ (60\%)$

Precision and Recall: An Example (cont'd)



Recall - Precision Curves

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



Interpolation

- In case the set Rq of relevant docs includes less than 10 docs, use interpolation
  - $P(r_j)$: precision at recall level r_j

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

Precision and Recall: Another Example

- Rq = {d3, d56, d129}

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

P=0.33, R=0.33
P=0.25, R=0.67
P=0.20, R=1.00

Interpolation: Example

- $P(r_j)$: precision at recall level r_j

$$P(r_j) = \max_{r_j \le r} P(r)$$

$r_j = 0.1$ maximum precision at $r \geq 0.1$

$P(0.1) = 0.33$

$P(0.2) = 0.33$

$P(0.3) = 0.33$

P=0.33, R=0.33
P=0.25, R=0.67
P=0.20, R=1.00

[review] Interpolation: Example (cont'd)

- $P(r_j)$: precision at recall level r_j

$$P(r_j) = \max_{r_j \le r} P(r)$$

$r_j = 0.4$ maximum precision at $r \geq 0.4$

$P(0.4) = 0.25$

$P(0.5) = 0.25$

$P(0.6) = 0.25$

P=0.25, R=0.67
P=0.20, R=1.00

Interpolation: Example (cont'd)

- $P(r_j)$: precision at recall level r_j

$$P(r_j) = \max_{r_j \le r} P(r)$$

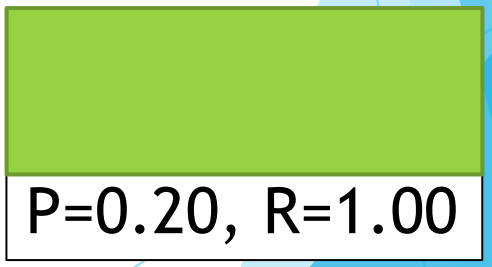
$r_j = 0.7$ maximum precision at $r \geq 0.7$

$P(0.7) = 0.20$

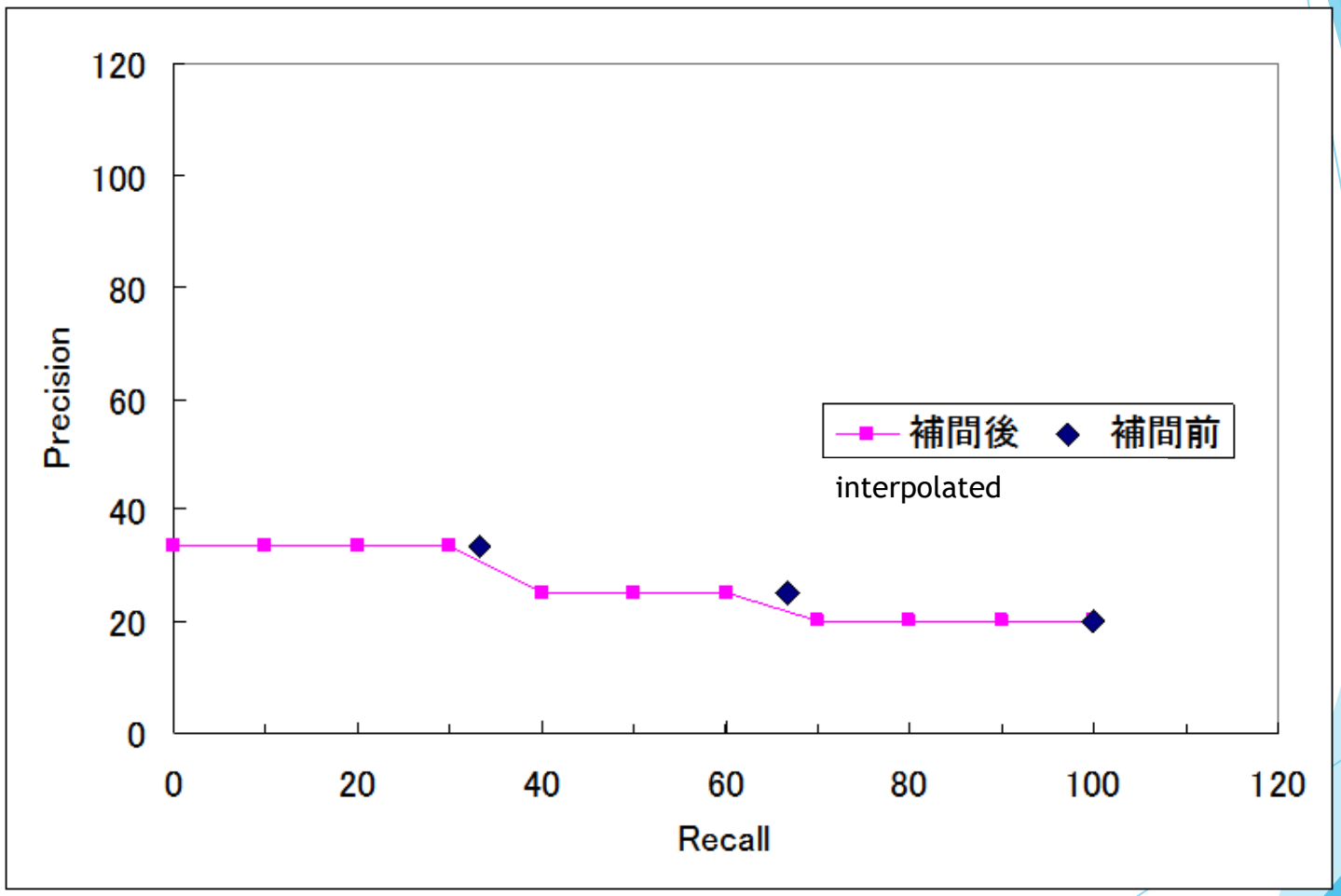
$P(0.8) = 0.20$

$P(0.9) = 0.20$

$P(1.0) = 0.20$



Interpolation Example (cont'd)



Single Value Summaries

- How to evaluate retrieval performance over individual queries?

- Use a single number to summarize retrieval performance for each query

F-measure

- $F = \frac{2Precision \cdot Recall}{Precision + Recall}$
- Harmonic mean of recall and precision

Average Precision

- $AP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{n}$

- n: the number of relevant documents
- $P(k)$: the precision at cut-off k in the list
- $rel(k)$:
  - if the item at rank k is a relevant document, $rel(k) = 1$
  - otherwise $rel(k) = 0$

Average Precision Example

- Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

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

$$AP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{n}$$

$$n = 10$$

Average Precision Example (cont'd)

- Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

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

$$AP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{n}$$

$$n = 10$$

$$k = 1,3,6,10,15 \rightarrow rel(k) = 1, \text{otherwise } rel(k) = 0$$

Average Precision Example (cont'd)

- Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}

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

$$AP = \frac{\sum_{k=1}^{n}(P(k) \times rel(k))}{n} \qquad n = 10$$

$$k = 1,3,6,10,15 \rightarrow rel(k) = 1, \text{otherwise } rel(k) = 0$$

$$AP = \frac{P(1) + P(3) + P(6) + P(10)}{10}$$

$$= \frac{1 + \frac{2}{3} + \frac{3}{6} + \frac{4}{10}}{10} = \frac{1 + 0.67 + 0.5 + 0.4}{10} = 0.26$$

11 point AP

- 11 point average precision
- P(r): precision at recall level r

$$M(P) = \frac{\sum_{i=0}^{10} P(0.1 \times i)}{11}$$

R-Precision

- R: the total number of relevant docs for a query q
- R-Precision: precision at the point at which exactly R docs have been examined

R-Precision: An Example

- Retrieval results (an answer set)

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

- Rq = {d3, d5, d9, d25, d39, d44, d56, d71, d89, d123}
- R=10
- R-Precision = 4/10 = 0.4

R-Precision: Another Example

- Retrieval results (an answer set)

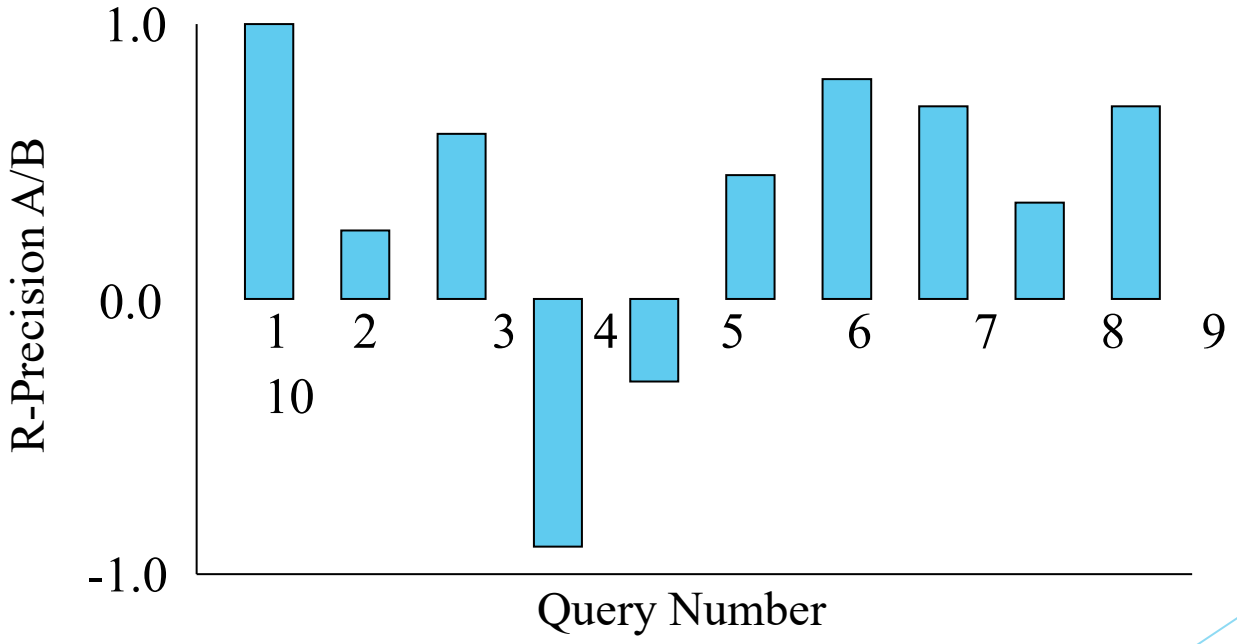
1	d123	6	d9	11	d38
2	d84	7	d511	12	d48
3	d56	8	d129	13	d250
4	d6	9	d187	14	d113
5	d8	10	d25	15	d3

- Rq = {d3, d56, d129}
- R=3
- R-Precision = 1/3 = 0.33

Precision Histogram

- Two retrieval algorithms A and B.
- $RP_A(i)$: R-precision for algorithm A for the ith query
- $RP_B(i)$: R-precision for algorithm B for the ith query

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



Reference Collections

TREC Collections

- 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-6 Collection

- Trec-6 document collection
  - 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-6 Collection (cont'd)

- Documents 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>

Topics at TREC Collections

- 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>

Tasks at TREC-6

- General
  - Ad hoc
  - Routing
- Specific
  - Chinese
  - 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)

Recent TREC

Call to TREC 2020

TREC home

TREC Statement on Product Testing and Advertising

CALL FOR PARTICIPATION

TEXT RETRIEVAL CONFERENCE (TREC) 2020

February 2020 - November 2020

Conducted by:
National Institute of Standards and Technology (NIST)

The Text Retrieval Conference (TREC) workshop series encourages research in information retrieval and related applications by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results. Details about TREC can be found at the TREC web site, http://trec.nist.gov.

You are invited to participate in TREC 2020. TREC 2020 will consist of a set of tasks known as "tracks". Each track focuses on a particular subproblem or variant of the retrieval task as described below. Organizations may choose to participate in any or all of the tracks. Training and test materials are available from NIST for some tracks; other tracks will use special collections that are available from other organizations for a fee.

https://trec.nist.gov/pubs/call2020.html

TREC2020 Tracks (1)

- Deep Learning Track
  - The Deep Learning track focuses on IR tasks where a large training set is available, allowing us to compare a variety of retrieval approaches including deep neural networks and strong non-neural approaches, to see what works best in a large-data regime.
- Fair Ranking Track
  - The Fair Ranking track focuses on building two-sided systems that offer fair exposure to ranked content producers while ensuring high results quality for ranking consumers.
- Health Misinformation Track
  - The Health Misinformation track aims to (1) provide a venue for research on retrieval methods that promote better decision making with search engines, and (2) develop new online and offline evaluation methods to predict the decision making quality induced by search results. Consumer health information is used as the domain of interest in the track. (This track was called the Decision Track in TREC 2019.)

TREC2020 Tracks (2)

- Incident Streams Track
  - The Incident Streams track is designed to bring together academia and industry to research technologies to automatically process social media streams during emergency situations with the aim of categorizing information and aid requests made on social media for emergency service operators.
- News Track
  - The News track features modern search tasks in the news domain. In partnership with The Washington Post, the track develops test collections that support the search needs of news readers and news writers in the current news environment.
- Podcasts Track
  - A new track for 2020. The aim of the Podcasts track is to develop methods for information retrieval and content understanding from open-domain podcast transcripts and audio.
- Precision Medicine Track
  - The Precision Medicine track focuses on building systems that use data (e.g., a patient's past medical history and genomic information) to link oncology patients to clinical trials for new treatments as well as evidence-based literature to identify the most effective existing treatments.

NTCIR

- Japanese project that is similar to TREC
  - Organized by NII (National Institute of Informatics)

http://research.nii.ac.jp/ntcir/ntcir-15/index-ja.html

NTCIR-15 Tasks (1)

- CORE TASKS
  - DialEval-1
    - The task is a continuation of the Dialogue Quality (DQ) and Nugget Detection (ND) subtasks run at the NTCIR-14 STC-3 task (See References).
  - FinNum-2
    - FinNum is a task for fine-grained numeral understanding in financial social media data - to identify the linking between the target cashtag and the target numeral.
  - QA Lab-PoliInfo-2
    - In NTCIR15 QA Lab-PoliInfo2, we propose three tasks (Stance Classification task, Dialog Summarization task and Entity Linking task) to solve political issues by natural language processing.
  - SHINRA2020-ML
    - SHINRA is a resource creation project started in the year 2017, aiming to structure the knowledge in Wikipedia. SHINRA2020-ML is the first shared-task of text categorization in SHINRA project, tackling the problem of classifying 30 language Wikipedia entities in fine-grained categories.

NTCIR-15 Tasks (2)

- CORE TASKS (cont'd)
  - WWW-3
    - WWW-3 is an adhoc web search task for Chinese and
    - English.
- PILOT TASKS
  - Data Search
    - NTCIR-15 Data Search is a shared task on ad-hoc retrieval for governmental statistical data. The first round of Data Search focuses on the retrieval of a statistical data collection published by the Japanese government (e-Stat), and one published by the US government (Data.gov).
  - MART
    - MART (Micro-activity Retrieval Task) is an NTCIR-15 workshop data challenge task. The NTCIR15-MART pilot task aims to motivate the development of a first generation of techniques for high-precision micro-activity detection and retrieval of micro-activities of daily living, to support identification and retrieval of activities that occur over short time-scales, such as minutes, rather than the long-duration event segmentation tasks of the past work.

CFC Collection

- 1,239 documents indexed with the term *cystic fibrosis* in the National Library of Medicine's MEDLINE
- Each document record is composed of
  - MEDLINE accession number
  - author
  - title
  - source
  - major subjects
  - minor subjects
  - abstract
  - references
  - citations

CFC Collection (cont'd)

- 100 information requests with extensive relevance judgments
  - 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 (cont'd)

- Small and nice collection for experimentation
- Number of information requests is large relative to the collection size
- Good relevance judgments

That's it today

今日はここまで

近いうちに課題を出します / Assignments will be given in few weeks (not today)