Received: 5 May 2022 | Accepted: 17 May 2022

DOI: 10.1002/aaai.12050

EDITORIAL

Recommender systems: Trends and frontiers

Dietmar Jannach[1] | Pearl Pu[2] | Francesco Ricci[3] | Markus Zanker[1,3]

[1]Faculty of Technical Sciences, University of Klagenfurt, Klagenfurt, Austria

[2]Faculty of Computer and Information Sciences, EPFL, Lausanne, Switzerland

[3]Faculty of Computer Science, Free University of Bozen-Bolzano, Bozen-Bolzano, Italy

Correspondence
Dietmar Jannach, Faculty of Technical Sciences, University of Klagenfurt, Klagenfurt, Kärnten, Austria.
Email: Dietmar.Jannach@aau.at

Abstract

Recommender systems (RSs), as used by Netflix, YouTube, or Amazon, are one of the most compelling success stories of AI. Enduring research activity in this area has led to a continuous improvement of recommendation techniques over the years, and today's RSs are indeed often capable to make astonishingly good suggestions. With countless papers being published on the topic each year, one might think the recommendation problem is almost solved. In reality, however, the large majority of published works focuses on algorithmic improvements and relies on data-based evaluation procedures which may sometimes tell us little regarding the effects new algorithms will have in practice. This special issue contains a set of papers which address some of the open challenges and frontiers in RSs research: *(i)* building interactive and conversational solutions, *(ii)* understanding recommender systems as socio-technical systems with longitudinal dynamics, *(iii)* avoiding abstraction traps, and *(iv)* finding better ways of assessing the impact and value of recommender systems without field tests.

RECOMMENDER SYSTEMS – A SUCCESS STORY, MOSTLY

Personalized suggestions for items to buy, news to read or movies to watch are nowadays ubiquitous on the web. Recommender systems are software solutions that generate these suggestions, commonly with the help of statistical models and machine learning techniques. Given their widespread use in practice, their often astonishingly good recommendations, and their proven value for consumers and providers, it is no surprise that research on recommender systems is flourishing. Today, we are witnessing a constantly growing interest in the topic both in academia and industry, with countless papers being published every year.

Given this continued research interest, the use of latest deep learning technology also in industry, for example (Steck et al. 2021), and the high quality of many deployed systems, one might think that the recommendation problem is almost solved. However, looking closer at today's published research on recommender systems, we find that the community seems to mainly focus on a rather narrow part of the overall problem setting. Specifically, the overwhelming majority of published works proposes new machine learning models to create better "one-shot" relevance rankings of items for a given user profile and application scenario. From a methodological viewpoint, these new models are then in most cases evaluated with the help of data-based (offline) experiments which do not involve humans in the loop.

Determining suitable item rankings is certainly an important problem in any recommender system. However, the ranking algorithm is still only one of several parts of the larger sociotechnical system that recommender systems represent (Jannach et al. 2016; Xiao and Benbasat 2007). Given today's predominance of algorithm research, it becomes apparent that we as a research community seem to focus too much only on a specific part of the problem, and in particular on one where increasingly complex algorithms may only lead to diminishing returns. Moreover, since we mainly optimize our models for abstract computational metrics (e.g., precision and recall) and not for key performance indicators in industry (e.g., increased sales or retention), it remains unclear if our improved models would indeed lead to better systems in practice (Jannach and Bauer 2020).

Besides these questions that relate to our predominant research methodology, there are a number of additional important challenges and corresponding opportunities that we currently face in our field. We discuss these challenges next, calling for action to push current frontiers in recommender systems research.

PUSHING THE FRONTIERS

In this section, we first discuss two key directions and research paths for the next generation of recommender systems. Afterwards, we shed light on two main areas where we may have to revise our research practices to make our research more impactful.

Towards human-like conversational recommendation

Most research today focuses on an application scenario where recommending is a non-interactive process. In such scenarios, the system monitors the consumer behavior, maybe collects some explicit feedback, and then provides appropriate recommendations. In most cases, users cannot give feedback on recommendations, ask for alternative suggestions, state specific preferences, correct the system's assumptions, or ask for an explanation. In real life, when people make recommendations to each other, all of these types of interactions—and probably many more—may happen. And they may even be required until a person seeking a recommendation is confident to make a decision.

The promise of *conversational recommender systems* (CRS) is to enable such more natural conversations, see (Jannach et al. 2021) for a survey. Early research in CRS was published more than two decades ago. Due to the limitations of natural language processing (NLP) at that time, these systems were often based on web interfaces with forms and buttons, and they were often knowledge-based (Burke 1999; Jannach 2004; Shimazu 2002). Today, with the advances in NLP and machine learning in general, much more natural and dynamic conversational systems have become possible (Chen et al. 2019; Christakopoulou, Radlinski, and Hofmann 2016; Zhou et al. 2020).

Much more work is however still needed, as today's "end-to-end learning" systems, which are trained on large dialog corpora, still have major limitations. In way too many cases, they return non-meaningful responses to users and they typically support only a small set of often pre-defined user intents. Most of today's systems for example cannot answer requests for explanations (Manzoor and Jannach 2021). Ultimately, we may seek to implement a sci-fi version of a human-like "recommender bot", which is emphatic and behaves socially, can engage in chit-chat, and is able provide persuasive arguments to recommendation seekers. This may be a grand challenge for AI and require inter-disciplinary research within and outside computer sciences. Still, this vision will certainly help us identify relevant research questions to address next.

Understanding recommenders as sociotechnical systems

The predominant machine learning perspective on the recommendation problem described above leads to the problem that a variety of important questions are underexplored in the research literature. In reality, recommender systems are not just algorithms, but sociotechnical systems, which are operated in a certain environment for longer periods of time and with a defined purpose. Humans interact with these systems and the system's outputs may affect individuals, organizations and even society. Recent research work on topics such as multi-stakeholder recommendation, system biases, fairness and various potentially negative effects of recommender systems started to address these important questions (Abdollahpouri et al. 2020; Deldjoo et al. 2021; Ekstrand et al. 2021). However, still too often these problems are mainly addressed from a purely algorithmic perspective, typically with the goal of balancing some competing abstract optimization objectives or of meeting some pre-defined target distribution. Moreover, many computational studies only consider a single point in time, but do not address longitudinal effects potentially emerging over time when the different actors of a sociotechnical system interact (Zhang et al. 2019; Hazrati and Ricci 2022).

Thus, to truly advance the field in the future, a more holistic and interdisciplinary approach is required to obtain results that are more impactful in the real world (Jannach and Zanker 2022). Within *computer science*, research on human-computer interaction (HCI) aspects seem to be explored too little compared to algorithms (Konstan and Terveen 2021), even changes in the user interface of recommenders may have significant impact on the acceptance and effectiveness of a system (Garcin et al. 2014; Steck, van Zwol, and Johnson 2015). In the neighboring *information systems* field, a more holistic approach is taken to study phenomena related to recommender systems, acknowledging that the underlying technology is important but that a system often cannot be studied independent of its context of use. Finally, some research questions may require building on insights from, for instance, psychology, consumer behavior or marketing. How to explain recommendations in an effective way or how to persuade and nudge users into a specific direction are typical examples of questions that cannot be solely answered from a computer science perspective (Miller 2019; Yoo, Gretzel, and Zanker 2013).

Avoiding abstraction traps

Approaching problems only from a computer science perspective and using only computational metrics for our evaluations may ultimately lead us to different *abstraction traps*, as identified by Selbst et al. (2019) in the area of fair machine learning. Abstraction certainly is one of the key principles of computer science, and most recommender systems research aims at developing general-purpose, domain-independent algorithms. In our field, the canonical abstraction of the recommendation problem is assuming we are given a dataset of past user-item interactions and the goal to fit a function to these noisy data without overfitting. This commonly-agreed operationalization has led to a mostly standardized scheme of evaluating recommendation algorithms and such an approach in principle should also ensure reproducibility and progress. Ultimately, the quest for general-purpose algorithms has to some extent led to a "leaderboard chasing" culture, where every published work on algorithms has to demonstrate that it advances the *state-of-the-art*. To demonstrate this progress, research papers typically provide empirical results for at least two datasets, preferably from two different domains. Unfortunately, the choice of the datasets, as well as most other particularities of the experimental setup (including preprocessing, baselines, or metrics) are in most cases almost arbitrary. As a result, any claim of advancing the state-of-the-art in general seems largely overstated. All that is shown is that a new model is better than a selection of previous algorithms in a very particular experimental setting chosen by the researcher (Cremonesi and Jannach 2021).

A more fruitful approach, in contrast, would be to evaluate a recommender system in its context of use, considering it as a sociotechnical system as described above. Many of the open challenges mentioned above, for example, the impact and value of a recommender on multiple stakeholders or whether it is fair or not, cannot be addressed with our predominant research operationalization and with abstract computational metrics such as precision and recall. Alternative and typically non-validated computational proxies, for example, for fairness, do not help here either. In contrast, to study such aspects, one has to first understand the idiosyncrasies of a given application domain, the social context in which a system is deployed, and what purpose a recommender system should fulfill in this context (Burke and Ramezani 2011; Gunawardana and Shani 2015; Jannach and Bauer 2020). Only when we understand under which circumstances a system is effective in a set of comparable environments, we may want to start to find suitable abstractions and make conclusions regarding the generalizability of our findings.

Improving offline evaluation

Several important facets of recommender systems cannot be studied with our established offline experimental procedures. Nonetheless, offline evaluations will remain to be an important research method, for example, to pre-select algorithms for inclusion in A/B tests based on their accuracy on historical data (Gunawardana and Shani 2015). Nonetheless, there are a number of ways in which we may improve and extend the scope of data-based approaches.

First, instead of only asking if algorithm A is better than algorithm B for a given accuracy metric, we could use the available data more often for analytical research. Such analytics could concern both domain-independent and domain-specific aspects. As a domain-independent question, we could aim at understanding the key factors of a dataset impacting algorithm performance. Or we could use offline experiments to compare algorithms in terms of their tendency to favor popular items. Whether recommending popular items is desirable may, however, depend on the particular application (Gunawardana and Shani 2015). In terms of application-specific analytics, a typical question could be to obtain a deeper understanding in which cases a recommendation was successful, see (Jannach, Ludewig, and Lerche 2017) for an example from the fashion domain.

Independently of what we aim at studying with offline experiments, we should more often follow a research approach that is guided by clear and explicit hypotheses, which then determine the experimental design and in particular the used metrics. In current research, as mentioned above, major parts of the chosen experimental configuration are often not justified beyond the fact that others used a similar configuration in previous research. Typical implicit hypotheses that a newly proposed deep learning architecture should be better than a previous one is often also rather vague and usually not informed by theoretical considerations.

In terms of the used metrics, significant and important research was done in the past years with respect to *beyond-accuracy* metrics, which cover aspects such as diversity, novelty, serendipity and, most recently, fairness (Castells, Hurley, and Vargas 2015). While the proposed metrics are certainly plausible and intuitive, many of them were not validated. For instance, it may not be entirely clear if a particular diversity metric correlates well with user perceptions. To be useful in practice, such a validation step, for example, with the help of user studies, is however important. Unfortunately, similar problems exist for commonly used accuracy metrics. There are a number of reports signifying that improved offline accuracy does not translate to improvements in terms of key performance indicators of a deployed application, see also for a discussion of experiences at Netflix (Gomez-Uribe and Hunt 2015; Steck et al. 2021).

Another shortcoming of common train-test evaluation setups is that such settings are not suited to study sequential or longitudinal dynamics that may emerge over time. Therefore, in recent years new offline evaluation approaches were proposed, which address some of these limitations. Most importantly, a number of simulation environments as well as counterfactual reasoning approaches were put forward to study algorithms that are based on reinforcement learning, see, for example (Li et al. 2010; Rohde et al. 2018; Shi et al. 2019). The promise of such approaches is to narrow the gap between offline and online experimentation. An alternative to these approaches, which mainly aim at comparing algorithms, is to use simulation techniques to study complex longitudinal effects of recommenders, for example, in terms of usefulness of recommendations for consumers over time or long-term effects that an algorithm may have on consumer trust or the profitability of providers (Hazrati and Ricci 2022; Ghanem, Leitner, and Jannach 2022; Zhang et al. 2019).

PAPERS IN THIS ISSUE

The six papers in this special issue push the current frontiers in recommender systems and address several of the challenges of open questions outlined above. In their article, Jannach and Chen (2022) elaborate why building a conversational recommender system is difficult, and consider such systems a "Grand AI Challenge". Moreover, they discuss the challenges that come with the evaluation of conversational systems and outline a number of future directions in the area. The works by Sonboli et al. (2022) and Adomavicius et al. (2022) look beyond the computer science perspective and consider the sociotechnical environment of recommender systems. Sonboli et al. (2022) address the important topic of fairness in recommender systems and in particular address its multisided nature when various stakeholders should be considered. Adomavicius et al. (2022), on the other hand, study the effect of "preference pollution", which may occur when the available item ratings upon which recommender system operates are biased and not representative of the true user preferences. Two other papers, those by Afchar et al. (2022) and Massimo and Ricci (2022) focus on specific application domains, music and tourism, and thus aim at improving our understanding of particular problems in these areas, such as, the impact of item popularity in the recommender systems and in the users' evaluation of recommendations. Afchar et al. (2022) specifically address the problem of explainability of music recommendations and discuss questions of how to integrate such explanations within a large-scale industrial music streaming platform. Massimo and Ricci (2022), on the other hand, investigate the problem of recommending the next point-of-interest (POI) to tourists. They in particular discuss the specific constraints and idiosyncrasies of the problem setting and raise a discussion about evaluation methods for recommender systems and importance of understanding the true needs of users. In the last work in this issue, finally, Castells and Moffat (2022) reflect on the current state of offline evaluation, discuss the singularities that differentiate recommender system evaluation from information retrieval principles, and provide a survey on current developments, for example, in terms of considering potential evaluation biases with respect to the use of simulation approaches.

CONFLICT OF INTEREST

The authors declare that there is no conflict.

ORCID

Dietmar Jannach https://orcid.org/0000-0002-4698-8507
Pearl Pu https://orcid.org/0000-0001-8841-4416
Francesco Ricci https://orcid.org/0000-0001-5931-5566
Markus Zanker https://orcid.org/0000-0002-4805-5516

REFERENCES

Abdollahpouri, H., G. Adomavicius, R. Burke, I. Guy, D. Jannach, T. Kamishima, J. Krasnodebski, and L. Pizzato. 2020. "Multistakeholder Recommendation: Survey and Research Directions." *User Modeling and User-Adapted Interaction* 30: 127–58.

Adomavicius, G., J. Bockstedt, S. Curley, and J. Zhang. 2022. "Recommender Systems, Ground Truth, and Preference Pollution." *AI Magazine* 43(2): 177–89.

Afchar, D., A. B. Melchiorre, M. Schedl, R. Hennequin, E. V. Epure, and M. Moussallam. 2022. "Explainability in music recommender systems." *AI Magazine* 43(2): 190–208.

Burke, R. 1999. "The Wasabi Personal Shopper: A Case-based Recommender system." In *AAAI '99/IAAI '99*, 844–9.

Burke, R. D., and M. Ramezani. 2011. "Matching Recommendation Technologies and Domains." *Recommender Systems Handbook*, edited by F. Ricci, L. Rokach, B. Shapira, and P. B. Kantor, 367–86. New York, NY: Springer.

Castells, P., N. J. Hurley, and S. Vargas. 2015. "Novelty and Diversity in Recommender Systems." *Recommender Systems Handbook*, edited by F. Ricci, L. Rokach, and B. Shapira, 881–918. New York: Springer.

Castells, P., and A. Moffat. 2022. "Offline Recommender System Evaluation: Challenges and New Directions." *AI Magazine* 43(2): 225–38.

Chen, Q., J. Lin, Y. Zhang, M. Ding, Y. Cen, H. Yang, and J. Tang. 2019. "Towards Knowledge-based Recommender Dialog System." In *EMNLP-IJCNLP '19*, 1803–13.

Christakopoulou, K., F. Radlinski, and K. Hofmann. 2016. "Towards Conversational Recommender Systems." In *KDD '16*, 815–24.

Cremonesi, P., and D. Jannach. 2021. "Reproducibility and Progress in Recommender Systems Research: Crisis? What Crisis?" *AI Magazine* 42(3): 43–54.

Deldjoo, Y., V. W. Anelli, H. Zamani, A. Bellogin, and T. Di Noia. 2021. "A Flexible Framework for Evaluating User and Item Fairness in Recommender Systems." *User Modeling and User-Adapted Interaction* 31: 1–47.

Ekstrand, M. D., A. Das, R. Burke, and F. Diaz. 2021. "Fairness and Discrimination in Information Access Systems." *CoRR* abs/2105.05779.

Garcin, F., B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber. 2014. "Offline and Online Evaluation of News Recommender Systems at swissinfo.ch." In *RecSys '14*.

Ghanem, N., S. Leitner, and D. Jannach. 2022. "Balancing Consumer and Business Value of Recommender Systems: A Simulation-based Analysis." *arXiv preprint arXiv:2203.05952*.

Gomez-Uribe, C. A., and N. Hunt. 2015. "The Netflix Recommender System: Algorithms, Business Value, and Innovation." *Transactions on Management Information Systems* 6(4): 13:1–19.

Gunawardana, A., and G. Shani. 2015. "Evaluating Recommender Systems." *Recommender Systems Handbook*, edited by F. Ricci, L. Rokach, and B. Shapira, 265–308. New York: Springer.

Hazrati, N., and F. Ricci. 2022. "Recommender Systems Effect on the Evolution of Users' Choices Distribution." *Information Processing & Management* 59(1): 102766.

Jannach, D. 2004. "ADVISOR SUITE – A Knowledge-based Sales Advisory System." In *ECAI '04*, 720–4.

Jannach, D., and C. Bauer. 2020. "Escaping the McNamara Fallacy: Towards More Impactful Recommender Systems Research." *AI Magazine* 41(4): 79–95.

Jannach, D., and L. Chen. 2022. "Conversational Recommendation: A Grand AI Challenge." *AI Magazine* 43(2): 151–63.

Jannach, D., M. Ludewig, and L. Lerche. 2017. "Session-based Item Recommendation in E-commerce: On Short-term Intents, Reminders,Trends, and Discounts." *User-Modeling and User-Adapted Interaction* 27(3–5): 351–92.

Jannach, D., A. Manzoor, W. Cai, and L. Chen. 2021. "A Survey on Conversational Recommender Systems." *ACM Computing Surveys* 54(5): 1–26.

Jannach, D., P. Resnick, A. Tuzhilin, and M. Zanker. 2016. "Recommender Systems - Beyond Matrix Completion." *Communications of the ACM* 59(11): 94–102.

Jannach, D., and M. Zanker. 2022. "Impact and Value of Recommender Systems." *Recommender Systems Handbook*, edited by F. Ricci, B. Shapira, and L. Rokach, New York: Springer.

Konstan, J. A., and L. G. Terveen. 2021. "Human-Centered Recommender Systems: Origins, Advances, Challenges, and Opportunities." *AI Magazine* 42(3): 31–42.

Li, L., W. Chu, J. Langford, and R. E. Schapire. 2010. "A Contextual-bandit Approach to Personalized News Article Recommendation." In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, 661–70.

Manzoor, A., and D. Jannach. 2021. "Conversational Recommendation Based on End-to-end Learning: How Far Are We?" *Computers in Human Behavior Reports* 4: 100139.

Massimo, D., and F. Ricci. 2022. "Building Effective Recommender Systems for Tourists." *AI Magazine* 43(2): 209–24.

Miller, T. 2019. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence* 267: 1–38.

Rohde, D., S. Bonner, T. Dunlop, F. Vasile, and A. Karatzoglou. 2018. "RecoGym: A Reinforcement Learning Environment for the Problem of Product Recommendation in Online Advertising." arXiv preprint arXiv:1808.00720.

Selbst, A. D., D. Boyd, S. A. Friedler, S. Venkatasubramanian, and J. Vertesi. 2019. "Fairness and Abstraction in Sociotechnical Systems." In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, 59–68.

Shi, B., M. G. Ozsoy, N. Hurley, B. Smyth, E. Z. Tragos, J. Geraci, and A. Lawlor. 2019. "PyRecGym: A Reinforcement Learning Gym for Recommender Systems." In *Proceedings of the 13th ACM Conference on Recommender Systems*, 491–5.

Shimazu, H. 2002. "ExpertClerk: A Conversational Case-based Reasoning Tool for Developing Salesclerk Agents in E-Commerce Webshops." *Artificial Intelligence Review* 18(3-4): 223–44.

Sonboli, N., R. Burke, M. Ekstrand, and R. Mehrotra. 2022. "The Multisided Complexity of Fairness in Recommender Systems." *AI Magazine* 43(2): 164–76.

Steck, H., L. Baltrunas, E. Elahi, D. Liang, Y. Raimond, and J. Basilico. 2021. "Deep Learning for Recommender Systems: A Netflix Case-Study." *AI Magazine* 42(3): 7–18.

Steck, H., R. van Zwol, and C. Johnson. 2015. "Interactive Recommender Systems: Tutorial." In *RecSys '15: Proceedings of the 9th ACM Conference on Recommender Systems*, 359–60.

Xiao, B., and I. Benbasat. 2007. "E-commerce Product Recommendation Agents: Use, Characteristics, and Impact." *MIS Quarterly* 31(1): 137–209.

Yoo, K.-H., U. Gretzel, and M. Zanker. 2013. *Persuasive Recommender Systems: Conceptual Background and Implications*. New York: Springer.

Zhang, J., G. Adomavicius, A. Gupta, and W. Ketter. 2019. "Consumption and Performance: Understanding Longitudinal Dynamics of Recommender Systems via an Agent-based Simulation Framework." *Information Systems Research* 31(1): 76–101.

Zhou, K., W. X. Zhao, S. Bian, Y. Zhou, J.-R. Wen, and J. Yu. 2020. "Improving Conversational Recommender Systems via Knowledge Graph Based Semantic Fusion." In *KDD '20*, 1006–14.

AUTHOR BIOGRAPHIES

Dietmar Jannach (dietmar.jannach@aau.at) is a professor of computer science at the University of Klagenfurt, Austria. His research is focused on practical applications of artificial intelligence, with a focus on recommender systems. He is also the leading author of

the first textbook on recommender systems published with Cambridge University Press.

Pearl Pu (pearl.pu@epfl.ch) is a senior researcher at the Ecole Polytechnique Fédérale de Lausanne, where she founded the human–computer interaction group. She has served on many editorial boards such as *IEEE Multimedia*, *ACM TIIS*, *UMUAI*, *the AI Magazine*, and *the Harvard Data Science Review*. She was program chair of the ACM conferences on Electronic Commerce, Recommender Systems, and Adaptive Hypermedia, and track and area chair at many other scientific conferences. Pearl was named a distinguished speaker of the Association for Computing Machinery between 2016-2020.

Francesco Ricci (fricci@unibz.it) is a professor of computer science at Free University of Bozen-Bolzano, Italy. His research focuses on recommender systems and decision support, in context-aware and group recommendation scenarios, with a particular attention to the tourism application domain. He co-edited *the Handbook of Recommender Systems* (2015, Springer).

Markus Zanker (markus.zanker@unibz.it) is a professor at the Faculty of Computer Science at Free University of Bozen-Bolzano and he is, in addition, associated with University of Klagenfurt. His research focuses on information systems supporting decision making processes such as personalized information filtering and retrieval as well as product recommendation.

How to cite this article: Jannach, D., P. Pu, F. Ricci, and M. Zanker. 2022. "Recommender systems: Trends and frontiers." *AI Magazine* 43: 145–50. https://doi.org/10.1002/aaai.12050