Practitioners' Evaluation Roundtable

Jochen L. Leidner[1,2] Ingo Frommholz[3]

[1]Coburg University of Applied Sciences and Arts

[2]University of Sheffield

[3]University of Wolverhampton

BCS Search Solutions
November 23, 2021

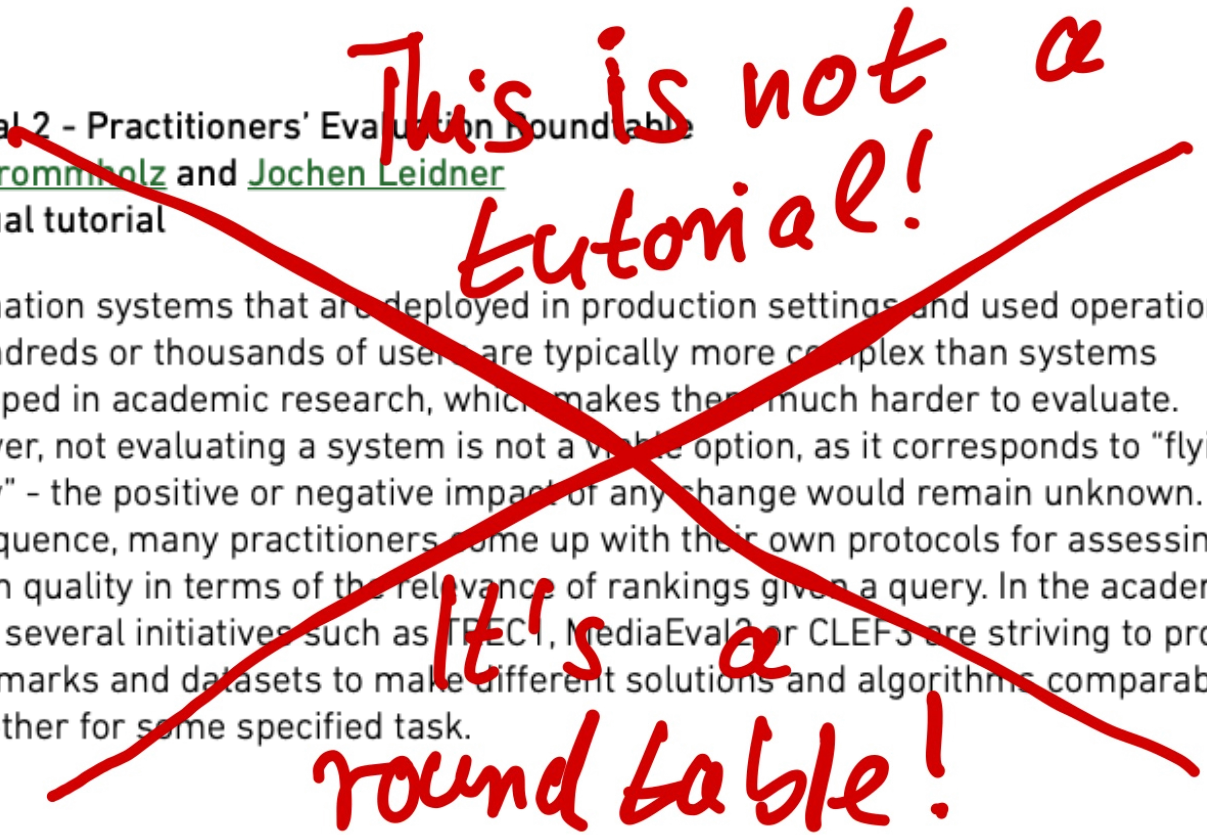
About the Organisers

Ingo Frommholz is a Reader in Data Science at the University of Wolverhampton. His research interests include interactive Information Retrieval and Digital Libraries, for example DB+IR integration, Bibliometric-enhanced IR and quantum-inspired IR models. He received his PhD in 2008 from the University of Duisburg-Essen in Germany in on the topic of probabilistic, logic-based models for annotation-based retrieval.

Jochen L Leidner FRGS is the Professor of Explainable and Responsible Artificial Intelligence at Coburg University of Applied Sciences and a Visiting Professor of Data Analytics at the University of Sheffield. His research areas include NLP, IR and applied Machine Learning. He holds an MA in computational linguistics, English language and linguistics and computer science from Friedrich-Alexander-Universität Erlangen-Nuremberg, an MPhil in computer speech, text and internet technology from the University of Cambridge and a PhD in Informatics from the University of Edinburgh.

Table of Contents

Table of Contents



Tutorial 2 - Practitioners' Evaluation Roundtable

Ingo Frommholz and Jochen Leidner

A virtual tutorial

Information systems that are deployed in production settings and used operationally by hundreds or thousands of users are typically more complex than systems developed in academic research, which makes them much harder to evaluate. However, not evaluating a system is not a viable option, as it corresponds to "flying blindly" - the positive or negative impact of any change would remain unknown. As a consequence, many practitioners come up with their own protocols for assessing system quality in terms of the relevance of rankings given a query. In the academic world, several initiatives such as TREC[1], MediaEval[2] or CLEF[3] are striving to provide benchmarks and datasets to make different solutions and algorithms comparable to each other for some specified task.

Schedule and Agenda

3.00pm	A brief introduction of IR systems evaluation – Ingo Frommholz & Jochen Leidner
3.45pm	Discussion & Lightning talks: Methods, metrics, challenges – how do practitioners evaluate their systems so far? – all participants
4.30pm	Break
4.45pm	Discussion/Breakout Groups: Evaluation in "real-world" environments — all participants
5.30pm	Discussion of results/wrap up – all participants
6.00pm	Closing

Table of Contents

Why Do We Evaluate?[1]

To answer questions such as:

- What should I do to improve the quality of my system?
- What works well, what doesn't?
- Which retrieval model give me the best results?
- Which system/search engine is better?
- Which system should I buy?
- How is 'quality' defined?
- How can I measure quality?

[1] Inspired by the lecture notes of Norbert Fuhr's Information Retrieval lecture

Evaluation Characteristics

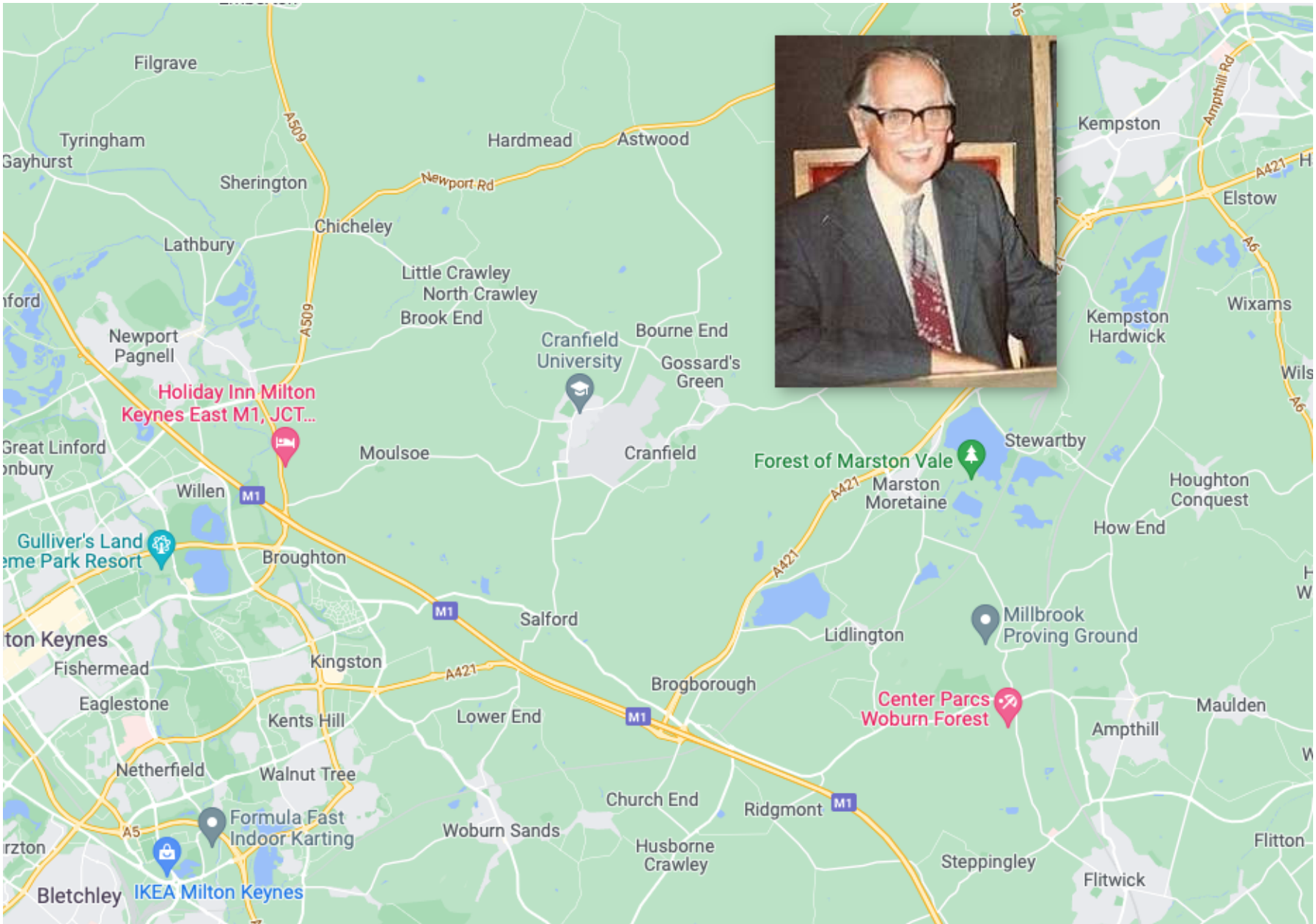
- Reliability
  - Reproducibility
  - Sufficient documentation
  - Representative datasets (documents and users)
  - Remove potential bias
  - Open source code and data (if possible); Open Science
- Validity
  - Reflect ‘real’ circumstances

Evaluation Criteria

- Efficiency
  - How quickly can a user solve a task?
  - System's response time
- Effectiveness
  - Quality of results
  - Information Retrieval deals with *vagueness* and *uncertainty*
  - Results are rarely *correct* (everything retrieved is also relevant)
  - Results are rarely *complete* (everything relevant is retrieved)
  - Focus of the *Cranfield Paradigm*
- Satisfaction of the user with the system
  - Are users happy with how the system supported their task?

Table of Contents

Cyril Cleverdon's Cranfield Studies



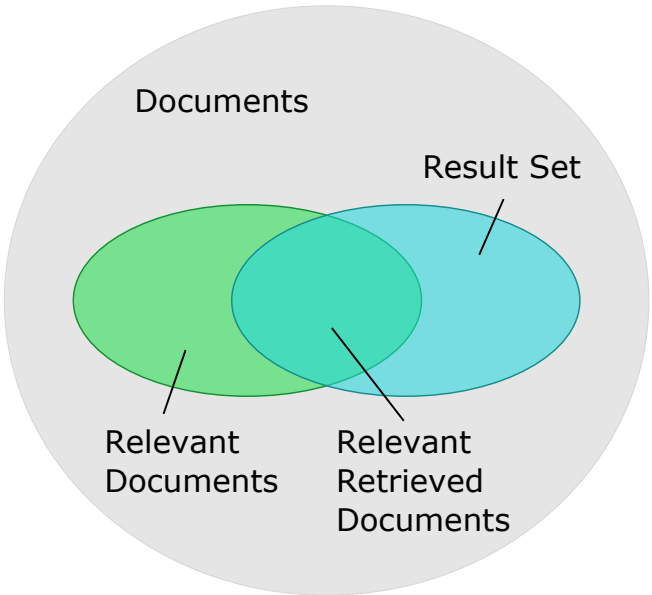
Retrieval Experiments, Cranfield Style [Har11] Information Retrieval Test Collection

- A collection of *documents*
- A set of *queries*
- A set of *relevance judgement*

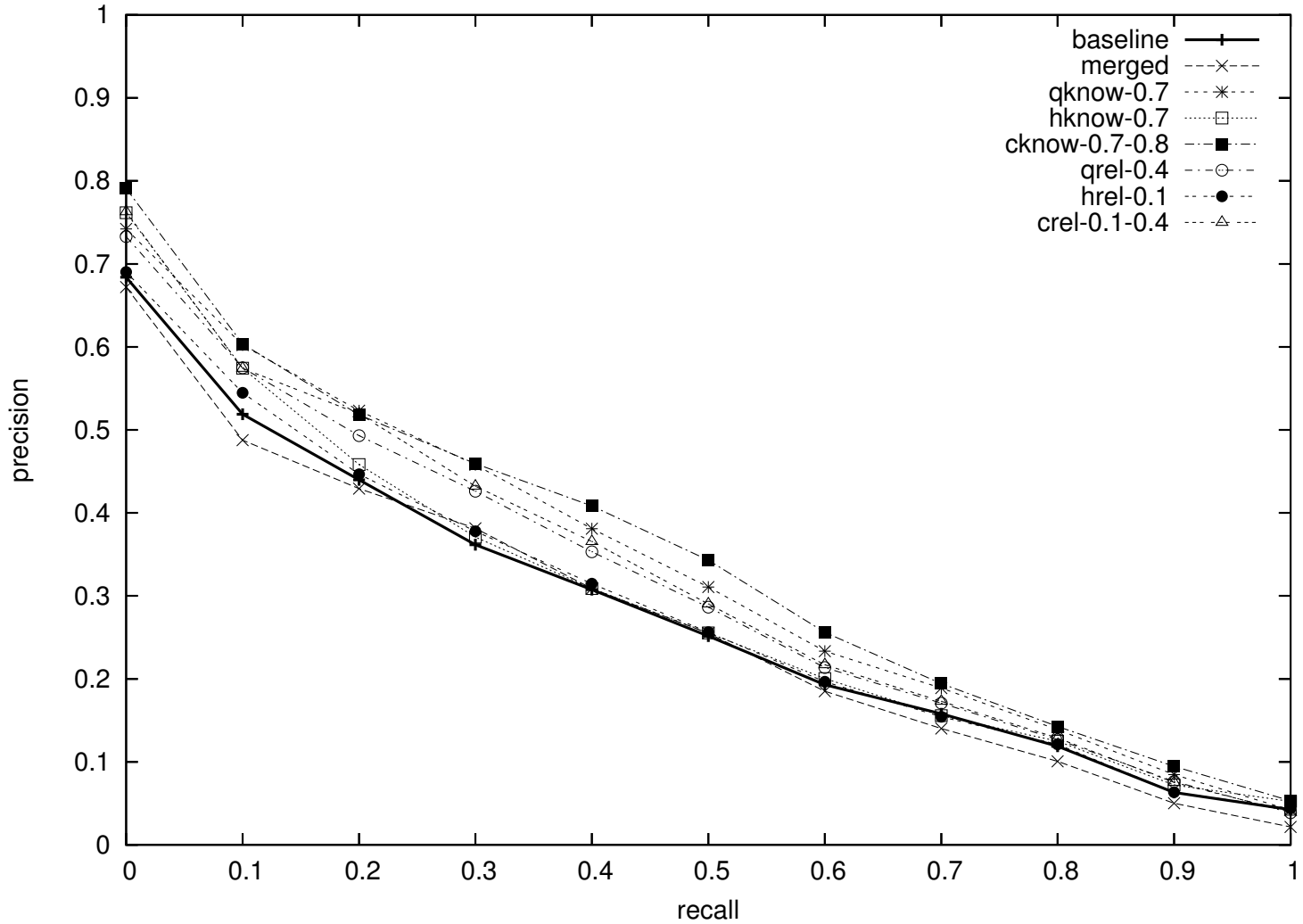
Ranking, Recall and Precision

$$\text{Precision } p = \frac{\#\text{relevant retrieved}}{\#\text{ retrieved}}$$

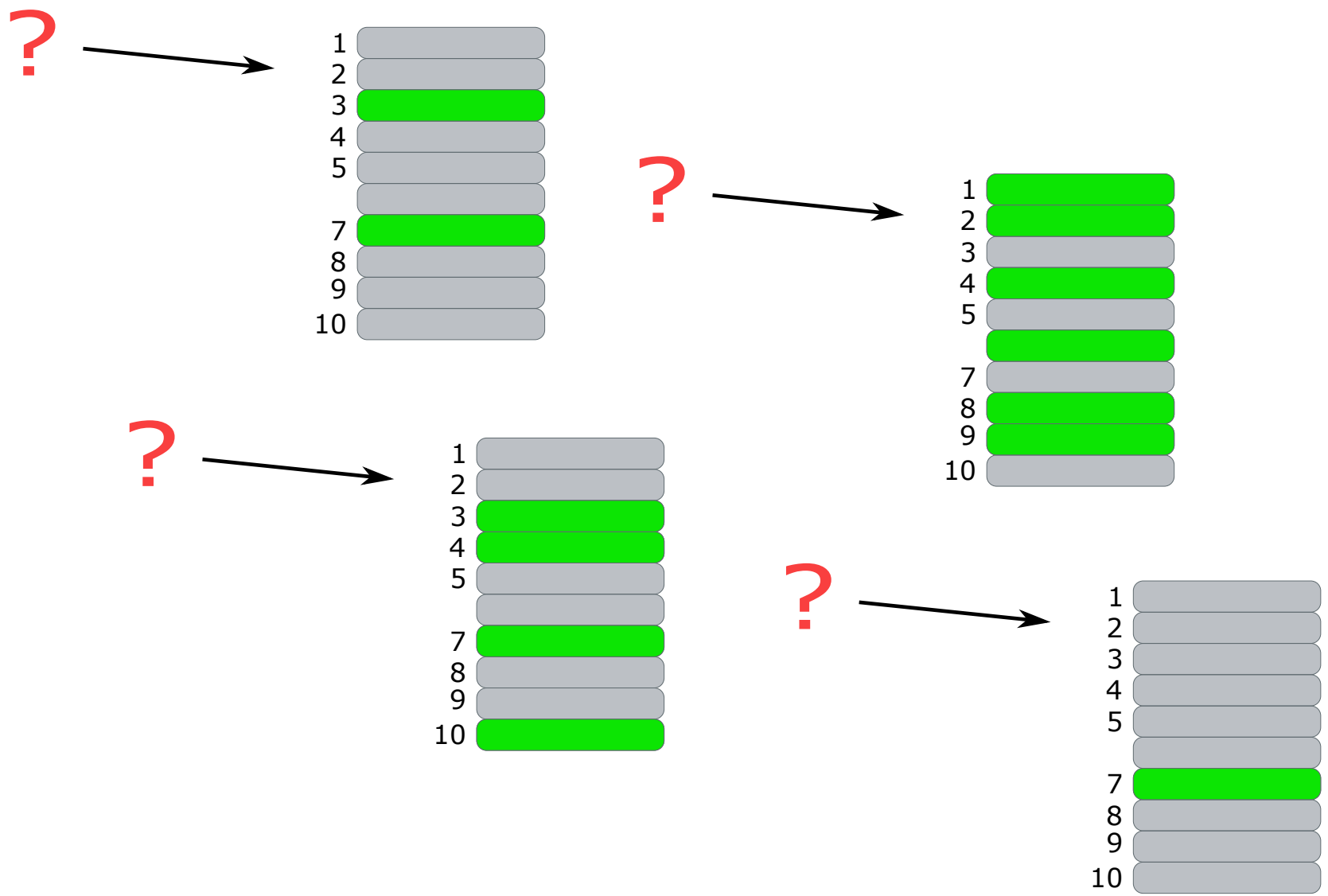
$$\text{Recall } r = \frac{\#\text{relevant retrieved}}{\#\text{ relevant}}$$



Recall–Precision Graph



Several Topics/Queries



F-measure, Mean average Precision

F_1 measure: harmonic mean of precision and recall

$$F_1 = 2 \cdot \frac{p \cdot r}{p + r}$$

General F_β score:

$$F_\beta = (1 + \beta^2) \cdot \frac{p \cdot r}{(\beta^2 \cdot p) + r}$$

Mean average precision (MAP) [BV05]:

1. Measure precision after each relevant document
2. Average over the precision values to get *average precision* for *one* topic/query
3. Average over topics/queries

System- vs user-oriented measures

- So far, these measures are used to compare entire rankings produced by different search systems.
- They are *system-oriented* measures with no assumption on user behaviour.
- Other measures are *user-oriented* as they make certain simple assumptions on user behaviour.
  - Seen n documents
  - Seen n relevant documents
  - Seen n non-relevant documents
  - Seen n non-relevant documents consecutively

Precision at n

Precision at n ($P@n$): Precision after looking at n documents

- For example, Web searchers usually look at 10 documents (first page) $\rightsquigarrow$ $P@10$

Mean Reciprocal Rank (MRR): Position of the first correct (relevant) result (e.g. question answering):
R_i ranking of query q_i
N_q number of queries
S_{corr} position of the first correct answer in the ranking

$$MRR = \frac{1}{N_q} \sum_{1}^{N_q} \frac{1}{S_{corr}(R_i)}$$

Normalized Discounted Cumulative Gain (NDCG)

- Sometimes, relevance is non-binary (e.g., 'not relevant', 'marginally relevant', 'fully relevant') $\rightsquigarrow$ graded relevance
- Highly relevant documents at the top of the ranking should be preferred over those at the end

Normalized Discounted Cumulative Gain (NDCG)

- *Gain vector* G_i for ranking R_i
- *Cumulative gain vector* CG_i

$$CG_i[j] = \begin{cases} G_i[1] & \text{if} \quad j = 1 \\ G_i[j] + CG_i[j-1] & \text{if} \quad j > 1 \end{cases}$$

- *Discounted cumulative gain vector* DCG_i: taking the position j into account (discount factor $\log_2 j$)

$$DCG_i[j] = \begin{cases} G_i[1] & \text{if} \quad j = 1 \\ \frac{G_i[j]}{\log_2 j} + DCG_i[j-1] & \text{if} \quad j > 1 \end{cases}$$

		G_i	CG_i
1	1	1	1
2	0	0	1
3	0	0	1
4	3	3	4
5	1	1	5
6	2	2	7
7	0	0	7
8	1	1	8
9	0	0	8
10	0	0	8

Normalized Discounted Cumulative Gain (NDCG)

- *Ideal gain vector*

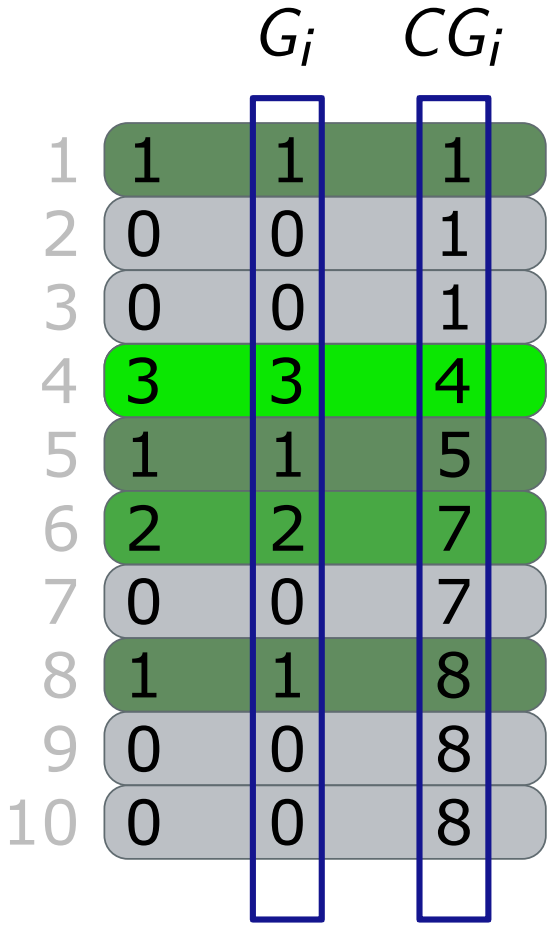
$$IG_i = (3, 2, 1, 1, 1, 0, 0, 0, 0, 0)$$

- ICG_i and $IDCG_i$ analogously
- Average over all queries to compute NDCG:

$$\overline{DCG[j]} = \frac{1}{N_q} \sum_{i=1}^{N_q} DCG_i[j]$$

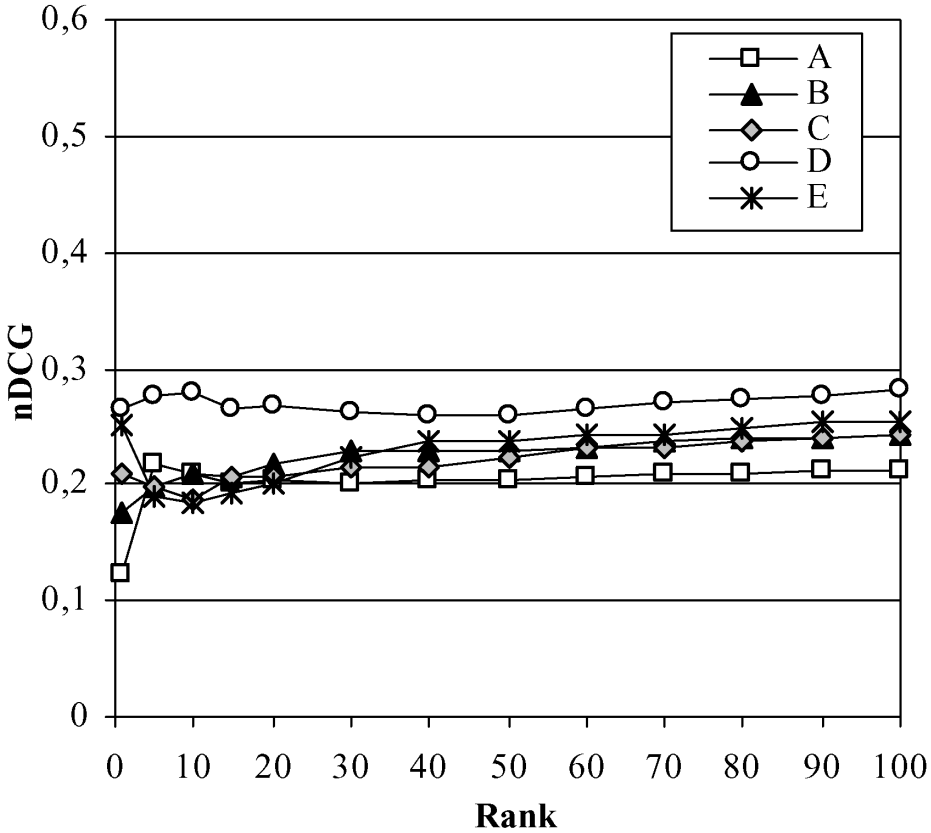
$$\overline{IDCG[j]} = \frac{1}{N_q} \sum_{i=1}^{N_q} IDCG_i[j]$$

$$NDCG[j] = \frac{\overline{DCG[j]}}{\overline{IDCG[j]}}$$



NDCG – Comparing Systems

- Compare NDCG curves
- Compare at a given position, e.g., *NDGC*[10] analogously to *P*@10



(Taken from [JK02])

Some Evaluation Metrics

- Accuracy
- **Precision (p)**
- **Recall (r)**
- Fall-out (converse of Specificity)
- **F-score** (F-measure, converse of Effectiveness) (F_β)
- **Precision at k ($P@k$)**
- R-precision (RPrec)
- **Mean average precision (MAP)**
- **Mean Reciprocal Rank (MRR)**
- **Normalized Discounted Cumulative Gain (NDCG)**
- Maximal Marginal Relevance (MMR)
- Other Metrics: bpref, GMAP, ...

Some metrics (e.g., MRR) are controversially discussed [Fuh17; Sak20; Fuh20].

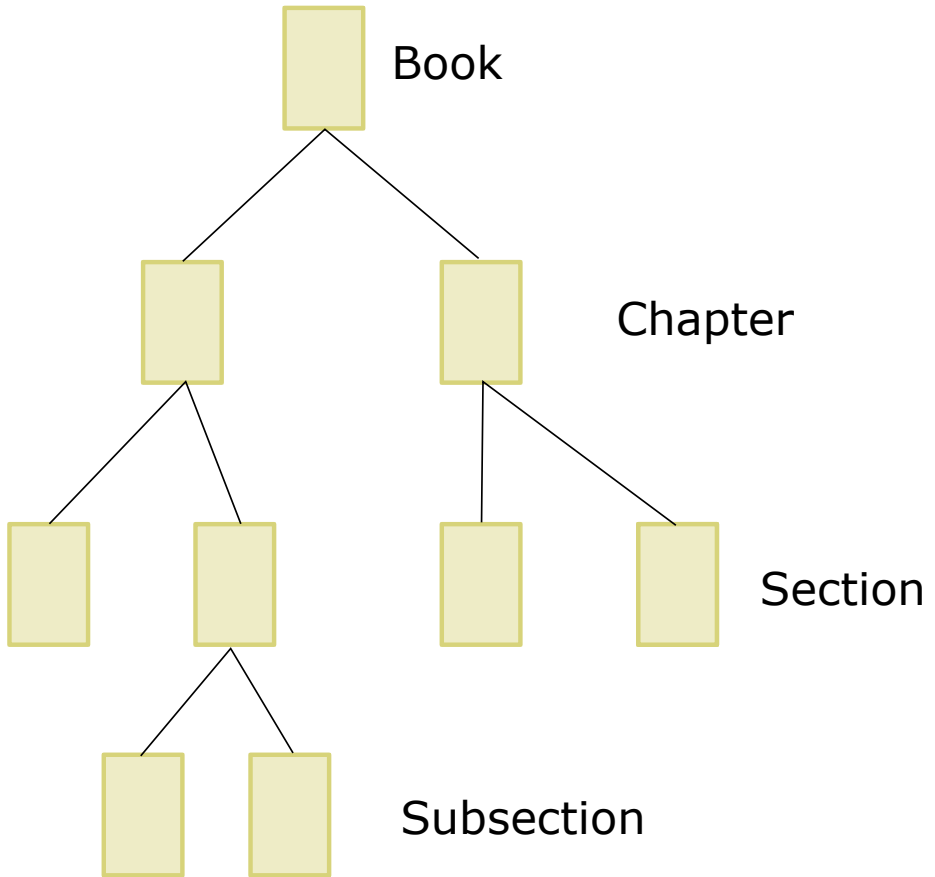
Precision- and Recall-oriented tasks

- Many task are *precision-oriented*, i.e. we seek a high precision
  - Web search
  - Mobile search
  - Question answering
- Other tasks are *recall-oriented*, i.e. we seek a high recall
  - Patent search for prior art (to check for novelty of an application)
  - Systematic reviews
  - Investigative journalism
- Real life may force you to customize the metric to use!
- E.g.: Is Precision or Recall more important for *your* task? (If equal, then use F_1, otherwise $F_{0.5}$ or F_5 may be more suitable.)

"Hacking Your Measures" – Evaluating Structured Document Retrieval

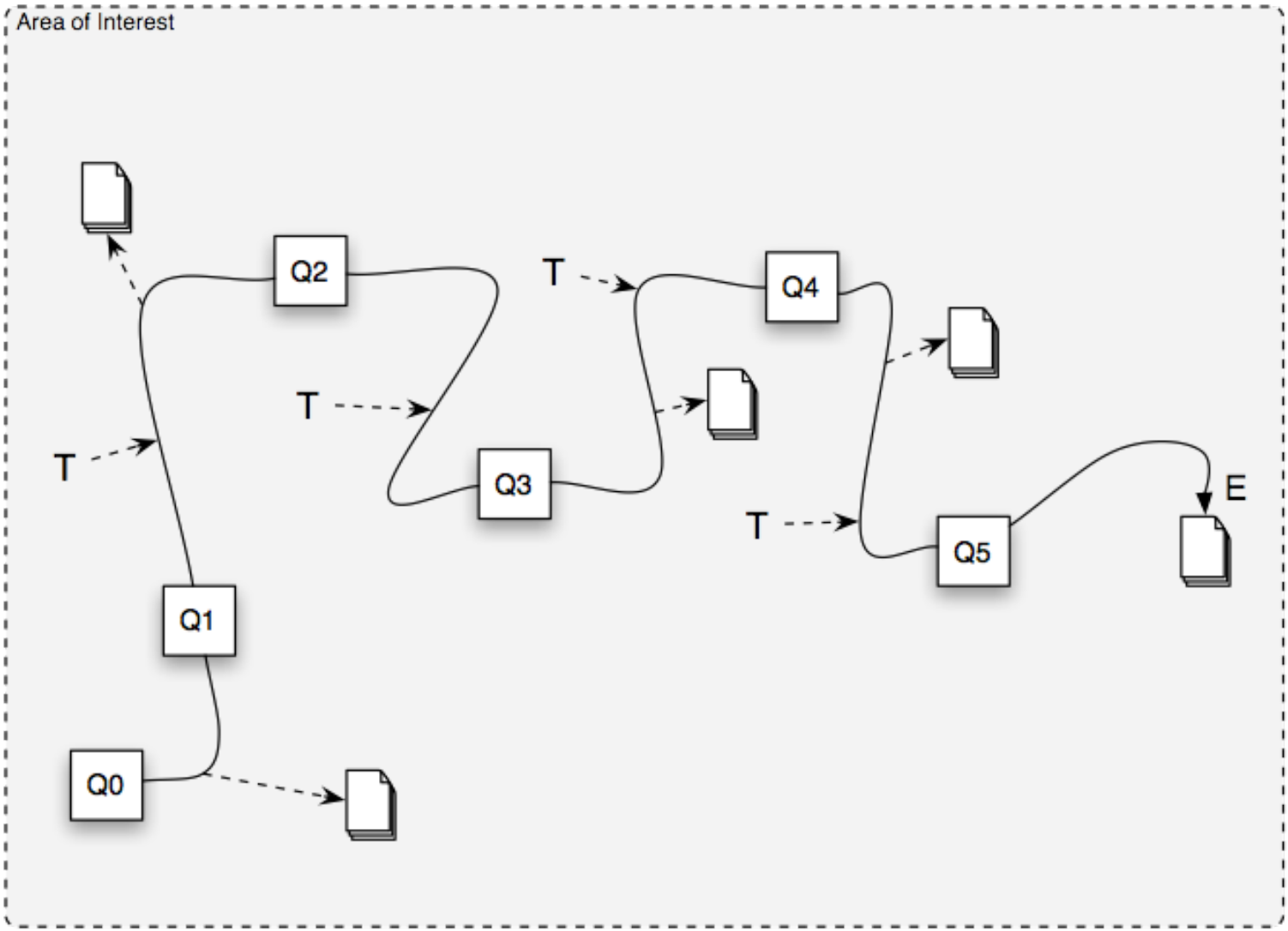
INEX – INitiative for the Evaluation of XML Retrieval [LÖ09]

- Find smallest component that is highly relevant
- INEX created long discussion threads on suitable evaluation measures
- Specificity: Extend to which a document component is focused on the information need
- Exhaustivity: Extend to which the information contained in a document component satisfies the information need



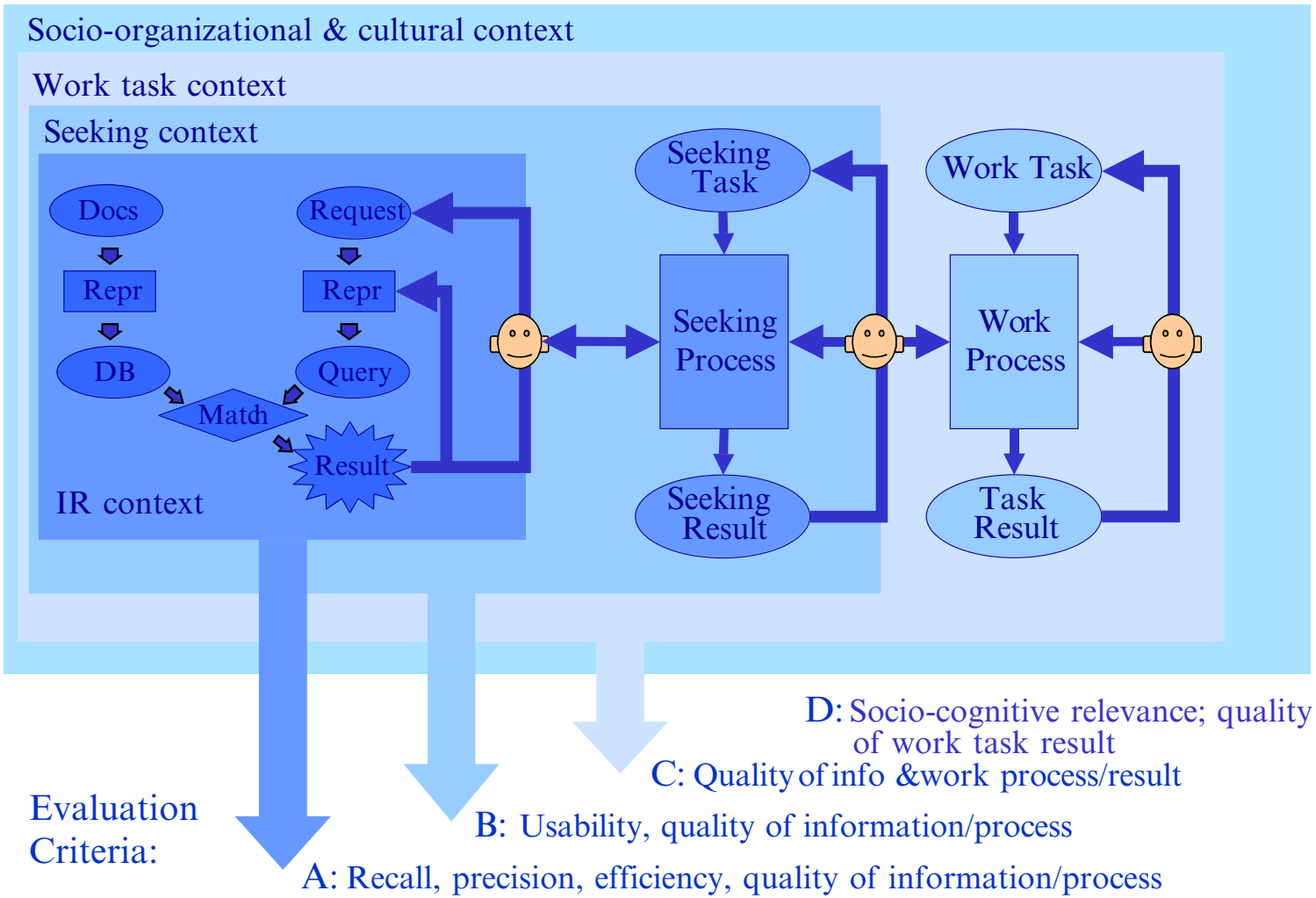
Beyond Cranfield

Search is an iterative process ("berrypicking" [Bat89])



Beyond Cranfield – Information Seeking & Searching

Task-based Information Seeking an Searching [IJ05]



Beyond Cranfield – Interactive IR evaluation

Borlund's Simulated Tasks [Bor03]

- Use of realistic scenarios
- *Simulated work tasks*:

  Simulated work task situation: *After your graduation you will be looking for a job in industry. You want information to help you focus your future job seeking. You know it pays to know the market. You would like to find some information about employment patterns in industry and what kind of qualifications employers will be looking for from future employees.*

  Indicative request: *Find, for instance, something about future employment trends in industry, i.e., areas of growth and decline.*

- Simulated work tasks should be realistic (e.g., not "imagine you're the first human on Mars...")
  - However, a less relatable task may be outweighed by a topically very interesting situation
- Taking into account situational factors (*situational relevance*)

Table of Contents

What to Evaluate?

- relevance (retrieval effectiveness) ⟵ most IR academics focus here
- coverage (e.g. percentage of actual user queries answerable)
- speed (throughput, responsiveness)
- user interface quality (UX)
- cost to build
- time to build
- cost to operate
- task completion time
- scalability (order of magnitude of number of docs.)
- memory requirements (transient, persistent)
- index freshness (how long before I can retrieve it)
- user's subjective satisfaction

Table of Contents

Common Evaluation Initiatives

- TREC [VH99; VH05]: US DARPA funded public benchmark and associated workshop series (1992-);
- NTCIR [SOK21]: Japanese language initiative (1999-)
- CLEF [FP19]: a European initiative of originally volunteer researchers interested in monolingual and cross-lingual search (2000-);

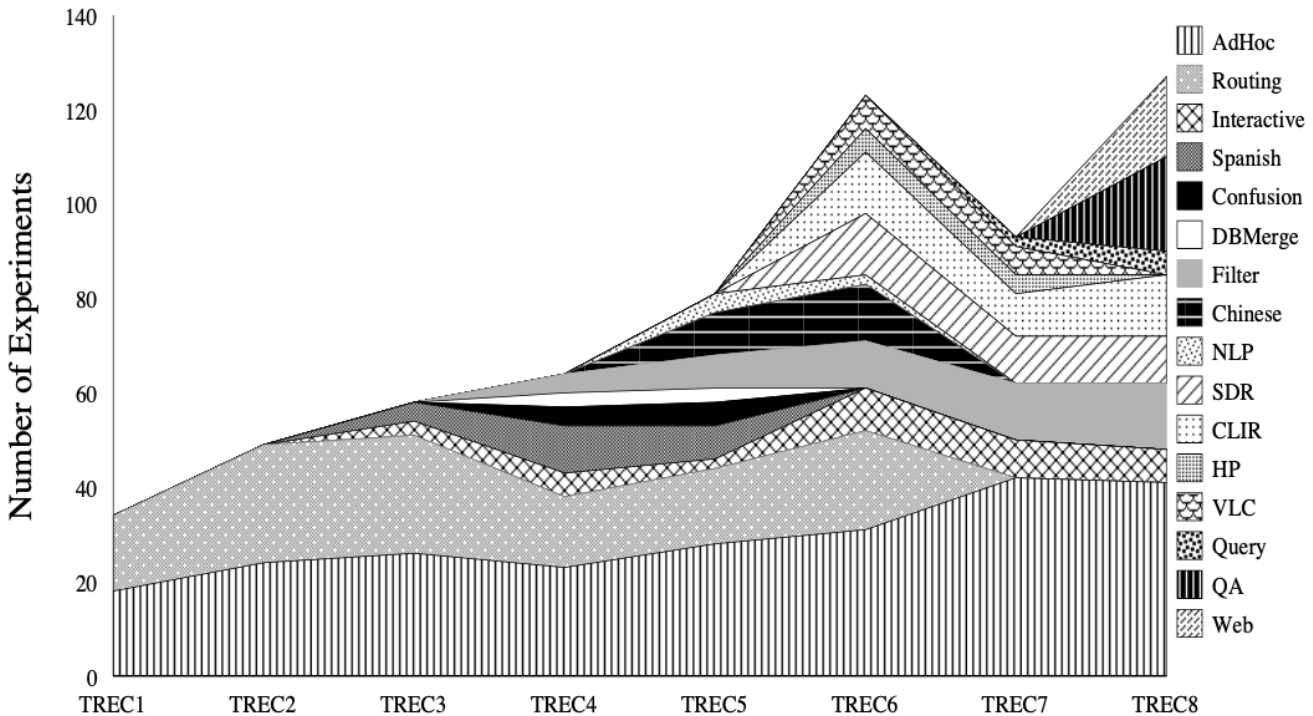


Table of Contents

Challenges to Practitioners

- Legacy: old code-base with strange/non-standard/non-effective evaluation metrics in use combined with a resistance to change;
- Knowledge gap: lack of skills/expertise/experience in the core team;
- Resources: no gold data available;
- Planning: evaluation was not budgeted for;
- Awareness: team consists of traditional managers and software engineers that do not realise quantitative evaluation is a thing;
- Infrastructure: evaluation needs to be done "in vivo" as there is no second system instance available; and
- Scaffolding: search component is deeply embedded in the overall system and cannot be run as a batch script.
- standard evaluation metrics do not assess domain peculiarities [LC14]

Table of Contents

Where is the User?

- The client, customer or user should be at the centre.
- Not a single stakeholder:
  - People are different (user diversity)
  - People search differently
  - People have different information needs (objective diversity)
  - People hold different professional roles (role diversity)
  - People differ with respect to skills & experiences (skill diversity)
- The user also cares about the **user experience (UX)**

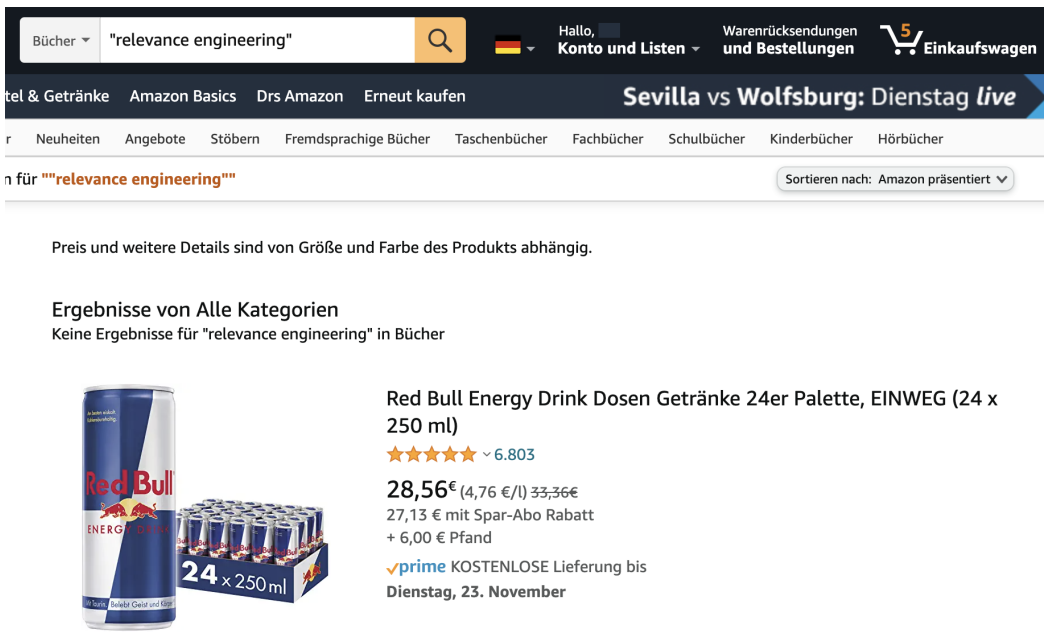
Table of Contents

What is Different in Industry?

- Effectiveness is only one of many concerns (sometimes even forgotten!)
- Functionality-oriented view (search seen as one of many functions, no or little awareness of effectiveness issues)
- Search is often (and wrongly) considered a "solved" problem
- Frequently held developer sentiment: "Just use Elastic, Solr or Lucene, and we're done."
- Developers unskilled in IR may integrate libraries with default settings inappropriate for a given use case
- Evaluation often done online (on a running system) - **A/B testing** [KL17]

Sustained Relevance as a Priority

- Where relevance is a priority, it will be sustained *after* a project to build a search function has been completed.
- It's a *process*, not a *project*!
- Role of the **relevance engineer**



(Source: Amazon (DE book search, 2021-11-21))

That's It from Us – Over to You!

- So much for introduction; now the floor is Yours!
- This is *not* a tutorial, so we hope we have reminded you of some situations in your professional lives where evaluation was hard or missing...
- ...so now please share, discuss & have fun!

(Source: Wikimedia Commons)

References and Further Reading

- Introduction to Information Retrieval - Evaluation (slides)
  https://web.stanford.edu/class/cs276/handouts/EvaluationNew-handout-6-per.pdf
- Online Controlled Experiments and A/B Testing
  https://www.researchgate.net/profile/Ron-Kohavi/publication/316116834_Online_Controlled_Experiments_and_AB_Testing
- Enterprise Search – Evaluation (Chapter 4 of [KH])
  https://www.flax.co.uk/wp-content/uploads/2017/11/ES_book_final_journal_version.pdf
- TREC Conferences https://trec.nist.gov
- CLEF Initiative http://www.clef-initiative.eu
- NTCIR Workshops
  http://research.nii.ac.jp/ntcir/data/data-en.html

Bibliography I

[Bat89] M J Bates. "The design of browsing and berrypicking techniques for the online search interface". In: *Online Review* 13.5 (1989), pp. 407–424.

[Bor03] Pia Borlund. "The IIR evaluation model: a framework for evaluation of interactive information retrieval systems". In: *Information Research* 8.3 (2003). URL: http://informationr.net/ir/8-3/paper152.html.

[BV05] Chris Buckley and Ellen Vorhees. "Retrieval System Evaluation". In: *TREC Exp. Eval. Inf. Retr.* Ed. by Ellen Vorhees and Donna Harman. MIT Press, 2005, Chapter 3.

[FP19] Nicola Ferro and Carol Peters, eds. *Information Retrieval Evaluation in a Changing World: Lessons Learned from 20 Years of CLEF*. Cham, CH: Springer-Nature, 2019.

Bibliography II

[Fuh17] Norbert Fuhr. "Some Common Mistakes In IR Evaluation, And How They Can Be Avoided". In: *SIGIR Forum* 33 (2017). URL: http://www.is.informatik.uni-duisburg.de/bib/pdf/ir/Fuhr_17b.pdf.

[Fuh20] Norbert Fuhr. "SIGIR Keynote: Proof By Experimentation? Towards Better IR Research". In: *SIGIR Forum* 54.2 (2020), pp. 1–4. URL: http://sigir.org/wp-content/uploads/2020/12/p04.pdf.

[Har11] Donna Harman. *Information Retrieval Evaluation*. Synthesis Lectures on Information Concepts, Retrieval, and Services. San Francisco, CA, USA: Morgan & Claypool, 2011. DOI: https://doi.org/10.2200/S00368ED1V01Y201105ICR019.

[IJ05] Peter Ingwersen and Kalvero Järvelin. *The turn: integration of information seeking and retrieval in context*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2005. ISBN: 140203850X.

Bibliography III

[JK02] Kalvero Järvelin and Jaana Kekäläinen. "Cumulated Gain-Based Evaluation of IR Techniques". In: *ACM Trans. Inf. Syst.* 20.4 (2002), pp. 422–446. URL: https://www.cc.gatech.edu/$%5Csim$zha/CS8803WST/dcg.pdf.

[KH] Udo Kruschwitz and Charlie Hull. *Searching the Enterprise*. Foundations and Trends in Information Retrieval. Now Publishers. ISBN: 978-1680833041. DOI: http://dx.doi.org/10.1561/1500000053.

[KL17] Ron Kohavi and Roger Longbotham. "Online Controlled Experiments and A/B Testing". In: *Encyclopedia of Machine Learning and Data Mining*. Ed. by Claude Sammut and Geoffrey I. Webb. Boston, MA, USA: Springer, 2017, pp. 922–929. DOI: 10.1007/978-1-4899-7687-1_891.

[LC14] Bo Long and Yi Chang. *Relevance Ranking for Vertical Search Engines*. Morgan Kaufmann, 2014. ISBN: 978-0124071711.

Bibliography IV

[LÖ09] Ling Liu and M. Tamer Özsu. "XML Information Retrieval". In: *Encyclopedia of Database Systems*. Ed. by Ling Liu and M. Tamer Özsu. Springer, 2009. DOI: https://doi.org/10.1007/978-0-387-39940-9_4084.

[Sak20] Tetsuya Sakai. "On Fuhr's Guideline for IR Evaluation". In: *SIGIR Forum* 54.1 (2020).

[SOK21] Tetsuya Sakai, Douglas W. Oard, and Noriko Kando, eds. *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact*. The Information Retrieval Series, 43. Heidelberg, Germany: Springer, 2021. ISBN: 978-9811555534.

[VH05] Ellen M. Voorhees and Donna K. Harman, eds. *TREC - Experiment and Evaluation in Information: Experiment and Evaluation in Information Retrieval*. Cambridge, MA, USA: MIT Press, 2005.

Bibliography V

[VH99] Ellen M. Voorhees and Donna K. Harman. “The Text REtrieval Conference (TREC): history and plans for TREC-9”. In: *SIGIR Forum* 33.2 (1999), pp. 12–15. DOI: 10.1145/344250.344252.