Towards More Impactful Recommender Systems Research

A Working Paper

Dietmar Jannach
University of Klagenfurt, Austria
dietmar.jannach@aau.at

Oren Sar Shalom
Intuit AI, Israel
Oren_SarShalom@intuit.com

Joseph A. Konstan
University of Minnesota, USA
konstan@umn.edu

ABSTRACT

A number of ways exist in which a recommender system can have an impact on users and business. However, only a small number of them can be reasonably addressed with today's predominant and narrow research approach based on offline experimentation and accuracy measures. It sometimes even stands to question if small increases in prediction accuracy will actually lead to a better systems in any of the ways in which a recommender system can impact users and create value. We therefore argue for a more *impact-oriented* approach to research in the field of recommender systems. With such a refocused lens, we hope that the corresponding research results are also more *impactful* and relevant in reality. To foster such research, we present in this work a first taxonomy describing the various facets to consider when developing impact-oriented research, ranging from the expected value of a recommender for different stakeholders to the potential risks that come with such applications.

CCS CONCEPTS

• **Information systems** → **Recommender systems**; • **Social and professional topics** → *Socio-technical systems.*

KEYWORDS

Impact of Recommender Systems, Evaluation

1 INTRODUCTION

In the field of Computer Science, academic research on recommender systems is dominated by the hunt for algorithmic improvements. The main goal of most research works in this context is to demonstrate that the newly proposed technical approach is better than previous works at predicting held-out user preferences for unseen items. The underlying, implicit assumption of researchers is that when we can estimate the relevance of an item more reliably, the actually relevant items will be placed higher in the recommendation lists presented to the users. As a result, users can be positively impacted by the recommender system, e.g., as they can more easily find what they are looking for.

Framing the recommendation task as a "curve-fitting" problem is attractive for researchers for different reasons, for example, because it helps us to abstract from domain specifics. More importantly, however, it also relieves us from the hassle of creating and arguing for a specific experimental design for our research. As long we rely on our implicit assumption described above, it is considered sufficient to show that a new algorithm is better, by a few percent, in an experimental configuration of (i) baselines, (ii) datasets, and (iii) evaluation procedure and (iv) performance measure.The experimental configuration can be chosen by the researcher with some degrees of freedom. Since, for example, our techniques are usually domain-agnostic, the choice if the dataset—as one part of the experiment setup–can almost be done arbitrarily.

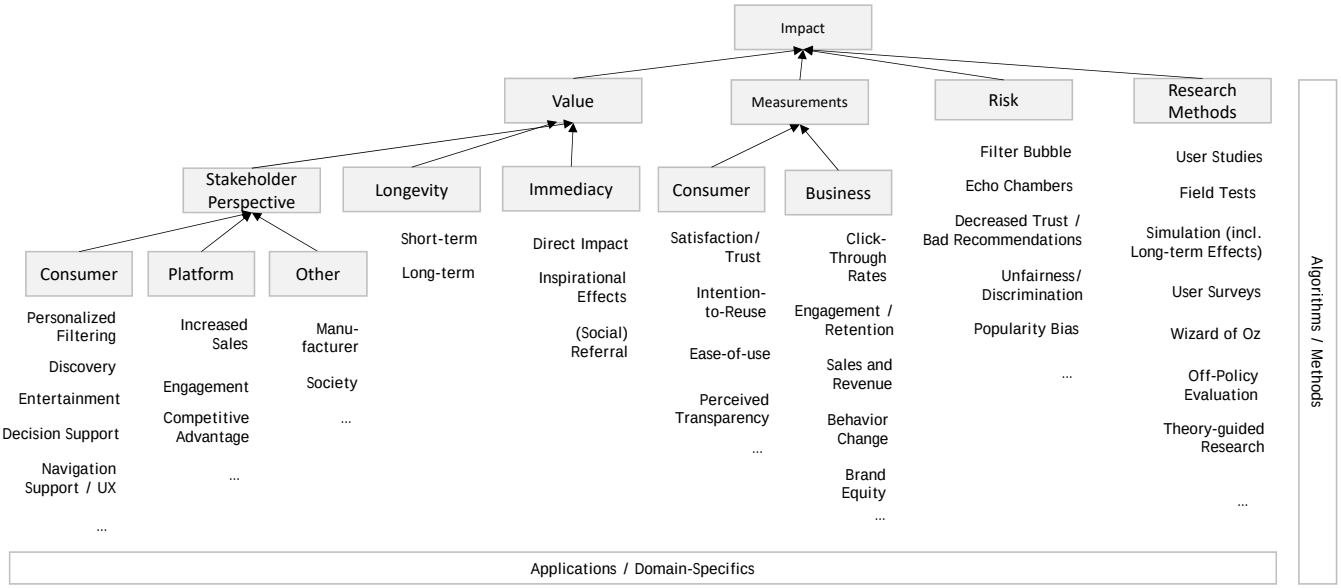
Generally, there is no doubt that the choice and the design of the recommendation algorithm matters a lot in practice. In [10], various reports on practical deployments of recommender systems are summarized. These reports show that recommender systems impact users in different ways, e.g., regarding which choices they make or how engaged they are with the site. This, in turn, can lead to a number of effects that are desirable from the perspective of the business, e.g., in terms of increased sales or customer retention.

However, in almost all of these results that were observed in A/B tests, the comparison was made between algorithms that were quite different, e.g., a simple popularity-based method vs. a machine learning approach, or a purchase-based vs. a view-based collaborative filtering technique [14]. This is quite some contrast to the academic world, where we often compare algorithm variants of very similar types (e.g., two matrix factorization techniques that were trained on the same data). Since the differences are often small, it remains to question if small increases will actually make a difference with respect to the users' *perception* of the recommendations. A number of user studies could in fact *not* show that higher offline accuracy led to a better quality perception by users [2, 4, 6, 21]. At the same time, Gomez-Uribe and Hunt mention in [7] that in the case of the Netflix recommender system the outcomes of offline experiments were not too predictive of online "success" either.

Generally, focusing only on one measure, in our case prediction accuracy, is well known to be "not enough" [18] and might actually hamper progress in the field of recommender systems. Various additional quality factors were discussed over the past ten years, e.g., with respect to diversity, novelty or serendipity. However, as long as these aspects are only investigated through offline experiments and computational measures, it remains unclear if users actually like more diversified recommendation lists or if they will lead to more business value. In some ways, with the predominant algorithm-centric approach, a certain stagnation in certain subfields of recommender systems research can be identified. Rendle et al., for example, recently found that—even in the traditional core area of making better rating predictions—the techniques published in the last few years were actually not better than what we had years ago. Similar phenomena were found in the area of information retrieval recently [15] and ten years ago [1].

ImpactRS Workshop at ACM RecSys '19, September 19, 2019, Copenhagen, Denmark

Figure 1: Conceptual Model of Aspects of Impact-Oriented Research in Recommender Systems.



With this position paper and with the *1st Workshop on the Impact of Recommender Systems* at the ACM Conference on Recommender Systems, our goal is to raise the awareness among researchers regarding the importance of *impact-oriented* research. Specifically, we fear that certain parts of today's academic research might have very limited impact in practice. To outline possible areas of future research, we present a first taxonomy of impact-oriented research in the next section.

2 ASPECTS OF IMPACT-ORIENTED RECOMMENDER SYSTEMS RESEARCH

The limitations of almost mono-thematic, algorithms- and accuracy-centric "leaderboard chasing" research on recommender systems in the field of Computer Science has been discussed previously in the literature, see, e.g., [13], a paper which in particular focuses on the *user experience* or [11], where the authors argue for a *systems*-oriented research approach. The Information Systems literature is much richer in this respect, see e.g., the conceptual framework proposed in [24], where the algorithms are only one of many ingredients to consider when designing recommender systems.

In the spirit of this latter work, we propose in this present work a taxonomy (see Figure 1) as a conceptual model of relevant aspects to consider for *impact-oriented*, and thus hopefully more impactful, research on recommender systems.

Value and Purpose. Research today often focuses on a small set of potential ways in which a recommender system can create value for the *consumer*. A recommender system can, however, serve various purposes and, furthermore, be beneficial for various stakeholders, including and service providers, online retailers or manufacturers [8, 22]. This value can be created *directly*, e.g., through increased sales, or more *indirectly*, e.g., through increased engagement and customer trust. In our networked society, recommendations can furthermore create value through *referral*, e.g., on social media. Another differentiation is related to the longevity of the value, i.e., if it is a *short-term* or a lasting *long-term effect*. Overall, to be more impactful, it is important that researchers consider such details of such value-related aspects and also investigate potential trade-offs between stakeholder goals in more depth in the future.

Risks and Responsible Recommendations. Recommender systems can also have undesired impacts and risks. An obvious one is that individual users might be dissatisfied with the quality of the recommendations and stop using the service. But also on the organizational or even societal level recommenders can have an effect. They can, for example, can lead to actually decreased diversity and popularity reinforcement effects [14] for an e-commerce site or create filter bubbles [19] and echo chambers in a social network or society. Other risks include discrimination or the lack of fairness [5], topics which received increased interest in recent years, and which require novel mechanisms for responsible recommendation.

Research Methods and Measurements. Today's methodological repertoire can sometimes appear very narrow, mostly consisting of offline experiments as mentioned above, which can tell us little about impact. Given the many limitations of such a research approach, more user-centric research is needed, and many tools for this type of research are readily available [12, 20]. In some areas, e.g., in conversational recommendation, even more foundational research is needed to understand how humans interact and how an effective computerized recommendation dialog should be designed. Here, observational studies or Wizard-of-Oz experiments are promising means for future research.

But there is also potential for better offline evaluation. Alternative approaches, e.g., based on the simulation of long-term effects on user behavior, could for instance be explored to obtain a better understanding of the effects of recommenders without the need

for a field test, see e.g., [9, 17, 25]. Recently, also techniques like *off-policy* evaluation [3] and counterfactual reasoning were investigated as a means to obtain a more realistic assessment of effects from offline data. While having solid theoretical justification, off-policy evaluation is focus on accuracy and hence lack the ability to predict various additional forms of impact. One possible future line of work could for example line in the extension of such approaches to also consider the phenomena of diversity.

Impact-Oriented Algorithms. As emphasized in the introduction, the majority of the algorithmic proposals that are published today focus solely on offline accuracy and do not take any of the discussed value perspectives directly into account. To be impactful, future research should focus more on the intended purpose of a system and its expected impact [8]. These could be, for example, algorithms that are able to consider business goals together with consumer value [16], persuasive or explanatory approaches that help users make better-informed decisions or increase their choice satisfaction, or approaches which better understand the user's current contextual situation (e.g., the phase in the decision process) when generating item suggestions.

Applications. Last but not least, domain-specific and application-specific aspects are far too often not taken into account. Ultimately, recommender systems research is mostly very applied, i.e., there mostly are no theories or hypotheses in algorithms-related research. The search for the "best" model across domains appears mostly futile, because many aspects are depending on the intended purpose of the system. Recommending items that are already popular can, for example, be desirable in one domain to increase revenue, and not desirable in another, where discovery support is the main goal.

Generally, the hunt for the best model, see also [23], can lead to a "leaderboard chasing" research approach where already tiny improvements are considered as a publishable result, and where the underlying reasons for the improvements become secondary. Furthermore, it can be assumed that there actually is no "best model" across domains as algorithm rankings typically depend on a variety of factors, including baselines, datasets, data pre-processing, optimization procedures, evaluation protocols and metrics.

Impact-oriented research therefore has to take these domain-specifics into account, both in the way the algorithms are designed and in the way the effects are evaluated. This, of course, does not mean that researchers should only focus on individual domains and not strive for generalizable approaches. It remains important that researchers develop techniques that generalize and are useful beyond the individual case.

3 SUMMARY

Current research is mostly focused on a very specific technical part of a recommender system. It furthermore often relies on a very abstract problem formulation and evaluation approaches, which make it difficult to judge if the obtained improvements are actually relevant in the real work. With this position paper and the workshop on the impact of recommender systems, we hope to raise awareness that more encompassing research approach is required to achieve insights that matter.

REFERENCES

[1] Timothy G. Armstrong, Alistair Moffat, William Webber, and Justin Zobel. 2009. Improvements That Don't Add Up: Ad-hoc Retrieval Results Since 1998. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*. 601–610.

[2] Jöran Beel and Stefan Langer. 2015. A Comparison of Offline Evaluations, Online Evaluations, and User Studies in the Context of Research-Paper Recommender Systems. In *Proceedings TPDL '15*. 153–168.

[3] Stephen Bonner and Flavian Vasile. 2018. Causal embeddings for recommendation. In *Proceedings of the 12th ACM Conference on Recommender Systems*. ACM, 104–112.

[4] Paolo Cremonesi, Franca Garzotto, and Roberto Turrin. 2012. Investigating the Persuasion Potential of Recommender Systems from a Quality Perspective: An Empirical Study. *Transactions on Interactive Intelligent Systems* 2, 2 (2012), 1–41.

[5] Michael D. Ekstrand, Robin Burke, and Fernando Diaz. 2019. Fairness and Discrimination in Retrieval and Recommendation. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2019*. 1403–1404.

[6] Florent Garcin, Boi Faltings, Olivier Donatsch, Ayar Alazzawi, Christophe Bruttin, and Amr Huber. 2014. Offline and Online Evaluation of News Recommender Systems at Swissinfo.Ch. In *Proceedings RecSys '14*. 169–176.

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

[8] Dietmar Jannach and Gediminas Adomavicius. 2016. Recommendations with a Purpose. In *Proceedings of the 10th ACM Conference on Recommender Systems (RecSys '16)*. 7–10.

[9] Dietmar Jannach and Gediminas Adomavicius. 2017. Price and Profit Awareness in Recommender Systems. *CoRR* abs/1707.08029 (2017). http://arxiv.org/abs/1707.08029

[10] Dietmar Jannach and Michael Jugovac. 2019. Measuring the Business Value of Recommender Systems. *CoRR* abs/1908.08328 (2019). arXiv:1908.08328

[11] Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, and Markus Zanker. 2016. Recommender Systems - Beyond Matrix Completion. *Commun. ACM* 59, 11 (2016), 94–102.

[12] B.P. Knijnenburg, M.C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. 2012. Explaining the user experience of recommender systems. *User Modeling and User-Adapted Interaction* 22, 4 (2012), 441–504.

[13] Joseph Konstan and John Riedl. 2012. Recommender systems: from algorithms to user experience. *User Modeling and User-Adapted Interaction* 22, 1-2 (2012), 101–123.

[14] Dokyun Lee and Kartik Hosanagar. 2014. Impact of Recommender Systems on Sales Volume and Diversity. In *Proceedings of the 2014 International Conference on Information Systems (ICIS '14)*.

[15] Jimmy Lin. 2019. The Neural Hype and Comparisons Against Weak Baselines. *SIGIR Forum* 52, 2 (Jan. 2019), 40–51.

[16] Xiao Lin, Hongjie Chen, Changhua Pei, Fei Sun, Xuanji Xiao, Hanxiao Sun, Yongfeng Zhang, Wenwu Ou, and Peng Jiang. 2019. A Pareto-efficient Algorithm for Multiple Objective Optimization in e-Commerce Recommendation. In *Proceedings of the 13th ACM Conference on Recommender Systems (RecSys '19)*. 20–28.

[17] Wei Lu, Shanshan Chen, Keqian Li, and Laks V. S. Lakshmanan. 2014. Show Me the Money: Dynamic Recommendations for Revenue Maximization. *Proc. VLDB Endow.* 7, 14 (2014), 1785–1796.

[18] 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)*. 1097–1101.

[19] Eli Pariser. 2011. *The Filter Bubble: What the Internet Is Hiding from You.* Penguin Group , The.

[20] Pearl Pu, Li Chen, and Rong Hu. 2011. A User-centric Evaluation Framework for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems (RecSys '11)*. 157–164.

[21] Marco Rossetti, Fabio Stella, and Markus Zanker. 2016. Contrasting Offline and Online Results when Evaluating Recommendation Algorithms. In *Proceedings RecSys '16*. 31–34.

[22] Özge Sürer, Robin Burke, and Edward C. Malthouse. 2018. Multistakeholder recommendation with provider constraints. In *Proceedings of the 12th ACM Conference on Recommender Systems, RecSys 2018*. 54–62.

[23] Kiri Wagstaff. 2012. Machine Learning that Matters. In *Proceedings ICML '12*. 529–536.

[24] Bo Xiao and Izak Benbasat. 2007. E-commerce Product Recommendation Agents: Use, Characteristics, and Impact. *MIS Quarterly* 31, 1 (2007), 137–209.

[25] Jingjing Zhang, Gediminas Adomavicius, Alok Gupta, and Wolfgang kKetter. 2019. Consumption and Performance: Understanding Longitudinal Dynamics of Recommender Systems via an Agent-Based Simulation Framework. *Information Systems Research* forthcoming (2019).