Item Familiarity Effects in User-Centric Evaluations of Recommender Systems

Dietmar Jannach
TU Dortmund, Germany
dietmar.jannach@tu-dortmund.de

Lukas Lerche
TU Dortmund, Germany
lukas.lerche@tu-dortmund.de

Michael Jugovac
TU Dortmund, Germany
michael.jugovac@tu-dortmund.de

ABSTRACT

Laboratory studies are a common way of comparing recommendation approaches with respect to different quality dimensions that might be relevant for real users. One typical experimental setup is to first present the participants with recommendation lists that were created with different algorithms and then ask the participants to assess these recommendations individually or to compare two item lists. The cognitive effort required by the participants for the evaluation of item recommendations in such settings depends on whether or not they already know the (features of the) recommended items. Furthermore, lists containing popular and broadly known items are correspondingly easier to evaluate.

In this paper we report the results of a user study in which participants recruited on a crowdsourcing platform assessed system-provided recommendations in a between-subjects experimental design. The results surprisingly showed that *users found non-personalized recommendations of popular items the best match for their preferences.* An analysis revealed a measurable correlation between item familiarity and user acceptance. Overall, the observations indicate that item familiarity can be a potential confounding factor in such studies and should be considered in experimental designs.

Keywords

Recommender Systems; User-centric Evaluation; Bias

1. INTRODUCTION

Studies with users in a controlled environment are a powerful means to assess qualities of a recommendation system which can often not be evaluated in offline experimental designs. A common setup in the research literature is that the participants of an experiment use a software tool that implements two or more variations of a certain recommendation functionality. After interacting with the system, the participants are asked to explicitly evaluate certain aspects of the system, including, e.g., the suitability or the perceived diversity of the recommendations or other aspects like the value of system-provided explanations [1, 2, 3, 5].

In the recent studies presented in [2] and [5], the subjects were asked to assess the presented movie recommendations in dimensions such as diversity, novelty or perceived accuracy and the participants had to either evaluate lists of recommended movies individually or make side-by-side comparisons. One typical problem in such setups is that the recommendation *lists contain both movies that the users already know and movies unknown to the participants.* In the second case, additional information about the movies is often provided [3, 5] and users have to make their assessment based on plot summaries or movie trailers. This situation may in turn lead to two possible effects. First, in case unknown movies are displayed, the cognitive load for the participants to assess, e.g., the suitability of the recommended movies, is higher, which can result in a reduced overall satisfaction with the system. Second, an assessment like "Would I enjoy this movie?" based only on the meta-information or a trailer could be an unreliable predictor of the assessment of a movie after a participant has actually watched it. To our knowledge, how item familiarity can impact the users' perception of a recommendation system in different dimensions has not been discussed explicitly in the literature before. The study in [5] does not consider item familiarity as a factor; the authors of [2] cover item familiarity in their "novelty" construct but base it on the self-reported familiarity with the recommendation list as a whole and do not explicitly ask users to indicate if they know the individual movies.

2. EXPERIMENT

We conducted a user study[1] in the style of [2] and [5]. The participants were first asked to rate a set of movies known to them using a specifically designed web application based on MovieLens data. In the second step, they were presented with movie recommendations created with five different algorithms including Matrix Factorization (Funk-SVD), Bayesian Personalized Ranking (BPR), SlopeOne, a content-based technique (CB), and a non-personalized popularity-based baseline (PopRank). The participants had to rate the presented movies individually (based on meta-information) and furthermore assessed the lists as a whole regarding factors like diversity, transparency, or surprise. For each presented movie, the users had to state if they already knew the movie or not. The participants were recruited via Mechanical Turk. From the 175 "Turkers" we filtered unreliable ones through different automated and comparably strict measures. At the end 96 participants (about 20 per treatment) were considered as being reliable.

3. OBSERVATIONS

Accuracy. Fig. 1(a) shows how the participants answered the question how well the presented list of movies

[1] Details are described in [4].

RecSys 2015 Poster Proceedings, September 16-20, 2015, Austria, Vienna.

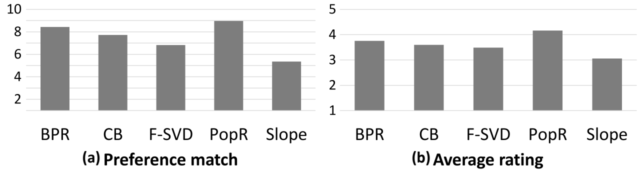

Figure 1: Self-reported preference match and avg. ratings.

as a whole matched their preferences; Fig. 1(b) displays the average rating assigned to the recommended movies.

To some surprise, the *popularity-based method PopRank is perceived by the users as the most accurate method*, followed by the BPR technique, which has a comparably strong bias to recommend popular items to everyone. Movie recommendations that contained only blockbusters – about 94% of the recommendations made by these two methods were known to the users – were considered the best preference match for the participants (significant at $p < 0.05$).

In comparison, recommendation lists that were actually personalized and contained various niche items[2] received lower scores. *The preference match is inversely related to the number of items known to the user.* Funk-SVD users knew about 50% of the items and SlopeOne users even less.

We contrasted these findings with an offline accuracy analysis of the underlying MovieLens dataset, which led to the expected superiority of the Funk-SVD method in terms of precision and the RMSE. We then computed the accuracy of the different algorithms for those recommended items that were rated by the participants. We took individual measurements for the set of known and unknown movies for the algorithms which did not only contain popular movies (Fig. 1). The "offline" accuracy measurement – except for SlopeOne – shows to be a comparably good predictor for the movies that the users already knew ("*seen*"). When applied to the unseen movies, the predictions made by the algorithms largely deviate from the user's ratings[3].

Table 1: Offline accuracy analysis (RMSE).

RMSE	CB	Funk-SVD	SlopeOne
MovieLens (offline)	1.92	1.65	1.72
Survey, all	2.19	3.46	4.03
Survey, only *not seen*	2.90	4.55	4.45
Survey, only *seen*	1.87	1.99	2.59
% of *seen* movies	0.74	0.52	0.27

Diversity, Surprise, Transparency. Fig. 2 shows the averaged questionnaire answers regarding perceived diversity, surprise and transparency. Again, we see unexpected results, in particular that *the content-based (CB) recommendations were perceived to be diverse.* When measuring the inverse Intra-List-Similarity (ILS) of the recommendations using TF-IDF vectors of the movie descriptions, the CB method as expected led to the lowest diversity, *which raises the question if the ILS measure is a suitable proxy for perceived diversity.* The surprise factor for the popularity-biased methods was low, as expected. Finally, users felt that they could understand the logic of the recommendations ("transparency") in particular when popular items were presented (both in a non-personalized and personalized way).

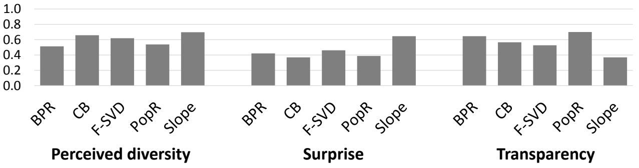

Figure 2: Perceived Diversity, Surprise and Transparency.

User Acceptance. Figure 3 finally reports the average answers regarding user acceptance in terms of ease-of-use, intention-to-reuse, and intention to recommend the system to a friend. "Ease of use" is generally high, but users had more trouble using the system (assigning ratings) when unfamiliar movies were presented. The other two satisfaction indicators in Fig. 3 are correlated with the assessment of the preference match shown in Fig. 1.

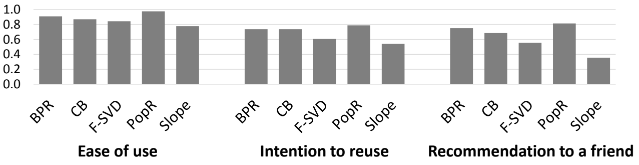

Figure 3: User Acceptance Results.

4. DISCUSSION

Recommending popular and familiar items has shown to be a well-suited strategy in this user study to achieve high satisfaction with the system and the presented recommendations, even though in practice recommending only popular items is typically of limited value.

Our preliminary study – experiments with more participants, non-Turkers, and a more specific questionnaire focusing on item familiarity are still required – suggests that *item familiarity can be a possible confounding factor in user studies.* Specifically, lab experiments in which users are asked to assess items unknown to them might have limited predictive power with respect to the true usability of the tested system.

5. REFERENCES

[1] P. Cremonesi, F. Garzotto, and R. Turrin. User-centric vs. system-centric evaluation of recommender systems. In *Proc. INTERACT 2013*, pages 334–351, 2013.

[2] M. D. Ekstrand, F. M. Harper, M. C. Willemsen, and J. A. Konstan. User perception of differences in recommender algorithms. In *Proc. RecSys '14*, pages 161–168, 2014.

[3] F. Gedikli, D. Jannach, and M. Ge. How should I explain? A comparison of different explanation types for recommender systems. *Int. J. Hum.-Comput. Stud.*, 72(4):367–382, 2014.

[4] D. Jannach, L. Lerche, and M. Jugovac. Item familiarity as a possible confounding factor in user-centric recommender systems evaluation. *i-com Journal for Interactive Media*, 14(1):29–40, 2015.

[5] A. Said, B. Fields, B. J. Jain, and S. Albayrak. User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm. In *Proc. CSCW '13*, pages 1399–1408, 2013.

[2] In [2], unpopular items recommended by Funk-SVD were filtered.

[3] The absolute RMSE values are comparably high as the participants only had to rate 15 items in the first phase. The observations for precision are comparable; RMSE values for PopRank and BPR are missing as these methods generate no rating predictions.