# INFORMATION RETRIEVAL

## L02. EVALUATION

SUZAN VERBERNE 2022

Universiteit Leiden

# I HAVE TO LEAVE AT 10.45 TODAY

INVITATION

To attend the public defence
of the dissertation

Digging
in
Documents

by Alex Brandsen

15-02-2022 @ 11.15

via: universiteitleiden.nl/
wetenschappers/livestream-
promotie

PARANYMPHS


Universiteit Leiden

# OUTLINE

- Evaluation in IR: the Cranfield paradigm
- Set evaluation measures
- Evaluation of ranked lists
- Evaluation with multi-level judgments
- Practical issues in evaluation

# EVALUATION IN IR

Universiteit Leiden

# EXERCISE

- Suppose that you developed a search engine for the archives of a large city. All documents are in one index. Through a standard query interface citizens can search for information contained in the archive. The results are presented in a list, ranked by relevance.
- The client asks you what the quality of the search engine is. How would you answer that question?
- 3 minutes, discuss with your neighbour.
- Write down some keywords of your solution

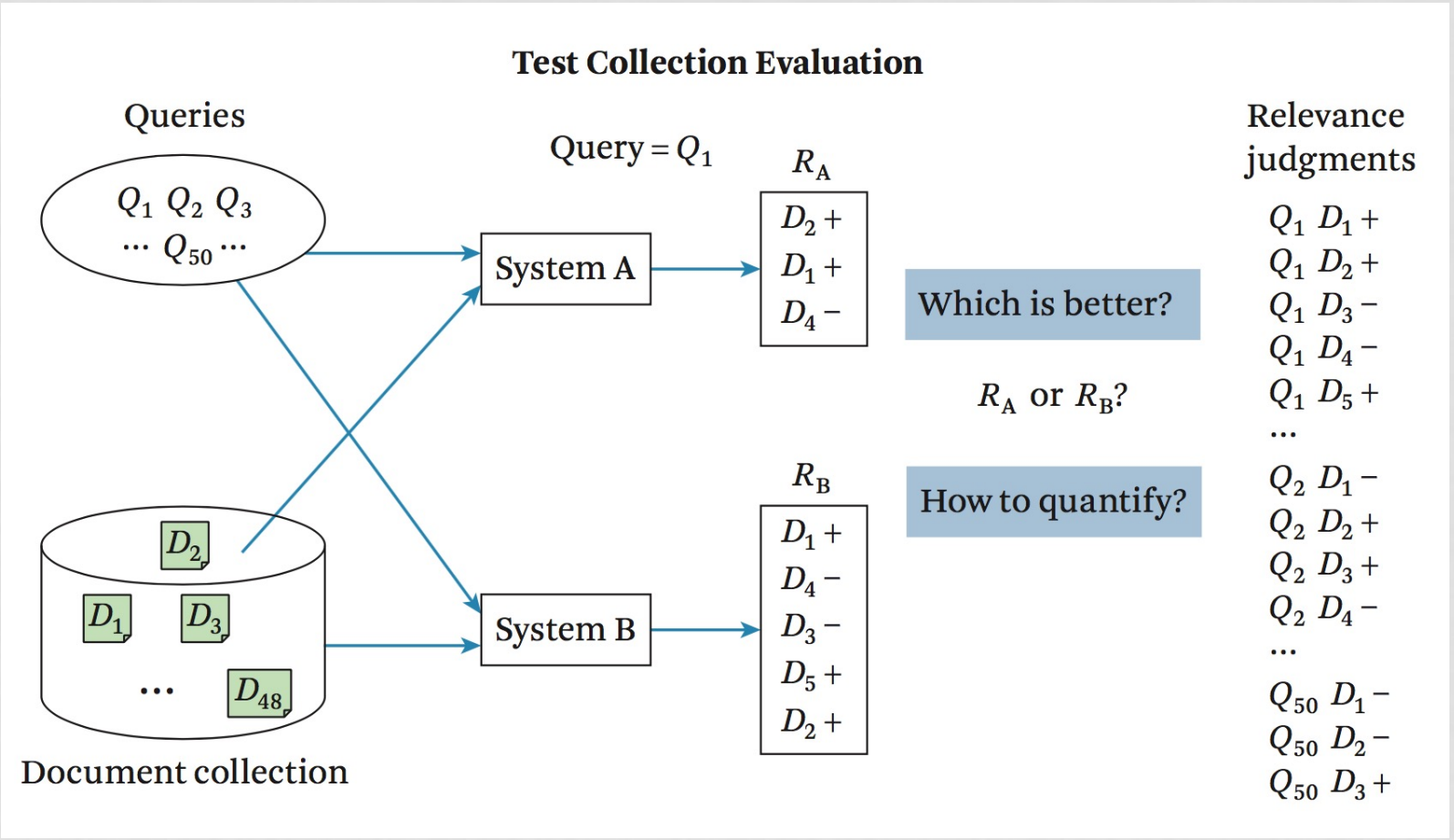
# WHAT IS NEEDED FOR THE EVALUATION OF A SEARCH ENGINE?

- What do you want to measure?
  - Effectiveness/accuracy
  - Efficiency
  - Usability
- Would you ask users?
- Test collection:
  - Search tasks, information needs, queries
  - Relevance judgments (manually!)
- A quantification of the quality: evaluation metrics
  - Precision, recall
  - An evaluation of the ranking: at which rank is/are the relevant result(s)

# THE CRANFIELD PARADIGM

- The Cranfield evaluation methodology was developed in the 1960s and is a strategy for laboratory testing of system components
- Idea:
  - build reusable test collections
  - define evaluation metrics for these collections
- Content of Information Retrieval test collection:
  - collection of documents similar to a real document collection in a search application
  - a sample set of queries or topics that simulate the user's information need
  - relevance judgments/assessments: which documents should be returned for which queries

# THE CRANFIELD PARADIGM




Zhai & Massung, 2016

Universiteit Leiden

# PRACTICAL NOTE

- Where we say ‘documents’ we mean ‘retrieval units’

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

# SET EVALUATION MEASURES

Universiteit Leiden

# PRECISION AND RECALL

- In the Cranfield paradigm, we have 2 sets of documents for each query:
  - the set of relevant documents (according to the relevance judgments)
  - the set of retrieved documents (by the search engine)
  - (we for now assume no ranking)

- Two central evaluation measures:

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

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

# PRECISION AND RECALL

| | | 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)

Universiteit Leiden

# F-MEASURE

- There is a tradeoff between precision and recall → combine them

$$F_\beta = \frac{\left(\beta^2 + 1\right) P * R}{\beta^2 P + R}$$

- β = 1: harmonic mean of precision and recall (most used):

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

- (why not the arithmetic mean? – find the answer in the book)

Universiteit Leiden

# EXERCISE

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

  R R N R R N N R N N

1. Calculate precision and recall.
2. Show the calculation of F1

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

# LIMITATIONS OF PRECISION AND RECALL FOR SEARCH ENGINE EVALUATION

- Can you think of limitations of precision and recall for search engine evaluation?
  - Relevance assessments tend to be incomplete → recall is unknown
  - Web search: we can do relevance assessments for the first 10 retrieved results → precision@10. Recall still unknown.
  - Ranking is not taken into account: a document retrieved in position 1000 will not be seen by the user
  - … and many more, but that is beyond the Cranfield paradigm

# EVALUATION OF RANKED LISTS

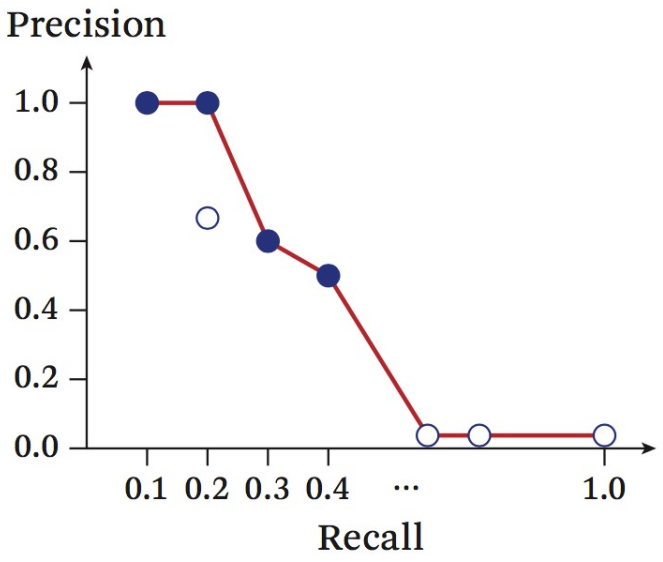
Universiteit
Leiden

# PRECISION-RECALL CURVE

**Evaluating Ranking: Precision–Recall (PR) Curve**

Total number of relevant documents in collection: 10

| | Precision | Recall |
|---|---|---|
| $D_1$ + | 1/1 | 1/10 |
| $D_2$ + | 2/2 | 2/10 |
| $D_3$ – | 2/3 | 2/10 |
| $D_4$ – | | |
| $D_5$ + | 3/5 | 3/10 |
| $D_6$ – | | |
| $D_7$ – | | |
| $D_8$ + | 4/8 | 4/10 |
| $D_9$ – | | |
| $D_{10}$ – | ? | 10/10 |

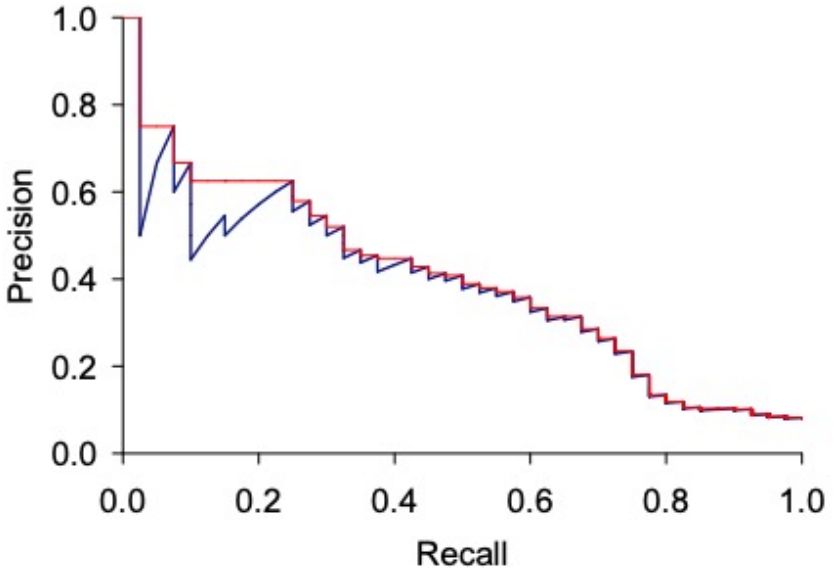



Zhai & Massung, 2016

Universiteit Leiden

# PRECISION-RECALL CURVE

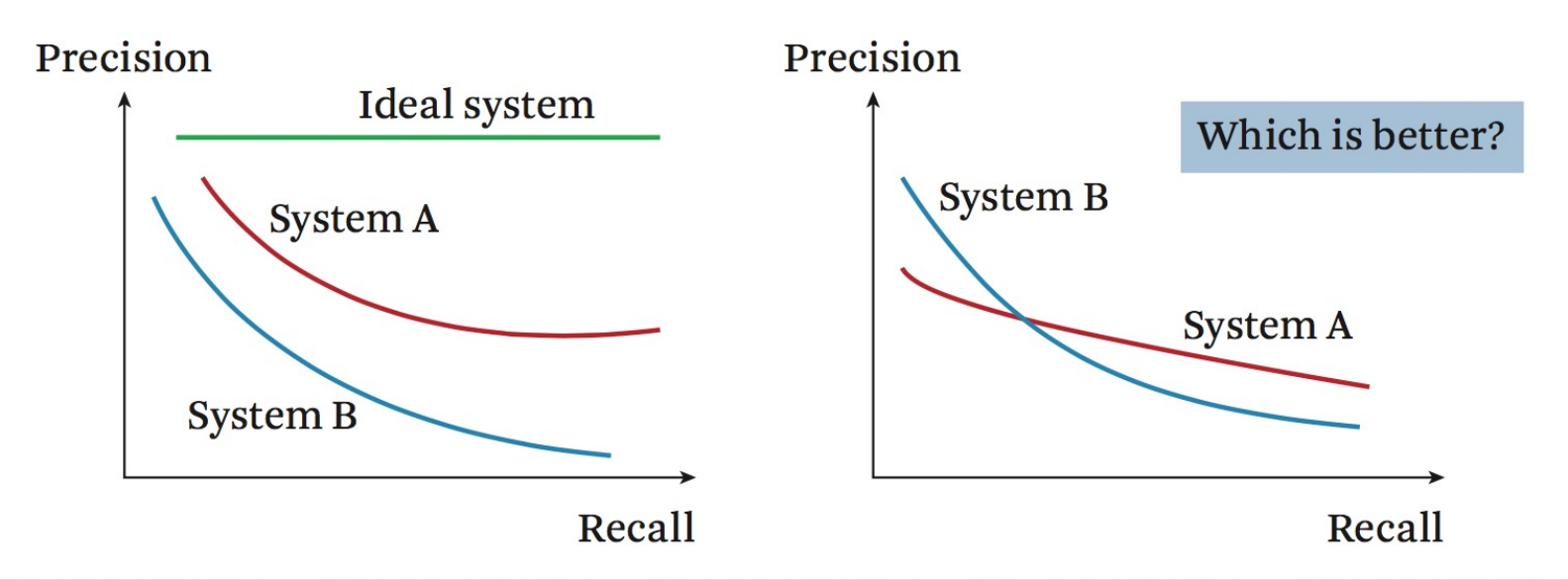
- The P-R curve as a continuously decreasing function, using interpolated precision: the highest precision found for any recall level (red line)




► **Figure 8.2** Precision/recall graph.

Universiteit Leiden

# PRECISION-RECALL CURVES




- The right picture is realistic, but makes it difficult to answer the question “Which is better?”
- Which system is better depends on the user’s task
- When comparing systems, we want 1 measure to summarize the PR-curve

# AVERAGE PRECISION (AP)

- Calculate precision at the position of each relevant retrieved document (at each point in the ranked list where recall increases)
- Sum over these precision scores
- Divide by the total number of relevant documents in the collection
- Definition:

$$\text{AveP} = \frac{\sum_{k=1}^{n} \left( P(k) \times \text{rel}(k) \right)}{\#\text{ of relevant documents in collection}}$$

- where n is the number of results in the ranked list that we consider (AP@n), P(k) is precision at position k, and rel(k) is an indicator function equaling 1 if the item at rank k is a relevant document, zero otherwise

# EXERCISE

- We have a collection with 16 relevant documents.
- Our search engine retrieves the following top-10:
- R R N R R N N R N N
- Show the calculation of Average Precision

$$\text{AveP} = \frac{\frac{1}{1}+\frac{2}{2}+\frac{3}{4}+\frac{4}{5}+\frac{5}{8}}{16} \approx 0.26$$

| | k | R | P |
|---|---|---|---|
| * | 1 | 1/16 | 1/1 |
| * | 2 | 2/16 | 2/2 |
| | 3 | 2/16 | 2/3 |
| * | 4 | 3/16 | 3/4 |
| * | 5 | 4/16 | 4/5 |
| | 6 | 4/16 | 4/6 |
| | 7 | 4/16 | 4/7 |
| * | 8 | 5/16 | 5/8 |
| | 9 | 5/16 | 5/9 |
| | 10 | 5/16 | 5/10 |

# EVALUATION FOR MULTIPLE QUERIES

- Calculate Average Precision (AP) per query
- Mean Average Precision (MAP) = the mean over all queries

Universiteit Leiden

# EVALUATION WITH MULTI-LEVEL JUDGMENTS

Universiteit Leiden

# CUMULATIVE GAIN

- What if the relevance judgments are not binary?
  - not relevant – relevant
- But multi-level?
  - not relevant – slightly relevant – relevant – highly relevant
  - or some other formulation:

> We set up an annotation interface in which we showed a question together with one retrieved video and the question "How well does the video answer the question?" Relevance labeling was done on a four-point scale: (3) Excellent answer; (2) Good answer; (1) May be good; (0) Not relevant. In order to validate

# CUMULATIVE GAIN

- What measure do we use for evaluation?
  - Assumption 1: Highly relevant results contribute more than slightly relevant results
- → Cumulative Gain
  - = sum of relevance judgments of retrieved documents

$$CG(L) = \sum_{i=1}^{n} r_i$$

where n is the number of results in the ranked list that we consider (CG@n) and $r_i$ = relevance grade for result i

Universiteit Leiden

# DISCOUNTED CUMULATIVE GAIN

- Assumption 2: the gain of a document should degrade with its rank
- The lower in the list, the lower the probability that the user sees it
- → Discounted cumulative gain

$$DCG(L) = r_1 + \sum_{i=2}^{n} \frac{r_i}{\log_2 i}$$

# NORMALIZED DISCOUNTED CUMULATIVE GAIN

- The effect of  of DCG depends on the number of relevant documents for a query → different scale per query
- Assumption 3: Better would be to have a 0-1 scale so that scores for queries can be compared
- → Normalized Discounted Cumulative Gain

$$nDCG(L) = \frac{DCG(L)}{iDCG}$$

- iDCG = the DCG for the ideally ranked list.
  - First all highly relevant documents, then the relevant documents, then the slightly relevant documents, and then the non relevant documents.

# CUMULATIVE GAIN

| | Gain | Cumulative gain |
|---|---|---|
| $D_1$ | 3 | 3 |
| $D_2$ | 2 | $3+2$ |
| $D_3$ | 1 | $3+2+1$ |
| $D_4$ | 1 | $3+2+1+1$ |
| $D_5$ | 3 | ... |
| $D_6$ | 1 | |
| $D_7$ | 1 | |
| $D_8$ | 2 | |
| $D_9$ | 1 | |
| $D_{10}$ | 1 | |

Relevance level: $r = 1$ (non-relevant), 2 (marginally relevant), 3 (very relevant)

Assume: there are 9 documents rated "3" in total in the collection

# CUMULATIVE GAIN

| | Gain | Cumulative gain | Discounted cumulative gain |
|---|---|---|---|
| $D_1$ | 3 | 3 | 3 |
| $D_2$ | 2 | 3 + 2 | 3 + 2/log **2** |
| $D_3$ | 1 | 3 + 2 + 1 | 3 + 2/log **2** + 1/log **3** |
| $D_4$ | 1 | 3 + 2 + 1 + 1 | ⋯ |
| $D_5$ | 3 | ⋯ | |
| $D_6$ | 1 | | |
| $D_7$ | 1 | | |
| $D_8$ | 2 | | |
| $D_9$ | 1 | | |
| $D_{10}$ | 1 | | |

$$\text{Normalized DCG} = \frac{\text{DCG@10}}{\text{IdealDCG@10}}$$

$$\text{DCG@10} = 3 + 2/\log \mathbf{2} + 1/\log \mathbf{3} + \cdots + 1/\log \mathbf{10}$$

$$\text{IdealDCG@10} = \mathbf{3} + \mathbf{3}/\log 2 + \mathbf{3}/\log 3 + \cdots + \mathbf{3}/\log 9 + \mathbf{2}/\log 10$$

Relevance level: $r = 1$ (non-relevant), 2 (marginally relevant), 3 (very relevant)

Assume: there are 9 documents rated "3" in total in the collection

# PRACTICAL ISSUES IN EVALUATION

Universiteit Leiden

# SET UP THE EVALUATION OF A RETRIEVAL SYSTEM

- Suppose that you developed a search engine for the archives of a large city. All documents are in one index. Through a standard query interface citizens can search for information contained in the archive. The results are presented in a list, ranked by relevance.
- The client asks you what the quality of the search engine is. How would you answer that question?

- → We need queries, documents and a set of relevance assessments
- What are the challenges?

# CHALLENGES

- Queries and documents:
  - Sufficient number of queries (dozens, preferably hundreds)
  - Queries and documents need to be representative of real users' information need
  - Sufficient relevant documents per query (number depends on task)

- 'Complete' relevance judgments:
  - Ideal: a judgment for each document in the collection for each query
  - Is infeasible with real collections
  - Alternative: create a pool of (say, 200) documents per query, retrieved by multiple (baseline) retrieval systems and have those judged

# ISSUES REGARDING RELEVANCE JUDGMENTS

- Manually creating relevance judgments is labor-intensive
- Use crowd-sourcing
  - Assessors who are not the 'owner' of the information need cannot always reliably judge the relevance of a document
  - Best is to use real users with real queries and let them do the judgments
- Automatically derive relevance judgments from click log data
  - A click is not necessarily an indicator of relevance
  - Click + 30 seconds dwell time? (still noisy evidence of relevance)
  - (we come back to that in lectures 9 and 11)

# SIGNIFICANCE TESTING

- Suppose we have the results of two retrieval systems on the same test collection. Which system is better? Are the differences significant?
- We have the same set of queries and one score per query
- → we can use a paired test
- Continuous data
  e.g. AP, nDCG
  → paired t-test
- Categorical data
  e.g. hit rate@10
  → McNemar's test

| query | AP system A | AP system B |
|---|---|---|
| 1 | 0.20 | 0.18 |
| 2 | 0.23 | 0.32 |
| 3 | 0.56 | 0.48 |
| 4 | 0.44 | 0.46 |
| 5 | 0.03 | 0.12 |
| 6 | 0.31 | 0.39 |
| MAP | 0.30 | 0.33 |

# BEYOND THE CRANFIELD PARADIGM

- Evaluation with real users:
  - A-B testing: have two versions of the ranking algorithm. Show one of the two (randomly) to a visitor. Measure the visitor's success
  - Users in a lab setting: observe behavior, measure success with questionnaire
- Evaluation with log data (queries, clicks)
- Evaluation with simulation

# CONCLUSIONS

Universiteit Leiden

# HOMEWORK

- Literature: IIR book, chapter 8
- Homework exercises on Brightspace (Assignments -> week 2)
  - Submit your answers through Brightspace before or on Sunday, February 20, 23.59
  - Submit 1 PDF file (any formatting is good)
- The solutions of the homework of week 1 are on Brightspace (Content -> Week 1)

# AFTER THIS LECTURE…

- You can design an evaluation study for a search engine following the Cranfield paradigm
- You can give the definitions of Precision, Recall, F1 and Mean Average Precision
- You can calculate Precision, Recall, F1 and Mean Average Precision for a given result set
- You can create and interpret Precision-Recall graphs
- You can explain the purpose and meaning of nDCG