Beyond-Accuracy Goals, Again

Maarten de Rijke
University of Amsterdam
Amsterdam, The Netherlands
m.derijke@uva.nl

ABSTRACT

Improving the performance of information retrieval systems tends to be narrowly scoped. Often, better prediction performance is considered the only metric of improvement. As a result, work on improving information retrieval methods usually focuses on improving the methods' accuracy. Such a focus is myopic. Instead, as researchers and practitioners we should adopt a richer perspective measuring the performance of information retrieval systems. I am not the first to make this point (see, e.g., [4]), but I want to highlight dimensions that broaden the scope considered so far and offer a number of examples to illustrate what this would mean for our research agendas.

First, trustworthiness is a prerequisite for people, organizations, and societies to use AI-based, and, especially, machine learning-based systems in general, and information retrieval systems in particular. Trust can be gained in an intrinsic manner by revealing the inner workings of an AI-based system, i.e., through explainability. Or it can be gained extrinsically by showing, in a principled or empirical manner, that a system upholds verifiable guarantees. Such guarantees should obtained for the following dimensions (at a minimum): (i) accuracy, including well-defined and explained contexts of usage; (ii) reliability, including exhibiting parity with respect to sensitive attributes; (iii) repeatable and reproducible results, including audit trails; (iv) resilience to adversarial examples, distributional shifts; and (v) safety, including privacy-preserving search and recommendation.

Second, in information retrieval, our experiments are mostly conducted in controlled laboratory environments. Extrapolating this information to evaluate the real-world effects often remains a challenge. This is particularly true when measuring the impact of information retrieval systems across broader scales, both temporally and spatially. Conducting controlled experimental trials for evaluating real-world impacts of information retrieval systems can result in depicting a snapshot situation, where systems are tailored towards that specific environment. As society is constantly changing, the requirements set for information retrieval systems are changing as well, resulting in short-term and long-term feedback loops with interactions between society and information retrieval systems.

CCS CONCEPTS

• **Information systems → Information retrieval**; **Recommender systems**.

KEYWORDS

Evaluation, Measurement, Performance

ACM Reference Format:
Maarten de Rijke. 2023. Beyond-Accuracy Goals, Again. In *Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining (WSDM '23), February 27-March 3, 2023, Singapore, Singapore.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3539597.3572332

BIOGRAPHY

Maarten de Rijke is a Distinguished University Professor of Artificial Intelligence and Information Retrieval at the University of Amsterdam. He is also the scientific director and a co-founder of the national Innovation Center for Artificial Intelligence (ICAI), a large-scale public-private collaboration in the Netherlands that is focused on (i) talent development in AI, (ii) joint development and execution of shared AI research agendas by knowledge institutes, industry, governmental and societal organizations, and (iii) learning-by-doing. His research is focused on designing trustworthy technology to connect people to information, particularly search engines, recommender systems, and conversational assistants. His work targets two key questions: How can we create intrinsic trust in information retrieval systems, that is, align their reasoning process with human expectations? And how can we establish extrinsic trust in information retrieval systems, that is, establish verifiable guarantees on their behavior?

ACKNOWLEDGMENTS

The talk benefited from discussions and joint work with Gabriel Bénédict, Romain Deffayet, Philipp Hager, Maartje ter Hoeve, Ana Lucic, Clemencia Siro, and Ruqing Zhang.

This research was supported by the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organization for Scientific Research, https://hybrid-intelligence-centre.nl. All content represents the opinion of the author, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

REFERENCES

[1] Romain Deffayet, Jean-Michel Renders, and Maarten de Rijke. To appear. Evaluating the Robustness of Click Models to Policy Distributional Shift. *ACM Transactions on Information Systems* (To appear).

[2] Romain Deffayet, Thibaut Thonet, Jean-Michel Renders, and Maarten de Rijke. 2022. Offline Evaluation for Reinforcement Learning-based Recommendation: A Critical Issue and Some Alternatives. *SIGIR Forum* 56, 2 (December 2022).

[3] Thomas G. Dietterich. 2018. Robust Artificial Intelligence and Robust Human Organizations. *arXiv preprint arXiv:1811.10840* (2018).

WSDM '23, February 27-March 3, 2023, Singapore, Singapore
© 2023 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-9407-9/23/02.
https://doi.org/10.1145/3539597.3572332

[4] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. In *RecSys 2010: Proceedings of the Fourth ACM Conference on Recommender Systems*. ACM, 257–260.

[5] Philipp Hager, Maarten de Rijke, and Onno Zoeter. 2023. Contrasting Neural Click Models and Pointwise IPS Rankers. In *ECIR 2023: 45th European Conference on Information Retrieval*. Springer.

[6] Ana Lucic, Sheeraz Ahmad, Amanda Furtado Brinhosa, Q. Vera Liao, Himani Agrawal, Umang Bhatt, Krishnaram Kenthapadi, Alice Xiang, Maarten de Rijke, and Nicholas Drabowski. 2022. Towards the Use of Saliency Maps for Explaining Low-Quality Electrocardiograms to End Users. In *ICML 2022 Workshop on Interpretable ML in Healthcare*. Also in IJCAI 2022 Workshop on Explainable Artificial Intelligence (XAI).

[7] Clemencia Siro, Mohammad Aliannejadi, and Maarten de Rijke. 2022. Understanding User Satisfaction with Task-Oriented Dialogue Systems. In *SIGIR 2022: 45th international ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2018–2023.

[8] Maartje ter Hoeve, David Grangier, and Natalie Schluter. 2022. High-Resource Methodological Bias in Low-Resource Investigations. *arXiv preprint arXiv:2211.07534* (2022).

[9] Chen Wu, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. To appear. PRADA: Practical Black-Box Adversarial Attacks against Neural Ranking Models. *ACM Transactions on Information Systems* (To appear).