Web Information Retrieval

Evaluation

(SOSE 2023)

Frank Hopfgartner, Stefania Zourlidou
Institute for Web Science and Technologies

Objectives of the lecture

universität koblenz
weiter:denken

- Why is an evaluation needed?
- What is the *Cranfield* paradigm?
- How is the relevance assessment created?
- What is *precision*, *recall* and *F1-score*?
- Which metrics evaluate a ranked result list?

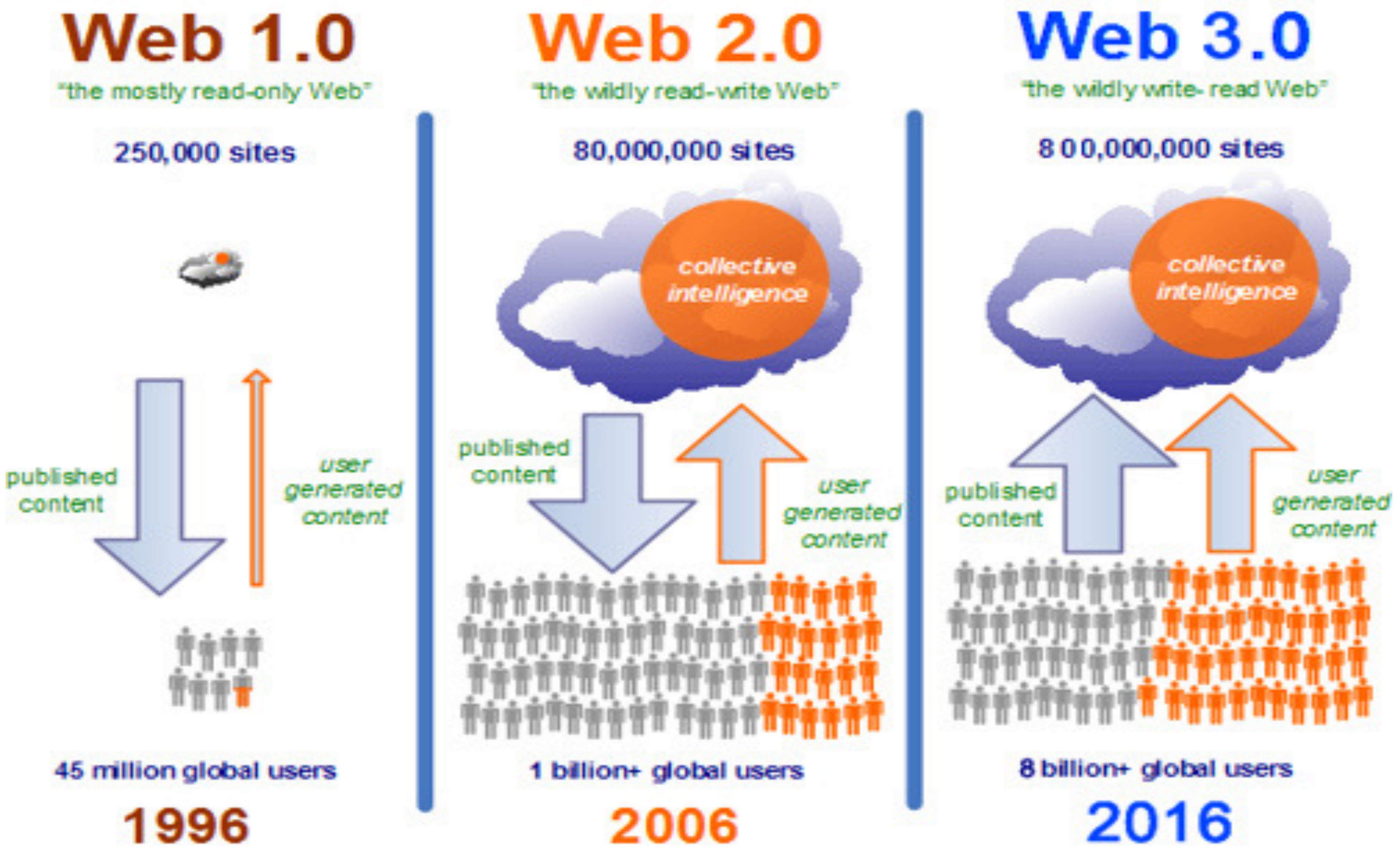
universität koblenz
weiter:denken

What makes WIR special?

What makes WIR special (different)?

- Larger than traditional information resources
- Presence of hyperlinks
- Data in semi-structured
- Evolves significantly
- Multiple content types (text, images, and even tables) + application
- Quality of document

Web evolution



How a web search engine works

- Web corpus collection (crawling)
- Preprocessing
- Indexing
- Document retrieval

Crawling the web

- Start from an initial page
- Retrieve all linked pages
- Iterate on new pages
- Do not visit the same page twice
- Avoid conflict and overlapping when crawling with parallel machines
- Crawl important pages (avoid leaving important pages)

Indexing

- It is the key to the effectiveness of a search engine
  - Retrieving relevant result quickly
- Avoids linear scanning of texts for each query

Evaluation of Information Systems

- General measures for software systems
  - Completeness, covering all requirements
  - Efficient use of resources (runtime, RAM, disk space bandwidth)
  - Usability
- Measures for database systems
  - Runtime indexing
  - Runtime querying
  - Max number of parallel users

Evaluation of WIR Systems

universität koblenz
weiter:denken

The key measure is user satisfaction

What is user satisfaction?

- Factors include
  - Speed of response
  - Size of index
  - Uncluttered UI
  - Most important: relevance
  - Free
- Note that none of these alone is sufficient to satisfy the user
  - Fast response & irrelevant result
  - Free & very small size of index
- How can we quantify user satisfaction?

Relevance

universität koblenz
weiter:denken

- The retrieved resource is relevant if it is appropriate to the information need (not a query). Otherwise, it is nonrelevant
- Types
  - Actual relevance: hard to estimate
  - Subjective relevance/ Pertinence: Relevance to a particular user
  - Objective relevance: External assessor(s)
  - System relevance: determined by an IR system
    - RSV (Retrieval Status Value)

Evaluation for IR Systems

universität koblenz
weiter:denken

- Particular to Retrieval
- Effectiveness in supporting the search for information
- Ease of finding (all) useful documents

How to evaluate an IR system?

universität koblenz
weiter:denken

- Given a test collection consisted of
  - A collection of resources, e.g. documents
  - A set of informations needs
    - Topics that are expressible as queries
  - A set of relevance judgements
    - typically a binary assessment being of either relevant or nonrelevant
  - Assessors
- Evaluate retrieval effectiveness
  - One assessor per resource/information need
  - Binary assessment
  - No agreement among assessors is required

Collection

- The assessments are called gold standards or ground truth
- The outcome of the evaluation is highly variable for different resources and information needs.
  - The test collection should be of reasonable size

Example Query/Topic (TREC 8)

- <num> Number: 412
- <title> airport security
- <desc> Description
  - What security measures are in effect or are proposed to go into effect in airports?
- <narr> Narrative
  - A relevant document could identify a specific airport and describe the security measures already in effect or proposed for use at that airport. Relevant items could also describe a failure of security that was cited as a contributing cause of a tragedy which came to pass or which was later averted. Comparisons between and among airports based on the effectiveness of the security of each are also relevant.

Corpora

- Classical corpora
  - Small, first testing

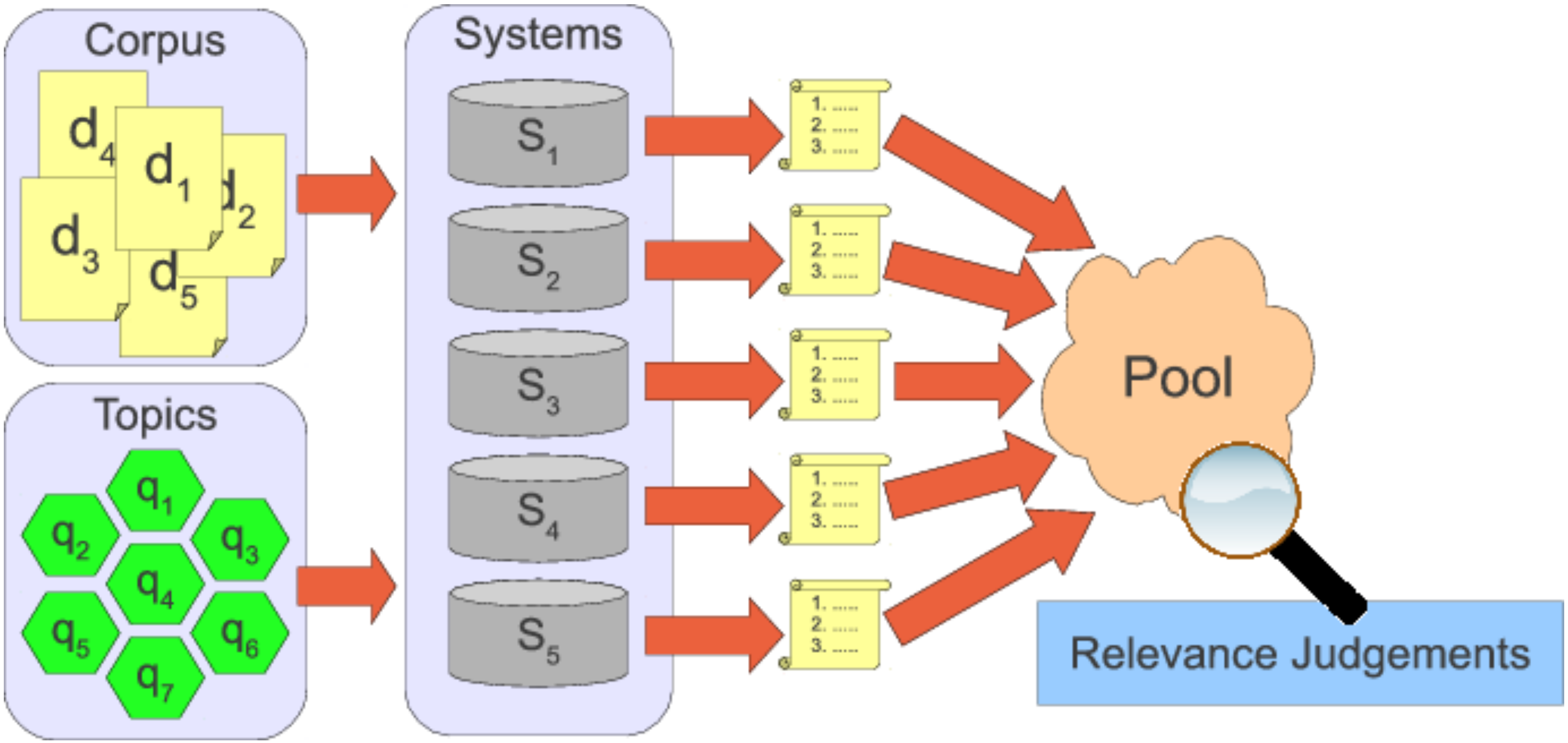
Corpus	Composition	Docs	Topics
Cranfield	Articles on aerodynamics	1,400	225
MED	Biomedical articles	1,033	30
TIME	News	425	83
CACM	Computing science papers	3,204	52

- Modern corpora
  - TREC, CLEF
  - Large, Very large
  - Different tasks
- Reuters CV1, CV2

Creating Relevance Assessments

- Assessor
  - Specialists
  - Computer support
  - Fast document scanning
- Old collections
  - Complete judgements
- But: TREC Terabyte Ad hoc Track 2005
  - 25.000.000 Documents, 50 Topics
  - Required time (theoretic)
    - 40 assessors, 10s / document, 8h /day
    - Total: 29.7 years
- Solution: Pooling

Pooling


Corpus
d_4 d_1 d_2 d_3 d_5
Topics
q_1 q_2 q_3 q_4 q_5 q_6 q_7
Systems
S_1 S_2 S_3 S_4 S_5
1. ......
2. ......
3. ......
Pool
Relevance Judgements

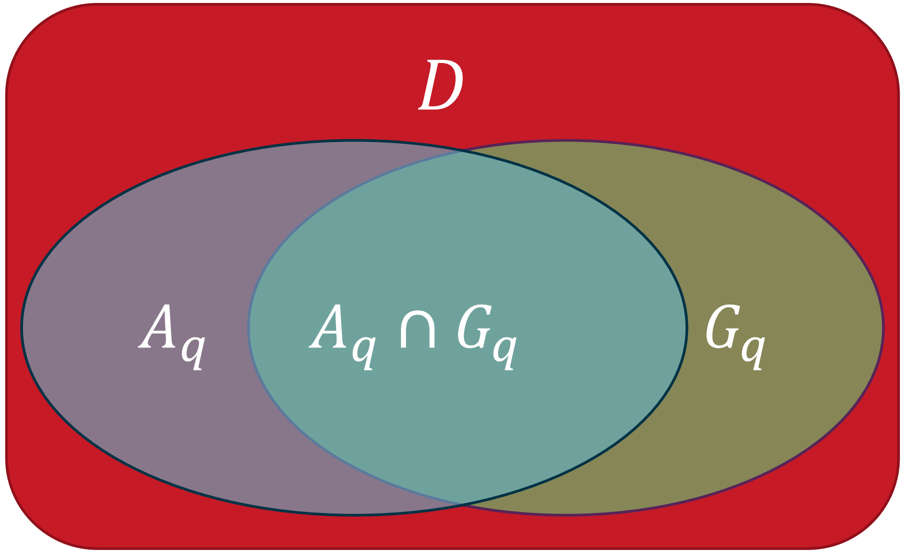
Crowdsourcing Relevance Judgements

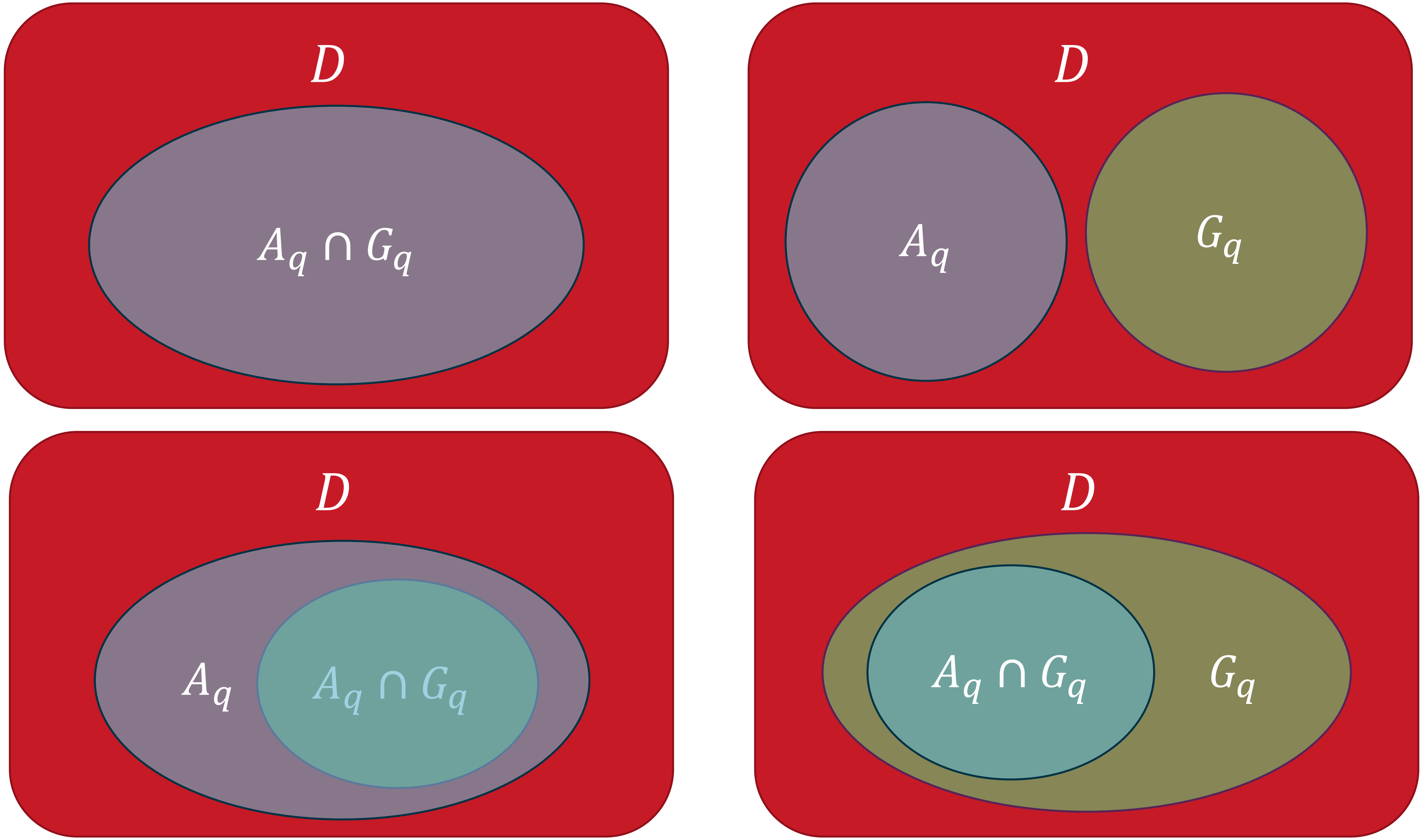
- Use non-professional assessors
  - Massive parallel assessments
  - Established platform: Amazon Mechanical Turk
- Benefits
  - Fast
  - Cheap: 0.01 to 0.05 cents per judgement
- Issues
  - Agreement of assessors
  - Spam
  - User interface

Metrics ignoring the ranking

A typical retrieval scenario

- $D = \{d_1, d_2, \ldots, d_N\}$ is the collection of N resources
- q is the query
- G_q is the gold standard set that corresponds to q
- A_q is the retrieved result given q




D
A_q ∩ G_q
D
A_q
G_q
D
A_q
A_q ∩ G_q
D
A_q ∩ G_q
G_q

Confusion matrix

- Each document d is either retrieved or not, and either relevant or not. This induces the following confusion matrix:

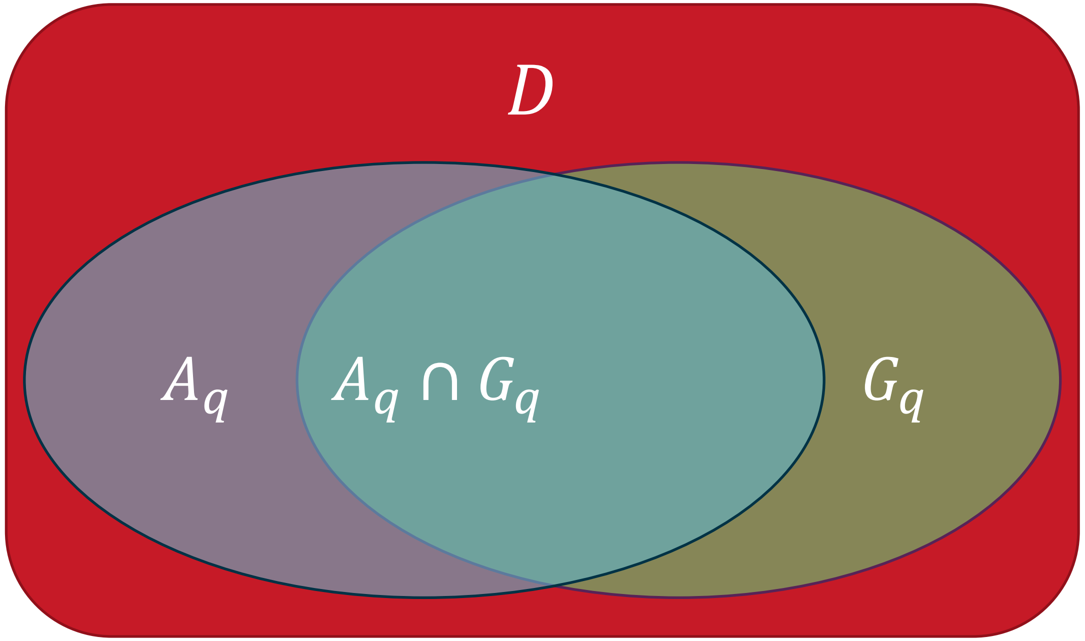
$$\begin{array}{r c} & \begin{array}{cc} \text{relevant} & \text{not relevant} \end{array} \\ \begin{array}{r} \text{retrieved} \\ \text{not retrieved} \end{array} & \begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix} \end{array}$$

$$\begin{array}{r c} & \begin{array}{cc} \text{relevant} & \text{not relevant} \end{array} \\ \begin{array}{r} \text{retrieved} \\ \text{not retrieved} \end{array} & \begin{pmatrix} hits & noise \\ misses & rejected \end{pmatrix} \end{array}$$

Confusion matrix

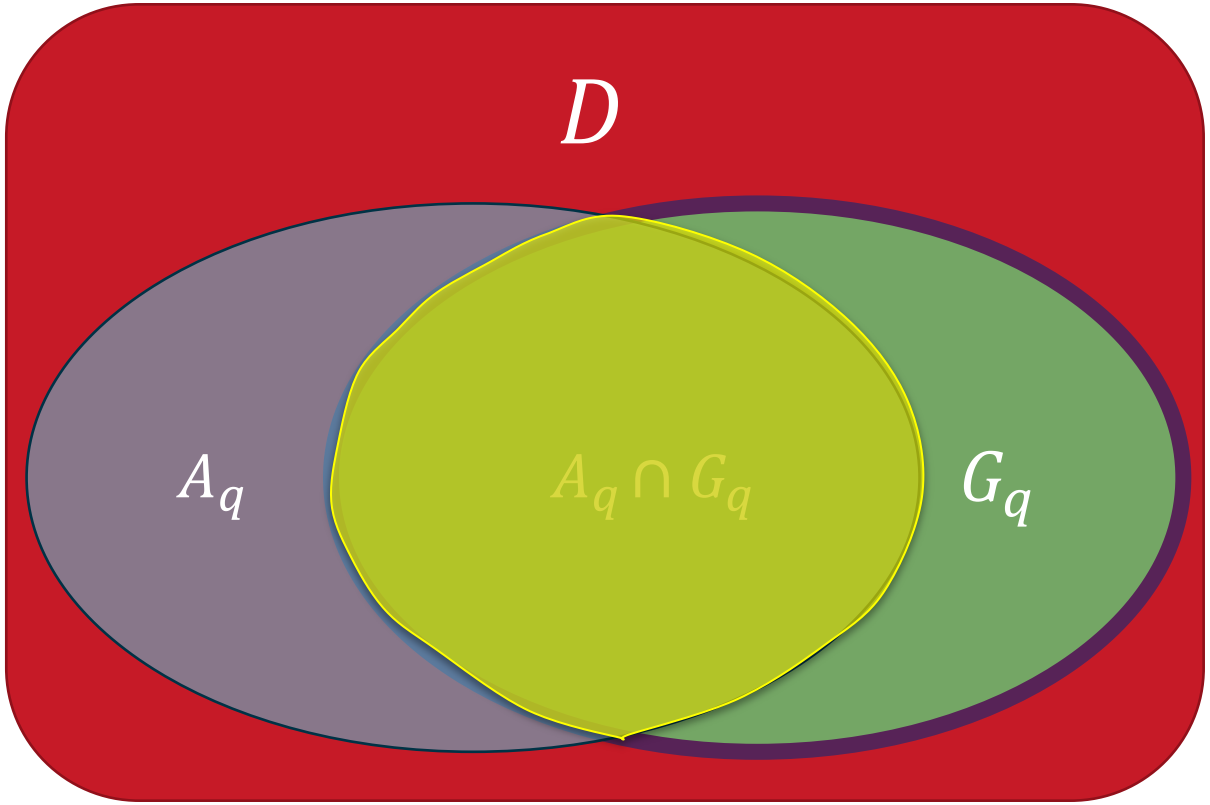
$$\begin{array}{r c c} & \text{relevant} & \text{not relevant} \\ \begin{array}{r}\text{retrieved} \\ \text{not retrieved}\end{array} & \multicolumn{2}{c}{\begin{pmatrix} TP & FP \\ FN & TN \end{pmatrix}} \end{array}$$

- $A_q \cap G_q = TP$
- $G_q = TP + FN$
- $A_q = TP + FP$



Recall

- Among all relevant resources, which fraction is retrieved?
- $r = \frac{|A_q \cap G_q|}{|G_q|}$
- $r = \frac{TP}{TP+FN}$

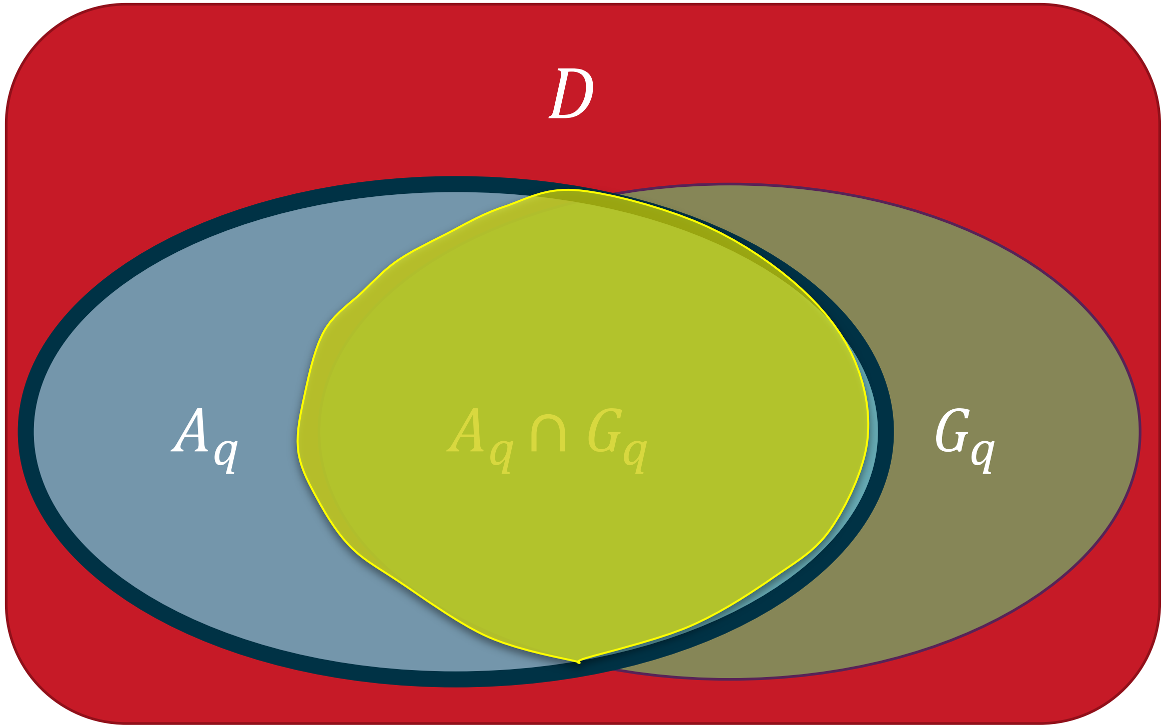


Recall: an example

- Given
  - the collection $D = \{d_1, d_2, \dots, d_{100}\}$
  - a query q
  - the corresponding gold standard $G_q = \{d_2, d_3, d_6, d_8, d_{10}, d_{14}, d_{17}, d_{29}\}$
  - the corresponding retrieved set $A_q = \{d_2, d_3, d_4, d_7, d_8, d_{10}, d_{12}, d_{17}, d_{20}, d_{29}\}$
- Recall
  - $r = \frac{|A_q \cap G_q|}{|G_q|} = \frac{|\{d_2, d_3, d_8, d_{10}, d_{17}, d_{29}\}|}{|\{d_2, d_3, d_6, d_8, d_{10}, d_{14}, d_{17}, d_{29}\}|} = \frac{6}{8} = 0{,}75$

Precision

- Among all retrieved resources, which fraction is relevant?
- $p = \frac{|A_q \cap G_q|}{|A_q|}$
- $p = \frac{TP}{TP+FP}$



Precision: an example

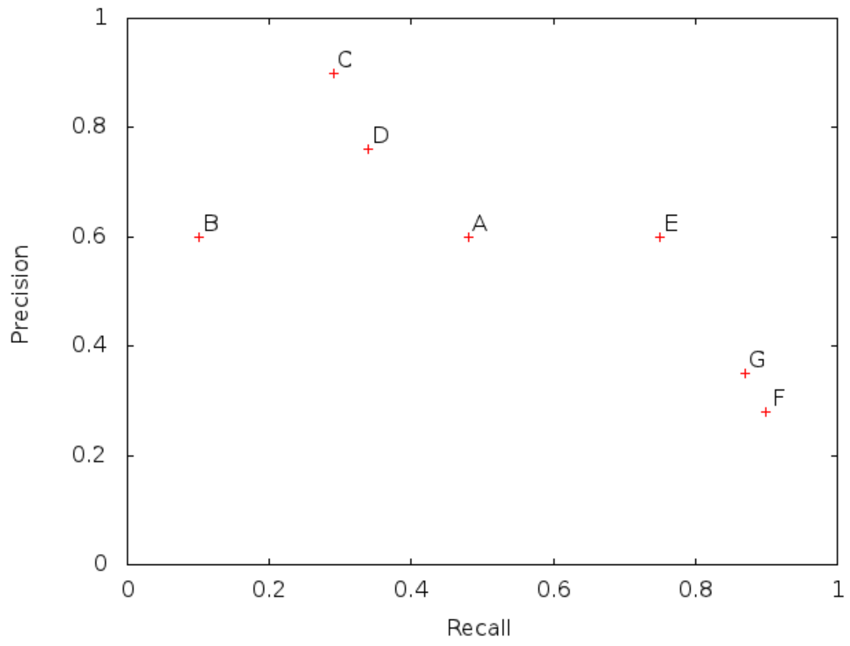
- Given
  - the collection $D = \{d_1, d_2, \dots, d_{100}\}$
  - a query q
  - the corresponding gold standard $G_q = \{d_2, d_3, d_6, d_8, d_{10}, d_{14}, d_{17}, d_{29}\}$
  - the corresponding retrieved set $A_q = \{d_2, d_3, d_4, d_7, d_8, d_{10}, d_{12}, d_{17}, d_{20}, d_{29}\}$
- Precision
  - $p = \frac{|A_q \cap G_q|}{|A_q|} = \frac{|\{d_2, d_3, d_8, d_{10}, d_{17}, d_{29}\}|}{|\{d_2, d_3, d_4, d_7, d_8, d_{10}, d_{12}, d_{17}, d_{20}, d_{29}\}|} = \frac{6}{10} = 0{,}6$

Properties of Precision and Recall

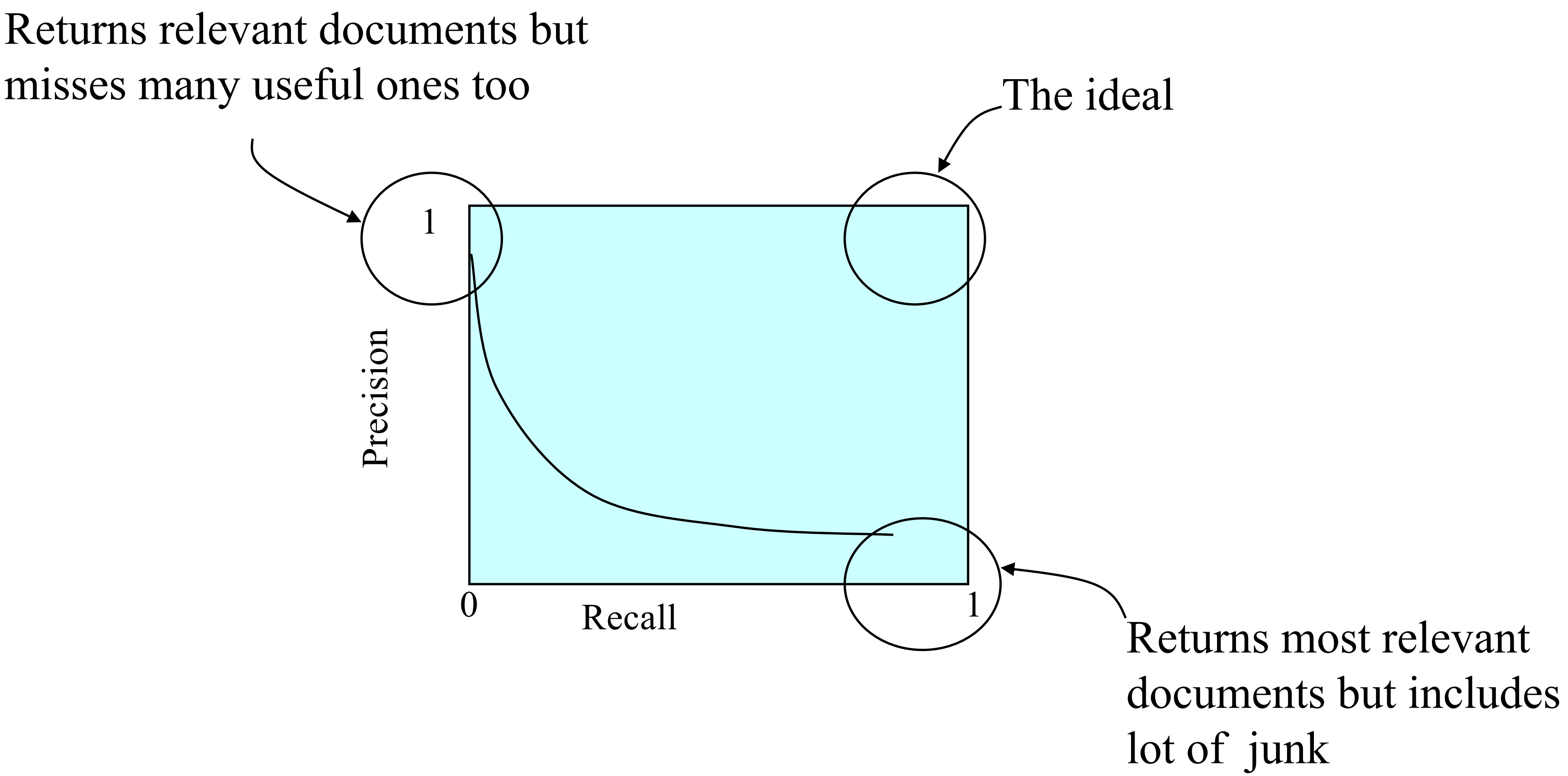
universität koblenz
weiter:denken

- Range [0,1]
- High values are better
  - Recall of 1 can always be obtained
  - High precision can be influenced
- Values are opposed
- How to compare Systems?
  - Application might dictate preference of recall or precision

System	Recall	Precision
A	0.48	0.60
B	0.10	0.60
C	0.29	0.90
D	0.34	0.76
E	0.75	0.60
F	0.90	0.28
G	0.87	0.35



Trade-offs



Returns relevant documents but
misses many useful ones too
The ideal
1
Precision
0
Recall
1
Returns most relevant
documents but includes
lot of junk

F-Measure

- Combines recall and precision (weighted harmonic mean)

$$H_\alpha(r,p)=\frac{1}{\alpha\frac{1}{p}+(1-\alpha)\frac{1}{r}}$$

- Typically formulated as F-Measure:

$$F_\beta=(\beta^2+1)\frac{rp}{\beta^2 p+r}$$ by setting $$\alpha=\frac{\beta^2}{\beta^2+1}$$

- (Nearly) always used with β=1 : $$F_1=\frac{2rp}{p+r}$$

- This means that the precision and recall are equaly important
- If precision is more important than recall, we set $\beta < 1$. Otherwise, we set $\beta > 1$

van Rijsbergen [8]

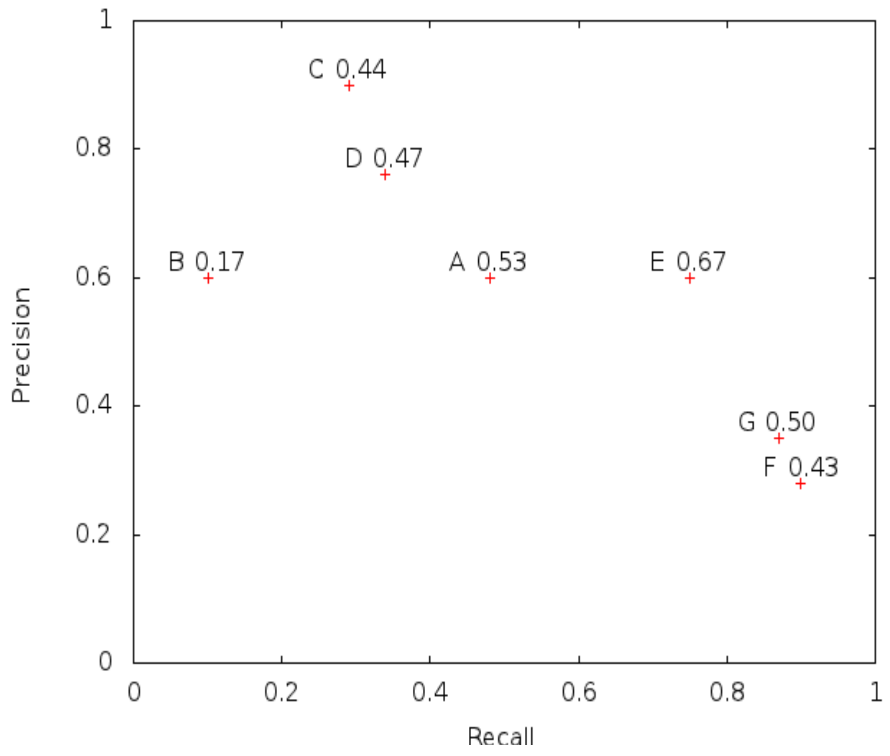
F1-Score: an example

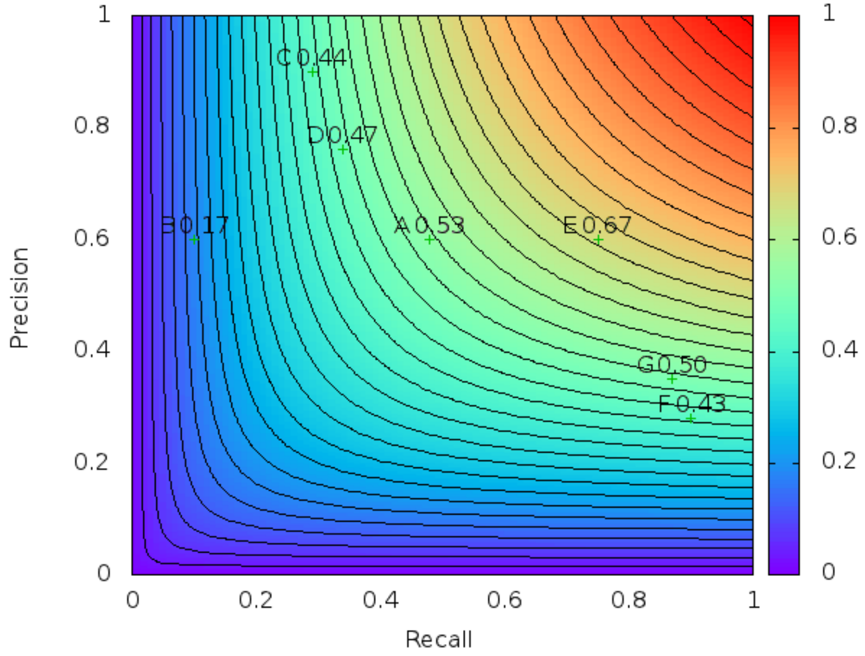
- Given
  - the collection $D = \{d_1, d_2, \dots, d_{100}\}$
  - a query q
  - the corresponding gold standard $G_q = \{d_2, d_3, d_6, d_8, d_{10}, d_{14}, d_{17}, d_{29}\}$
  - the corresponding retrieved set $A_q = \{d_2, d_3, d_4, d_7, d_8, d_{10}, d_{12}, d_{17}, d_{20}, d_{29}\}$
- F1-score
  - $F1 = 2\frac{rp}{p+r} = \frac{2\times 0{,}6\times 0{,}75}{0{,}6+0{,}75} = 0{,}667$

Properties of F1

- Range [0,1]
- High values are better

System	Recall	Precision	F1
A	0.48	0.60	0.53
B	0.10	0.60	0.17
C	0.29	0.90	0.44
D	0.34	0.76	0.47
E	0.75	0.60	0.67
F	0.90	0.30	0.43
G	0.87	0.32	0.50

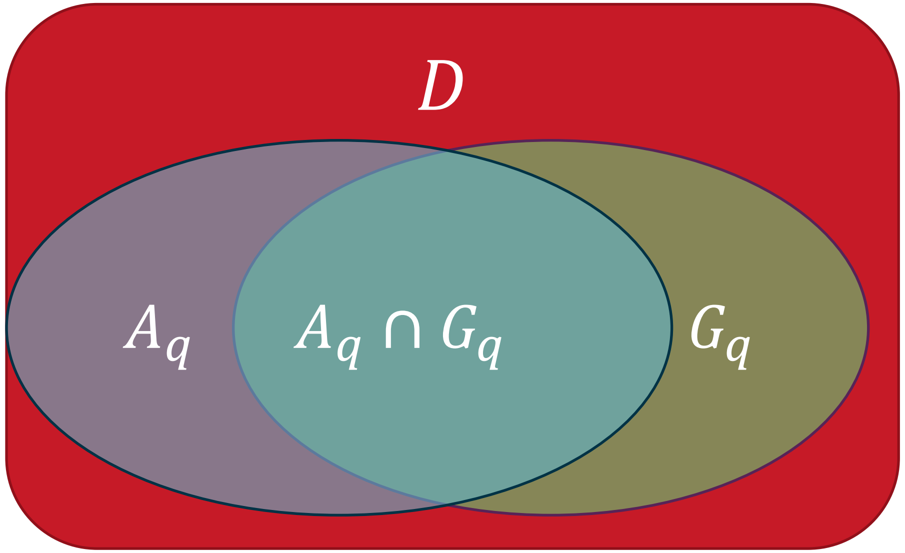




Accuracy

universität koblenz
weiter:denken

- Accuracy is the fraction of correct decisions
  - $Acc = \frac{TP+TN}{TP+TN+FP+FN}$
  - $= \frac{|A_q \cap G_q| + |D \setminus \{A_q \cup G_q\}|}{|D|}$



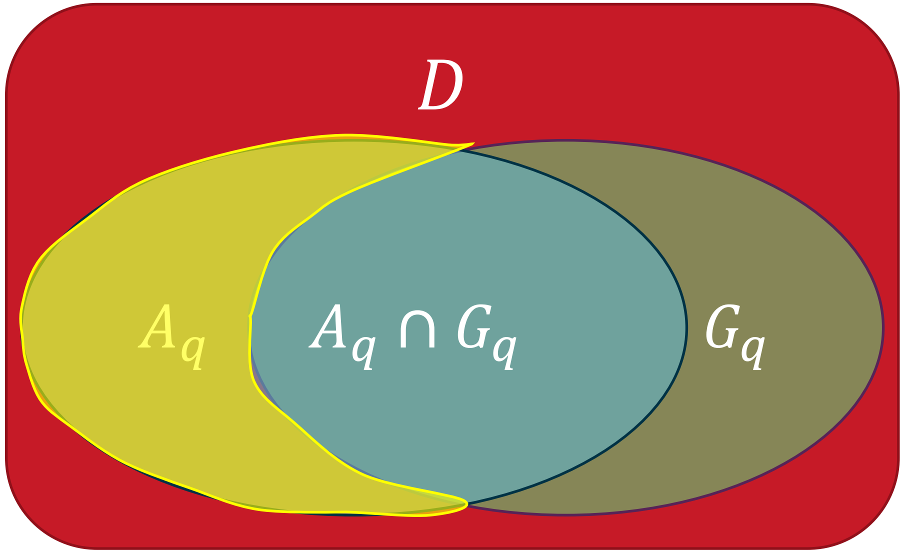
- Considering the size of D, accuracy is not a good measure for IR systems.
- If for every query, a ($a \rightarrow |D|$) resources are not relevant, a system which does not retrieve anything will get an accuracy $= a/|D|$

Accuracy: an example

- Given
  - the collection $D = \{d_1, d_2, \dots, d_{100}\}$
  - a query q
  - the corresponding gold standard $G_q = \{d_2, d_3, d_6, d_8, d_{10}, d_{14}, d_{17}, d_{29}\}$
  - the corresponding retrieved set $A_q = \{d_2, d_3, d_4, d_7, d_8, d_{10}, d_{12}, d_{17}, d_{20}, d_{29}\}$
- Accuracy
  - $Acc = \frac{A_q \cap G_q + D \setminus \{A_q \cup G_q\}}{D} = \frac{|\{d_2,\ d_3,\ d_8,\ d_{10},\ d_{17},\ d_{29}\}| + |\{d_1, \dots\}|}{\{d_1, d_2, \dots, d_{100}\}} = \frac{6+88}{100} = 0{,}94$

Fallout

- Fallout is the fraction of the noise that the system exposes to the user
  - $Fallout = \frac{|A_q \setminus G_q|}{|D \setminus G_q|}$



- Considering the size of D, fallout is of little use to evaluate IR systems

Fallout: an example

- Given
  - the collection $D = \{d_1, d_2, \dots, d_{100}\}$
  - a query q
  - the corresponding gold standard $G_q = \{d_2, d_3, d_6, d_8, d_{10}, d_{14}, d_{17}, d_{29}\}$
  - the corresponding retrieved set $A_q = \{d_2, d_3, d_4, d_7, d_8, d_{10}, d_{12}, d_{17}, d_{20}, d_{29}\}$
- Fallout
  - $Fallout = \frac{|A_q \setminus G_q|}{|D \setminus G_q|} = \frac{|\{d_4, d_7, d_{12}, d_{20}\}|}{\{d_1, d_4, \dots\}} = \frac{4}{92} = 0{,}043$

Evaluation

- Precision, Recall, F-score are good for evaluating the performance of Boolean retrieval systems (Relevant and Non-relevant)
- They cannot evaluate rankings
- For example, [R,R,N,N] and [N,N,R,R] will be evaluated similarly by these measures
  - R: relevant
  - N: Non-relevant

Ranking Aware Metrics

Typical Ranked Retrieval Setting

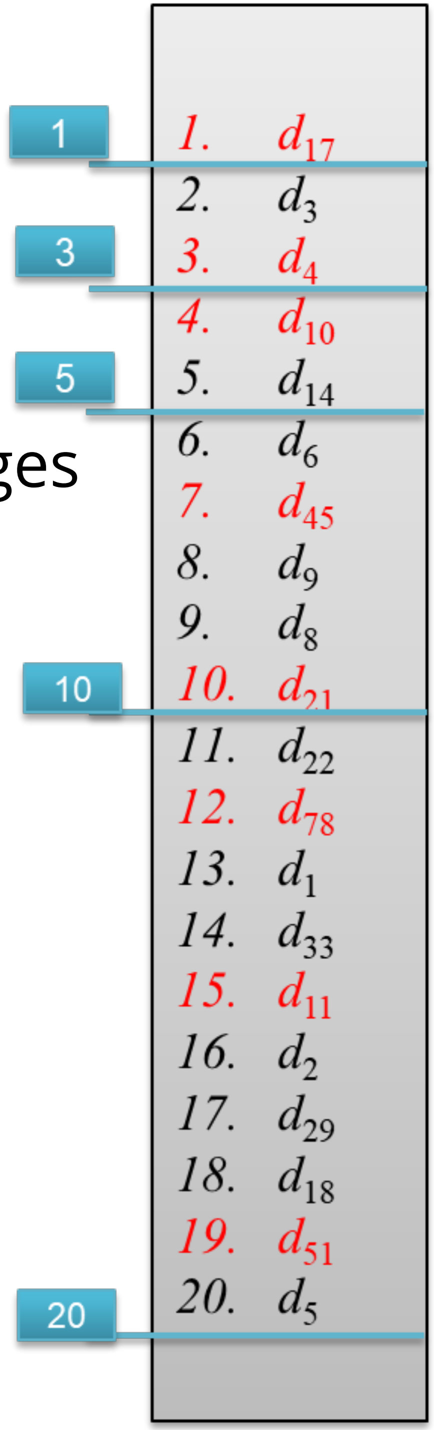
- $D = \{d_1, d_2, \ldots, d_N\}$ is the collection of N resources
- q is the query
- G_q is the gold standard set that corresponds to q
- L_q is the ordered retrieved result given q
  - Order of relevance
- Example
  - $G_q = \{d_4,\ d_{10},\ d_{11},\ d_{17},\ d_{21},\ d_{45},\ d_{51},\ d_{78}\}$

G_q	$\{d_4,\ \ d_{10},\ \ d_{11},\ \ d_{17},\ \ d_{21},\ \ d_{45},\ \ d_{51},\ \ d_{78}\}$
L_q	$\{d_{17}, d_3, d_4, d_{10}, d_{14}, d_6, d_{45}, d_9, d_8, d_{21}, d_{22}, d_{78}, d_1, d_{33}, d_{11}, d_2, d_{29}, d_{18}, d_{51}, d_5\}$
	$\{d_{17}, d_3, d_4, d_{10}, d_{14}, d_6, d_{45}, d_9, d_8, d_{21}, d_{22}, d_{78}, d_1, d_{33}, d_{11}, d_2, d_{29}, d_{18}, d_{51}, d_5\}$

Precision at k (p@k)

- Fixed cutoff (k) in results list
- Motivation from UI
  - Systems deliver chunks of result list as pages
  - Users rarely go beyond first page
- Determine precision at cutoff (p@k)
- Example

k	# relevant docs	p@k
1	1	1.000
3	2	0.667
5	3	0.600
10	5	0.500
20	8	0.400



universität koblenz
weiter:denken

R-Precision

- Problem of p@k
  - Choice of k?
  - Less than k relevant documents
  - Stability
- R-Precision
  - Flexible cutoff at $|G|$
  - Precision-recall break-even: $|G| = |A|$
- Example
  - $G_q = \{d_4,\ d_{10},\ d_{11},\ d_{17},\ d_{21},\ d_{45},\ d_{51},\ d_{78}\}$
  - $p_R = \frac{4}{8}$

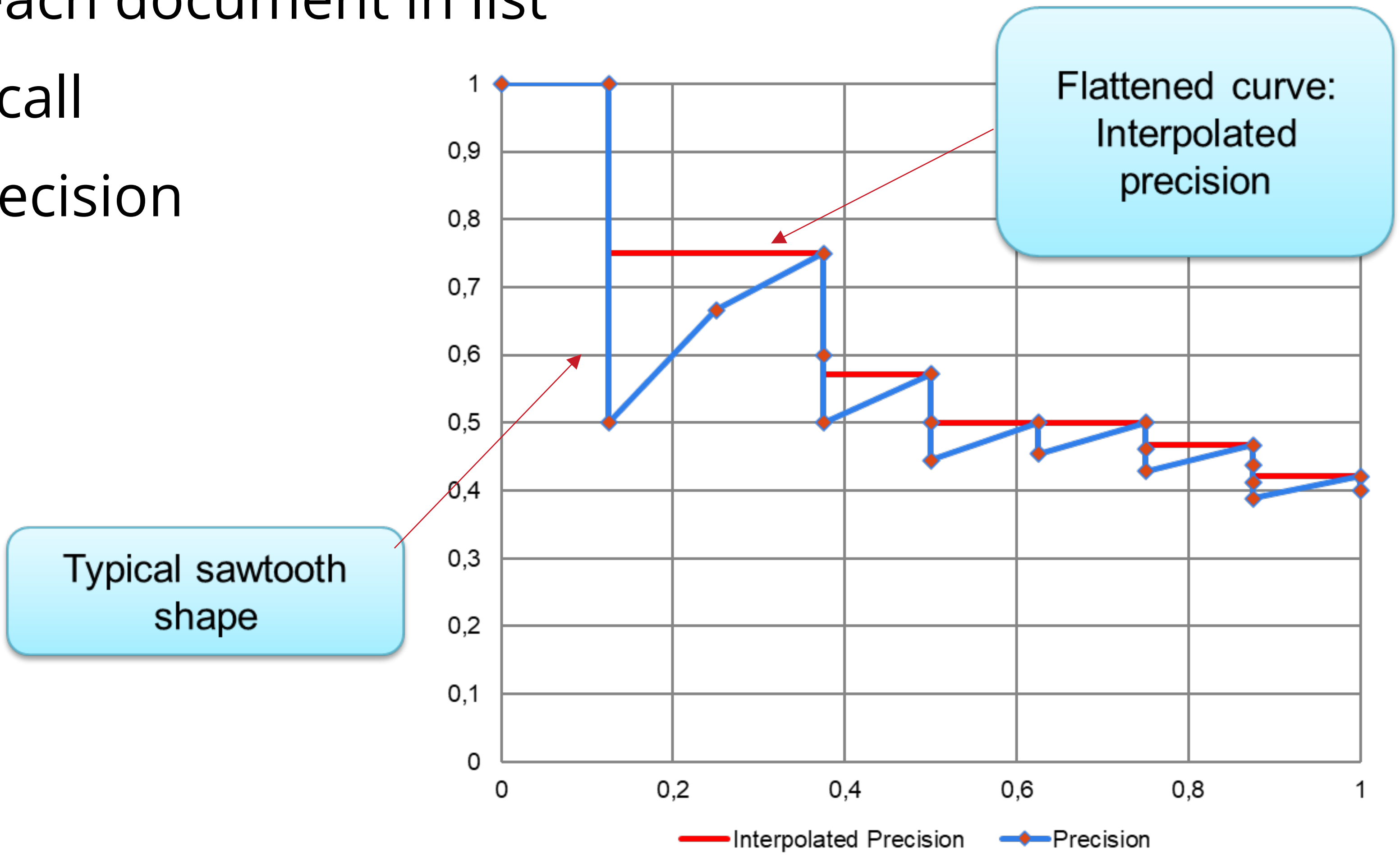
1. d_{17}
2. d_3
3. d_4
4. d_{10}
5. d_{14}
6. d_6
7. d_{45}
8. d_9 |G|
9. d_8
10. d_{21}
11. d_{22}
12. d_{78}
13. d_1
14. d_{33}
15. d_{11}
16. d_2
17. d_{29}
18. d_{18}
19. d_{51}
20. d_5

Precision Recall Graph

- Plot evolution of recall and precision in result list (no function)
- For each document in list

  x: recall

  y: precision

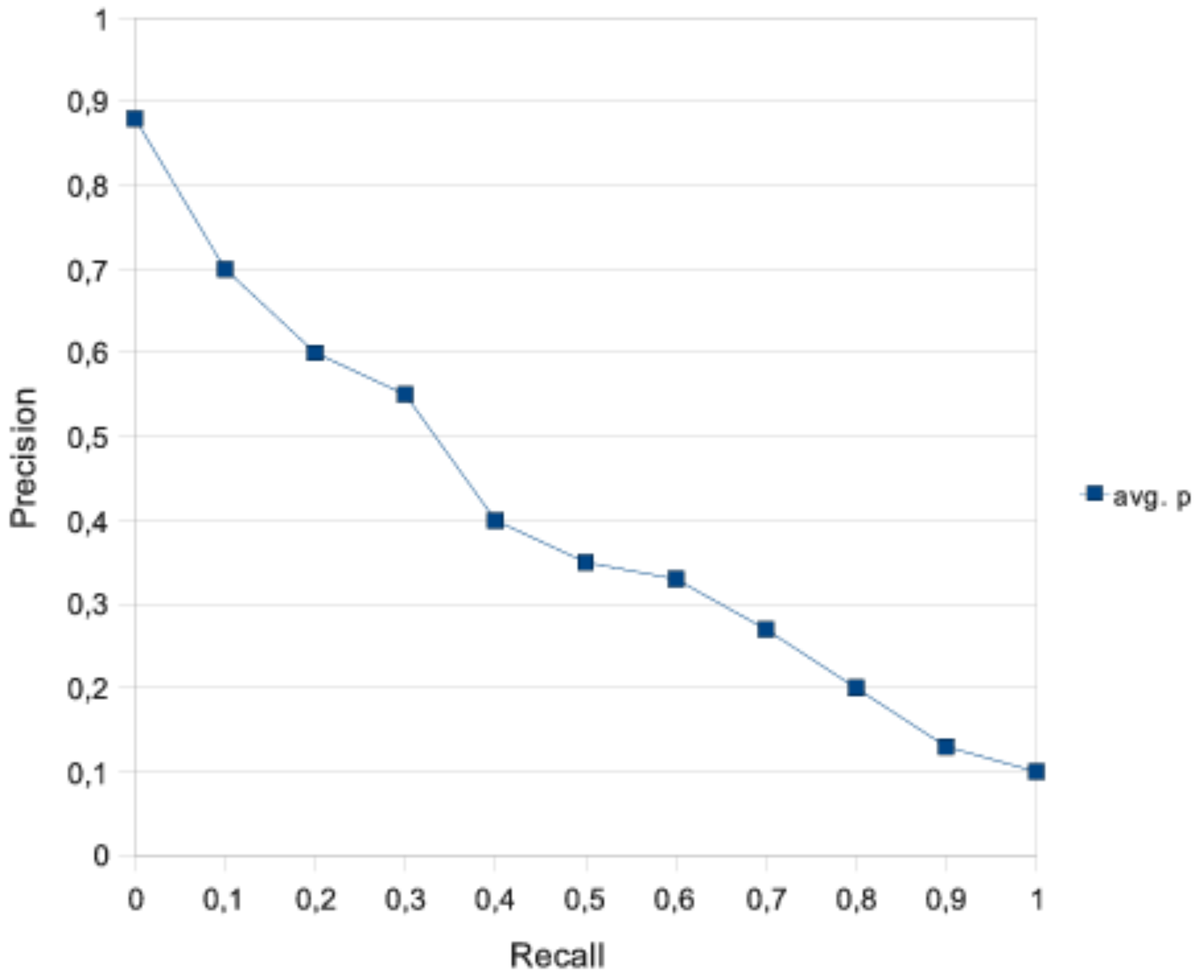


1. d_{17}
2. d_3
3. d_4
4. d_{10}
5. d_{14}
6. d_6
7. d_{45}
8. d_9
9. d_8
10. d_{21}
11. d_{22}
12. d_{78}
13. d_1
14. d_{33}
15. d_{11}
16. d_2
17. d_{29}
18. d_{18}
19. d_{51}
20. d_5

11-Point Precision Recall Graph

- Fixed set of recall values
  - 0 to 1, steps 0.1
  - Interpolated precision

    p_interp(r) = max_{r' >= r} p(r')



Mean Average Precision

- Mean Average Precision:

$$MAP = \frac{1}{|Q|} \sum_{i=1}^{|Q|} \frac{1}{m_i} \sum_{j=1}^{m_i} P\left(k_{ij}\right)$$

- One integrated value for the quality of a ranking
  - m_i number of relevant documents for query q_i
  - k_{ij} position of the j-th relevant document for query q_i
  - $P(k_{ij})$ precision @ k_{ij} for query q_i (set to 0 if document is not in the result list)

MAP – an example

- Average precision (AP) for one query

Document	Position	Precision
d_{17}	1	1.000
d_4	3	0.667
d_{10}	4	0.750
d_{45}	7	0.571
d_{21}	10	0.500
d_{78}	12	0.500
d_{11}	15	0.467
d_{51}	19	0.421
Average Precision		***0.609***

- MAP: Mean over AP for several queries

1. d_{17}
2. d_3
3. d_4
4. d_{10}
5. d_{14}
6. d_6
7. d_{45}
8. d_9
9. d_8
10. d_{21}
11. d_{22}
12. d_{78}
13. d_1
14. d_{33}
15. d_{11}
16. d_2
17. d_{29}
18. d_{18}
19. d_{51}
20. d_5

universität koblenz
weiter:denken

MAP – an example

- Assume two documents are missing in the result set

Document	Position	Precision
d_{17}	1	1.000
d_4	3	0.667
d_{10}	4	0.750
d_{45}	7	0.571
d_{21}	10	0.500
d_{78}	12	0.500
d_{11}	15	0.467
d_{51}	19	0.421
d_{73}	-	0
d_{39}	-	0
Average Precision		***0.488***

1. d_{17}
2. d_3
3. d_4
4. d_{10}
5. d_{14}
6. d_6
7. d_{45}
8. d_9
9. d_8
10. d_{21}
11. d_{22}
12. d_{78}
13. d_1
14. d_{33}
15. d_{11}
16. d_2
17. d_{29}
18. d_{18}
19. d_{51}
20. d_5

Further Evaluation Approaches

Indirect Measures

- User behaviour when seeking information
  - Time
  - Number of interactions
  - Viewed documents
  - Query modifications
  - Methods:
    - Clickstream mining
    - Lab tests, observation
- User surveys
  - Ask for satisfaction
  - A/B testing

Evaluation at large search engines

- Search engines have test collections of queries and hand-ranked results
- Recall is difficult to measure on the web
- Search engines often use precision at top k, e.g., k = 10
- . . . or measures that reward you more for getting rank 1 right than for getting rank 10 right.
  - NDCG (Normalized Cumulative Discounted Gain)
- Search engines also use non-relevance-based measures
  - Clickthrough on first result
    - Not very reliable if you look at a single clickthrough … but pretty reliable in the aggregate
  - Studies of user behavior in the lab
  - A/B testing

A/B testing

universität koblenz
weiter:denken

- Purpose: Test a single innovation
- Prerequisite: You have a large search engine up and running.
- Have most users use old system
- Divert a small proportion of traffic (e.g., 1%) to the new system that includes the innovation
- Evaluate with an "automatic" measure like clickthrough on first result
- Now we can directly see if the innovation does improve user happiness
- Probably the evaluation methodology that large search engines trust most
- In principle less powerful than doing a multivariate regression analysis, but easier to understand

Summary

Summary

- At the end of this lecture, you are expected to
  - understand how to evaluate an IR system
  - understand the difference between evaluation measures that ignore the ranking and those that consider the ranking

References and credits

[1]: https://olat.vcrp.de/url/RepositoryEntry/2565867182

[2]: https://videoakademie.ko-ld.de/Panopto/Pages/Sessions/List.aspx?folderID=07fcee3e-4b21-482c-90ab-ab9500ec2019

[3]: http://west.uni-koblenz.de/studying/ws1920/machine-learning-and-data-mining

[4]: https://videoakademie.ko-ld.de/Panopto/Pages/Viewer.aspx?id=545f21ba-a671-4ada-b137-ab9500f21941

[5]: Schütze, H., Manning, C. D., & Raghavan, P. (2008). Introduction to information retrieval (Vol. 39, pp. 1041-4347). Cambridge: Cambridge University Press.

[6]: https://www.gartner.com/

Credit for these slides

These slides have been adapted from

- Web IR (Zeyd Boukhers-WeST, SOSE 2020)