QUARE: 1st Workshop on Measuring the Quality of Explanations in Recommender Systems

Alessandro Piscopo
Datalab, BBC
London, United Kingdom
alessandro.piscopo@bbc.co.uk

Oana Inel
University of Zurich
Zurich, Switzerland
inel@ifi.uzh.ch

Sanne Vrijenhoek
University of Amsterdam
Amsterdam, the Netherlands
s.vrijenhoek@uva.nl

Martijn Millecamp
AE NV
Bruges, Belgium
martijn.millecamp@hotmail.com

Krisztian Balog
Google
Stavanger, Norway
krisztianb@google.com

ABSTRACT

QUARE[1]—measuring the QUality of explAnations in REcommender systems—is the first workshop that aims to promote discussion upon future research and practice directions around evaluation methodologies for explanations in recommender systems. To that end, we bring together researchers and practitioners from academia and industry to facilitate discussions about the main issues and best practices in the respective areas, identify possible synergies, and outline priorities regarding future research directions. Additionally, we want to stimulate reflections around methods to systematically and holistically assess explanation approaches, impact, and goals, at the interplay between organisational and human values. The homepage of the workshop is available at: https://sites.google.com/view/quare-2022/.

CCS CONCEPTS

• **Information systems → Recommender systems**; • **Human-centered computing → HCI design and evaluation methods**;

KEYWORDS

Recommender systems, explanation quality, explanation goals, explanation evaluation

ACM Reference Format:
Alessandro Piscopo, Oana Inel, Sanne Vrijenhoek, Martijn Millecamp, and Krisztian Balog. 2022. QUARE: 1st Workshop on Measuring the Quality of Explanations in Recommender Systems. In *Proceedings of the 45th Int'l ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '22), July 11–15, 2022, Madrid, Spain.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3477495.3531699

1 MOTIVATION AND RELEVANCE

Recommendations are ubiquitous in many contexts and domains due to a continuously growing adoption of decision-support systems. For example, recommendations are often provided to help us decide on items to buy, news items to read or watch, or even educational institutions or job positions to apply to. Explanations may be provided along with recommendations with the reasoning behind suggesting a particular item [4]. However, explanations may also significantly affect a user's decision-making process by serving a number of different goals [8], such as transparency, persuasiveness, and effectiveness, among others.

In the last few years, a growing number of papers has been published on the explainability of recommender systems, at venues such as but not limited to SIGIR [1–4, 9] and RecSys [10]. Despite a vast amount of research [11], the evaluation of recommendation explanations is still an area where significant gaps remain. For example, as of yet, there is no consensus if there exists a one-size-fits-all good explanation or how to measure its quality [7]. Furthermore, the relationship between the quality and effects of explanations has not been investigated in depth yet [1, 5].

The lack of established, actionable methodologies to evaluate explanations for recommendations and the lack of evaluation datasets, hinder cross-comparison between different explainable recommendations approaches and hamper the widespread adoption of explanations in industry settings. A public service broadcaster may want to support its audience-facing recommender systems with explanations whose main intents are to explain how the system works and ensure users have confidence in it. The same broadcaster may want to build an internal tool to allow scrutiny from its editorial team. On the other hand, engagement maximisation purposes through diverse content recommendations may drive a commercial media platform to focus more on persuasive and efficient explanations.

Conversely, end-users of a recommender system may be bearers of different values, and explanations can affect them differently [6]. For instance, if users value transparency and trust and expect these from an organisation, they may be put off by explanations that primarily aim to persuade them to consume more content. In general, however, different organisational values may require a different combination of explanation goals; also, within the same organisation, some combinations of goals may be more appropriate for some use cases or some user groups and less for others. Therefore, understanding whether explanations are fit for their intended goals and users is key to subsequently implementing them in a production stage.

[1] In Latin, *why, from what cause.*

SIGIR '22, July 11–15, 2022, Madrid, Spain.

ACM ISBN 978-1-4503-8732-3/22/07.
https://doi.org/10.1145/3477495.3531699

This workshop aims to extend existing work in the field by bringing together and facilitating the exchange of perspectives and solutions from industry and academia and bridging the gap between academic design guidelines and the best practices in the industry regarding the implementation and evaluation of explanations in recommender systems, with respect to their goals, impact, potential biases, and informativeness. With this workshop, we provide a platform for discussion among scholars, practitioners, and other interested parties.

2 THEMES AND TOPICS

The motivation of the workshop is to promote discussion upon future research and practice directions of evaluating explainable recommendations, by bringing together academic and industry researchers and practitioners in the area. We focus, in particular, on real-world use cases, diverse organizational values and purposes, and different target users. Therefore, we encourage submissions that study different explanation goals and combinations of those, how they fit various organization values, and different use cases and target users. Furthermore, we welcome submissions that propose and make available for the community high-quality datasets and benchmarks.

- Evaluation
  - Relevance of explanation goals for different use cases;
  - Soliciting user feedback on explanations;
  - Implicit vs. explicit evaluation of explanations and goals;
  - Reproducible and replicable evaluation methodologies;
  - Online vs. offline evaluations.
- Personalisation
  - User-modelling for explanation generation;
  - Evaluation approaches for personalised explanations (e.g., content, style);
  - Evaluation approaches for context-aware explanations (e.g., place, time, alone/group setting, exploratory/transaction mode).
- Presentation
  - Explanation of different explanation modalities (e.g., text, graphics, audio, hybrid);
  - Evaluation of interactive explanations.
- Datasets
  - Generation of datasets for evaluation of explanations;
  - Evaluation benchmarks.
- Values
  - Evaluation of explanations in relation to organisational values;
  - Evaluation of explanations in relation to personal values.

3 KEY OUTCOMES

The overarching outcome of QUARE is to create a venue for researchers and practitioners to discuss, disseminate, and advance research and practice of measuring the quality of explanations in recommender systems, by:

- Fostering connections between practitioners and researchers to create a fertile ground for empirical studies that address real-world use cases.
- Bringing together practitioners and researchers to discuss and outline best practices regarding the evaluation of goals, impact, potential biases, and informativeness of explanations for recommender systems.
- Bringing together different viewpoints on evaluating explanations in recommender systems to identify and list future research directions.
- Summarising the workshop discussions and contributions in a technical report co-authored by organisers and participants, to be submitted to SIGIR Forum.

4 WORKSHOP FORMAT

The workshop is a half-day event composed of paper presentation sessions, keynotes, invited and featured talks, and interactive and engaging sessions to foster discussion and potential collaborations among workshop participants. In addition, we plan to have a final discussion session to summarise the workshop's outcome and look forward.

We welcome the following types of contributions:

- *Position or perspective papers* (2-4 pages): original ideas, perspectives, research vision, and open challenges in the area of evaluation approaches for explainable recommender systems;
- *Demonstration papers* (2-4 pages): original or already published prototypes and operational evaluation approaches in the area of explainable recommender systems;
- *Featured talks*: either already published papers or papers summarising existing publications in leading conferences and high-impact journals that are relevant to the topic of the workshop. With these contributions, we would like to ensure that we attract interested audiences and foster high-quality discussions.

The submissions will be evaluated and selected based on their relevance to the workshop, innovation, and research potential. Finally, we plan to publish the accepted contributions on the website of our workshop: https://sites.google.com/view/quare-2022/.

4.1 Proposed Workshop Activities

Before the workshop day:

- *Authors' talks*: We plan to ask authors to record a short presentation of their contribution (3 - 5 minutes) to be published on the workshop's website.
- *Dilemmas of explainable recommendations*: We will ask prospective workshop participants to think about dilemmas in the area of explanations and recommender systems. These dilemmas will be further discussed during the workshop.

The half-day workshop will consist of several sessions:

- *Welcome and Introduction*: A brief welcome and overview of planned activities.
- *Invited Talk*: An invited talk from an industry practitioner or academic researcher in the area of recommender systems and explanations, to set out the most pressing issues, challenges, and opportunities related to the topic and outline the main points of the discussion that will follow.

- *Lighting Talks*: The authors of position, perspective, demonstration, and featured contributions will give a lightning talk (3-5 minutes); the lighting talks are intended to stimulate the interaction in the discussion sessions that will be the core session of the day.
- *Discussion Session*: Drawing inspiration from the contributions received and the dilemmas proposed by participants, the workshop organizers and attendees will discuss these pressing matters during the workshop. Depending on the number of participants, the discussion will either take place in a plenary fashion or in breakout rooms/groups. The notes taken during these sessions will be made available online.
- *Final Discussion*: A plenary session with all workshop attendees, summarising the main points discussed in the groups.
- *Wrap*: Concluding thoughts and future plans.

After the workshop day:

- *Technical report*: Based on the submissions received and the discussions during the workshop, we plan to write a technical report, soliciting comments and contributions from the workshop participants. The report is to be submitted to SIGIR Forum.

5 ORGANISING COMMITTEE

Alessandro Piscopo is a Principal Data Scientist at the BBC. His team, Datalab, focuses on developing recommendation engines across the organisation, and has so far deployed live recommenders on products such as BBC Sounds, BBC World Service, and the BBC News app. His previous research is situated at the intersection of peer-production communities, collaborative knowledge engineering, and data quality. He received his PhD from the University of Southampton in 2019. Website: https://www.linkedin.com/in/alessandro-piscopo-ds.

Oana Inel is a Postdoctoral Researcher at the University of Zurich. Currently, Oana is investigates the use of explanations to provide transparency for decision-support systems and foster reflective thinking in people. Previously, she was a Postdoctoral Researcher at TU Delft. She did her PhD at the Vrije Universiteit Amsterdam, where her research focused on detecting and representing events and their semantics for understanding knowledge on the Web. She has co-organised several workshops and tutorials in the area of explanations, human computation, and semantic web at TheWebConf, UMAP, ISWC. Website: https://oana-inel.github.io.

Sanne Vrijenhoek is a Project Researcher with a background in Artificial Intelligence at the University of Amsterdam's Institute of Information Law. She works in an interdisciplinary project on assessing diversity in news recommendations. An important part of this project is translating normative notions of diversity into concrete concepts that can be used to inform recommender system design, and has as such extensive experience in bridging the gap between computer science and the social sciences. She has organised a number of non-scientific workshops, facilitating discussions between computer scientists and news editors. Website:https://www.uva.nl/profiel/v/r/s.vrijenhoek/s.vrijenhoek.html.

Martijn Millecamp is a UX engineer at AE NV, Belgium. As a consultant, he guides companies to develop user-centered explanations for several machine learning algorithms and recommender systems. He obtained a PhD at the Augment group at the KU Leuven investigating how explanations for a music recommender system should be personalized to different personal characteristics and how learning analytics dashboards should be adapted to fit a different context. During his PhD, he has organised a number of both scientific and non-scientific workshops, mostly facilitating discussions between computer scientists, teachers and students. Website: https://www.linkedin.com/in/martijn-millecamp/.

Krisztian Balog is a Staff Research Scientist at Google and a Full Professor at the University of Stavanger. At Google, he works on models and evaluation methodology for explaining user models and items in conversational recommender systems. He has co-organised numerous workshops on evaluation at SIGIR and CIKM, as well as large-scale benchmarking efforts at TREC and CLEF. He served as the general co-chair of ICTIR'20, program co-chair of the CIKM'21 short paper track, and general co-chair of ECIR'22. Website: https://krisztianbalog.com.

6 RELATED WORKSHOPS

The Explainable User Models and Personalized Systems workshop (ExUM),[2] collocated with UMAP in 2019–2022, studies the role of explanations as a means to provide users with control in personalised and adaptive systems. The Workshop on ExplainAble Recommendation and Search (EARS)[3] collocated with SIGIR in 2018–2020 addressed explainability issues in recommender systems and search. However, we bring several novel aspects in our **QUARE** workshop: (1) focusing on an open challenge in the field, namely the evaluation of explanation quality; (2) close interactions and collaboration between academic and industry research; (3) touching upon multi-disciplinary aspects of evaluating explanations for recommender systems—besides technical aspects, we are also interested in integrating human and organisational values; and (4) a setup that fosters creativity, lays down future research directions, and potentially provides actionable points to bridge the gap between academic design guidelines and the best practices in the industry regarding the implementation and evaluation of explanations in recommender systems.

REFERENCES

[1] Krisztian Balog and Filip Radlinski. 2020. Measuring Recommendation Explanation Quality: The Conflicting Goals of Explanations. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. ACM, 329–338.

[2] Krisztian Balog, Filip Radlinski, and Shushan Arakelyan. 2019. Transparent, Scrutable and Explainable User Models for Personalized Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '19)*. ACM, 265–274.

[3] Zuohui Fu, Yikun Xian, Ruoyuan Gao, Jieyu Zhao, Qiaoying Huang, Yingqiang Ge, Shuyuan Xu, Shijie Geng, Chirag Shah, Yongfeng Zhang, and Gerard de Melo. 2020. Fairness-Aware Explainable Recommendation over Knowledge Graphs. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '20)*. ACM, 69–78.

[4] Deepesh V. Hada, Vijaikumar M, and Shirish K. Shevade. 2021. ReXPlug: Explainable Recommendation using Plug-and-Play Language Model. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. ACM, 81–91.

[5] Chen He, Denis Parra, and Katrien Verbert. 2016. Interactive Recommender Systems: A survey of the State of the Art and Future Research Challenges and Opportunities. *Expert Syst. Appl.* 56 (2016), 9–27.

[2] http://www.di.uniba.it/~swap/exum/index.html
[3] https://ears2020.github.io

[6] Martijn Millecamp, Cristina Conati, and Katrien Verbert. 2022. "Knowing Me, Knowing You": Personalized Explanations for a Music Recommender System. *User Model. User-adapt. Interact.* 32 (2022), 215–252.
[7] Ingrid Nunes and Dietmar Jannach. 2017. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Model. User-adapt. Interact.* 27, 3-5 (2017), 393–444.
[8] Nava Tintarev and Judith Masthoff. 2015. Explaining Recommendations: Design and Evaluation. In *Recommender Systems Handbook*. Springer, 353–382.
[9] Khanh Hiep Tran, Azin Ghazimatin, and Rishiraj Saha Roy. 2021. Counterfactual Explanations for Neural Recommenders. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '21)*. ACM, 1627–1631.
[10] Yikun Xian, Tong Zhao, Jin Li, Jim Chan, Andrey Kan, Jun Ma, Xin Luna Dong, Christos Faloutsos, George Karypis, S. Muthukrishnan, and Yongfeng Zhang. 2021. EX3: Explainable Attribute-aware Item-set Recommendations. In *Fifteenth ACM Conference on Recommender Systems (RecSys '21)*. ACM, 484–494.
[11] Yongfeng Zhang and Xu Chen. 2020. Explainable Recommendation: A Survey and New Perspectives. *Found. Trends Inf. Retr.* 14, 1 (2020), 1–101.