NWI-I00041 Information Retrieval Evaluation

Dr. ir. Faegheh Hasibi
f.hasibi@cs.ru.nl

Nijmegen, September 28th, 2020

Discussion

- Suppose that you are the VP of Search at a giant internet retail company. Your boss brings in her nephew, who claims to have built a better search engine for the company.
  - How would you react to this statement?
  - How do you assess the quality of the new system?

groups of 2-3; write down a few keywords to capture your solution

What to measure?

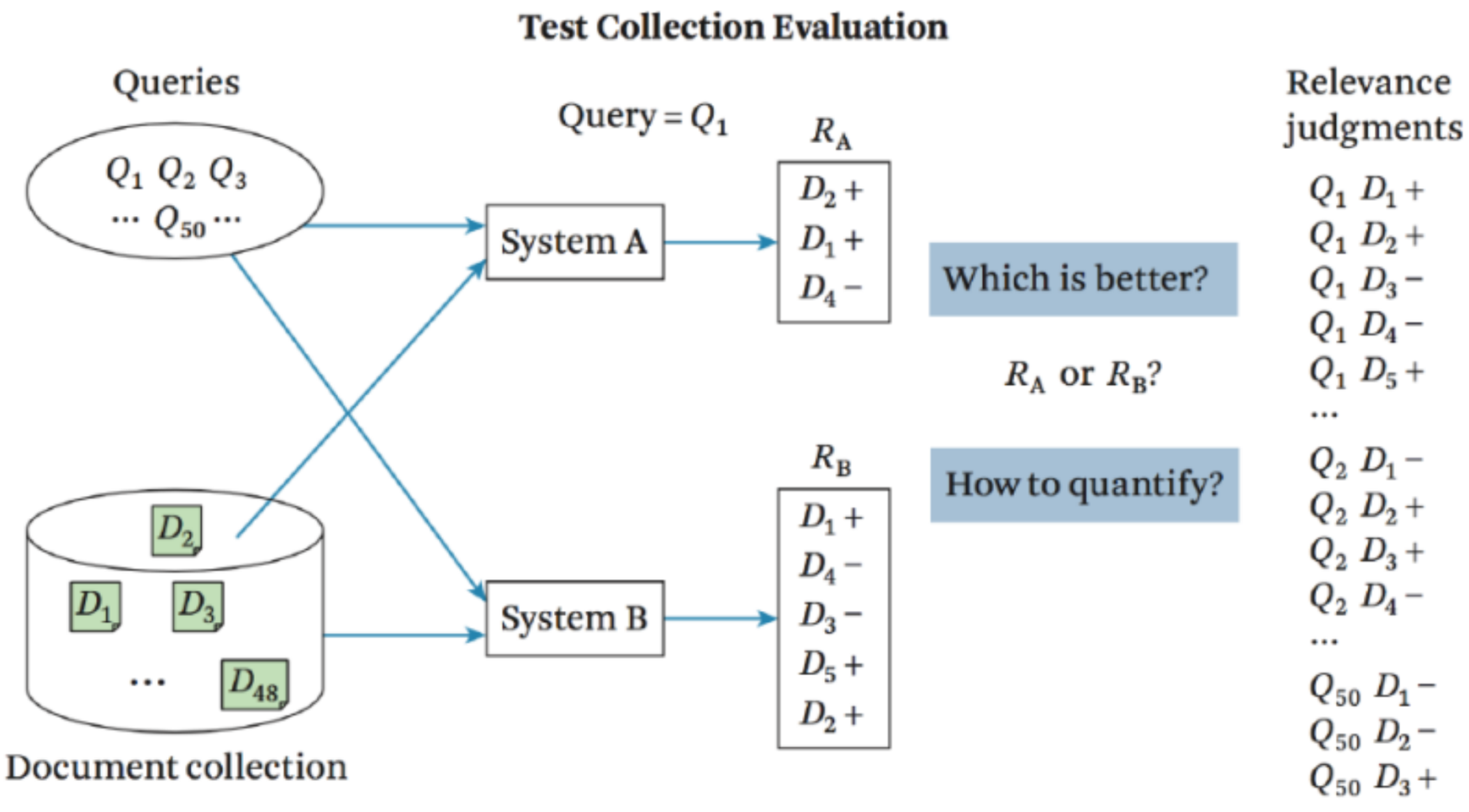
- **Effectiveness**
  - How accurate are the search results?
  - I.e., the system's capability of ranking relevant documents ahead of non-relevant ones

- Efficiency
  - How quickly can a user get the results?
  - I.e., the response time of the system

- Usability
  - How useful is the system for real user tasks?

“Cranfield paradigm”

Central idea:
Build **re-usable test collection** and define **measures**



“Cranfield paradigm”

- An IR test collection ingredients:
  - A collection of documents
  - A set of search requests
  - Relevance assessments for each search request against the collection of documents

- Measures:
  - Quantify the performance of a system
  - Compare the system’s results with relevance judgements

Document: “Retrieval Unit”

Collection	Retrieval unit
The web	Web page
E-mail archive	Message
Twitter	Tweet
City archives	Any document
Web forum	Thread
…	…

Relevance Judgments

Ground truth labels for *query-item pairs*

- **Binary**
  - 0: non-relevant
  - 1: relevant
- **Graded; e.g.:**
  - 0: non-relevant
  - 1: somewhat relevant
  - 2: relevant
  - 3: highly relevant / perfect match

Obtaining Relevance Judgements

- Expensive
- Time-consuming process

- Two approaches:
  - Crowdsourcing
  - Expert judgements

Crowdsourcing

- “Microtasks", performed in parallel by large, paid crowds

- Platforms:
  - Amazon Mechanical Turk
  - Figure Eight (Crowdflower)
  - etc.

- Control strategies need to be implemented
  - E.g., intake quiz, hidden tests
  - Fair payments

Expert Judgments

- Experts are trained about the task
- Each query-item pair requires few assessments
  - often less than crowdsourcing
- Agreement between annotators is generally high

Text Retrieval Conference (TREC)

- Yearly benchmarking cycle
- Multiple tracks
- Organized by the US National Institute of Standards and Technology (NIST)

TREC assessors at work

Example TREC Topic

<num> Number: EX52

<title>ontology engineering</title> Short query

<desc> Description:
Find individuals with expertise regarding ontology engineering.
</desc> Longer description of the query

<narr> Narrative:
This topic attempts to find individuals with expertise regarding to ontology engineering. Ontology engineering concerns the whole life-cycle of ontologies, such as ontology construction, ontology learning, ontology mapping, and ontology evolution. We want people with expertise about ontology engineering rather then other things related to ontology.
</narr>

A complete description of relevance for assessors

Challenges

- ‘Complete’ relevance judgments:
  - Idealized setting: a judgment for each document in the collection for each query.
  - **Infeasible with real collections!**

  - **Alternative:** Pooling Method
    - Create a pool of (say, 200) documents per query, retrieved by multiple (baseline) retrieval systems and have those judged

“Cranfield paradigm”

- An IR test collection ingredients:
  - A collection of documents
  - A set of search requests
  - Relevance assessments for each search request against the collection of documents

- Measures:
  - Quantify the performance of a system
  - Compare the system’s results with relevance judgements

Set-based Retrieval Evaluation

		Action: Retrieved	Action: Not retrieved
Doc	Relevant	*a*	*b*
	Not relevant	*c*	*d*

$$\text{Precision} = \frac{a}{a+c}$$

Ideal results: precision = recall = 1.0

$$\text{Recall} = \frac{a}{a+b}$$

In reality, high recall tends to be associated with low precision

a = true positives (tp)
b = false negatives (fn)
c = false positives (fp)
d = true negatives (tn)

Precision and Recall

$$\text{Precision} = \frac{\#\text{ of retrieved documents that are relevant}}{\#\text{ of retrieved documents}}$$

$$\text{Recall} = \frac{\#\text{ of relevant documents that are retrieved}}{\#\text{ of relevant documents}}$$

Precision and Recall

- Trade-off between precision and recall
  - High recall is easy to achieve when precision is ignored
  - High precision can still mean a bad system, considering that important relevant results may be missed

F-measure

- Harmonic mean of recall and precision

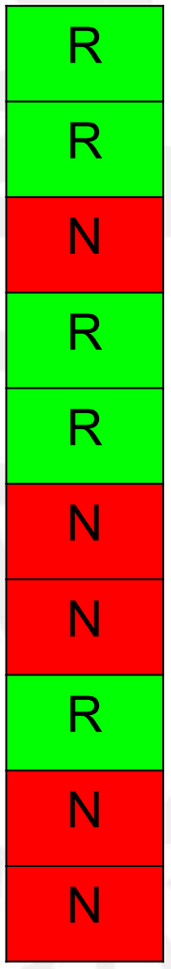
$$F_\beta = \frac{(\beta^2 + 1)P * R}{\beta^2 P + R} \qquad F_1 = \frac{2PR}{P + R}$$

- Harmonic mean penalize cases with high value only for one of them (rewards when both are roughly similar)

- Arithmetic mean is affected more by large values

Exercise

- We have a collection with 16 known relevant documents
- Our search engine retrieves the following top-10:
  R: relevant
  N: not-relevant



- Calculate precision, recall, F1

- Precision = 5/10 = 0.5
- Recall = 5/16 = 0.3125
- F1 = 2*(5/10*5/16)/(5/10+5/16) ≈ 0.38

Rank-based Retrieval Evaluation

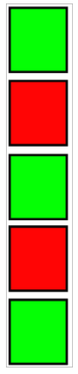
- Binary relevance
  - Precision@K (P@K)
  - Recall@K (R@K)
  - Mean Average Precision (MAP)
  - Mean Reciprocal Rank (MRR)

- Multiple levels of relevance
  - Normalized Discounted Cumulative Gain (NDCG)

Precision@K

- Set a rank threshold K
- Compute precision for top-K retrieved results
- Ignores documents ranked lower than K

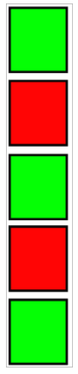
- Example:
  - Prec@3:
  - Prec@4:
  - Prec@5:



Precision@K

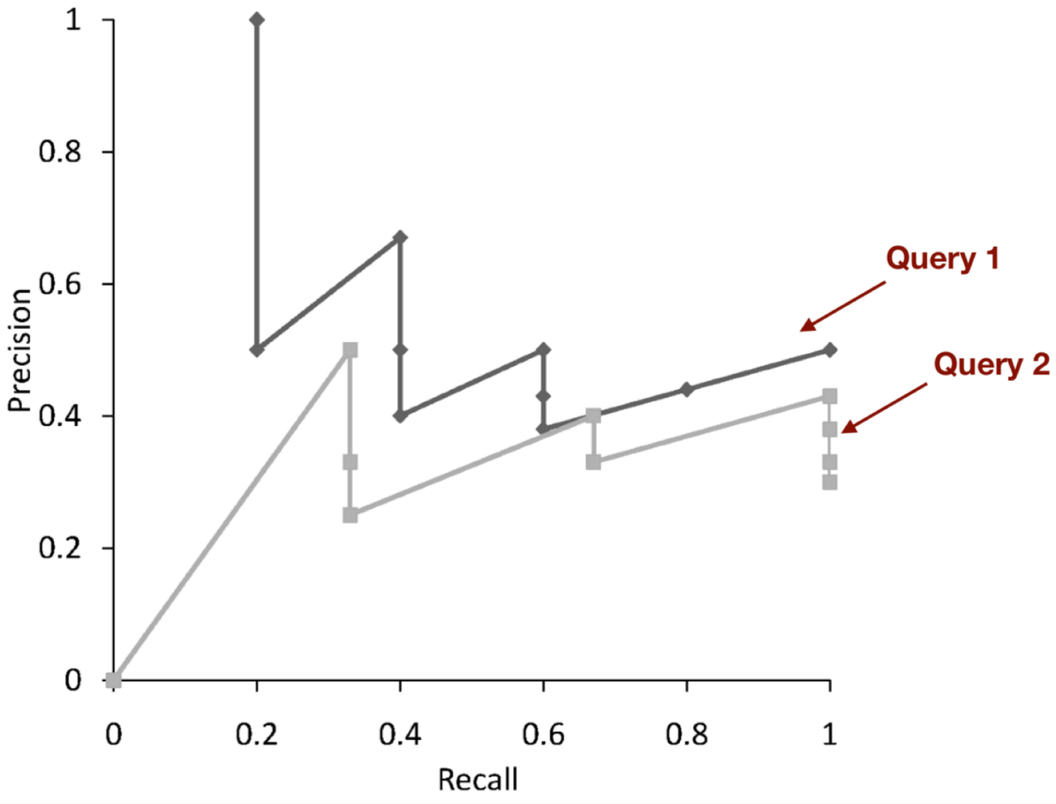
- Set a rank threshold K
- Compute precision for top-K retrieved results
- Ignores documents ranked lower than K

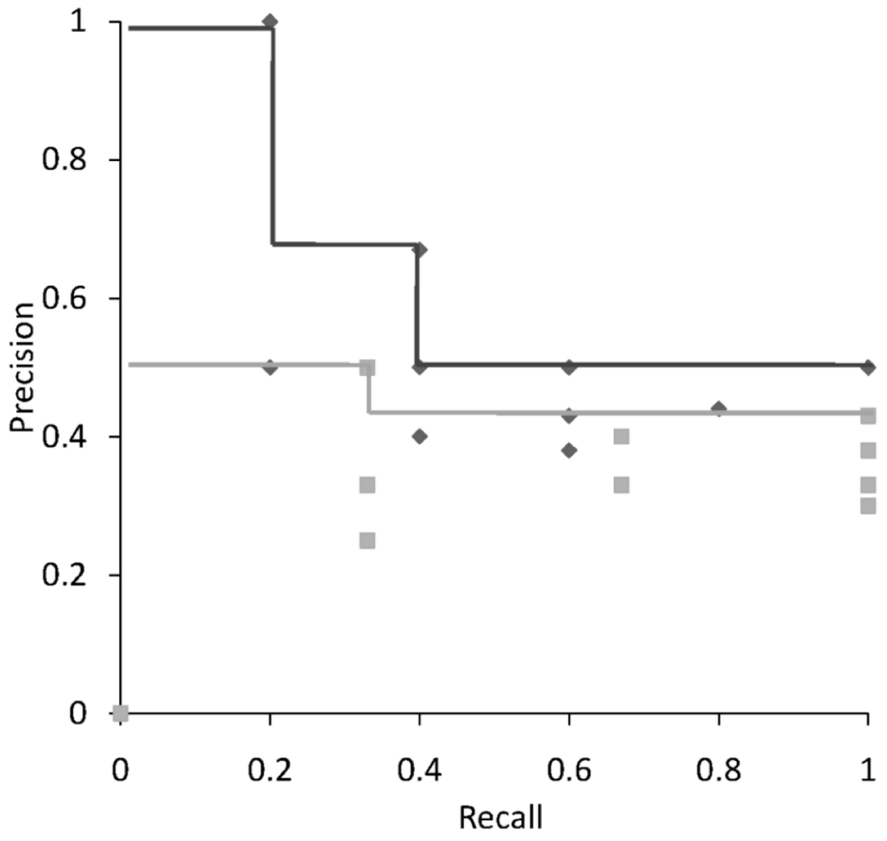
- Example:
  - Prec@3: 2/3
  - Prec@4: 2/4
  - Prec@5: 3/5



Precision-Recall Graph

- Precision-Recall graph produces a step function
- Interpolated Precision is used:
  - *Interpolated precision* at a certain recall level `r` is the highest precision found for any recall level `r'>= r`





MAP: Mean Average Precision

- Compute precision@K for each K_1, K_2, … K_R
  - I.e. precision at each point in the ranked list where recall increases)

- Average precision = average of P@K
  - Sum over these precision scores
  - Divide by the *total number of relevant documents in the collection*

- Take mean over all queries/rankings

Exercise

- We have a collection with 16 known relevant documents
- Our search engine retrieves the following top-10

- Sow the calculation of average precision

	Rank	R	P
R	1	1/16	1/1
R	2	2/16	2/2
N	3	2/16	2/3
R	4	3/16	3/4
R	5	4/16	4/5
N	6	4/16	4/6
N	7	4/16	4/7
R	8	5/16	5/*8*
N	9	5/16	5/9
N	10	5/16	5/10

$$\mathrm{AP} = \frac{1}{16}\left(\frac{1}{1} + \frac{2}{2} + \frac{3}{4} + \frac{4}{5} + \frac{5}{8}\right) = 0.26$$

MRR: Mean Reciprocal Rank

- **Assumption:** There is only one known relevant document

- Take reciprocal rank
  - 1/r where r is the position of the first relevant document

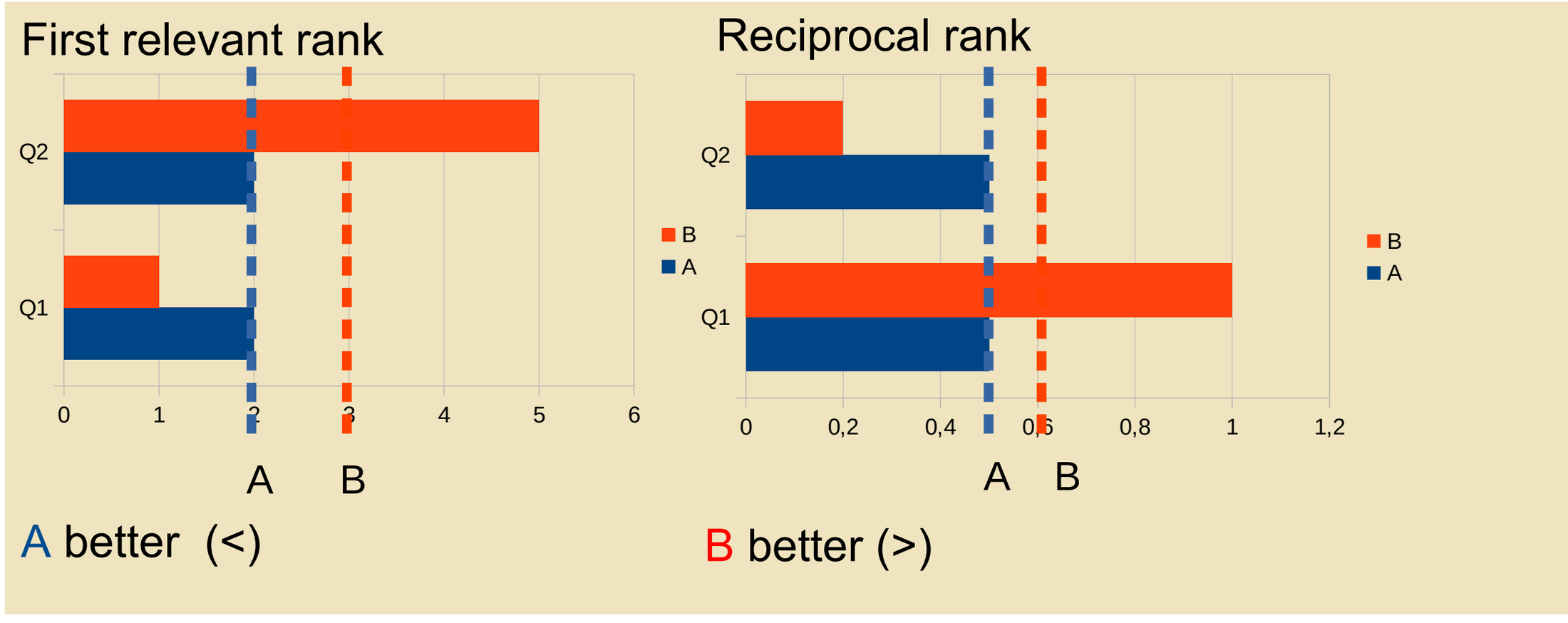
- Take average of all reciprocal ranks over all queries

- Theoretical problems (ratio scale, cannot average), but in wide use nevertheless.

MRR – a critical view

Reciprocal Rank

UNIVERSITÄT DUISBURG ESSEN
Open-Minded



Slides link: https://www.is.inf.uni-due.de/bib/pdf/talks/Fuhr_20ta.pdf