Offline Performance vs. Subjective Quality Experience: A Case Study in Video Game Recommendation

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

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

ABSTRACT

Research in the field of recommender systems is largely based on offline experimentation on historical datasets. Several recent works however suggest that models optimized for accuracy measures are not necessarily those that lead to the best user experience or perceived system utility. In this work we first determine the offline performance of different algorithms in the domain of video game recommendation and then investigate the perceived recommendation quality through a user study. The offline results show that learning-to-rank methods optimized for implicit feedback situations as expected perform best in terms of accuracy, where higher accuracy often comes with a stronger tendency of the algorithms to recommend mostly popular items. In the user study, however, methods that also consider the similarity between items in their algorithms perform at least equally well in terms of accuracy, which could not be expected from the offline experiment. Such content-enhanced methods were also slightly favored by users in terms of perceived transparency.

CCS Concepts

•**Information systems** → **Recommender systems;**

Keywords

Recommender Systems; Offline Performance; User Experience; Case Study; Video Games

1. INTRODUCTION

Significant advances were made in recent years in terms of improving the capability of recommendation algorithms to predict the relevance of items for individual users or to rank items according to the estimated preferences of the users. The predominant research approach in the field is based on offline experimentation on historical data, where some of the known user preferences are held out and the goal is to estimate the hidden preferences. To assess the quality of different algorithms, various accuracy measures like RMSE or Precision and Recall are typically used in the field.

This research approach has a number of advantages as it does for example not depend on a particular application context or domain-specific considerations and enables the comparison of different research results [6]. This abstraction from the problem domain and the use of problem-independent performance measures on the other hand raise the question if algorithms that were optimized for prediction accuracy in an offline process are effective when deployed in practice.

In fact, a number of works in recent years suggest that higher offline prediction accuracy does not necessarily equate to online success in real-world use or the highest quality perception in laboratory studies [1, 4, 7, 11]. In addition to the general relevance of an item for a user, other quality aspects might be equally relevant, including the user's current contextual situation, the recency or general popularity of the presented items, the diversity of the list, the user's familiarity with the items, etc. Finding measures that predict online success through offline experimentation is in fact one of the main challenges mentioned in [5] in the context of the Netflix recommendation system.

With this work we aim to contribute additional evidence that it can be misleading to choose a recommendation algorithm for a production environment only based on offline accuracy. We demonstrate this in the context of recommending video games. Specifically we use data from the popular "Steam" game platform[1] to benchmark different algorithms in an offline setting and in addition assessed the quality perception of users through a user study. In contrast to some previous works like [1], which followed similar goals, the user preference information in our scenario is given solely in terms of implicit feedback (game ownership). We correspondingly include learning-to-rank algorithms based on Bayesian Personalized Ranking (BPR) [16] in our experiments that are optimized for such problem settings. Since algorithms of this type can have a tendency to focus mostly on popular items – which are in turn often familiar to users [10] – we expected that these algorithms would also perform well in the user study. Interestingly, however, the best performing approaches according to the offline study, BPR++ [12] and GBPR [13], were not perceived as being better in terms of accuracy in the user study.

Next, in Section 2, we will describe the used recommendation dataset and present the results of the offline analysis. Section 3 then contains the observations of the user study.

[1]http://store.steampowered.com

SAC 2017, April 03 - 07, 2017, Marrakech, Morocco

ISBN 978-1-4503-4486-9/17/04…$15.00

DOI: http://dx.doi.org/10.1145/3019612.3019758

2. OFFLINE EVALUATION

2.1 Data

Steam is a popular digital marketplace for video games and includes a social network component for players to discover, discuss, and share content. We created a dataset for recommendation purposes by using their public API and to a smaller extent by scraping additional information from the website. Specifically, the core of our dataset consists of information about which games each user owns and has played. Among various other types of information, we know each games' release date, genre and set of tags attached to it from a predefined collection of 314 tags.

For our evaluations we randomly selected a number of active users from Steam's social network. We only considered users that (a) have played games for at least 10 hours, (b) played 3 games at least for 2 hours, and (c) played another 3 games at least for 1 hour. In our view, these somewhat arbitrarily selected conditions ensure that the users can be considered as active. The dataset comprises 5,184 users who on average own 57 games, of which they played 37 at least once. At the time of data collection, 7,776 games[2] were available. Around 60% of these games were never purchased by any user in our dataset. We selected those 1,845 games for the subset that were owned by at least 20 users.

2.2 Algorithms

In our offline experiments we evaluated the predictive accuracy of several recommender algorithms, including learning-to-rank approaches and content-based strategies that use the metadata included in our dataset. The algorithms were implemented on top of the Recommender101 framework [8] and the algorithm parameters were manually fine-tuned.

PopRank. This non-personalized baseline method recommends the best-selling games on the platform to every user.

BPR++. This extension of the BPR (Bayesian Personalized Ranking) algorithm [16] can handle graded implicit feedback signals [12]. In addition to considering if a game is owned by a user or not, our BPR++ configuration takes the similarity of a game's tags to the tags of all games owned by a user into account, which leads to a higher accuracy than BPR alone.[3] Since BPR and consequently also BPR++ have some known tendency to recommend popular items, we implemented the sampling strategy in the learning process from [10] to counteract the popularity bias without a loss of accuracy.

GBPR. Group Bayesian Personalized Ranking [13] is another extension to BPR [16], which takes implicitly created user groups into account that are formed based on co-purchases. [4]

SimilarTags. This content-based recommender considers the tags assigned to each game as binary vectors (with 314 dimensions) to calculate the distance to user profiles. A user profile is represented as the arithmetic mean of all tag vectors of the games a user owns. As a distance metric, the cosine similarity is used and the games with the shortest distance to the user profile are recommended.

SimilarGames. This strategy simulates one of the recommenders on the Steam website. As we have no access to the individual recommendations of each Steam user, we approximate these recommendations using the "more like this" list on the product pages and create personalized recommendation lists for a user u by scoring each recommendable item i as follows. Let $L(j)$ be the list of n items that are similar to an item j according to the Steam platform. Furthermore, let $pos_i(L(j))$ return the position of item i in such a list of similar games. $pos_i(L(j))$ returns 0 if i is not in $L(j)$.

In our calculation scheme, we look at each item j in the user's set of owned games I_u and compute an aggregate score for each recommendable item i as follows:

$$score_{u,i} = \sum_{j \in I_u} (n + 1 - pos_i(L(j)))$$

As a result, those items receive a high score that appear often at the top of Steam-provided similar-games lists of the games that a user already owns.

Hybrid: SimilarTags & GBPR. A hybrid strategy that combines the normalized scores of each method with equal weights. The recommendations are by design equally influenced by the characteristics of both techniques, i.e., the weight parameter is not optimized for some metric.

Rating Prediction Methods: In addition to these *ranking* algorithms, we included different *rating prediction* algorithms in our experiments, namely Funk's matrix factorization (MF) method [14], an item-based k-nearest neighbors method with the Jaccard index as the similarity metric, and a method based on Factorization Machines (FM) [15] using Alternating Least Squares (ALS) optimization. Again, all algorithm parameters were fine-tuned for optimal accuracy.

2.3 Evaluation Protocol and Measures

The task of the recommender algorithms is to predict game ownership (purchases). As usual, the given 0/1 matrix of game ownerships is split into training set (80%) and test set (20%). The data was split on a per-user basis and a five-fold cross-validation procedure is applied.

To measure accuracy and ranking performance, we choose to report Precision, Recall, and the Mean Reciprocal Rank using recommendation lists of size 10, though other metrics like the NDCG would be applicable as well. We furthermore check if certain algorithms have a bias toward popular items by reporting the average popularity of the recommended items. Additionally, we analyze the catalog coverage of each algorithm by counting the overall number of different items that appear in the top-10 lists of all users and by reporting the Gini index[5], see [10]. We also show the diversity as the average number of different tags assigned to games in a recommendation list. Finally, as a domain-specific aspect we report the average release date of the recommended games, to see if some algorithms tend to focus on more recent games.

2.4 Results

The offline evaluation results are given in Table 1. In each column, the "best" result is highlighted with bold font if it is significantly higher/lower than the other values.[6]

[2]We excluded additionally downloadable game content, trial versions, and other available software and media content.

[3]We also evaluated other possible graded feedback signals, which did however not lead to better accuracy.

[4]We also tried to use the explicit friend lists from the platform as groups, which did however not lead to better results.

[5]The Gini coefficient ranges from 0 (equal distribution of items) to 1 (concentration on one single item).

[6]To determine statistical significance, we use Student's t-test with $p < 0.01$ and Bonferroni correction.

Table 1: Results of the offline analysis sorted by Precision.

Algorithm Short Name	Precision	Recall	MRR	Popularity	Coverage	Gini	Diversity	Release
GBPR	**0.206**	**0.321**	**0.084**	974	865	0.937	77	May '11
Hybrid: SimilarTags & GBPR	0.189	0.294	0.077	735	870	0.922	65	Jul '11
BPR++	0.166	0.288	0.069	697	**1,522**	**0.835**	71	Sep '11
PopRank	0.104	0.163	0.043	**1,351**	41	0.989	77	Sep '09
SimilarGames	0.092	0.147	0.035	346	1,010	0.875	66	Jan '13
Factorization Machines	0.076	0.140	0.033	1,045	31	0.989	70	Jun '11
SimilarTags	0.045	0.076	0.016	201	826	0.926	49	Oct '12
Funk-SVD	0.010	0.023	0.004	187	18	0.989	68	**Nov '13**
Item-kNN	0.009	0.012	0.002	146	204	0.988	**85**	Jul '11

2.4.1 *Accuracy Results*

Not surprisingly, the BPR extensions, which are designed for one-class collaborative filtering problems and optimize a ranking criterion, led to the best results in our comparison. The highest values for all accuracy measures were obtained with the group-based GBPR method, but the BPR++ method was also consistently better than the popularity-based baseline. It was also better than the "pure" BPR method, which we do not explicitly list in this comparison.

The SimilarGames method, which is based on and simulates one recommendation strategy of the Steam platform, works fairly well, but is not better than recommending popular items to everyone. Although the Item-KNN shares some similarities with the SimilarGames method, the latter performs much better, which is likely caused by the additional domain knowledge about the item neighbors extracted from the Steam platform.

Clearly, recommending only the most popular items to everyone might be of limited value for users in reality, but in such offline experiments this strategy can represent a comparably strong baseline [2, 10]. The content-based SimilarTags method leads to mediocre results in this comparison. Except for the Factorization Machines technique, none of the rating prediction algorithms led to competitive accuracy results in this setup. Finally, combining the GBPR scores with the content-based SimilarTags strategy did not lead to a further accuracy improvement.

2.4.2 *Analysis of Popularity Biases*

The tendency of algorithms to mostly recommend popular items can be undesired in some domains, because it can have an impact on the user's perception of the recommendation quality, see [9]. Users might already know the most popular items recommended to them. This in turn means that the recommendation system does not help users to find relevant new items and therefore, from the provider's viewpoint, does not help to stimulate additional sales.

The PopRank technique by design only recommends the most popular items to everyone, leading to an "average-item-popularity" score of 1,351, i.e., each recommended item was purchased on average by 1,351 users of our dataset. Looking at this aspect for the different BPR variants and the hybrid, we see that recommending more popular items seems to translate into higher accuracy. This means, while GBPR recommends with the highest accuracy, it also focuses strongly on comparably popular items. The recommendations of BPR++ are less popular due to the implemented popularity-accuracy trade-off sampling scheme.

The content-based recommender as well as the Item-kNN and Funk-SVD techniques do not exhibit such strong biases, although Funk-SVD seems to recommend many niche items, which was also observed in [3]. An exception is again the Factorization Machine approach, which achieves comparably good accuracy values, but at the cost of a stronger bias. This observation corroborates, e.g., the results from [10].

2.4.3 *Analysis of Concentration Biases*

The concentration bias is reported here as an algorithm's tendency to focus its top-n recommendations on a small set of the catalog items. Again, this can be undesired in some domains as it might, for example, lead to "rich-get-richer" effects and limited discovery support for users.

The results show that BPR++ includes the largest number of different games in the recommendations and that the other BPR variants, while still covering a good part of the item spectrum, focus on significantly smaller sets of items. In terms of the tags of the games, it is less diverse than GBPR, but more diverse than the GBPR & SimilarGames hybrid. The SimilarGames strategy, which mimics the behavior of the Steam recommender, and the content-based SimilarTags strategy also recommend a variety of different items while being less diverse than the learning-to-rank strategies. Finally, the rating prediction techniques often use small sets of items which they recommend to everyone.

The Gini index summarizes the concentration bias in a different way, and we see that not only the rating prediction algorithms focus on a small set of items but also the GBPR method and the hybrid. Although GBPR recommends a number of different items at least once for some users, the majority of the recommendations seems to consist of the same items, which is captured by a high Gini value.

2.4.4 *Release Dates*

Finally, we report the average year of release for the recommended games in Table 1 as the freshness of a recommended game might impact the user's quality perception. PopRank on average recommends the oldest games, which can be expected as often sold games generally have been available for a long time. The average release year of many other algorithms is mainly around 2011. Exceptions are the SimilarGames method, which seems to focus slightly on newer releases, and Funk-SVD, which has an even stronger tendency to recommended newer games. The algorithm focuses on a tiny set of games and for this dataset these coincidentally happen to include a number of little-known "indie" games from 2015 and 2016.

3. USER STUDY

3.1 Goals and Study Design

While the first part of our analysis showed that the recent GBPR method worked best in terms of typical accuracy measures, the goal of the subsequently described user study was to test if higher offline accuracy translates into better user perception.

3.1.1 Study Design

The study consisted of three phases and was conducted online using a web application that was created for the purpose.

1. Preference Acquisition. First, we collected the user's preferences for video games. Users could either specify a list of at least six games that they own and like, or they could provide their Steam account ID and we would then retrieve their owned games through the public Steam API. The large majority of the participants used this second option.

2. Individual Evaluation of Recommendations. The participants were then provided with four different recommendation lists in a within-subjects design, where each of the lists was created with a different algorithm. The order of the algorithm was randomized across participants. To not overwhelm the users, each list contained six games. We furthermore provided a short description and a link to the Steam store for the game. For each recommendation list the users had to individually rate every recommended item, and indicate (a) if they knew the game before the study and (b) if they have already played it (see Figure 1). For each list, the participants had to answer four questions relating to their quality perception on a 7-point Likert scale. To keep the cognitive load for the users low, we decided to use direct questions to assess the user's quality perception in terms of the perceived accuracy, diversity, novelty, and transparency of the recommendation process.

3. Ranking Task and General Profile. After the participants had evaluated each recommendation list individually, they were asked to rank the recommendation lists based on their subjective quality experience, i.e., how much they liked each list. We also asked participants questions about their age, if playing video games is one of their hobbies and if they play video games from different genres.

3.1.2 Algorithm Selection

The four algorithms included in our comparison were the GBPR method, the SimilarTags & GBPR hybrid, BPR++, and SimilarGames, as they had distinctive characteristics according to our offline analysis.

Since the BPR-based methods would require retraining the model when a new participant enters his or her preferences, we implemented a neighborhood-based approach to generate recommendations in real-time. Given the preferences of a new participant, we first identified the single existing Steam user who was most similar to the new user based on the Jaccard index of owned video games. Then we used the recommendations of this existing user to create the ranked item lists. To ensure that the characteristics of the proxy-based recommendations are representative of those that the user would receive based on his or her own preferences, we repeated the offline analysis with this proxy-based scheme. The detailed results, which we omit here for space reasons, showed that although the absolute accuracy values decrease when using the proxy-based method, we can still observe the same relative ranking of the algorithms as well as similar popularity and concentration biases.

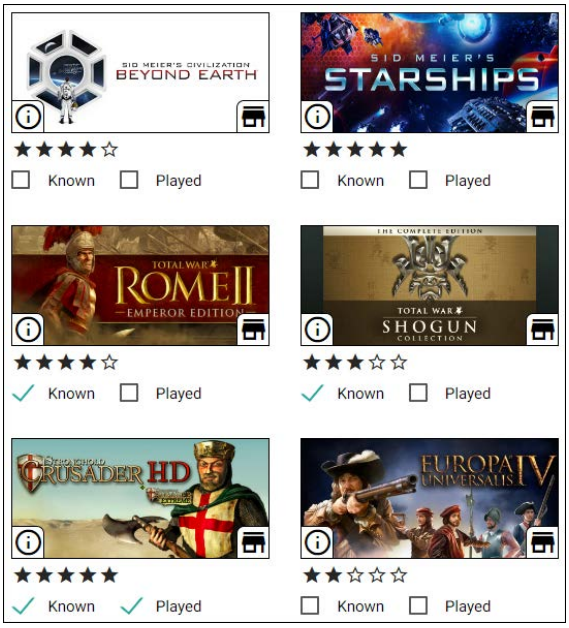

Figure 1: Video Game Rating Interface.

3.1.3 Participants

We recruited 158 participants by posting invitations on different game-related and general message boards and by inviting participants of university classes. Most of the participants were between 20 and 29 years old, about two thirds of them were German-speaking. Most of the participants expressed strong interest in playing games from different genres. Also, more than half of them stated that playing video games is one of their main hobbies. 104 of the participants completed the entire study.

3.2 Observations

For the subsequent evaluation we used the feedback of 97 users, because for seven of the participants the proxy user from our dataset was too different from the user profile of the respective participant. This happens when the set of owned games of a user only includes niche items which we filtered out when creating our dataset.

Figure 2 (left) shows how the participants answered the questions regarding the different quality dimensions accuracy, novelty, diversity, and transparency.

Accuracy / Preference Match: When asked if the presented recommendations "match their preferences", no statistically significant difference between the algorithms can be observed. The average answer scores on the 7-point scale were somewhere between 3.5 and 4. The slight superiority of GBPR from the offline experiments was therefore not replicated in the user study. In fact, the SimilarGames method performed equally well as the BPR-variants.

Novelty / Item Discovery: The general novelty level for all algorithms was mediocre and the lowest perceived novelty was observed for the GBPR method. The highest but still not very strong novelty level was reported for

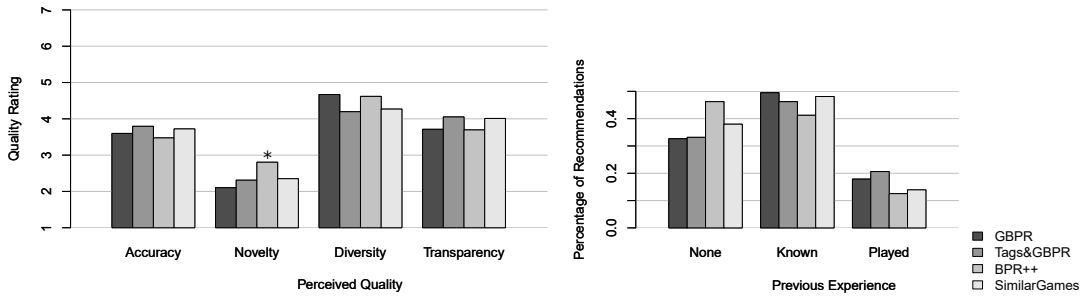

Figure 2: Left chart: perceived quality by the 97 participants on a scale from 1 (worst) to 7 (best). Right chart: previous experience with the games recommended by each algorithm to the 97 participants.

BPR++. It is significantly higher than the other three methods with $p < 0.05$. Among the methods with high accuracy, BPR++ also exhibited the lowest popularity bias in our offline analysis.

The SimilarGames strategy did not lead to a high novelty perception, even though it recommends less popular items than BPR++ according to our offline analysis. This indicates that item popularity in terms of item ownership (past purchases) does not fully cover aspects of novelty. In our user study, we had differentiated for each item if a user *knows* a recommended item or if he or she *owns* it. In Figure 2 (right), we show the results for these categories for the different algorithms. Looking in particular at the results for the SimilarGames strategy, we observe that many participants actually knew quite a lot of the recommended items, even though they were not the most popular ones in terms of sales in this community. These observations explain the limited novelty of the SimilarGames strategy.

A side result of this analysis is that GBPR "wins" in terms of recommending popular items that people know, but adding the content-based SimilarTags component to GBPR is even slightly better in terms of predicting game ownership in this comparison. This corroborates the finding of [7] obtained in a field study on mobile game recommendation that many users tend to purchase multiple games of similar genres.

Perceived Diversity The two methods that use similarity information to generate the recommendations (SimilarTags & GBPR, SimilarGames) not surprisingly led to the lowest perceived diversity, which corresponds to the results of our offline evaluation, see Table 1. For our hybrid method this means that adding a content-based component leads to a significantly higher level of perceived diversity compared to GBPR ($p < 0.05$), but not a decrease in perceived accuracy, which one could suspect when looking at the offline experiment alone.

Transparency When asked to what extent the participants believed that they understood *why* certain items were recommended, they were slightly more sure for the two methods based on item similarity. This means that they probably noticed that the recommended games were of similar genres like the games that they played in the past. However, the differences are not statistically significant.

Algorithm Ranking Finally, we determined the "most liked" recommendations in the user study by merging the algorithm rankings provided by the participants with the Borda count metric[7], a common scheme to determine the outcome of a ranking vote. In contrast to the accuracy results of the offline evaluation, the SimilarGames strategy led to the best overall quality perception in this study with a score of 255. It is ranked significantly higher than BPR++ (224) according to Welch's t-test with $p < 0.05$. However, SimilarGames is not ranked significantly higher than GBPR (239) or SimilarTags & GBPR (252).

3.3 Discussion

Overall, our results indicate that the superior offline performance of GBPR did not translate to a better experience in this domain. In fact, the GBPR-based recommendations for some users did not match the user's interests at all. One participant, for example, commented that he only received online multiplayer recommendation, while his user profile contained mostly single-player games. The recommendations of the GBPR method were also considered to be less novel than those by BPR++. As reported in [9], recommending novel items can hurt the overall perception of the recommendations and here BPR++ was also ranked last.

The SimilarGames method, which has a tendency to recommend more recent releases and items that are similar to a user's past purchases, led to a better quality perception in the user study. This indicates that "freshness" and "familiarity" can also be relevant quality criteria in this domain. We plan to investigate these aspects in future works.

A side observation of our study is that measuring novelty by using the item popularity in terms of past purchases of an item as a proxy is only a rough estimate and can be insufficient in some domains.

4. PREVIOUS WORKS

Limited research exists in the academic literature that addresses the question to what extent algorithms that are optimized for high prediction accuracy on historical data lead to "online success". The reasons are not only that academic researchers rarely have access to real-world systems to run A/B test, but also that the definitions of online success can be diverse and the measurement is often done in a domain-specific way, e.g., in terms of customer retention rates [5].

However, a few works exist that compare these two experimental conditions in different application domains, e.g., [1, 4, 7, 11]. Similar to our work, these papers provide indications and evidence that higher offline accuracy does not

[7]Each recommender receives one point for being ranked last by a user, two for being ranked next-to-last, and so on.

necessarily mean that a certain algorithm also generates the most useful recommendations for its users.

The offline/online comparison of several recommendation algorithms in [1] for example showed that no algorithm performed significantly better or worse than any other in terms of global satisfaction, although in terms of offline accuracy there were significant differences. These results are therefore similar to those of our study. In addition, one of the techniques that worked best in the user study in [1] did not even rely on personalization.

In [4], an offline analysis and an A/B field test were used to compare different algorithms for news recommendations. In the offline setup, popularity-based recommendations yielded the best accuracy values. In the A/B test, however, a context-based technique generated the best business value in terms of click-through rates and visit durations. News recommendation is also the problem domain in [11]. Again, the authors report that the results of the offline experiments do not match the online performance of different algorithms on a large-scale news site and recommending content similar to past user interests worked comparably well. The authors of both works argue that creating "good" recommendations is very domain dependent. However, the algorithms that dominate offline experiments tend to recommend mostly popular items, as was also observed in our study.

The results of an A/B test of different mobile game recommendation strategies are reported in [7] and a content-based strategy also led to the largest increase in revenue. Collaborative filtering in contrast only led to high click-through rates and therefore, customer attention, but not to increased sales. In offline experiments that were conducted before the field test, this aspect could not be observed.

Finally, the work in [5] reviews the recommendation system used by the Netflix video-on-demand platform. The authors report that offline experiments help to find promising recommendation techniques that might work in practice. However, to determine the "best" approaches, all of them must be evaluated and fine-tuned online. Employing large-scale A/B tests remains essential, although the outcomes of such tests can also be influenced by external factors.

5. CONCLUSIONS

With this work we provide additional evidence that recommendations that are generated by algorithms optimized for "offline accuracy" are not necessarily the most relevant ones for users in deployed applications. Specifically in situations when implicit feedback signals are used, high accuracy can be obtained by algorithms that mainly recommend popular items, such as the BPR-based learning-to-rank strategies. This however leads to limited novelty for the user in reality and, depending on the domain, to a reduced business value of the recommendation system.

Offline experimentation will nonetheless remain to be an important instrument for researchers and practitioners to design new recommendation strategies, e.g., because of the high costs of user studies and field tests. Therefore, this calls for the design of additional evaluation methods and metrics for offline experimental setups in recommender systems. In particular, methods are required that are able to predict not only whether a user will like a recommended item, but if the recommendation is *useful*, e.g., in terms of item discovery.

6. ACKNOWLEDGEMENTS

We would like to thank Florian Will who contributed to this work with the results of his master's thesis.

7. REFERENCES

[1] P. Cremonesi, F. Garzotto, S. Negro, A. Papadopoulos, and R. Turrin. Comparative evaluation of recommender system quality. In *CHI EA'11*, pages 1927–1932, 2011.

[2] P. Cremonesi, Y. Koren, and R. Turrin. Performance of algorithms on top-n recommendation tasks. In *RecSys '10*, pages 39–46, 2010.

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

[4] F. Garcin, B. Faltings, O. Donatsch, A. Alazzawi, C. Bruttin, and A. Huber. Offline and online evaluation of news recommender systems at swissinfo.ch. In *RecSys '14*, pages 169–176, 2014.

[5] C. A. Gomez-Uribe and N. Hunt. The Netflix recommender system: Algorithms, business value, and innovation. *ACM TMIS*, 6(4):13:1–13:19, 2015.

[6] D. Jannach and G. Adomavicius. Recommendations with a purpose. In *RecSys '16*, 2016.

[7] D. Jannach and K. Hegelich. A case study on the effectiveness of recommendations in the mobile internet. In *RecSys '09*, pages 205–208, 2009.

[8] D. Jannach, L. Lerche, F. Gedikli, and G. Bonnin. What recommenders recommend - an analysis of accuracy, popularity, and sales diversity effects. In *Proc. UMAP '13*, pages 25–37, 2013.

[9] D. Jannach, L. Lerche, and M. Jugovac. Item familiarity effects in user-centric evaluations of recommender systems. In *RecSys '15 Posters*, 2015.

[10] D. Jannach, L. Lerche, I. Kamehkhosh, and M. Jugovac. What recommenders recommend - an analysis of recommendation biases and possible countermeasures. *UMUAI*, 25(5):427–491, 2015.

[11] E. Kirshenbaum, G. Forman, and M. Dugan. A live comparison of methods for personalized article recommendation at Forbes.com. In *ECML/PKDD '12*, pages 51–66, 2012.

[12] L. Lerche and D. Jannach. Using graded implicit feedback for bayesian personalized ranking. In *RecSys '14*, pages 353–356, 2014.

[13] W. Pan and L. Chen. GBPR: group preference based bayesian personalized ranking for one-class collaborative filtering. In *IJCAI '13*, pages 2691–2697, 2013.

[14] S. Funk. (pen name). Try this at home. http://sifter.org/˜simon/journal/20061211.html, 2006.

[15] S. Rendle. Factorization machines with libFM. *ACM Transactions on Intelligent Systems Technology*, 3(3):57:1–57:22, 2012.

[16] S. Rendle, C. Freudenthaler, Z. Gantner, and L. Schmidt-Thieme. BPR: Bayesian personalized ranking from implicit feedback. In *UAI '09*, pages 452–461, 2009.