Measuring the Impact of Recommender Systems – A Position Paper on Item Consumption in User Studies

Benedikt Loepp
University of Duisburg-Essen
Duisburg, Germany
benedikt.loepp@uni-due.de

Jürgen Ziegler
University of Duisburg-Essen
Duisburg, Germany
juergen.ziegler@uni-due.de

ABSTRACT

While participants of recommender systems user studies usually cannot experience recommended items, it is common practice that researchers ask them to fill in questionnaires regarding the quality of systems and recommendations. While this has been shown to work well under certain circumstances, it sometimes seems not possible to assess user experience without enabling users to consume items, raising the question of whether the impact of recommender systems has always been measured adequately in past user studies. In this position paper, we aim at exploring this question by means of a literature review and at identifying aspects that need to be further investigated in terms of their influence on assessments in users studies, for instance, the difference between consumption of products or only of related information as well as the effect of domain, domain knowledge and other possibly confounding factors.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Recommender Systems; Experimentation; User Studies

1 THE PROBLEM WITH USER STUDIES

Questionnaires for assessing quality of recommendations and user experience of recommender systems (RS) have been, for instance, proposed in [4, 5, 8]. These established instruments are often employed in academic user studies, where participants usually have to first use a RS and are subsequently asked to fill in a questionnaire. However, recommended items in these scenarios are almost always represented through "proxy presentations", i.e. items are only shown to users by means of images, descriptive texts, metadata, etc. The actual consumption of items is in contrast to real-world situations rarely possible. There, it is mostly required to have, for instance, bought a product, visited a hotel, or watched a movie, before being even able to provide an opinion.

Previously, we have investigated whether the consumption of items during user studies has an impact on the succeeding assessment of recommendations by means of questionnaires [6]. In other studies, e.g. on explanations [1, 9], the impact of consumption has never been directly addressed. However, we found, among others, that it strongly depends on domain as well as type and amount of presented information whether it is possible for participants to adequately assess recommendation quality and aspects related to user experience without being able to consume recommended items. Accordingly, depending on certain circumstances, allowing participants to experience items may be a necessity for ensuring the validity of RS user studies.

While we were able to derive important conclusions for future user studies (e.g. results appear to provide at least a lower bound), there are many open questions that are strongly related, but go beyond what we already have investigated in the music and movie domain [6]. Regardless of the success of A/B tests in industry, user studies are especially important in academia, where they become more and more acknowledged as indispensable means for holistically capturing the qualities of RS [3]. Considering this and the generally increasing efforts towards reproducibility, it thus seems to be of particular interest to study the impact of item consumption and of other possibly confounding factors on the assessment of recommendations in user studies in more depth.

2 LITERATURE REVIEW

First, for putting our findings from [6] more into context, we have performed a literature review. We analyzed all 46 papers accepted to the five editions of the *Joint Workshop on Interfaces and Human Decision Making for Recommender Systems (IntRS)*[1] which were held from 2014 to 2018 in conjunction with the *ACM Conference on Recommender Systems (RecSys)*. In 66 % of these papers, a user study had been reported (there were a few more, which however did not focus on recommendation issues but e.g. "only" on comparing different interfaces). In some of the papers without a user study, applying such an evaluation method would not have been appropriate for investigating the respective research question (or even impossible). Accordingly, this number seems actually quite high, especially considering that user studies are still rarely used in broader recommender research [3].

Taking a closer look at the procedure of the user studies however sheds a bit different light: As far as we were able to grasp the details from the papers, it was possible only in 44 % of the reported user studies to actually consume products (i.e. in 30 % of all papers; see Figure 1). Admittedly, in some papers, this would have made no sense or consumption would have been unrealistic (e.g. hotel or date recommendations). Sometimes, it simply was not necessary for answering the underlying research question. We decided not to count in consumption of movie trailers [2] or song excerpts [7], but included cases where, for instance, recommended research papers were only accessible via a link to an external website [10], making it less likely that many participants took that chance. In summary, with the smaller number of user studies presented at

ImpactRS '19, September 19, 2019, Copenhagen, Denmark

[1] Website of this year's edition: https://intrs19.wordpress.com/

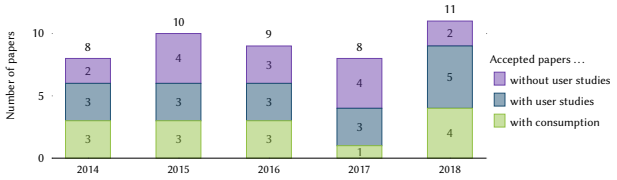

Figure 1: Results from our literature review showing how many papers were accepted to past IntRS workshop editions, how many of these papers contained user studies, and in how many studies item consumption was possible.

less user-centric venues in mind that likely allowed consuming items in even fewer cases, the question arises whether evaluation results would have been overall the same if item consumption had been possible. While our literature review is indeed limited, the impact of RS has most likely not always been measured accurately since participants might not have had everything they needed to adequately assess recommendation quality and user experience.

3 ASPECTS TO INVESTIGATE

With the importance of item consumption in mind, the literature review points out possible omissions in past research, emphasizing the need to take this aspect more into account when designing future user experiments. For doing so, a number of research questions still need to be answered. This may help to decide in a more structured manner, for example, whether it is necessary to provide participants with the possibility to consume items at all, or which substitutes may be used otherwise. It may also indicate which factors that might confound the assessment and thus lead to a distorted impression of the recommender's impact need to be considered—for planning the study, analyzing results and drawing conclusions.

The following (non-exhaustive) list contains aspects we think are generally important and possibly mediate the effects of item consumption. Concretely, we suggest investigating the influence:

- of item consumption also in other domains, depending on domain knowledge of participants as well as *product* type and attributes (e.g. search vs. experience products),
- of presenting different kinds of *information* (subjective vs. objective item descriptions) as possible substitutes for item consumption at varying level of detail (only metadata or additional content descriptions, other item-related information such as user reviews, system-generated explanations, paper abstracts, song excerpts, movie trailers, etc.),
- of *user* characteristics such as personality or decision-making style (making decisions in either a rational or intuitive way might affect the need for actual item consumption),
- and of the point in *time* assessments take place (since the effect of item consumption might diminish over time).

Beyond that, there are certainly many other aspects that may influence study results when trying to quantify the impact of RS. For instance, the improvements made regarding user experience in the past couple of years led to higher perceived recommendation quality without any changes to recommendations [3]. However, apart from these attempts intended to positively affect the impact of RS, some aspects may unintentionally cause differences due to the specific characteristics of user experiments. First, the experimental situation itself (e.g. presence of supervisor, lab study), with systems specifically designed for the purpose of the study, thus also limited to this purpose, might affect ecological validity: The assessment might be different compared to when a recommendation set is integrated into a real-world e-commerce platform. Among others, economic reasons (real money needs to be spent) or different relationships between students and researchers vs. customers and commercial system providers might affect e.g. reported purchase intention or perceived trustworthiness. Also, questionnaires might interfere with internal validity as item formulations can be ambiguous, e.g. regarding whether recommended products are novel (recently released vs. only new to the participant) or the set appears well-chosen (because products fit together or actually represent the participant's taste). More generally, using such instruments at all might be an issue as they might provoke a more conscious assessment, possibly affecting decision-making (i.e. participants could settle for different items if not confronted with a questionnaire).

4 CONCLUSIONS AND PERSPECTIVES

We have now positioned our work on the effects of item consumption [6] in context of the broader question of how the impact of RS can adequately be measured by means of user studies in academia. We identified a number of aspects that still need to be investigated in order to pursue the superordinate goal of deriving a set of guidelines for promoting validity of future experiments and fostering reproducibility. Currently, we are planning a study to investigate the influence of the aspects listed in the previous section. In addition, we would like to address the questions beyond and encourage others to do so as well, possibly also by employing unprecedented means for assessing the impact of RS. For instance, developing methods that use eye-tracking to determine which items are of most interest for participants might help to avoid interventions and make them switch decision-making styles. Overall, the insights that may be gained could also have broader impact, for example, by finding solutions for algorithms to adequately deal both with ratings provided in real-world systems without previously experiencing the products (e.g. "This recipe sounds awesome" → 5-star rating) and ratings resulting from actual consumption.

REFERENCES

[1] M. Bilgic and R. J. Mooney. 2005. Explaining recommendations: Satisfaction vs. promotion. In *Proc. Beyond Personalization Workshop*.
[2] M. P. Graus and M. C. Willemsen. 2016. Can trailers help to alleviate popularity bias in choice-based preference elicitation?. In *Proc. IntRS '16*. 22–27.
[3] B. P. Knijnenburg and M. C. Willemsen. 2015. *Recommender Systems Handbook*. Springer US, Chapter Evaluating recommender systems with user experiments, 309–352.
[4] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, and C. Newell. 2012. Explaining the user experience of recommender systems. *User Model. User-Adap.* 22, 4-5 (2012), 441–504.
[5] B. P. Knijnenburg, M. C. Willemsen, and A. Kobsa. 2011. A pragmatic procedure to support the user-centric evaluation of recommender systems. In *Proc. RecSys '11*. ACM, New York, NY, USA, 321–324.
[6] B. Loepp, T. Donkers, T. Kleemann, and J. Ziegler. 2018. Impact of item consumption on assessment of recommendations in user studies. In *Proc. RecSys '18*. ACM, New York, NY, USA, 49–53.
[7] F. Lu and N. Tintarev. 2018. A diversity adjusting strategy with personality for music recommendation. In *Proc. IntRS '18*. 7–14.
[8] P. Pu, L. Chen, and R. Hu. 2011. A user-centric evaluation framework for recommender systems. In *Proc. RecSys '11*. ACM, New York, NY, USA, 157–164.
[9] N. Tintarev and J. Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Model. User-Adap.* 22, 4-5 (2012), 399–439.
[10] K. Verbert, D. Parra, and P. Brusilovsky. 2014. The effect of different set-based visualizations on user exploration of recommendations. In *Proc. IntRS '14*. 37–44.