Report on the 1st Workshop on Measuring the Quality of Explanations in Recommender Systems (QUARE 2022) at SIGIR 2022

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

Explainable recommenders are systems that explain why an item is recommended, in addition to suggesting relevant items to the users of the system. Although explanations are known to be able to significantly affect a user's decision-making process, significant gaps remain concerning methodologies to evaluate them. This hinders cross-comparison between explainable recommendation approaches and is one of the issues hampering the widespread adoption of explanations in industry settings. The goal of QUARE '22 was to promote discussion upon future research and practice directions around evaluation methodologies for explanations in recommender systems. To that end, we brought together researchers and practitioners from academia and industry in a half-day event, co-located with SIGIR 2022. The workshop's program included two keynote talks, three sessions of technical paper presentations in the form of lightning talks followed by panel discussions, and a final plenary discussion session. Although the area of explanations for recommender systems is still in its early stages, QUARE saw the participation of researchers and practitioners from several fields, laying the groundwork for the creation of a community around this topic and indicating promising directions for future research and development.

Date: 15 July, 2022.

Website: https://sites.google.com/view/quare-2022/home.

1 Introduction

Recommenders are widely used to support users in finding relevant items in contexts where a large number of options may be available. Recommendations may help users decide which items

to buy, which news items to read or watch, or even to which educational institutions or job positions to apply. Alongside recommendations, a system may provide explanations of why an item or a set of items was suggested [Hada et al., 2021]. We are aware that these explanations may affect a user's decision-making process in a range of different ways [Tintarev and Masthoff, 2015]. Nevertheless, the evaluation of recommendation explanations—which would allow us to measure and understand the effects of explanations—is still an area where significant gaps remain. For example, there is no consensus if one-size-fits-all explanations exist or how to measure their quality [Nunes and Jannach, 2017]. This lack of established, actionable methodologies to evaluate explanations for recommendations, as well as evaluation datasets, hinders cross-comparison between different explainable recommendation approaches and is one of the issues hampering their widespread adoption in industry settings.

The main aim of this workshop was to bring together researchers and practitioners from industry and academia, facilitate the exchange of perspectives and solutions, and bridge the gap between academic design guidelines and the best practices in the industry regarding the implementation and evaluation of explanations in recommender systems. QUARE '22 was a half-day workshop, co-located with SIGIR '22, which welcomed previously published (featured papers) and ongoing research, vision, and position papers around the following themes:

- Design and generation of explanations;
- Evaluation of explanations;
- Personalisation;
- Presentation.

We had a total of 14 accepted submissions, with 12 featured papers, one demo paper, and one position paper; these submissions are summarised below. Authors of each successful contribution were invited to give a lightning talk, which served as a starting point for subsequent discussions among authors and participants. In addition, the workshop featured two keynotes from researchers and practitioners in industry and academia.

2 Keynotes

2.1 *Evaluating Explainability in Recommendations: Current Practices and Opportunities* by Xiting Wang

Xiting Wang (Microsoft Research Asia) was the first keynote speaker at the workshop. In her talk, she critically presented different evaluation methods for explanations in recommender systems, by focusing on their limitations and challenges, and reflected on how to choose the most appropriate one. While explanations can make users feel recommendations more "human-like" and can increase clicks even up to 40%, their effects are often underestimated by practitioners, who, instead, typically invest more efforts into building a new recommender that may improve accuracy in a much smaller percentage. The cases of successful deployment of explainable recommendations are still limited in real-world products, even though they are already adopted across several commercial websites, such as Facebook, Amazon, and Last.FM. Furthermore, as of yet, the understanding of goals and advantages of explanations is still patchy and a benchmark for

objective, cross-comparable evaluation is not available. In terms of challenges, the diversity of goals, the lack of well-defined foundations, and the fact that explanations are personalised by nature further hinder their evaluation. Before delving into explanation evaluation approaches, Xiting distinguished two types of explainable recommender methods, *self-explainable models* and *post-hoc explanations*, and compared them in terms of *recommendation accuracy*, *fidelity*, and *readability and usefulness*. Self-explainable models aim to ensure good accuracy while keeping the additional benefits of explainability. Post-hoc explanation approaches primarily aim to ensure that the explanations faithfully reveal the working mechanisms of the original model. When it comes to gauging the performance of self-explainable models, recommendation accuracy might not reach optimal levels and should be evaluated first, whereas, in post-hoc approaches, the fidelity of explanations should take priority. Self-explainable models may have lower accuracy than black box models, but may naturally be more robust to noise, due to human-understandable constraints [Lee et al., 2022]. Hence, it is worth using, besides common accuracy metrics (e.g., AUC, NDCG, precision, recall), methods to evaluate robustness (i.e., verify whether recommendations remain accurate in the presence of increased noise or adversarial attacks).

The keynote continued by presenting four types of evaluation methods for assessing the readability and usability of explainable recommenders: case study, automatic quantitative experiment, user study, and online experiment. Case studies are able to provide people with an intuitive feeling of how explanations look like and the speaker strongly advised everyone working on empirical research on explanations to perform at least an evaluation with this method [Chen et al., 2021] (i.e., show a few lines from actually generated explanations [Chen et al., 2021] or compare them with explanations generated by an existing approach [Chen et al., 2019]). Quantitative approaches may be used to supplement case studies. In these cases, metrics such as BLEU or ROUGE may be used to compute how close explanations are to, for instance, user reviews, which can be treated as ground-truth explanations. Alternatively, automatic heuristic criteria may be created to quantify readability and usefulness. Wang et al. [2018] compare sentiment consistency, coherency, and conciseness between different approaches, whereas Liu et al. [2021] look at informativeness, explanation quality, and efficiency of a news explainable recommender. Overall, quantitative approaches are easy to design and cheap to evaluate, but finding suitable datasets and heuristics may be difficult and cumbersome. User studies may add substantial value to the evaluation of an explanation approach, even with a relatively small number of participants [Yang et al., 2022], since they allow to evaluate any motivation or goal with a pool of real users. However, the additional time and budget required by user studies may be a concern and should be defined early in any project. The fourth, and most difficult evaluation method is the online experiment, which requires deploying a model and carrying out an experiment in a live product [Wang et al., 2021]. Nevertheless, online experiments are able to include much larger pools of users and measure directly the effect of explanations, e.g., in terms of impressions, clicks, etc.

In her closing statements, Xiting stressed the importance of finding common evaluation criteria, as opposed to focusing on the optimisation of current explanation approaches, to move beyond the present early stages of explainability research. She also pointed out that there is no fixed answer to evaluate explainability and that the priority of evaluation design should be consistent with the motivation behind a new approach or a piece of research. Finally, she advocated for the creation of explanation benchmarks and interdisciplinary studies, drawing from psychology, cognitive science, natural language processing, and computer vision.

2.2 *Explaining Recommendations to Users: Simple or Complex?* by Dietmar Jannach

The second keynote of the workshop was delivered by Dietmar Jannach (University of Klagenfurt, Austria). Recommendations may be explained for different purposes (transparency, scrutability, trust, effectiveness, persuasiveness, efficiency, and satisfaction [Tintarev and Masthoff, 2015]) and along various dimensions (type of information to show, form, personalisation, availability, etc. [Nunes and Jannach, 2017]). Explanations may be *complex*, such as the following example taken from the Pandora system by Tintarev and Masthoff [2022]: "*Based on what you've told us so far, we're playing this track because it features solo strings, mystical qualities, minor key tonality, melodic songwriting and intricate melodic phrasing.*" Here, information is presented about assumed item features and (assumed) user preferences in a text-based and rather detailed form; the explanation is personalised and is available upon request. Alternatively, a *simple* explanation could be "*Everybody likes Ed Sheeran*" or "*Our recommendations are based on the trendiness of the artist.*" The question then arises: which of these is better and according to which purpose? To answer this question, Dietmar first reviewed early work (2000–2014) on comparing different explanation styles for movie [Herlocker et al., 2000; Gedikli et al., 2014] and book recommendations [Bilgic and Mooney, 2005]. Common to these studies is that they investigate the effects of explanations with respect to one or multiple explanation purposes based on user studies, often using nearest-neighbour techniques as the baseline recommendation method.

Next, Dietmar reviewed more recent work (2015–present), which is often driven by the desire for transparency due to advances in deep learning methods [Chen et al., 2022]. Algorithmic studies are commonly accompanied by case studies on explainability/interpretability, but these are not designed for, nor evaluated with, end users (i.e., no treatment and control groups with human judges who are representative of target users). Focusing on natural language explanations, user reviews are often considered as ground truth and compared against the generated explanations using machine translation measures (BLEU and ROUGE). There are several issues with this approach. First, it assumes a single ground truth, while in reality there may be many good explanations. Also, it ignores important aspects, such as personalisation and grammar. More generally, it is unclear to what extent these metrics correspond to human perceptions (cf. [Liu et al., 2016]) and whether these types of explanations serve any of the potential explanation purposes.

In the final part of his talk, Dietmar discussed conversational recommender systems, which offer a natural test-bed for explanations. However, most of today's systems cannot answer 'why' questions and explanations are also missing from dialogue datasets. He emphasised the need for more studies on end-user perceptions and considering the questions of user control: explanations can serve as an entry point to let users give feedback or correct the system. Finally, explanations to a large extent are a problem of human-computer interaction (HCI)—showing explanations is only the beginning, there are several exciting future questions that require collaboration between data science and HCI.

3 Session 1: Explanation generation approaches

The first technical session of the QUARE workshop focused on the topic *explanation generation approaches* and consisted of five *featured* contributions.

3.1 Papers

Pairwise Review-Based Explanations for Voice Product Search (featured in Penha et al. [2022])

by ***Gustavo Penha***, *Eyal Krikon and Vanessa Murdock*

Gustavo Penha presented two methods for generating explanations for the voice product search domain, which is very challenging due to the limited amount of information that can be disclosed at a time. The work was done in collaboration with Alexa, Amazon. The first explanation generation method proposed was called pointwise and consisted of selecting the top-1 most helpful sentence from the top relevant review of the recommended product. The pairwise explanation generation method consisted of selecting a diverse pair of products from the top-k ranked products for a query and choosing a sentence from a helpful review for each product to be included in the explanation. In a user study evaluation, it was found that both pointwise and pairwise explanations help users make a good decision (i.e., effective) and explain how the system works (i.e., transparency), but they neither convince users to buy the products (i.e., persuasiveness) nor allow them to tell the system is wrong (i.e., scrutability).

ELIXIR: Learning from User Feedback on Explanations to Improve Recommender Models (featured in Ghazimatin et al. [2021])

by ***Azin Ghazimatin***, *Soumajit Pramanik, Rishiraj Saha Roy and Gerhard Weikum*

In the second talk of the session, Azin Ghazimatin talked about the role of explanations in enhancing the quality of recommender models and proposed ELIXIR as a method for pairing recommendations with explanations and using these pairs to better learn user preferences. In short, ELIXIR extends existing recommender models by requesting feedback on pairs of recommendations and explanations and then it incorporates the feedback in the recommender models to improve future recommendations and better learn user-specific preferences. The explanations provided are at the level of the item, minimal, and counterfactual. Through user studies in the movie and book recommendation domains, the authors showed that this pair-level feedback improves recommendations and it is more discriminative than item-level feedback.

Counterfactual Explanations for Neural Recommenders (featured in Tran et al. [2021])

by *Khanh Hiep Tran, Azin Ghazimatin and* ***Rishiraj Saha Roy***

Counterfactual explanations show changes in sets of recommended items, as they would appear if parts of a user history were removed or modified. Rishiraj Saha Roy presented ACCENT, a

general framework to find counterfactual explanations for neural recommenders. ACCENT adapts prior counterfactual approaches, by combining some of their features—notably the contribution scores for each item in a user history obtained using Personalised PageRank (PRINCE) and the approximate influence scoring mechanism from the fast influence analysis (FIA)—and expanding them with a pairwise comparison between the recommended items scores and their possible replacements. The results show that ACCENT can be used to produce effective explanations for end users, based only on their previous interaction history, and can be applied to a large number of neural recommenders.

Automatic Generation of Natural Language Explanations (featured in Costa et al. [2018])

by ***Aonghus Lawlor***, *Sixun Ouyang, Peter Dolog and Felipe Costa*

Several explanation generation approaches rely on user reviews as ground truth. However, this data is sparse and noisy, i.e., may contain irrelevant or hard to interpret parts for a machine. The approach presented by Aonghus Lawlor aims to generate readable reviews, by focusing only on the important features within the text. It consists of a generative model which encodes the input text and attributes into high-dimensional embeddings, aligns those, and outputs personalised reviews for each user-item pair. The authors compare the explanations generated by their model with human-made reviews, computing a sentiment value for each word and highlighting high- and low-polarity words. The explanations generated by this approach perform on par with human reviews in terms of accuracy and readability, prompting the authors to suggest further testing of the model in larger and more varied review domains.

Exploring the Role of Local and Global Explanations in Recommender Systems (featured in Radensky et al. [2022])

by ***Marissa Radensky***, *Doug Downey, Kyle Lo, Zoran Popović and Daniel Weld*

In the final paper of the session, Marissa Radensky explored the role of local and global explanations in the domain of research paper recommender systems. The work is motivated by the fact that there has been little investigation regarding the purposes and benefits to the users for using local and global explanations, even though it is known that explanations improve transparency. The proposed local explanations (i.e., why a particular item is being recommended) show the top relevant terms for retrieving a particular research paper together with their global weights. The proposed global explanations (i.e., how items are recommended overall) show the recommendation feed's top terms and their weights. In exploratory and controlled user studies, the authors found that the combination of the two types of explanations is more helpful than either of the two alone for helping people understand how the recommender can improve. However, the global explanations alone are the most efficient in helping people identify false positive and false negative recommendations.

3.2 Discussion

The discussion following the first technical session of the workshop focused on the interplay between explanation generation approaches and human perception of good or useful explanations. Overall, all panelists agreed that in order to design a good explanation for a particular recommender system, we need to reason and reflect on the goals that the explanation should serve, on the user needs with regard to what they expect or prefer to see in the explanations. According to Azin Ghazimatin, explanations should always provide an amount of information that is manageable by users. Furthermore, when dealing with counterfactual explanations for neural recommendations, Rishiraj Saha Roy pointed out that it is difficult to evaluate the effects or quality of the explanations at each iteration, and thus it is customary to have post-hoc evaluations and focus on the conciseness and scrutability of the explanations, as well as on the actionability or level of control that users can have through explanations to influence further recommendations. In the discussion, it was also pointed out, by Aonghus Lawlor, that many times explanations are designed based on the needs of the algorithm, or aspects that appear natural from the domain of the systems, e.g., review alike explanations or keyword-based explanations. However, it is unfortunate that human evaluations or inputs are not yet a mainstream requirement.

Marissa Radensky also mentioned that explanations are highly related to algorithmic transparency, which in turn can help people become more literate in algorithmic decision-making. In addition, when dealing with explanations for recommender systems that are used in production, from an industry perspective, persuasiveness is always an important goal, but it is imperative to also build user trust with the system, according to Gustavo Penha.

4 Session 2: Evaluation approaches for explanations

The second technical session of the QUARE workshop focused on the topic *evaluation approaches for explanations* and consisted of one *perspective* paper, one *demo* paper, and two *featured* papers.

4.1 Papers

Evaluating the Explanation Tags of Online Food Delivery Recommendation: A Position Paper (perspective)

by ***Yurou Zhao****, Ruidong Han, Fei Jiang, Lu Guan, Xiang Li, Wei Lin, Yiding Sun and Jiaxin Mao*

Yurou Zhao presented a collaboration study with the food ordering system Meituan. Though Meituan has different types of tags as explanations for their food recommendations, the space available in its interface is limited, and as such only the most accurate tags can be shown to the users. While previous work focused on the *persuasiveness* of a tag, this work proposes to also consider *usefulness* and *faithfulness* as measures of explanation quality. To measure the usefulness of an explanation CTR (Click-Through Rate) and CVR (ConVersion Rate) are combined, and for faithfulness, a proxy model is built that takes the explanation tags as input and outputs a food recommendation. A tag is then considered faithful to the degree to which it contributed a (correct) recommendation. As a limitation of this approach, the authors note that users may

perceive tags differently, and that in the future, user studies should be conducted to measure the gap between the user's objective evaluation result and the computed score.

Investigating Users' Preferences for Explanation Tags on Online Delivery Platform (demo)

by ***Yiding Sun***, *Yurou Zhao, Ruidong Han, Fei Jiang, Lu Guan, Xiang Li, Wei Lin and Jiaxin Mao*

In the second presentation of the session, Yiding Sun presented another collaboration study with the food delivery platform Meituan, in which they conducted a user study on users' personal preferences for explanation tags. Given that Meituan utilises a large number of fine-grained types, they first conducted a pre-study where they asked participants to first review and summarise the tags and then created a taxonomy. This yielded five high-level tag categories. They then tested the overall performance of the tags ("will having this type of tag help you make more accurate decisions?"), and the user's personal preferences for types of tags in light of different explanation goals (effectiveness, efficiency, trust, transparency, and satisfaction). They found that all types of explanations yielded positive results and performed better with statistical significance compared to having no explanations. In future work, they aim to conduct a follow-up study with a larger group of participants.

Post-Processing Recommender Systems with Knowledge Graphs for Recency, Popularity, and Diversity of Explanations (featured in Balloccu et al. [2022])

by ***Giacomo Balloccu***, *Ludovico Boratto, Gianni Fenu and Mirko Marras*

In the first lightning talk of the session, Giacomo Balloccu introduced knowledge graphs for generating explanations. These knowledge graphs leverage "paths" between items, such as movies sharing actors, to make recommendations based on items a user has interacted with before. They introduce three new explanation quality metrics: recency (time since linking interaction took place), popularity (of the linking entity), and diversity (of the type of explanations in the top-k recommendations). Based on these metrics they propose a post-processing approach. They found that while optimising the recommendations for these metrics greatly benefited the actual metrics, they had a minor influence on recommendation utility compared to traditional policy-guided path reasoning (PGPR).

Argumentative explanations for interactive recommendations (featured in Rago et al. [2021])

by ***Antonio Rago***, *Oana Cocarascu, Christos Bechlivanidis, David Lagnado and Francesca Toni*

During the last talk of the session, Antonio Rago described their work on argumentative explanations for interactive recommendations. In this work they focused on the transparency and satisfaction of the explanations, simplifying the predicted ratings of the system to the point that faithfulness and scrutability were guaranteed by design. They also posited that when it comes

to satisfaction, different users would prefer different styles of explanations. They created an argumentative 'scaffolding', which allowed for different types of explanations to be tested. While performance was below state-of-the-art, reasonable accuracy was achieved. In order to test the user's satisfaction with the system, a study with restricted feedback and predetermined actions was set up. In terms of transparency, it was found that all styles of explanation improved how well users understood the system significantly. However different users preferred different styles, with a few users having no preference, leading to different levels of satisfaction.

4.2 Discussion

The discussion following the second technical session first focused on the tradeoff between explanation utility and explainability. Giacomo Balloccu noted that there were specific patterns observed, with explainability consistently increasing and NDCG decreasing, but also that this effect was stronger in some instances than in others. Conducting an online study would be a good next step to gain more insight into these patterns. Next, the presenters were asked whether they thought their work could be useful to different domains and contexts. All presenters agreed that there were certain parts of their work that were very application-specific, but that some of the underlying frameworks could be translated to other domains. For Yurou Zhao, this would be the approach of using a tag as an explanation, whereas for Yiding Sun, this would be the framework of evaluating the user's preferences. Giacomo Balloccu noted that the main limitation lied in the need to build and populate a knowledge graph every time, which could be difficult for domains in which collecting the required data is not as straightforward. Lastly, Antonio Rago noted that their domain was uniquely low-risk, which made it less problematic when there were few earlier interactions to work with. This would likely be different in high-risk domains, such as medicine or finance.

5 Session 3: Effects of explanations

The third technical session of the QUARE workshop focused on the topic *effects of explanations* and consisted of five *featured* multi-disciplinary contributions.

5.1 Papers

A Survey on Effects of Adding Explanations to Recommender Systems (featured in Vultureanu-Albişi and Bădică [2022])

by ***Alexandra Vultureanu-Albisi*** *and Costin Badica*

Alexandra Vultureanu-Albisi presented the main findings of a survey paper investigating the implications of explainability applied to recommender systems. The work was motivated by two factors regarding the evaluation of explanations for recommender systems: insufficient human-centred evaluation and lack of conceptualisation of explainable AI terminology with regard to recommender systems. The main contribution of the survey is a process for evaluating the explanations of a recommender system using human-centred metrics such as user mental models,

performance of human-machine tasks, user satisfaction with the explanation, user trust, and confidence. The proposed process is guided by various effects that influence how people perceive the explanations and properties that should be achieved in order for the explanation to be properly accepted and interpreted.

Why trust technology? Combining legal and individuals' perspectives to enable trust in news personalisation (featured in Drunen et al. [2022])

by ***Max van Drunen****, Brahim Zarouali and Natali Helberger*

In the second featured paper, Max van Drunen stated that trust has two-fold importance in news recommender systems: from a recommender system perspective, it expresses how individuals decide whether they can rely on it for recommendations, while from a recommender system explanation perspective, it allows individuals to determine whether the recommender is trustworthy. To promote trust in news recommender systems and help users decide when they could rely on a news recommender system, Drunen et al. [2022] took a legal perspective. The goal was to understand what kind of trust should be promoted and which measures are found important by individuals with regard to trust in news recommender systems. The main outcome of the study was that individuals want and care about control and transparency when interacting with news recommender systems regardless of whether they already trust the media. Furthermore, while control seems to be more important than transparency for trust in news recommender systems, neither of them is sufficient—trust is not only an individual's concern, but it also affects media companies, policymakers, and the government, among others.

The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images (featured in Dominguez et al. [2019])

by ***Denis Parra****, Vicente Domínguez, Pablo Messina and Ivania Donoso-Guzmán*

Denis Parra presented an approach for recommending and providing explanations for artistic images in collaboration with UGallery, a platform focusing on promoting artists. In their work, in an online user study, Dominguez et al. [2019] studied the interplay between accuracy and explainability in three different recommendation interfaces. More precisely, they studied the effect of explaining recommendations of images using two algorithms—one with high predictive accuracy but unexplainable features (DNN) and one with lower accuracy but high potential for explanations (AVF) in three different UI conditions: no explanations (baseline), textual explanations and top-3 similar images (explainable recommendation), and features' bar chart and top-1 similar image (explainable and transparent recommendation). The study showed that explanations have, overall, a positive effect on users in terms of usefulness. In addition, the perception of recommendation accuracy improves just by adding explanations, but only for the DNN algorithm—it is perceived better than AVF in most dimensions, showing that there are interaction effects between the algorithm and the explainable interfaces. In conclusion, user interfaces and recommender algorithms should not be studied in isolation, since users need both good algorithmic accuracy and reasonable explanations, which are highly dependent on how they are presented.

Nudging towards news diversity: A theoretical framework for facilitating diverse news consumption through recommender design. (featured in Mattis et al. [2022])

by ***Nicolas Mattis***, *Philipp Masur, Judith Moeller and Wouter van Atteveldt*

Nicolas Mattis introduced the concept of nudging users towards reading and consuming more diverse news recommendations, with the aim of benefiting democracy. As highlighted by the paper, there are two different approaches for motivating diversity nudges. The first approach consists of developing a news recommender system that optimises for diversity. In time, after interacting and being exposed to such a system, users might be more likely to consume diverse recommendations. The second approach involves the presentation aspects of diverse recommendations. This second approach is motivated by the fact that people have various selection criteria (i.e., interests, attitudes) for news and the way news is presented can positively impact their interaction and consumption of diverse news. Thus, Mattis et al. [2022] proposed social norm interventions, which consist of labels assigned to news articles that help make diversity values salient and thus encourage people to consume more diverse news.

Towards Explainable AI: Assessing the Usefulness and Impact of Added Explainability Features in Legal Document Summarization (featured in Norkute et al. [2021])

by ***Milda Norkute***, *Nadja Herger, Leszek Michalak, Andrew Mulder and Sally Gao*

Milda Norkute talked about the effects of explainability in the context of legal document summarisation, at Thomson Reuters Labs. More precisely, they proposed a legal text summarisation system to support the editorial team in monitoring and collecting court cases and automatically writing a short summary of these documents. To support editors in verifying the suitability of these AI-generated summaries, Norkute et al. [2021] proposed two approaches to accommodate the requirements of the editorial team and explain the summary: 1) attention highlights (shown as different shades of a colour) consisting of a model-specific approach that uses the attention scores produced by the model to show which parts of the documents the model paid attention to when generating the summary and 2) source highlights (shown in a single colour) consisting of a model agnostic approach that identifies sentences in the input texts that had an influence on the summary. These explanation approaches were compared to a baseline approach (i.e., no explanation given) in a user study with the editorial team. The main finding of the study was that the editors were significantly faster when seeing attention highlights, due to their presentation characteristics—different shades of blue indicated different importance values as well as the certainty that the AI inspected the entire case. However, this was not the case for the source highlights, where the presentation was less intuitive and less trustworthy (a single shade of blue).

5.2 Discussion

The discussion that followed the third technical session touched upon three important aspects that we summarise below.

First of all, Nicolas Mattis, Max van Drunen, and Alexandra Vultureanu-Albisi mentioned that theoretical frameworks for defining, designing, and presenting explanations offer the foundation

for studying explanations, their quality, and ultimately their effects. However, there is still a disconnect between the theoretical and technical aspects of generating explanations. During the workshop, much has been argued that in order to properly support users with quality-driven recommendations and explanations, we need multi-disciplinary approaches. Second, Milda Norkute and Denis Parra mentioned that the effects of explanations should still be studied with actual system users in order to provide the best functionality and also personalisation strategy. Third, while there seems to be some tension between two explanation goals discussed during the session—nudging into consuming more diverse content and providing control, Max van Drunen indicated that these goals are not mutually exclusive. More precisely, they are two approaches that allow users to make different choices. Furthermore, explanations or nudges can be designed in such a way as to actually provide more control to users.

6 Final Discussion

The workshop concluded with a final discussion among all organisers, authors, and participants. It became clear early on that "*explanations*" is an overloaded term and agreeing on its definition is not trivial. Instead of aiming for one clear-cut definition, a more useful approach would be to characterise what explanation are able to enable. On the one hand, there are several organisations that are committed to bringing more transparency to their AI systems, and explanations are a supporting mechanism for that. On the other hand, explanations can provide more control to end-users, to steer the recommendations they receive. There is also a trade-off between the two that needs to be balanced, with additional legal requirements (e.g., privacy) to be considered. Overall, the questions of when explanations are needed and what purpose they should serve seem to be highly application specific.

A recurring theme was the question of evaluation. Inevitably, there is a gap between academic studies and real-world applications. Currently, user studies are the primary evaluation instrument, but it is unclear whether findings translate to real-world settings. For example, users' decision-making processes might be very different between low-cost and high-cost scenarios; it is doubtful whether study participants were to behave the same way in reality as in the imaginary scenario they are presented with. Also, even though it sounds obvious that user study participants need to be representative of the general population, in practice it is often not ensured or validated. Adding explanations to recommendations might change how users behave, and there is currently little understanding of what these changes may look like in the short and long term—this is difficult to capture in user studies.

During the discussions, it became obvious that investigating explanations in recommender systems is a multi-faceted research problem that requires insights and expertise from other disciplines beyond IR and RecSys, including those of human-computer interaction and social sciences. Throughout the workshop and in particular during the final, plenary discussion session, there was plenty of thoughtful engagement on the part of participants and a desire for continued discussion was expressed.

7 Summary

QUARE '22 aimed at bringing together researchers and practitioners to identify outstanding issues, priorities, and next steps in research and developments within the explainable recommendation field. Around 30 on-site and remote participants contributed to the success of the workshop, sharing the results and perspectives deriving from their previous research and practice. The two keynote speakers looked at the topic of the workshop from different angles, which helped shape and stimulate the subsequent discussion.

Overall, explanations for recommender systems are gaining traction as a way to improve the effectiveness of recommendations and achieve better transparency. However, the lack of shared evaluation methods and benchmarks makes the comparison between explanation approaches harder and prevent organisations from adopting them more widely. One of the possible solutions, which came up at different times during the workshop, is the establishment of cross-disciplinary research and practices, able to fill the gap between the theoretical and technical aspects of generating and evaluating explanations for recommenders. QUARE represents an important step in that direction, having connected people across academia and industry from different disciplines, such as IR, HCI, Social Sciences, and Law. We hope to build upon these connections to create a community around the topic of explanations for recommender systems in future editions of the workshop.

Acknowledgements

We thank all speakers and workshop participants for their contributions. We would also like to thank our program committee members: Alexandra Vultureanu-Albisi (University of Craiova), Oscar Alvarado (Katholieke Universiteit Leuven), Vito Walter Anelli (Politecnico di Bari), Costin Badica (University of Craiova), Ludovico Boratto (University of Cagliari), Robin Burke (University of Colorado (Boulder), Xu Chen (Renmin University of China), Li Chen (Hong Kong Baptist University), Eelco Herder (Radboud University Nijmegen), Dietmar Jannach (University of Klagenfurt), Styliani Kleanthous (Open University of Cyprus and Research Centre on Interactive Media (Smart Systems and Emerging Technologies), Bart Knijnenburg (Clemson University), Benedikt Loepp (University of Duisburg-Essen), Cataldo Musto (Dipartimento di Informatica - University of Bari), Marco Polignano (Università degli Studi di Bari Aldo Moro), Polina Proutskova (BBC), Arash Rahnama (Amazon), Zhaochun Ren (Shandong University), Cagatay Turkay (University of Warwick), Markus Zanker (Free University of Bozen-Bolzano and University of Klagenfurt), Yongfeng Zhang (Rutgers University).

References

Giacomo Balloccu, Ludovico Boratto, Gianni Fenu, and Mirko Marras. Post processing recommender systems with knowledge graphs for recency, popularity, and diversity of explanations. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 646–656, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3532041. URL https://doi.org/10.1145/3477495.3532041.

Mustafa Bilgic and Raymond Mooney. Explaining recommendations: Satisfaction vs. promotion. In *Proceedings of Beyond Personalization 2005: A Workshop on the Next Stage of Recommender Systems Research at the 2005 International Conference on Intelligent User Interfaces*, 2005.

Xu Chen, Yongfeng Zhang, and Ji-Rong Wen. Measuring "why" in recommender systems: A comprehensive survey on the evaluation of explainable recommendation. *arXiv*, cs.IR/2202.06466, 2022.

Zhongxia Chen, Xiting Wang, Xing Xie, Tong Wu, Guoqing Bu, Yining Wang, and Enhong Chen. Co-attentive multi-task learning for explainable recommendation. In Sarit Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 2137–2143. ijcai.org, 2019. doi: 10.24963/ijcai.2019/296. URL https://doi.org/10.24963/ijcai.2019/296.

Zhongxia Chen, Xiting Wang, Xing Xie, Mehul Parsana, Akshay Soni, Xiang Ao, and Enhong Chen. Towards explainable conversational recommendation. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2994–3000, 2021.

Felipe Costa, Sixun Ouyang, Peter Dolog, and Aonghus Lawlor. Automatic generation of natural language explanations. In *Proceedings of the 23rd international conference on intelligent user interfaces companion*, pages 1–2, 2018.

Vicente Dominguez, Pablo Messina, Ivania Donoso-Guzmán, and Denis Parra. The effect of explanations and algorithmic accuracy on visual recommender systems of artistic images. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*, pages 408–416, 2019.

Max van Drunen, Brahim Zarouali, and Natali Helberger. Recommenders you can rely on: A legal and empirical perspective on the transparency and control individuals require to trust news personalisation. *JIPITEC*, 13, 2022.

Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. How should I explain? A comparison of different explanation types for recommender systems. *Int. J. Hum. Comput. Stud.*, 72(4):367–382, 2014.

Azin Ghazimatin, Soumajit Pramanik, Rishiraj Saha Roy, and Gerhard Weikum. Elixir: learning from user feedback on explanations to improve recommender models. In *Proceedings of the Web Conference 2021*, pages 3850–3860, 2021.

Deepesh V. Hada, Vijaikumar M, and Shirish K. Shevade. 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, pages 81–91. ACM, 2021.

Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, CSCW '00, pages 241–250, 2000.

SeungEon Lee, Xiting Wang, Sungwon Han, Xiaoyuan Yi, Xing Xie, and Meeyoung Cha. Self-explaining deep models with logic rule reasoning. *CoRR*, abs/2210.07024, 2022. doi: 10.48550/arXiv.2210.07024. URL https://doi.org/10.48550/arXiv.2210.07024.

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, EMNLP '16, pages 2122–2132, November 2016.

Danyang Liu, Jianxun Lian, Zheng Liu, Xiting Wang, Guangzhong Sun, and Xing Xie. Reinforced anchor knowledge graph generation for news recommendation reasoning. In Feida Zhu, Beng Chin Ooi, and Chunyan Miao, editors, *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 1055–1065. ACM, 2021. doi: 10.1145/3447548.3467315. URL https://doi.org/10.1145/3447548.3467315.

Nicolas Mattis, Philipp Masur, Judith Möller, and Wouter van Atteveldt. Nudging towards news diversity: A theoretical framework for facilitating diverse news consumption through recommender design. *new media & society*, page 14614448221104413, 2022.

Milda Norkute, Nadja Herger, Leszek Michalak, Andrew Mulder, and Sally Gao. Towards explainable ai: Assessing the usefulness and impact of added explainability features in legal document summarization. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021.

Ingrid Nunes and Dietmar Jannach. A systematic review and taxonomy of explanations in decision support and recommender systems. *User Model. User-adapt. Interact.*, 27(3-5):393–444, 2017.

Gustavo Penha, Eyal Krikon, and Vanessa Murdock. Pairwise review-based explanations for voice product search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 300–304, 2022.

Marissa Radensky, Doug Downey, Kyle Lo, Zoran Popovic, and Daniel S Weld. Exploring the role of local and global explanations in recommender systems. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pages 1–7, 2022.

Antonio Rago, Oana Cocarascu, Christos Bechlivanidis, David Lagnado, and Francesca Toni. Argumentative explanations for interactive recommendations. *Artificial Intelligence*, 296:103506, 2021.

Nava Tintarev and Judith Masthoff. Explaining recommendations: Design and evaluation. In *Recommender Systems Handbook*, pages 353–382. Springer, 2015.

Nava Tintarev and Judith Masthoff. Beyond explaining single item recommendations. In *Recommender Systems Handbook*, pages 711–756. Springer, 2022.

Khanh Hiep Tran, Azin Ghazimatin, and Rishiraj Saha Roy. Counterfactual explanations for neural recommenders. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1627–1631, 2021.

Alexandra Vultureanu-Albişi and Costin Bădică. A survey on effects of adding explanations to recommender systems. *Concurrency and Computation: Practice and Experience*, page e6834, 2022.

Xiting Wang, Yiru Chen, Jie Yang, Le Wu, Zhengtao Wu, and Xing Xie. A reinforcement learning framework for explainable recommendation. In *2018 IEEE international conference on data mining (ICDM)*, pages 587–596. IEEE, 2018.

Xiting Wang, Xinwei Gu, Jie Cao, Zihua Zhao, Yulan Yan, Bhuvan Middha, and Xing Xie. Reinforcing pretrained models for generating attractive text advertisements. In Feida Zhu, Beng Chin Ooi, and Chunyan Miao, editors, *KDD '21: The 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Virtual Event, Singapore, August 14-18, 2021*, pages 3697–3707. ACM, 2021. doi: 10.1145/3447548.3467105. URL https://doi.org/10.1145/3447548.3467105.

Ruichao Yang, Xiting Wang, Yiqiao Jin, Chaozhuo Li, Jianxun Lian, and Xing Xie. Reinforcement subgraph reasoning for fake news detection. In Aidong Zhang and Huzefa Rangwala, editors, *KDD '22: The 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, Washington, DC, USA, August 14 - 18, 2022*, pages 2253–2262. ACM, 2022. doi: 10.1145/3534678.3539277. URL https://doi.org/10.1145/3534678.3539277.