Leveraging Multi-Method Evaluation for Multi-Stakeholder Settings

Christine Bauer
Johannes Kepler University Linz, Austria
christine.bauer@jku.at

Eva Zangerle
University of Innsbruck, Austria
eva.zangerle@uibk.ac.at

ABSTRACT

In this paper, we focus on recommendation settings with multiple stakeholders with possibly varying goals and interests, and argue that a single evaluation method or measure is not able to evaluate all relevant aspects in such a complex setting. We reason that employing a multi-method evaluation, where multiple evaluation methods or measures are combined and integrated, allows for getting a richer picture and prevents blind spots in the evaluation outcome.

KEYWORDS

recommender systems, evaluation, multi-methods, multi-stakeholder, user experience

1 INTRODUCTION

In recommender systems (RS) research, we observe a strong focus on advancing systems such that they accurately predict items that an individual user may be interested in. The approach of evaluating an RS is thereby largely focused on system-centric methods and metrics (e.g., recall and precision in leave-n-out analyses [16]). By employing such an evaluation approach and aiming at optimizing these metrics, the following crucial components in the ecosystem are neglected [6, 9]: (i) multiple stakeholders are embedded in the ecosystem, but current research largely considers merely the role of the end consumer; (ii) the stakeholders typically have diverging interests and objectives for an RS; however, accurately predicting a user's interests is the predominant focus in current RS research; and (iii) with taking a mainly accuracy-driven, system-centric approach to evaluation, many aspects that determine a user's experience with an RS are not considered [16]. This results in an incomplete picture of user experience, leaving "blind spots" that are not captured in the quality evaluation of an RS. Although studies [4] could show that a lower accuracy rate may increase the business utility (e.g., revenue) without any significant drop in user satisfaction, the objectives and interests of stakeholders other than the user are typically not the focus of research in academic settings in the RS community.

In this paper, we call for considering the multiple stakeholders in RS evaluation and postulate that only taking a multi-method evaluation approach allows for capturing and assessing the various interests, objectives, and experiences of these very stakeholders; thus, contributing to eliminating the blind spots in RS evaluation.

To illustrate the opportunities of multi-method evaluations in evaluating an RS from multiple stakeholders' perspectives, we exemplarily focus on the music domain. In Section 2, we outline related work. Section 3 points out the multitude and diversity of stakeholders to be considered in the evaluation of music RS. In Section 4, we exemplify for selected stakeholders how the integration of multiple evaluation methods contributes to eliminating blind spots in RS evaluation while balancing the stakeholders' interests.

2 RELATED WORK

The idea to combine different research methods is not a new one. The concept of mixed methods research [8], for instance, combines quantitative and qualitative research approaches. It has been termed the third methodological paradigm, with quantitative and qualitative methods representing the first and second paradigm respectively [23]. Yet, it seems that mixed methods research appears to attract considerable interest but is rarely brought into practice [3]. From a practical point of view, the reasons for the low adoption of evaluations leveraging multiple methods are manifold, including higher costs, higher complexity, wider skill requirements compared to adopting one method only [7].

For RS research, Gunawardana and Shani [10] point out that there is an extensive number of aspects that may be considered when assessing the performance of a recommendation algorithm. Indeed, already early research on RS pointed towards the wide variety of metrics available for system-centric RS evaluation, including classification metrics, predictive metrics, coverage metrics, confidence metrics, and learning rate metrics [22]. As accuracy-driven evaluation has been shown not to be able to capture all the aspects that are relevant for user satisfaction [18], more user-relevant metrics and measures have been introduced and considered over time [15] (so-called "quality factors beyond accuracy" [13]). This wider range of objectives includes qualities such as novelty, serendipity [11], or diversity [15].

Kohavi et al. [17] stress the importance of applying multiple metrics also in the field of A/B testing and online experiments, pointing out that different metrics reflect different concerns. For A/B testing in RS research, Ekstrand and Willemsen [9] emphasize the need to include methods and metrics that go beyond the typical A/B behavior metrics. They argue that the currently dominating RS evaluation based on implicit feedback and A/B testing (they refer to this combination as "behaviorism") is often very limited in its ability to explain why users acted in a particular way. They emphasize that experiments need to be thoroughly grounded in theory and point to the advantages of collecting subjective responses from users which may help to explain their behavior.

Jannach and Adomavicius [12] point out that academic research in the field of RS tends to focus on the consumer's perspective

with the goal to maximize the consumer's utility (measured in terms of the most accurate items for a user), while maximizing the provider's utility (e.g., in terms of profit) appears to be neglected. While industry research on RS will naturally build around the provider's perspective, publications in this area are scarce [25].

3 DIGITAL MUSIC STAKEHOLDERS

Various stakeholders are involved in the digital music value chain [1]. From songwriters who create songs; to performers (e.g., (solo) artists, bands); to music producers who take a broad role in the production of a track; to record companies, including the three major labels; to music platform providers with huge repositories of music tracks, acting at the interface to music consumers; and hundreds of millions of music consumers with different music preferences and various objectives for using RS (e.g., discovering previously unknown items, rediscovering items not having listened to in a while); to society at large with its social, economic, and political objectives and needs.

Some stakeholders focus on user experience, where the goal is to propose "the right music, to the right user, at the right moment" [19]. Other stakeholders have business-oriented utility functions [1]. For instance, artists will most likely want to have their own songs recommended to consumers. While some artists may be fine with any of their songs being recommended, others may prefer to increase the playcount of a particular song (e.g., to reach the top charts, which would open an opportunity to draw an even broader audience; or some song may generate higher revenues than others due to contract rules). Achieving additional 1,000 playcounts will not get apparent for highly popular artists with yearly playcounts of several billions, but could be an important milestone for a comparatively less popular (local) artist.

4 BALANCING STAKEHOLDER INTERESTS IN EVALUATION

In the following, we aim to make the case for multi-method evaluations that contribute to identifying the strong and weak spots of a music RS for the stakeholders involved, where we focus on the users' and artists' perspectives in this section.

From a user perspective, recommendations that are adequate in terms of system-centric measures—e.g., the predictive accuracy of recommendation algorithms—do not necessarily meet a user's expectations [18]. User-centric evaluation methods, in contrast, involve users who interact with an RS [16] to gather user feedback [5] either implicitly or explicitly (depending on the concrete evaluation design). Such methods measure a user's perceived quality of the RS at the time of recommendation, e.g., by established questionnaires [21]. Still, relying solely on user-centric methods does not reveal the accuracy of the recommendations, because, given the vast amount of items, users are not able to judge whether a given recommendation was indeed the most relevant one [5].

Measuring accuracy does not capture the recommendations' usefulness for users, because higher accuracy scores do not necessarily imply higher user satisfaction [20]. For instance, a user's most favorite song is an accurate prediction; still, repeating the same song five times is, though accurate, likely not a satisfying experience. Hence, we argue that for evaluating the user's perspective of a

RS—the user being only one of the many stakeholders involved—multiple evaluation methods and measures are required. This may include combining a set of different measures (ranging from recall and precision to serendipity, list diversity or novelty) or integrating different evaluation methods (ranging from leave-n-out offline experiments to user studies and A/B testing). Furthermore, although A/B-testing using user's implicit feedback is effective for testing the impact of different algorithms or designs on user behavior—and is, thus, frequently considered the "golden standard" for recommender evaluation—, it has limited ability in explaining why users acted in a particular way [9]. Additional information (e.g., users' subjective responses) is necessary to allow for explaining behavior.

In short, sticking to a single evaluation method narrows our view on the RS, literally having blinders on, while devising and evaluating RS. We can borrow from social and behavioral sciences, where, e.g., mixed-methods research combines quantitative and qualitative evaluations using different designs [8]. Creswell's proposed designs include—among others—the convergent parallel design and the sequential design. In the convergent parallel design, two evaluation methods are first applied in parallel, and finally integrated into a single interpretation. The sequential design uses sequential timing, employing the methods in distinct phases. The second phase of the study, using the second method, is designed such that it follows from the results of the first phase. Depending on the research goal and the concrete choice of methods, researchers may either interpret how the second phase's results help to explain the initial results (explanatory design) or they build on the exploratory results of the first phase to subsequently employ a different method (in the second phase) to test or generalize the initial findings (exploratory design). For instance, Kamehkhosh and Jannach [14] showed that in the field of music RS, the results of a conducted offline evaluation could be reproduced with online studies assessing the users' perceived quality of recommendations. Similarly, for the Recommender Systems Challenge 2017, participants firstly evaluated their prototypes in offline evaluations, before actually deploying them and evaluating them in the live system utilizing A/B tests [2], showing that many of RS who performed well in offline evaluations were able to repeat this in online experiments. However, some of the devised RS also performed substantially worse in online experiments—highlighting a shortcoming that was not revealed by evaluating from solely an offline perspective. Along the same lines, Ekstrand and Willemsen [9] state that utilizing behaviorism for evaluation purposes (e.g., through A/B tests) is not sufficient to understand *why* users act in a particular way and, for instance, like a particular recommendation.

While academic research in the field of RS tends to focus on maximizing the users' utility, some authors (e.g., [12]) emphasize the importance of profit (or value) maximization. Profit maximization may not only be a goal for platform providers, but also for artists who are the content providers for music platforms. From an artist's perspective, a good RS recommends the respective artist's songs sufficiently frequently, which may ultimately lead to playcounts, likes, purchases, profit maximization, etc. Evaluating for profit may, though, leave blind spots. For example, depending on the chosen strategy, an artist may want to emphasize other values such as expanding the audience (thus, reaching new listeners) or increasing the listening or purchase volume within the current fan base. Hence, metrics such as number of unique listeners per artist,

the sum of playcounts over all songs of an artist, and metrics such as profit-per-audience type may be valuable for RS optimization and need to be considered in the RS evaluation strategy. Accordingly, evaluation efforts need to elicit and integrate the artists' goals and preferences need to be elicited and integrated into the evaluation efforts. While evaluation on a per-artist-basis might be interesting for the individual artists (e.g., for a comparison between platforms and their integrated RS), it may not be adequate for an overall RS evaluation. Still, an RS needs to be evaluated for its ability to serve the various strategies and for revealing potential tendencies towards the one or other strategy. As the targeted strategy might correlate with artist characteristics (e.g., top-of-the-top artists vs. "long tail" artist; early career vs. come-back phase vs. long-term career; mainstream artists vs. niche genres), it might be in the society's interest to evaluate for and ensure a balance.

Having given these examples, we emphasize that, due to interdependencies between the RS and the various stakeholders' actions, the entire RS ecosystem has to be taken into account in the evaluation. For instance, low accuracy of recommendations and low user experience are not likely to continuously increase profits for the platform provider and all kinds of artists; high accuracy does not automatically imply high user experience and may not contribute to profit maximization.

5 CONCLUSIONS

In this position paper, we exemplarily focused on the digital music ecosystem to illustrate that multiple stakeholders are impacted by music RS, and discussed the opportunities of multi-method evaluations to consider the multiple stakeholders' perspectives. We emphasize that—irrespective of the application domain—there are always multiple stakeholders involved in recommendation settings. Hence, there are always multiple—and possibly diverging—perspectives and goals of these very stakeholders which need to be considered in evaluating an RS. Consequently, multiple evaluation methods and criteria have to be combined and possibly also weighted.

Multi-method evaluations allow for gathering a richer and more integrated picture of the quality of a RS and contributes to understanding the various phenomena involved in a multi-stakeholder setting, for which one method in isolation would be insufficient [24].

ACKNOWLEDGMENTS

REFERENCES

[1] Himan Abdollahpouri and Steve Essinger. 2017. Multiple stakeholders in music recommender systems. In *1st International Workshop on Value-Aware and Multistakeholder Recommendation at RecSys 2017 (VAMS '17)*. arXiv:1708.00120
[2] Fabian Abel, Yashar Deldjoo, Mehdi Elahi, and Daniel Kohlsdorf. 2017. Recsys challenge 2017: Offline and online evaluation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, New York, NY, USA, 372–373.
[3] Pär J Ågerfalk. 2013. Embracing diversity through mixed methods research. *European Journal of Information Systems* 22, 3 (2013), 251–256. https://doi.org/10.1057/ejis.2013.6
[4] Amos Azaria, Avinatan Hassidim, Sarit Kraus, Adi Eshkol, Ofer Weintraub, and Irit Netanely. 2013. Movie Recommender System for Profit Maximization. In *Proceedings of the 7th ACM Conference on Recommender Systems (RecSys '13)*. ACM, New York, NY, USA, 121–128. https://doi.org/10.1145/2507157.2507162
[5] Joeran Beel, Marcel Genzmehr, Stefan Langer, Andreas Nürnberger, and Bela Gipp. 2013. A Comparative Analysis of Offline and Online Evaluations and Discussion of Research Paper Recommender System Evaluation. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation (RepSys '13)*. ACM, New York, NY, USA, 7–14. https://doi.org/10.1145/2532508.2532511
[6] Robin Burke. 2017. Multisided fairness for recommendation. In *4th Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML '17)*. arXiv:1707.00093
[7] Ilknur Celik, Ilaria Torre, Frosina Koceva, Christine Bauer, Eva Zangerle, and Bart Knijnenburg. 2018. UMAP 2018 Intelligent User-Adapted Interfaces: Design and Multi-Modal Evaluation (IUadaptMe). In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization (UMAP '18)*. ACM, New York, NY, USA, 137–139. https://doi.org/10.1145/3213586.3226202
[8] John W. Creswell. 2003. *Research design: qualitative, quantitative, and mixed methods approaches* (2nd ed.). Sage Publications, Thousand Oaks, CA, USA.
[9] Michael D. Ekstrand and Martijn C. Willemsen. 2016. Behaviorism is not enough: better recommendations through listening to users. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. ACM, New York, NY, USA, 221–224. https://doi.org/10.1145/2959100.2959179
[10] Asela Gunawardana and Guy Shani. 2015. Evaluating Recommender Systems. In *Recommender Systems Handbook* (2nd ed.), Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, Boston, MA, USA, 265–308. https://doi.org/10.1007/978-1-4899-7637-6_8
[11] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Transaction on Information Systems* 22, 1 (Jan. 2004), 5–53. https://doi.org/10.1145/963770.963772
[12] Dietmar Jannach and Gediminas Adomavicius. 2017. Price and profit awareness in recommender systems. In *1st International Workshop on Value-Aware and Multistakeholder Recommendation at RecSys 2017 (VAMS '17)*. arXiv:1707.08029
[13] Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, and Markus Zanker. 2016. Recommender Systems — Beyond Matrix Completion. *Commun. ACM* 59, 11 (2016), 94–102. https://doi.org/10.1145/2891406
[14] Iman Kamehkhosh and Dietmar Jannach. 2017. User Perception of Next-Track Music Recommendations. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization (UMAP '17)*. ACM, New York, NY, USA, 113–121. https://doi.org/10.1145/3079628.3079668
[15] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Transactions on Interactive Intelligent Systems* 7, 1, Article 2 (2016), 42 pages. https://doi.org/10.1145/2926720
[16] Bart P. Knijnenburg and Martijn C. Willemsen. 2015. Evaluating Recommender Systems with User Experiments. In *Recommender Systems Handbook* (2nd ed.), Francesco Ricci, Lior Rokach, and Bracha Shapira (Eds.). Springer, Boston, MA, USA, 309–352. https://doi.org/10.1007/978-1-4899-7637-6_9
[17] Ron Kohavi, Alex Deng, Brian Frasca, Toby Walker, Ya Xu, and Nils Pohlmann. 2013. Online Controlled Experiments at Large Scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '13)*. ACM, New York, NY, USA, 1168–1176. https://doi.org/10.1145/2487575.2488217
[18] Joseph A. Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction* 22, 1 (2012), 101–123. https://doi.org/10.1007/s11257-011-9112-x
[19] Audrey Laplante. 2014. Improving music recommender systems: what can we learn from research on music tags?. In *15th International Society for Music Information Retrieval Conference (ISMIR '14)*. International Society for Music Information Retrieval, 451–456.
[20] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems (CHI EA '06)*. ACM, New York, NY, USA, 1097–1101. https://doi.org/10.1145/1125451.1125659
[21] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-centric Evaluation Framework for Recommender Systems. In *Proceedings of the 5th ACM Conference on Recommender Systems (RecSys '11)*. ACM, New York, NY, USA, 157–164. https://doi.org/10.1145/2043932.2043962
[22] Alan Said, Domonkos Tikk, Klara Stumpf, Yue Shi, Martha Larson, and Paolo Cremonesi. 2012. Recommender Systems Evaluation: A 3D Benchmark. In *Proceedings of the Workshop on Recommendation Utility Evaluation: Beyond RMSE (RUE '12)*, Vol. 910. CEUR Workshop Proceedings, 21–23. http://ceur-ws.org/Vol-910/
[23] Charles Teddlie and Abbas Tashakkori. 2009. *Foundations of mixed methods research: Integrating quantitative and qualitative approaches in the social and behavioral sciences*. Sage Publications, Thousand Oaks, CA, USA.
[24] Viswanath Venkatesh, Susan A. Brown, and Hillol Bala. 2013. Bridging the qualitative-quantitative divide: Guidelines for conducting mixed methods research in information systems. *MIS Quarterly* 37, 1 (2013), 21–54.
[25] Markus Zanker, Laurens Rook, and Dietmar Jannach. 2019. Measuring the impact of online personalisation: Past, present and future. *International Journal of Human-Computer Studies* (2019). https://doi.org/10.1016/j.ijhcs.2019.06.006