Recommender Systems: Past, Present, Future*

Dietmar Jannach, Pearl Pu, Francesco Ricci, Markus Zanker

The origins of modern recommender systems date back to the early 1990s when they were mainly applied experimentally to personal email and information filtering. Today, thirty years later, personalized recommendations are ubiquitous and research in this highly successful application area of AI is flourishing more than ever. Much of the research in the last decades was fueled by advances in machine learning technology. However, building a successful recommender system requires more than a clever general-purpose algorithm. It requires an in-depth understanding of the specifics of the application environment and the expected effects of the system on its users. Ultimately, making recommendations is a human-computer interaction problem, where a computerized system supports users in information search or decision-making contexts. This special issue contains a selection of papers reflecting this multi-faceted nature of the problem and puts open research challenges in recommender systems to the forefront. It features articles on the latest learning technology, reflects on the human-computer interaction aspects, reports on the use of recommender systems in practice, and it finally critically discusses our research methodology.

The 1990s laid many foundations of modern-day recommender systems. In 1992, the concept of ''Collaborative Filtering'' was introduced with an experimental mail system called Tapestry [Goldberg et al. 1992], where users could write mail filtering rules that, among other aspects, could relate to the opinions and behavior of others. Soon later, in 1994, the GroupLens news filtering system [Resnick et al. 1994] was presented, which aimed at automating the rule-based collaborative filtering process of the Tapestry system. With GroupLens, one of the first systems was proposed, which *(i)* operated on the basis of explicit ratings provided by a community of users and *(ii)* which employed machine learning to make predictions if a user will like specific unseen messages.

With the rapid development of the World Wide Web in the 1990s, more and more application areas for recommender systems emerged. Even before the decade

*Appeared in AI Magazine, Vol. 42(3), 2021

ended, a number of success stories regarding the use of recommender systems in e-commerce were reported [Schafer et al. 1999] with Amazon.com being one of the first adopters of recommendation technology at large scale [Linden et al. 2003]. Today, personalized recommendations are an ubiquitous element of our online experience, and many more reports on the business value of such recommendations were published over the years [Jannach and Jugovac 2019].

From a technical perspective, the early GroupLens system framed the recommendation task as a ''matrix filling'' problem, where the input to the machine learning algorithm is a sparse user-item rating matrix and the goal is to predict the missing entries. This problem abstraction is still predominant today, with the main difference that in real problem settings *(i)* the matrix entries are more often implicit feedback signals than ratings and *(ii)* the ranking of items is more relevant than the accurate prediction of relevance scores.

The technical solutions for such prediction and ranking problems changed however significantly over the years. The early GroupLens system relied on a comparably simple nearest-neighbor approach. Since then, however, all sorts of machine learning methods were applied or tailored to the problem setting. For many years, matrix factorization techniques, also proposed first to be used for collaborative filtering in the 1990s [Billsus and Pazzani 1998], dominated the landscape. Later, research on the use of machine learning algorithms for rating prediction and item ranking was supercharged by the Netflix Prize (2006 to 2009), where the goal was to make accurate movie rating predictions. Today, fifteen years after the Netflix Prize was launched, research of a very similar nature is booming, this time fueled by the broad success and use of deep learning in many application areas of machine learning.

After at least twenty-five years of algorithms research, one might assume that the recommendation problem is solved, not only because of algorithmic improvements, but also because scholars have identified and addressed a variety of limitations of the original matrix-completion problem abstraction and evaluation approach. For example, considering only the accuracy of individual predictions does not allow us to assess potentially desired quality factors of entire *lists* of recommendations, such as the diversity of the recommendations or the novelty of the identified items. These insights led to the development of a number of ''beyond-accuracy'' measures in algorithms research that relate, for example, to the novelty or serendipity of a set of recommendations. Moreover, given the possible limitations of only utilizing the available ratings, researchers proposed to consider all sorts of side information---e.g., item-related data such as tags or meta-data, multi-criteria ratings, the user's social network, contextual information, or time---within hybrid recommendation approaches. Finally, in recent years, researchers started to more frequently consider an alternative problem abstraction, where the main input is not a user-item rating matrix but a sequential log of recorded user sessions, termed ''sequence-aware recommendation'' [Quadrana et al. 2018] and the recommendations should offer relevant ''next items'' to explore to the user.

Nonetheless, even though the field has matured over the last decades, and even though many algorithmic proposals are published every year, the recommendation problem is far from being solved. This, to a large extent, has to do with the

predominant way of how research is done in this field. Today, the research community to an overwhelming extent relies on data-centric ''offline'' experiments that do not involve the human in the loop. These offline experiments are however based on a number of assumptions. In particular, it is assumed that the used computational metrics (e.g., Precision and Recall) are suitable proxies for the effectiveness of an algorithm whenever it will be deployed online. In reality, however, higher precision obtained by an algorithm in an offline experiment does not necessarily mean that it will lead to the desired impact or business value, e.g., in terms of sales or user engagement.

As a result, the danger exists that much of the research done in this area is relying on simplifications and is making assumptions that are generally too strong. As a result, it is important that we re-focus our research efforts to ensure that we investigate problems that matter. More research is needed that aims at understanding how recommender systems affect the individual and collective behavior of humans and what this behavior change means for organizations and societies. To truly understand these phenomena, it is often important to understand the idiosyncrasies of the particular application, as the intended effects of using a recommender system---and thus the relevant performance metrics---largely depend not only on a particular domain or application, but even on the business model of the provider.

Therefore, we should focus much more often on the problem of understanding how systems affect both organizations and entire user experience journey than on minor improvements in prediction accuracy on historical datasets. In fact, various questions regarding the design of the user experience are largely unexplored in the literature. How may users reveal their preferences and would they worry about privacy? How long and diverse should a recommendation list be? What leads to a perception of diversity, novelty or familiarity? How should a system explain its recommendations? In areas such as fitness and health, how should new applications offer joyful experience while addressing fears of privacy and overpersuasion? More and more applications appear that may change the behavior of individuals, thinking of apps for personal fitness or health. Still, little research seems to be done regarding the design choices for such systems.

This special issue contains a selection of papers that address many of the above mentioned issues and topics. It features *(i)* papers that emphasize on the human-centric perspective of recommender systems, *(ii)* up-to-date reports of successful deployments and open challenges of recommendation technology in industry, and *(iii)* works that critically reflect current research practices and outline future directions in offline evaluation.

Understanding the HCI Side of Recommendation Recommendation is---to a large extent---a problem of human-computer interaction (HCI) and user experience design. Being able to estimate that a user will most probably like a movie may often not be enough. It might for example happen that a recommendation algorithm returns a collection of rather niche items such as a Spanish black-and-white movie from the 1930s, see also the study by [Ekstrand et al. 2014].

In such a case, a user might not even try out this recommendation, even if she or he would have loved the movie. In such a situation, users might only follow such recommendations if they have already developed trust into the system over time. Or, they might only explore such a recommendation if the system by some other means such as explanations increases the user's interest in the movie or their confidence in the system's recommendations.

Generally, recommender systems are often characterized as tools that help users in their decision-making process. How a system can support users in this process in the best possible way is a central question in HCI research in recommender systems. Typical questions in that context related, for example, to the number of options that should be presented to the user, or how and when they should be presented. Moreover, HCI research---differently from offline experimentation---allows us to explore if users are satisfied with recommendations, if they discovered something new, or if they would like to continue receiving recommendations. In [Konstan and Terveen 2021], the authors review the last 25 years of recommender systems research from a human-centered perspective and look at the challenges and opportunities that come with the recent developments in machine learning with respect to the design of effective recommender systems.

The Impact and Value of Recommender Systems Academic research, as mentioned above, is often done based on a very abstract problem formalization and evaluated with the help of domain-independent computational metrics. The consideration of the idiosyncrasies of a particular recommender system deployment is however crucial in any practical application. Typically, building an effective system requires an in-depth understanding of *(i)* which types of recommendations create value, both for the consumer and the provider of the recommendations, *(ii)* how the impact and success of the recommender can be measured, *(iii)* and if there are any risks associated with the recommendations.

Understanding such aspects is also crucial for academic researchers. While academic research generally strives for generalizable approaches, it remains important to address problems that matter in practice. This special issue features three papers that report on successful applications and open challenges in different practical settings.

In [Steck et al. 2021], the authors report on the journey of Netflix adopting deep learning technology for their various recommendation problems. One key insight of their work is that this technology can be particularly helpful when side information beside the implicit and explicit user feedback signals is considered. [Gulla et al. 2021], on the other hand, reports on challenges of building recommender systems for the news domain and how technology has changed an entire industry. They in particular also report on the difficulties that medium sized and traditional media companies face when adopting personalization and recommendation technology.

New Directions in Offline Evaluation Despite their limitations, offline evaluations will remain a valuable means to investigate certain aspects of

recommendation algorithms, e.g., if they have a tendency to recommend mostly popular items, and to compare different machine learning models. Several researchers however argue that a shift is needed in terms of how we do offline evaluations. Like in other application areas of machine learning, we often observe a hyper-focus on benchmark datasets [Wagstaff 2012] and accuracy measures. Moreover, various works report that the predominant offline evaluation approach used today---predicting held-out user interactions---is suited to estimate how an algorithm would behave in practice.

Two papers in this issue address questions related to the evaluation of recommender systems. In [Cremonesi and Jannach 2021], the authors report on existing problems of today's research practice, including a certain lack of reproducibility and methodological issues that may prevent the field from moving forward. As an alternative way of building and evaluating recommender systems [Joachims et al. 2021] propose to consider ''*recommendations as treatments*''. They propose to adopt an *interventional* view on recommender systems and highlight in which ways off-policy evaluation based on counterfactual estimators may help to overcome limitations of today's offline evaluation procedures and to more accurately predict how an alternative algorithm (policy) would fare in an online A/B test.

References

[Billsus and Pazzani 1998] D. Billsus and M. J. Pazzani. Learning collaborative information filters. In *Proceedings of the Fifteenth International Conference on Machine Learning*, ICML '98, page 46–54, 1998.

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

[Ekstrand et al. 2014] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, and J. A. Konstan. User perception of differences in recommender algorithms. In *Proceedings of the 2014 ACM Conference on Recommender Systems (RecSys '14)*, pages 161--168, 2014.

[Goldberg et al. 1992] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using Collaborative Filtering to Weave an Information Tapestry. *Commun. ACM*, 35 (12):61–70, 1992.

[Gulla et al. 2021] J. A. Gulla, R. D. Svendsen, L. Zhang, and J. Frøland. Recommender Systems in Online News Personalization. *AI Magazine*, 42(3), 2021.

[Jannach and Jugovac 2019] D. Jannach and M. Jugovac. Measuring the business value of recommender systems. *ACM Transactions on Management Information Systems (TMIS)*, 10(4):1--23, 2019.

[Joachims et al. 2021] T. Joachims, B. London, Y. Su, A. Swaminathan, and L. Wang. Recommendations as Treatments. *AI Magazine*, 42(3), 2021.

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

[Linden et al. 2003] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing*, 7(1):76--80, 2003.

[Quadrana et al. 2018] M. Quadrana, P. Cremonesi, and D. Jannach. Sequence-aware recommender systems. *ACM Computing Surveys*, 51:1--36, 2018.

[Resnick et al. 1994] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, CSCW '94, page 175–186, 1994.

[Schafer et al. 1999] J. B. Schafer, J. Konstan, and J. Riedl. Recommender systems in e-commerce. In *Proceedings of the 1st ACM Conference on Electronic Commerce*, EC '99, page 158–166, 1999.

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

[Wagstaff 2012] K. Wagstaff. Machine learning that matters. In *Proceedings of the International Conference on Machine Learning, ICML '12*, pages 529--536, 2012.

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. His research focuses on information systems supporting decision making processes such as personalized information filtering and retrieval and product recommendation.